



T.C.

SELÇUK ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

VERİ MADENCİLİĞİ UYGULAMALARI İÇİN VERİ İNDİRGEME ALGORİTMALARININ GELİŞTİRİLMESİ VE RESİM MADENCİLİĞİNE UYGULANMASI

Onur İNAN

DOKTORA TEZİ

Bilgisayar Mühendisliği Anabilimdalı

Haziran-2015
KONYA

TEZ BİLDİRİMİ

Bu tezdeki bütün bilgilerin etik davranış ve akademik kurallar çerçevesinde elde edildiğini ve tez yazım kurallarına uygun olarak hazırlanan bu çalışmada bana ait olmayan her türlü ifade ve bilginin kaynağına eksiksiz atıf yapıldığını bildiririm.

DECLARATION PAGE

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Onur İNAN

Tarih:

ÖZET

DOKTORA TEZİ

VERİ MADENCİLİĞİ UYGULAMALARI İÇİN VERİ İNDİRGEME ALGORİTMALARININ GELİŞTİRİLMESİ VE RESİM MADENCİLİĞİNE UYGULANMASI

Onur İNAN

Selçuk Üniversitesi Fen Bilimleri Enstitüsü
Bilgisayar Mühendisliği Anabilim Dalı

Danışman: Doç.Dr. Nihat YILMAZ

2015, 164 Sayfa

Jüri
Danışmanın Doç.Dr. Nihat YILMAZ
Prof.Dr. Ahmet ARSLAN
Prof.Dr. Saadetdin HERDEM
Doç.Dr. Rıdvan SARAÇOĞLU
Yrd.Doç.Dr. Ali ÖZTÜRK

Teknolojinin ilerlemesi ve yaygınlaşmasıyla birlikte çok sayıda veri üretilmekte ve depolanmaktadır. Genelde kendi başlarına değersiz olan bu verilerin hızlı bir şekilde analiz edilerek anlamlı bilgilere dönüştürülmesi gerekir. Veriler büyüdükçe, verilerin tutarlılığı bozulmaya başlamış, içinde gürültü diye tabir edilen tutarsız ve yanlış veri miktarı artmaya başlamıştır. Bu gürültülü veriler, anlamlı bilgiye ulaşılmasını güçleştirmekte, bazı durumlarda imkânsız hale getirmektedir. Bu tez çalışmasında, özellikle medikal ve endüstriyel alandaki görüntü verileri üzerinde veri madenciliği ve sınıflama işlemlerinde kullanılacak verilerin analizinde karşılaşılan gürültülü, yanlış elde edilmiş ve sınıflanmış verilerin elenmesi, daha az nitelikle bilgiyi daha hızlı bir şekilde elde etmeyi amaçlayan sistemler geliştirilmiştir. Gürültülü verilerin elenmesinde, kümeleme algoritmalarından olan k-ortalamalar algoritması kullanılmış ve geliştirilmiştir. Sınıflama işlemlerinde güvenirliliği artırmak için kullanılan k-kat çaprazlama doğrulama algoritması tüm eğitme süreçlerine entegre edilmiştir. Niteliklerin azaltılması işleminde ise birliktelik algoritmalarından Apriori algoritması ve verinin en önemli karakteristik niteliklerini tespit eden Temel Bileşen Analizi (TBA) birleştirilerek oluşturulan hibrid bir sistem ile daha az nitelikle yararlı bilginin elde edilmesi sağlanmıştır. Geliştirilen bu sistemler literatürde yaygın olarak kullanılan veri kümeleriyle test edilmiş ve elde edilen yüksek başarı değerleri literatürdeki sonuçlarla karşılaştırılmıştır. Ayrıca endüstride yüksek hızlı termal transfer yazıcılarda kullanılan yazıcı başlıklarının termal hücre arızalarının miktarını özel tasarlanmış bir mikroskop sistemi ile elde edilen görüntülerinden tespit eden bir sistem geliştirilmiştir. Bu sistemin başarısının artırılması amacıyla gürültülü verilerin elenmesinde geliştirdiğimiz algoritmalar uygulanmış ve sistemin başarısının yükseldiği gözlenmiştir.

Anahtar Kelimeler: Birliktelik, Görüntü analizi, Gürültülü veri tespiti, Kümeleme, Nitelik azaltımı, Sınıflandırma, Temel Bileşen analizi, Veri madenciliği

ABSTRACT
Ph.D THESIS

DEVELOPMENT OF DATA REDUCTION ALGORITHMS FOR DATA MINING APPLICATIONS AND IMPLEMENTATION OF THESE ALGORITHMS TO IMAGE MINING

Onur İNAN

THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCE OF SELÇUK UNIVERSITY
THE DEGREE OF DOCTOR OF PHILOSOPHY
IN COMPUTER ENGINEERING

Advisor: Assoc.Prof.Dr. Nihat YILMAZ
2015, 164 Pages

Juri
Advisor Assoc.Prof.Dr. Nihat YILMAZ
Prof.Dr. Ahmet ARSLAN
Prof.Dr. Saadetdin HERDEM
Assoc.Prof.Dr. Rıdvan SARAÇOĞLU
Assistant Prof. Dr. Ali ÖZTÜRK

By the advancement and spread of technology, lots of data are produced and stored. These data, usually worthless by themselves, must be quickly analyzed and transformed into meaningful information. As the data expands, its consistency started to be deteriorated, and the amount of inconsistent and incorrect data inside it, which is called as "noise", began to increase. The noise makes it difficult and in some cases impossible to reach meaningful information. In this thesis study, in order to eliminate the noisy data that is untruly gotten and classified, which is met in the analysis of data used in data mining and classification especially on image data in the medical and industrial fields, it is developed the systems that aim at getting less information more quickly. In the elimination of noisy data, it is used and developed K-means algorithm, which is one of the clustering algorithms. K-fold cross-validation algorithm, which is used to increase reliability in the classification operations, is integrated in all training processes. In the process of data reduction, it is ensured to achieve less information by a hybrid system formed by the union of Apriority Algorithm, which is one of the integration algorithms, and Principal Component Analysis (PCA), which determines the most important characteristics of data. The systems developed are tested by the data clusters that are commonly used in the literature, and the high values of success obtained are compared to the other conclusions in the literature. In addition, it is developed a system that determines the amount of thermal cell failures of print heads, which are used in the high-speed thermal transfer printers in the industry, by the images that get via a special-designed microscope system. The success of the system is boosted by the method that we develop to eliminate the noisy data by the aim of strengthening the success of this system.

Keywords: Association rules, image analysis, noisy data detection, clustering, Attributes reduction, classification, Principal Component Analysis, Data Mining

ÖNSÖZ

Çalışmalarım esnasında desteğini ve yardımlarını esirgemeyen, bilgi ve deneyimlerinden yararlandığım başta danışmanım Doç. Dr. Nihat YILMAZ' a ve tez çalışmam sırasında tezin oluşmasındaki yönlendirmelerinden dolayı Tez izleme komitesi üyelerinden Prof. Dr. Ahmet ARSLAN ve Prof. Dr. Saadetdin HERDEM'e teşekkürlerimi sunarım. Ayrıca her türlü manevi desteği benden esirgemeyen, çok sevdiğim aileme ve eşime şükranlarımı sunarım.

Onur İNAN
KONYA-2015

İÇİNDEKİLER

ÖZET ..... **i**
ABSTRACT ..... **ii**
ÖNSÖZ ..... **iii**
İÇİNDEKİLER ..... **iv**
ÇİZELGELER ..... **viii**
ŞEKİLLER ..... **xi**
SİMGELER VE KISALTMALAR ..... **xiv**
1 GİRİŞ **1**
2 LİTERATÜR ARAŞTIRMASI **7**
3 VERİ MADENCİLİĞİ **15**
3.1 Veri nedir 15
3.2 Veri madenciliği ve bilgi keşfi aşamaları 15
3.3 Veri Madenciliği Teknikleri 17
3.3.1 Tanımlama ve Ayrımlama 18
3.3.2 Birliktelik analizi 18
3.3.3 Sınıflandırma ve Regresyon 19
3.3.4 Kümeleme analizi 19
3.3.5 Sıradışılık analizi 20
3.3.6 Evrimsel analiz 20
3.4 Veri Madenciliğinde Karşılaşılan Problemler 21
3.4.1 Veri Tabanının Boyutu 21
3.4.2 Gürültülü Veri 21
3.4.3 Boş Değerler 22
3.4.4 Eksik Veri 22
3.4.5 Artık Veri 23
3.4.6 Dinamik Veri 23
3.5 Birliktelik Kuralları 24
3.5.1 Birliktelik Kuralları Temel Kavramları 25
3.5.2 Apriori Algoritması 28

3.6 Kümeleme 37
3.6.1 Kümeleme analizinin gereksinimleri 38
3.6.2 Kümelemenin temel adımları 38
3.6.3 Kümeleme Analizinin Uygulama Aşamaları 39
3.6.4 Kümeleme Analizinde Değişken seçimi 39
3.6.5 Uzaklık ölçüleri 40
3.6.6 Kümeleme Metotları 42
3.6.7 Hiyerarşik Kümeleme Metotları 42
3.6.8 Hiyerarşik olmayan kümeleme yöntemleri 43
3.7 Normalizasyon Metotları 43
3.7.1 Literatürde Kullanılan Normalizasyon Metotları 46
3.8 Örüntü 48
3.9 Öznitelik Çıkarımı 49
3.10 Öznitelik Seçimi 49
3.11 Temel Bileşen Analizi 51
3.11.1 Kovaryans 53
3.11.2 Temel Bileşenler Analizinin Adımları 53
3.11.3 Temel Bileşenler Analizinin Geometrik Anlamı 56
3.12 Sınıflandırma 57
3.12.1 Sınıflama başarılarının Ölçülmesi 59
3.12.2 Sınıflandırma Doğruluğu 61
3.12.3 Karmaşıklık (Bozulma) Matrisi 61
3.12.4 Hassasiyet ve Seçicilik Analizi 62
4 YAPAY SİNİR AĞLARI 63
4.1 Yapay Sinir Ağlarının Tanımı 63
4.2 Yapay Sinir Ağlarının Genel Özellikleri 64
4.3 Yapay Sinir Ağlarının Avantajları 64
4.4 Yapay Sinir Ağlarının Dezavantajları 65
4.5 Biyolojik Sinir (Nöron) Yapısı 66
4.6 Biyolojik Sinir Ağları 66
4.7 Yapay Sinir Hücreleri 68
4.8 Yapay Sinir Ağlarının Temel Elemanları 70
4.8.1 Katmanlar 70

4.8.2 Gizli Katman 70
4.8.3 Ağırlıklar 71
4.8.4 Birleştirme Fonksiyonu 72
4.8.5 Etkinlik (Aktivasyon) Fonksiyonu 74
4.8.6 Hata Fonksiyonu 76
4.9 Yapay Sinir Ağı Mimarileri 76
4.9.1 İleri beslemeli yapay sinir ağları: 77
4.9.2 Geri beslemeli yapay sinir ağları: 78
4.10 Yapay Sinir Ağlarında Öğrenme 79
4.10.1 Hebb Kuralı 80
4.10.2 Hopfield Kuralı 81
4.10.3 Delta Kuralı 82
4.10.4 Kohonen Kuralı 83
5 DESTEK VEKTÖR MAKİNELERİ 84
5.1 Destek Vektör Makinelerinin Avantajları 85
5.2 Destek Vektör Makinelerinin Dezavantajları 86
5.3 Doğrusal Destek Vektör Makineleri 86
5.3.1 Doğrusal ayrılma durumu 87
5.3.2 Doğrusal Olmayan Destek Vektör Makineleri: Haritalama Yaklaşımı ve Çekirdek Fonksiyonlarının Kullanımı (Kernel trick) 92
5.3.3 Çekirdek (Kernel) Fonksiyonları 95
5.4 Ardışık Minimal Optimizasyon Algoritması (SMO) 97
6 GELİŞTİRİLEN ALGORİTMALARDA KULLANILAN VERİ KÜMELERİ 100
6.1 Göğüs Kanseri Veri kümesi 100
6.2 Kalp Hastalığı Veri kümesi 101
6.2.1 Cleveland Kalp Hastalıkları Veri Kümesi 102
6.2.2 Spect Veri kümesi 103
6.3 Karaciğer rahatsızlığı tanımı 105
6.3.1 Karaciğer rahatsızlığı veri kümesi 105
6.4 Hepatit Veri Kümesini 106
6.5 Diabet hastalığı 107
6.5.1 Diabetes Veri Kümesi 107

7 GELİŞTİRİLEN ALGORİTMALAR.....109
7.1 Apriori ve TBA ile Oluşturulmuş Hibrid Özellik Seçim Sistemi 109
7.1.1 Sınıflandırma Parametreleri 114
7.1.2 Deneysel Sonuçlar ve Tartışmalar.....114
7.2 Veri setleri için modifiye edilmiş kümeleme algoritmalarına dayalı veri hazırlama yöntemi 118
7.2.1 Sınıflandırma Parametreleri 121
7.2.2 Deneysel Sonuçlar ve Tartışmalar.....122
7.3 Veri setleri için k-kat çapraz doğrulama yöntemiyle entegre edilmiş kümeleme tabanlı veri eliminasyon stratejisiyle sınıflama performansının geliştirilmesi 128
7.3.1 Sınıflandırma Parametreleri 131
7.3.2 Deneysel Sonuçlar ve Tartışmalar.....131
7.4 Yüksek hızlı termal transfer yazıcı kafa hasarlarının belirmesini sağlayan görüntü işleme ve veri madenciliği teknikleri tabanlı uzman sistem tasarımı 138
7.4.1 Sistem tasarımı 140
7.4.2 Sınıflandırma Parametreleri 149
7.4.3 Deneysel Sonuçlar ve Tartışmalar.....149
8 SONUÇ VE ÖNERİLER.....151
KAYNAKÇA.....156
ÖZGEÇMİŞ.....165

ÇİZELGELER

Çizelge 3.1 Birliktelik kurallarını apriori algoritmasıyla bulunması amacıyla örnek bir veri seti için hareketler ve ürünler.....33
Çizelge 3.2 Tekli birlikteliklerin destek değerleri (C1 nesneler kümesi) 34
Çizelge 3.3 Minimum destek değerini sağlayan ürünler (L1 nesneler kümesi) .34
Çizelge 3.4 İkili birliktelikler ve destek değerleri 34
Çizelge 3.5 İkili birlikteliklerden destek değerini sağlayan setler.....35
Çizelge 3.6 Üçlü birliktelikler ve destek değerleri 35
Çizelge 3.7 Üçlü birlikteliklerden destek değerini aşan ürün setleri 36
Çizelge 3.8 Üçlü birlikteliklerden birliktelik kurallarının çıkarılması.....36
Çizelge 3.9 Karışıklık matrisinde yer alan dört sınıflama performans endeksleri 61
Çizelge 4.1 İnsan beyni ile bilgisayar sistemlerinin karşılaştırılması (Crone 2004).....67
Çizelge 4.2 : Biyolojik sinir sistemi ile YSA'nın yapısal benzerlikleri.....68
Çizelge 4.3 İstatistik ve YSA karşılaştırılması (Sagıroğlu ve ark 2004). 69
Çizelge 4.4 Birleştirme fonksiyonları (Öztemel 2006).....73
Çizelge 4.5 Etkinlik Fonksiyonları (Emir 2013) 75
Çizelge 5.1 DVM'de kullanımı uygun olan çekirdek fonksiyonları (Ayhan 2013) 96
Çizelge 6.1 Göğüs kanser tespitinde kullanılan nitelikler.....101
Çizelge 6.2 Cleveland Kalp hastalıkları veri kümesi de bulunan nitelikler ve aralıkları.....102
Çizelge 6.3 Liver veri kümesi de bulunan nitelikler ve aralıkları (Blake ve Merz 1998).....105
Çizelge 6.4 Hepatit veri kümesi de bulunan nitelikler ve aralıkları 106
Çizelge 6.5 Diabetes veri kümesi de bulunan niteliklerin değerleri ve aralıkları 108
Çizelge 7.1 Güvenlik seviyesine göre uygun seçilen kurallar.....110
Çizelge 7.2 DVM için sınıflandırma parametreleri listesi 114
Çizelge 7.3 MLP mimarisi ve eğitim parametreleri 114

Çizelge 7.4 AP+YSA kullanan göğüs kanseri tespiti için doğru sınıflandırma oranı. 114
Çizelge 7.5 AP+DVM kullanan göğüs kanseri tespiti için doğru sınıflandırma oranı. 115
Çizelge 7.6 Önerilen sistemin elde edilen sınıflama doğruluk oranı ve literatürdeki diğer çalışmalarla karşılaştırılması 115
Çizelge 7.7 AP + YSA kullanarak göğüs kanseri tespiti için performans sonuçları (2 hariç) 116
Çizelge 7.8 Önerilen Yöntemin Karışıklık matrisler çapraz doğrulama sonuçları 116
Çizelge 7.9 Sınıflandırma parametreleri listesi 122
Çizelge 7.10 k-ortalamalar algoritması ile veri setleri için ön işleme sonuçları 122
Çizelge 7.11 Modifiye k-ortalamalar algoritması ile veri setleri için ön işleme sonuçları. 122
Çizelge 7.12 Statlog Kalp Hastalığı veri seti için sınıflandırma performansı ... 123
Çizelge 7.13 Bizim metodumuzun ve literatürdeki diğer metotların Statlog kalp hastalığı veri seti üzerinde elde edilmiş sınıflandırma doğrulukları 123
Çizelge 7.14 SPECT görüntü veri seti için sınıflandırma performansı 125
Çizelge 7.15 Bizim metodumuzun ve literatürdeki diğer metotların SPECT görüntü veri seti üzerinde elde edilmiş sınıflandırma doğrulukları 125
Çizelge 7.16 Diabetes veri seti için sınıflandırma performansı 126
Çizelge 7.17 Bizim metodumuzun ve literatürdeki diğer metotların diabetes veri seti üzerinde elde edilmiş sınıflandırma doğrulukları 126
Çizelge 7.18 Sınıflama Parametreleri Listesi 131
Çizelge 7.19 Veri setleri üzerinde algoritmanın uygulanması 131
Çizelge 7.20 Hepatitis veri seti için sınıflandırma performansı 132
Çizelge 7.21 Bizim metodumuzun ve literatürdeki diğer metotların Hepatitis veri seti üzerinde elde edilmiş sınıflandırma doğrulukları 132
Çizelge 7.22 Liver Disorders veri seti için sınıflandırma performansı 133
Çizelge 7.23 Bizim metodumuzun ve literatürdeki diğer metotların Liver Hastalığı veri seti üzerinde elde edilmiş sınıflandırma doğrulukları 134
Çizelge 7.24 SPECT görüntüleri veri seti için sınıflandırma performansı 134

Çizelge 7.25 Bizim metodumuzun ve literatürdeki diğer metotların Spect görüntüleri veri seti üzerinde elde edilmiş sınıflandırma doğrulukları 135
Çizelge 7.26 Statlog Heart Disease veri seti için sınıflandırma performansı 136
Çizelge 7.27 Bizim metodumuzun ve literatürdeki diğer metotların Statlog kalp hastalığı veri seti üzerinde elde edilmiş sınıflandırma doğrulukları 136
Çizelge 7.28 Kamera 590CU marka CCD kameranın teknik özellikleri 142
Çizelge 7.29 Sınıflamada kullanılan parametre listesi 149
Çizelge 7.30 Termal yazıcı kafalarından elde ettiğimiz görüntülerden oluşturulmuş veri seti için sınıflandırma performansı 149
Çizelge 7.31 Termal yazıcı kafalarından elde ettiğimiz görüntülerden oluşturulmuş veri seti için saf hali ve ön işlemeden sonraki sonuçları 150

ŞEKİLLER

Şekil 3.1 Veri tabanlarında bilgi keşfi aşamaları(Han ve Kamber 2001) 16
Şekil 3.2 Apriori algoritması akış diyagramı(Gülce 2010).....29
Şekil 3.3 Apriori algoritmasının birleştirme özelliği(Gülce 2010).....29
Şekil 3.4 Apriori budama işleminin grafiksel gösterimi(Gülce 2010).....30
Şekil 3.5 Apriori Algoritmasının basitleştirilmiş kodu.....31
Şekil 3.6 Canberra uzaklık ölçüsü(Erdoğan 2004).....42
Şekil 3.7 K-ortalamalar akış şeması(Sevinç 2009).....45
Şekil 3.8 SFS için basitleştirilmiş kod (Ladha ve Deepa 2011).....50
Şekil 3.9 SBS için basitleştirilmiş kod (Ladha ve Deepa 2011).....51
Şekil 3.10 Ana Bileşenler analizi yönteminin grafiksel gösterimi (Bozik 2011)52
Şekil 3.11 Öz değerler ve temel bileşenler(Yalçın 2010) 55
Şekil 3.12 Temel bileşenlerin varyans yüzdesine göre sıralanması(Yalçın 2010) 56
Şekil 3.13 Temel bileşen ve herhangi bir noktanın gösterimi(Yalçın 2010) 56
Şekil 3.14 Temel Bileşen(Yalçın 2010).....57
Şekil 3.15 Temel bileşen düzlemi(Yalçın 2010) 57
Şekil 3.16 *k*-kat çapraz doğrulama yönteminin grafiksel gösterimi 60
Şekil 3.17 k-kat çapraz doğrulama yöntemin basitleştirilmiş kodu.....60
Şekil 4.1 Bir canlının biyolojik sinir hücresinin yapısı (Freeman ve Skapura 1991).....67
Şekil 4.2 Biyolojik ve yapay sinir ağı.....68
Şekil 4.3 Yapay sinir hücresinin yapısı (Tarassenko 1998).....69
Şekil 4.4 İleri ve Geri yönelimli Ağ mimarileri(Emir 2013) 77
Şekil 4.5 İleri yönelimli çok katmanlı ağ(Emir 2013).....78
Şekil 4.6 Geri beslemeli çok katmanlı yapay sinir ağı(Emir 2013).....79
Şekil 4.7 Hebb net'in mimarisi(Fausett 1994) 80
Şekil 4.8 Hebb algoritmasının basitleştirilmiş kodu.....81
Şekil 4.9 : Hopfield Net Modeli(Cinsdikici 2015) 82
Şekil 5.1 DVM'nin genel yapısı (Ayhan 2013) 84
Şekil 5.2 2-sınıflı veri setini ayıran farklı düzlemlere ilişkin örnekler(Ayhan 2013).....86

Şekil 5.3 2-sınıflı problem örneği için Destek Vektör Makinaları(Ayhan 2013)87
Şekil 5.4 2-sınıflı problem için doğrusal ayrılabilme durumu(Ayhan 2013).....89
Şekil 5.5 2-sınıflı problem için doğrusal ayrılamama durumu(Ayhan 2013) 92
Şekil 5.6 Doğrusal olmayan haritalama yaklaşımı örneği(Ayhan 2013).....93
Şekil 5.7 Girdi uzayının özellik uzayına dönüştürülmesi gösterilmiştir (Ayhan 2013).....94
Şekil 5.8 İki Lagrange çarpanının gösterilmesi (Elmas 2012) 98
Şekil 5.9 SMO algoritmasının küçültülmüş kodu (Kuan ve ark 2012).....99
Şekil 6.1 Kalp SPECT görüntülerinde serpilme: (a) Normal serpilme; (b) normal olmayan serpilme. 103
Şekil 6.2 Kalp SPECT görüntülerinden özellik el edilmesinin grafiksel gösterimi(Kurgan ve ark 2001).....103
Şekil 6.3 Kadın hastalarda normal LV Şablon-modeli (Kurgan ve ark 2001)..105
Şekil 6.4 Liver görüntüsü.....105
Şekil 7.1 AP+YSA metodu blok şeması 110
Şekil 7.2 Basitleştirilmiş Apriori algoritma kodu.....113
Şekil 7.3 Önerilen sistemin blok diyagramı 119
Şekil 7.4 Modifiye edilmiş k-ortalamalar algoritmasının basitleştirilmiş kodu 120
Şekil 7.5 Önerilen veri hazırlama sistemi 121
Şekil 7.6 Önerilen veri hazırlama yönteminin basitleştirilmiş kod sunumu 121
Şekil 7.7 Önerilen veri eleme yönteminin basitleştirilmiş kod sunumu 129
Şekil 7.8 Önerilen sistemin blok diyagramı 130
Şekil 7.9 a) Termal transfer baskı Sürekli stili b) Baskı kafası yapısı.....140
Şekil 7.10 Meiji MX9430 serisi mikroskop 141
Şekil 7.11 TPH Analiz Sistemi Tasarım 142
Şekil 7.12 Tasarlanan sistemin ön işleme algoritması Sonuçları 143
Şekil 7.13 Her ısıtma hücresinin yatay pozisyonlarının tespiti 144
Şekil 7.14 Tüm ısıtma hücrelerinin bölütlemesi için basitleştirilmiş kodu (Pseudo-Code)(Yilmaz ve ark 2012).....145
Şekil 7.15 Geliştirilen Sistemin Veri İşleme Blok Şeması.....146
Şekil 7.16 Son pixel görüntüleri 146
Şekil 7.17 Termal baskı kafalarından elde edilen örnek görüntüler.....147
Şekil 7.18 Baskı kafalarının sisteme alınamadığı görüntülerinden örnekler 147

Şekil 7.19 Önerilen veri eleme yönteminin basitleştirilmiş kodunun sunumu..148

SİMGELER VE KISALTMALAR

AIRS	Ağırlıklandırılmış Bulanık Sistem İle Yapay Bağışıklık Tanıma Sistemi
AIS	Yapay Bağışıklık Sistemleri
AWAIS	Özellik Ağırlıklı Yapay Bağışıklık Sistemi
DHMM	Ayrık Gizli Markov Modeli
DVM	Destek Vektör Makinaları
FN	Yanlış Negatif
FP	Yanlış Pozitif
FS	Özellik Seçimi
FS	İleri Seçim
FSM	Özellik Uzayı Haritalama
KFFS	F-Skor Özellik Seçimini
KFFS	Çekirdek F-Skor Özellik Seçimini
k-NN	K-En Yakın Komşu
LDA	Doğrusal Diskriminant Analizi
LFDA	Yerel Ficher Ayırma Analizi
LV	Sol Kapakçık
LVQ	Learning Vector Quantization
NASA	Ulusal Havacılık Ve Uzay Dairesi
NN	Sinir Ağı
PCA	Temel Bileşen Analizi
PSO	Parçacık Sürüsü Optimizasyonu
SA	Benzetimli Tavlama Metodunu
SBFS	Ardışık Geri Yönde Kayan Seçim
SBS	Ardışık Geri Yönde Seçim
SFFS	Ardışık İleri Yönde Kayan Seçim
SFS	Ardışık İleri Yönde Seçim
SNR	Düşük Sinyal-Gürültü Oranı
SSV	Ayrılabilen Bölünmüş Değerler
SVDF	Destek Vektör Karar Fonksiyonu

SVM	Support Vektor Machine(Destek Vektör Makinaları)
TBA	Temel Bileşenler Analizi
TN	Gerçek Negatif
TP	Gerçek Pozitif
UCI	University Of California, Irvine
VHR	Yüksek Uzaysal Çözünürlüğü
VTBK	Veri Tabanlarında Bilgi Keşfi
YSA	Yapay Sinir Ağları

1 GİRİŞ

Bilişim alanındaki teknolojik gelişmeler sayesinde bilgisayar sistemlerinin donanımsal kapasiteleri hızla artmakta ve bu teknolojik cihazların maliyetleri düşerek her sektörde kullanımı yaygınlaştırmaktadır. Bu yaygınlaşma ve teknolojinin ilerlemesiyle birlikte veri tabanlarında ve diğer bilgi depolama birimlerinde çok sayıda bilgi depolanmaya başlanmıştır. Veri tabanı sistemlerinin kullanımındaki artış ile toplanan veriler, kendi başlarına değersizdir. Bu verileri insanlık için değerli hale dönüştürmek için, veriler içerisinden hedefe uygun faydalı bilgiye ulaşmak gerekir. Verinin bilgiye dönüştürülme işlemi veri analizi olarak tanımlanır. Saklanan verilerden hem ticari olarak, hem de o zamana kadar yapılan işlemlerden gelecekle ilgili anlamlı tahminler çıkaracak sistemler geliştirilmeye başlanmıştır. Elde edilen veriler büyüdükçe veriden anlamlı bilgiler çıkarılması güçleşir. Geleneksel analiz yöntemleriyle bu büyük veri yığınlarından gizli örüntüler bulmak oldukça güçtür. Bu nedenle bu örüntüler bilgi keşfi ya da veri madenciliği olarak bilinen teknikler yardımıyla çözümlenebilir. Veri madenciliği, büyük veri yığınları veya veri tabanları içerisinden desenlerin, ilişkilerin, düzensizliklerin, kuralların keşfedilmesiyle gelecekle ilgili tahminde bulunabilmemizi sağlayabilecek bağıntıların bir bilgisayar programı kullanarak aranmasıdır.

Verilerin kalitesi veri madenciliği uygulaması için önemlidir. Çeşitli kalite önlemleri arasında doğruluk ve tutarlılık veri kalitesi için en önemli iki parametredir. Gerçek hayatta elde edilen veriler içerisinde, sıklıkla bu iki parametreye uygun olmayan veriler bulunur. Veri tabanları büyüdükçe pek çok niteliklerin değeri yanlış girilmiş olabilir. Hatalar, veri girişi sırasında yapılan insan hataları ya da girilen değerin yanlış ölçülmesinden kaynaklanır. Veri girişi ya da veri toplanması sırasında oluşan sistem dışı hatalara gürültü adı verilir.

Gürültülü veriler ticari yazılımların çalıştığı veri tabanlarında da ciddi problemler oluşturmaktadır. Veri tabanlarından alınan veri kümelerinde bulunan gürültülü veriler üzerinde yapılan problem araştırmaları, tümevarımsal karar ağaçlarında uygulanan yöntemler bağlamında kapsamlı bir biçimde yapılmıştır. Buradan çıkan sonuca göre özellikle tanıma sistemlerinin eğitiminde gürültülü verilerin ayrıştırılması ve ihmal edilmesi gerektiği ortaya çıkmıştır. Quinlan, gürültünün sınıflama üzerindeki

etkisini araştırmak için bir dizi deneyler yapmıştır (Quinlan 1986). Çıkan sonuçlarda etiketli öğrenme de gürültülü verilerin öğrenme algoritmasının başarımını doğrudan kötü yönde etkilediği görülmüştür.

Çağımız hız çağıdır. Elde edilen verilerden hem anlamlı bilgiler çıkarılması hem de bunların hızlı bir biçimde elde edilmesi gerekir. Bu amaçla tez çalışmasında veri tabanlarında oluşan gürültülü, eksik ya da yanlış sınıflanmış veriler tespit edilip, elenerek sınıflama başarısı yükseltilmiştir. Bununla birlikte verilerin kullanılmayan özelliklerini tespit edilen bir sistemde geliştirilerek anlamlı bilgiye dönüştürme süreleri kısalmıştır. Bu özelliklerin bir daha kullanılmayacağı düşünüldüğünde, bu niteliklerin verilerini elde etmek için kullanılan iş gücü ve zamandan da tasarruf sağlanmış olacaktır.

İlişkisel veri tabanlarında depolanan, metin ağırlıklı verilerden bilgiyi keşfetmek için araştırmacılar sıklıkla veri madenciliği yöntemlerinden yararlanmışlardır. Bununla birlikte depolama ürünlerinin gelişmesi ve daha düşük maliyetle yüksek kapasitelerin kullanılması ile birlikte görüntü gibi standart olmayan verilerinde depolanıp anlamlı bilgiye dönüştürülmesi problemleri ortaya çıkmaya başlamıştır. Görüntü verilerinden değerli bilgilerin çıkarılması veri madenciliği için yeni bir odak noktası olmaya başlamıştır. Görüntü madenciliğinde anlamlı bilgiler içeren resimler ele alınır. Bu resimler analiz edilir. Hem bu analizler sırasında hem de analizler sonucunda çıkan veriler üzerinde veri madenciliği çalışmaları yapılır.

Tez çalışmasında kullanılan görüntü verileri tıp ve endüstriyel cihazlarda elde edilen görüntülerden elde edilmiştir.

Endüstriyel alanda, termal transfer baskı makinesinin, termal kafalarının mikroskop görüntüleri alınıp, bu kafaların bir uzmana gerek olmadan bozuk olup olmadığını ve bozukluk oranlarını tespit eden bir sistem tasarlanmıştır. Kurulan bu sistemin daha başarılı sonuçlar elde edebilmesi için sistemde bulunan gürültülü, bozuk ve yanlış sınıflanmış verileri tespit eden ön işleme yöntemi geliştirilmiş ve bu sayede sistemin tespit başarısı artırılıp daha hızlı karar alınmasına yardımcı olunmuştur. Çalışmada kullanılan resimler bir mikroskop yardımıyla alınmakta ve belli bir ön işleme işleminden geçtikten sonra termal baskı kafalarının görüntüleri piksel piksel sınıflandırıcıya verilmektedir. Sınıflandırıcı tarafından termal baskı kafalarının bozuk olup olmadığını tespit eden sistemin başarı oranı 10-kat çapraz doğrulama kullanılarak

%96.45 oranında tespit edilmiştir. Sistemin ön işleme yöntemiyle geliştirilmesi sonucunda başarı oranı 10-kat çapraz doğrulama yöntemi kullanılarak %99.88 bulunmuştur. İşlemin güvenirliliği için 10-kat çapraz doğrulamanın 500 kere çalıştırılmasıyla elde edilen ortalama değer %99.75 olarak tespit edilmiştir.

Tıbbi cihazların günümüzde gelişmesi ve sayısal hale gelmesine bağlı olarak hastane bilgi sistemleri de gelişmiştir. Bunun sonucu olarak da tıp alanı veri birikmesinin en çok yaşandığı alanlardan biri olmuştur. Bu sayısal verilerden yararlanarak çıkan sonuçları doktorların ya da sağlık çalışanlarının tek başlarına yorumlaması oldukça güçtür. Ayrıca bu yorumları yapabilecek kadar deneyimli personel her sağlık kuruluşunda bulunmayabilmektedir. Doktorlar tanı koyma işlemlerini belli tahliller ve elde ettikleri verilere dayanarak teşhis koymaya çalışırlar. Hastalar için kullanılan verilerin artması ve karmaşıklaşması veri analizi, geleneksel yöntemler ve istatistiksel analizlerin yanı sıra bu tip problemlerin çözümü için veri madenciliği teknikleri araştırmacı ve doktorların yardımına yetişmiştir.

Bu tez çalışmasında uygulanan yöntemlerden biride tıbbi metin ve görüntü verilerinin sınıflandırılması ve yeni bir hastayla ilgili tanının hızlı bir şekilde uzmana gerek duyulmadan elde edilebilmesinin sağlanmasıdır. Bu tür verilerle çalışırken veri elde edilen cihazlardaki arızalar, eksikliklerle yeterli verinin elde edilememesi, ölçümlerdeki hatalar veya eksiklikler neticesinde sonuca etki edecek birçok problemler ortaya çıkmaktadır. Bu problemlerin çözümü içinde bu yanlış verilerin temizlenmesi gerekmektedir. Tez çalışması sırasında bu tür bilgiler geliştirilen yöntem ile temizlenmiştir. Gerekmeyen özelliklerde tespit edilip hastalardan o verilerin alınmadan da tanı konulması sağlamıştır. Ayrıca ortaya konan sistem ile muhtemelen yanlış tanı konulmuş hastaları da tespit edilerek bir kontrol mekanizması olarak da kullanılabilmektedir.

Gürültülü bilgilerin elenmesi için uygulanan sistem iki türlü tasarlanmıştır. İlk yöntemde sınıflandırıcının hemen önüne k-ortalamalar algoritmasıyla bir ön işleme yapılarak sınıflamayı bozan veriler tespit edilmiştir. Bu veriler elenerek sınıflandırıcıya verilmiştir. Bu sistemi geliştirmek için ayrıca k-ortalamalar algoritmasında uzaklık ölçütü olarak kullanılan Öklid formülüne ağırlık katsayısı eklenerek sistemi iyileştirici etki yapması sağlanmıştır. İkinci yöntem sınıflandırıcının güvenini sağlamak için kullanılan 10-kat çaprazlama yönteminde ayrılan eğitim verilerin de bu eleme işlemi

kullanılmıştır. Test için ayrılan verilerde bir işlem yapılmayarak bu tür sistemi bozan bilgilerin gelmeye devam etmesiyle, sistemi bozan bilgileri eleme yaptıktan sonra sınıflandırıcı da ne kadar etki yaratacağı ölçülmüştür. Burada da k-ortalamalar algoritmasına ağırlık katsayısı eklenerek iyileştirme yapılmıştır. Sınıflandırıcının başarısının arttığı gözlenmiştir. Eğitim işlemleri daha kaliteli veriler üzerinde yapılırken elde edilen eğitilmiş tanıma sistemin genelleme yeteneği ham haldeki test verileri üzerinden denenmiştir.

Sınıflandırma, yeni bir nesnenin özelliklerini açıklamak ve bu yeni nesnenin daha önceden tanımlanmış gruplardan birine atanmasını sağlama işleminin yapılmasıdır. Veri madenciliğinde önemli bir bölümü olan sınıflandırma işlemi önceden belirlenmiş bir sınıfa ait olan verilerin oluşturduğu veri kümesine dışardan yeni bir veri verildiğinde hangi sınıfta olduğuna karar verecek modelin bulunmasıdır (Han ve Kamber 2001).

Sınıflandırma problemlerinde ve birçok öğrenme algoritmalarında konu ile ilgili olan, en faydalı ve en önemli nitelikler seçilerek veri tabanına ait nitelik sayısının azaltılması önemlidir. Bunun amacı daha az hesaplama yükü ile daha yüksek başarı elde edilebilmesidir. Büyük veri tabanlarında nitelikler arasındaki ilişkileri tespit etmek için birçok yöntem kullanılmaktadır. Bu yöntemlerden birisi de birliktelik kurallarıdır. Birliktelik kuralının bu özelliği, özellik seçimi yöntemi olarak kullanılabileceğini göstermektedir. Birliktelik kuralları arasındaki en yaygın kullanılan algoritmalardan birisi de Apriori algoritmasıdır. Eğer bir veya birkaç nitelik kümesi belli bir destek ve güven değerinde birbiri ile ilişkili ise bu niteliklerin bir altkümesi bir başka nitelik altkümesini belirtir.

TBA ise veri tabanındaki verinin en önemli karakteristik niteliklerini belirlerken bir taraftan da verinin boyutunu azaltarak ayırt edici nitelikleri içeren bilginin kullanılmasını sağlamaktadır. Bu yöntem var olan özelliklere bir dönüşüm uygulayarak yeni bir uzaya çevirmekte ve istenirse özellik sayısı da azaltılmaktadır. Çalışmamızda birliktelik kuralları ve TBA birleştirilerek bir hibrid sistem geliştirilmiştir. Bu hibrid sistem sınıflandırıcının önünde bir öznitelik seçme yöntemi olarak kullanılmıştır. Hibrid sistem sayesinde daha az öznitelikle, hızlı bir şekilde daha başarılı sonuçlar elde edilmesi amaçlanmıştır. Kurulan sistem, literatürde sıkça kullanılan UCI veri tabanından alınan Wisconsin göğüs kanseri veri kümesinde test edilmiştir.

Yapılan bu tez çalışması 8 bölümden oluşmaktadır.

Birinci bölümde; Teze genel bir bakış açısı kazandırmak için veri madenciliğiyle ve tezde uygulanan yöntemlerle ilgili temel bilgilere giriş yapılmıştır.

İkinci bölümde; Veri madenciliği, resim madenciliği, gürültülü verilerin temizlenmesi, özellik seçimi ile ilgili literatürde yapılan çalışmalar hakkında bilgiler verilmiştir.

Üçüncü bölümde; Veri madenciliği ve veri madenciliği ile önemli tanımlamalar yapılmıştır. Literatürde ve bu tezde kullanılan teknikler genel hatlarıyla anlatılmıştır. Veri madenciliğinde karşılaşılan problemlere yer verildikten sonra tezde üzerine odaklandığımız birliktelik kuralları, kümeleme yöntemleri, Temel Bileşen Analizi (TBA) ve normalizasyon metotları hakkında daha fazla detaya yer verilmiştir. Sınıflama işlemleri sırasında kullanılan değerlendirme ölçütleri hakkında bilgi verilmiştir

Dördüncü bölümde, tezde kullanılan sınıflandırıcılardan Yapay Sinir Ağları (YSA) ve Beşinci bölümde ise Destek Vektör Makinaları (DVM) detayları olarak anlatılmıştır.

Altıncı bölümde tez boyunca tasarlanan sistemleri test etmek için kullandığımız veri setlerinin tanıtımı ve verilerin içeriklerinin anlatımı yapılmıştır.

Yedinci bölümde, geliştirdiğimiz sistemlerin anlatımı yapılmıştır. Bu bölümde toplam dört farklı uygulamaya yer verilmiştir. Veri setlerinde, veriyi temsil eden niteliklerin tamamı bütün problemlerin kullanımı için uygun olmayabilir. Uygulamalar için önem arz etmeyen öznitelikler bulunabilir. Birinci uygulamada, büyük veri tabanlarında öznitelikler arasındaki ilişkiyi bulmakta kullanılan birliktelik kurallarından biri olan Apriori algoritması ile ilişkili değişkenler içeren veri setinin boyutlarını daha az boyuta indirgenmesini sağlayan bir dönüşüm tekniği olan TBA (PCA) birleştirilerek bir hibrid öznitelik seçme yöntemi oluşturulmuştur. Yöntemin başarısını ölçmek için YSA ve DVM sınıflandırıcılarında denenmiştir. Test işlemlerinde dünya çapında sık kullanılan veri tabanı olan Wisconsin göğüs kanseri veri seti kullanılmıştır.

İkinci uygulamada; gürültülü, tutarsız ve yanlış sınıflanmış verilerin tespit edilip sisteme zarar veren bu verilerin elenmesi için önişleme yapılmıştır. Bu işlemde kullanılmak üzere modifiye edilmiş k-ortalamalar Algoritması ve sınıflandırıcı olarak Destek Vektör Makinalarını (Support Vector Machines) kullanan hibrid bir yaklaşım sunulmuştur. Geliştirilen yaklaşım insan yaşam kalitesini düşüren, toplumda oldukça sık

görülen ve önemli ölüm sebepleri arasında yer alan kalp ve diyabet hastalıklarının teşhisinde test edilmiştir.

Üçüncü uygulamada; Yanlış tamamlanan eksik verilerin, gürültülü ve tutarsız verilerin eliminasyonunun, doğrulama yöntemleri ile uyumlu bir şekilde sınıflama sistemine uygulanması bu çalışmanın ana temasını oluşturmuştur. Geliştirilen bu sistemde DVM sınıflandırıcısı ile k-ortalamalar algoritması bütünleştirilerek eğitim sürecinin gürültülü ve tutarsız verilerinden arındırılması sağlanmıştır. Test verileri üzerinde herhangi bir işlem yapılmamıştır. Sınıflandırıcıda eğitim için ayrılan verilerde ön işleme yapılmıştır. Sınıflandırıcının başarısını ölçmek için ayrılan test verilerinde herhangi bir işlem yapılmayıp ham halde test edilmiştir. Uygulanan yöntemin başarısını göstermek için literatürde sıkça kullanılan veri tabanlarından elde edilmiş Hepatitis, Liver Disorders, SPECT görüntüleri ve Statlog (Heart) veri setleri kullanılmıştır.

Dördüncü uygulamada; görüntü işleme ve veri madenciliği yöntemleri ile termal transfer baskı makinesinin, baskı kafalarının zarar görmüş bölgelerini tespit eden bir sistem tasarlanmıştır. Bu sistemin doğruluğunu artırabilmek için sınıflama sistemine zarar veren gürültülü ve yanlış sınıflanmış veriler elenerek sınıflama başarısı artırılmıştır.

Son bölümde; tez çalışmasında uygulanan yöntemlerin sonuçları değerlendirilmiş, literatürdeki diğer çalışmalardan üstünlükleri ortaya konmuştur. Ayrıca çalışmanın ilerde geliştirilmesi için çeşitli önerilerde bulunulmuştur.

2 LİTERATÜR ARAŞTIRMASI

Geliştirdiğimiz sistemlerde veri setlerindeki gürültülü, hatalı sınıflanmış verileri tespit etmek için uzaklık ölçütünde düzenleme yapılan k-ortalamalar algoritması kullanılmıştır. Bu algoritma tutarlı ya da tutarsız referans noktalarının belirlemesi için bir araç olarak kullanılmıştır. Veri eliminasyonu için k-ortalamalar algoritması kullanan literatürde bazı dikkat çekici uygulamalar şunlardır:

Tang ve Khoshgoftaar, NASA yazılım projelerinden elde edilen verilerin üzerine k-ortalamalar algoritması kullanarak, kümelemeye dayalı bir gürültü algılama yaklaşımı sunmuşlardır. Gürültülü veriler çıkarılarak kalan veriler C4.5 algoritması ile sınıflandırılmıştır. C4.5 algoritması ile yapılan sınıflandırma işlemi, k-ortalamalar algoritmasından sonra performansının arttığı görülmüştür. Düşük sinyal-gürültü oranı (SNR) ve yüksek uzaysal heterojenite, yüksek uzaysal çözünürlüğü (VHR) olan hiperspektral görüntülerin sınıflamasını olumsuz yönde etkilemiştir. Olumsuz etkiyi kurdukları sistem ile azaltmışlardır (Tang ve Khoshgoftaar 2004).

Patil ve arkadaşları tarafından Tip-2 diyabet verileri üzerinde klasik k-ortalamalar algoritması ile geliştirdikleri yöntem ile orijinal veri deseni üzerinde yanlış sınıflandırılmış verileri elemişlerdir. Eleme işlemi sırasında yanlış ve doğruların olduğu iki küme kullanılmıştır. Eledikleri verilerden kalanları k-kat çapraz doğrulama yöntemi ile C4.5 algoritmasıyla sınıflandırmışlardır (Patil ve ark 2010). Yine Patil ve arkadaşları meme ve diyabet verileri üzerinde doğru ve yanlış iki küme kullanarak klasik k-ortalamalar algoritmasıyla gürültülü verileri elemiş DVM ve Naive Bayes sınıflandırıcılarda başarı oranlarını artırmışlardır (Patil ve ark 2010).

Zhang ve arkadaşları tarafından geleneksel k-ortalamalar algoritmasına komşuluk endeksi ekleyerek komşuluk sınıflama, k-ortalamalar algoritmasını geliştirmişlerdir (Zhang ve ark 2013).

Teknolojinin gelişmesiyle oluşan veri yığınlarının çözümü için geliştirilen algoritmaların test edilmesi için en önemli veriler tıbbi verilerdir. Bu algoritmaların performansının ölçülmesi için tüm bilim adamlarının ortak kullanımına açık olduğu hastalık verileri üzerinde test edilmektedir. Bu amaçla kullanılan en popüler veri tabanlarından biri UCI veri kümesidir. Biz de çalışmalarımızda geliştirdiğimiz

yöntemleri test edip, karşılaştırmak için Statlog (Kalp), SPECT görüntüleri, Göğüs kanseri, hepatit ve Pima Kızılderilileri Diyabet veri kümeleri seçilmiştir. Bu veri setleri üzerine yapılan çalışmaların literatür taraması aşağıdaki gibidir.

Hepatitis veri tabanı kullanarak karaciğer hastalığı tanısı için yapılan çalışmalar şunlardır;

Polat ve Güneş, 2006 yılında hepatit hastalığının tanısı için bulanık kaynak ayırımı mekanizmasıyla yapay bağışıklık sistemi tanımasını (AIRS) kullanan hibrid bir yöntem ile özellik seçimi (FS) önermişlerdir. Önerilen sistemin elde edilen sınıflandırma doğruluğu %92.59 bulmuşlardır (Polat ve Gunes 2006).

Polat ve Güneş, 2007 yılında hibrid bir özellik seçim yöntemi sunmuştur. Hibrid sistem, önişleme ile Ağırlıklandırılmış Bulanık Sistem ile Yapay Bağışıklık Tanıma Sistemi'nden (AIRS) oluşturmuşlardır. Sistemin hepatit hastalığı veri seti için %50-50 eğitimi-test verisi şeklinde ayrılarak sınıflandırma doğrulukları %81.82 elde etmişledir (Polat ve Gunes 2007).

Polat ve Güneş, 2007 yılında hepatit veri setini C4.5 karar ağacı algoritmasıyla 19 özelliği 10 özelliğe düşüren özellik seçimi uygulamışlardır. Azaltılmış hepatit veri kümesi bulanık ağırlıklı ön işleme uygulanmış ve AIRS sınıflandırıcısıyla sınıflama işlemi yapılmıştır. Sistemden elde edilen doğruluk oranını %94.12 bulmuşlardır (Polat ve Gunes 2007).

Diğer bir çalışmada Polat ve Güneş, 2007 yılında hepatit hastalığının tahmini için Temel bileşen analizi (PCA) ve yapay bağışıklık tanıma sistemi (AIRS) ile bir sistem gerçekleştirmişlerdir. Önerilen sistemin elde edilen sınıflandırma doğruluğu 10-kat çapraz doğrulama kullanarak %94.12 çıkmıştır (Polat ve Gunes 2007) .

Kahramanlı ve Allahverdi, eğitilmiş hibrid yapay sinir ağ kurallarını çıkarmak için Yapay Bağışıklık Sistemleri (AIS) algoritması kullanan bir yöntem önermişlerdir. Önerilen yöntemin Hepatit veri kümesi için elde edilen doğruluk oranı %96.8 dir (Kahramanli ve Allahverdi 2009).

Doğantekin ve arkadaşları, Hepatit hastalığı için Bulanık Çıkarım Sistemine (ANFIS) dayanarak Adaptif Ağa ve Doğrusal Diskriminant Analizi (LDA) temelli otomatik teşhis sistemi önermişlerdir. LDA-ANFIS adını verdikleri otomatik hepatit hastalığı teşhis sisteminin sınıflama doğruluğunu %94.16 bulmuşlardır (Dogantekin ve ark 2009).

Bascil ve Temurtas, Levenberg-Marquardt eğitim algoritması ile Sinir Ağı yapısı kullanılarak hepatit hastalığının tanısını koyacak bir sistem önermişlerdir. Önerilen sistemin elde edilen sınıflandırma doğruluğu 10-kat çapraz doğrulama yoluyla %91.87 bulmuşlardır (Bascil ve Temurtas 2011).

Chen ve arkadaşları hepatit hastalığı teşhisi için Yerel Ficher Ayırma Analizi (LFDA) olarak adlandırılan bir özellik çıkarma yöntemi ile bir sınıflandırma algoritmasının entegre edilip adına LFDA_SVM adını verdikleri yeni bir hibrid sistem önermişlerdir. En iyi sınıflandırma doğruluğu %96.77 elde etmişlerdir (Chen ve ark 2011).

Sartakhti ve arkadaşları Destek Vektör Makinesi (DVM) ve Benzetimli Tavlama metodunu (SA) birleştirerek yeni bir hibrid makine öğrenme yöntemi önermişlerdir. Önerilen sistemin elde edilen sınıflandırma doğruluğunu 10-kat çapraz doğrulama kullanarak %96.25 bulmuşlardır (Sartakhti ve ark 2012).

Karaciğer Hastalıkları veri kümesini kullanarak karaciğer hastalığının teşhisi için çalışmalar aşağıdaki gibi sıralanabilir;

Lee ve arkadaşları tarafından rastgele bir çekirdek kullanılarak desen sınıflamasında, destek vektör makinalarını sorunsuz çözüm oluşturması için yeniden formüle etmişlerdir. Yeniden formüle edilmiş sorunsuz bir destek vektör makinesi (SSVM) olarak adlandırmışlardır. Önerilen sistemin karaciğer bozuklukları veri setinde 10-kat çapraz doğrulama ile sınıflandırma performansı %70.33 bulmuşlardır (Lee ve Mangasarian 2001).

Van Gestel ve arkadaşlarının makalesinde, Bayes ifadesi çerçevesinde en küçük kareler destek vektör makinesi (LS-SVM) sınıflandırıcı formülasyonu birleştirmişlerdir. Önerilen sistemin karaciğer bozuklukları veri kümesinde 10-kat çapraz doğrulama ile doğru sınıflandırma oranı % 69.7 bulmuşlardır (Van Gestel ve ark 2002).

Goncalves ve arkadaşları, veri tabanlarından kural çıkarma ve özellikle kayıt sınıflandırma için oluşturulan adına Ters Hiyerarşik Nöro-Bulanık BSP Sistemi (HNFB) adını verdikleri yeni bir nöro-bulanık modeli önermişlerdir. Önerilen sistemin doğru sınıflandırma oranını karaciğer bozuklukları veri kümesi için %73.33 olarak bulmuşlardır (Goncalves ve ark 2006).

Özşen ve Güneş, 2008 yılında yapay bağışıklık sisteminin (AIS) önemine katkıda bulunmayı amaçlamışlardır. Bu amaçla; yalın AIS ile kullanılan uzaklık

ölçütleri Euclidean, Manhattan ve hibrid benzerlik ölçütünden yararlanmışlardır. Öznitelik ağırlıklı yapay bağışıklık sistemine sahip olan sistemin karaciğer bozukluğu için doğru sınıflandırma oranı %70.17, hibrid benzerlik ölçüsüyle ile %60.57,Manhattan uzaklığı ile %60.21, Euclidean uzaklığı ile %60.00 bulmuşlardır (Ozsen ve Gunes 2008).

Li ve arkadaşları küçük bir veri kümesi için asıl verideki öznitelik değerlerinden sınıflandırma ile ilgili bilgileri genişletmek için bir bulanık tabanlı doğrusal olmayan dönüşüm yöntemi önermiştir. Sınıflandırıcı olarak bir destek vektör makinesi (DVM) kullanmışlardır. Önerilen sistemin karaciğer bozuklukları için sınıflandırma doğruluk oranını %70.85 bulmuşlardır (Li ve ark 2011).

Chen ve arkadaşları parçacık sürüsü optimizasyonu (PSO) ve bir en yakın komşu (1-NN) yöntemini entegre ederek analitik bir yaklaşım önermişlerdir. Önerilen sistemin karaciğer bozuklukları veri kümesi için 5 kat çapraz doğrulama ile sınıflandırma doğruluk oranı %68.99 bulmuşlardır (Chen ve ark 2012).

Chang ve arkadaşları tıbbi verileri sınıflandırmak için bir sorun-tabanlı tümevarımsal yaklaşımı, Parçacık Sürüsü Optimizasyon modeline entegre ederek geliştirdikleri hibrid bir model önermişlerdir. Karaciğer bozuklukları için ortalama tahmin doğruluğu % 76.8 bulmuşlardır (Chang ve ark 2012).

SPECT kalp veritabanını kullanan kalp hastalığı teşhisi üzerinde yapılan çalışmalar aşağıda sıralanmıştır.

Polat ve Güneş, 2009 yılında tıbbi veri setleri sınıflandırılmasında ön-işleme adımı olarak kullanılan yeni bir özellik seçme yöntemi olarak adlandırılan çekirdek F-skor özellik seçimini (KFFS) önermiştir. Önerilen sistem (KFFS+ LS-SVM) ile elde edilen en yüksek sınıflama doğruluk oranı %83.46 elde etmişlerdir (Polat ve Guenes 2009).

Zaman ve Karray tarafından bulanık sonuç çıkarma modeli yaklaşımı ile İleri Seçim (FS) ve Destek Vektör Karar Fonksiyonu (SVDF) temel alan özellik seçimi için yeni bir algoritma önermişlerdir. Seçilen özellikler temel alınarak elde edilen sınıflandırma doğruluğu DVM kullanılarak %76.73 bulmuşlardır (Zaman ve Karray 2009).

McSherry, iNN (k) adlandırılan etkileşimli vaka-tabanlı akıl yürütme (CBR) için bir algoritma geliştirmiştir. Bu algoritmada özellik seçimi, hedef sınıf lehine özelliğin

ayırt edicilik gücü ölçüp bu değere göre hedefe sınıf ataması yapılarak, üyeliğinin onaylanması prensibine göre işletilir. iNN(2) algoritmasında, SPECT kalp verisi üzerinde en yüksek doğruluk oranı %84.3 elde edilmiştir (McSherry 2011).

Saraçoğlu, özellik azaltımı için PCA ile sınıflandırma için ayrık gizli Markov modeli (DHMM) kullanan hibrid bir yöntem önermiştir. Önerilen sistemin Spect kalp verisi için doğru sınıflandırma oranı %72.2 bulmuştur (Saracoglu 2012).

Statlog veri kümesini kullanarak kalp hastalığı teşhisi yapan literatürdeki çalışmalar aşağıda verilmiştir.

Duch ve arkadaşları, Manhattan ile k-En Yakın Komşu (k-NN), Özellik Uzayı Haritalama (FSM), Ayrılabilen Bölünmüş Değerler (SSV) algoritmaları ile k-NN kullanımını önermişlerdir. Statlog Kalp verisinde k-NN kullanılarak doğruluk oranı %85.6 elde etmişlerdir (Duch ve ark 2001).

Sahan ve arkadaşları, 2005 yılında şekil-uzayı temsilinde Öklid mesafesi hesaplanırken tüm nitelikleri alarak olumsuz, etkilerini ortadan kaldırmak için Özellik Ağırlıklı Yapay Bağışıklık Sistemi (AWAIS) olarak adlandırılan yeni bir yapay bağışıklık sistemini önermişlerdir. Elde edilen sınıflandırma doğruluğu 10-kat çapraz doğrulama yöntemi ile %82.59 bulmuştur (Sahan ve ark 2005).

Eick ve arkadaşları k-ortalamalar kümeleme algoritması kullanarak yeni bir yaklaşım uygulamıştır. UCI' den elde edilen yedi veri seti üzerinde denemeler yapmışlardır. Bir eğitim setinin tüm noktalarının yerine öğrenilen mesafe fonksiyonu ve küme merkezlerini kullanan NCC ismini verdikleri yöntemi geleneksel 1- En Yakın Komşu (1-NN) sınıflandırıcı ve sıkıştırılmış 1-En Yakın Komşu sınıflandırıcında denemişlerdir. Kalp veri kümesinde en iyi elde edilen sınıflandırma doğruluğu 10-kat çapraz doğrulama yöntemi ile %81.07 bulmuşlardır (Eick ve ark 2006).

Özşen ve Güneş, yapay bağışıklık sisteminin (AIS) önemine katkıda bulunmayı amaçlamışlardır. Bu amaçla; yalın AIS ile kullanılan uzaklık ölçütleri Euclidean, Manhattan ve hibrid benzerlik ölçütünden yararlanmışlardır. Önerilen sistemin Statlog kalp verisi için 10-kat çapraz doğrulama yöntemi ile %83.95 bulmuşlardır (Ozsen ve Gunes 2008).

Kahramanlı ve Allahverdi, yapay sinir ağı (YSA) ve bulanık sinir ağı (FNN) içeren bir hibrid sinir ağı sunmuşlardır. Önerilen yöntemin Statlog kalp veri seti için doğruluk oranını %86.8 bulmuşlardır (Kahramanli ve Allahverdi 2008).

Polat ve Güneş, 2009 yılında tıbbi veri setleri sınıflandırılmasında ön-işleme adımı olarak kullanılan yeni bir özellik seçme yöntemi olarak adlandırılan çekirdek F-skor özellik seçimini (KFFS) önermişlerdir. Önerilen sistem (KFFS+LS-SVM) ile elde edilen Statlog Kalp veri seti için en yüksek sınıflama doğruluk oranı %83.70 bulmuşlardır (Polat ve Guenes 2009).

Abonyi ve Szeifert, denetimli bulanık kümeleme uygulanması ile 10-kat çapraz doğrulama yöntemi kullanarak göğüs kanseri verisin de %95.57 sınıflandırma doğruluğu elde etmişlerdir (Abonyi ve Szeifert 2003).

Akay, göğüs kanseri teşhisi için özellik seçimi ile birleştirilmiş DVM tabanlı bir yöntem kullanmıştır. Önerdiği metodun başarı oranları çapraz doğrulama olmadan beş özellik kullanılarak eğitim-test ayrımı %50-%50,%70-%30,%80-%20 ye göre çıkan sonuçlar sırası ile sınıflamanın en yüksek doğruluğa göre %98.53, %99.02, %99.51 bulmuştur (Akay 2009).

Yeh ve arkadaşları göğüs kanseri örüntüsü üzerinde veri madenciliği için ayrık parçacık sürüsü optimizasyonu ve istatistiksel yöntem kullanılarak yeni bir hibrid yaklaşım önermişlerdir. Bu yaklaşımın doğruluğu %98.71'e ulaşmıştır (Yeh ve ark 2009).

Karabatak ve İnce göğüs kanseri teşhisinde kullanılmak üzere AR + YSA yöntemini önermişlerdir. Bu yöntem, iki aşamadan oluşur. İlk aşamasında, giriş özellik vektörü boyutu birliktelik kuralları kullanılarak azaltılır. Gereksiz özellikler elenerek, giriş değerleri azaltılır. İkinci aşamada sinir ağına göğüs kanseri verisindeki bu azaltılmış girişler verilerek sınıflandırmada kullanılır. Test aşamasında 3-kat çapraz doğrulama yöntemi uygulamışlardır. Önerilen sistemin ortalama doğruluk oranı dört giriş özelliği için %95.6 ve sekiz giriş özelliği için %97.4 bulmuşlardır (Karabatak ve Ince 2009).

Marcano-Cedeño ve arkadaşları Göğüs Kanserini Sınıflandırmak için biyolojik meta plasite özelliğine dayalı bir yapay sinir ağı sunmuşlardır. Yapay meta plasite çok katmanlı algılayıcı (AMMLP) ile elde edilen çapraz doğrulama olmadan sınıflama doğruluğu %99.26'dır. (Marcano-Cedeno ve ark 2011)

Chen ve arkadaşları tarafından göğüs kanseri teşhisi için kaba küme tabanlı destek vektör makinalı bir sınıflandırma önermişlerdir. Önerdikleri metodun başarı oranları 5-kat çapraz doğrulamayla ve beş özellik kullanılarak eğitim-test ayrımı %50-

%50,%70-%30,%80-%20 göre çıkan sonuçlar sırası ile sınıflamanın en yüksek değerine göre %99.41, %100, %100 bulunmuş, ortalama doğruluk değeri ise sırası ile %96.55, %96.72, %96.87 bulmuşlardır (Chen ve ark 2011).

Uzer ve arkadaşları, göğüs kanserinin hızlı ve etkili tespitinde kullanılmak üzere, Kuadratik Diskriminant Analizi (Quadratic Discriminant Analysis =QDA) sınıflandırma algoritmasını kriter olarak kullanılarak geliştirilen Ardışık İleri Yönde Seçim (Sequential Forward Selection =SFS) ve Ardışık Geri Yönde Seçim (Sequential Backward Selection=SBS) ile Temel Bileşen Analizi (Principal Component Analysis=PCA)' nın birleştirilmesiyle oluşturulmuş iki tane hibrid öznitelik seçim yöntemi (SFSP ve SBSP) sunmuşlardır. Öznitelik seçiminde kriter olarak kullanılan algoritmada 10-kat çapraz doğrulama metodu uygulanmıştır. Geliştirilen iki hibrid öznitelik seçimi sayesinde giriş öznitelik uzayının boyutu 9' dan 4' e düşürülmüştür. Sınıflandırıcı olarak Yapay Sinir Ağları (YSA) kullanılmıştır. Geliştirilen iki hibrid sistemin 10-katlı çapraz–doğrulama kullanılarak elde edilen ortalama sınıflandırma doğrulukları sırasıyla SFSP+YSA için %97.56 ve SBSP+YSA için %98.57 bulunmuştur. Göğüs kanseri verisi için çapraz–doğrulama yöntemiyle yapılan çalışmalar içerisinde SBSP+YSA sonucunun çok umut verici olduğunu belirtmişlerdir. Çıkan sonuçlar, boyut azaltımı yapılarak uygulanan bu hibrid sistemin, hastalıkların daha hızlı ve başarılı teşhis edilmesi için kullanılabilir bir sistem olduğunu ortaya koymuşlardır (Uzer ve ark 2013).

Uzer ve arkadaşları çalışmalarında yeni bir özellik seçim algoritması önermiştir. Geliştirilen algoritma, arıların yiyecek arama davranışını taklit eden Yapay Arı Kolonisi(ABC) algoritmasına yeni bir uygulama alanı sağlamıştır. ABC algoritması özellik seçimi için, kümeleme ve ileri özellik seçimi algoritmaları ile birleştirilmiştir. Özellik seçiminde kullanılan bu algoritmanın 100 defa çalıştırılıp çıkan sonuçların ortalamaları alınarak, elenecek ve sınıflandırıcının girişine verilecek öznitelikler seçilmiştir. Sınıflandırıcı için Destek Vektör Makinalarını (Support Vector Machines=SVMs) kullanarak yeni bir hibrid yaklaşım sunmuşlardır. Bu makalenin amacı, veri setlerindeki önemsiz ve gereksiz özelliklerin elimine edilmesinin sınıflandırma başarısı üzerindeki etkisini DVM sınıflandırıcısı ile test etmektir. Geliştirilen yaklaşım, insan yaşam kalitesini düşüren ve toplumda yaygın olarak görülen

karaciğer ve diabet hastalıklarının teşhisinde kullanılmıştır. Bu hastalıkların teşhisi için UCI veri tabanından alınan Hepatitis, Liver Disorders ve Diabetes veri setleri kullanılmış ve önerilen sistem, bu veri kümelerinde sırasıyla %94.92, %74.81 ve %79.29 10-kat çapraz doğrulama kullanılarak sınıflandırma doğruluğuna ulaşmışlardır (Uzer ve ark 2013).

3 VERİ MADENCİLİĞİ

3.1 Veri nedir

Veri, sezgisel anlamı bulunan kaydedilmiş bilinen gerçeklerdir. Veri tek başına değersizdir, hiçbir anlam ifade etmez (Kahramanli 2008). Günlük hayatta veri, bilgi ile aynı anlamda olarak kullanılır. Ancak, düzenlenmemiş bir ölçüm olarak nitelendirilebilecek veri düzenlendiğinde bilgiye dönüşmektedir. Alınan "145" verisi ürünün ID'si mi, tutarı mı, yoksa ürün ağırlığı mı diye bilinmiyorsa, bu veri bilgi içermez. İsteğimiz amacımız doğrultusunda bilgidir. Bilgi, bir amaca yönelik işlenmiş veridir. (Alpaydın 2000). Bilgi, verinin kendisi ve açıklamasının toplamından elde edilir.

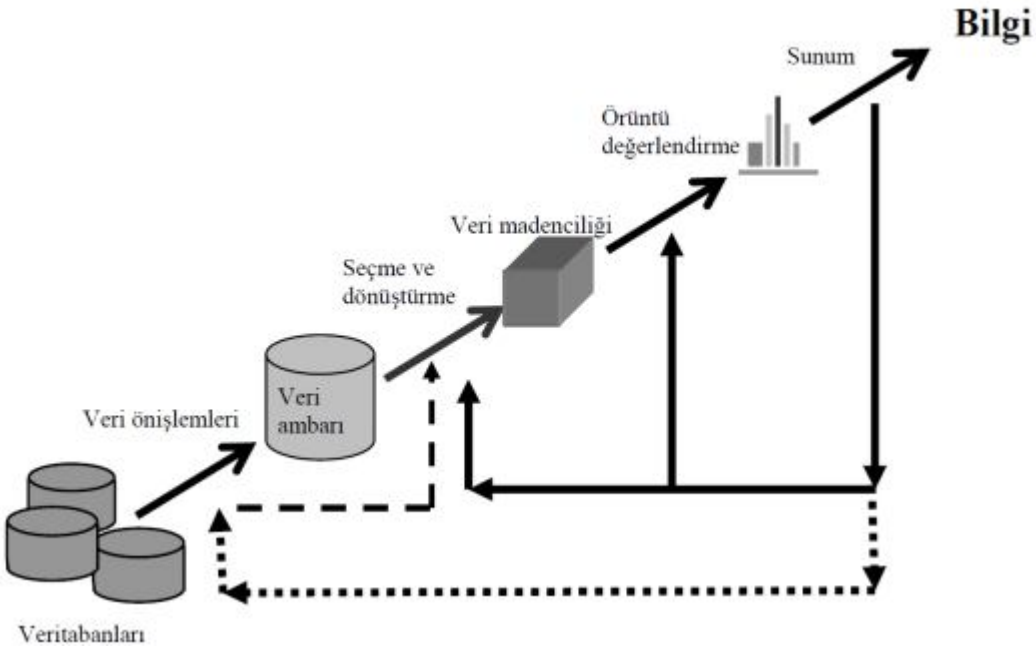
Veriyi bilgiye çevirmeye veri analizi denir. Veriyi oluşturan sayılar, harfler ve onların anlamı metaveri (metadata, üstveri) olarak bilinir (Döslü 2008).

3.2 Veri madenciliği ve bilgi keşfi aşamaları

İşlenmemiş verinin bilgiye çevrilmesi insanlık için bir problem olmuştur. Son yıllarda teknolojinin hızla gelişmesiyle büyük verilerin elde edilmesi ve bunların saklamasını sağlayan yüksek kapasiteli depolama araçları çok hızlı bir şekilde gelişmiştir. 1995 yılında birincisi düzenlenen Knowledge Discovery in Databases konferansı bildiri kitabı sunusunda, enformasyon teknolojilerinin oluşturduğu veri dağları şu cümleler ile vurgulanmaktadır. "Dünyadaki enformasyon miktarının her 20 ayda bir ikiye katlandığı tahmin edilmektedir. Bu ham veri seli ile ne yapmamız gerekmektedir? İnsan gözleri bunun ancak çok küçük bir kısmını görebilecektir. Bilgisayarlar bilgelik pınarı olmayı vaat etmekte, ancak veri sellerine neden olmaktadır".

Büyük miktardaki bu ham verilerden, nasıl faydalanacağı sorunu ortaya çıkmıştır. İstatistiksel yöntemler, geleneksel sorgu veya raporlama araçları bu veri yığılanlarından anlamlı bilgiye ulaşmada yetersiz kalmaya başlanmıştır. Bunun sonucunda yeni arayışlarla birlikte, veri tabanlarında bilgi keşfi (VTBK) (KDD-Knowledge Discovery in Databases) adı altında yaklaşımlar doğmuştur.

Veri madenciliği, VTBK içerisindeki modelin kurulması ve değerlendirilmesi aşamasına verilen genel bir ad olarak verilmektedir.



Şekil 3.1 Veri tabanlarında bilgi keşfi aşamaları(Han ve Kamber 2001)

VTBK sürecini oluşturan aşamaları Şekil 3.1'de gösterilmiştir. Bu aşamalar aşağıdaki gibi tanımlanabilir.

a) Veri Önişlemleri: Veriler üzerindeki tutarsızlıklar, düzensizlikler ve gürültülerin giderildiği işleme veri temizleme (Data Cleaning) denir. Keşfedilecek bilgilerin kalitesi bu şekilde artırılır. Bu aşamada ayrıca birden fazla kaynak varsa kaynaklardan gelen verilerin tek bir veri ambarında toplanabilmesi için gerekli uyumluluk, normalizasyon ve genelleme işlemlerinin yapıldığı veri birleştirme (Data integration) işlemi uygulanır.

b) Veri Seçme ve Dönüştürme: Konu ile ilgi verilerin bulunup seçilmesidir. Bu adım, birkaç veri kümesini birleştirerek sorguya uygun örneklem kümesini elde etmeyi gerektirir. Veri madenciliği ile ilgili bilgi seçimi ve veri türlerinin belirlenmesi, hiyerarşik yapı ve genellemelerin verilerde belirlenmesi, yenilik ve ilginçlik ölçümü yöntemlerinin belirlenmesi, sonuçta bulunan değerler için sunum araçlarını belirlenmesi gibi önişlem adımlarından geçirilir.

c) Veri Madenciliği: İnsanoğlu için veriden örüntüler çıkarmak için kullanılan çeşitli yöntemleri içeren en önemli aşamadır.

d) Örüntü Değerlendirme: Bu aşamada veri tabanından gerçekte ilginç ve doğru olan verilerin ne kadar ilginç ve yararlı olduğu tespit edilir.

e) Bilgi Sunumu: Çeşitli araçlarla keşfedilen ve elde edilen bilgilerin geçerlilik, yenilik, yararlılık ve basitlik kıstaslarına göre değerlendirilmesi ve kullanıcılara sunulmasıdır.

VBTK işleminin bütünlüğü açısında veri madenciliği kadar diğer yöntemler de önemlidir (Fayyad ve ark 1994).

3.3 Veri Madenciliği Teknikleri

Veri madenciliği teknikleri eldeki veri türüne ve elde edilen sonuçların kullanım amacına göre Tanımlayıcı ve Öngörüsel olmak üzere iki kategoride incelenebilir (Han ve Kamber 2001):

Tanımlayıcı kategori, verinin veri yığınları içindeki genel karakterini ve o anki durumuna yönelik yöntemleri içerir. Öngörüsel kategori ise eldeki verilere göre gelecekte neler olabileceğine dair tahminler, keşifler yapma, sonuç çıkarma yöntemlerini içerir.

Kullanıldıkları tekniklere göre veri madenciliği, veri yapılarına ve keşfedebildikleri örüntü biçimlerine göre kategorilere ayrılır. Veri madenciliği teknikleri için birçok gruplandırma yönteminden en çok kullanılan J.Han'a göre aşağıdaki şekilde kategorilere ayrılır.

a) Tanımlama ve Ayrımlama
b) Birliktelik Analizi
c) Sınıflandırma ve Öngörü
d) Kümeleme Analizi
e) Sıradışılık Analizi

f) Evrimsel Analiz

3.3.1 Tanımlama ve Ayrımlama

Verilerin ortak özelliklerine göre genelleştirilmiş sınıflara ayrılabilirler. Bir eğitim kurumu öğrencilerin aldığı not durumlarına göre belirli bir ortalamanın üstünde olan öğrenciyi "Çalışkan", ortalama düzeyinde olan öğrenciye "orta", ortalamanın altında olan öğrenciye "zayıf" olarak tanımlayabilir. Genellemeler verilerin ortak özellikleri veya farklı veri tabanlarının diğer veri kümelerinden farklılıklarını yansıtacak şekilde olmalıdır.

a) Tanımlama: Verilerin genel özelliklerini özetlemek için kullanılır. Örneğin "bir eğitim kurumunda bu yıl başarı oranı, %15'in üzerinde artan dersler" ifadesi bir Tanımlama işlemidir.
b) Ayrımlama: Verilerin farklarını ortaya çıkarmak için kullanılır. Örneğin "bu yıl başarı oranı %10 artan dersler ile başarı oranı %20 azalan derslerin karşılaştırılması" Ayrımlama tabanlı veri madenciliğidir.

3.3.2 Birliktelik analizi

Veri kümesindeki kayıtlar arasındaki kendiliğinden, sıklıkla gerçekleşen, birlikte ya da aynı süre içinde alınma, oluşma etkileri gibi bağlantıları arayan denetimsiz veri madenciliğidir. Birliktelik analizi, müşterilerinin satın alma davranışlarını ortaya koymak için çoğu zaman perakende sektöründe kullanıldığından "pazar sepeti analizi" olarak da adlandırılır.

Sepet analizinde amaç ürün ile ürünü alanlar arasındaki ilişkileri bulmaktır. Eğer X malını alanların Y malını da çok yüksek olasılıkla aldıklarını biliyorsak, X malını alan her müşterinin Y malını almak için potansiyel bir müşteri olduğunu bilebiliriz. Birliktelik analizi yalnızca mal ve hizmetlerin birlikte satın alınması için değil, aynı zamanda hangi koşulları sağlayan müşterilerin hangi ürünleri alacağı hakkında da çözümler getirmektedir.

3.3.3 Sınıflandırma ve Regresyon

Sınıflandırma işlemi insan düşünce yapısına en uygun ve veri madenciliği tekniklerinde en çok kullanılanıdır. Mevcut verilere dayanarak gelecekteki durumlarla ilgili öngörü yapılmasında ve yeni eklenen veri elemanının daha önceden belirlenmiş sınıflara atamak için kullanılır. Sınıflandırma, bir veri öğesini, önceden tanımlı sınıflardan birine tasnif ederken, regresyon süreklilik gösteren değerlerin tahmin edilmesinde kullanılır.

Sınıflandırma işlemine, bir okulda yeni gelen öğrencilerin hangi sınıfta eğitim görmesi gerektiğinin belirlenmesi örnek olarak verilebilir. Regresyon işlemine örnek olarak deprem tahmini verilebilir.

Sınıflama ve regresyon modeli birbirine giderek yaklaştığı için aynı tekniklerden yararlanılması mümkün olmaktadır.

Sınıflama ve regresyon modellerinde kullanılan başlıca teknikler şunlardır;

a) Karar Ağaçları
b) Yapay Sinir Ağları
c) Genetik Algoritmalar
d) K-En Yakın Komşu
e) Bellek Temelli Nedenleme
f) Lojistik Regresyon
g) Destek Vektör Makineleri

3.3.4 Kümeleme analizi

Veri tabanındaki verileri alt kümelere ayıran yöntemde, her kümede yer alan veriler birbirlerine çok benzerler, özellikleri farklı olan veriler ise farklı kümelerde bulunmaktadır. Kümeleme analizinde en önemli özellik, sınıf içi benzerliği maksimum, sınıflar arası benzerliği minimumda tutmaya çalışmaktır (Han ve Kamber 2001).

Sınıflandırma ve regresyon işleminin aksine kümeleme yönteminde, veri kümesini önceden kategorilere ayrılmaz, bunun yerine veriler dağılımlarına göre araştırılarak doğal sınıflandırmalar oluşturulur. Kümeleme analizinin sınıflandırma işleminden en önemli farkı önceden belirlenmiş sınıflar ya da sınıf tanımları (etiketleri)

olmamasıdır. Bu yüzden kümeleme işlemi gözetimsiz (unsupervised) veri madenciliği yöntemidir.

Kümelemede, genellikle bir kayıt kendisine en yakın kümeye atanır ve bu kümeyi tanımlayan değeri değiştirir. Optimum çözüm bulununcaya kadar kayıtlar yeniden atanır ve küme merkezleri ayarlanır. En yaygın kullanılan kümeleme algoritması "k ortalamalar algoritması" dır (Akbulut 2006).

Kümeleme analizi sadece veri madenciliğinde değil, örüntü tanıma, görüntü işleme, coğrafi bilgi sistemleri gibi birçok alanda yoğun olarak kullanılmaktadır.

3.3.5 Sıradışılık analizi

Sıradışılık analizi, veri tabanındaki verilerin genel özelliklerinden veya veri dağılımlarından farklılık gösteren nesnelere sıra dışı denir. Birçok veri madenciliği yöntemi istisnaları gürültü veya aşırı durumlar olarak görüp dikkate almasa da bazı durumlarda istisna noktalar genele uyan verilere göre çok daha fazla bilgi içerebilir. Kredi kartı veya sigorta sahtekârlıklarının tespitinde normal veriler haricinde uyuşmayan veriler daha çok işe yaramaktadır. Aynı zamanda tıp biliminde yeni bir hastalığın başlangıcını tespit etmede de sıradışılık analizi kullanılabilir.

Sıradışılık analizinde istatistik ve yoğunluk tabanlı yöntem olmak üzere iki yöntem mevcuttur (Han ve Kamber 2001). İstatistik tabanlı yöntem, çok büyük veri yapılarında yoğun hesaplama gerektirdiği için performansı düşüktür. Standart sapma gibi istatistik yöntemleriyle istisnalar tespit edilir. Yoğunluk tabanlı yöntem de ise her noktanın çevresindeki komşularıyla ilgili yakınlıkları hesaplanır. Yeterince komşu olmayan noktalar tespit edilir (Dinçer 2006).

3.3.6 Evrimsel analiz

Nesnelerin zamanla davranışlarının sisteme uygunluğunu ya da eğilimlerini bulmayı amaçlar (Han ve Kamber 2001).

Evrimsel analiz tanımlama, diğer veri madenciliği tekniklerini içerse de asıl amacı verinin zaman ile ilişkisini ortaya çıkarmaya çalışır. Birçok kaynakta bağımsız bir kategoride değerlendirilse de J.Han tarafından veri madenciliği içinde kategorilendirilmiştir.

3.4 Veri Madenciliğinde Karşılaşılan Problemler

Veri madenciliği uygulamalarında, veri tabanından olabilecek birçok problemle karşılaşmak mümkündür. Veri tabanlarının dinamik, eksiksiz ve net veri içermesi gerekir. Aksi durumda sağlıksız bilgiler içeren analizler yanlış stratejilerin oluşturulmasına neden olacaktır. Veri tabanları, normalizasyon şartlarını taşımalıdır.

Boyut olarak küçük veri tabanlarında çalışan veri madenciliği uygulamaları hızlı ve doğru bir biçimde çalışırken, çok büyük veri tabanlarında tam tersi davranabilir.

Küçük ya da büyük veri tabanlarında bulunan gürültü sistemlerin başarısını ve hızını kötü biçimde etkileyebilir. Veri madenciliğinde genel olarak karşılaşılan problemler aşağıda verilmiştir (Alataş 2002) (Karabatak 2008):

a) Veri tabanının boyutu
b) Gürültülü veri
c) Boş değerler
d) Eksik veri
e) Artık veri
f) Dinamik veri

3.4.1 Veri Tabanının Boyutu

Teknolojinin gelişmesiyle veri tabanı boyutları inanılmaz bir hızla artmaktadır. Makine öğrenme algoritmalarının pek çoğu küçük örneklemleri ele alabilecek biçimde tasarlanmıştır. Bu algoritmaları büyük örneklemlerle kullanmak oldukça zordur.

Bu yüzden veri madenciliği sistemlerinde karşılaşılan en önemli sorunlardan biri veri tabanı boyutunun çok büyük olmasıdır. Büyük veri tabanları taranırken arama uzayı sezgisel olarak taranabilir. Bir başka yol ise örneklemlerin, yatay/dikey olarak indirgenmesidir. Özellik seçimi yöntemleri ya da nitelik bağımlılık çizelgesi uygulanarak yapılır.

3.4.2 Gürültülü Veri

Gerçek hayatta elde edilen veriler içinde mutlaka bazı sorunlar bulunur. Veri tabanları büyüdükçe pek çok niteliğin değeri yanlış olabilir. Hatalar, veri girişi sırasında

yapılan insan hataları ya da girilen değerin yanlış ölçülmesinden kaynaklanır. Veri girişi ya da veri toplanması sırasında oluşan sistem dışı hatalara gürültü adı verilir.

Gürültülü veriler ticari yazılımların çalıştığı veri tabanlarında da ciddi problemler oluşturmaktadır. Veri tabanlarından alınan veri kümelerinde bulunan gürültülü veriler üzerinde yapılan problem araştırmaları, tümevarımsal karar ağaçlarında uygulanan yöntemler bağlamında kapsamlı bir biçimde yapılmıştır. Buradan çıkan sonuca göre gürültülü verilerin tanımlanması ve ihmal edilmesi gerektiği ortaya çıkmıştır. Quinlan, gürültünün sınıflama üzerindeki etkisini araştırmak için bir dizi deney yapmıştır. Çıkan sonuçlarda etiketli öğrenme de gürültülü verilerin öğrenme algoritmasının başarımını doğrudan kötü yönde etkilediği görülmüştür (Quinlan 1986).

Buna karşın çalışmada eğitim kümesindeki nesnelerin özellikleri/nitelikleri üzerinde en çok %10'luk gürültü miktarı ayıklanabilmiştir. Gürültünün etkisini analiz etmek için istatistiksel yöntemlerde kullanmışlardır (Chan ve Wong 1991, Karabatak 2008).

3.4.3 Boş Değerler

Veri tabanlarında her bir satırın yerine vekil olarak atanabilecek birincil anahtar değer, haricinde diğer nitelikler boş değer alabilir. İlişkisel veri tabanlarında tasarımdan kaynaklanan nedenlerle nitelik değerleri boş olarak bırakılabilir. Boş değer içeren nitelikler tamamıyla ihmal edilir ya da niteliğe olası en yakın değer atanır (Quinlan 1986).

Lee, tarafından ilişkisel veri tabanını genişletmek üzere boş değeri; uygulanamaz, bilinmeyen ve her ikisini den biri olacak şekilde öne sürmüştür. Boş değerleri sistemde devre dışında bırakmak yerine bu değerlerin ne olabileceği yönünde de çalışmalar yapılmıştır. (Grzymala-Busse , Lee 1992, Luba ve Lasocki 1994)

3.4.4 Eksik Veri

Veri tabanları genel olarak veri madenciliği dışındaki amaçlar için tasarlanmışlardır. Bu nedenle, öğrenme görevini kolaylaştıracak bazı özellikler bulunmayabilir. Veri tabanlarında her nesne ayrıntılı biçimde tanımlansaydı, sınıflama işlemi basitçe nesnelerin altkümelerinden faydalanılarak yapılırdı. Veri tabanları o anki

yapılacak işlemlere göre tasarladığı için alınan bilgiler gerçek hayatı yeterince yansıtmayabilir. Örneğin bir hastalığı tanımlamak için orta yaşlı insanların bulunduğu veri kümelerinin tanısından çocuklar için aynı hastalığın tanısı konması pek sağlıklı olmaz.

3.4.5 Artık Veri

Veri madenciliği çalışması için veri tabanlarından elde edilen bilgilerde uygun olmayan veya artık nitelikler olabilir. Veri madenciliği çalışmaları sırasında veri tabanından alınan bilgiler çalışma yapılacak konuyla ilgili nitelikler alındığı gibi ilgisiz niteliklerde alınır. Artık nitelik olarak adlandırılan bu bilgileri elemek için geliştirilmiş algoritmalar özellik seçimi olarak adlandırılmaktadır. Özellik seçiminde çok sayıda niteliğe sahip olan bir veri tabanında konu ile ilgili olan en önemli niteliklerin seçilmesi amaçlanır. Bu sayede daha az işlem yükü ile yüksek başarım elde edilebilmektedir. Özellik seçimi yalnız arama uzayını küçültmekle kalmayıp, sınıflama işleminin kalitesini de arttırmaktadır (Almuallim ve Dietterich 1991, Kira 1992).

3.4.6 Dinamik Veri

Veri tabanları içerikleri her zaman ekleme, çıkarma, düzeltme gibi işlemler ile değişim halindedir. Bu dinamik yapıda veri içerikleri sürekli değiştiğinden veri madenciliği yöntemlerinde kuralların hala aynı kalıp kalmadığı ve istikrarlılığı problemi ortaya çıkmaktadır.

Mevcut veri tabanı ile birlikte veri madenciliği uygulamaları çalıştırıldığın da mevcut uygulamalarında performansını düşürür (Karabatak 2008). Eğer veri tabanından ayrı olarak çalıştırıldığında veri tabanındaki bilgiler değişmeyeceği için, değişen verinin elde edilen örüntülere yansıtılması gerekmektedir. Veri tabanları “triger” gibi tetikleme mekanizmalarına sahiptir. Bu özellik sayesinde farklı bir yerdeki bilgilerimiz sürekli güncellenerek mevcut uygulamada sorun çıkarmadan bilgi keşif işlemleri yapılabilmektedir (Karabatak 2008).

3.5 Birliktelik Kuralları

Günümüzde, birçok alandaki veriler bilgisayarlarda ve veri tabanları üzerinde saklanmaktadır. Son zamanlarda, otomatik tanıma ve veri toplama uygulamalarındaki gelişmeler sayesinde büyük verilerin elde edilmesi ve bunların saklamasını sağlayan yüksek kapasiteli depolama araçları çok hızlı bir şekilde gelişmiştir. Biriken geçmiş verilerin analiz edilerek bu veriler içindeki birliktelik davranışlarının tespiti ile geleceğe yönelik çalışmalar yapılmasını destekleyen yaklaşıma birliktelik kuralları denir.

Birliktelik kuralı, birçok alanda geniş kullanım alanına sahiptir ve nesnelerin veya niteliklerin bir arada olma durumlarını belirlemede kullanılmaktadır. Birliktelik kuralı bulma işlemi, yoğun nesne kümesi hesaplamaya dayalı bir işlem olup büyük veri tabanları üzerinde uygulanması oldukça zaman alıcı bir işlemdir (Koyuncu 2004).

Birliktelik kuralları büyük veri tabanlarının analiz edilerek ilginç birliktelik örüntülerini keşfederek karar verme, pazarlama ve iş yönetimine oldukça fayda sağlamaktadır. Bundan dolayı veri tabanlarında bilgi keşif işlemlerinde birliktelik kuralları konusu oldukça popüler olmaktadır. Bu kurallar anlaşılması basit ve kullanışlı olduğundan finans, telekomünikasyon, pazarlama, perakendecilik ve online ticaret gibi alanlarda geniş bir alanda kullanılmaya başlanmıştır (Han ve Fu 1995, Chen ve ark 2006).

Birliktelik kurallarının kullanıldığı en belirgin örnek market sepet analizidir. Bu işlem müşterinin sepetine koyduğu farklı ürünler arasındaki birliktelikleri bularak müşterinin satın alma alışkanlıklarını analiz etmektedir. Elde edilen sonuçlar perakendecilere birlikte satılacak ürünlerden daha iyi kazançlar elde etme konusunda pazarlama stratejileri geliştirme imkânı sağlamaktadır (Chen ve Lin 2007). Perakende satışlarında, yerleşim düzenin satın alma davranışlarını etkileyen önemli faktörlerden biri olduğu görülmektedir. Bu yüzden ürünlerin tiplerine ve markasına göre yerlerinin ayarlanması oldukça önemlidir. Sepet analizleri, müşteri dağılım bilgileri de içerdiğinden indirim, mağaza düzeni, ürün alımı, reklam, planlama yatırımları için karar alanları yardımcı olmaktadır (Yang ve Lai 2006) (Griffith 2005).

3.5.1 Birliktelik Kuralları Temel Kavramları

Agrawal, Imielinski ve Swami tarafından 1993 yılında ifade edilen matematiksel modele göre; $I = \{i_1, i_2, ..., i_m\}$ nesnelerin kümesi ve D işlemler kümesi olarak ifade edilir. D veri tabanın da her hareket (transaction) T olarak, her i, bir nesne (ürün) olarak adlandırılır. Nesnelerin kümesi $T \subseteq I$ olacak şekilde tanımlanarak, her hareket bir tanımlayıcı alan olan *TID* ile temsil edilir.

Nesnelerin kümeleri olarak A ve B'yi tanımlarsak, T işlemler kümesi ancak ve ancak $A \subseteq T$ ise (A, T'nin alt kümesi) A'yı kapsıyor denir. Bir birliktelik kuralı $A \Rightarrow B$ şeklinde ifade edilir. A önce ve B sonuç olarak adlandırılır. Burada, $A \subseteq I$, $B \subseteq I$ ve $A \cap B = \varnothing$ dir.

T işleminin $A \Rightarrow B$'yi içerme olasılığına destek denir. $A \Rightarrow B$ kuralı için d olasılığı ile kuralın destek değeri, ilk olarak tanımlanır. Daha sonra, $A \Rightarrow B$ kuralının g ile gösterilen güven değeri tanımlanır. Bu değer, T işleminin A'yı ve aynı zamanda B'yi içermesidir.

Kuralın destek ve güven değerleri matematiksel olarak aşağıdaki gibi ifade edilir;

$$Destek(A \Rightarrow B) = P(A \cup B) \tag{3.1}$$

$$Güven(A \Rightarrow B) = P(B / A) \ veya \tag{3.2}$$

$$Güven(A \Rightarrow B) = Destek(A \Rightarrow B) \ / \ Destek(A) \tag{3.3}$$

Şeklinde gösterilir. Burada *Destek (A) = Destek ($A \Rightarrow A$)*'dır. Destek ve güven değerleri ile de ifade edilebilir;

$$Destek(A) = A \,/\, |D| \tag{3.4}$$

$$Destek(A \Rightarrow B) = A.B \,/\, |D| \tag{3.5}$$

$$Güven(A \Rightarrow B) = Destek(A \Rightarrow B) / \ Destek\left(A\right) \quad (3.6)$$

ile gösterilir.

A: İncelenen kayıtlardaki A ürününü içeren işlemlerin sayısını

B: İncelenen kayıtlardaki B ürününü içeren işlemlerin sayısını

A.B: İncelenen kayıtlardaki A ve B ürünlerini birlikte içeren işlemlerin sayısını

D: Veri tabanındaki bütün işlemlerin sayısını ifade etmektedir (Döslü 2008).

Kuralların ilginçliğini ifade eden destek ve güven değerleri, keşfedilen kuralların kullanışlılığını ve doğruluğunu ifade eder (Han ve Kamber 2001). A ve B nesnelerinin aynı anda bulunma ihtimalini destek değeri gösterir. A'yı içeren kayıtların B'yi de içereceğini güven değeri gösterir (Han ve Kamber 2001).

Yüksek güven ve destek değerine sahip kurallara güçlü (strong) kurallar adı verilir (Agrawal ve ark 1993). Kullanıcı tarafından asgari destek eşik değeri (min_destek) ve asgari güven eşik değeri (min_güven) belirlenir. Bu belirlenen eşik değerlerini aşan birliktelik kuralları dikkate alınır. Genel olarak eşik değerleri yüzdelik olarak ifade edilir ve %0 - %100 aralığında ifade edilmektedir (Han ve Kamber 2001).

Birliktelik Kurallarının amacı, bir D işlemler kümesinde uygulayıcı tarafından belirlenen asgari destek ve güven değerlerinden büyük A⇒ B gibi kuralların bulunmasını sağlamaktır.

Örneğin bir A ürününü satın alan müşteriler aynı alışveriş içinde B ürününü de satın alıyorlarsa, bu durum aşağıdaki birliktelik kuralı ile gösterilir.

$$A \Rightarrow B \ \left[destek \ = \ \%5, \ güven \ = \ \%82\right] \quad (3.7)$$

Denklem 3.7'deki destek ve güven ifadeleri kuralın enteresanlıklarını tanımlar ve şu şekilde yorumlanır. Market alışverişinin %5'inde A ve B ürünlerinin aynı anda alındığı ve A ürününü satın alan müşterilerin %82'si, aynı market alışverişinde B ürününü de satın aldığını ifade etmektedir (Han ve Kamber 2001). Nesnelerin kümesi, nesne küme (itemset) olarak ifade edilmektedir. k tane nesne içeriyor ise bu küme "k-nesne küme" olarak ifade edilir. Eğer bir nesne küme eşik destek değerini sağlıyor ise bu nesne küme yaygın nesne küme olarak adlandırılır. k-nesne kümenin yaygın kümeleri şeklinde ifade edilmektedir (Han ve Kamber 2001).

Sepet analizi yapılırken analiz sonunda bulunan bağıntının rastlantısal olup olmadığını anlaşılamaması gibi problemlerle karşılaşılabilmektedir. Örnek verecek olursak, markette süt alan müşterilerin %35'inin yoğurt aldığı gözlemlenirse, bu bilgi hemen bir anlam içermeyebilir. Market alışverişlerinin tümünün %35'i zaten yoğurt alışverişlerini içeriyorsa, süt ve yoğurt arasındaki bağıntı sadece rastlantısaldır diyebiliriz. Market alışverişlerinin %10'u yoğurt alırken, süt alanların %35'i süt alıyorsa analizin sonucunu çok daha farklı etkileyecektir. Bu duruma ürünlerin birbirini çekmesi denilmektedir. Problemlerin bir kısmı sepet analizi tekniğinin bağıntılar arasındaki ne kadar rastlantısal olduğunu bilmiyor olmasından kaynaklanmaktadır. Sepet analizi, bilgisayar performansına çok bağımlı olması ve analiz yapılacak ürünlerin doğru olarak seçilmesi gerekliliği zayıf yönlerindendir. Güçlü yönleri ise kolay anlaşılır sonuçlar vermesi ve denetimsiz veri madenciliği yöntemini başarı ile uygulamasıdır.

Birliktelik kuralı çıkarım problemi, minimum destek ve güvenirlik değerlerini sağlayan iki adıma ayrılmıştır (Agrawal ve ark 1993, Agrawal ve Srikant 1994) (Karabatak 2008).

a-) Kullanıcı tarafından verilmiş minimum destek değerini sağlayan ürün gruplarının bulunup, sık geçen nesne gruplarının veya yoğun nesne gruplarının adı verilmektedir. Alınan numunelerde *N* adet ürün içeriyorsa, potansiyel olarak *2N* adet yoğun nesne grubu olabilir. Yoğun nesne gruplarını bulmak için üstel arama uzayı etkili biçimde kullanılmalıdır.

b-) Minimum güvenlik kıstasını sağlayan birliktelik kurallarının bulunmasında yoğun nesne grupları kullanılır. Yoğun olan her *l* nesne grubu için, *l*'nin tüm boş olmayan alt grupları üretilir. Bu boş olmayan alt gruplar a ile temsil edilir. *a* gruplarının hepsi için $a \Rightarrow (l-a)$ gerektirmesi, *l* grubunun destek ölçütünün a grubunun destek ölçütüne oranı minimum güvenirlik eşiği ölçütünü sağlıyorsa $a \Rightarrow (l-a)$ birliktelik kuralı olarak üretilir. Böylece minimum destek eşiğine göre üretilen çözüm uzayında, minimum güvenirlik eşiğine göre taranarak bulunan birliktelikler kullanıcının ilgilendiği ve potansiyel olarak önemli bilgi içeren birlikteliklerdir. İlk adım birliktelik kurallarını belirleyen algoritmaların başarısını belirler (Ding 2001, Karabatak 2008).

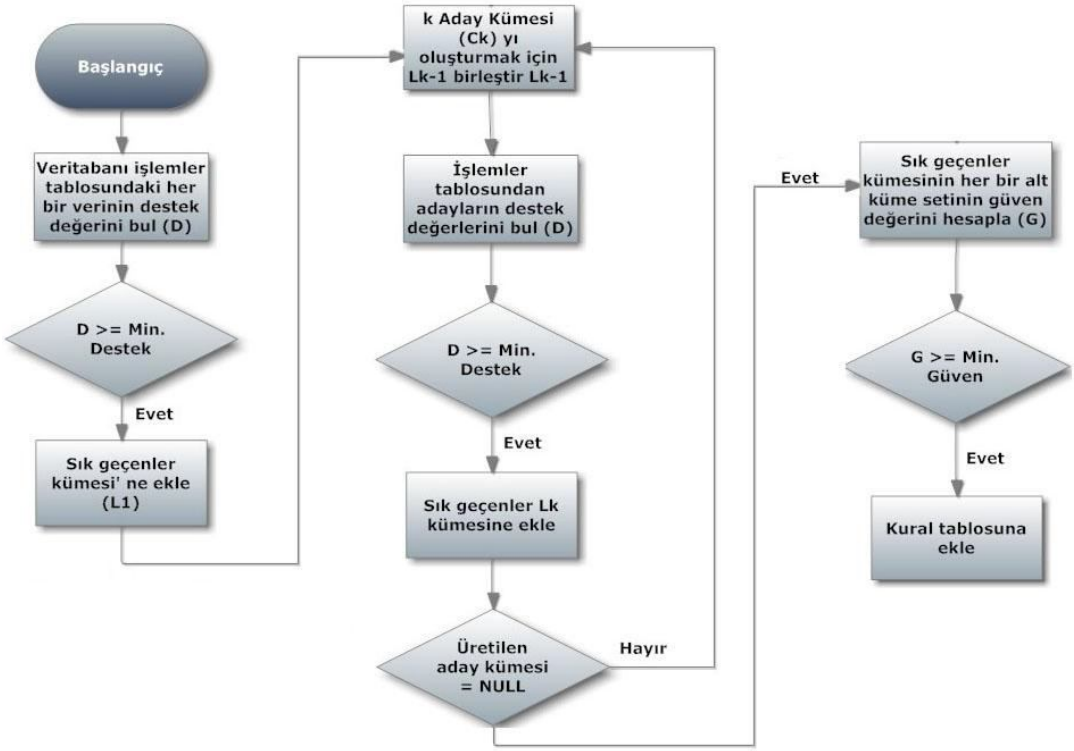
3.5.2 Apriori Algoritması

Apriori algoritması, Agrawal ve diğerleri tarafından 1994 yılında geliştirilen en çok bilinen birliktelik kuralları algoritmalarındandır. Algoritmanın ismi, bilgileri bir önceki adımdan almasından dolayı "önceki (prior)" anlamında "apriori" konmuştur. Algoritma sık kullanılan öğe kümelerinin tüm altkümeleri de sık kullanılan olmalıdır kuralına dayanmaktadır. Bu kümeleri bulmak için birçok kez veri tabanını taramak gerekmektedir. Minimum destek ölçütünü sağlayan tek elemanlı sık geçen öğe kümelerini ilk taramada bulmaktadır. Diğer taramalarda, bir önceki taramada tespit edilen sık geçen kümeler, yeni sık geçen öğe kümelerini üretmek için kullanılmaktadır. Tarama sırasında aday kümelerin destek değerleri hesaplanır. Minimum destek değerini sağlayan kümeler o kısımda üretilen sık geçen öğe kümeleri olmaktadır. Sık geçen öğe kümeleri, yeni bir sık geçen öğe kümeleri bulunmayana kadar bir sonraki geçiş için aday küme olmaktadırlar (Han ve Kamber 2001) .

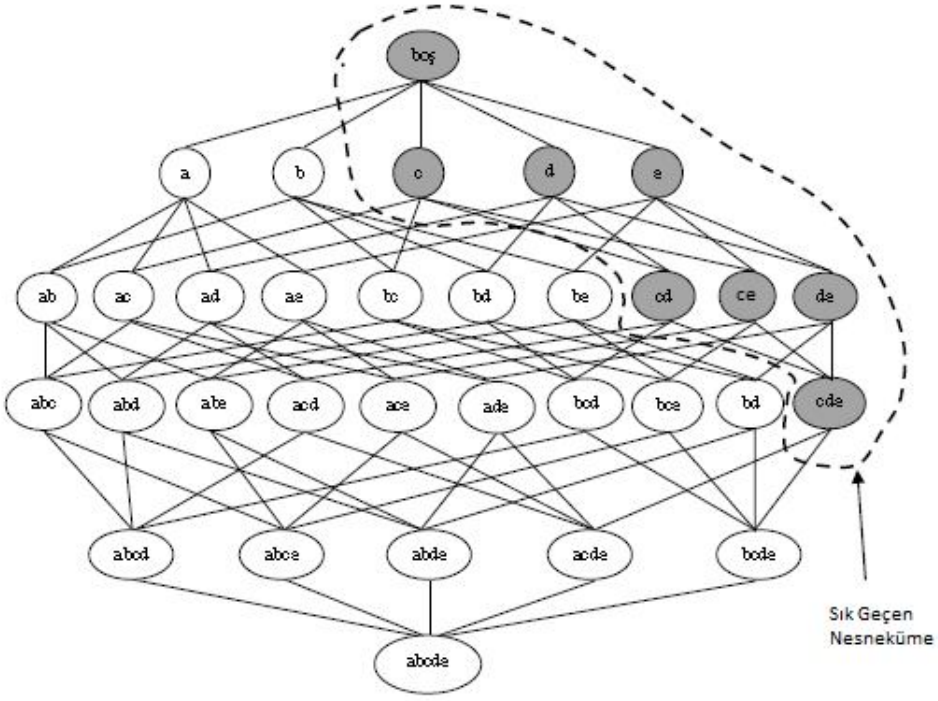
Apriori algoritmasının temel yaklaşımı bir öğe kümesi minimum destek ölçütünü sağlarsa bu kümenin alt kümeleri de minimum destek ölçütünü sağlamalıdır kuralına dayanır. Bakılan öğeler kümesinin destek değeri daima alt kümelerinin destek değerlerinden büyük olmaktadır.

Eğer bir nesneküme sık geçen ise bu kümenin boş olmayan altkümeleri de sık geçen nesne küme olmalıdır. Eğer *X* sık geçmeyen nesne kümesine *B* nesnesi eklenirse, kümenin son hali *X U B* olur. *X U B* kümesi, *X* kümesi sık geçen olmadığı için sık geçen değildir (Sever ve Oğuz 2002). Şekil 3.2'de Apriori algoritmasının akış diyagramı görülmektedir.

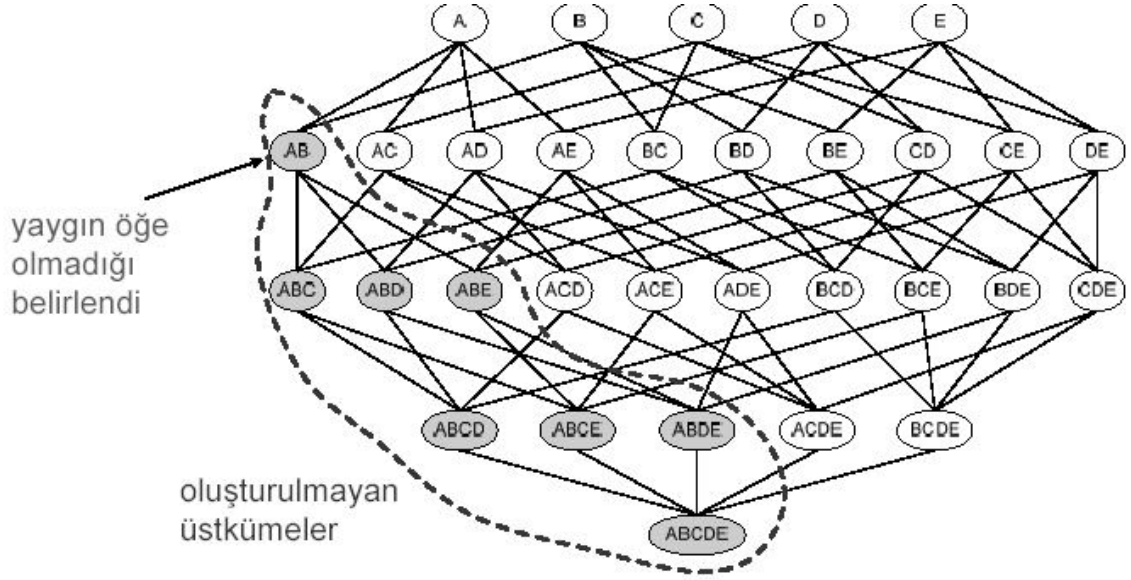
Apriori özelliği, Şekil 3.3'de görüldüğü gibi ;{c, d, e} kümesinin bir sık geçen nesne küme olduğu varsayılırsa, her işlemin bu kümenin alt kümeleri olan {c, d}, {c, e},{d, e}, {c}, {d} ve {e} kümelerini de içerdiği görülür. Sonuç olarak da, {c, d, e} sık geçen nesneküme ise bütün alt kümeleri de sık geçen nesneküme olmalıdır (Gülce 2010)



Şekil 3.2 Apriori algoritması akış diyagramı(Gülce 2010)



Şekil 3.3 Apriori algoritmasının birleştirme özelliği(Gülce 2010)



Şekil 3.4 Apriori budama işleminin grafiksel gösterimi(Gülce 2010)

Minimum destek değerinin altında kalarak sık olmayan bir nesneküme tespit edilirse, bütün üst kümeleri de sık olmayan nesneküme olmuş olur. Şekil 3.4'te gösterildiği gibi {a, b} kümesinin sık olmayan küme olarak tespit edildiğinde, {a, b} kümesini içeren bütün üst kümeler hemen budanabilir (Tan ve ark 2006).

```
Giriş: Database (D), minimum destek min_sup.
Çıkış: L, D de sık olan nesne kümeleri
(1) L1=find_frequent_1-itemsets(D) //large 1-nesnekümeler
(2) for (k=2;L(k-1)≠0;k++) {
(3) Ck=apriori_gen(L(k-1), min_sup) // yeni adaylar
(4) for each transaction t∈D{ //sayım için D yi tara
(5) Ct=subset(Ck,t); // Aday olan t' lerin alt kümeleri elde edilir
(6) for each candidate c∈Ct
(7) c.count++;
(8) }

(9) Lk={ c∈Ck | c.count ≥min_sup }
(10) }
(11) return L=Uk Lk;

procedure apriori_gen (L(k-1);min_sup) // (L k–1: sık geçen (k–1)-nesnekümeler)
(1) for each itemset l1∈L(k-1)
(2) for each itemset l2∈L(k-1)
(3) if (l1[1]= l2[1] )∧(l1[2]=l2[2]) ∧…∧ (l1[k-1]=l2[k-1]) then {
(4) c=l1x l2; // aday nesnelerin üretilmesi
(5) if has_infrequent_subset(c,L(k-1)) then
(6) delete c; // gereksiz adayların silinmesi
(7) else add c to Ck;
(8) }
(9) return Ck;

procedure has_infrequent_subset(c,L(k-1));
// (c: aday k-nesneküme;L k–1: sık geçen (k-1)-nesnekümeler); // prior bilgisini kullan
(1) for each (k-1)-subset s of c
(2) if s∉L(k-1) then
(3) return TRUE
(4) return FALSE
```

Şekil 3.5 Apriori Algoritmasının basitleştirilmiş kodu

Şekil 3.5'te Apriori algoritması ve bu algoritmaya ait alt programların basitleştirilmiş kodu görülmektedir. Basitleştirilmiş kodlar incelediğinde sık geçen öğe kümelerinin tespit edebilmek için birçok kez veri tabanı tarandığı görünmektedir. Madencilik uygulanacak veri tabanı taranarak nesnelerin kaç adet hareket kaydı içinde bulunduğu tespit edilir. Her nesne için tespit edilen bu değere destek sayıcı denir. Destek sayıcı minimum destek değerine eşit veya büyük olan nesnelerin sık geçen 1-

nesne kümesi olarak belirlendiği varsayılarak işleme başlanır. Bu işlem algoritmanın 1. satırında 1-elemanlı aday nesneler yani L1 bulunmaktadır.

Özet koddaki döngü ile sık geçen öğe kümesinin nesneleri ikili kombinasyonuna benzer bir biçimde ($L_1 \infty L_1$) birleştirilerek yeni bir küme oluşturulur. Bu oluşan aday öğe kümelere C harfi ile simgelenir. İki adet öğeden oluştuğu için bu aday kümeler C2 ifadesi ile isimlendirilir.

Apriori-gen fonksiyonu ile aday kümeler budanır. Ele alınan nesnelerin tüm alt kümeleri subset fonksiyonu ile bulunarak, öğe kümesinin içinde yer almayan küme elemanları aday kümesinden silinir. Algoritmanın uygulandığı veri seti bir kez daha taranarak budama işlemi uygulanan C2 kümesinin elemanlarına ait alt kümelerinin L1 öğe kümesi içinde yer almayan küme elemanları C2 aday kümesinden atılır. Algoritmada kullanılan veri seti tekrar taranarak C2 aday kümesi elamanlarının destek sayıcı bulunması için kaç kere geçtikleri tespit edilir. Destek sayacı, minimum destek değerinin altında olan C2 aday kümesi elemanları elenir. Sık geçen öğe kümelerinin test edilmesi işlemi de has_infrequent_subset alt programı ile gerçekleştirilmektedir. Kalan elemanlar L2 sık geçen öğe kümesini oluşturur. Budama işlemi Şekil 3.4'te gösterilmiştir. L2 kümesinin üçlü kombinasyonlarını oluşturulmak üzere döngü algoritmaya devam eder.

Bu kombinasyonlarla C3 ifadesi ile sembolleşen yeni aday kümeler oluşturulur. Önceki adımlar yeni aday küme içinde tekrarlanarak minimum destek seviyesinin üstünde kalan elemanlarla L3 sık geçen öğe kümesi oluşturulur. Bu süreç yeni bir sık geçen öğe kümesi bulunamayana kadar döngü her dönüşünde öğe sayısı artarak devam eder.

Veriler üzerinde sık geçen öğe kümeleri tespit edildikten sonra bu öğe kümeleri ile çok güçlü kurallar oluşturulabilmektedir. Minimum destek ve güven değerlerinin her ikisini de sağlayan bu kurallar aşağıdaki güven denklemleri kullanılarak elde edilebilir (Han ve Kamber 2001).

$$Güven(A \Rightarrow B) : P(B \setminus A) = \frac{Destekdeğeri(A \bigcup B)}{Destekdeğeri(A)} \tag{3.8}$$

Destek değeri (A U B), A ve B'nin birlikte bulunduğu öğe sayısıdır.

Destek Değeri (A), sadece A'yı içeren öğe sayısıdır.

Bu denklemden, aşağıdaki gibi birliktelik kuralları elde edilebilir.

• l öğesini tara ve l'nin boş olmayan alt kümelerini bul.

• Her boş olmayan altküme nesnesi için elde edilen kural;

$$"s \Rightarrow (l-s)"if \frac{destek(l)}{destek(s)} \geq \min_guv \quad (3.9)$$

Denklem 3.9'daki min_guv değeri, minimum güven değeridir. Kurallar, sık geçen öğe kümelerinden üretildiği için her bir kural minimum destek değerini sağlamaktadır. Çizelge 3.1'de birliktelik kurallarını apriori algoritmasıyla bulunması amacıyla örnek bir veri seti verilmiştir.

Çizelge 3.1 Birliktelik kurallarını apriori algoritmasıyla bulunması amacıyla örnek bir veri seti için hareketler ve ürünler

Müşteri Numarası	Aldığı Ürünler
01	Fındık, Fıstık, Leblebi, Badem
02	Nohut, Badem, Ceviz, Fıstık
03	Fındık, Badem, Nohut
04	Fıstık, Leblebi, Badem, Ayçekirdeği,
05	Fındık, Badem, Ceviz, Leblebi
06	Kuru üzüm, Badem, Leblebi
07	Fındık, Mısır, Badem
08	Ceviz, Kayısı, Leblebi
09	Kayısı, Fıstık, Leblebi
10	Leblebi, Kuru üzüm, Fındık, Ceviz, Badem

1-) Minimum destek sayısı 3 olarak ve minimum güven değeri 70 olarak belirlenir.

Çizelge 3.2'de göründüğü gibi algoritmanın ilk iterasyonunda bir elemanlı nesneler kümesi C1 olarak elde edilir. Algoritma, her nesnenin sayısını bulmak için veri tabanını basitçe taramaktadır.

Çizelge 3.2 Tekli birlikteliklerin destek değerleri (C1 nesneler kümesi)

Ürün Seti	Destek Değeri
Fındık	5
Fıstık	4
Leblebi	7
Ceviz	4
Nohut	2
Badem	8
Ayçekirdeği	1
Kuru üzüm	2
Kayısı	2
Mısır	1

Çizelge 3.3'te, L1 tablosunun oluşturulması için, destek değeri, minimum destek değeri olan 3 den küçük olan öğeler devre dışı bırakılır. Bu ürünler Nohut, Ayçekirdeği, Kuru üzüm, Kayısı, Mısır'dır.

Çizelge 3.3 Minimum destek değerini sağlayan ürünler (L1 nesneler kümesi)

Ürün Seti	Destek Değeri
Fındık	5
Fıstık	4
Leblebi	7
Ceviz	4
Badem	8

Çizelge 3.4'te, C2 tablosunun oluşturulması için, tekli birliklerden ikili kombinasyonlar elde edilir.

Çizelge 3.4 İkili birliktelikler ve destek değerleri

Ürün Seti	Destek Değeri
Fındık, Fıstık	1
Fındık, Leblebi	3
Fındık, Ceviz	2
Fındık, Badem	5
Fıstık, Fındık	1
Fıstık, Leblebi	3
Fıstık, Ceviz	2
Fıstık, Badem	3
Leblebi, Fındık	2
Leblebi, Fıstık	3
Leblebi, Ceviz	3

Leblebi, Badem	5
Ceviz, Fındık	2
Ceviz, Fıstık	2
Ceviz, Leblebi	3
Ceviz, Badem	3
Badem, Fındık	5
Badem, Leblebi	5
Badem, Ceviz	3
Badem, Fıstık	3

Çizelge 3.5 'te, L2 tablosunun oluşturulması için, destek değeri, minimum destek değeri olan 3 den küçük olan 2'li öğeler devre dışı bırakılır.

Çizelge 3.5 İkili birlikteliklerden destek değerini sağlayan setler

Ürün seti	Destek Değeri
Fındık, Leblebi	3
Fındık, Badem	5
Fıstık, Leblebi	3
Fıstık, Badem	3
Leblebi, Ceviz	3
Leblebi, Badem	5

Çizelge 3.6 'da,C3 tablosunun oluşturulması için, üçlü kombinasyonlar meydana getirilir.

Çizelge 3.6 Üçlü birliktelikler ve destek değerleri

Ürün Seti	Destek Değeri
Fındık, Leblebi, Fıstık	1
Fındık, Leblebi, Badem	3
Fındık, Badem, Fıstık	2
Fındık, Badem, Leblebi	3
Fındık, Badem, Ceviz	2
Fıstık, Leblebi, Fındık	1
Fıstık, Leblebi, Badem	2
Fıstık, Badem, Fındık	1
Fıstık, Badem, Leblebi	2
Fıstık, Badem, Ceviz	1
Leblebi, Ceviz, Badem	2
Leblebi, Ceviz, Fındık	2
Leblebi, Ceviz, Kayısı	1
Leblebi, Badem, Fıstık	2
Leblebi, Badem, Fındık	3
Leblebi, Badem, Ceviz	2

L3 tablosunun oluşturulması için, minimum destek değeri olan 3 den küçük olan 3'lü öğeler devre dışı bırakılır.

Çizelge 3.7 Üçlü birlikteliklerden destek değerini aşan ürün setleri

Ürün Seti	Destek Değeri
Fındık, Leblebi, Badem	3

Çizelge 3.7'de oluşan üçlü veri seti minimum destek değerini sağladığı ve ikili alt kümeleri olan (Fındık, Leblebi), (Fındık, Badem) ve (Leblebi, Badem) de L2 tablosunda yer aldığı için devre dışı bırakılmadan (Fındık, Leblebi, Badem) veri seti kabul edilir.

L3 kümesinde tek bir nesneküme olduğundan dolayı C4 boş kümedir ve algoritma sık geçen bütün nesnekümeleri bulduğundan sonlandırılır.

Çizelge 3.8 Üçlü birlikteliklerden birliktelik kurallarının çıkarılması

Birliktelik	Açıklama	Güven
Fındık & Leblebi => Badem	Fındık ve Leblebinin bulunduğu sette Bademin olma olasılığı	3/3=%100
Fındık & Badem =>Leblebi	Fındık ve Bademin bulunduğu sette Leblebinin olma olasılığı	3/5=%60
Leblebi & Badem =>Fındık	Leblebi ve Bademin bulunduğu sette Fındığın olma olasılığı	3/5=%60
Leblebi =>Fındık & Badem	Leblebinin bulunduğu sette Fındık ve Bademin olma olasılığı	3/7 = %42
Badem =>Fındık & Leblebi	Bademin bulunduğu sette Leblebi ve Fındığın olma olasılığı	3/8 = %38
Fındık =>Leblebi &Badem	Fındığın bulunduğu sette Leblebi ve Bademin olma olasılığı	3/5 = %60

Çizelge 3.8'den çıkan sonuçlara göre minimum güvenilirlik değeri olan %70 barajını geçemeyen kurallar elenir. Güvenlik değerini geçen tek kural (Fındık & Leblebi=>Badem) çıkmıştır. Bu kuralın tersi olan (Badem =>Fındık & Leblebi) kuralının destek değeri %70'in altında olduğu için geçerli bir kural değildir. Kuralları açıklamak gerekirse, Fındık ve Leblebi alan herkes Badem almıştır. Ama Badem alan herkes Fındık ve Leblebi almamıştır.

3.6 Kümeleme

Kümeleme analizi ilk kez 1939 yılında Tryon tarafından kullanılmıştır. Yaygın biçimde kullanılması da 1960'lı yıllardan sonra olmuştur (Anderberg 1973).

Veri madenciliğinde önemli işlemlerden birisi, büyük veri setlerini homojen olarak ayırmaktır. Kümeleme analizinin çeşitli tanımları mevcuttur. Bunlardan bazıları şu şekilde ifade edilir; bir araştırmada incelenen öğeleri aralarındaki benzerliklerine göre belirli gruplar içine alarak sınıflandırma yapan, bu sınıflarla ilgili genel tanımlamalar yapmak için öğelerin ortak özelliklerini çıkaran bir yöntemdir (Kaufman ve Rousseeuw 1990).

Diğer bir tanımda ise; verileri belirlenen yöntemlerle analiz ederek daha önceden etiketleri belli olmayan yüksek düzeyde küme içi homojenlik ve kümeler arası heterojenlik gösteren gruplara ayrılması işlemidir (Kantardzic 2011) (Günay Atbaş 2008).

Benzer birimlerin sahip oldukları karakteristiklerden yola çıkarak tanımlama amacını güden kümeleme analizi sonucunda, küme içindeki her birim önceden belirlenmiş bir kritere göre yine küme içindeki diğer birimlere çok benzer. Böylece oluşan kümelerde kümeler içi yüksek homojenlik ve kümeler arası yüksek heterojenlik sağlanır. Sınıflandırma başarılı olursa, geometrik olarak kümeler grafiğe yerleştirildiğinde küme içi birimler birbirine çok yakın iken, farklı kümelerdeki birimler birbirlerinden çok uzakta olacaktır.

Kümelemeyi sınıflamadan ayıran en önemli özellik, önceden bilenen ve hedeflenen bir sonuç olmamasıdır (Larose 2005). Ayrıca kümeleme büyük verilerin içerisindeki bilgiye kolay erişilmesi sağlayan bir teknik olmasının yanında verilerin içinde gizli kalmış örüntülerin bulunmasını da sağlamaktadır (He ve ark 2004).

Kümeler, potansiyel nesne sınıfları bulmak için de kullanılmışlardır. Bazı örnekler aşağıda sunulmaktadır (Tan ve ark 2006).

Biyoloji: Yaşayan canlıları sınıflandırmak için biyologlar çok zaman harcamışlardır. Kümelemeyi, şu an mevcut olmayan çok miktardaki genetik bilgiyi analiz etmek için kullanmışlardır. Yapılmış olan çalışmalar, bu canlıların sınıflandırılması için yaratılan matematiksel modellerin oluşturulmasına ışık tutmuştur.

Psikoloji ve ilaç: Bir hastalığın çok çeşitli varyasyonlarını farklı alt sınıflarda tanımlamak için kümeleme analizleri kullanılır.

İş: Mevcut ve potansiyel müşteriler üzerinde toplanan veriler üzerinde, kümeleme işlemleriyle müşteri analizleri ve müşterilerin işlem analizleri yapılır.

İnternet : Web üzerindeki sayfaların sınıflandırılmasında kullanılır.

Deprem araştırmaları: Deprem merkezlerinin gruplandırılması için belirlenmiş sürekli fay hatlarından yararlanılır.

Önceden tespit edilmiş kümeler üzerinde sınıflama yapan diğer algoritmalar için bir ön işlem adımı olarak da kullanır (Sevinç 2009).

3.6.1 Kümeleme analizinin gereksinimleri

Ölçülebilirlik: Çoğu kümeleme algoritmaları küçük veri setlerinde iyi çalışsa da büyük veri setlerine de uygulanabilmektedir.

Farklı özelliklerle uygulanabilme: Kümeleme algoritmalarının birçoğu sayısal verilerle işlem yapılmak üzere tasarlanmasına rağmen farklı veri tipleri içinde uygulanabilmektedir.

Farklı Biçimdeki kümelerin Keşfi: Kümeleme algoritmaları genel olarak küçük boyutlu veya küresel kümeleri bulmaya yöneldiği için Öklid ve Manhattan uzaklık ölçütünü kullanır. Kümelerin keşfinde önemli olan keyfi şekilde gruplara ayıran algoritmaların geliştirilmesidir.

3.6.2 Kümelemenin temel adımları

Kümeleme işlemlerinde gerçekleşmesi gereken adımlar aşağıdaki gibi özetlenmektedir (Jain ve Dubes 1988).

a) Örüntü seçimi: Bu adımda küme sayısının belirlenmesi, örüntü kümesi büyüklüğünün belirlenmesi sağlanır. Ayrıca kümeleme algoritmasında kullanılabilecek kayıt tipleri, niteliklerinin sayıları gibi bilgilerin belirlenmesi işlemleri gerçekleştirilir.
b) Özellik seçimi: Kümeleme işleminde en uygun sonuçları bulmak için, veri setlerinden en uygun özellikler seçilerek kümele işleminde kullanılacak veriler oluşturulur.

c) Benzerlik yöntemi seçimi: Kümeleme de örüntü içerisindeki en önemli konulardan biri veri çiftlerinin birbirlerine olan benzerliklerinin ya da aykırılıklarının belirlenmesi için bir uzaklık fonksiyonu seçilmesidir. Kaynaklarda farklı uzaklık fonksiyonları kullanılmasına rağmen iki nokta arasındaki uzaklığın bulunması için en sık kullanılan fonksiyon Öklid uzaklık fonksiyonudur (Michalski ve ark 1983).

d) Kümeleme işlemi: Kümeleme temelde giriş verisi kesin sınırlarla kümelere ayrılacak şekilde kümelendirilenler ve her örüntü elemanının her kümeye ne kadar yakın olduğu belirlenerek bulanık olarak kümelendirilenler olmak üzere iki farklı şekilde gerçekleştirilmektedir.

e) Sonuçların özetlenmesi ve saklanması: Kümeleme sonuçlarının sunulması aşamasıdır. Uzman kişiler tarafından her kümeyi karakterize eden kuralların özeti çıkarılır. Ya da başka bir algoritmada giriş verisi olarak kullanılır (Diday ve Simon 1976).

3.6.3 Kümeleme Analizinin Uygulama Aşamaları

Kümeleme Analizinin uygulama aşamaları aşağıdaki gibidir (Günay Atbaş 2008);

1. Aşama: Değişkenlerin seçilmesi ve veri matrisinin belirlenmesi
2. Aşama: Öğelerin birbirleriyle olan benzerlik ya da uzaklıklarını gösterecek uygun bir benzerlik/uzaklık ölçüsü ile benzerlik/uzaklık matrisinin oluşturulması
3. Aşama: Uygun bir kümeleme tekniği ile benzerlik/uzaklık matrisine göre öğelerin uygun sayıda kümelere ayrılması
4. Aşama: Oluşturulan bu kümelerin yorumlanması, gerekli analitik yöntemlerin uygulanması.

3.6.4 Kümeleme Analizinde Değişken seçimi

Kümeleme analizinde değişkenler, nesnelerin tanımlanan özelliklerini kapsayarak gösterdikleri özellikleriyle karşılaştırılırlar. Bu yönüyle diğer çok değişkenli analiz yöntemlerinden ayrılırlar.

3.6.5 Uzaklık ölçüleri

Benzerlikler ve farklılıklar, öğeleri gruplamak ya da ayırmak için kullanılan bir takım kurallardır. Benzerlik kullanılırken belirlenen değişken değerleri aynı standartta alınmalıdır. Benzerlik ya da uzaklık ölçüleri tek boyutlu veya çok boyutlu yapılabilir. Her boyut öğeleri gruplandırmak için kullanılır (Günay Atbaş 2008).

3.6.5.1 Öklid uzaklığı

Uzaklık ölçütleri arasında en sık kullanılan ölçüttür. İki nokta arasındaki doğrusal uzaklık olarak tanımlanır. İki boyutlu düzlemlerde daha kolay uygulanır. Çok boyutlu uzayda hesaplama süresi artsa da uygulanabilir. Bu uzaklık ölçütünde değişkenlerin birbirinden bağımsız olduğunu varsaymaktadır. Ölçütün uygulanabilmesi için öğelerin aralıklı ölçütte ölçülmüş olması gerekir.

$$d_{ij} = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2} \qquad (3.10)$$

biçiminde hesaplanır (Tatlıdil 1996) ,burada

d_{ij} ; i. ve j. birimin birbirine olan uzaklığı

x_{ik} ; i. birimin k. değişken değeri

x_{jk} ; j. birimin k. değişken değeri

p değişken ve n birim sayısı olmak üzere; i =1, . . . , n ; j =1, . . . , n ve k =1, . . . , p'dir.

Karesel Öklid uzaklığı ise

$$d_{ij}^{\,2} = \sum_{k=1}^{p} (x_{ik} - x_{jk})^2 \qquad (3.11)$$

biçimindedir (Tatlıdil 1996).

3.6.5.2 Manhattan (City-Block) uzaklığı

Öğeler arasındaki mutlak uzaklıkların toplamını alarak hesaplayan bir uzaklık ölçüsüdür. Aşağıdaki gibi ifade edilir.

$$d_{ij} = \sum_{k=1}^{p} \left| x_{ik} - x_{jk} \right| \tag{3.12}$$

Manhattan uzaklık ölçüsünde en belirgin sorun olarak uzaklık ölçüsünün değişkenleri arasında ilişki (korelasyon) olmadığını varsaymasıdır. Eğer araştırma konusunda değişkenler arasında ilişki varsa anlamlı kümelemeler bulunamayacaktır (Johnson ve Wichern 1988).

3.6.5.3 Minkowski uzaklığı

Minkowski uzaklık ölçüsü Öklid ve Manhattan uzaklık ölçülerinin genelleştirilmiş halidir. Minkowski uzaklık ölçüsü aşağıdaki şekilde formüle edilmiştir.

$$d_{ij} = \left[\sum_{k=1}^{p} \left| x_{ik} - x_{jk} \right|^{q} \right]^{1/q} \tag{3.13}$$

Minkowski uzaklık ölçüsü q =1 için bir sonraki kesimde verilecek olan Manhattan uzaklık ölçüsüne, q = 2 için ise Öklid uzaklık ölçüsüne eşit olacaktır (Anderberg 1973).

3.6.5.4 Mahalanobis uzaklığı

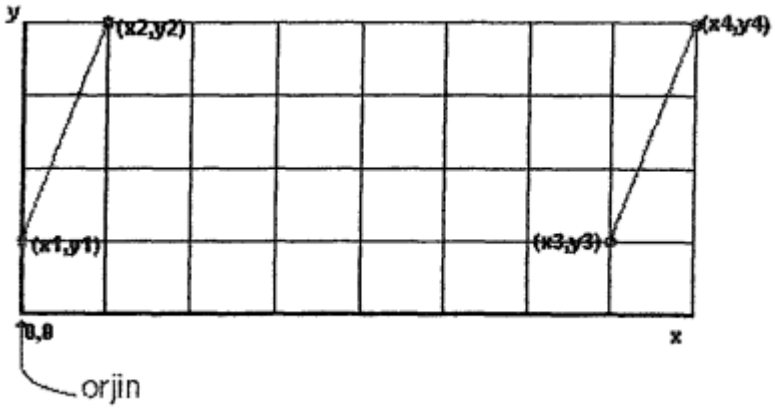
Diskriminant analiz içerisindeki olaylarda kullanılır. Özellik uzayındaki her grubun merkez noktası ile veri arasındaki uzaklıktır. Veriler her grup için bir uzaklık değerine sahip olur. Veriler, uzaklığa göre en çok benzer olduğu kümeye dâhil olur.

$$Md_{ik} = (x_i - x_k)' S^{-1} (x_i - x_k) \tag{3.14}$$

Bu uzaklık aykırı noktaları da hesaplaması bir avantajdır. Bu yönüyle uzaklık ölçüleri arasında avantajlı bir konuma sahiptir (Günay Atbaş 2008).

3.6.5.5 Canberra Uzaklığı

Benzer düzlemsel değerler arasındaki oranların düzeltilmesini yapmaktadır. İki nokta arasındaki uzaklık değerleriyle beraber bu noktaların orijin noktasıyla ilişkilerini de dikkate alır. Yapay sinir ağları ve bioinformatik gibi alanlarda kullanılır. Şekil 3.6'da Canberra uzaklık ölçüsünün şekilsel anlatımı gösterilmektedir.



Şekil 3.6 Canberra uzaklık ölçüsü(Erdoğan 2004)

3.6.6 Kümeleme Metotları

Kümeleme yöntemleri hiyerarşik kümeleme ve hiyerarşik olamayan kümeleme olarak iki ana grupta incelenebilir.

3.6.7 Hiyerarşik Kümeleme Metotları

Aşama sıralı kümeleme yöntemi olarak da bilinen bu yöntem gruplayıcı ve bölücü olmak üzere iki çeşidi mevcuttur (Hubert 1974) (Günay Atbaş 2008).

3.6.7.1 Gruplayıcı hiyerarşik yöntem

Bu yöntemde başlangıçta her veri, bir küme olarak kabul edilir. Yöntemin diğer adımında en yakın iki küme yeni bir kümede toplanarak birleştirilir. Her adımda küme sayısı bir azaltılır. Temelde bu algoritmaya dayanan önemli hiyerarşik kümeleme yöntemleri ise şunlardır (Ünlükaplan 2008) :

Tek Bağlantı Kümeleme Yöntemi: Kümenin oluşturulması, küme elemanları arasındaki en düşük uzaklık değeri temel alınarak yapılan yöntemdir.

Ortalama Bağlantı Kümeleme Yöntemi: Tek bağlantı ve tam bağlantı tekniği ile paralel olarak, kümeler arasındaki ortalama uzaklığın en düşük değerini temel alır.

Tam Bağlantılı Kümeleme: İki kümenin elemanları arasındaki en fazla uzaklığın minimum değerini temel alır.

Küresel Ortalama Bağlantı Kümeleme Yöntemi: Bir kümeyi oluşturan verilerin ortalamalarını esas alır. Kümede sadece bir merkez varsa onun değeri merkez olarak kabul edilir.

Ward Bağlantı Bağlantı Kümeleme Yöntemi: Minimum varyans yöntemi olarak da bilinen bu yaklaşım, bir kümenin ortasına düşen verinin, aynı kümenin içinde bulunan verilerden ortalama uzaklığını ele alır ve toplam sapma karelerinden yararlanır.

3.6.7.2 Bölücü hiyerarşik yöntem

Gruplayıcı hiyerarşik yöntemin tersi bir süreç işlemektedir. Bu yöntemde verilerin hepsinden oluşan büyük bir küme ile işe başlanır. Benzer olmayan veriler ayıklanarak daha küçük kümeler oluşturulur (Everitt ve ark 2001). Uygulamalarda çoğunlukla gruplayıcı hiyerarşik kümeleme yöntemi kullanılmaktadır.

3.6.8 Hiyerarşik olmayan kümeleme yöntemleri

Küme sayısının önceden belli olduğu, bu küme sayısına göre çözümler üretilen kümeleme yöntemidir. Kümeleme yapılmadan önce verilerin kaç kümeye bölüneceği hakkında ön bilgi veya görüş varsa çok uzun zaman alan hiyerarşik yöntemler yerine hiyerarşik olmayan yöntemler kullanılmaktadır (Anderberg 1973).

Hiyerarşik olmayan yöntemlerde belirli ilk kısım veya bir başlama noktası ile başlanır. İlk kısım ile başlanırsa ana verilerden belirli bir örneklem seçilir ve ilk kısım elde edilene kadar küme üyeleri değiştirilir. Bu yöntemin dezavantajı küme sayısının önceden bilinmesi gerektiği veya küme sayısının keyfi olarak seçilmesi gerekliliğidir (Blashfield ve Aldenferder 1978).

Bu kümeleme yöntemi altında çeşitli algoritmalardan söz edilebilir ancak bunlardan en sık kullanılanı k-ortalamalar algoritmasıdır.

3.6.8.1 k-ortalamalar Algoritması

Kümeleme algoritmaları arasında en sık kullanılan algoritmalardan biri olan k-ortalamalar algoritması J. MacQueen tarafından 1967'de sunulmuştur. Algoritma çalışmaya başlamadan K adı verilen sabit küme sayısına ihtiyaç duymaktadır. Kümeler, elemanları birbirine yakınlıklarına göre, her verinin sadece bir kümeye ait olabilmesine izin veren k adet kümeden oluşur.

Algoritma bir hiyerarşik olmayan kümeleme algoritmasıdır. Algoritma, tekrarlı bölümleyici yapısıyla, her veri nesnesinin ait olduğu kümeye olan uzaklıkları toplamını küçültmektedir. Büyük ölçekli verileri hızlı ve etkin şekilde kümeleyebilmesinden dolayı da sık kullanılmaktadır (Kohei ve Barakbah 2007).

Algoritmada öncelikle ilk küme merkezlerini temsil edecek k adet rasgele nokta belirlenir. Daha sonra gelen her veri k adet küme merkezinden hangisine yakın ise o kümeye dâhil edilir. Eklenen verilere göre küme merkezinin ağırlığı tekrar hesaplanarak yeni bir küme merkezi belirlenir. Bu küme merkezi diğer işlemlerde kümeyi temsil eder. Küme elemanları değişmeyene kadar bu işlemler devam eder.



Şekil 3.7 K-ortalamalar akış şeması(Sevinç 2009)

k-Ortalamalar algoritmasının adımları (Sevinç 2009)

1-)Veriler içinden, küme merkezlerini temsil edecek olan k adet rastgele nokta (veri) belirlenir. k dışardan kullanıcıdan alınan bir parametredir. Dışarıdan alınan k adette küme merkezi seçilir. Küme merkezleri genelde birbirinden uzak rastgele değerlerden seçilir. İlk küme merkezlerini oluşturan bu elemanlar tek verili başlangıç kümelerini oluştururlar (MacQueen 1967).

2-) Kümeler arasındaki sınırlar belirlenerek küme merkezi haricinde kalan elemanlar kendisine yakın olan kümeye dâhil edilir. İki boyutlu verilerde iki küme merkezi bir doğruyla birleştirilip, doğrunun orta noktasından geçen ve doğruyu dik kesen başka bir doğru daha geçirilir, bu doğru iki kümenin sınırı olarak kabul edilir. Boyut sayısı arttığında doğru yerine düzlem kullanılır. Bilgisayar programları ile geliştirilen k-ortalamalar algoritmalarında ise, noktalar arasındaki uzaklıklar hesaplanarak, noktaların merkeze yakınlığı dikkate alınmaktadır. Uzaklık hesaplamalarında genellikle Öklid uzaklık bağlantısı kullanılmaktadır.

3-) Elemanlar, küme merkezlerine yakınlıklarına göre dağıldıktan sonra, her bir boyutundaki bütün küme elemanlarının ağırlıklı ortalaması hesaplanarak yeni küme merkezleri bulunur. İkinci oluşan küme merkezleri, ilk elemanlar gibi verinin içinden bir değer değildir. Algoritmanın her döngüsünde elemanlar yeni oluşan küme merkezlerine göre küme değiştirebilirler. Değişim işlemi küme sınırlarının değişimi bitene kadar devam etmektedir. Algoritmaya ait akış şeması Şekil 3.7'de görülmektedir.

3.7 Normalizasyon Metotları

Normalizasyon yapay zekâ performansını artırmak için mxn veri boyutuna sahip bir veri setini bir uzaydan başka bir uzaya taşıyan oldukça popüler bir ön işleme tekniğidir. Normalizasyon işlemi yapay zekânın transfer karakteristiğine uygunsa, yeni maximum ve minimum noktalar oluşturarak veri setinin mxn olan boyutunu değiştirmeden performansını artıran bir çözüm sunar. Yapay sinir ağı tabanlı uygulamalarda bu dönüşümle çok önemli performans artışları gözlenmektedir (Widrow ve A. 1990) (Akdemir 2009).

YSA transfer fonksiyonlarından en yaygın olarak kullanılan LOGSIS' de verilerin normalizasyonu -1 ile 1 veya 0 ile 1 arasında olması mecburiyeti getirilmiştir (Vong ve ark 2005).

Veri setlerine ve uygulanacak algoritmaya göre normalizasyon uygulamaya karar verildikten sonra uygulanacak normalizasyonun seçilmesi gerekir. Daha sonra veriler üzerine nasıl uygulanacağına karar verilmesi gerekir. Sayısal olmayan verilerden, sayısal ortama dönüştürülebilen verilerde normalizasyon işlemine tabi tutulabilir. Örneğin; Az-çok, iyi-kötü, hasta- hasta değil gibi veriler 0-1 gibi ifadelerle sayısal bilgilere dönüştürülerek kullanılabilir.

Normalizasyon türüne bakılmadan farklı özellikleri olan veri setlerinde, veriler sütun sütun birbirlerinden bağımsız olarak normalize edilmektedir. Bu tür verilerde her sütun ayrı bir özelliği gösterir ve özellikler ayrı ayrı normalize edilir.

3.7.1 Literatürde Kullanılan Normalizasyon Metotları

Birçok normalizasyon metodu olması rağmen literatürlere bakıldığında yaygın olarak üç tanesinin kullanıldığı görülür. Bunlar (Akdemir 2009);

- Minimum-maksimum metodu
- Ondalık ölçekleme metodu
- Z-skor metodu

dur.

3.7.1.1 Minimum-maksimum metodu

Minimum-Maksimum metodu, verileri doğrusal olarak normalize eder. Minimum; veri sütununda bir verinin alabileceği en düşük değer iken, maksimum; veri sütununda verinin alabileceği en yüksek değeri ifade eder. Bu metot çoğunlukla mühendislik çalışmalarında, mühendislik çalışmalarında kullanılan tıbbi veriler, görüntü verileri gibi veri setlerinde de yaygın olarak kullanılmaktadır. Minimum-maksimum modelinde yeni değerler Denklem 3.15'te hesaplanabilir.

$$v' = \frac{v - \min_A}{maks_A - \min_A}(yeni_maks_A - yeni_\min_A) + yeni_\min_A \quad (3.15)$$

Burada

v : Giriş verisini

v' : Normalize edilmiş bilgiyi

min_A : Veri içerisindeki minimum değeri,

maks_A :Veri içerisindeki maksimum değeri

yeni_maks_A :Verinin taşınacağı aralığın üst değeri

yeni_min_A :Verinin taşınacağı aralığın alt değeri

göstermektedir. Burada veri eğer 0 ile 1 arasına taşınmak isteniyorsa yukarıdaki ifade Denklem 3.16 'ya dönüşür.

$$v' = \frac{v - \min_A}{maks_A - \min_A} \quad (3.16)$$

Bu metot veri setlerinin içindeki iki değer arasındaki orantısallığı değiştirmez. Sayısal veriler içerisinde negatif değerler olsa bile normalizasyon sonunda ortadan kalkar ve değerler pozitif hale gelirler.

3.7.1.2 Ondalık ölçekleme metodu

Ondalık ölçekleme metodu, sayıları 1 den küçük hale getirmek için 10 veya 10'un kuvvetlerine bölme tekniğine dayanmaktadır. Denklem 3.17'den de görülebileceği üzere ondalık normalizasyonun tanımı j de yatmaktadır ve j, A değerini birden küçük yapan en küçük değer olarak tanımlanabilir.

$$A' = \frac{A}{10^j} \tag{3.17}$$

3.7.1.3 Z-skor metodu

Z-skor metodunun uygulanabilmesi için verilerin standart sapma ve ortalama değerleri ilk olarak hesaplanması gerekir. Z-skor verilerin ortalama değeri ile bu değerlerin ortalamadan olan uzaklıkları ile ilgili yeni değerler üretir (http://en.wikipedia.org/wiki/Standard_score 2015). Z-skor normalizasyondan sonra ortalaması sıfır olan bir gauss dağılımı sunar. Standart sapma, varyansın kareköküdür ve Denklem 3.18'daki gibi hesaplanabilir.

$$S = \left(\frac{1}{n-1} \sum_{i=1}^{n} \left(x_i - \bar{x} \right)^2 \right)^{1/2} \tag{3.18}$$

3.8 Örüntü

Olay veya nesnelerin sistemli bir biçimde birbirini takip ederek gelişmesi olarak ifade edilir (Türk Dil Kurumu 2008). Örüntüyü, kendini tekrarlayan olay veya nesnelere karşılık gelen ifade olarak da belirtebiliriz. Örüntü tanımanın en önemli amaçları; bilinmeyen örüntü sınıflarına belirli bir şekil vermek ve bilinen bir sınıfa ait olan örüntüyü teşhis etmektir (Turkoğlu 2002). Örüntüye örnek vermek gerekirse; parmak izi, ses sinyali, insan yüzü, retina, bir metin içerisindeki karakterler ve biyomedikal cihazlardan elde edilen görüntüler örüntüye dair bazı örneklerdir.

Günlük hayatımızda örüntü tanıma olayını geçmiş tecrübemiz ve yaşantımız ile sürekli gerçekleştirmekteyiz. Örüntü tanıma alanı uzun bir çalışma tarihine sahiptir. İstatistik alanında teorik olarak 1960'ların öncesine kadar çalışılmıştır. Teknolojinin

gelişmesi ile son yıllarda bu alanda, ortak özelliğe sahip veya aralarında bir ilişki kurulabilen karmaşık olayları ve nesneleri çeşitli yöntemler vasıtasıyla tanıma ve sınıflandırmanın sağlanabilmesi için makina ortamında örüntü tanıma kavramı ortaya çıkmıştır. Günümüz teknolojisi ile örüntü tanıma görevini makinelere de yaptırılmaktadır. Son yıllarda yapılan uygulamalarla bu teknoloji hızlı bir büyüme sergilemiştir.

Örüntü tanıma; öznitelik çıkarımı, öznitelik seçimi ve sınıflandırma olmak üzere üç ana öğeden oluşmaktadır. Bu öğelerin yanında eldeki verilere birtakım, önişlem (gürültü azaltma v.b.) ve son işlemlerde süreç içerisinde yer alabilir (Karabulut 2009).

3.9 Öznitelik Çıkarımı

Bir boyut indirgeme işlemi olarak ifade edilir. Elde edilen örüntüler sınıflandırma için kullanılacak bilgiden genellikle fazlasını içerirler. Bu durum, sınıflandırmanın süresini uzattığı gibi hassasiyetini de düşmesine sebep olur. Sınıflandırmaya zararlı etkisi olan bu durumu önlemek için örüntüyü daha düşük miktarda temsil edilen başa bir veriye çevrilir. Örüntüyü temsil eden gereksiz verinin çıkarılıp, tüm veriden daha az sayıdaki karakteristik verinin elde edildiği bu değişime öznitelik çıkarımı adı verilir.

Kendine özgün özniteliklerin çıkarımı ayırt edici ve veriye uygun en az sayıda olması, tanıma sürecinin daha kolay sınıflandırıcılarla, daha yüksek hassasiyetle ve daha kısa zamanda yapılmasını sağlar.

3.10 Öznitelik Seçimi

Örüntüyü temsil etmeyen bilgilerin elenip, özgün özelliklerin elde edilmesiyle belli oranda boyut indirgeme öznitelik çıkarımı ile sağlanır. Bu öznitelik çıkarımı işleminden sonra da sınıflama işlemi için verinin özniteliklerinin hepsi gerekli olmayıp, önemli özniteliklerin seçilmesi kullanılan sınıflandırma algoritmalarının daha az bellek kapasitesiyle daha hızlı çalışmasını sağlayacaktır. Öznitelik seçimindeki amaç, netliklerin ayırt edicilikleri incelenerek olan öznitelik kümesinden daha ayırt edici, yüksek ayırım gücü ve yüksek güvenilirliğe sahip esas bilgiyi koruyan bir altküme

oluşmasını sağlanmaktır. Örüntüyü temsil eden öznitelik seçim yöntemi problemin çözümünü hafifletmektedir (Theodoridis ve Koutroumbas 2003, Günal 2010) .

Öznitelik seçim işlemi; özniteliklerin ayırt ediciliklerini belirlemede kullanılan değerlendirme ölçütü ve arama taktikleri olmak üzere iki bileşenden oluşur Tüm öznitelikler kümesi N öznitelik içeriyorsa, aday alt kümelerinin toplam sayısı 2N'dir. Bu nedenle özniteliğin fazla bulunduğu durumlarda tüm bilgiler üzerinde arama uygun olmayabilir. Bu gibi durumlarda, en uygun öznitelikler alt kümesinin bulunması için çeşitli araştırma teknikleri vardır. Tüm veriler üzerinde aramaya göre çok daha kısa olan ileri ya da geri yönde ilerleme prensibine dayalı arama teknikleri tercih edilir. Ardışık ileri yönde seçim (SFS), Ardışık Geri Yönde Seçim (SBS) ve Ardışık İleri Yönde Kayan Seçim (SFFS) bunlardan bir kaçıdır. Bu metotlarda değişim sadece alt kümeden özniteliklerin silinmesi veya eklenmesi yoluyla gerçekleşir. Ardışık İleri Yönde Seçim (SFS), elemanı olmayan bir küme ile başlar ve her döngüde küme içerisine eklemek için en iyi özellik seçilir. SFS için kaba kodu Şekil 3.8'de verilmiştir. Burada x^+ ; amaç fonksiyonuna göre en iyi özelliği temsil eder, Y_k seçilmiş özelliklerin tutulduğu kümeyi temsil eder. Burada x^- ; amaç foksiyonuna göre en kötü özelliği temsil eder. Bunun tam tersi Ardışık Geri Yönde Seçim (SBS), tüm özniteliklerin olduğu kümeden başlar ve her bir döngü de küme den en kötü bir özelliği siler. SBS için basitleştirilmiş kodu Şekil 3.9'da verilmiştir. SFS algoritması genişletilerek aşağıdan yukarı araştırma metodu SFFS algoritması geliştirilmiştir. Bu algoritmada geri yönde izleme dinamik olarak kontrol edilerek alt küme sonuçları bir öncekinden daha iyi sonuçlandığı sürece her ileri adımdan sonra bir geri adım uygulamasını içerir. Sonuç olarak performans geliştirilemezse geri adım yoktur. Benzer uygulama Ardışık Geri Yönde Kayan Seçim (SBFS) prosedüründe de geçerlidir. Bu algoritmada parametre ayarlarına ihtiyaç duyulmaz (Pudil ve ark 1994).

Adım 1. Boş küme $Y_0 = \{\varnothing\}$ ile başlayın
Adım 2. Sonraki en iyi özelliği $x^+ = \arg\max[J(Y_k + x)]; x \notin Y_k$ ile seçin
Adım 3. Güncelleyin $Y_{k+1} = Y_k + x^+; k = k + 1$
Adım 4. Adım 2'ye gidin

Şekil 3.8 SFS için basitleştirilmiş kod (Ladha ve Deepa 2011)

Adım 1. Tam küme $Y_0 = X$ ile başlayın
Adım 2. En kötü özelliği $x^- = \arg\max[J(Y_k - x)]; x \notin Y_k$ ile kaldırın.
Adım 3. Güncelleyin $Y_{k+1} = Y_k - x^-; k = k + 1$
Adım 4. Adım 2'ye gidin

Şekil 3.9 SBS için basitleştirilmiş kod (Ladha ve Deepa 2011)

Öznitelik seçimi için literatürde süzgeçlemeden, sarıcı yaklaşımlara (Kohavi ve John 1997) kadar pek çok yöntem bulunmaktadır. Bu yöntemlerin arasında küçük ve orta büyüklükteki öznitelik kümeleri için bile oldukça yüksek işlem süresi olsa da yalnızca tam kapsamlı arama (ES) ve dal-ve-sınır yöntemi (Narendra ve Fukunaga 1977) en iyi öznitelik altkümesine ulaşmayı garantiler. İşlem süreleri yüzünden daha kısa sürelerde sonuç veren alt en iyi öznitelik seçim yöntemleri olan Ardışık İleri Yönde Seçim (SFS) (Whitney 1971), Ardışık Geri Yönde Seçim (SBS) (Marill ve Green 1963), Artı l – Çıkar Seçim (PTA) (Stearns 1976) , Ardışık İleri Yönde Kayan Seçim (SFFS) , Ardışık Geri Yönde Kayan Seçim (SBFS) (Pudil ve ark 1994), ve Genetik Seçim (Siedlecki ve Sklansky 1989) (Yang ve Honavar 1998)daha çok ön plana çıkmaktadır (Günal 2008).

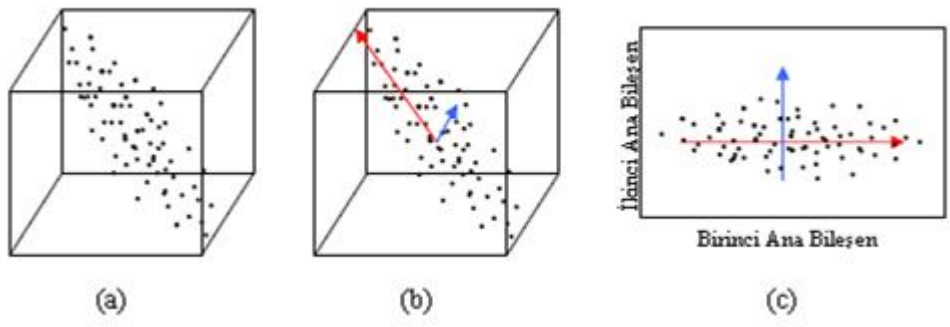
Öznitelik seçim yöntemlerinin tamamı, özniteliklerin ilgililiğine ya da ilgisizliğine karar verebilmek için bir ölçüt fonksiyonu (ayrılabilirlik ölçüsü) kullanır. Yaygın olarak tercih edilen ölçütler, hatalı sınıflandırma olasılığı, ve Uzaksaklık, Dönüşmüş Uzaksaklık, Bhattacharyya, Jeffries – Matusita gibi olasılıksal uzaklık ölçüleridir (Webb 2002, Theodoridis ve Koutroumbas 2003). Hatalı sınıflandırma olasılığı, sınıflandırıcı türüne bağlı bir ölçüttür. Bu ölçüt ile özniteliklerin seçiminde, hatalı sınıflandırma olasılığını en küçükleyecek öznitelik altkümesine ulaşılması hedeflenir (Günal 2008).

3.11 Temel Bileşen Analizi

Temel Bileşenler Analizi ilk kez 1901'lü yılların başında Karl Pearson tarafından çalışılmaya başlanmış, 1933 yılında Hotelling tarafından korelasyon yapılarının analizinde belirli amaçlar için geliştirilmiştir. 1938 yılında Girschick, 1958 yılında Anderson temel bileşen dağılımları ve örnek özelliklerini incelemişlerdir.

Çok değişkenli istatiksel analizde *n* tane nesneye ilişkin *p* tane özellik incelenmektedir. Bu özelliklerin birbiriyle ilişkili ve değişken sayısının *(p)* çok fazla olması sebebiyle küçültmek ve böylece matematiksel işlemlere uygun hale getirmek için uygulanan yöntemler den bir tanesi de temel bileşen analizidir. Başka bir ifade ile bir değişken kümesini, büyük oranda temsil eden daha küçük boyuttaki bir veri kümesine doğrusal olarak dönüştürebilen bir istatistiksel analiz tekniğidir (Bozik 2011).

Değişken sayılarının yani özelliklerin çok fazla olması, değerlendirmeleri güçleştirdiğinden temel bileşenler analizi, değişkenler arasındaki korelasyon yapısının yok edilmesi ve/veya veri içerisinde var olan değişimlerin mümkün olduğunca korunarak daha az boyuta indirgenmesi amacıyla kullanılan bir tekniktir. Ayrıca bir analiz yöntemi olduğu gibi, diğer analiz yöntemleri için veri hazırlama tekniği olarak da kullanılmaktadır.



Şekil 3.10 Ana Bileşenler analizi yönteminin grafiksel gösterimi (Bozik 2011)

İşlem yapılacak veri setini Şekil 3.10(a)'da gösterildiği gibi bir küme oluşturduğu düşünülürse, bu veri setini karakterize etmek için temel bileşim analizi uygulandığında temel bileşenler, Şekil 3.10(b) ve Şekil 3.10(c)'de görüldüğü gibi veri yoğunluğunun yüksek olduğu eksenler boyunca uzanacaktır. İlk ana bileşen veri setini büyük oranda tanımlayan yatay eksen boyunca çizilen doğrudur. Diğer bileşen ise ilk bileşene dik olan veri setine dikey olarak tanımlanan eksen olarak tanımlanarak veri seti iki boyuta getirilmiş olur (Bozik 2011).

Temel bileşen analizi uygulandıktan sonra elde edilen veri seti, ilk veri setinin temel bileşenleri olmaktadırlar. Varyans değeri en büyük olan vektör ilk temel bileşen olmaktadır. Diğer bileşenlerde varyans değerlerinin büyüklüklerine göre dizilirler. Temel bileşen analizinin avantajı, hafıza ve kapasite ihtiyaçlarının azalması, gürültüye karşı duyarlılığın azalması olarak gösterilir.

3.11.1 Kovaryans

Temel bileşenler analizinde, özvektör ve özdeğerlerin bulunması aşamasında kullanılır. İki değişkenin birlikte ne kadar değiştiği hakkında bilgi verir. İşlem yapılacak verilerin ikiden fazla ise bu değişkenlerin ikili kombinasyonları şeklinde tanımlanır.

Denklem 3.19'de x vektörüne ait kovaryans matrisi gösterilmiştir. Burada ele alınacak değişken X vektörü ile X vektörünün ortalaması ise X_m şeklinde tanımlanmıştır.

$$C_x = E((x - x_m)(x - x_m)^T) \tag{3.19}$$

Kovaryans matrisleri, öz değerleri pozitif gerçel sayılardan oluşmaktadır (Çatalbaş 2014).

Çok boyutlu uzaydaki bir noktayı, düşük boyutlu uzaya indirgemek için hatanın özdeğer, özvektör ve kovaryans matrisiyle ilişkisi matematiksel olarak aşağıdaki biçimde ifade edilir.

$x \in R, x = (x_1, x_2, ...x_n)$ noktalar ifade edilir. Taban vektör kümesi $u = (u_1, u_2, ...u_k)$ birimdik temelli düşünüldüğünde; *i* ve *j*'nin eşit olmadığında $u_i.u_i = 1$ ve $u_i.u_j = 0$ taban seçimi yapılır. Uzayın göreli konumunu tanımlanarak merkezi seçilir. Düşük boyutlarda noktasal çarpımlarla tanımlanarak en iyi koordinatlar bulunur: $(z_1, z_2, ...z_k)$

$$z_i = \left(x - \bar{x}\right).u_i \tag{3.20}$$

Ortalama karesel hatayı (MSE) en küçük olduğu, en büyük değişikliğe sahip Z_i doğrultusunu bulunması için çalışılır (Yalçın 2010).

3.11.2 Temel Bileşenler Analizinin Adımları

X veri matrisi, n adet değişken ve m adet gözlemden oluşan çok sayıda noktadan oluşan bir topluluktur. Değişkenler arasında tam bağımsızlık olsa ve eksenler birbirine dik olsa daha fazla bilgiye sahip olunur. Fakat gerçek hayatta elde edilen bilgiler birbirinden tam bağımsız değildir. Bu amaçla, eksenlerin birbirine dik olması için

dönüştürme yapılarak noktaların sahip oldukları toplam değişinti ilk eksen boyunca değişmemesi sağlanarak yeni eksenlerin birbirine dik olması sağlanır.

1. Evre: Veri, X mxn boyutlu matris haline getirilir.

X veri matrisi, n adet değişken ve m adet gözlemden oluşan çok sayıda noktadan oluşan bir topluluktur.

2. Evre:

Sütunlar, verilerin sıfır ortalaması olacak şekilde düzenlenir.

$$X_c = X - \overline{X} \tag{3.21}$$

X veri uzayında bir noktaya veri değişken ortalaması denir. Temel bileşenlerin tümü ortalama değerinden geçer. Bu ortalama değeri eşitlik (3.21)'de ifade edildiği gibi veri matrisinden çıkarılır.

3.Evre: Kovaryans matrisi elde edilmek için hesaplanır.

$$\sum = \frac{1}{m}\sum_{i=1}^{m}\left(X - \overline{X}\right)\left(X - \overline{X}\right)^T \tag{3.22}$$

4.Evre: Hesaplanan Kovaryans matrisinin ($\sum$) öz vektörü ve öz değerleri bulunur.

a) Öz değer ve öz vektör Kavramı

$T : V \to V$ Doğrusal dönüşümü verilsin.

$v \in V$, olduğunda sıfırdan farklı bir x vektörü için T (v) = λ v denklemini doğrulayan bir λ sayısı varsa, λ sayısına T dönüşümünün öz değeri, v vektörüne de λ öz değerine karşılık gelen öz vektörü denir (Lay 2000). Öz değer ve öz vektör arasındaki bağıntılar Denklem 3.23-3.25 arasındaki denklemlerde gösterilmektedir.

X kare matrisi ise aşağıdaki denklemler uygulanır;

$$Xv = \lambda v \tag{3.23}$$

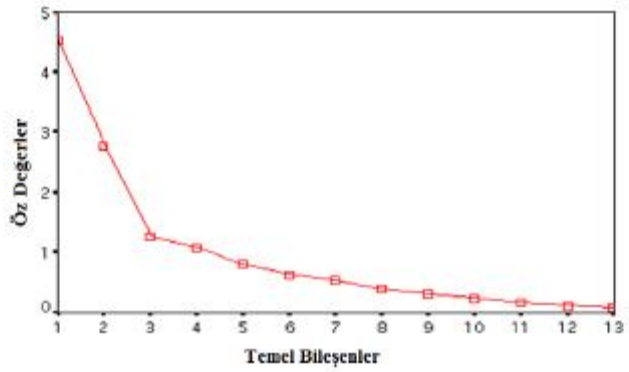
$$Xv - \lambda Iv = 0 \tag{3.24}$$

$$(X - \lambda I)v = 0 \qquad (3.25)$$

v sıfıra eşit olmadığı durumda $X - \lambda I=0$ olur.

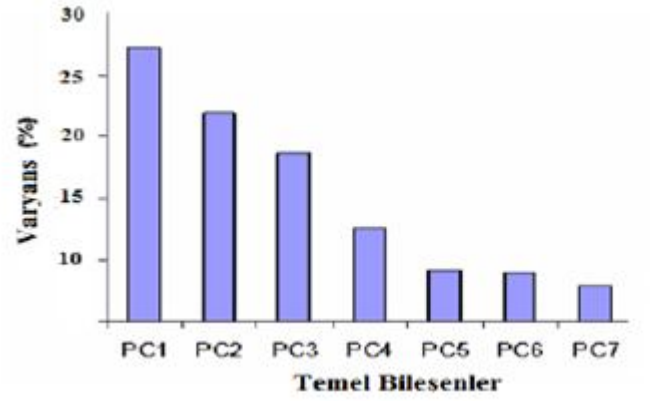
5. Evre: Temel Bileşenler belirlenir: Temel bileşen olarak yüksek öz değerli öz vektörler seçilir (Yalçın 2010).

A) 1.yöntem: Öz değerlerin genliklerine göre öz vektörler sıralanır. Değişinti düzeyi %80-90 olan öz vektörler temel bileşen olarak seçilir. Şekil 3.11'de görüldüğü gibi, temel bileşenlerin sayısı arttıkça, öz değerlerin genlikleri azalır.



Şekil 3.11 Öz değerler ve temel bileşenler(Yalçın 2010)

Şekil 3.12'de değişinti yüzdesi yüksek düzeyde olan öz değerlere ait öz vektörler temel bileşen olarak belirlendiği görünmektedir.

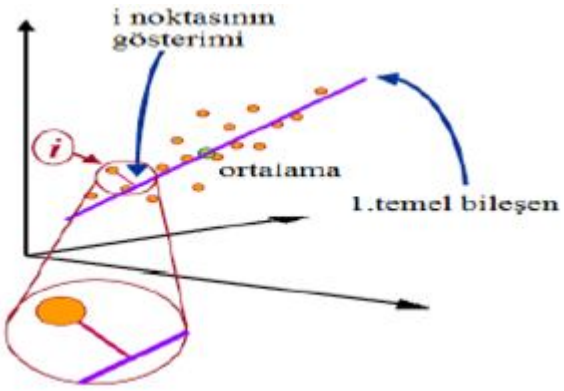


Şekil 3.12 Temel bileşenlerin varyans yüzdesine göre sıralanması(Yalçın 2010)

B) 2.yöntem: öz değerlerin, birden büyük olanlarının sayısı m, değişken sayısını da n olarak alırsak; $\sum_{i=1}^{m} \frac{\lambda_i}{n} \geq \frac{2}{3}$ koşulunun sağlandığı en küçük m değeri temel bileşen sayısını önemli görülenini belirlemektedir (Yalçın 2010).

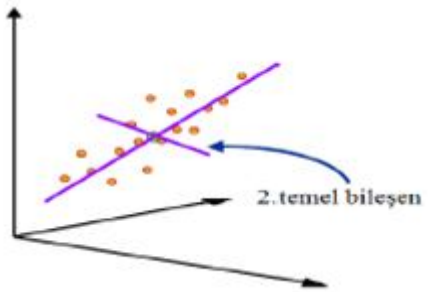
3.11.3 Temel Bileşenler Analizinin Geometrik Anlamı

1.temel bileşenin Şekil 3.13'te veri dağılımına bakıldığında veri değişimi maksimum olduğu ve yüksek değişintili veri dağılımına sahip olduğu görülür. Bu eksen ortalama değerinden geçmektedir. Temel bileşen ekseninin yönü yük vektörü (p1) ile belirlenir. Noktaların konumu ise skorlar (ti) belirlenir.

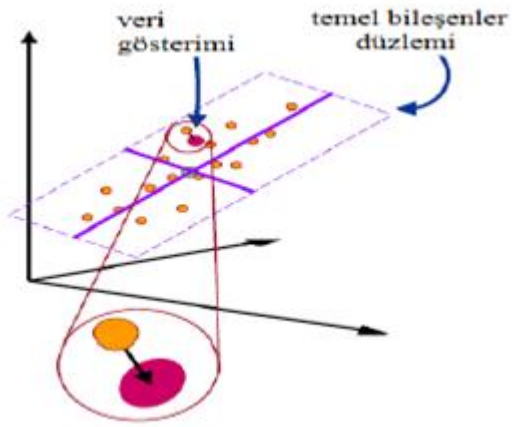


Şekil 3.13 Temel bileşen ve herhangi bir noktanın gösterimi(Yalçın 2010)

Şekil 3.14'de görüldüğü gibi 1.temel bileşene dik, veri dağılımı merkezden geçen 2.temel bileşen bulunur. Bu bileşenin temel bileşen ekseni boyunca veri değişimi 1.'ye göre daha sınırlıdır (Yalçın 2010) . Temel bileşen düzlemenin oluşumu Şekil 3.15' de gösterilmiştir.



Şekil 3.14 Temel Bileşen(Yalçın 2010)



Şekil 3.15 Temel bileşen düzlemi(Yalçın 2010)

3.12 Sınıflandırma

Sınıflandırma, yeni bir nesnenin özelliklerini açıklamak ve bu yeni nesnenin daha önceden tanımlanmış gruplardan birine atanmasını yapmaktır.

Sınıf değerleri { 0, 1 } ,{doğru, yanlış} gibi farklı etiket değerleri alarak kümenin bir sınıfa üye olması, diğer sınıftan farklı olduğunu gösteren iki farklı etikete sahip olur. Bu tanım sadece ikili sınıflandırma problemi için belirtilmiş olmasına rağmen uygun sayıda farklı etiketler belirtilerek çoklu sınıflandırma için de tanım düzenlenebilir (Hamel 2009).

$L(y, f(x,\omega))$ negatif değer alamayan kayıp (loss) fonksiyonuyla öğrenme makinesinin kalitesi ölçülür. Küçük pozitif değerler iyi bir yaklaşmayı gösterir. Değer büyüdükçe yaklaşma zayıflar.

Denklem 3.26 ile "risk fonksiyoneli" olarak adlandırılan kaybın beklenen değeri belirlenir.

$$R(\omega) = \int L(y, f(x,\omega))dP(x,y)dxdy \tag{3.26}$$

parametre kümesine bağlı R(ω) risk değeri seçilen yaklaşan fonksiyonu için kullanılır.

P (x,y) olasılık dağılımını bilmeden eğitim kümesini kullanarak öğrenme makinesi tarafından desteklenen fonksiyonlar kümesi üzerinde R(ω) risk fonksiyonelini minimize edecek $f(x,\omega_0)$ dayanak fonksiyonunu tahmin süreci öğrenme olarak tanımlanır. Belli sayıdaki eğitim verisi ile $f(x,\omega_0)$ fonksiyonunun sonucu kesin olarak belirlenemeyeceğinden, eğitim kümesinden bir öğrenme yöntemiyle $f(x,\omega^*)$ yaklaşan fonksiyonu kullanılır (Kantardzic 2011).

İkili sınıflandırma probleminde sistemin çıktısı sadece sınıflara karşılık gelen y = { — 1, + 1 } veya y = { 0, 1 } gibi değerleri alabilir. Bu nedenle, $f(x,\omega), \omega \in \Omega$ fonksiyonlar kümesi "gösterge (indicator)" fonksiyonlar kümesi haline gelir.

$$L(y, f(x,\omega)) = \begin{cases} 0, & y = f(x,\omega) ise \\ 1, & y \neq f(x,\omega) ise \end{cases} \tag{3.27}$$

İkili sınıflandırma problemi için yaygın olarak kullanılan kayıp fonksiyonu sınıflandırma hatasını ölçer (Cherkassky ve Mulier 2007).

(Denklem 3.26) de gösterilen R(ω) risk fonksiyonelinde kullanıldığında, risk fonksiyoneli yanlış sınıflandırma olasılığını (3.27) de gösterilen kayıp fonksiyonu ölçer. Öğrenme işlemi, yalnızca eldeki eğitim verisiyle, yanlış sınıflandırma olasılığını minimize eden $f(x,\omega_0)$ gösterge fonksiyonunun tahmini problemi haline gelir (Kantardzic 2011).

Sınıflandırma oranlarının ölçümünde verilerde bazı bağımlılıklar yoksa her bir sınıflandırma kararının birbirinden bağımsız ve eşit önemde olduğu kabul edilir.

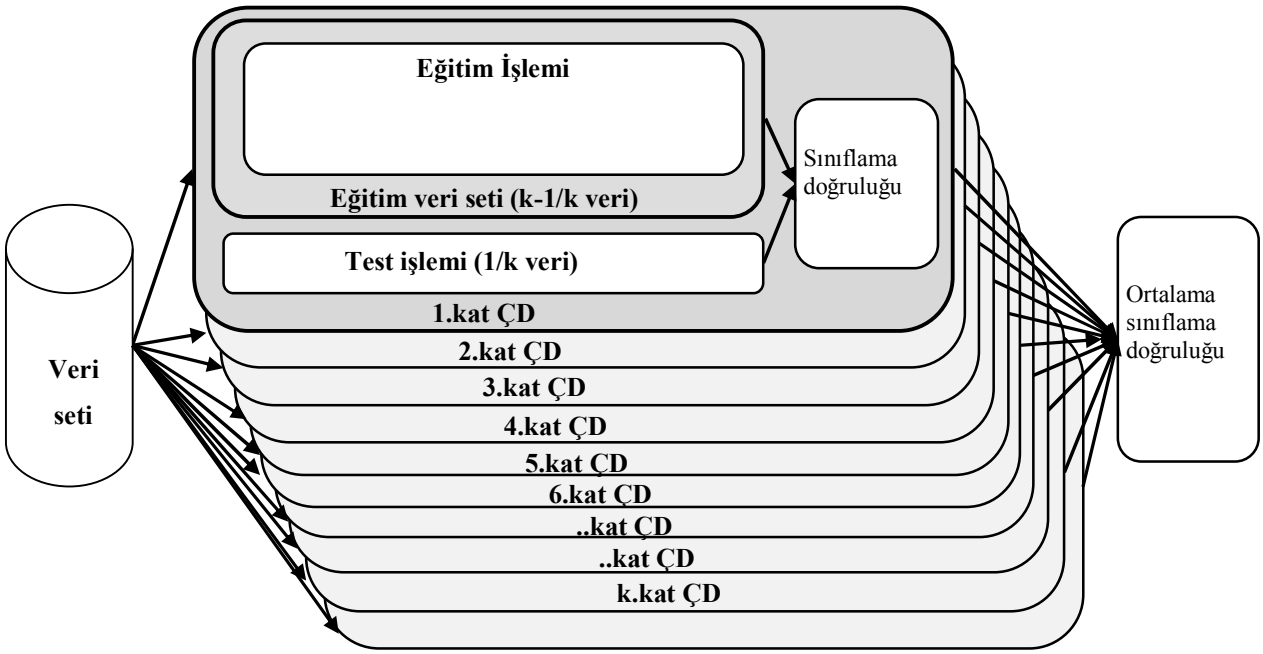
Bağımlılık varsa bu kabule uyulmayabilir. Verinin içinde bir grup olduğu görülürse, bu grubun bitip diğer grubun başladığı noktayı tespit etmeye çalışarak daha yüksek doğruluk oranına ulaşmaya çalışır.

Uzman bir doktor soğuk algınlığına yakalanmış hastalar sıklıkla hastaneye gelmeye başladığında birkaç hastayı gördükten sonra, büyük olasılıkla bir sonraki hastayı belirtileri önceki hastalar kadar açık olmasa bile soğuk algınlığı olarak sınıflandırmaya eğilim gösterir. Bu gibi bağımlılıklar sıklıkla meydana gelebilmektedir (Dietterich 2003).

Sınıfların eşit önemde olması gerekliliğini ihmal edilebilecek durumlar ortaya çıkabilir. Örneğin bir sınıflandırıcının laboratuvar sonuçlarına göre bir kişinin hasta olup olmadığını sınıflandırdığını varsayarsak, kişinin hasta oluğu halde hasta değil şeklinde sınıflandırmanın, kişinin hasta olmamasına rağmen hasta gibi sınıflandırılmasından daha maliyetli olduğundan dolayı öğrenme algoritması daha az "hasta olabilecek bir kişiyi sağlıklı" tespit etmeye eğilimli olarak tasarlanabilir.

3.12.1 Sınıflama başarılarının Ölçülmesi

Sınıflama başarılarının ölçülmesi için en yaygın ve güvenilir yöntemlerden birisi *k*-kat çapraz doğrulama yöntemidir. Bu çapraz doğrulama yönteminde veri kümesi X ile ifade edersek, k adet eleman sayısı yaklaşık aynı olan ve içinde rastgele atanmış $X_1, X_2, \ldots, X_k$ alt kümeye bölünür. Sınıflandırma işlemi k-1 adet veri eğitim için ayrılır. Bir tane alt kümede test işleminde kullanılır. Test için ayrılan bir alt küme her seferinde değiştirilerek k kere sınıflandırma işlemi yapılır. Çıkan doğruluk değerlerinin aritmetik ortalaması alınarak sınıflandırma doğruluğu bulunmuş olur. K- kat çapraz doğrulama sisteminin temsili grafiksel gösterimi Şekil 3.16'da gösterilmiştir. Şekil 3.17'de basitleştirilmiş kodu verilmiştir.



Şekil 3.16 *k*-kat çapraz doğrulama yönteminin grafiksel gösterimi

Eğitim ve test işlemi aşağıdaki gibi yapılmaktadır:

Adım 1 l=1 yapılarak bir rastgele alt küme seçilir.

Adım 2 l. Seçilen alt küme hariç diğer kümeler eğitim verisini oluşturur.

Adım3 Oluşturulan eğitim kümesi ile sınıflandırıcının eğitim işlemi yapılır.

Adım 4 geride kalan Xl kümesiyle test işlemi yapılır. Doğruluk oranı bl kaydedilir.

Adım 5 l bir artırılır.

Adım 6 l ≤ k ise adım 2-5 arası devam edilir.

Adım 7 b lerin toplamı k ya bölünürek ortalama değer bulunur.

Şekil 3.17 k-kat çapraz doğrulama yöntemin basitleştirilmiş kodu

Çalışmalar göstermiştir *k*-kat çapraz doğrulama yönteminde *k* için en iyi değer 10'dur (Breiman ve ark 1984, Kohavi 1995) (Kahramanli 2008).

Sistemin başarısı aşağıdaki gibi ölçülmektedir:

3.12.2 Sınıflandırma Doğruluğu

Veri kümeleri için sınıflandırma doğrulukları aşağıdaki denklemde görüldüğü gibi ölçülür:

$$doğruluk(T) = \frac{\sum_{i=1}^{N} değerlendirme(t_i)}{N}, \quad t_i \in T \tag{3.28}$$

$$değerlendirme(t_i) = \begin{cases} 1, \ if \ \text{sınıflandırma}(t_i) \equiv doğrusınıflandırma, \\ 0, \ bunundışında, \end{cases} \tag{3.29}$$

T: Test kümesi için Veri öğelerinin kümesini sınıflandırır.

N:Veri setindeki test örneklerinin sayısı (Yilmaz ve ark 2014)

3.12.3 Karmaşıklık (Bozulma) Matrisi

Makine öğrenme alanında karışıklık matrisi bir algoritmanın performansını görünür hale getirmek için özel bir tablo yapısı oluşturur (Stehman 1997). Karışıklık matrisi dört sınıflama performansından oluşur.

Çizelge 3.9'da göründüğü gibi; Doğru olumlu, yanlış olumlu, yanlış olumsuz ve doğru olumsuz olmak üzere dört ifadeden oluşur. Bu performans hesabı sadece iki sınıflı sınıflama işlemlerinde kullanılır (Xu ve ark 2012).

Çizelge 3.9 Karışıklık matrisinde yer alan dört sınıflama performans endeksleri

Gerçek Sınıf	Öngörülen Sınıf	
	Pozitif	Negatif
Pozitif	Doğru pozitif (TP)	Yanlış negatif (FN)
Negatif	Yanlış pozitif (FP)	Doğru negatif (TN)

Doğru negatif; örneğin sonucu negatif olan doğru tahminler sayısıdır. Yanlış pozitif ise bir örneğin pozitif yanlış tahminler sayısıdır. Yanlış negatife, bir örneğin

negatif olan yanlış tahminler sayısıdır. Doğru pozitif ise bir örneğin pozitif olduğu doğru tahminler sayısıdır (Polat ve Gunes 2007).

3.12.4 Hassasiyet ve Seçicilik Analizi

Aşağıdaki denklemler; hassasiyet, özgünlük, pozitif öngörü değeri ve negatif öngörü değerinin ölçülmesini ifade etmektedir (Polat ve Gunes 2007).

$$Hassasiyet(\%) = \frac{TP}{TP + FN} \times 100 \tag{3.30}$$

$$Seçicilik(\%) = \frac{TN}{TN + FP} \times 100 \tag{3.31}$$

$$Pozitif\ Öngörü\ Değeri(\%) = \frac{TP}{TP + FP} \times 100 \tag{3.32}$$

$$Negatif\ Öngörü\ Değeri(\%) = \frac{TN}{TN + FN} \times 100 \tag{3.33}$$

a) Gerçek pozitif (TP) : hasta olarak tanınmış, gerçekte hasta kişi.
b) Gerçek negatif (TN): sağlıklı olarak tanınmış, gerçekte sağlıklı kişi.
c) Yanlış pozitif (FP): hasta olarak tanınmış, gerçekte sağlıklı kişi.
d) Yanlış negatif (FN): sağlıklı olarak tanınmış, gerçekte hasta kişi (Polat ve Gunes 2007).

4 YAPAY SİNİR AĞLARI

4.1 Yapay Sinir Ağlarının Tanımı

Günümüz araştırmacıları, bilgisayarlara insan zekâsı gibi düşünce yeteneği kazandıracak programlar ile yorum yapıp, akıl yürütebilmek için çaba sarf etmektedir. Bu amaçla beynin temel işlem elemanı olan nöronun çalışma prensibine dayanan öğrenme özelliklerine sahip yapay zekâ sistemleri oluşturmaya çalışmışlardır.

Yapay sinir ağları için birkaç tanımda tanıtacak olursak; Beynimizin çalışma prensiplerini gözeterek öğrenme, analiz, kavrama, genelleme yapma gibi yetenekleri sinirlerin işleyiş ve çalışma mekanizmasını taklit ederek yapan sistemdir (Grossberg 1988).

Bir sinir ağını bilgisayar sistemine benzetirsek, bir işlem yapan birimden oluşan, daha önceden yapılmış işlemlerden kaynaklı deneyimleri kazanım olarak kaydedip, gerektiğinde kaydedilen deneyimleri kullanılmasını işlemleri paralel olarak yapan bir işlemcidir.

Bilgiye ulaşmak için ağ yardımıyla öğrenme işlemi gerçekleştirmesi ve ulaşılan bilgilerin depolanması için nöronlar arası bağlantı güçleri kullanması yönüyle insan beyini ile benzerlik göstermektedir. (Haykin 1999) (Bozik 2011).

YSA ile basit biyolojik sinir sisteminin çalışma şekli simüle edilir (benzetilir). Simüle edilen sinir hücreleri nöronlar içerirler ve bu nöronlar çeşitli şekillerde birbirlerine bağlanarak ağları oluştururlar. Bu ağlar öğrenme, hafızaya alma ve veriler arasındaki ilişkiyi ortaya çıkarma kapasitesine sahiptirler. Diğer bir ifadeyle, YSA'lar, normalde bir insanın düşünme ve gözlemlemeye yönelik doğal yeteneklerini gerektiren problemlere çözüm üretmektedir. Bir insanın, düşünme ve gözlemleme yeteneklerini gerektiren problemlere yönelik çözümler üretebilmesinin temel sebebi ise insan beyninin ve dolayısıyla insanın sahip olduğu yaşayarak veya deneyerek öğrenme yeteneğidir (wikipedia 2015).

4.2 Yapay Sinir Ağlarının Genel Özellikleri

Bu ağlarda bir problemin temsil ettiği alan, beynimizin yalnızca bir parçası kadardır. Bu ağlarda işlem birimlerinin birçoğu aynı anda eş zamanlı olarak çalışabilirler ve aynı anda aktive olabilirler. YSA'ların bir başka özelliği de genelleme yapabilmeleridir. Ağ eğitimleri sırasında sayısal bilgileri eşleştirip, bu eşleştirme işlemi sırasında da bilgileri ayırma özellikleri vardır. Ağ fonksiyonları doğrusal olmayabilir.

YSA'lar, donanımsal olarak da gerçekleştirilebilirler. Yapay Sinir Ağı'nın genel özellikleri şu şekilde sıralanabilir (Küçükönder 2011):

a) Makine öğrenmesi gerçekleştirirler.
b) Klasik programlama yöntemlerine benzemez.
c) Bilgiyi depolar.
d) Örnekleri kullanarak öğrenirler.
e) Ağın eğitimine ve performansına bağlı olarak güvenle çalıştırılır.
f) Genelleme özellikleri sayesinde o ana kadar değerlendirilmemiş örnek hakkında bilgi üretebilirler.
g) Algılamaya yönelik olaylarda kullanılabilirler.
h) Şekil ilişkilendirebilme yapabilirler.
i) Kendisinin öğrenme yeteneği vardır.
j) Eksik bilgi ile çalışabilme özelliği vardır.
k) Hata ve gürültüye karşı müsamahaya sahiptir.
l) Bozulmayı dereceli olarak gösterirler.
m) Dağıtık belleğe sahiptirler.
n) Sadece sayısal bilgiler ile çalışabilirler.
o) Doğrusal ve doğrusal olmayan problemleri cevaplanmasında kullanabilirler (Sagıroğlu ve ark 2004).

Bu özellikleri YSA'ların günlük hayatta daha da fazla kullanılarak problemlerimizin çözümünde daha çok kullanılmasına olanak sağlamıştır.

4.3 Yapay Sinir Ağlarının Avantajları

Yapay sinir ağlarının önemli görünen avantajları şunlardır (Küçükönder 2011);

a) Karmaşık ve matematiksel olarak ifade edilme yeteneği zayıf olan problemleri çözüme kavuşturma olanağı sağlar (Elmas 2003).

b) Problemleri sadece örneklerle çözüme ulaştırırlar. Her problem için ayrı ayrı bilgiye ihtiyaç duymazlar.
c) Problemler için belli bir kural tabanı gerektirmeden örneklerden yola çıkarak genelleme yapabilir.
d) Yapay sinir ağlarıyla olayların içindeki etki ve ilişkilerin ortaya çıkarılmasını kolaylaştırır.
e) YSA'lar olaylardaki doğrusal olmayan ilişkileri geleneksel sistemlere nispeten daha gerçekçi sonuç ve çözümler üretirler.
f) YSA'lar sadece olayın çözümü için kullanılacak örneklemler ve basit bir programlama ile çözülebileceğinden ekonomik ve pratiktir.
g) YSA'lar diğer geleneksel yöntemlere nazaran problemin çözümünde zaman bakımından daha verimlidir.
h) YSA'lar problemler için yeni veriler oluştuğu takdirde yeniden eğitilerek problemin çözümünde tekrar kullanılabilirler.
i) YSA'lar eş zamanlı uygulamalar için paralel olarak çalıştırılabilirler.

4.4 Yapay Sinir Ağlarının Dezavantajları

YSA uygulamalarında dezavantajları sayacak olursak (Küçükönder 2011);

a) Uygulamalar da ağ topolojisi ve model seçimi için belirli kurallar olmayıp kullanıcıya bağlıdır. Kullanıcının daha önce yaşadığı tecrübeleri bu seçimlerde önem kazanır, bu durum kişiye bağlı olarak iyi sonucun bulunmasında zorluklar ortaya çıkmasını sebep olur.
b) Sistem çalışırken içerisinde neler olduğu bilenememektedir.
c) Kararlılık analizi bazı ağ yapıları hariç yapılamamaktadır.
d) Diğer sistemlerle entegrasyon işlemlerinde problemlere göre güçlükler çıkabilmektedir.
e) Problemlerin çözümünde kullanılacak örneklemlerin hazırlanması kişilere bağlıdır. Kullanılacak örneklerin oluşturulması amacıyla yapay sinir ağlarında bir kural yoktur.
f) YSA'da öğrenme hızı problemlere göre değişkenlik gösterebilir (Efe ve Kaynak 2004).

g) Yapay sinir ağlarının yapısının esnek olması sebebiyle eğitim aşamasında verilere aşırı uyumda ve gürültülü veriler sebebiyle sonuçlarda performansı kötü yönde etkileyebilmektedir (Küçükönder 2011).

Araştırmalar sonucunda geleneksel ve yapay sinir ağları ile uygun sonuçlar bulunsa dahi bunların en iyi çözüm olduğunun belirtilmesi yanlış olur. Bulunan sonuçlar uygulanan veriler için oluşacak sonuçların çözüm kümesinde bir elemanı temsil eder. Bir problemin çözümü için oluşturacak örneklerin yanlış seçilmesi sonuçların sağlıklı ve doğru çıkmasını engelleyebilir.

4.5 Biyolojik Sinir (Nöron) Yapısı

Yapay sinir ağlarını anlayabilmek için insan beynindeki sinir hücrelerinin tanınmasıyla başlanması gerekir. Biyolojik sinir sistemimizin merkezinde olaylar hakkında bilgiyi alıp, yorumlayan onun hakkında hüküm üreten bir sistemdir (Haykin 1999).

4.6 Biyolojik Sinir Ağları

Biyolojik sinir ağları, çevreden gelen verileri elektriksel sinyallere çevirip beyne ileten Alıcı sinirler, beyin tarafından üretilen elektriksel sinyallerde uygun tepkilere dönüştüren tepki sinirleri ve bu alıcı ile tepki sinirleri arasında koordinasyon için ileri ve geri besleme yaparak uygun tepkiler üreten merkez sinir olarak tanımlanan 3 katmandan oluşur (Saraç 2004).

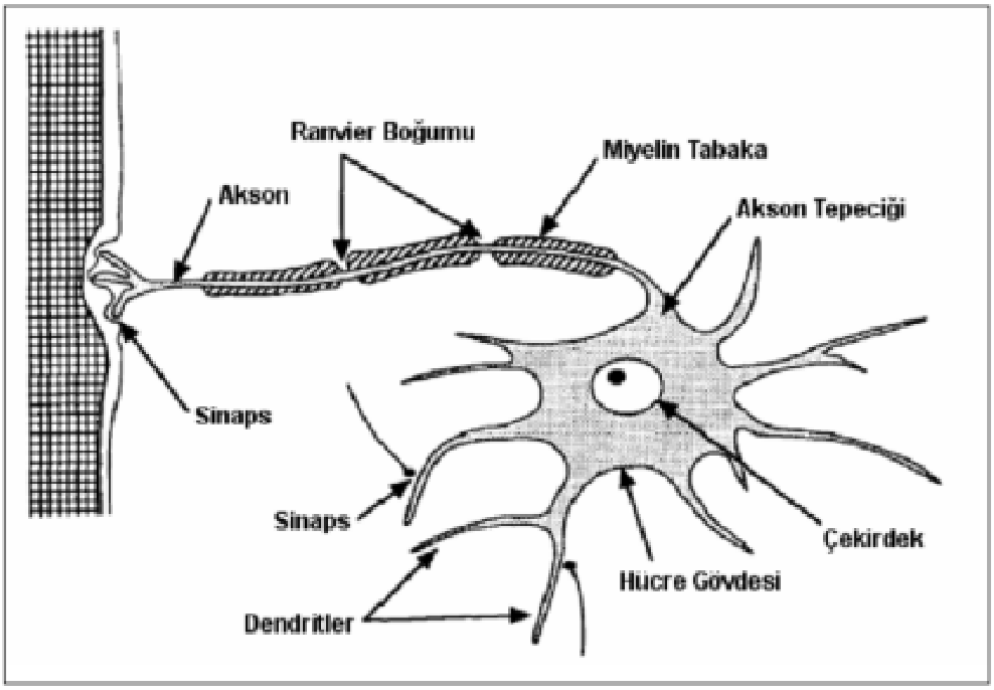
Bu sinir ağları beş duyu organından gelen bilgiler ışığında geliştirdiği algılama ve anlama mekanizmalarını çalıştırarak olaylar arasındaki ilişkileri öğrenir. İnsan beyninde, sinir sisteminin temelini oluşturan vücuda sinyal yayma işleminde kullanılan tıpta adına nöron adı verilen özel hücreler vardır.

Bir sinir hücresi; sinaps, **hücre gövdesi (**soma), **akson** (axon) ve dentrit'lerden oluşmaktadır. Sinapssinir hücreleri arasındaki elektrik sinyallerini geçmesini sağlayan boşluklardır. Hücre gövdesi gelen elektrik sinyalleriyle oluşan uyartıyı yorumlar, bu gövdenin çevresinde, saç teline benzeyen ve duyulardan gelen elektrik sinyalleriyle oluşan uyartıları almakla görevli **dentrit** (sinyal alıcıları - dendrite) bulunur. Hücre

gövdelerine bağlı olan ve aldığı sinyalleri diğer nöronlara ileten kuyruk gibi akson (axon) da sinir hücrelerinde bulunmaktadır.

Dentritler üzerinden duyulardan gelen elektrik sinyalleri alınarak hücre gövdesi tarafından doğrusal olmayan bir şekilde işlenmektedir. İşlenen sinyalleri, iletme hızını artıran bir çeşit yalıtım maddesi olan **miyelin tabaka** (myelin sheath) kaplı olan aksonlar üzerinden çıkışa taşınır.

Bu miyelin tabaksının üzerinde giriş sinyallerini belli aralıklarla yeniden üreten birkaç milimetrede bir **ranvier boğumu** (ranvier node) bulunur. Dendrit- akson bağlantısını sinir hücreleri arasındaki elektrokimyasal bağlantıyı da yapan **sinaps** (synapse) yapar (Baş 2006). Bir canlının biyolojik bir sinir hücresinin temel yapısı Şekil 4.1'de gösterilmiştir. Çizelge 4.1'de insan beyniyle bilgisayar sisteminin karşılaştırılma çizelgesi verilmiştir.



Şekil 4.1 Bir canlının biyolojik sinir hücresinin yapısı (Freeman ve Skapura 1991).

Çizelge 4.1 İnsan beyni ile bilgisayar sistemlerinin karşılaştırılması (Crone 2004).

Özellikler	**İnsan Beyni**	**Bilgisayar**
İşlemci Hızı	10^{-3} ms (0,25 MHz)	10^{-9} ms (2500 MHz)
Nöron / Transistör	10^{11} ve 10^{33} bağlantı	10^{9} (Yonga)
Ağırlık	1500 gr.	Kilogramlarca

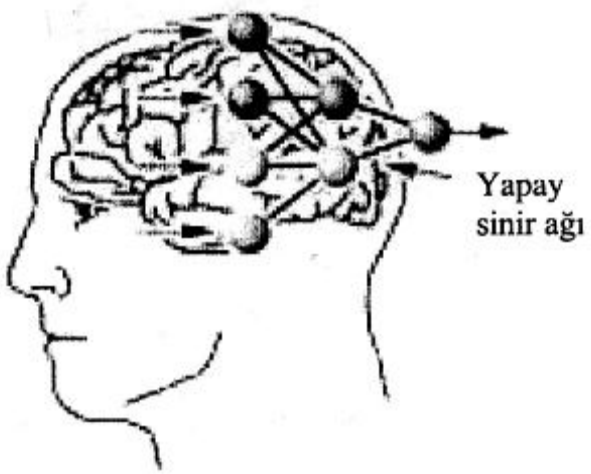
Enerji Tüketimi	10^{-16} Joule	10^{-6} Joule
Hesaplama	100 Adım	Milyarlarca Adım

4.7 Yapay Sinir Hücreleri

Canlılarda bulunan biyolojik sinir ağlarında olduğu gibi yapay sinir ağları da birbiriyle bağlantılı birçok hücreden oluşur. (Şekil 4.2) Yapay sinir hücreleri, biyolojik sinir hücrelerinin basit bir modeli olduğu için yapısal olarak da birbirlerine benzerdirler. Bu benzerlikler Çizelge 4.2'de gösterilmiştir (Sagıroğlu ve ark 2004).

Çizelge 4.2 : Biyolojik sinir sistemi ile YSA'nın yapısal benzerlikleri

Biyolojik Sinir Sistemi	Yapay Sinir Ağları
Sinir Sistemi	Sinirsel Hesaplama Sistemi
Nöron (Sinir Hücresi)	İşlemci Eleman (Yapay Sinir Hücresi)
Dendrit	Toplama Fonksiyonu
Hücre gövdesi (Soma)	Transfer (Aktivasyon) Fonksiyonu
Aksonlar	İşlemci Eleman Çıktısı
Sinapslar	İşlemci Elemanlar Arasındaki Bağlantı Ağırlıkları



Şekil 4.2 Biyolojik ve yapay sinir ağı

Birden fazla da giriş sinyalini aynı anda toplayabilen, bağlantı ağırlık değerlerine, eşik değerine, giriş sinyaline ve aksiyon fonksiyonuna göre yapay sinir ağlarının içindeki sinir hücreleri bir çıkış değerine dönüştürür (Hertz ve ark 1991). Çıkış değeri bir başka sinir hücresine iletildiği gibi, sinir ağının dışına da verilebilir.

Yapay sinir ağlarında genel olarak beş ana öğe oluşmaktadır (Baş 2006).

a) · Girdiler (Inputs)

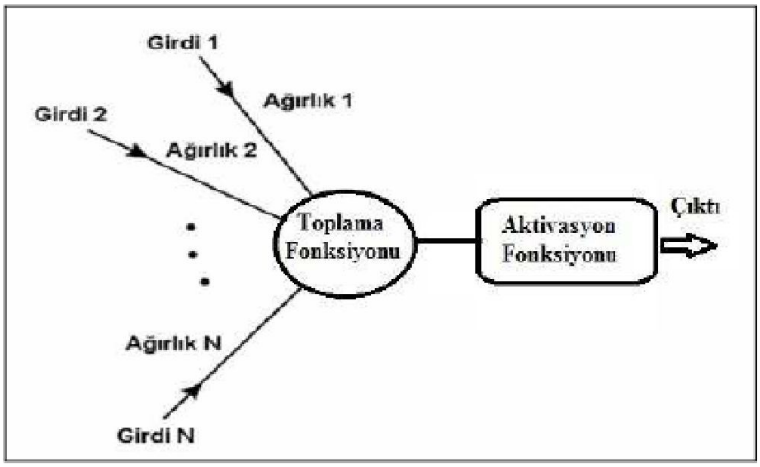
b) · Ağırlıklar (Weights)

c) · Toplama Fonksiyonu

d) · Aktivasyon (Transfer) Fonksiyonu

e) · Çıktı (Output)

Şekil 4.3'de yapay sinir hücresi ana bileşenleri yapısı gösterilmektedir.



Şekil 4.3 Yapay sinir hücresinin yapısı (Tarassenko 1998)

Çizelge 4.3'te istatistik ve YSA'nın karşılaştırılmasıyla ilgili çizelge verilmiştir.

Çizelge 4.3 İstatistik ve YSA karşılaştırılması (Sagıroğlu ve ark 2004).

İstatistik	YSA Sistemi
Model	Ağ
Öngörü	Öğrenme
Regresyon	Danışmanlı öğrenme
İnterpolasyon	Genelleştirme
Gözlem	Öğrenme Algoritması
Parametre	Ağ Parametreleri (ağırlıklar/katsayılar)
Bağımsız değişken	Girdiler
Bağımlı değişken	Çıktılar
Sınır regresyonu	Ağırlık budama işlemi

4.8 Yapay Sinir Ağlarının Temel Elemanları

Yapay sinir ağlarında bulunan ana öğeler aşağıda belirtilmiştir.

4.8.1 Katmanlar

Yapay sinir ağlarının basit topolojik yapıları vardır. Ağların birçoğu üç katmandan oluşur. Bu katmanlar girdi, gizli ve çıktı katmanıdır. Gizli katman sayısı için genellikle tek kullanılsa da ihtiyaca göre daha fazla katmanda kullanılabilmektedir. Birimler arasında bağlantıların *w* ile belirtilen ağırlık değerleri vardır (Larose 2005).

Veri setindeki özelliklerin sayısı ve tipine göre girdi birimlerinin sayısı değişir. Gizli katman sayısı ve gizli katmandaki birim sayısını kullanıcı tarafından belirlenir. Problemin özelliğine göre çıktı katmanı sayısı belirlenir.

4.8.1.1 Girdi Katmanı

Veri setinden özellik değerlerini alıp, gizli katmana ileten hayali bir katmandır. Bu katmanın üzerinde hiçbir işlem yapılmadığı için diğer katmanlardaki ayrıntılı bir yapısı yoktur.

4.8.2 Gizli Katman

Bu katman bilginin işlendiği katmandır. Giriş katmanından aldığı bilgiyi işleyerek çıkış katmanına iletir. Diğer katmanlara göre genel olarak nöron sayısı fazladır. Bu katmandaki nöron sayıları problemlere göre değişlik gösterebilir. Bu katmandaki birimler doğrusal olmayan davranışlar içerirler. Gizli katmanlarda karmaşık örüntülerin çözümü için çok fazla düğüm kullanılmak istense de problem için gerekli katmandan daha büyük gizli katman eğitim veri setinin ezberlenmesine neden olur. Böylece istenmeyen aşırı öğrenme gerçekleştirilmiş olur. Öğrenme sonuçları istenilen seviyede olabilmesi için gizli katmandaki birim sayıları değiştirilerek öğrenme sonuçları istenilen düzeye çıkarılabilir (Mehrotra 1997).

Yeni gizli katmanlar ilave edildiğinde, eğitim sürecinde hata fonksiyonunun takılabileceği en düşük noktalar da ilave edilir. Ağ oluşturulurken önce tek gizli katman

ile başlanır. Ağ yetersiz ise problem için kullanılabilecek minimum sayıda yeni gizli katmanlar ilave edilir (Kriesel 2007). Problemler için en iyi gizli katman sayısını veren bir formül geliştirilmemiştir. Gizli katman sayısının belirlenmesi kullanıcının tecrübesi ve deneme yanılmayla en uygun değer tespit edilir. (Munakata 2008). *"Doğrusal ayrılabilir (linearly separable)"* problemler için çok katmanlı modeller ile tek katmanlı modeller arasında üstünlük olmadığı için tek katmanlı modeller kullanılır. (Kecman 2001).

4.8.2.1 Çıktı Katmanı

Gizli katmandan gelen bilgiyi, giriş katmanından gelen bilgiye göre işleyip sonucu sistem dışına gönderir. Her nöron tek bir çıkışa ait olacak şekilde oluşturulur.

Sınıflandırma problemlerinde tek bir çıktı birimi varsa çıktı katmanında sınıfları ayıracak bir önsel olarak atanan tek birim olan "eşik" değeri yaygın olarak kullanılır. Bu sınıfları birbirinden ayırt edebilmek için uzmanların belirlediği her bir sınıfın eşik değerinin belirlenmesi gerekir. Sınıf sayısı az ise bayrak değişkenleri de kullanılabilir.

Birçok ağda, bilgi işlem birimleri aşağıdaki şekilde çıktı üretirler.

$$y = f\left(\sum_i w_i x_i\right) \tag{4.1}$$

Formülde;

x_i : birimlerin çıktılarını veya dışarıdan gelen girdi değerlerini

w_i : Bağlantı ağırlıklarını

f(.) : Doğrusal olmayan bir fonksiyonu gösterir.

Bütün birimler kendine gelen ağırlıklandırılmış girdilerin toplamını hesaplarlar. Hesaplanan değerler skaler bir çıktı oluşturulması için genellikle azalmayan, doğrusal olmayan f fonksiyonuna gönderilir (Reed ve Marks 1999).

4.8.3 Ağırlıklar

Yapay sinir hücresine gelen bilgiler giriş bölümünden değerlendirme kısmına ulaşmadan geldikleri bölümün ağırlığı ile çarpılarak iletilir. Bu şekilde girdilerin

üretilecek çıktı üzerindeki etkisi ayarlana bilinmektedir. Ağırlıklar, YSA'nın öğrenme yeteneği ile ilgilidir ve aynı zamanda sinir ağının zekâsını oluşturur. Yapay sinir ağlarında bilginin sergilenmesi ağırlıkla yapılır. Bilginin sergilenmesi ağın tasarım ve performansını etkiler. Bu amaçla dört temel kural kullanır (Haykin 1999).

Kural 1: Girdilerin, Euclid uzaklığı, iç çarpım veya Mahalanobis uzaklığı gibi ölçütler ile benzerlikleri tespit edilip benzer girdiler aynı kategoride yer alacak şekilde sınıflandırılır.

Kural 2: İlk kuralın tersi olarak ayrı kategoride sınıflandırılacak girdilerin gösteriminde farklılıklar çok fazla olmalıdır.

Kural 3: Özniteliklerin önemlilik oranlarında birim sayısı artırılır. Bu artırılma karar vermede yüksek kesinlik ve hata durumlarına karşı müsamaha sağlar.

Kural 4: Önsel bilgiler, basit bir ağ oluşturmak için ağ tasarımı sırasında kullanılmalıdır. Önsel bilgiyi kullanmak için *"ağırlık paylaşımı (weight sharing*)" gibi yaklaşımlarla ağırlıkları seçimi kısıtlanır. Bu yaklaşımlar kullanmak ağ tasarlayan bırakılmıştır. Bu yaklaşımı kullanmak ağ için kullanıcıdan istenen parametre sayısı düşürülmüş olur (Bishop 1995).

4.8.4 Birleştirme Fonksiyonu

Bu fonksiyon bütün birimler için ilgili birimin girdi değerleri ile bu birimin ağırlıklarının doğrusal birleşiminden bir değer üretir. Bu değer *net girdi* olarak adlandırılır.

Birleştirme fonksiyonu olarak en sık kullanılan formül, her gelen girdilerin kendilerinin bağlı olduğu ağırlık değeri ile çarpılıp elde edilen değerlerin toplandığı ağırlıklı toplamdır.

Denklem 4.2'de bir j düğümü için net girdiyi ağırlıklı toplam kullanılarak bulunur.

$$Net\ girdi_j = \sum_{i=1} w_{ij} x_{ij} = w_{0j} x_{0j} + w_{1j} x_{1j} + ... + w_{ij} x_{ij} \tag{4.2}$$

Formülde belirtilen;

x_{ij} *: j* düğümüne gelen *i.* girdiyi göstermektedir.

w_{ij} : j düğümü ile i. girdi arasındaki ağırlığı göstermektedir. Toplam $l+1$ adet girdi bulunur.

$x_1 + x_2 , \ldots x_l$ birime gelen girdileri temsil eder. Her çıktı ve gizli katman biriminde x_{0j} =1 ile temsil edilen ve değeri 1 olan ekstradan bir girişe sahiptir (Rumelhart 1986).

Problemlere göre en uygun birleştirme fonksiyonunu deneme yanılma ya da kullanıcının deneyimlerine dayanarak bulunur. Bu konuda probleme göre geliştirilmiş bir formül yoktur. Çizelge 4.4'te kullanılabilecek birleştirme fonksiyonları bulunmaktadır (Öztemel 2006).

Çizelge 4.4 Birleştirme fonksiyonları (Öztemel 2006)

Birleştirme Fonksiyonu	**Net Girdi**	
Ağırlıklı toplam	$Net\ girdi = \sum_{i=1}^{N} x_i w_i$	Ağırlık değerleri girdiler ile çarpılır ve bulunan değerler birbirleriyle toplanarak Net girdi hesaplanır.
Çarpım	$Net\ girdi = \prod_{i=1}^{N} x_i w_i$	Ağırlık değerleri girdiler ile çarpılır ve daha sonra bulunan değerler birbirleriyle çarpılarak Net Girdi Hesaplanır.
Maksimum	$Net\ girdi = maks(x_i w_i)$	n adet girdi içinden ağırlıklar girdilerle çarpıldıktan sonra içlerinden en büyüğü Net girdi olarak kabul edilir.
Minimum	$Net\ girdi = \min(x_i w_i)$	n adet girdi içinden ağırlıklar girdilerle

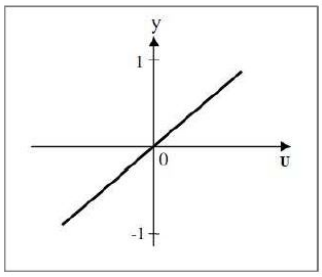
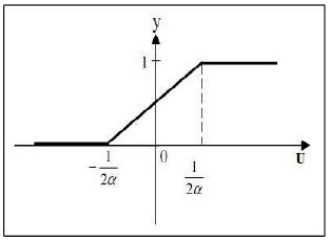
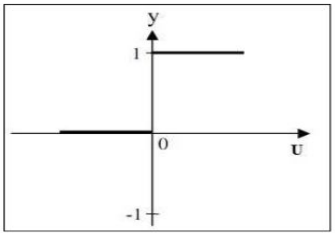
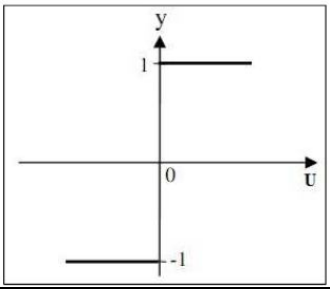
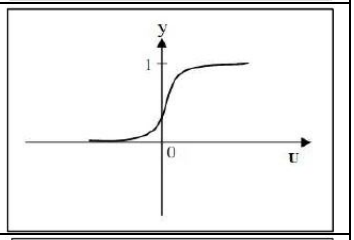
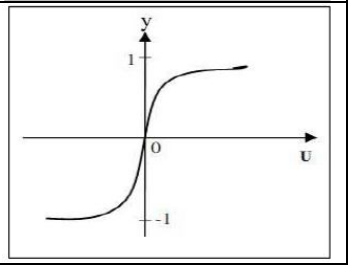
		çarpıldıktan sonra içlerinden en küçüğü Net girdi olarak kabul edilir.
Çoğunluk	$Net\ girdi = \sum_{i=1}^{N} \text{sgn}(x_i w_i)$	n adet girdi içinden girdilerle ağırlıklar çarpıldıktan sonra pozitif ile negatif olanların sayısı bulunur. Büyük olan sayı hücrenin net girdisi olarak kabul edilir.
Kümülatif toplam	$Net\ girdi = net(eski) + \sum_{i=1}^{N} x_i w_i$	Hücreye gelen bilgiler ağırlıklı olarak toplanır. Daha önce hücreye gelen bilgilere yeni hesaplanan girdi değerleri eklenerek hücrenin net girdisi hesaplanır.

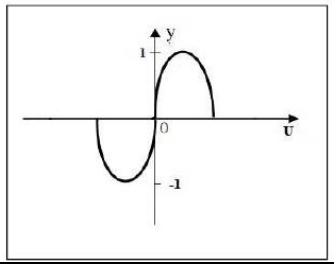
4.8.5 Etkinlik (Aktivasyon) Fonksiyonu

Etkinlik fonksiyonu yapay sinir ağlarında kullanılan birimlerin davranışlarını belirler. Bu fonksiyon nörona gelen net girdiyi işleyerek nöronun bu girdiye karşılık üreteceği çıktıyı belirler. Biyolojik nöronlarda, ağırlıklandırılmış girdilerinin toplamı belli bir eşik değerini aşan nöronlar diğer nöronlara bir elektrik sinyali gönderdiği gözlenmiştir. Bu davranışı yapay sinir ağlarında oluşturmak için aktivasyon fonksiyonu genellikle doğrusal olmayan bir fonksiyon seçilir. Ayrıca bu fonksiyonun türevinin kolay hesaplanması tercih edilir.

En sık kullanılan etkinlik fonksiyonu sigmoid fonksiyonudur. İlgili birimin çıktısını oluşturmak için net girdi değeri sigmoid fonksiyonuna girdi olarak verilir. Oluşan çıktı değeri çıkış katmanına ağırlandırılmış olarak gönderilir (Zurada 1992) . Sıklıkla kullanılan etkinlik fonksiyonları Çizelge 4.5'te formülleri ve grafiksel gösterimleriyle tanıtılmıştır.

Çizelge 4.5 Etkinlik Fonksiyonları (Emir 2013)

Aktivasyon fonksiyonu	Formül	Grafik
Doğrusal (Lineer) Fonksiyon	$U = \sum_{i=1}^{n} x_i w_i + \theta$ Veya $U = \sum_{i=1}^{n} x_i w_i - \theta$	
Parçalı Doğrusal Fonksiyon	$y = F(U) = \begin{Bmatrix} 0 & U \le -\frac{1}{2a} \\ aU + \frac{1}{2} & \lvert U \rvert < \frac{1}{2a} \\ 1 & U \ge \frac{1}{2a} \end{Bmatrix}$	
Basamak (Step) Fonksiyon Tek Kutuplu (Unipolar) Basamak Fonksiyonu	$y = F(U) = \begin{Bmatrix} 1 & U \ge 0 \\ 0 & U < 0 \end{Bmatrix}$	
Basamak (Step) Fonksiyon Çift Kutuplu (Bipolar) Basamak Fonksiyonu	$y = F(U) = \begin{Bmatrix} +1 & U \ge 0 \\ -1 & U < 0 \end{Bmatrix}$	
Sigmoid Fonksiyon	$y = F(U) = \frac{1}{1+e^{-U}} = \frac{1}{2}\left[\tanh(U/2) - 1\right]$	
Tanjant Hiperbolik Fonksiyon	$y = \tanh(\beta U) = \frac{1-e^{-2U}}{1+e^{2U}}$	

Sinüs Tipi Fonksiyon	$y = F(U) = \sin U$	

4.8.6 Hata Fonksiyonu

Hata fonksiyonu olarak genel olarak eğitim kümesindeki tüm gözlemlerin tüm çıktı birimlerindeki tahmin hatalarının karelerinin toplamının alındığı *"Hata Kareleri Toplamı (HKT)"* kullanılır (Zurada 1992).

$$HKT = \sum_{girdiler} \sum_{\substack{Çıktı \\ Birimleri}} \left(\begin{matrix} Bekelenen \\ Çıktı \end{matrix} - \begin{matrix} gerçekleşen \\ Çıktı \end{matrix} \right)^2 \tag{4.3}$$

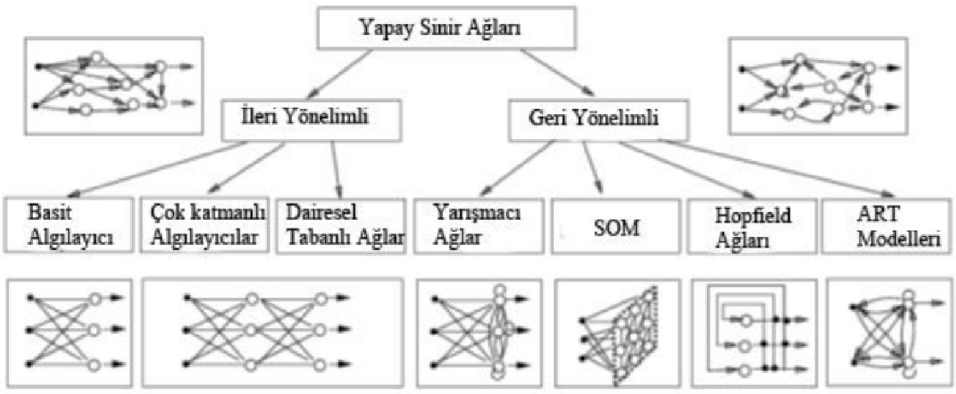
Problem çözümlerinde hata kareleri toplamına ek olarak, HKT değerini eğitimde kullanılan gözlem sayısı ve ağın çıktı sayısını kullanarak normalize edilerek oluşturulan *"Ortalama Hata Kareleri (OHKT)" de* kullanılır. Ağ sonucunda elde edilecek çıktının olasılık tahmini vermesinin istenildiğinde hata fonksiyonu olarak *"çapraz entropi (cross entropy)"* fonksiyonu kullanılır (Mitchell 1997).

Hata fonksiyonlarını seçimi problemlere göre değişmektedir. Genel olarak kolaylıkla türevlenebilir olmaları, oluşan hataların önceki hataların eğilimlerinden ve büyüklüklerinden etkilenmemeleri, ayrıca girdilerden bağımsız olarak eşit hataların maliyetlerinin eşit olması sebebiyle, standart hata fonksiyonları kullanılır (Reed ve Marks 1999).

4.9 Yapay Sinir Ağı Mimarileri

Yapay sinir ağları içerdiği nöronların (hücrelerin) birbirine bağlanış şekline göre ileri ve geri beslemeli olmak üzere iki temel mimari söz konusudur. Tahmin, fonksiyon kestirme, örüntü tanıma gibi problemlerde ileri yönelimli ağlar, kümeleme, optimizasyon problemlerinde ise geri yönelimli ağlar daha çok kullanılır (Cura 2008).

Şekil 4.4, ileri ve geri yönelimli mimariye sahip temel YSA modellerini göstermektedir (Emir 2013).



Şekil 4.4 İleri ve Geri yönelimli Ağ mimarileri(Emir 2013)

4.9.1 İleri beslemeli yapay sinir ağları:

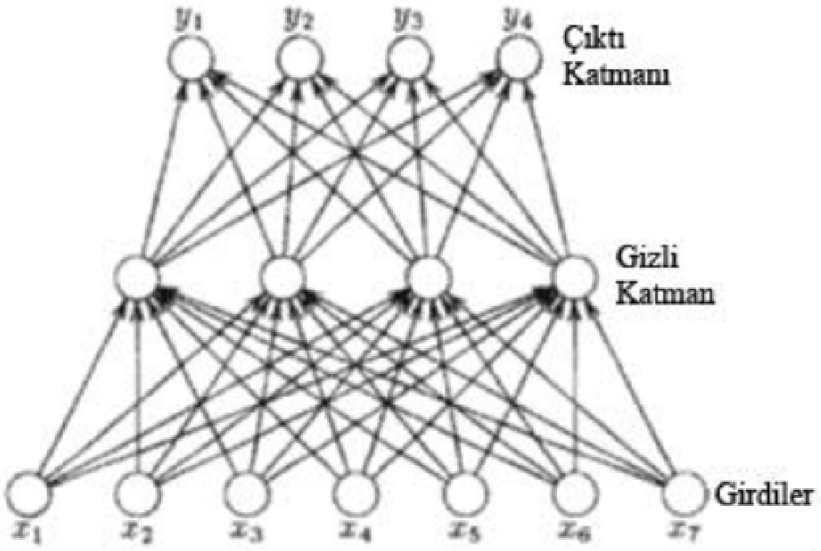
Bu tür yapay sinir ağlarında nöronlar bir katmandan diğer bir katmana bağlantı kurarken, aynı katmanda ya da önceki katmanlara bağlantı kurmazlar. Bilgiler sadece ileri yönde iletilir.

Bir katmandaki hücrelerin çıkışları bir sonraki katmana ağırlıklar üzerinden giriş olarak verilir.

Çok katmanlı algılayıcı, LVQ (Learning Vector Quantization) ağları, radyal tabanlı sinir ağları ileri beslemeli ağlara örnek olarak verilebilir.

İleri yönde veri akışı olan sistemlerde ara ve çıkış katmanındaki hücreler, kendisinin bulunduğu katmanın sadece bir önceki katmandan değil de, kendisinden diğer katmanlardan da besleniyorsa kaskat bağlı yapay sinir ağı olarak tanımlanır.

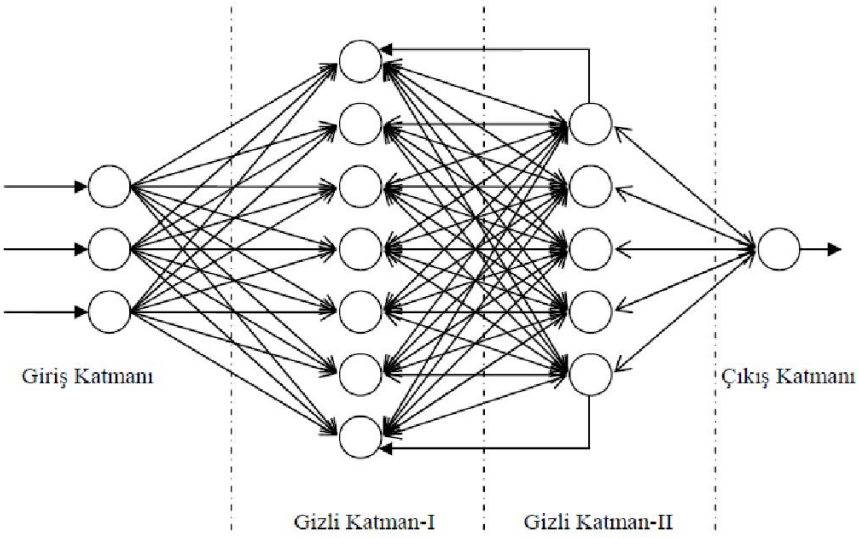
İleri beslemeli YSA'ları sinyal işleme, fonksiyon yaklaştırma (function fitting) ve örüntü tanıma (pattern recognition) problemlerinin çözümünde yaygın olarak kullanılır. Şekil 4.5'te ileri yönelimli çok katmanlı ağa ile ilgili bir örnek model gösterilmiştir.



Şekil 4.5 İleri yönelimli çok katmanlı ağ(Emir 2013)

4.9.2 Geri beslemeli yapay sinir ağları:

Giriş katmanı hariç diğer katmanların kendisinden önceki katmanlara geri beslemesi vardır. Girişler ileri ve geri yönde bilgi aktarımı yapacak şekilde tasarlanır. Geri beslemeli ağlarda o anki çıkış değerinin, o anki giriş ve o andan önceki girişlerin etkisi vardır. Bu tip beslemelerde, besleme aynı katmandaki nöronlar arasında olduğu gibi farklı katmanlardaki nöronlar arasında da olabilir. Geri beslemeli ağlar yapılış şekline göre Hopfield, SOM, Elman ve Jordan ağlarında olduğu gibi farklı ağ yapıları elde edilebilir. Bu sinir ağları doğrusal olmayan dinamik davranış gösterirler. Şekil 4.6'da geri beslemeli çok katmanlı ağ ile ilgili bir örnek model gösterilmiştir.



Şekil 4.6 Geri beslemeli çok katmanlı yapay sinir ağı(Emir 2013)

4.10 Yapay Sinir Ağlarında Öğrenme

Yapay sinir ağları öğrenme ve hatırlama olmak üzere iki işleve sahiptir. Girişler arasındaki etki derecesini ayarlayan ağırlıkların istenen çıkış değerleriyle o anki çıkan sonuçların değerlendirilmesiyle sonucu düzenlemesi işlemleri için geçen süreye öğrenme adı verilir. Öğrenme işlemi bittikten sonra oluşan eşik ve ağırlık değerlerine göre belirli bir veri girişi için uygun çıkış üretmesine hatırlama adı verilir. Öğrenme işlemi gerçekleştikten sonra öğrenme işlemi sırasında kullanılmayan giriş verileri sisteme verilir. Sinir ağının çıkış değerleriyle gerçek yaklaşımlar incelendiğinde doğru yaklaşımlar elde ediliyorsa sinir ağı öğrenmiş demektir. Sinir ağlarına verilen giriş değeri sayısı uygun sayının üstündeyse öğrenme işlemi değil ağ ezberleme işlemi yapmış olur. Çıkış değerleriyle gerçek yaklaşımlar incelenip karşılaştırılarak hatlar tespit edilir. Öğrenmenin en iyi olduğu zaman hatanın en küçük olduğu zamandır.

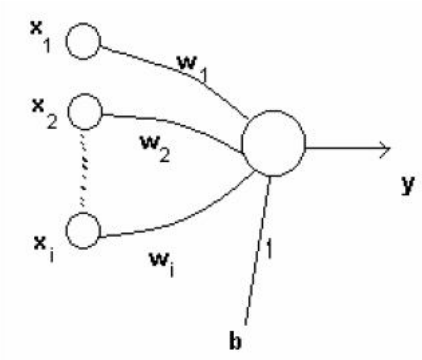
Yapay sinir ağlarında ağın öğrenmesinde kullanılacak olan yöntemin doğru olarak belirlenmesi oldukça önemlidir. Problem çözümlerinde yaygın olarak kullanılan öğrenme kurallarını Hebb kuralı, Hopfield kuralı, Delta ve Kohonen Kuralları kullanılır. Bu kurallarla ilgili açıklamalar aşağıda tanımlanmıştır.

4.10.1 Hebb Kuralı

1949 tarihinde Donald Hebb tarafından biyolojik olarak beynin öğrenme sistemine dayanarak geliştirilen ve beynin sinir hücrelerinin yapısını ve işleyişini temel alan öğrenme kuralıdır. İki sinir hücresinin arasındaki durumu inceler. Bu hücrelerin ikisi de aktif ise bağlantı o kadar güçlü olmuş olur. Şekil 4.7'de Hebb net'in mimarisi gösterilmiştir.

Hebb kuralında giriş yapılarının birbiriyle ilişkisiz ve dikey olduğu durumlarda daha uygun sonuçlar üretir (Özşahin 2009).

Hebb kuralı iki sinir hücresinin girdi alışverişleri sırasında iki hücrenin matematiksel olarak işaretleri incelenmesiyle bağlantı durumlarının düşük ya da yüksek dereceli olduğu tespit edilebilmektedir. Matematiksel olarak aynı işarete sahip iki hücrenin bağlantısı kuvvetlenir, zıt olan hücrelerin bağlantısı zayıflar (Küçükönder 2011).



Şekil 4.7 Hebb net'in mimarisi(Fausett 1994)

İki hücre aynı aktiflikteyse "w" ağırlıkları artırılır, farklı ise bu ağırlık değeri azaltılır. Girdinin örüntü tipine göre değiştirilmez. Girdi olarak ikili(0 veya 1) veri tipi kullanılırsa, ağırlık olarak ifade edilen "w" ler farklı aktiflikteyseler azaltılmaya uğratılmazlar. Kullanılan örüntü iki kutuplu ise ağılık olarak ifade edilen "w" değeri azaltılır. Bu ifadelere göre oluşan Denklem 4.4'de belirtilmiştir.

$$w_i\left(yeni\right) = w_i\left(eski\right) + x_i * y \tag{4.4}$$

Yukarıda belirtildiği gibi ikili veri tipleri kullanılıyorsa x_i ve y nin etkinlikleri ters ise w_i değişmeyecektir. x_i ve y ikisi birden pasif etkinlikte ise (0,0) "w_i" yine değişmez. Bu kuralda w ilerin kuvvetlendirilmesi gerekir. Bu sebeple girdi verileri Hebb kuralında iki kutuplu alınması gerekir. Hebb kuralı algoritmasının basitleştirilmiş kodu Şekil 4.8'de gösterilmiştir. (Fausett 1994);

Adım 0: i=1,2,3,…..n $w_i = 0$ Bütün ağırlıkların başlangıç değerlerine sınıf atanır.

Adım 1: $x_i=s_i$ i=1,2,3,…..n Girdi örüntüsünden (s) her bir girdi elemanı alınır.

Adım 2: y=t Çıkması gereken sonucu(t) ata

Adım 3: Ağırlıklar ve dengeler ayarlanır.

$w_i(yeni)=w_i(eski) + x_i*y$ (i=1,2,3…..n)

b(yeni)=b(eski) + y

Adım 4 : Adim 1-3 arasını tekrarla.

Şekil 4.8 Hebb algoritmasının basitleştirilmiş kodu

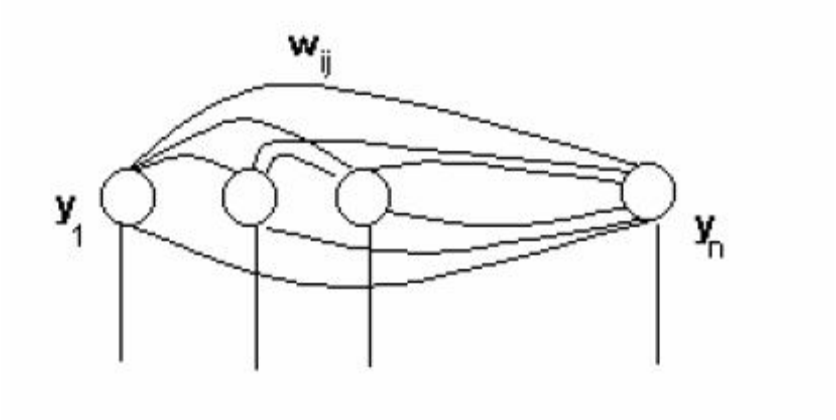
Hebb algoritmasıyla, ağın eğitimden önce ağırlıklarının sıfırlanması gerekir.

4.10.2 Hopfield Kuralı

Bu kural, Hebb kuralının temelinde geliştirilmiştir. Hebb kuralından belli noktalarda farklılıkları vardır. Tek katmanlı ve geri dönüşümlü bir yapıya sahiptir (B. 2006). Şekil 4.9'da Hopfield net'in modeli gösterilmiştir.

YSA'da işleme girecek elemanlar girdi ya da çıktı elemanı olabilir. Bağlantılarda kullanılan ağırlık değerinde yapılan değişiklikler için genel olarak öğrenme katsayısı kullanılır.

Çıktı için hedeflenen değer ile girişlerin ikisi de aktif veya pasif olmasına göre 0 ile 1 arasında bulunan ve araştırmacılar tarafından belirlenen pozitif sayı olan öğrenme katsayısı, bağlantı ağrılıkları kadar azaltılır ya da artırılır (Tebelkıs 1995).



Şekil 4.9 : Hopfield Net Modeli(Cinsdikici 2015)

Giriş değerlerinde ikili(binary) değerler yerine iki kutuplu değerlerin kullanılması daha uygun olur. Bu yaklaşıma göre örüntü için Denklem 4.5'te gösterildiği gibi öğrenme ifade edilir.

$$W = \{w_{ij}\} \quad \text{ve} \quad w_{ij} = \sum_{p} s_i(p) * s_j(p) \quad , i \neq j \tag{4.5}$$

Bu formülde ifade edilen Hopfield öğreneme formülü Hebb kuralından sağlanmaktadır.

P adet girdinin her bir elemanı "n" hücreden oluşmaktadır("i=j=1,2,....n.") Bu girdilerin ağırlık matrisi oluşturmaktadır. Tüm girdi örüntüleri kullanılarak "w_{ij}" üzerinde öğrenmenin tamamlanmış olduğu belirtilir (Fausett 1994) (Cinsdikici 2015).

4.10.3 Delta Kuralı

Widrow ve Hoff tarafından, Hebb kuralını geliştirerek ortaya koydukları bir kuraldır (Baş 2006). Bu kuralda, hedeflenen sonuç ile sinir ağından oluşan çıkış değerleri arasındaki farkı azaltmak için ağırlıklar değiştirilerek sinir ağından en iyi sonuç alınmaya çalışılır (Çorumluoğlu ve ark 2005). Kuralda hücrenin belirlenen gerçek çıkış değeri ile istenilen çıkış değerleri arasındaki farkı azaltmak için giriş bağlantı değerleri güçlendirmek için sürekli değiştirilir. Delta kuralı, bağlantı ağırlık değerlerini

değiştirerek ortalama karesel hatayı minimize amaçladığından dolayı en küçük kareler kuralı olarak da bilinmektedir (Least-Mean-Square Rule LMS). Kuralda minimize edilme işlemi hatanın geriye dağıtılma işlemi ile gerçekleşir. Hata, ağ içinde bir önceki katmana göre çoğaltılır, bu işlemin ilk katmana doğru ilerlemesi ile sonlanır.

Delta kuralı,

$$\Delta W_{ij} = 2\mu x_j \tag{4.6}$$

Bu formülde; Eğitim oranını μ ile i-inci sinapsın j-inci ağırlığındaki değişim ΔW_{ij} ile i-inci sinapsın giriş değerleri x_{ij} ile ifade edilir (Özşahin 2009) (Küçükönder 2011).

4.10.4 Kohonen Kuralı

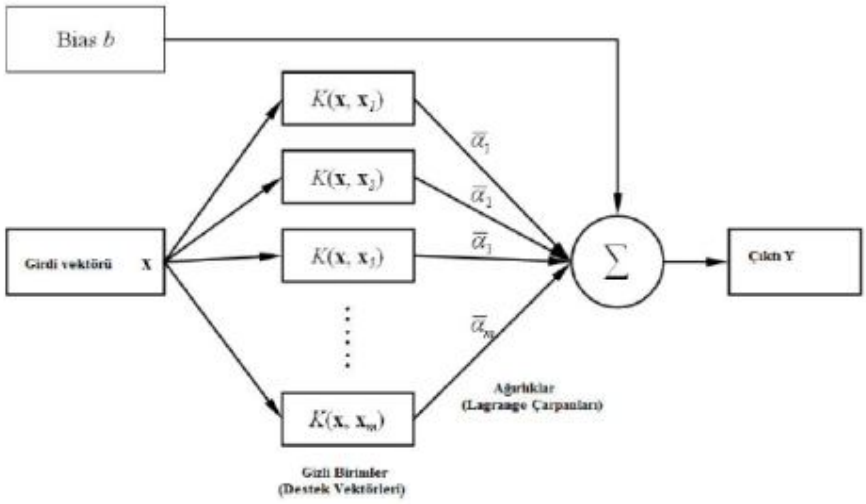
1982 yılında Kohonen tarafından geliştirilen kurala göre ağ kendini eğitebildiği Kohonen tabakasındaki süreç elemanlarının birbiriyle yarış halinde olmasına dayanır. Kohonen tabakasında, işlemci elemanlar ağırlıkların ayarlanması için yarışırlar. İşlemci elemanın referans vektörü girdi vektörüne ne kadar yakın ise o yarışı kazanır (Küçükönder 2011).

5 DESTEK VEKTÖR MAKİNELERİ

Destek vektör makinaları Vladimir Vapnik ve Alexey Chervonenkis tarafından 1960'lı yılları sonunda istatiksel öğrenme teorisini diğer bir ifadeyle Vapnik-Chervonenkis (VC) teorisini temel alan genellikle sınıflandırma, kümeleme yoğunluk tahmini gibi işlemlerde kullanılan bir makine öğrenme algoritmasıdır. Çok güçlü bir teoriğe dayanmasına rağmen 1990'lı yıllara kadar pek takdir görmemiştir. DVM'ler günümüzde birçok işlemlerde istatistiksel yöntemler ve yapay sinir ağlarından daha başarılı ve hızlı sonuç vermektedir (Elmas 2012).

Örnek uzayı lineer olmayan problemlerdeki verilerin, lineer olarak ayrılabilecek en yüksek boyuta aktarılarak, farklı verilerin arasındaki maksimum sınırın bulunmasına dayanan danışmanlı bir öğrenme yöntemidir.

Problemlerde, değişkenler arasındaki örüntülerin bilinmediği veri setlerindeki sınıflama problemleri için önerilmiş bir yöntemdir.



Şekil 5.1 DVM'nin genel yapısı (Ayhan 2013)

Şekil 5.1'de; girdi vektörü (x), girdilerin iç çarpımlarının hesaplandığı çekirdek fonksiyonları (K(xi , xj)), ağırlıkların gösterildiği Lagrange çarpanları (α), çıktı (y) olarak verilen ağ yapısı görünmektedir.

Destek vektör makinalarında sonuç değerini, girdilerin iç çarpımları ile Lagrange çarpanlarının bağımsız kombinasyonu toplamlarına eşittir (Ayhan 2013).

DVM için x yüksek boyutlu giriş vektörü ise ve y çıkış ise x ve y arasında y=f(x) ile ifade edebileceğimiz doğrusal olmayan bağımlılık oluşumu gibi öğrenme problemleri vardır. Ortak olasılık fonksiyonları hakkında bilgi olmadığı için tek bir serbest öğrenme dağılımı gerçekleştirilir.

DVM'lerde klasik istatiksel çıkarımların aksine parametreler eğitim verisinin sayısına göre değiştiğinden önceden belirtilmezler. Modelin hacmini tanımlayacak parametreler hacimle bağlantı kuracak şekilde veri tabanlıdır.

Problem verilerinin birbirinden ayırmak için en kabul edilebilir fonksiyonun tahmin edilmesi esasına dayanır (Özkan 2008). Sınıflandırma mantığının esası, her iki sınıfın uç noktasında ve eğitme örnekleri arasında seçilen destek vektörlerdir. Düzlemin en uygun değeri bulmasıyla maksimum genelleme yeteneğine erişilmiş olur.

DVM'ler genellikle iki sınıflı sınıflandırma problemlerinde kullanılsa da çok sınıflı sınıflandırma problemlerinde kullanılan bir algoritmadır. Çalışılan veri setlerinin doğrusal olarak ayrıla bilirliğine göre *doğrusal* ve *doğrusal olmayan* DVM olmak üzere iki bölüme ayrılır.

5.1 Destek Vektör Makinelerinin Avantajları

a) Sınıflandırma problemini kareli optimizasyon problemine dönüştürüp problemin çözümüne ilişkin öğrenme aşamasında işlem sayısı azalmakta ve diğer algoritmalara göre daha hızlı çözüme ulaştırmaktadır (Osowski ve ark 2004).
b) Doğruluk tahminlerini diğer sitemlere göre çok daha kolay bulur.
c) Anlaşılması zor karar sınırlarını biçimlendirme becerisi vardır
d) Birden çok bağımsız değişkenle görev yapabilmektedir.
e) Doğrusal olarak ayrılan ve ayrılamayan verilere uygulanabilir.
f) Uygun bir dönüşüm ile her zaman veri bir hiper düzlem ile iki sınıfa ayrılabilir.
g) En iyiyi bulma işleminde dışbükey amaç fonksiyonu kullanılarak global en küçük garanti edilir.
h) Genelleme becerisi yüksektir (Elmas 2012).

5.2 Destek Vektör Makinelerinin Dezavantajları

Destek vektör makinalarının; çekirdek seçiminde sınırlama getirilmesi ve verilerin boyutları arttıkça hızının azalması gibi dezavantajları vardır (Elmas 2012).

5.3 Doğrusal Destek Vektör Makineleri

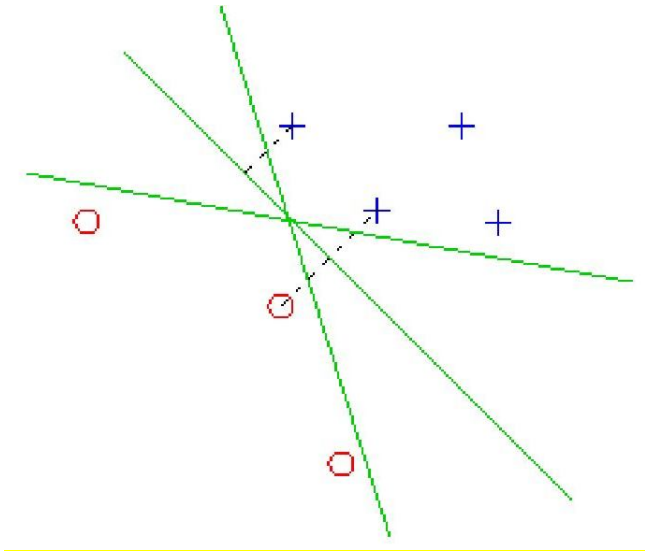
Tanım: Doğrusal Destek Vektör Makineleri

Bir problemde her x_i örneği, niteliğe sahip *girdi* olarak p adet alınırsa, $y_i \in \{-1,1\}$ örneklerin ait olduğu sınıfı temsil eden *çıktı* ve $x \in \mathbb{R}^p$ yüksek boyutlu *girdi vektörü* ise;

w, ağırlık vektörü ve hiper düzlemin normali, b'yide sabit olarak tanımlanırsa; n hacimli (x_i,y_i) ikililerinden oluşan eğitim kümesinde $D = \{(x_1,y_1), (x_2,y_2), (x_n, y_n)\}$ verildiğinde, örneklerin hangi sınıflara ait en iyi şekilde ayıracak,

$$w * x + b = 0 \tag{5.1}$$

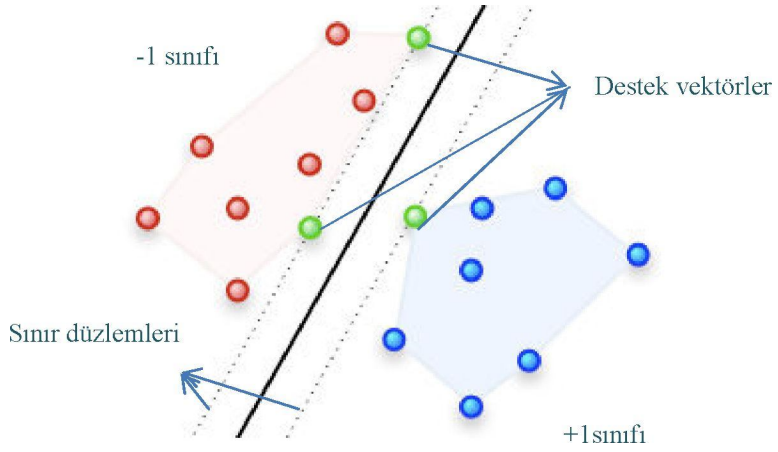
doğrusal hiper düzleminin bulunmasına yardımcı olan bir algoritmadır. Bu algoritma denetimli öğrenme sınıfına ait makine öğrenme algoritmasıdır (Soman ve ark 2011).



Şekil 5.2 2-sınıflı veri setini ayıran farklı düzlemlere ilişkin örnekler(Ayhan 2013)

Şekil 5.2'de örnek bir problemde kullanılan veri setini ayıran düzlemlere ilişkin geometrik sunum belirtilmiştir. Burada görüldüğü farklı sınıfların örneklerini ayıran bir çok doğrusal düzlem olsa da destek vektörleri arasındaki uzaklığı maksimize eden ayırma işlemi hiper düzlemin bulunması işlemini destek vektör makinaları yapar.

Sınır düzlemleri, destek vektörlerinin üzerinde bulunan noktalarla belirtilmiş düzlemler olarak tanımlanır. Hiper düzlem ise oluşan sınır düzlemlerinin ortasından geçen ve her iki düzlemede eşit mesafede olan düzlemdir. Şekil 5.3'te iki boyutlu ve iki sınıfı olan problem örneği için doğrusal DVM'nin geometrik sunumu belirtilmiştir. Ayırma hiper düzlemine en yakın olan iki sınıfa ait en yakın noktalar olarak şekilde yeşil renkte gösterilen noktalar destek vektörleri olarak tanımlanır. Destek vektörler kendi sınıflarının sınırların da hiper düzleme paralel bir düzlem oluşturur (Burges 1998).



Şekil 5.3 2-sınıflı problem örneği için Destek Vektör Makinaları(Ayhan 2013)

5.3.1 Doğrusal ayrılma durumu

DVM'ler eğitim veri setlerini doğrusal olarak ayırabilmek için maksimum sınıra sahip ayırma hiper düzlemini bulmaya çalışır. Ayırma hiper düzlemi için veri setinin örnekleri Denklem 5.2 ve 5.3'ü sağlaması gerekir (Soman ve ark 2011).

$$f(x_i) = \langle w, x_i \rangle + b \geq +1, \qquad y_i = +1, \tag{5.2}$$

$$f(x_i) = \langle w, x_i \rangle + b \leq -1, \qquad y_i = -1, \tag{5.3}$$

Yukarıda belirtilen formüller Denklem 5.4'te belirtildiği gibi tek formülde birleştirilebilir.

$$\forall i \;\; için \;\; y_i(\langle w, x_i \rangle + b) - 1 \geq 0 \tag{5.4}$$

Normali w ile temsil edersek, orijinden dik uzaklığı $\frac{|b|}{\|w\|}$ olan, w.x + b = 0 düzlemine hiper düzlem olarak tanımlanır (Schölkopf ve Smola 2002).

$H_1 = (w, x_i) + b = +1$ düzlemi ve $H_2 = (w, x_i) + b = -1$ düzlemi denklem 5.2 ve 5.3'tende anlaşılacağı üzere paralel düzlemlerdir. *H1* düzleminde *w* normali, $\frac{|1-b|}{\|w\|}$ orjinden dik uzaklığı gösterir. H_2 düzleminde *w* normali, $\frac{|-1-b|}{\|w\|}$ orjinde dik uzaklığı gösterir. *H1* ve H_2 paralel düzlemleri arasında bir eğitim örneği bulunmayıp, düzlem üzerinde eğitim örnekleri bulunur. Bu örnekler hiper düzleme en yakın destek vektörleridir (Burges, 1998). Her iki sınıfın destek vektörleri arasındaki uzaklığı maksimum yapan sınırın ortasından geçen düzlem ayırma düzlemi denir.

$$d(w, b; x) = \frac{|(w, x_i) + b|}{\|w\|} \tag{5.5}$$

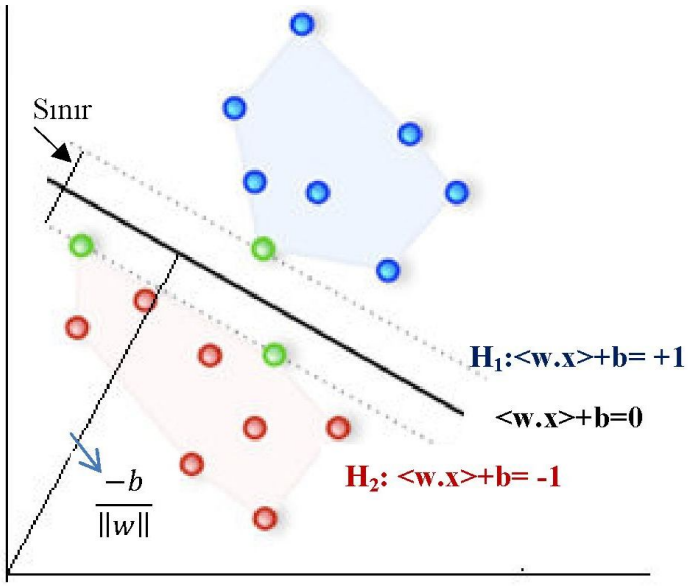
Bir eğitim örneğinin hiper düzleme uzaklığı Denklem 5.5'de verildiği şekilde alındığında sınır değeri olan p Denklem 5.6'daki gibi elde edilir (Gunn 1998).

$$\begin{aligned} p(w,b) &= \min_{x:y=-1}\left(\frac{|\langle w, x_i \rangle + b|}{\|w\|}\right)\min_{x:y=+1} = \left(\frac{|\langle w, x_i \rangle + b|}{\|w\|}\right) \\ &= \frac{1}{\|w\|}(\min_{x:y=-1}|\langle w, x_i \rangle + b| + \min_{x:y=+1}|\langle w, x_i \rangle + b|) = \frac{2}{\|w\|} \end{aligned} \tag{5.6}$$

F0ormülde ağırlık vektörünün olarak belirtilen w normal düzleminin normu ||w|| ile ifade edilir. Hiper düzlememin en yakının da olan örneklerin hiper düzleme olan

uzaklığının tersi onların ağırlık vektörünün normuyla aynı olmalıdır (Schölkopf ve Smola 2002) (Gunn 1998). Belirttiğimiz teoreme göre en iyi hiper düzlem eğitim örneklerini ayıran Denklem 5.7'de belirtilen eşitliliği en küçükleyendir.

$$\Phi(w) = \frac{1}{2}\|w\|^2 \tag{5.7}$$



Şekil 5.4 2-sınıflı problem için doğrusal ayrılabilme durumu(Ayhan 2013)

Şekil 5.4'te, 2-sınıflı sınıflandırma problemine ilişkin doğrusal ayrılabilme durumu gösterilmiştir. Denklem 5.7'deki en küçükleme işlemi, aynı zamanda VC boyutunun üst sınırını en küçükleyerek, yanlış sınıflandırma olasılığını da düşürür (Soman ve ark 2011).

Denklem 5.8' de yukarıdaki bilgilerden ortaya çıkan sonuçlara göre; Denklem 5.7'deki en küçükleme işlemi ile Denklem 5.4'teki kısıt altındaki kareli optimizasyon problemi olarak ifade edilir (Cortes ve Vapnik 1995) (Fletcher 2009).

$$Min\ \frac{1}{2}\|w\|^2$$

$$\text{s.t.} \quad y_i(\langle w, x_i \rangle + b) - 1 \geq 0 \quad \forall i \text{ için} \tag{5.8}$$

Sınıflara ait destek vektörleri arasındaki uzaklığı maksimuma çıkarmak için en uygun ayırma düzlemini Denklem 5.8'deki problemin optimal olarak çözülmesi sonucunda elde edilir (Gunn 1998) (Cortes ve Vapnik 1995).

Destek vektör makinalarında en iyi ayırma hiper düzleminin bulunması için kareli optimizasyon sorununun çözümünde Lagrange çarpanları kullanılır. Bu çarpanlar en küçükleme tipinde problemin daha kolay çözülmesi için çift probleme dönüştürür (Gunn 1998).

Denklem 5.9'da problem çözümünde kullanılan Lagrange fonksiyonu sunulmuştur.

$$L_p(w,b,a) = \frac{1}{2}\|w\|^2 - a\left[y_i(\langle w, x_i\rangle + b) - 1\right] \tag{5.9}$$

$$L_p(w,b,a) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n} a_i\left[y_i(\langle w, x_i\rangle + b) - 1\right] \tag{5.10}$$

$$L_p(w,b,a) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n} a_i\left[y_i(\langle w, x_i\rangle + b)\right] + \sum_{i=1}^{n} a_i \tag{5.11}$$

Yukarıda kullanılan formüllerdeki, her bir a_i (a_i > 0 olmak üzere)Lagrange çarpanı olarak belirtilir. a_i , a_i > 0 olmak üzere dual bir değişken, sabit b'yi en küçük alınmasını sağlayan, w ağırlık vektörü kullanan L_p fonksiyonudur (Burges 1998).

Denklem 5.12 ve 5.13'de Karush-Kuhn-Tucker (KKT) koşullarını elde etmek için Lagrange fonksiyonunun w ve b'ye göre kısmi türevleri alınır.

$$\frac{\partial L_p(w,b,\alpha)}{\partial w} = w - \sum_{i=1}^{n} y_i\alpha_i x_i = 0 \Rightarrow w = \sum_{i=1}^{n} y_i\alpha_i x_i \tag{5.12}$$

$$\frac{\partial L_p(w,b,\alpha)}{\partial b} = 0 \Rightarrow \sum_{i=1} y_i\alpha_i = 0 \tag{5.13}$$

Denklem 5.12'de, daha önce elde edilen eşitlikler Lagrange fonksiyonunda uygun yerlere koyularak problem en büyüğü bulma işlemi için çift Lagrange problemine $(L_D(a))$ çevrilir.

$$MaxL_D(a) = L_D(w,a,b) = \sum_{i=1}^{n} a_i \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j y_i y_j \left\langle x_i x_j \right\rangle$$

$$\text{s.t.} \quad \sum_{i=1}^{n} a_i y_i = 0 \text{ ve } \forall a_i \tag{5.14}$$

Denklem 5.14'de verilen problem çözülerek çift Lagrange $L_D(a)$yı en yüksek yapan a_i değeri elde edilerek en iyi hiper düzlem belirlenir. Burada destek vektörleri, a_i Lagrange çarpanlarından negatif olmayan ve sıfırdan büyük değer alan eğitim örnekleridir. Lagrange çarpanlarının negatif olmayan ve sıfırdan büyük değer alanları en iyi ayırma düzlemini belirler (Burges 1998) (Gunn 1998).

Denklem 5.15'te a_i sonucu bulunduktan sonra en uygun hiper düzlemin ağırlık vektörü w hesaplanır. Ağırlık değeri de bulunduktan sonra Denklem 5.16 ile b sabit değeri bulunur.

$$w^* = \sum_{i=1}^{n} y_i a_i x_i \tag{5.15}$$

$$b^* = -\frac{1}{2} \left\langle w^*, x_s \right\rangle \tag{5.16}$$

Denklem 5.17'te hiper düzleme bağlı sınıflandırıcı elde edilmiştir

$$f(x) = \text{sgn}(\left\langle w^*, x_i \right\rangle + b^*) = sign(\sum_{i=1}^{n} y_i a_i \left\langle x_i x_j \right\rangle) \tag{5.17}$$

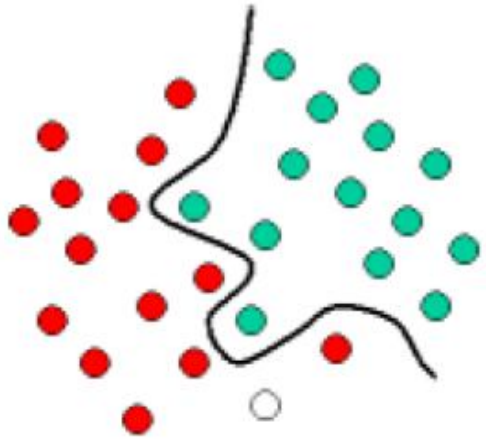
Denklem 5.17' teki sınıflandırıcıdan gerçek değerler elde edilmesinden dolayı Denklem 5.18 ile sınıflandırmak işleminin yapılması daha uygun olur (Gunn 1998).

$$f(x)=h(\langle w^*,x_i\rangle+b^*)\Rightarrow h(z)=\begin{cases}-1 & z<-1\\ z & -1\le z\le+1\\ +1 & z>+1\end{cases} \tag{5.18}$$

5.3.2 Doğrusal Olmayan Destek Vektör Makineleri: Haritalama Yaklaşımı ve Çekirdek Fonksiyonlarının Kullanımı (Kernel trick)

Verilerin belirli bir hata oranıyla ya da tam olarak doğrusal bir fonksiyonla ayrılamadığı durumlarda doğrusal olmayan destek vektör makinaları kullanılır. Uygulamalarımızı teoriden çıkarıp gerçek veriler üzerinde uygulamaya başladığımızda elimizdeki verilerin hiper düzlemle doğrusal olarak ayrılması güç olduğundan sınıflara ayırma işlemi, ayırma eğrisinin tahmin edilmesi güç de olsa tahmin yöntemiyle yapılır.

Şekil 5.5'te bir veri grubunun doğrusal ayrılamama halinin geometrik olarak gösterilmektedir.



Şekil 5.5 2-sınıflı problem için doğrusal ayrılamama durumu(Ayhan 2013)

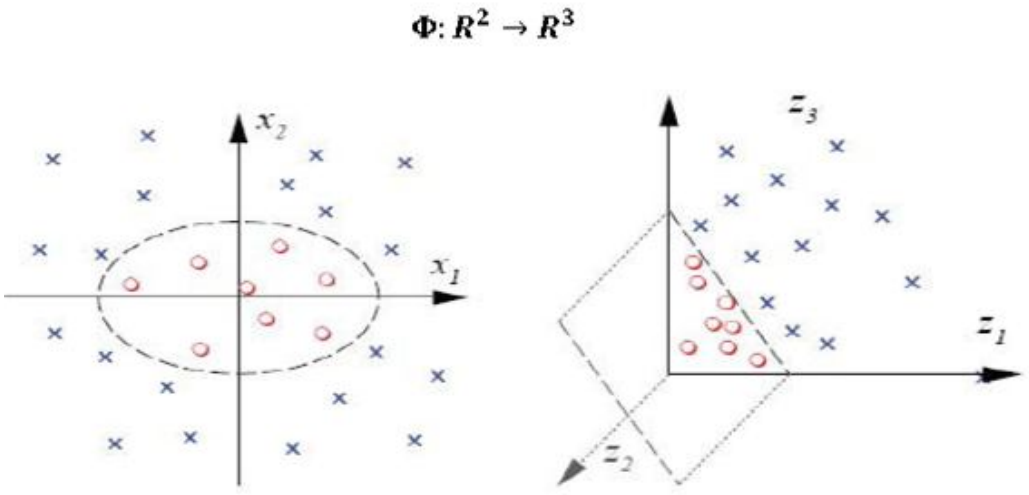
Verilerin doğrusal olmama durumunda girdi boyutu *p* olan vektörü **x**'in *P*-boyutlu özellik vektörü Φ'ye çevrilmesi gerekmektedir (Cortes ve Vapnik 1995).

p-boyutlu girdi vektörü **x** 'in *P*-boyutlu özellik vektörü Φ'ye dönüştürülebilmesi

x, *p*-boyutlu girdi vektörünü temsil ettiği durumda, x'in *P*-boyutlu özellik vektörü Φ'ye dönüştürülebilmesi için en uygun ayırma düzleminin özellik uzayında gerçekleşmesi için doğrusal olmayan haritalama yönteminden yararlanılır (Busuttil 2003).

Orijinal girdi uzayını x ile temsil edersek yüksek boyutlu F özellik uzayı olan Hilbert uzayına dönüştürülerek doğrusal olmayan verilerin doğrusal ayırma işlemine tabi tutulmasına sağlayan yaklaşıma "doğrusal olmayan haritalama" yöntemi denir (Suykens 2002). Hilbert uzayı pozitif skaler çarpıma sahip ve ögeleri fonksiyonlardan oluşan tam i çarpım uzayları olarak ifade edilmektedir (Çakar 2007).

Şekil 5.6'da iki boyutlu doğrusal olmayan verilerin haritalama yaklaşımı ile doğrusal olmayan haritalama yaklaşımı iki boyutlu veri setinin üç boyutlu özellik boyutuna taşınarak doğrusal ayrımın geometrik olarak sağlanması gösterilmiştir.



Şekil 5.6 Doğrusal olmayan haritalama yaklaşımı örneği(Ayhan 2013)

Buradaki (X_1 , X_2) ile temsil edilen iki boyutlu girdi vektörü ve (Z_1,Z_2,Z_3) ile temsil edilen üç boyutlu özellik uzayının fonksiyonlarının Denklem 5.19 ile eşitliği gösterilmektedir. Denklem 5.20'de ise özellik uzayının haritalanmış girdi vektörlerinin

iç çarpımlarının elde edilmesi formülüze edilir. Bu işlemler sonucunda iki boyutlu uzaydan üç boyutlu uzaya dönüştürülmüş olur (Soman ve ark 2011).

$$\Phi : R^2 \to R^3 \qquad (x_1, x_2) \to (z_1, z_2, z_3) := (x_1^2, \sqrt{2}x_1x_2, x_2^2) \tag{5.19}$$

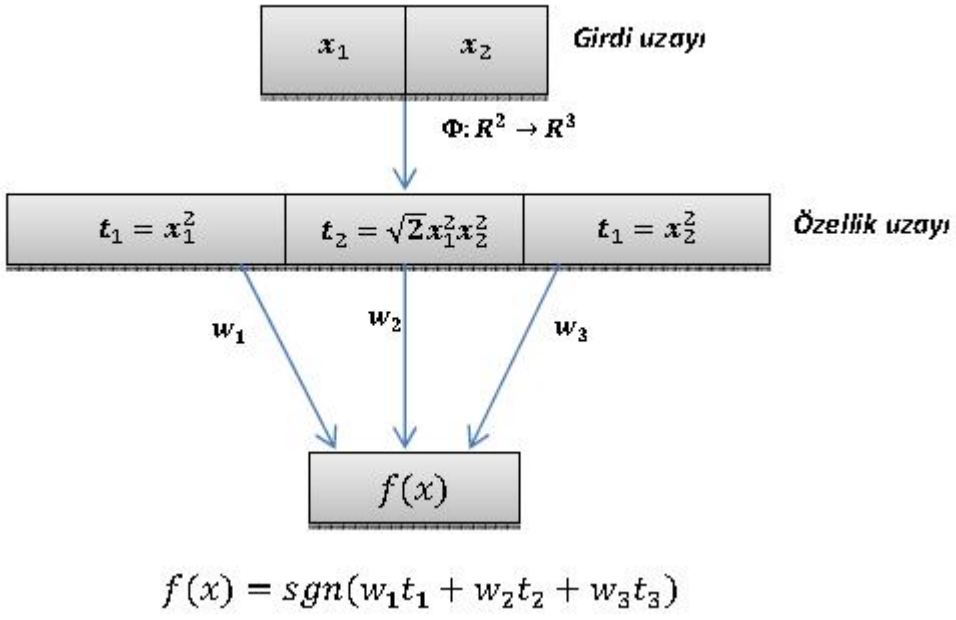
$$\Phi(x)^T \Phi(x) = (x_1^2, \sqrt{2}x_1x_2, x_2^2)(x_1^2, \sqrt{2}x_1x_2, x_2^2)^T$$

$$= (x_1^2 x_1'^2 + 2x_1x_1'x_2x_2' + x_2^2x_2'^2)^2 \tag{5.20}$$

$$f(x) = sign(\langle w^*, x_i \rangle + b^*) = sign(\sum_{i=1}^{n} y_i a_i \langle \phi(x_i), \phi(x_j) \rangle) \tag{5.21}$$

Denklem 5.21 ile destek vektör makinaları için özellik uzayında ayırma hiper düzleminin sınıflandırıcı karar fonksiyonu tanımlanmıştır.

Şekil 5.7'de destek vektör makinalarında haritalama yaklaşımı kısaca anlatılmıştır.



Şekil 5.7 Girdi uzayının özellik uzayına dönüştürülmesi gösterilmiştir (Ayhan 2013).

Bir **x** girdi vektörü için Φ(x), k boyutlu Öklid vektör uzayında basit bir vektör olarak düşünülse de özellik uzayının boyutu, girdi uzayının boyutundan daha büyük olduğu için haritalama kısmının karmaşıklığı artarak hesaplamaları zorlaştırmaktadır.

$$K(X,X') = (X,X'^2) = ((x_1,x_2).(x_1,x_2)')^2 = (x_1x_1' + x_2x_2')^2$$

$$= (x_1^2x_1'^2 + 2x_1x_1'x_2x_2' + x_2^2x_2'^2)^2 \quad (5.22)$$

Özellik uzayında haritalanmış girdi vektörlerinin iç çarpımlarının elde edildiği Denklem 5.20 ile çekirdek fonksiyonunun elde edildiği Denklem 5.22 aynı sonucu verdiği ortaya çıkmaktadır. Bu sebeple Denklem 5.23 da özellik uzayında haritalanmış girdi vektörleri yerine çekirdek fonksiyonlarını n kullanılmasıyla ortaya çıkan karar fonksiyonu gösterilmiştir. Böylece hesaplamaları kolaylaştıracak, maliyetlerini düşürecek ve iç çarpımların yapılmasını kolaylaştıracak çekirdek fonksiyonlarının kullanılması bir çok literatür de önerilmiştir. Bu olay literatürde *kernel trick* olarak adlandırılır (Chapelle 2007).

$$f(x) = sign(\langle w^*, x_i \rangle + b^*) = sgn(\sum_{i=1}^{n} y_i a_i K(x_i, x_j)) \quad (5.23)$$

5.3.3 Çekirdek (Kernel) Fonksiyonları

Çekirdek fonksiyonları, sınıflandırma işlemleri için algoritmalarda benzerlik ölçütü olarak kullanılır. Doğrusal olarak ayrılmaya imkân vermeyen veri setlerinde, girdi uzayını yüksek boyutlu özellik uzayına dönüştürerek doğrusal ayrımın gerçekleştirmesine olanak sağlar (Eray 2008) (Li ve ark 2009). Çekirdeği, $K: X \, x \, X \rightarrow R$ olarak tanımlarsak X'teki tüm x_i ve x_j'ler için Denklem 5.24'deki formülü sağlayan simetrik fonksiyona çekirdek fonksiyonu olarak tanımlanır (Vanschoenwinkel ve Manderick 2004).

$$K(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle \quad (5.24)$$

Çok boyutlu uzayda yer alan örneklere ilişkin değerlerin ikili çarpımları hesaplamak yerine çekirdek fonksiyonunda örnek değeri yerine koyarak özellik uzayındaki değeri daha kolay hesaplanması sağlanır (Li ve ark 2009).

Destek vektör makinalarında Mercer teoremini sağlayan simetrik fonksiyonlar çekirdek fonksiyon olarak kullanılabilir. Bu yüzden çekirdek fonksiyonu olarak sürekli, simetrik, pozitif tanımlı olma koşulunu sağlayan fonksiyonlar kullanılır.

Simetrik matris, nxn boyutunda bir K matrisinin transpozu kendisine eşitse o matrise simetriktir denir. Pozitif tanımlı matris kavramında K simetrik matris olmak üzere, tüm elemanları sıfırdan farklı X vektörü için $X^T KX>0$ şartını sağlayan K matrisine denir. Bu eşitlikte $K>0$ ve $X=0$ olması durumunda, $X^T KX=0$ elde edilir. Pozitif yarı tanımlı olması durumu ise; $X^T KX>=0$ ise $K>=0$ olması ile gerçekleşir. $X^T KX=0$ eşitliğinde K, pozitif yarı tanımlı ise eşitliği sağlayan farklı bir X vektörü vardır (Wiskott 2006).

Çekirdek fonksiyonun pozitif tanımlıları birimler arasındaki benzerlik ölçütlüğü yaparken iç çarpımlarla hesaplanır. Bu hesaplama bir matrisle gösterilir. Bu matrise *GRAM matris* denir (Cristianini ve Shawe-Taylor 2000).

Bir *GRAM* matrisi simetrik ve pozitif yarı tanımlı olması gerekir. Girdi vektörünü $X=x_1, x_2, \ldots x_n$ alırsak, $K(x_i,x_j)$ fonksiyonu için

$$K = \begin{pmatrix} \langle \phi(x_1)\phi(x_1) \rangle & \cdots & \langle \phi(x_1)\phi(x_n) \rangle \\ \vdots & \ddots & \vdots \\ \langle \phi(x_n)\phi(x_1) \rangle & \cdots & \langle \phi(x_n)\phi(x_n) \rangle \end{pmatrix} \tag{5.25}$$

formülünden *GRAM matrisi* elde edilir (Ayhan 2013).

Optimizasyon problemlerinde çözümün tek olmasını sağlayan çözümlerde pozitif tanımlı çekirdeğin kullanılmasıdır (Soman ve ark 2011).

Çizelge 5.1'de Destek vektör makinaları uygulamalarında önemli rol oynayan çekirdek fonksiyonlarının en çok kullanılması önerilenleri formülleriyle açıklanmıştır.

Çizelge 5.1 DVM'de kullanımı uygun olan çekirdek fonksiyonları (Ayhan 2013)

Çekirdek Fonksiyonu	Matematiksel Gösterimi

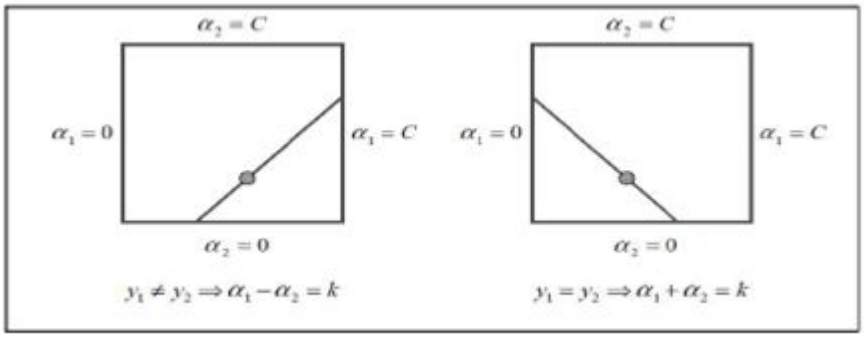
Doğrusal	$k(x,x') = x^T x' + c$
Polinomial	$k(x,x') = (ax^T x' + c)^d$
Gaussian	$k(x,x') = \exp(-\frac{\|x-x'\|^2}{2\sigma^2})$
Eksponansiyel	$k(x,x') = \exp(-\frac{\|x-x'\|}{2\sigma^2})$
Hiperbolik Tanjant (Sigmoid)	$k(x,x') = \tanh(ax^T x' + c)$
Anova	$k(x,x') = \sum_{k=1}^{n} \exp(-\sigma(x^k x'^k)^2)^d$
Wave (Dalga)	$k(x,x') = \frac{\theta}{\|x-x'\|} \sin\frac{\|x-x'\|}{\theta})$
Genelleştirilmiş t-Student	$k(x,x') = \frac{1}{1+\|x-x'\|^d}$

5.4 Ardışık Minimal Optimizasyon Algoritması (SMO)

Bu algoritma her safhada uygun olan en küçük optimizasyon çözümüne ulaşmaya çalışır. Her safhada iki Lagrange çarpanı için en optimum değeri bularak DVM güncellemeleri için en optimum değerleri gerçekleştirir.

SMO'lar matris algoritması kullanılmadığından hassaslığa daha az duyarlı olduğu için büyük eğitim verilerinde destek vektör makinaları işlemlerini daha az güç gerektiren bilgisayarlarda da çözülmesine olanak sağlar. Analitik iki Lagrange çarpanın çözebilmesi üstün yönlerinden bir diğeridir.

Şekil 5.8'de belirtilen iki Lagrange çarpanının sonucunun bulunmasında kullanılan analitik metot ve çarpanların optimizasyonunu sağlayan sezgisel metot olmak üzere SMO'lar iki bileşenden oluşur (Platt 1998).



Şekil 5.8 İki Lagrange çarpanının gösterilmesi (Elmas 2012)

Şekil 5.8'de eşitsizlik kısıtı Lagrange çarpanına neden olur. Doğrusal eşitlik kısıtlaması diyagonal bir çizgiye sebep olur. SMO bu çizgide amaç fonksiyonunun en iyisini bulmayı amaçlar (Platt 1998) (Elmas 2012).

Denklem 5.26, DVM ilkel optimizasyon sorununun ikili gösterimini gösterir.

$$\max_{\alpha} \Psi(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} y_i y_j k(x_i, x_j) \alpha_i \alpha_j$$
$$\text{tabi} \sum_{i=1}^{N} y_i \alpha_i = 0, \; 0 \leq \alpha_i \leq C, \quad ve \; i = 1, ..., n \tag{5.26}$$

burada, x_i bir eğitim örneği, $y_i \in \{-1,+1\}$ karşılık gelen hedef değer olması durumunda, α_i Lagrange çarpanı ve C gerçek değer maliyet parametresidir.

SMO algoritmasının basitleştirilmiş kodu Şekil 5.9'da gösterilmiştir. Birinci bölümde, değişkenlerin başlangıç değerleri atanır. Döngü üç temel işlem içerir. Bunlar; ortak optimizasyonunu takiben iki Lagrange çarpanının seçimleri ve DVM parametreleri güncellenmesidir (Kuan ve ark 2012).

İlk Parametreler
w=0, b=0 and all α_i=0, E=0

repeat

Adım 1. Ön İşlem adımları:

1.1. $k_{ii} = x_i^T x_i, k_{jj} = x_j^T x_j, k_{ij} = x_i^T x_j$

1.2. $\eta = 2k_{ij} - k_{ii} - k_{jj}$

1.3. Calculate_prediction_error E_i, E_j and $\alpha_i^{old}, \alpha_j^{old}, \gamma$.

1.4. Determine_feasible_range [L,H] kırpma için kullanılan $\alpha_j^{new,unc}$

Adım 2. Lagrange çarpanının güncelleme adımları:

2.1. $\alpha_j^{new,unc} = \alpha_j^{old} + \frac{y_j(E_j - E_i)}{\eta}$ and Clipping to L or H

2.2. $\alpha_i^{new} = \alpha_i^{old} + y_i y_j(\alpha_j^{old} - \alpha_j^{new})$

Adım 3.DVM Parametre Güncelleme adımları:

3.1 güncelleme w, b, Δb

3.2 Tahmini hata güncellemesi E.

until (all α satisfied KKT condition)

Şekil 5.9 SMO algoritmasının küçültülmüş kodu (Kuan ve ark 2012)

6 GELİŞTİRİLEN ALGORİTMALARDA KULLANILAN VERİ KÜMELERİ

Bu bölümde, Bölüm 7.1 ve Bölüm 7.3 arasında geliştirilen algoritmalarda kullanılan veri kümeleri açıklanmıştır. Bu çalışmalarda kullanılan veri kümeleri göğüs kanseri, kalp hastalığı, karaciğer hastalığı kümeleridir. Hastalığın temeli, hastalık hakkındaki biyolojik ve medikal bilgiye önem verir. Hastalığı tespit etmede kullanılan veri setlerinin niteliklerinin yanında hastalıklar hakkında da bu bölümde kısa bilgi verilmiştir. Veri tabanı bilgisi, veri tabanını büyüklüğü ve eksik özelliğe sahip örneklerin veri tabanında olup olmadığı, eksik verilere nasıl bir işlem uygulandığı bilgileri de anlatılmıştır. Bu bölümde tanıtılacak veri tabanları makine öğrenme yöntemlerinde sınıflama işlemlerinde araştırmacılar tarafından en çok kullanılan veri tabanlarıdır. Geliştirdiğimiz algoritmaları diğer araştırmacılarla karşılaştırmak için bu bölümde tanıtılan veri tabanları seçilmiştir. Veri tabanları Irvine'deki California Üniversitesi, Bilgisayar Bilimi ve Bilgi Bölümünden, makine öğrenmesi repertuarından elde edilip kullanılmıştır.

6.1 Göğüs Kanseri Veri kümesi

Göğüs kanseri, sıklıkla bir kütle dokusuyla başlayan ve göğüste büyümeye başlayan kötü huylu bir dokudur. Dünyada görülme sıklığı Akciğer kanserinden sonra, en yüksek olan kanser türüdür. Hayatının belirli bir zamanında, her 8 kadından birinin meme kanserine yakalanabileceği düşünülmektedir. Göğüs kanseri, erken tespit edilirse, hasta %96 ölmeden kurtulma ihtimali vardır. (http://tr.wikipedia.org/wiki/Meme_kanseri 2015). Erkeklerde çok küçük oranda da olsa görülme ihtimali vardır.

Bu kanser, göğüs kütlesi, göğüsün büyüklüğünde ve şeklinde değişiklik, göğüsün derisinin hassaslığı veya rengindeki değişiklik, meme ucunun görünümündeki değişik, göğüs ağrısı ve meme ucu değişiklikleri v.b. gibi belirtilerle ortaya çıkar.

Kullanılan göğüs kanseri veri tabanı, Wisconsin Hastanesinde Dr. William H. Wolberg tarafından oluşturulmuştur. Veri kümesi 699 örnekten oluşmaktadır. 9 tamsayı değerli özelliğe ve her bir veri 1' den 10' a aralık değerine sahiptir. Çizelge 6.1'de veri setinin nitelikleri ve aldıkları değer aralıkları verilmiştir. Veri kümesi özniteliklerinde

16 örnekte eksik değerlere sahiptir. Eksik değerler yerine işlemlerimizde kullanmak için kendi sınıf değeri içinde o nitelikte en çok kullanılan değeri atanmıştır.

Çizelge 6.1 Göğüs kanser tespitinde kullanılan nitelikler

Nitelik sayısı	Nitelik ismi	Özelliğin açıklaması ve değeri
1	Kütle kalınlığı	1-10
2	Hücre Büyüklüğünün Benzerliği	1-10
3	Hücre Şeklinin Benzerliği	1-10
4	Marjinal Yapışma	1-10
5	Tek Epitel Hücresi Büyüklüğü	1-10
6	Çıplak Çekirdek Kütlesi	1-10
7	Hafif Çekirdek Kütlesi	1-10
8	Normal Çekirdek Kütlesi	1-10
9	Mitoz	1-10

Bu nitelikler, dokuz farklı ölçekte dış görünüm ve iç kromozom değişikliklerinden elde edilen ölçümlerdir. Göğüs kanserinin sınıf değişkeninde 458 adet iyi huylu (benign) ve 241 adet kötü huylu (malignant) olmak üzere iki sınıf vardır. Bu sınıflar sırayla 0 ve 1 nümerik değerini alır.

6.2 Kalp Hastalığı Veri kümesi

Her yıl 17,3 milyon kişi kalp hastalıkları nedeniyle hayatlarını kaybeder (WH. 2012). Bu ölümlerin çoğu ilk kalp krizi sırasında meydana gelir. Birçok hasta o ana kadar kalp hastalığından habersizdir. Kalp hastalığı, genellikle kalbin kan odacıklarını veya kalp kasını etkiler. Kalp rahatsızlığı, kan odacığının darlaşması veya kan akışının blokajı gibi problemle sıklıkla kan odacıklarında oluşur. Kalp hastalığının göğüs ağrısı, aşın yorgunluk, yüksek kan basıncı ve kan şekeri gibi belirtileri vardır.

6.2.1 Cleveland Kalp Hastalıkları Veri Kümesi

Cleveland Kalp hastalıkları verisi, Long Beach ve Cleveland Klinik Kuruluşu Medikal Center da Dr. Robert Detrano tarafından kalp hastalığı teşhisinde kullanılabilecek 76 test sonucunun en önemli 13'ünün kullanılması ile oluşturulmuştur. Veri tabanı 303 örnek içermektedir. Örneklerden 165'i hasta ("0"), 138'i sağlıklı("1") sınıfına aittir. Veri kümesindeki örneklerin 13 niteliği, Çizelge 6.2'de verilmiştir:

Çizelge 6.2 Cleveland Kalp hastalıkları veri kümesi de bulunan nitelikler ve aralıkları

Nitelik sayısı	Nitelik ismi	Niteliğin açıklaması ve değeri
1	Yaş	29 – 77
2	Cinsiyet	0: Erkek , 1: Kadın
3	Göğüs Ağrı Tipi	1:anjin, 2: asympt, 3:notang, 4:abnang
4	Dinlenme kan basıncı	94 – 200
5	Serum kolesterol	126 – 564
6	Açlık kan sekeri > 120	1: doğru, 0 :yanlış
7	Dinlenme elektrokardiyografi	0: normal, 1: Anormal olan ST-T dalga oluşması, 2: Sol ventrikular hypertropy olasılığının gösterilmesi
8	Elde edilen maksimum nabız	71 – 202
9	Anjine sebep olan egzersiz	1 : evet , 0:hayır
10	Dinlenmeye kıyasla egzersizle ortaya çıkan ST depresyonu	0 – 6.2
11	Egzersiz ST segment tepesinin egimi	1: Eğimin artması, 2: Düz ,3: Eğimin azalması
12	Floroskopiyle renklendirilen ana damarların sayısı	0-3
13	Thal	3: Normal, 6: Değişmez hasar,7: Eski hasar

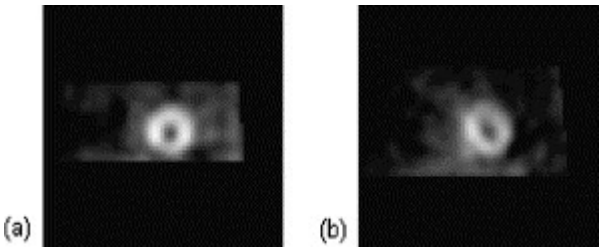
Bu veri kümesinde, %50 çap daralmasından daha az veya %50 çap daralmasından daha fazla olmasından oluşan kalp hastalığının teşhisi için iki çıkış vardır ve sırayla "0" veya "1" ile gösterilir.

6.2.2 Spect Veri kümesi

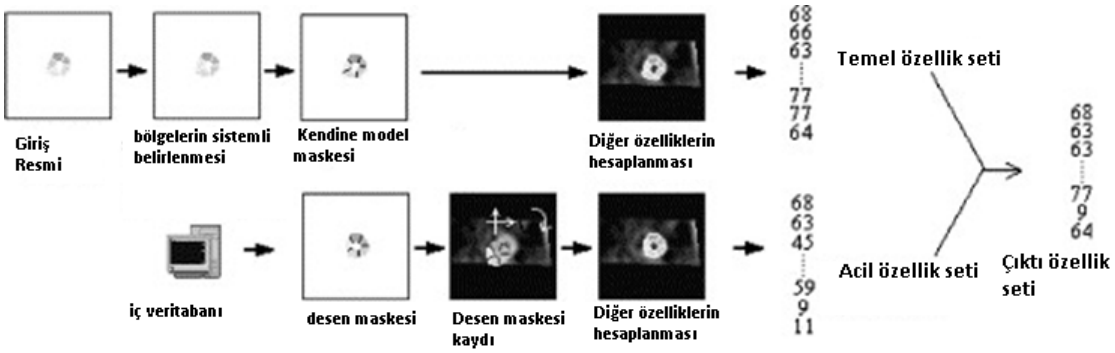
SPECT görüntülerinden elde edilen veri kümesi kalp hastalıklarının tanısı ile ilgilidir. Kalp Tek Proton Emisyon Bilgisayarlı Tomografi (SPECT) görüntülerinden teşhis verilerinin tanımlandığı kümedir. Kalbin bozukluğunun tespiti için, ilk aşamada ventrikül kaslarının bölgesel teşhisleri yapılır. İkinci aşamada sol ventrikül bölge teşhisinden tüm kalp ile ilgili sonuca varılmıştır. Veri seti oluşturulurken son olarak kalbin normal veya çalışmasının normal olmadığı yönünde iki sınıfa ayrılmıştır.

Normal 55 (0) ve normal olmayan 212 (1) veriyi içerir. Orjinal Spect görüntülerinden nitelikler çıkartılarak 267 adet Spect görüntü kümesinden veri seti oluşturulmuştur. SPECT görüntülerinden elde edilmiş desen veri kümesinde 22 nitelik, 0 veya 1 ile temsil edilmiştir.

Kalp spect görüntülerinin serpilmesi Şekil 6.1'de , görüntünün elde edilmesi ve dönüştürülmesi Şekil 6.2'de gösterilmiştir.



Şekil 6.1 Kalp SPECT görüntülerinde serpilme: (a) Normal serpilme; (b) normal olmayan serpilme.



Şekil 6.2 Kalp SPECT görüntülerinden özellik el edilmesinin grafiksel gösterimi(Kurgan ve ark 2001)

Görüntü analizi algoritması amacı kalp SPECT çalışmalarında 44 nitelikleri çıkarmaktır. Her nitelik belirli bir "ROI" de LV (sol kapakçık) maskesinin perfüzyonunu temsil eden radyoaktif sayıları ölçen bir tek sayıdır.

Her bir nitelik, aşağıdaki denkleme göre hesaplanır:

$$feature_i = \frac{\sum_R C(m,n)}{N(R)} \times \frac{100}{m_j}(\%) \qquad (6.1)$$

R, geçerli bölge (ROI biri) X,

N (R) R ait piksel sayısı

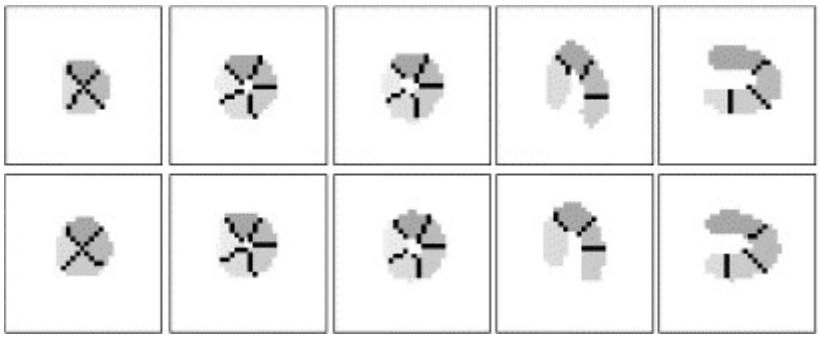
C(m, n) : R nin (m, n) koordinatının ait olduğu piksel değeri (Pikselin parlaklığı)

mj: LV iki ilgili görüntülerinden alınan pikselin maksimum değeri.

Özelikler, şablon tarafından tanımlanan o anki ROI de bulunan piksel ortalama sayısı olarak hesaplanır. Bu değer dinlenme ve stres çalışması olarak tanımlanan görüntülerin en yüksek parlaklık yüzde değeriyle kalibre edilir.

Normal LV modeli oluşturmak için veri tabanından kardiyolog tarafından daha önce tanısı normal olarak konan 20 vaka (10 kadın ve 10 erkek) alınmıştır. Model sırasıyla, erkek hastaların 10 resim setleri ve kadın hastaların 10 resim setleri ortalamasını alarak inşa edilmiştir. Normal LV Modeli, tüm ROI lokalizasyonu hakkında bilgi içerir. Bu teşhis işlemi sırasında kardiyolog tarafından seçilen 10 dilim karşılık gelen 10 görüntü oluşmuştur. Şekil 6.3'te kadın için normal bir LV şablon modelini gösterilmektedir.

Görüntü analizi algoritması çıkış sırasıyla, stres ve dinlenme çalışmaları için, 22 ROI'lardaki serpilme temsil eden 44 nitelikler kümesi oluşur. Veriler elde edilirken, ROI'lerin birinden serpilmeleri açıklayan 22 kısmi tanı özelliği oluşturulmuştur(Kurgan ve ark 2001).



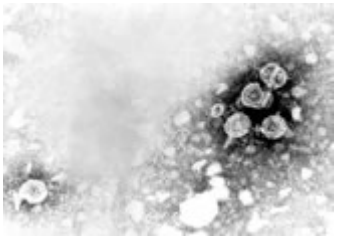
Şekil 6.3 Kadın hastalarda normal LV Şablon-modeli (Kurgan ve ark 2001).

6.3 Karaciğer rahatsızlığı tanımı

Karaciğer rahatsızlıkları tüm organizma üzerinde ağır etkiler oluşturabilir. Bu rahatsızlıklar genellikle, vücudumuzda fazla alkolün bulunmasıyla meydana gelir.

6.3.1 Karaciğer rahatsızlığı veri kümesi

Bu veri tabanı, BUPA Medikal Araştırma Şirketinde, Richard S.Forsyth tarafından oluşturulmuştur. Veri tabanı toplam olarak 345 örnekten oluşmaktadır. Veriler içinde hiçbir eksik nitelik veya değer yoktur. Veri tabanı, ilk beş özelliği kan testi olan 6 nitelikten oluşmaktadır (Blake ve Merz 1998). Çizelge 6.3 Liver veri kümesinde yer alan nitelikler ve onların aldığı değer aralıkları verilmiştir. Veri tabanı oluşturulurken kullanılan kan testleri, fazla alkol tüketiminden meydana gelebilen karaciğer rahatsızlıklarına duyarlıdır. Şekil 6.4'te bilgilerin elde edildiği karaciğerlerden birinin görüntüsü verilmiştir.



Şekil 6.4 Liver görüntüsü

Çizelge 6.3 Liver veri kümesi de bulunan nitelikler ve aralıkları (Blake ve Merz 1998)

Nitelik sayısı	Nitelik ismi	Özelliğin Açıklaması	Özelliğin Değer Aralığı
1	mcv	Ortalama korpüskulor hacim	65-103
2	alkphos	Safra salgısının hücrelerinde bulunan protein	23-138
3	sgpt	Hepatocellular nekrozda kan seviyesi yükselen bir çeşit aminotransferase.	4-155
4	sgot	Hepatocellular nekrozda kan seviyesi yükselen bir çeşit aminotransferase.	5-82
5	gammagt	Kandaki GGT enzim miktarını ölçen bir test	5-297
6	drinks	Günlük alman alkol miktarı	0-20

Bu veri tabanının normal ve hasta olmak üzere iki sınıfı vardır. Normal sınıf değeri "0" ile hasta sınıf değeri "1" ile temsil edilir.

6.4 Hepatit Veri Kümesini

Hepatit veri kümesi, Jozef Stefan Institute'den Bojan Cestnik tarafından oluşturulmuştur. Veri kümesini oluşturmaktaki amaç bu hastalığa sahip insanların yaşamaya devam edip edemeyeceğini araştırmaktır. Bu veri kümesi hepatit testlerinden 19'unun kullanılması ile oluşturulmuştur. Veri kümesi, 32'si ölmüş, 123'ü yaşamına devam toplam 155 örnekten oluşturulmuştur.

Sonuçlar oluşturulurken iki sınıfla temsil edilir. Kişinin yaşamına devam ettiğini "0", öldüğü "1" temsil etmektedir. Çizelge 6.4'te veri tabanında kullanılan niteliklerin açıklamaları ve aldığı değerler (değer aralıkları) görülmektedir.

Çizelge 6.4 Hepatit veri kümesi de bulunan nitelikler ve aralıkları

Nitelik sayısı	Nitelik ismi	Özelliğin Değer Aralığı
1	Yaş	7-78
2	Cinsiyet	Erkek , Kadın
3	Steroid	Evet, Hayır
4	Antivirals	Evet, Hayır
5	Halsizlik	Evet, Hayır
6	Kırıklık	Evet, Hayır

7	İştahsızlık	Evet, Hayır
8	Karaciğer büyümesi	Evet, Hayır
9	Karaciger sertlesmesi	Evet, Hayır
10	Dalak Palpabl	Evet, Hayır
11	Deri üzerinde örümceğe benzer ben	Evet, Hayır
12	Asit	Evet, Hayır
13	Varis	Evet, Hayır
14	Bilirubin	0.3-8
15	Alkalen fosfataz	26-295
16	Sgot	14-648
17	Albumin	2.1-6.4
18	Protime	0-100
19	Histoloji	Evet, Hayır

6.5 Diabet hastalığı

Genellikle genetik ve çevresel etkenlerin birleşimi ile oluşur. Kan glukoz seviyesinin normalden daha fazla yükselmesiyle (hiperglisemi) neticelenen metabolik bir bozukluktur (http://tr.wikipedia.org/wiki/Diyabet 2015). Kan şekerinin, vücutta düzenlenmesi çok sayıda hormonun ve kimyasalın karmaşık etkileşimi sonucunda sağlanır. Şeker metabolizmasının düzenlenmesinde en önemli rol pankreasın beta hücrelerinden salgılanan insülin hormonudur. Diabetes, insülin salgılanmasındaki eksiklik veya insülinin etkisindeki bir düzgün olmayan durumun sonucunda ortaya çıkan yüksek kan şekerinin yol açtığı hastalıkları tanımlamak için kullanılan bir terimdir (http://tr.wikipedia.org/wiki/Diyabet 2015).

6.5.1 Diabetes Veri Kümesi

Diyabet hastalıkları veri kümesi Black C.L. tarafından oluşturulup kullanılmıştır (Blake ve M.C.J. 1998). Veri seti 8 nitelik içerip, 768 örnek üzerinde çalışılmıştır. Veri kümesindeki tüm hastalar en az 25 yaşında olan ABD'de Arizona eyaletinin Phoenix yakınlarında yaşayan Pima Hintli kadınlarından elde edilmiştir. Sınıflama sonucu diyabet olan hastalar "1", olamayanlar "0" ile temsil edilecek şekilde veri kümesine işlenmiştir. Sınıfta 268 diyabet hastası, 500 diabet olmayan hasta vardır (Polat ve ark

2008). Çizelge 6.5'te diabetes veri kümesinde yer alan nitelikler ve onların aldığı değer aralıkları hakkında bilgi verir.

Çizelge 6.5 Diabetes veri kümesi de bulunan niteliklerin değerleri ve aralıkları

Nitelik	Ortalama	Standart sapma	Min	Maks.
Kaç kez Gebe	3.8	3.4	0	17
Plazma glükoz konsantrasyonu, bir oral glükoz tolerans testinde 2 saatlik	120.9	32.0	0	199
Diastolik kan basıncı(mm Hg)	69.1	19.4	0	122
Triceps deri kıvrım kalınlığı (mm)	20.5	16.0	0	99
2 saatlik serum insülin (mu U/ml)	79.8	115.2	0	846
Vücut kitle indeksi (kg/m^2)	32.0	7.9	0	67.1
Diyabet soyağacı fonksiyonu	0.5	0.3	0.078	2.42
Yaş	33.2	11.8	21	81

7 GELİŞTİRİLEN ALGORİTMALAR

7.1 Apriori ve TBA ile Oluşturulmuş Hibrid Özellik Seçim Sistemi

Veri setlerinde, veriyi temsil eden niteliklerin tamamı bütün problemlerin kullanımı için uygun olmayabilir. Uygulamalar için önem arz etmeyen nitelikler bulunabilir. Bu ifadeler sınıflama işlemlerinde ve problem çözümlerinde dezavantaj yaratır. Geliştirdiğimiz sistem de veri setlerinde bulunan gereksiz niteliklerin elenmesiyle daha hızlı, daha az bellek kapasitesiyle işlem yapılması sağlanarak sınıflama başarısı artırılmıştır.

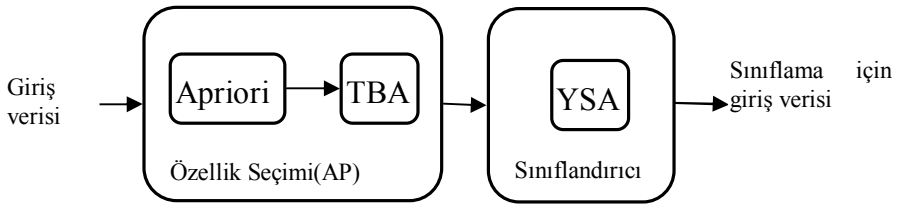
Büyük veri tabanlarında öznitelikler arasındaki ilişkiyi bulmakta kullanılan birliktelik kurallarından biri olan Apriori algoritması ile ilişkili değişkenler içeren veri setinin boyutlarını daha az boyuta indirgenmesini sağlayan bir dönüşüm tekniği olan TBA (PCA) birleştirilerek bir hibrid öznitelik seçme yöntemi oluşturulmuştur. Bu yöntemin ismine Apriori+TBA (AP) adı verilmiştir. Metot iki aşamadan oluşmaktadır. Birinci aşamada, kullanılacak niteliklerin seçimi için Apriori algoritmasından ilişkili birliktelikler tespit edilmiştir. Sınıflandırıcı için en az etkisi olan nitelik elenmiştir. Kalan nitelikler TBA algoritmasına verilerek farklı bir uzaya taşınmıştır. Bu uzayda da iki nitelik elenmiştir. Böylece önemli olan ilgili özniteliklerin seçilmesi ve öznitelik vektörü boyutunun azaltılması sağlanmıştır. İkinci aşamada, kalan nitelikler sınıflandırıcılara verilmiştir. Sınıflandırıcı olarak yapay sinir ağıları ve Destek vektör makinaları kullanılmıştır. Sınıflandırıcının güvenirliliğini artırmak için çapraz–doğrulama ile test edilmiştir.

Bu çalışmayla oluşturulan hibrid sistemin geliştirilme sebebi; Veri setlerinde var olan gereksiz girişler atılarak önemli öznitelikler seçilmiş, boyut olarak daha az verinin oluşması sağlanıp, sistemin sınıflandırma başarısını artırıp, az boyutlu uzaylarda daha etkin çalışması sayesinde bellek ve kapasite ihtiyaçlarının azaltılması sağlamıştır. İleride bu tür işlemlerde verinin işlenebilmesi için, kullanılmayan niteliklerin elde edilmeyeceğinden harcanan süre ve işlemlerden de tasarruf edilmiş olunacaktır.

Önerilen sistemin test edebilmek için Wisconsin göğüs kanseri verisi kullanılmıştır. Test aşamasında 10-katlı çapraz–doğrulama uygulanmıştır. Geliştirilen sistem ile AP ön işleminden geçtikten sonra YSA (NN) ve DVM sınıflandırıcılarına

giriş olarak verilerin başarıları ölçülmüştür. AP+YSA sisteminde ortalama sınıflandırma doğruluğu en yüksek değer olan %98.29 olarak bulunmuştur. Göğüs kanseri verisi için çapraz–doğrulama yöntemiyle yapılan çalışmalar içerisinde çalışmamız en iyi sınıflama başarılarından birini sağlamıştır. Çıkan sonuçlar gösteriyor ki boyut azaltımı yapılarak uygulanan bu hibrid sistemin, hastalıkların daha hızlı ve başarılı teşhis edilmesi için kullanılabilir bir sistem olduğu görülmüştür.

Bu çalışmada, AP+YSA metodu, göğüs kanseri probleminde kullanılmak üzere önerilmiştir. Önerilen sistemin blok şeması Şekil 7.1'de gösterilmiştir.



Şekil 7.1 AP+YSA metodu blok şeması

Çizelge 7.1 Güvenlik seviyesine göre uygun seçilen kurallar

Giriş	Değer	Güven değeri
1,2,3,7=>9	1,1,1,3=>1	100%
1,3,8,9=>2	1,1,1,1=>1	100%

Şekil 7.2'de uygulamada kullandığımız Apriori algoritması ve bu algoritmaya ait alt programların basitleştirilmiş kodu görülmektedir. Basitleştirilmiş kodlar incelediğinde sık geçen öğe kümelerinin tespit edebilmek için birçok kez göğüs kanseri veri kümesinin tarandığı görünmektedir. Göğüs kanseri veri tabanı taranarak nesnelerin kaç adet hareket kaydı içinde bulunduğu tespit edilmiştir. Her nesne için tespit edilen bu değere destek sayıcı denir. Destek sayıcı %15 olarak belirlenmiştir. Bu değerin altındaki minimum destek değerine eşit veya büyük olan nesnelerin sık geçen 1-nesne kümesi olarak belirlendiği varsayılarak işleme başlanmıştır. Bu işlem algoritmanın 1. satırında 1-elemanlı aday nesneler yani *L1* alınmıştır. Basitleştirilmiş koddaki döngü ile sık geçen öğe kümesinin nesneleri ikili kombinasyonuna benzer bir biçimde ($L_1 \infty L_1$)

birleştirilerek yeni bir küme oluşturulmuştur. Bu oluşan aday öğe kümeler C harfi ile simgelenmiştir.

Apriori-gen fonksiyonu ile aday kümeler budanmıştır. Ele alınan nesnelerin tüm alt kümeleri subset fonksiyonu ile bulunarak, öğe kümesinin içinde yer almayan küme elemanları aday kümesinden silinmiştir. Sık geçen öğe kümelerinin test edilmesi işlemi has_infrequent_subset alt programı ile gerçekleştirilmiştir. Süreç yeni bir sık geçen öğe kümesi bulunamayana kadar döngü her dönüşünde öğe sayısı artırılarak devam etmiştir.

Veriler üzerinde sık geçen öğe kümeleri tespit edildikten sonra bu öğe kümeleri ile çok güçlü kurallar oluşturulmuştur. Minimum destek değeri olan %15 ve güven değeri olan %100 sağlayan kurallar Çizelge 7.1'de verilmiştir.

AP tekniğinin ilk aşamasında, giriş parametreleri arasındaki birliktelik ilişkileri tespit edilmiştir. Bu ilişkiler, tespit edilirken yeterli destek değerine ve yüksek güvenlik değerine sahip kurallar çıkartılmıştır. Güvenlik seviyesi en yüksek çıkan ve öznitelik seçiminde kullanılabilecek nitelikteki iki kural Çizelge 7.1'de verilmiştir.

Birinci kurala göre; 1,2,3,7 giriş parametreleri aynı ifadedeyken, 9. giriş parametresi de sürekli aynı ifadededir. Bu durumda şunu söyleyebiliriz ki 9. giriş diğer girişlere bağlıdır. Bu yüzden yapay sinir ağı girişinde 9. giriş parametresini kullanılmamıştır. Geriye kalan girişler TBA uzayına dönüştürülmüş ve bu uzaydan iki giriş parametresi daha atılmıştır. Boyut olarak azaltılmış veriler sınıflandırıcıya giriş parametresi olarak verilmiştir.

İkinci kurala göre; 1,3,8,9 giriş parametreleri aynı değerdeyse, 2. giriş parametresi de aynı değerdedir. Bu durumda şunu söyleyebiliriz ki 2. Giriş, diğer girişlere bağlıdır. Bu yüzden yapay sinir ağı girişinde 2. giriş parametresi kullanılmamıştır. Geriye kalan girişler TBA uzayına dönüştürülmüş ve bu uzaydan iki giriş parametresi daha atılmıştır. Boyut olarak azaltılmış veriler sınıflandırıcıya giriş parametresi olarak verilmiştir.

Bu iki kural, iki ayrı sınıflandırıcıdaki başarıları test edilmiştir. Giriş değerleri sınıflandırıcıya girmeden önce normalizasyon işleminden geçirilmiştir. Sınıflandırıcı da test işlemleri sırasında güvenirliliği artırmak için k-kat çapraz doğrulama yöntemi kullanılmıştır. Deneysel çalışmalar göstermiştir ki, k-kat çapraz doğrulama yönteminde k için en iyi değer 10'dur (Breiman ve ark 1984, Kohavi 1995) (Kahramanli 2008). Bu yöntemde veriler 10 bölünüp 9'u eğitime verilir. Ayrılan kısım eğitimin testi için

kullanılır. Bu işlem her bir birim test için kalacak şekilde devam eder. Ortaya çıkan 10 tane doğruluk değerinin ortalaması alınarak başarı ortalaması bulunmuştur. YSA sonuçları sırasıyla %98.14% ve %98.29 bulunmuştur. DVM sınıflandırıcı sonuçları ise %97.85 ve %97.99 dir. Sınıflandırıcılar açısından en iyi sonuçlar YSA'da bulunmuştur. YSA sonuçlarına göre ikinci kuralın sonucu daha iyi olduğu için diğer yayınlarla karşılaştırma açısından ikinci kural seçilmiştir. Çizelge 7.4'te bu değerler görülmektedir. En iyi sonucun her bir kat için bulunmuş YSA performans sonuçları (duyarlılık, Özgünlük, pozitif öngörü değeri, negatif öngörü değeri, doğruluk oranı) Çizelge 7.7'de gösterilmiştir. Yöntemimizle elde edilen ağ sınıflandırma sonuçları karışıklık matrisle elde edilen sonuçlar Çizelge 7.8'de verilmiştir. Önerdiğimiz sistemle diğer sistemlerin karşılaştırılması Çizelge 7.6'de verilmiştir. Elde edilen sonuç, şimdiye kadar kanser verisi üzerine 10-kat çapraz doğrulama yöntemi kullanılarak yapılan çalışmalar arasında ortalama doğruluk sınıflandırma oranı umut vericidir.

Sınıflama için kullanılan parametreler DVM için Çizelge 7.2'de ve YSA için kullanılan MLP (ÇKP) mimarisi ve eğitim parametreleri Çizelge 7.3'te belirtilmiştir. Tüm verilerin uzunluğu 699 satırdan meydana gelir. Giriş nitelikleri; Kütle kalınlığı, Hücre Büyüklüğünün Benzerliği, Hücre Şeklinin Benzerliği, Marjinal Yapışma, Tek Epitel Hücresi Büyüklüğü, Çıplak Çekirdek Kütlesi, Hafif Çekirdek Kütlesi, Normal Çekirdek Kütlesi ve Mitoz'dur. Çıkış değerlerinin iyi huylu ve kötü huylu değer olmak üzere iki çıkış değeri vardır. Öncelikle, bu girişler özellik seçimi için birliktelik algoritmalarından apriori algoritması ile analiz edilip hücre büyüklüğünün benzerliği ve mitoz özelliklerinden biri elenir. Bu nedenle, giriş nitelik sayısı sekize düşürülmüştür. Düşürülen bu girişler TBA uzayına dönüştürülür. Bu uzaydan da iki özellik atılarak sınıflandırıcıların girişlerine 6 nitelik bilgisi verilmiştir. YSA için en iyi sonuç ikinci giriş ile TBA uzayından atılan 2 giriş ile birlikte 18. Gizli düğüm katmanında elde edilmiştir. Aktivasyon fonksiyonları olarak gizli katmanda Tangent-sigmoid kullanılmıştır. Çıkış katmanında ise Log-sigmoid kullanılmıştır. 10-kat çapraz doğrulama yapılarak ortalama en iyi başarı oranı %98.29 bulunmuştur. Çıkan sonuçlar çizelge 7.4' te verilmiştir.

DVM için SMO optimizasyon algoritması kullanılarak lineer kernel fonksiyonunda aynı giriş değerleri için %97.99 değeri elde edilmiştir. Çıkan sonuçlar

çizelge 7.5'te verilmiştir. Her iki sınıflandırıcıda da başlangıç değerleri rastgele alınmıştır.

```
Giriş: Database (D), minimum destek min_sup.
Çıkış: L, D de sık olan nesne kümeleri
(1) L1=find_frequent_1-itemsets(D) //large 1-nesnekümeler
(2) for (k=2;L(k-1)≠0;k++) {
(3) Ck=apriori_gen(L(k-1), min_sup) // yeni adaylar
(4) for each transaction t∈D{ //sayım için D yi tara
(5) Ct=subset(Ck,t); // Aday olan t' lerin alt kümeleri elde edilir
(6) for each candidate c∈Ct
(7) c.count++;
(8) }
(9) Lk={ c∈Ck | c.count ≥min_sup }
(10) }
(11) return L=Uk Lk;

procedure apriori_gen (L(k-1);min_sup) // (L k–1: sık geçen (k–1)-nesnekümeler)
(1) for each itemset l1∈L(k-1)
(2) for each itemset l2∈L(k-1)
(3) if (l1[1]= l2[1] )∧(l1[2]=l2[2]) ∧…∧ (l1[k-1]=l2[k-1]) then {
(4) c=l1x l2; // aday nesnelerin üretilmesi
(5) if has_infrequent_subset(c,L(k-1)) then
(6) delete c; // gereksiz adayların silinmesi
(7) else add c to Ck;
(8) }
(9) return Ck;

procedure has_infrequent_subset(c,L(k-1));
// (c: aday k-nesneküme;L k–1: sık geçen (k-1)-nesnekümeler); // prior bilgisini kullan
(1) for each (k-1)-subset s of c
(2) if s∉L(k-1) then
(3) return TRUE
(4) return FALSE
```

Şekil 7.2 Basitleştirilmiş Apriori algoritma kodu

7.1.1 Sınıflandırma Parametreleri

Çizelge 7.2 DVM için sınıflandırma parametreleri listesi

Parametreler	Değer
Metot	DVM
Optimizasyon algoritması	SMO
Doğrulama metot	k kat çapraz doğrulama (10-kat ÇD)
Çekirdek Fonksiyonu	Lineer
TolKKT	1.0000e-003
Maksimum Tekrarlama	15000
Kernel Cache Limit	5000
Başlangıç değeri	Rastgele

Çizelge 7.3 MLP mimarisi ve eğitim parametreleri

Mimari	
Katman sayısı	3
Katmanlar üzerinde nöron sayısı	Girdi: 6, Gizli: 18, Çıkış: 1
İlk ağırlıkları ve önyargıları	Rastgele
Aktivasyon fonksiyonları	Tangent-sigmoid (Gizli katman için) Log-sigmoid (Çıktı katmanı için)
Eğitim parametreleri	
Öğrenme kuralı	A gradient descent with momentum and adaptive learning rate backpropagation
Amaç	0.000001

7.1.2 Deneysel Sonuçlar ve Tartışmalar

Çizelge 7.4 AP+YSA kullanan göğüs kanseri tespiti için doğru sınıflandırma oranı.

Sınıflandırıcı	Devirler	Gizli katman nöron sayısı	Ortalama Doğru sınıflandırma oranı (%)
AP+NN (2 Hariç) (6, 18, 1)	325	18	98.29

AP+NN (9 Hariç) (6, 3, 1)	350	3	98.14

Çizelge 7.5 AP+DVM kullanan göğüs kanseri tespiti için doğru sınıflandırma oranı.

Performans Kriterleri	AP+DVM (9 Hariç)	AP+DVM (2 Hariç)
Doğruluk (%)	97.85	97.99
Hassasiyet (%)	97.29	97.11
Özgünlük (%)	98.57	99.63
Pozitif öngörü değeri (%)	95.40	94.91
Negatif öngörü değeri (%)	99.38	99.76

Çizelge 7.6 Önerilen sistemin elde edilen sınıflama doğruluk oranı ve literatürdeki diğer çalışmalarla karşılaştırılması

Yazar (Yıl)	Metot	Sınıflandırma Doğruluğu (%)
Goodman ve ark. (2002) (Goodman 2002)	Optimize- LVQ (10 × ÇD) Büyük- LVQ (10 × ÇD) AIRS (10 × ÇD)	96.70 96.80 97.20
Abonyi ve Szeifert (2003) (Abonyi ve Szeifert 2003)	Denetimli bulanık kümeleme (10 × ÇD)	95.57
Polat ve Güneş (2007)(Polat ve Gunes 2007)	LS-SVM (10 × ÇD)	98.53
Karabatak ve Ince (2009)(Karabatak ve Ince 2009)	AR1+NN (3 × ÇD) AR2+NN(3 × ÇD)	97.4 95.6
Yeh ve ark. (2009)(Yeh ve ark 2009)	DPSO(ÇD olamadan)	98.71
Marcano-Cedeño ve ark.(2011)(Marcano-Cedeno ve ark 2011)	AMMLP(ÇD olamadan)	99.26
Chen ve ark. (2011) (Chen ve ark 2011)	RS_SVM (5 × ÇD) (Eğitim: 50%-test: 50%) RS_SVM (5 × ÇD) (Eğitim: 70%-test: 30%) RS_SVM (5 × ÇD) (Eğitim: 80%-test: 20%)	96.55 96.72 96.87
Çalışma sonucu	AP+YSA (10 × ÇD)	98.29

Çizelge 7.7 AP + YSA kullanarak göğüs kanseri tespiti için performans sonuçları (2 hariç)

Sınıflandırıcı	ÇD parçaları	Duyarlılık (%)	Seçicilik (%)	Pozitif öngörü değeri (%)	Negatif öngörü değeri (%)	Sınıflama Doğruluğu(%)
AP+YSA (2 hariç) (6, 18, 1)	CV-1	100	100	100	100	100
	CV-2	100	100	100	100	100
	CV-3	100	92.59	95.56	100	97.14
	CV-4	95.45	96.15	97.67	92.59	95.71
	CV-5	97.78	100	100	96.15	98.57
	CV-6	97.92	100	100	95.65	98.57
	CV-7	97.50	100	100	96.77	98.57
	CV-8	97.73	100	100	96.30	98.57
	CV-9	97. 73	100	100	96.30	98.57
	CV-10	95.92	100	100	91.30	97.14
	Ortalama	98.03	98.87	99.32	96.51	98.29

Çizelge 7.8 Önerilen Yöntemin Karışıklık matrisler çapraz doğrulama sonuçları

ÇD parçası	Gerçek sınıf	"iyi huylu" olarak öngörülenlerin sayısı	"kötü huylu olarak öngörülenlerin sayısı"	ÇD parçası	Gerçek sınıf	"iyi huylu" olarak öngörülenlerin sayısı	"kötü huylu olarak öngörülenlerin sayısı"
ÇD-1	iyi huylu	49	0	ÇD-6	iyi huylu	47	1
	Kötü huylu	0	20		Kötü huylu	0	22
ÇD-2	iyi huylu	52	0	ÇD-7	iyi huylu	39	1
	Kötü huylu	0	18		Kötü huylu	0	30
ÇD-3	iyi huylu	43	0	ÇD-8	iyi huylu	43	1
	Kötü huylu	2	25		Kötü huylu	0	26
ÇD-4	iyi huylu	42	2	ÇD-9	iyi huylu	43	1
	Kötü huylu	1	25		Kötü huylu	0	26
ÇD-5	iyi huylu	44	1	ÇD-10	iyi huylu	47	2
	Kötü huylu	0	25		Kötü huylu	0	21

Sınıflandırma problemlerinde ve birçok öğrenme algoritmalarında konu ile ilgili olan, en faydalı ve en önemli nitelikler seçilerek veri tabanına ait nitelik sayısının azaltılması önemlidir. Bunun amacı daha az hesaplama yükü ile daha yüksek başarı elde edilebilmesidir.

Büyük veri tabanlarında nitelikler arasındaki ilişkileri tespit etmek için birçok yöntem kullanılmaktadır. Bu yöntemlerden biriside birliktelik kurallarıdır. Birliktelik kuralının bu özelliği, özellik seçimi yöntemi olarak kullanılabileceğini göstermektedir. Birliktelik kuralları arasındaki en yaygın kullanılan algoritmalardan biriside Apriori algoritmasıdır. Eğer bir veya birkaç nitelik kümesi belli bir destek ve güven değerinde birbiri ile ilişkili ise bu niteliklerin bir altkümesi bir başka nitelik altkümesini belirtir.

TBA ise, veri tabanındaki verinin en önemli karakteristik niteliklerini belirlerken bir taraftan da verinin boyutunu azaltarak ayırt edici nitelikleri içeren bilginin kullanılmasını sağlamaktadır. Bu yöntem var olan özelliklere bir dönüşüm uygulayarak yeni bir uzaya çevirmekte ve istenirse özellik sayısı da azaltılmaktadır. Çalışmamızda birliktelik kuralları ve TBA birleştirilerek bir hibrid sistem geliştirilmiştir. Bu hibrid sistem sınıflandırıcının önünde bir nitelik seçme yöntemi olarak kullanılmıştır. Hibrid sistem sayesinde az nitelikle, hızlı bir şekilde daha başarılı sonuçlar elde edilmesi amaçlanmıştır. Kurulan sistem literatürde sıkça kullanılan UCI veri tabanından alınan Wisconsin göğüs kanseri veri kümesinde test edilmiştir.

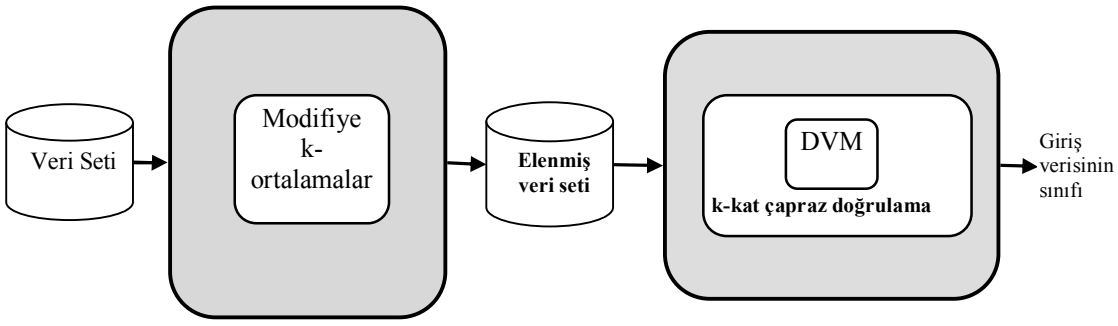
Bu çalışmada, tasarlanan hibrid sistem ile sınıflama başarısını daha az parametrelerle daha iyi doğruluk oranıyla tespit eden bir yapı ortaya konmuştur. Çalışmada göğüs kanseri gibi kadınlarda sık rastlanan literatürde sık kullanılan bir veri tabanında denenmiş ve Çizelge 7.6'da görüleceği gibi literatürdeki kat doğrulama kullanan çalışmalar arasında yüksek başarı oranı elde edilmiştir. Burada görünüyor ki sistemin belirlediği özellikler kullanılmadan daha hızlı bir tanı sistemi ortaya konmuştur. Ayrıca bu özelliklerin bulunması için gereken çaba ve maliyetten de tasarruf edilmiş olunacaktır.

Sistem ayrıca bir uzman gerektirmeden hastanın elindeki belli tahlil sonuçlarından hastalık teşhisi de yapabilmektedir. Erken tanının hayat kurtardığı kanser gibi hastalıklarla mücadele alanında sistemimizin daha hızlı ve daha güvenilir otomatik tanı sistemlerinin gelişmesine katkıda bulunacağına inanıyoruz.

7.2 Veri setleri için modifiye edilmiş kümeleme algoritmalarına dayalı veri hazırlama yöntemi

Örüntü tanımada, sistemin hızlı ve etkili çalışması engelleyen en önemli unsur, veri tabanında bulunan gürültülü ve tutarsız verilerdir. Bu çalışma gürültülü ve tutarsız verilerin eliminasyonu için ön işlemci olarak modifiye edilmiş k-ortalamalar Algoritması ve sınıflandırıcı olarak Destek Vektör Makinalarını (Support Vector Machines) kullanan hibrid bir yaklaşım sunmaktadır. Geliştirilen yaklaşım insan yaşam kalitesini düşüren, toplumda oldukça sık görülen ve önemli ölüm sebepleri arasında yer alan kalp ve diyabet hastalıklarının teşhisinde test edilmiştir. Bu hastalıkların teşhisinde kullanılan veri kümeleri, UCI veri tabanından alınan Statlog (Heart), SPECT görüntüleri ve Pima Indians Diabetes veri kümeleridir. Önerilen sistem, bu veri kümelerinde sırasıyla %97.9, %98.2 ve %96.7 sınıflama başarısına ulaşmıştır.

Veri tabanlarında, gürültüler nedeniyle benzer verilere göre tutarsız, birbirinin tekrarı ve bazı öznitelikleri belirsiz veriler bulunmaktadır. Bu verilerin tanıma sistemine girilmesi sistemin performansını önemli ölçüde düşürmektedir. Önerilen sistem ile modifiye edilmiş k-ortalamalar algoritması kullanılarak verilerdeki bu kusurlardan bazıları giderilmiştir. Böylece elde edilen yeni veri kümeleri sınıflandırıcının test başarısını ve sistemin kararlılığını artırmaktadır. Önerilen sistem, Şekil 7.3'te görüldüğü gibi iki aşamadan oluşmaktadır. İlk aşamada modifiye edilmiş k-ortalamalar algoritmasıyla gürültülü ve tutarsız veriler atılır. İkinci aşamada elde edilen arındırılmış veriler DVM sınıflandırıcısına verilerek doğruluk oranları tespit edilir. Sınıflandırıcının güvenirliliğini artırmak için k-kat Çapraz Doğrulama (ÇD) kullanılmıştır. Bu işlem 500 defa tekrar tekrar yapılarak performans değerlerinin ortalaması sınıflandırıcı başarısı etkisi test edilmiştir.



Şekil 7.3 Önerilen sistemin blok diyagramı

Modifiye k-ortalamalar algoritmasıyla oluşturulan kümeleme işleminde, algoritmaya eklenen bir ağırlık çarpanıyla daha kararlı kümeler oluşturulması hedeflenmiştir. Bu ağırlık çarpanı küme merkezleriyle örnekler arasındaki uzaklığın bulunduğu Öklid uzaklık ölçütüne çarpan olarak eklenmektedir.

$$G = 1 - \frac{std(x_j)}{\sum_{j=1}^{boyutu(giriş)} std(x_j)} \tag{7.1}$$

G ağırlık değeri, verilerden her bir sütunu temsil eden verilerin standart sapmalar tüm sütunların standart sapmaların toplamına bölünüp bir sayısından çıkarılarak elde edilmiştir. Böylece az değişen verilere yüksek ağırlık değeri, çok değişenlere az ağırlık değeri verilerek kümelerin kararlılıkları artırılmıştır.

Sonuç olarak, Modifiye k-ortalamalar algoritmasında her bir nesnenin merkez noktalara uzaklığını hesaplamak için kullanılan uzaklık ölçütü aşağıdaki gibidir.

$$uzaklık(x, y) = \left\{ \sum ((x_i - y_i) G_i)^2 \right\}^{1/2} \tag{7.2}$$

Burada, $uzaklık(x, y)$, x ve y noktaları arasındaki uzaklık, hata parametresidir. x ve y aralarındaki uzaklığı hesaplanan nesnelerin uzayda temsil edildiği noktalardır. G_i, ağırlık değeridir.

Başlangıç Fazı:

Adım 1. $\left(C_1, C_2, ..., C_k\right) = D$'

(Küme merkezleri rasgele olarak birbirleriyle örtüşmeyecek şekilde ayarlanır.)

Adım 2. G = Belirlenmiş ağırlık merkezi

İterasyon Fazı:

Adım 3.

1. **repeat**

3.1. $D_{ij} = G * \left(j \text{ kümesi ve } i \text{ örneği arasındaki uzaklık}\right)$;

3.2. $n_i = argmin_{1 \le j \le k}(D_{ij})$;

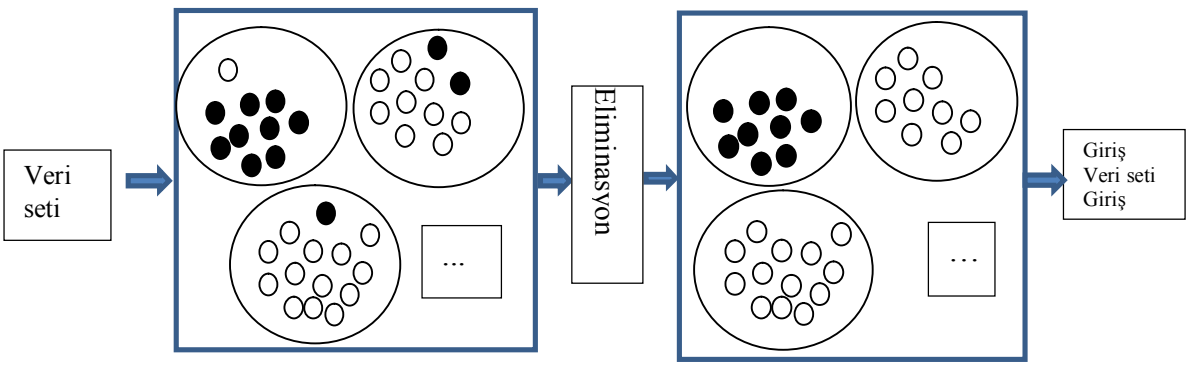
3.3. i örneğini n_i'nin kümesine ata.

3.4. Yukarıda yapılan değişikliklere göre küme merkezlerini (ortalamasını) yeniden hesapla.

until Kümelerin üyeliklerinde bir değişiklik yoksa iterasyonu tamamla aksi halde işlemleri tekrar et.

Şekil 7.4 Modifiye edilmiş k-ortalamalar algoritmasının basitleştirilmiş kodu

Geliştirilen Modifiye edilmiş k-ortalamalar algoritmasının basitleştirilmiş kodu Şekil 7.4'te verilmiştir. Bu kodda, ilk olarak birbirleriyle aynı olmayacak şekilde küme merkezleri atanır. Kullanılan veri tabanına göre her bir sütunun G ağırlık değeri hesaplanır. Bu ağırlık değeri küme merkezleri ve her bir örnek arasında hesaplanan Öklid uzaklık ölçütü ile çarpılarak D_{ij} uzaklık matrisi elde edilir. Bu elde edilen uzaklık matrisine göre her bir örnek en yakın olduğu küme merkezinin (C_k) bağlı bulunduğu kümeye atanır. Bu kümelere dahil olan tüm örneklerin ortalamaları alınarak yeni küme merkezleri hesaplanır. Bu işlem küme merkezlerinde veya küme üyeliklerinde bir değişiklik olmayana kadar devam eder.



Şekil 7.5 Önerilen veri hazırlama sistemi

```
Adım 1. Veriler organize edilir.( kolonlarda özellikler, satırlarda örnekler olmak üzere)
Adım 2. (kümelenmiş veri seti, dataindex)= Düzenlenmiş_k_means_algoritması(verisetinin girişleri)
Adım 3.
for i=1 to küme sayısı
  Adım 3.1. dağılım_değeri(i)=Frekans_dağılımı(kümelenmiş_veriseti(i)'deki veriseti(dataindex)'in çıkışı)
  Adım 3.2.
        if dağılım değeri(sınıf-1) > dağılım değeri(sınıf-2)
        then kümelenmiş_veriseti(i) deki sınıf-2 veriyi ele
        else kümelenmiş_veriseti(i) deki sınıf-1 veriyi ele)
end for
```

Şekil 7.6 Önerilen veri hazırlama yönteminin basitleştirilmiş kod sunumu

Kullanılan veri tabanları, sütunlarda özellikler, satırda ise örnekler olacak şekilde düzenlenir. Daha sonra her bir örnek modifiye k-ortalamalar algoritmasına uygulanır. N adet kümeye ayrılan veri tabanı, aynı kümede yer alan örneklerin aynı sınıfa ait olması gerektiği prensibi ile elenir. Buna göre hangi sınıf bir küme içerisinde baskın ise o veriler kalır. Diğerleri elenir. Eleme işlemi sırasında istenirse bir oran tespit edilip bu oranın altında olan verilerin atılması sağlanabilir. Bu sayede tutarsızlığın yüksek olduğu kümelerde veri eliminasyonu yapılmazken tutarsızlığı az olan kümelerde bozucu veriler elemine edilir. Bu sayede veri tabanında elimine edilecek verilerin sayısı da ayarlanabilmektedir. Bu işlem tüm kümeler için tekrarlanır. Elde edilen veri seti k-kat çapraz doğrulama kullanılan DVM sınıflandırıcısının girişine verilir. Aynı yöntem klasik k-ortalamalar algoritması ile de denenmiş tasarlanan sistemin başarısı karşılaştırılmıştır. Sistemin simüle edilmiş gösterimi şekil 7.5'te gösterilmiştir. Önerilen veri hazırlama yönteminin basitleştirilmiş kodu da şekil 7.6'da verilmiştir.

7.2.1 Sınıflandırma Parametreleri

Bu çalışmada, gürültülü ve hatalı ölçülmüş verilerin tespit edilip, sınıflandırma başarısını düşürmesindeki etkilerini azaltmak için tasarlanan sistem, kalp ve diyabet hastalıkları için hazırlanan veri setleri üzerinde DVM sınıflandırıcıyla test edilmiştir.

Sınıflandırıcının performans değerinin güvenirliliği 10-kat çapraz doğrulama yöntemi ile sağlanmıştır. Ayrıca sınıflandırma algoritmasını, 10-kat çapraz doğrulama yöntemi ile 500 kez çalıştırılarak, çıkan değerlerin ortalaması ve standart sapması da bulunmuştur. Bu sınıflandırıcı kullanırken, eğitim Çizelge 7.9'da verilen parametrelere göre yapılmıştır.

Çizelge 7.9 Sınıflandırma parametreleri listesi

Parametreler	Değer
Metot	DVM
Optimizasyon algoritması	SMO
Doğrulama metot	k- kat çapraz doğrulama (10-kat ÇD)
kN	500
Çekirdek Fonksiyonu	Lineer
TolKKT	1.0000e-003
Maksimum tekrarlama	15000
Kernel Cache Limit	5000
Başlangıç değeri	Rastgele

7.2.2 Deneysel Sonuçlar ve Tartışmalar

Çizelge 7.10 k-ortalamalar algoritması ile veri setleri için ön işleme sonuçları

Veri Seti	Sınıf Sayısı	Özellik Sayısı	Küme Sayısı	Örnekler	Temizlenmiş Örnek sayısı	10-Kat ÇD ile ortalama doğruluk	KN=500 ve 10-Kat ÇD için Ortalama Doğruluk	KN=500 ve 10-Kat ÇD için Standart Sapma
Statlog (Heart)	2	13	13	270	190	92.11	88.62	0.011
SPECT Görüntüleri	2	22	22	267	216	96.73	94.33	0.014
Diabetes	2	8	8	768	552	93.65	92.80	0.003

Çizelge 7.11 Modifiye k-ortalamalar algoritması ile veri setleri için ön işleme sonuçları.

Veri Seti	Sınıf Sayısı	Özellik Sayısı	Küme Sayısı	Örnekler	Temizlenmiş Örnek sayısı	10-Kat ÇD ile ortalama doğruluk	KN=500 ve 10-Kat ÇD için Ortalama Doğruluk	KN=500 ve 10-Kat ÇD için Standart Sapma
Statlog (Heart)	2	13	13	270	184	97.87	96.52	0.0074
SPECT Görüntüleri	2	22	22	267	221	98.18	96.92	0.0095
Diabetes	2	8	8	768	546	96.71	96.22	0.0027

Bozuk veriler sınıflandırıcı başarısını olumsuz etkileyen en önemli etmenlerden biridir. Geliştirilen sistem bu tip kötü verilerin eliminasyonunu sağlamaktadır. Veri eliminasyonu işlemi modifiye edilmiş k-ortalamalar algoritması ile gerçekleştirilmektedir. Veri eliminasyonuna tabii tutulmuş veri tabanları DVM kullanılarak sınıflandırılmaktadır. Kullanılan sınıflandırıcının performansını güvenilir kılmak için 10- kat çapraz doğrulama metodu kullanılmıştır. Çizelge 7.10'da klasik k-ortalamalar algoritması ile veri setlerine uygulanan ön işleme sonucunda çıkan sonuçlar verilmiştir. Çizelge 7.11'da ön işleme için modifiye ettiğimiz k-ortalamalar algoritmasının veri setlerine uygulanan sonuçlar verilmiştir. Sistem sonuçları klasik K-ortalamalar algoritması ve literatürde bu veri tabanlarına uygulanan diğer makaleler ile karşılaştırılmıştır. Önerilen sistemin sınıflandırma doğruluğu Statlog (Heart) veri seti, SPECT Heart veri seti ve Pima Indians Diabetes veri setleri için sırasıyla %97.9, %98.2, %96.7 olarak elde edilmiştir. Aynı datalar, sınıflandırma algoritmasının 500 defa çalıştırılıp performans değerlerinin ortalaması alındığında Statlog (Heart) veri seti, SPECT Heart veri seti ve Pima Indians Diabetes veri setleri için sırasıyla %96.5, %96.9, %96.2 olarak elde edilmiştir. Geliştirilen bu sistem diğer tüm desen tanıma uygulamalarında kullanılabilecek şekilde tasarlanmıştır.

Statlog Kalp Hastalığı veri kümesi için, sınıflandırma performans değerleri ve diğer sistemlerle karşılaştırmaları sırasıyla Çizelge 7.12 ve Çizelge 7.13'te verilmiştir.

Çizelge 7.12 Statlog Kalp Hastalığı veri seti için sınıflandırma performansı

Performans Kriterleri	DVM	k-ortalamalar +DVM	M. k-ortalamalar +DVM
Doğruluk (%)	83.33	92.11	97.87
Hassasiyet (%)	78.46	86.39	93.5
Özgünlük (%)	87.52	96.59	99.23
Pozitif öngörü değeri (%)	83.67	92.76	98.57
Negatif öngörü değeri (%)	83.51	91.36	98.08
Zaman tüketimi (saniye)	4.151391	1.620226	1.373317

Çizelge 7.13 Bizim metodumuzun ve literatürdeki diğer metotların Statlog kalp hastalığı veri seti üzerinde elde edilmiş sınıflandırma doğrulukları

Yazarlar (Yıl)	Metotlar	Sınıflama Doğruluğu (%)
Duch ve ark. (2001) (Duch ve ark 2001)	k-NN, k=28, 7 özellikler (10-kat ÇD)	84.6-85.6
	k-NN, k=28, Manhattan (10-kat ÇD)	82.2-83.4
	FSM, 27 bulanık kurallar	82
	SSV, 3 Kurallar	80.2-83.4
Sahan ve ark. (2005) (Sahan ve ark 2005)	AWAIS (10-kat ÇD)	82.59
Ozsen ve Gunes (2008) (Ozsen ve Gunes 2008)	Hibrid benzerlik ölçüsü ile AIS algoritması (10-kat ÇD)	83.95
Kahramanli ve Allahverdi (2008) (Kahramanli ve Allahverdi 2008)	YSA ve FNN kullanarak Hibrid sistem (10-kat ÇD)	86.8
Ozsen ve ark.(Ozsen ve ark 2009)	Çekirdek Tabanlı AIS (5-kat ÇD)	85.93
Tian J. et al. (2009) (Tian ve ark 2009)	CC-EBFNN	82.45
Polat ve Gunes (2009)(Polat ve Guenes 2009)	LS-SVM öznitelik seçimi olmadan sınıflandırıcı (Eğitim: 50%- test: 50%)	80.00
	Özellik seçimi olmadan LM_YSA (Eğitim: 50%- test: 50%)	71.11
	Özellik seçimi F-skoru ile LS-SVM birleşimi (Eğitim: 50%- test: 50%)	77.78
	Özellik seçimi F-skoru ile YSA birleşimi (Eğitim: %50- test: %50) 6 özellik azaltılmış.	77.61
	Özellik seçiminde F-skoru doğrusal çekirdek ile LS-SVM sınıflandırıcının birleşimi (Eğitim: %50- test: %50)	80.00
	Özellik seçimi F-skoru doğrusal çekirdek ile YSA birleşimi (Eğitim: 50%- test: 50%) 128 özellik azaltılmış	80.74
	Özellik seçiminde F-skoru RBF çekirdek ile LS-SVM sınıflandırıcının birleşimi (Eğitim: %50- test: %50) 117 özellik azalmış.	83.70
	Özellik seçimi F-skoru doğrusal çekirdek ile YSA birleşimi (Eğitim: 50%- test: 50%) 117 özellik azaltılmış	76.30
Subbulakshmi ve ark. (2012)(Sub bulakshmi ve ark 2012)	ELM	87.50
Karabulut E.M ve ark. (Karabulut ve Ibrikci 2012)(2012)	Rotation Forest, Levenberg-Marquardt (10-kat ÇD)	91.20
Ahmad F.ve ark.(2013) (Ahmad ve ark 2013)	Geliştirilmiş Genetik Al(Eğitim: 75%- test: 25%) 10 Kez çalışmanın ortalaması	86.30
Geliştirdiğimiz Metot	k-ortalamalar(10-kat ÇD)	92.11
Geliştirdiğimiz Metot	m. k-ortalamalar (10-kat ÇD)	97.87

SPECT görüntüleri veri kümesi için, sınıflandırma performans değerleri ve diğer sistemlerle karşılaştırmaları sırasıyla Çizelge 7.14 ve Çizelge 7.15'te verilmiştir.

Çizelge 7.14 SPECT görüntü veri seti için sınıflandırma performansı

Performans Kriteri	DVM	k-ortalamalar +DVM	M. k-ortalamalar +DVM
Doğruluk (%)	75.91	96.73	98.18
Hassasiyet (%)	56.08	89.17	90.00
Özgünlük (%)	80.45	98.36	98.97
Pozitif öngörü değeri (%)	51.75	91.67	93.00
Negatif öngörü değeri (%)	86.32	97.95	99.05
Zaman tüketimi (Saniye)	2.342839	0.652620	0.636283

Çizelge 7.15 Bizim metodumuzun ve literatürdeki diğer metotların SPECT görüntü veri seti üzerinde elde edilmiş sınıflandırma doğrulukları

Yazarlar (Yıl)	Metotlar	Sınıflama Doğruluğu (%)
Polat ve Güneş (2009) (Polat ve Guenes 2009)	Özellik seçimi olmadan LS-SVM sınıflama(Eğitim: 50%- test: 50%)	82.71
	Özellik seçimi olmadan LM_YSA (Eğitim: 50%- test: 50%)	77.44
	Özellik seçimi F-skoru ile LS-SVM birleşimi (Eğitim: 50%- test: 50%)	79.70
	Özellik seçimi F-skoru doğrusal çekirdek ile YSA birleşimi (Eğitim: 50%- test: 50%) 7 özellik azaltılmış.	75.70
	Özellik seçiminde F-skoru doğrusal çekirdek ile LS-SVM sınıflandırıcının birleşimi (Eğitim: 50%- test: 50%)	83.46
	Özellik seçimi F-skoru doğrusal çekirdek ile YSA birleşimi (Eğitim: 50%- test: 50%) 118 özellik azaltılmış	79.69
	Özellik seçiminde F-skoru RBF çekirdek ile LS-SVM sınıflandırıcının birleşimi (Eğitim: %50- test: %50) 118 özellik	83.46

	azalmış.	
	Özellik seçimi F-skoru doğrusal çekirdek ile YSA birleşimi (Eğitim: 50%- test: 50%) 118 özellik azaltılmış	79.69
Zaman ve Karray (2009) (Zaman ve Karray 2009)	Bulanık ESVDF	76.73
McSherry (2011) (McSherry 2011)	iNN(2) metot	84.3
Geliştirilen Metot	k-ortalamalar(10-kat ÇD)	96.73
Geliştirilen Metot	m. k-ortalamalar (10-kat ÇD)	98.18

Diabetes veri kümesi için, sınıflandırma performans değerleri ve diğer sistemlerle karşılaştırmaları sırasıyla Çizelge 7.16 ve Çizelge 7.17'de verilmiştir.

Çizelge 7.16 Diabetes veri seti için sınıflandırma performansı

Performans kriteri	DVM	k-ortalamalar +DVM	M. k-ortalamalar +DVM
Doğruluk (%)	77.60	93.65	96.71
Duyarlılık (%)	88.34	97.75	97.31
Özgünlük (%)	57.32	83.50	95.06
Pozitif öngörü değeri (%)	79.49	93.76	98.21
Negatif öngörü değeri((%)	74.28	93.15	92.05
Zaman tüketimi (Saniye)	6.716754	2.558991	1.815773

Çizelge 7.17 Bizim metodumuzun ve literatürdeki diğer metotların diabetes veri seti üzerinde elde edilmiş sınıflandırma doğrulukları

Yazarlar (Yıl)	Metotlar	Sınıflama Doğruluğu (%)
Sahan et al. (2005) (Sahan ve ark 2005)	AWAIS (10-kat ÇD)	75.87
Polat ve Güneş (2007) (Polat ve Gunes 2007)	PCA ve ANFIS birleştiren	89.47
Polat ve Gunes (2008)(Polat ve ark 2008)	LS-SVM (10-kat ÇD)	78.21
	GDA–LS-SVM (10-kat ÇD)	82.05
Kahramanli ve Allahverdi (2008) (Kahramanli ve	hibrid sistem (ANN ve FNN)	84.2

Allahverdi 2008)		
Tian J. ve ark. (2009) (Tian ve ark 2009)	CC-EBFNN	76.04
Patil (2010) (Patil ve ark 2010)	Hibrid Tahmin Modeli (HPM)	92.38
Isa ve Mamat (2011) (Isa ve Mamat 2011)	Kümelenmiş-HMLP	80.59
Aibinu ve ark.(2011)	AR1+NN (3-kat ÇD)	81.28
Chikh M.A ve ark. (2012) (Chikh ve ark 2012)	AIRS2 (10-kat ÇD)	82.69
	MAIRS2 (10-kat ÇD)	89.10
Ozcift A.(2012) (Ozcift 2012)	RBF+eACC, LADTree+ eACC (10-kat ÇD)	76.30
Ahmad F. ve ark. (Ahmad ve ark 2013) (2013)	Geliştirilmiş Genetik Al(Eğitim: 75%- test: 25%) 10-kez çalışmanın ortalaması	80.4
Geliştirilen Metot	k-ortalamalar (10-kat ÇD)	93.65
Geliştirilen Metot	m. k-ortalamalar (10-kat ÇD)	96.71

Önerilen sistemin, sınıflama doğruluk, hassasiyet ve özgünlük sonucu değerleri sırası ile Statlog (Heart) veri seti için Çizelge 7.12, Spect görüntüleri veri seti için Çizelge 7.14, Diabetes veri seti için Çizelge 7.16'da verilmiştir. Çalışmanın 10-kat çapraz doğrulama yöntemiyle elde edilen ortalama sonuçları literatürde kendi veri setleri üzerinde yapılan araştırmalarda elde edilen sonuçlara göre umut verici üstünlük göstermiştir.

Veri tabanları bulunan veriler içerisinde gürültülü, tutarsız ve eksik elde edilmiş olan veriler içerebilir. Böyle kusurlu veriler sınıflandırıcı başarısını olumsuz yönde etkileyen en önemli faktörlerdendir. Geliştirilen sistem ile kalitesiz bu tip veriler ön işleme ile elenerek sınıflama başarısı yüksek veriler elde edilmiştir. Veri ön işleme işlemi için k-ortalamalar kümeleme metodunun geliştirilmesiyle yapılmıştır. Test işlemleri global üç veri tabanında denenmiştir. Elde edilen sonuçlar literatürdeki sonuçlarla karşılaştırıldığında bu sonuçlardan daha başarılı olduğu gözlenmiştir. Veriler üzerindeki bu eleme yöntemini diğer tüm sınıflama işlemlerinde de kullanılarak sınıflama başarılarının arttırılabileceğine dair emareler alınmıştır.

7.3 Veri setleri için k-kat çapraz doğrulama yöntemiyle entegre edilmiş kümeleme tabanlı veri eliminasyon stratejisiyle sınıflama performansının geliştirilmesi

Veri kümelerinde bulunan eksik, tutarsız ve gürültülü veriler sınıflama sistemlerinin başarısını düşüren en önemli unsurlardandır. Yanlış tamamlanan eksik verilerin, gürültülü ve tutarsız verilerin eliminasyonunun, doğrulama yöntemleri ile uyumlu bir şekilde sınıflama sistemine uygulaması bu çalışmanın ana temasını oluşturmaktadır. Geliştirilen bu sistemde SVM sınıflandırıcısı ile k-ortalamalar algoritması bütünleştirilerek eğitim sürecinin gürültülü ve tutarsız verilerinden arındırılması sağlanmıştır. Test verileri üzerinde herhangi bir işlem yapılmamıştır. Bu sistemde güvenilirliğin ve karşılaştırıla bilirliğin sağlanması amacıyla k- kat çapraz doğrulama kullanılmıştır. Geliştirilen sistemde kullanılan veri kümeleri, UCI veri tabanından alınan Hepatitis, Liver Disorders, SPECT görüntüleri ve Statlog (Heart) verileridir. Önerilen sistem, bu veri kümelerinde sırasıyla %96.83, %76.85, %90.30 ve %90 sınıflama başarısına ulaşmıştır.

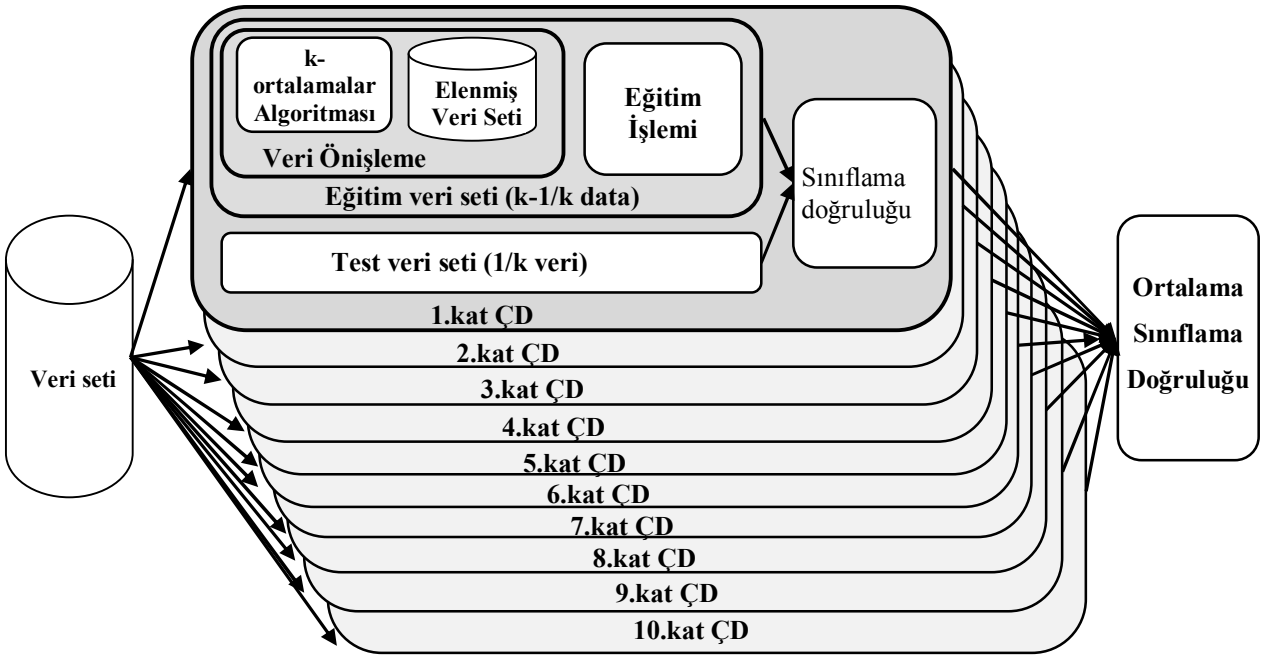
Veri tabanlarında, gürültüler nedeniyle benzer verilere göre tutarsız, birbirinin tekrarı ve bazı öznitelikleri belirsiz veriler bulunmaktadır. Bu verilerin tanıma sistemine girilmesi sistemin performansını önemli ölçüde düşürmektedir. Şekil 7.7'de basitleştirilmiş kodu gösterilen sistem ile k-ortalamalar algoritması kullanılarak eğitim verilerindeki bu kusurlardan bazıları giderilmiştir. Böylece elde edilen yeni veri kümeleri sınıflandırıcının test başarısını ve sistemin kararlılığını artırmaktadır. Gerçekleştirilen bu eleme işlemi sadece eğitim veri kümeleri üzerinde yapılmıştır. Test verileri üzerinden herhangi bir ön işlem yapılmamıştır. Böylece eğitim işlemi daha kaliteli veriler üzerinde yapılırken elde edilen eğitilmiş tanıma sistemin genelleme yeteneği ham haldeki test verileri üzerinden denenmiştir. Önerilen sistem Şekil 7.8'den de görülebileceği gibi veri kümesinin k- kat çapraz doğrulama metoduna göre k adet parçaya bölünmektedir. Bu parçalar sürekli yer değiştirecek şekilde bir parça test için ayrılır, geriye kalan k-1 parça eğitim amacıyla kullanılmaktadır. Test verileri aynen korunurken eğitim verileri üzerinde geliştirilen ön işleme algoritması uygulanarak sınıflandırıcıya eğitim için verilmektedir. Eğitilen sistemin eğitim ve test performansı hesaplanır. k defa bu işlemin tekrarlanması ile elde edilen performans değerlerinin

ortalaması alınarak sistem performansı olarak kaydedilmektedir. Bu sayede yüksek performans değerine sahip, güvenilir, yeni verilere uygulanması oldukça kolay bir sınıflama sistemi elde edilmiştir.

```
Adım 1. Veriler organize edilir.( kolonlarda özellikler, satırlarda örnekler olmak üzere)
Adım 2. (kümelenmiş veri seti, dataindex)= k_means_algoritması(verisetinin girişleri)
Adım 3.
for i=1 to küme sayısı
  Adım 3.1. dağılım_değeri(i)=Frekans_dağılımı(kümelenmiş_veriseti(i)'deki veriseti(dataindex)'in
çıkışı)
  Adım 3.2.
        if dağılım değeri(sınıf-1) > dağılım değeri(sınıf-2)
        then kümelenmiş_veriseti(i) deki sınıf-2 veriyi ele
        else kümelenmiş_veriseti(i) deki sınıf-1 veriyi ele)
end for
```

Şekil 7.7 Önerilen veri eleme yönteminin basitleştirilmiş kod sunumu

Kullanılan veri tabanları, sütunlarda özellikler, satırda ise örnekler olacak şekilde düzenlenmiştir. Eğitim için ayrılan veriler üzerinde önerilen bu veri eliminasyonu metodu uygulanır. Bu metotta N adet kümeye ayrılan eğitim verileri, aynı kümede yer alan örneklerin aynı sınıfa ait olması gerektiği prensibi ile elenir. Buna göre hangi sınıf bir küme içerisinde baskın ise o veriler kalır. Diğerleri elenir. Verilerin N adet kümeye ayırılma işlemi için k-ortalamalar algoritması kullanılmıştır. Eleme işlemi sırasında istenirse bir oran tespit edilip bu oranın altında olan verilerin atılması sağlanabilir. Bu sayede tutarsızlığın yüksek olduğu kümelerde veri eliminasyonu yapılmazken yüksek tutarlı kümelerde bozucu veriler elemine edilir. Bu sayede veri tabanında elimine edilecek verilerin sayısı da ayarlanabilmektedir. Elde edilen veri seti DVM sınıflandırıcısının eğitim verileri olarak kullanılır.



Şekil 7.8 Önerilen sistemin blok diyagramı

Bu çalışma ile literatürdeki yöntemlere nazaran daha güvenilir bir biyomedikal sınıflandırma sistemi geliştirilmiştir. Bu geliştirilen sistemde kullanılan veri tabanları içerdikleri bazı gürültülü, tutarsız ve eksik veriler nedeniyle sınıflandırıcı başarısını olumsuz etkilediği için bir veri eliminasyon işlemine tabi tutulmuşlardır. Eliminasyon süreci güvenilirliğini artırmak için k- kat çapraz doğrulama metodu ile entegre şekilde çalışabilecek şekilde yeniden düzenlenmiştir. Bu düzenlemede veri eliminasyonu işlemi toplam verilerin sadece eğitim için kullanılacak olanları üzerine uygulanmış, test verileri üzerinden bir ön işleme yapılmamıştır. Bu sayede geliştirilen DVM tabanlı sınıflandırıcı bilimsel tüm şüphelerden arındırılmış olarak çalıştırılmaktadır. Bu hali ile eğitilen sınıflandırıcı doğrudan gerçek bir sisteme uygulanabilecek bir nitelik kazanmıştır.

Önerilen sistemin başarısı UCI veri tabanındaki Hepatitis, Liver Disorders, SPECT görüntüleri ve Statlog (Heart) veri setleri kullanılarak gösterilmiştir. Bu denemeler sonucunda elde edilen sistem ile Kalp ve karaciğer hastalıkları için hekime yardımcı daha başarılı otomatik bir teşhisi sistemi elde edilmesine de olanak sağlamıştır. Sistem sonuçları literatürde bu veri tabanları üzerinde çalışılan makaleler ile karşılaştırılmıştır. Literatürdeki değerlerden daha iyi sonuçlara ulaşılmıştır.

7.3.1 Sınıflandırma Parametreleri

Bu çalışmada, gürültülü ve hatalı ölçülmüş verileri DVM sınıflandırıcısının eğitim algoritmaları kısmında tespit edip, sınıflandırma başarısını düşürmesindeki etkilerini azaltmak için tasarlanan sistemimizi, kalp ve diyabet hastalıkları için hazırlanan veri setleri üzerinde test edilmiştir. Test için ayrılan verilerde herhangi bir işlem uygulanmayıp orijinal hali korunmuştur. Sınıflandırıcının performans değerinin güvenirliliği 10-kat çapraz doğrulama yöntemi ile sağlanmıştır. Bu sınıflandırıcı kullanırken, eğitim Çizelge 7.18'de verilen parametrelere göre yapılmıştır.

Çizelge 7.18 Sınıflama Parametreleri Listesi

Parametreler	Değer
Metot	DVM (SVM)
Optimizasyon algoritması	SMO
Doğrulama metodu	k- kat çapraz doğrulama (10-kat ÇD)
Çekirdek Fonksiyonu	Lineer
TolKKT	1.0000e-003
MaxIter	15000
KernelCacheLimit	5000
Başlangıç değer	Rastgele

7.3.2 Deneysel Sonuçlar ve Tartışmalar

Seçilen veri kümeleri için bu çalışmada geliştirilen algoritmanın uygulama sonuçları Çizelge 7.19'da verilmiştir. Bu tablo içeriğinde, kümelerin sayısını, veri giriş - çıkış sayısı, eğitim ve test için ayrılmış örneklem sayısı ve veri eliminasyon işleminde her kat için ayrıntılı olarak eğitim ve test verileri için giriş-çıkış sayıları gösterilir.

Çizelge 7.19 Veri setleri üzerinde algoritmanın uygulanması

Veri			ÇD1	ÇD2	ÇD3	ÇD4	ÇD5	ÇD6	ÇD7	ÇD8	ÇD9	ÇD10
H E	Eğitim	Küme Sayısı	19	19	19	19	19	19	19	19	19	19

PATITIS		Giriş	140	139	139	139	139	139	140	140	140	140
		Çıkış	123	121	121	124	116	122	121	125	120	123
	Test	Ayrılmış Veri	15	15	15	15	15	15	15	15	15	15
LIVER	Eğitim	Küme Sayısı	30	30	30	30	30	30	30	30	30	30
		Giriş	311	310	310	310	310	310	311	311	311	311
		Çıkış	294	288	286	287	290	281	287	290	287	301
	Test	Ayrılmış Veri	34	35	35	35	35	35	34	34	34	34
SPECT	Eğitim	Küme Sayısı	45	45	45	45	45	45	45	45	45	45
		Giriş	241	240	240	240	240	240	240	240	241	241
		Çıkış	201	198	203	205	206	201	202	204	203	206
	Test	Ayrılmış Veri	26	27	27	27	27	27	27	27	26	26
HEART	Eğitim	Küme Sayısı	13	13	13	13	13	13	13	13	13	13
		Giriş	243	243	243	243	243	243	243	243	243	243
		Çıkış	191	178	196	190	184	195	193	193	200	193
	Test	Ayrılmış Veri	27	27	27	27	27	27	27	27	27	27

Hepatitis, Liver disorders, SPECT görüntüleri ve Statlog (Heart) veri kümeleri için geliştirilmiş k-ortalamalar-DVM metoduna ait test sonuçları, sırasıyla Çizelge 7.20, Çizelge 7.22, Çizelge 7.24 ve Çizelge 7.26'da verilmiştir. Bu testlerin sonuçları 10-kat ÇD teknikleri kullanılarak ortalama sonuç değerlerinin içermektedir. Hepatit veri kümesi için, sınıflandırma performans değerleri ve literatürde bulunan diğer sistemlerle karşılaştırılması sırasıyla Çizelge 7.20 ve Çizelge 7.21'de verilmiştir.

Çizelge 7.20 Hepatitis veri seti için sınıflandırma performansı

Performans kriteri	DVM	Tam entegre edilmiş k-ortalamalar + DVM
Doğruluk (%)	90.25	96.83
Duyarlılık (%)	94.59	97.71
Özgünlük (%)	78.81	86.67
Pozitif öngörü değeri (%)	93.49	98.62
Negatif öngörü değeri (%)	77.17	84.67

Çizelge 7.21 Bizim metodumuzun ve literatürdeki diğer metotların Hepatitis veri seti üzerinde elde edilmiş sınıflandırma doğrulukları

Yazarlar (Yıl)	Metotlar	Sınıflama Doğruluğu (%)

Polat ve Güneş (2006) (Polat ve Gunes 2006)	Bulanık ile FS-AIRS (10-kat ÇD)	92.59
Polat ve Güneş (2007) (Polat ve Gunes 2007)	FS-Bulanık-AIRS (%50 eğitim %50 test)	81.82
Polat ve Güneş (2007) (Polat ve Gunes 2007)	FS-Bulanık-AIRS (10 · ÇD)	94.12
Polat ve Güneş(2007)(Polat ve Gunes 2007)	AIRS (10-kat ÇD) PCA– AIRS (10-kat ÇD)	76.00 94.12
Kahramanlı ve Allahverdi (2009) (Kahramanli ve Allahverdi 2009)	Hibrid sistem(ANN ve AIS)	96.80
Dogantekin ve ark. (2009) (Dogantekin ve ark 2009)	LDA-ANFIS	94.16
Bascil ve Temurtaş (2011) (Bascil ve Temurtas 2011)	MLPNN (LM) (10 ÇD)	91.87
Chen ve ark. (2011)(Chen ve ark 2011)	LFDA_SVM	96.77
Sartakhti ve ark. (2012) (Sartakhti ve ark 2012)	SVM-SA(10-kat ÇD)	96.25
Geliştirilen Metot	Tam entegre edilmiş k-ortalamalar + DVM (10-kat ÇD)	96.83

Çizelge 7.21'de görülen sonuçlara göre, en yüksek doğruluk oranı bizim çalışmamıza aittir. Çalışmada, en yakın sonuçlar Kahramanlı ve Allahverdi ile Chen ve arkadaşlarına aittir. Her iki çalışmada sonucun güvenilirliği için kullanılan çapraz doğrulama tekniğinin kullanıldığından bahsedilmemiştir. Bizim çalışmamızda 10 kat çapraz doğrulama kullanılmasına rağmen sonuçları daha yüksek çıkmıştır. Yapılan testler göstermiştir ki, 10-kat çapraz doğrulama kullanılmadığı takdirde, elde edilen sonuçlar daha yüksek olmaktadır.

Liver Disorders veri kümesi için, sınıflandırma performans değerleri ve literatür de bulunan diğer sistemlerle karşılaştırılması sırasıyla Çizelge 7.22 ve Çizelge 7.23'de verilmiştir.

Çizelge 7.22 Liver Disorders veri seti için sınıflandırma performansı

Performans kriteri	DVM	Tam entegre edilmiş k-ortalamalar + DVM
Doğruluk (%)	69.59	76.85

Duyarlılık (%)	83.37	90.32
Özgünlük (%)	50.60	58.14
Pozitif öngörü değeri (%)	69.97	74.98
Negatif öngörü değeri (%)	70.61	82.39

Çizelge 7.23 Bizim metodumuzun ve literatürdeki diğer metotların Liver Hastalığı veri seti üzerinde elde edilmiş sınıflandırma doğrulukları

Yazarlar (Yıl)	Metotlar	Sınıflama Doğruluğu (%)
Lee ve ark. (2001) (Lee ve Mangasarian 2001)	SSVM (10-kat ÇD)	70.33
Van Gestel ve ark. (2002) (Van Gestel ve ark 2002)	LS-SVM + Bayes çercevesinde (10-kat ÇD)	69.70
Goncalves ve ark. (2006) (Goncalves ve ark 2006)	HNFB-1 metot	73.33
Özşen ve Güneş (2008) (Ozsen ve Gunes 2008)	AWAIS (10-kat ÇD) Hibrid benzerlik ölçüsü ile (10-kat ÇD) Manhattan mesafe ile (10-kat ÇD) Öklid mesafe ile (10-kat ÇD)	70.17 60.57 60.21 60.00
Li ve ark.(2011) (Li ve ark 2011)	Bir bulanık tabanlı doğrusal olmayan dönüşüm yöntemi +DVM	70.85
Chen ve ark.(2012) (Chen ve ark 2012)	(PSO) + 1-NN metot(5-kat ÇD)	68.99
Chang ve ark.(2012) (Chang ve ark 2012)	CBR + PSO(Eğitim: 75%- test: 25%)	76.80
Geliştirilen Metot	Tam entegre edilmiş k-ortalamalar + DVM (10-kat ÇD)	76.85

Liver disorders veri kümesi kullanan çalışmalar içerisinde en yüksek sonuç bizim çalışmamıza aittir. Bu çalışmaya oldukça yakın sonuç üreten Chang ve arkadaşları veri kümesini eğitim %75, test %25 olarak ayırmıştır. Bir başka deyişle kat doğrulama kullanılmamıştır. Diğer çalışmaların sonuçları ise elde ettiğimiz başarı oranından yaklaşık %4 daha düşüktür.

SPECT resim veri kümesinde geliştirilen sistemimizin performans değerleri Çizelge 7.24'te ve diğer sistemler yapılan karşılaştırılmalar Çizelge 7.25'te verilmiştir.

Çizelge 7.24 SPECT görüntüleri veri seti için sınıflandırma performansı

Performans kriteri	DVM	Tam entegre edilmiş k-ortalamalar + DVM
Doğruluk (%)	77.52	90.30
Duyarlılık (%)	55.26	64.52
Özgünlük (%)	84.11	96.71
Pozitif öngörü değeri (%)	46.49	83.48
Negatif öngörü değeri((%)	87.68	91.59

Çizelge 7.25 Bizim metodumuzun ve literatürdeki diğer metotların Spect görüntüleri veri seti üzerinde elde edilmiş sınıflandırma doğrulukları

Yazarlar (Yıl)	Metotlar	Sınıflama Doğruluğu (%)
Polat ve Güneş (2007) (Polat ve Gunes 2007)	FS-Fuzzy-AIRS (%50 eğitim %50 test)	92.59
Polat ve Güneş (2009) (Polat ve Guenes 2009)	Özellik seçimi olmadan LS-SVM sınıflama(Eğitim: 50%- test: 50%)	82.71
	Özellik seçimi olmadan LM_YSA (Eğitim: 50%- test: 50%)	77.44
	Özellik seçimi F-skoru ile LS-SVM birleşimi (Eğitim: 50%- test: 50%)	79.70
	Özellik seçimi F-skoru Lineer çekirdek ile YSA birleşimi (Eğitim: 50%- test: 50%) 7 özellik azaltılmış.	75.70
	Özellik seçiminde F-skoru Lineer çekirdek ile LS-SVM sınıflandırıcının birleşimi (Eğitim: 50%- test: 50%)	83.46
	Özellik seçimi F-skoru Lineer çekirdek ile YSA birleşimi (Eğitim: 50%- test: 50%) 118 özellik azaltılmış	79.69
Zaman ve Karray (2009) (Zaman ve Karray 2009)	Bulanık SVDF	76.73
McSherry (2011) (McSherry 2011)	iNN(2) metot	84.30
Saraçoğlu (2012) (Saracoglu 2012)	PCA-DHMM metot	72.20
Bizim çalışmamız	Tam entegre edilmiş k-ortalamalar + DVM (10-kat ÇD)	90.30

Spect veri kümesi kullanan çalışmalar içerisinde en yüksek değer Polat ve Güneş'e aittir. Bu sonuç ile bizim çalışmamız arasında %2.29 fark vardır. Yalnız Polat ve Güneş veri kümesini %50 eğitme %50 test olarak ayırmış ve herhangi bir doğrulama işlemi kullandığını belirtmemiştir. Bizim çalışmamızda ise elde ettiğimiz sonuçları doğruluğunu ve tutarlılığını gösterebilmek için mutlaka bir kat doğrulama metot

kullanılmıştır. Bu şartlar altında bile en yüksek sonuç olmasa da yeterli sayıda araştırmacını üzerinde çalışmadığı bu veri tabanı için ikinci iyi sonuca sahiptir.

Statlog Heart Disease veri kümesinde geliştirilen sistemimizin performans değerleri Çizelge 7.26'da ve diğer sistemler yapılan karşılaştırılmalar Çizelge 7.27'de verilmiştir.

Çizelge 7.26 Statlog Heart Disease veri seti için sınıflandırma performansı

Performans kriteri	DVM	Tam entegre edilmiş k-ortalamalar + DVM
Doğruluk (%)	82.22	90.00
Duyarlılık (%)	80.83	86.60
Özgünlük (%)	86.26	92.17
Pozitif öngörü değeri (%)	83.26	91.55
Negatif öngörü değeri((%)	83.06	89.97

Çizelge 7.27 Bizim metodumuzun ve literatürdeki diğer metotların Statlog kalp hastalığı veri seti üzerinde elde edilmiş sınıflandırma doğrulukları

Yazarlar (Yıl)	Metotlar	Sınıflama Doğruluğu (%)
Duch ve ark. (2001) (Duch ve ark 2001)	k-NN, k=28, 7 özellikler (10-kat ÇD)	84.60-85.60
	k-NN, k=28, Manhattan (10-kat ÇD)	82.20-83.40
	FSM, 27 Bulanık kurallar	82.00
	SSV, 3 kurallar	80.20-83.40
Şahan ve ark. (2005) (Sahan ve ark 2005)	AWAIS (10-kat ÇD)	82.59
Eick ve ark. (2006) (Eick ve ark 2006)	NCC (10-kat ÇD)	81.07
Özşen ve Güneş (2008) (Ozsen ve Gunes 2008)	Hibrid benzerlik ölçüsü ile AIS algoritması (10-kat ÇD)	83.95
Kahramanlı ve Allahverdi (2008)(Kahramanli ve Allahverdi 2008)	YSA ve FNN kullanarak Hibrid sistem (10-kat ÇD)	86.80
Polat ve Güneş (2009)(Polat ve Guenes 2009)	Özellik seçimi olmadan LS-SVM sınıflama(Eğitim: 50%- test: 50%)	80.00
	Özellik seçimi olmadan LM_YSA (Eğitim: 50%- test: 50%)	71.11
	Özellik seçimi F-skoru ile LS-SVM birleşimi (Eğitim: 50%- test: 50%)	77.78
	Özellik seçimi F-skoru ile YSA birleşimi	77.61

	(Eğitim: %50- test: %50) 6 özellik azaltılmış.	
	Özellik seçiminde F-skoru doğrusal çekirdek ile LS-SVM sınıflandırıcının birleşimi (Eğitim: %50- test: %50)	80.00
	Özellik seçimi F-skoru doğrusal çekirdek ile YSA birleşimi (Eğitim: 50%- test: 50%) 128 özellik azaltılmış	80.74
	Özellik seçiminde F-skoru RBF çekirdek ile LS-SVM sınıflandırıcının birleşimi (Eğitim: %50- test: %50) 117 özellik azalmış.	83.70
	Özellik seçimi F-skoru doğrusal çekirdek ile YSA birleşimi (Eğitim: 50%- test: 50%) 117 özellik azaltılmış	76.30
Bizim Çalışmamız	Tam entegre edilmiş k-ortalamalar + DVM (10-kat ÇD)	90.00

Bu sonuçlara göre, en yüksek ortalama doğruluk oranı bizim çalışmamıza aittir. Bu çalışmanın en yakın sonuçları Kahramanlı ve Allahverdi'ye aittir. Fakat onların çalışmaların da 10-kat çapraz doğrulama yönteminin kullanıldığından bahsedilmemiştir. Bu çalışma ve bizim çalışmamız arasında %3.2 fark vardır. Çizelge 7.27'da görüldüğü gibi, çalışmamız Statlog Kalp Hastalığı veri kümesi için en yüksek doğruluk oranı elde edilmiştir.

Önerilen sitemin, sınıflama doğruluk, hassasiyet ve özgünlük sonucu değerleri sırası ile Hepatitis için Çizelge 7.20, Liver Disorders veri seti için Çizelge 7.22, spect görüntüleri veri seti için Çizelge 7.24 Statlog Heart veri seti için Çizelge 7.26'da verilmiştir.

Çalışmanın 10-kat çapraz doğrulma yöntemiyle elde edilen ortalama sonuçları literatür de kendi veri setleri üzerinde yapılan araştırmalarda elde edilen sonuçlara göre umut verici üstünlük göstermiştir.

Bu çalışma ile verilerin literatürdeki diğer yöntemlere kıyasla daha iyi sınıflandırılmasına yardımcı olacak bir sistem geliştirilmiştir. Sistem 10-kat ÇD ile bir küme-tabanlı veri eliminasyonu ve DVM sınıflandırıcıya tamamen entegre olarak çalıştırılmıştır. Sınıflama işlemleri sırasında verilerin kalitesi önemlidir. Gerçek hayatta elde edilen veriler içersin de sıklıkla sorunlar bulunur. Veri girişi ya da veri toplanması sırasında oluşan sistem dışı hatalara gürültü adı verilir. Bu gürültüler verilerin toplanması sırasında insan, yazılım, donanım hatalarından kaynaklanabilir. Geliştirilen sistem ile kullanılan veri setlerinde sınıflandırıcıların başarısını olumsuz etkileyen,

gürültülü tutarsız ve eksik verilerin ortadan kaldırılması işlemine tabi tutulmuştur. Eliminasyon işlemi sınıflama işlemlerinde güvenilirlik için kullanılan 10-kat çapraz doğrulama yöntemi ile entegre bir şekilde çalışacak şekilde düzenlenmiştir. Bu çalışmada eliminasyon işlemi yalnızca, eğitim için kullanılacak verilerde uygulanır ve test verileri herhangi bir ön işleme sokulmayıp ham halde tutulur. Bu şekilde, gelişmiş DVM tabanlı sınıflandırıcı ham veriler üzerinde test edilerek gerçek verilerin hatalı gelmeye devam etmesi halinde bile sınıflandırıcıyı daha hızlı ve etkili çalıştırmayı başarmıştır. Bu tasarım sayesinde, eğitimli sınıflandırıcı bir gerçek sisteme doğrudan uygulanacak nitelik kazandırılmıştır. Önerilen sistemin başarısı global veri tabanlarından olan UCI'den Hepatit, karaciğer Bozuklukları, SPECT görüntüleri ve Statlog (Kalp) veri setlerini kullanarak gösterilmiştir. Bu çalışmaların bir sonucu olarak, elde edilen sistem ile kalp ve karaciğer hastalıklarının teşhisi için geliştirilmiş otomatik bir tanı sistemi de kazandırılmıştır. Sistem ile doktorlara yardımcı olacak aynı zamanda tanılarını kontrol edecekleri bir sistemin başarısını artıracak iyileştirilmeler yapılmıştır. Önerilen sistemin sınıflandırılması 10-kat çapraz doğrulama uygulanarak hepatit için %96.83 karaciğer Bozuklukları için %76.85 Spect görüntüleri için %90.30 ve Statlog (Kalp) veri kümesi için %90'na ulaşmıştır. Sistemin sonuçları bu veri setlerini araştıran literatürdeki makaleler ile de karşılaştırılmıştır. Elde edilen sonuçlara göre önerilen yöntemin sınıflama doğrulukları diğer sonuçlarına göre genellikle daha başarılı ve örüntü tanıma uygulamaları için çok umut verici olduğu görülmüştür. Geliştirilen bu sistem diğer tüm görüntü tanıma uygulamalarında da uygulanabileceği düşünülmektedir.

7.4 Yüksek hızlı termal transfer yazıcı kafa hasarlarının belirmesini sağlayan görüntü işleme ve veri madenciliği teknikleri tabanlı uzman sistem tasarımı

Bu çalışmada, görüntü işleme yöntemleri ile termal transfer baskı makinesinin, baskı kafasının zarar görmüş bölgelerini tespit eden bir uzman sistem tasarlanmış ve bu uzman sistemi geliştirmek için sistem içinde kullanılan sınıflama sistemine zarar veren gürültülü değerlerin elenmesi işlemi geliştirilmiştir. Gürültülü ve yanlış sınıflanan veriler elenerek sınıflama başarısı yükseltilmiştir. Böylece uzman sistemin başarısı artırılmıştır.

Geliştirilen algoritmanın amacı, yazdırma işlemi sırasında ısıtılan baskı kafası direnç bölgelerinde oluşan hasarlı bölgelerin bir uzmana gerek olmadan bir sistem tarafından tespit edilmesinin sağlanmasıdır. Çalışmada kullanılan fotoğraflar bir mikroskop ile alınır ve dirençli parçaların görüntüleri piksel piksel sınıflandırıcıya verilir.

Termal transfer baskı teknolojisi, şu anda özellikle paketler üzerine endüstriyel baskı yazıcısı olarak kullanılır. Bu yazıcılarda bir baskı kafası, şerit ve yazdırılabilir bir ortamda oluşan bir yazdırma ünitesi vardır. Küçük çaplı elektrotlar baskı kafasının gerekli bölgelerini ısıtarak yazdırması gereken yüzeye şerit üzerinden boyanmasını sağlar. Bir nokta bu yolla yazılır. Noktalardan resimler ve harfler oluşturulur. Elektrotlar bir noktaya milisaniyeler kadar ısı uygular (Pennington ve Crooks 1985).

Kalın ve ince filme baskı yapan iki tür termal baskı kafası vardır. İnce filme baskı yapan termal kafalar yüksek çözünürlükte baskı yapar. Kalın filme baskı yapan termal baskı kafası daha düşük enerji harcamasına rağmen verimi diğerine göre kötüdür. Bu dezavantaja, muhtemelen ısı elemanlarının koruyucu katmanı etkili olmaktadır (Ishizuka 1989).

Yüksek hızda termal transfer baskı için kullanılan matematiksel modelleri Uyhan R. Ve arkadaşları (Uyhan ve King-Hele 2008) ile Nolde K. ve arkadaşlarının(Nolde ve Morari 2010) yayınladığı yayınlarda tanımlanmıştır. Baskı kafaları bu modeller kullanılarak çalıştırılabilmektedir.

Baskı kafalarının hataları üç nedene dayanır.

a) Isı yalıtımı sağlayan sır tabakası düşük gevşeme sıcaklığına sahiptir. Yazıcı kafası kritik sıcaklığın üstünde ısıtıldığında sır tabakası yok olur.
b) Koruyucu tabaka aşınıp yok olduktan sonra baskı kafası da zedelenip iş göremez duruma gelir. Koruyucu tabakanın aşınması; bileşime, baskı basıncı ve termal kâğıt özelliklerine bağlıdır.
c) Baskı kafaları kritik sıcaklığının altında işletilenlerde yeterince kalın bir koruyucu tabakada olsa zamanla bu tabaka ısı dirençlerinin yavaş yavaş azalıp iletkenliklerini kaybederler. Bu bozulma asıl olarak çalışma sıcaklığı ve ısı darbe süresine bağlıdır (Tokunaga ve ark 1981).

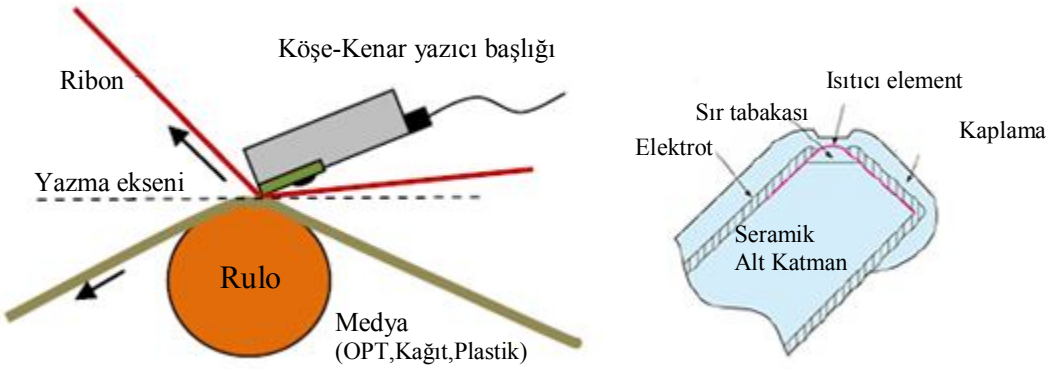
Bunların yanı sıra yazıcıların kirli yerlerde yanlış şekilde kullanılması ısıtma işlemleri sırasında baskı kafası yüzeyinde zararlar oluşturabilir. Bu çalışmada bozulan

bu bölümleri görüntü işleme ve sınıflama metoduyla tespit eden bir uzman sistem geliştirilmiştir. Daha sonra bu sistemi yanıltan kirli ve tutarsız verileri elemine edecek ön işleme algoritması geliştirilerek uzman sistemin başarısı artırılmıştır.

7.4.1 Sistem tasarımı

Termal yazıcı kafalarında bulunan ısı elemanları bir mikroskop buna bağlı kamera sistemi ile görüntülenmiştir. Düz bir yüzeye sahip bir tepsi bir adım motora bağlandı. Yazıcı kafası bir 10X 7.3mm mikroskop (Meiji-MT9430 Polarize Mikroskop) merceği altına bir mikrostep motor sürücüsü ile hareketlendirilen bir lineer hareket sistemine bağlanmıştır. Bu sistem sayesinde el değmeden yazıcı kafanın tüm ısı elemanlarının görüntülenmesi sağlanmaktadır. Tüm görüntüleme işlemi mikroskop ile uyumlandırılmış bir CCD kamera ile alınmıştır. Kameradan görüntüler USB2.0 portu üzerinden bilgisayara aktarılmaktadır. Bu sistem tarafından her baskı kafası için yaklaşık 80 görüntü alınıp ve bilgisayara aktarılmıştır.

Şekil 7.9'da, hızlı termal transfer yazıcılar için kullanılan ve çalışmamızda ele aldığımız kenar-baskı kafasının (Corner Edge Print Head) yapısı ve sürekli yazım için çalışma şekli verilmiştir.



Şekil 7.9 a) Termal transfer baskı Sürekli stili b) Baskı kafası yapısı

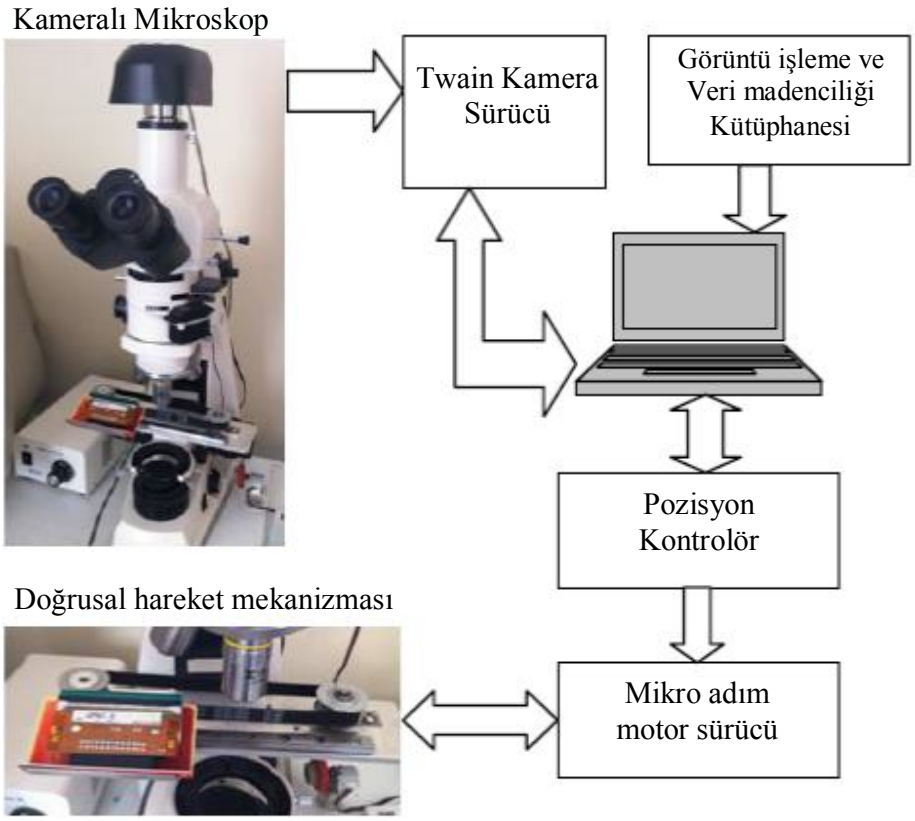
7.4.1.1 Tasarlanan Mekanik Sistemin Donanımı

Bu uygulama çalışması sırasında Şekil 7.10'da gösterilen Meiji marka MX9430 serisi, sabit kameralı trinoküler polarizan mikroskop kullanılmıştır.



Şekil 7.10 Meiji MX9430 serisi mikroskop

Meiji marka üst aydınlatma sistemine sahip sabit kameralı polarizan mikroskobu kullanarak incelemede kullanılmak üzere bilgisayar kontrollü otomatik elektromekanik bir sistem tasarımı yapılmış ve mikroskoba Şekil 7.11'deki gibi monte edilmiştir. Burada; TPH, mikroskop, kamera, hareket kontrol ünitesi ve bilgisayardan oluşan bir test sisteminin dizaynı gösterilmiştir.



Şekil 7.11 TPH Analiz Sistemi Tasarım

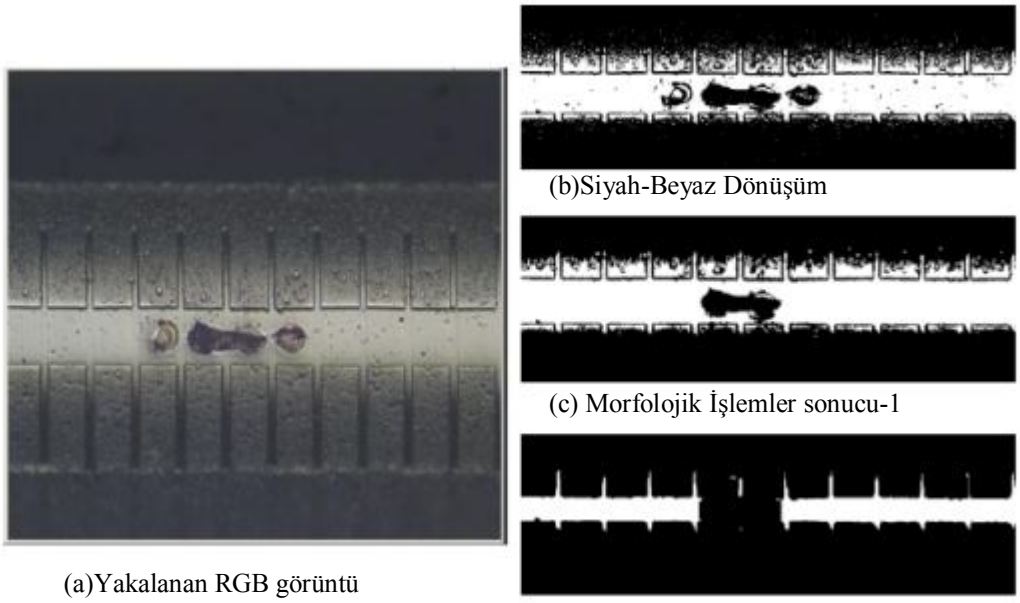
Çizelge 7.28 Kamera 590CU marka CCD kameranın teknik özellikleri

Özellik	Değer
Maksimum çözünürlük	5 megapiksel
Maksimum resim boyutu	2592*1944 piksel
Sensör tipi	CCD renkli
Sensör boyutu	½ inç
Piksel ölçüleri	3,2 * 3,2 mikron
Efektif piksel	2048 x 1536 piksel
Yakalama hızı	21 fps
Mikroskop bağlantı arayüzü	c
Bit derinlik değeri	3x8 bit
Pc monitör ön izleme	2048 x 1536 piksel
Oküler bağlantı aparatı	Opsiyonel
Harici kullanım	Yok
IR filtre	Var
Pc bağlantısı	Usb 2.0
İşletim sistemi uyumu	Win XP

Özellikleri Çizelge 7.28'de verilen USB bağlantısıyla mikroskoba monte edilmiş CCD kameradan alınan görüntüler, bilgisayara aktarılmaktadır. Motorların kontrolü, bir motor kontrol kartı üzerindeki PIC18F4620 mikro denetleyici aracılığıyla yapılmaktadır. Bilgisayardan gelen komutlar, motor kontrol kartına RS232/USB dönüştürücü ile iletilmektedir.

7.4.1.2 Görüntü İşleme ve Özelliklerin Çıkartılması

Baskı kafası hasar görmüş bölümlerin daha iyi belirlenmesi için çalışmada bir görüntü işleme algoritması geliştirilmiştir. İlk adım olarak, ısıtma bölgesindeki bir yazdırma kafasının programla fotoğrafı oluşturulur (Şekil 7.12a).



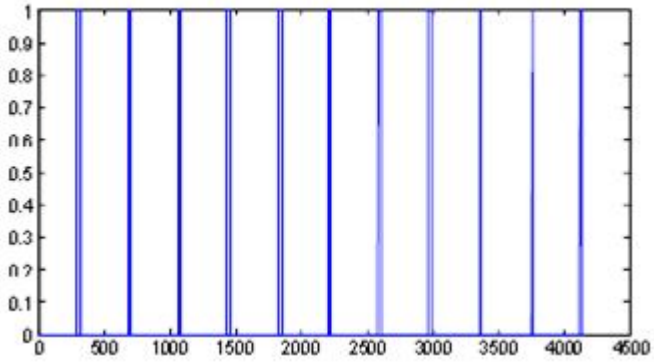
(a)Yakalanan RGB görüntü

(b)Siyah-Beyaz Dönüşüm

(c) Morfolojik İşlemler sonucu-1

(d) Morfolojik İşlemler sonucu-2

Şekil 7.12 Tasarlanan sistemin ön işleme algoritması Sonuçları



Şekil 7.13 Her ısıtma hücresinin yatay pozisyonlarının tespiti

Baskı kafasından elde edilen görüntü gri değerde RGB formatına dönüştürülür. Görüntü gri değere dönüştürüldükten sonra yatay ve gri değer tahminleri hesaplanarak yazıcının ana ısıtma hücre bölgesi bulunur. Şekil 7.12b'de, 11 ısıtma hücresinin gri değere dönüştürülmesi gösterilir. Şekil 7.13'de kendi yatay konumları gösterilir (Yilmaz ve ark 2012).

```
//Veri toplama:
  a=Resim okuma

//Ön İşleme:
   Gray_image=RGB to Gray(a)
   BW_image==image to BW(Gray_image)
   BW_image==BW_area_open(BW, 250)
   BW_image= image_erode(BW_image, Rect(150,3))

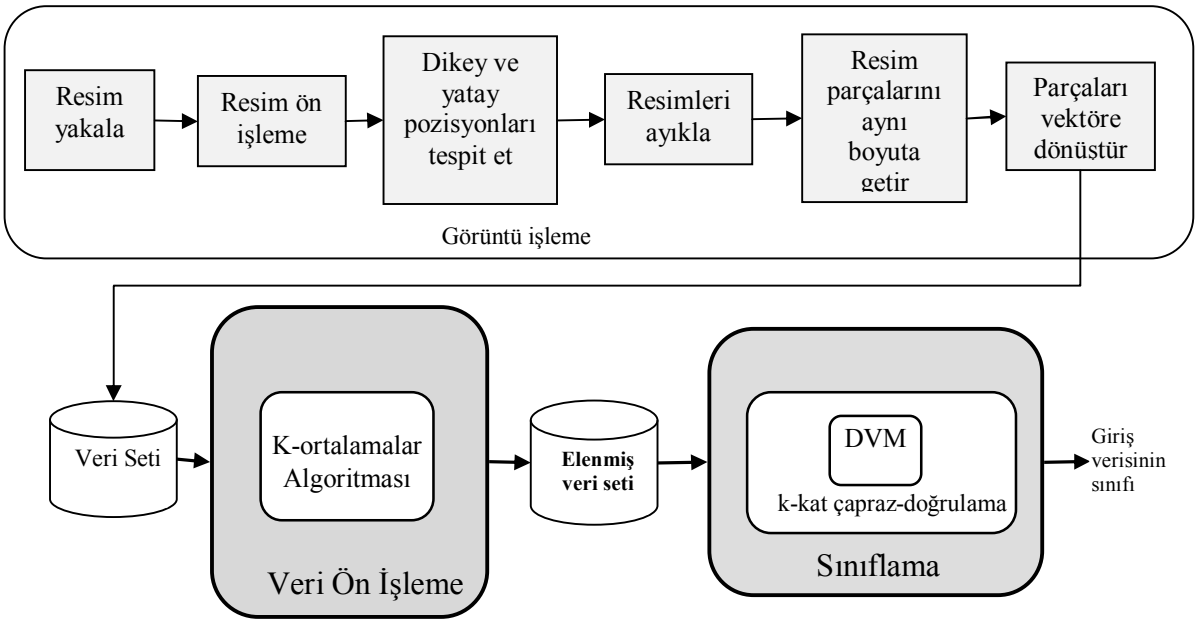
// Isıtma hücre dizisinin Dikey pozisyon hesaplaması
   PV=vertical_projection(BW_image)
   PV_peaks= PV< (max(PV)*0.8)
   Diff_PV_peaks=delayed_diff(PV_peaks,2)
   Vertical_min=find_position(Diff_PV_peaks=1)
   Vertical_max=find_position(Diff_PV_peaks=-1)
   Vert_pos=[ Vertical_min Vertical_max]

// Isıtma hücrelerinin Yatay Pozisyon hesaplama
   PH=Horizontal_projection(BW_image(Vertical_min-300:Vertical_min-50));
   PH_peaks=PH>(max(PH)*0.45);
   Diff_PH_peaks=delayed_diff(PH_peaks,2)
   Hor_Pos=find_position(Diff_PH_peaks)

// Tüm Isıtma hücrelerin Segmentasyon:
for i=1:size(Horizontal_Pos)-2
    if((Horizontal_Pos (i+1)- Horizontal_Pos (i))>250)
    new_image=a(Vert_pos±200, Hor_Pos(i):Hor_Pos(i+1));
   else new_image=a(Vert_pos±200,Hor_Pos(i+1):Hor_Pos(i+2)); end
   image_write(new_image);
 end for
```

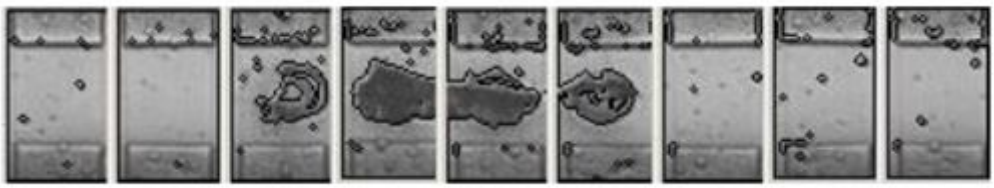
Şekil 7.14 Tüm ısıtma hücrelerinin bölütlemesi için basitleştirilmiş kodu (Pseudo-Code)(Yilmaz ve ark 2012)

Şekil 7.14'te termal kafa resimlerinin bölütlemesinin nasıl yapıldığı basitleştirilmiş kod olarak verilmiştir. Şekil 7.15'te ise geliştirilen sistem ile veriyi işleme adımları bloklar halinde gösterilmiştir. Buna göre ilk önce resim alınır. Resimdeki hücreler ön işleme adımlarından geçirilerek parazitlerden ve kusurlardan arındırılır. Her bir ısıtma elementinin dikey ve yatay pozisyonu bulur. Bu pozisyonlara uygun olarak ilk resim üzerinden her bir element resimden ayıklanır. Bu ayıklanmış element görüntüleri bir dizi haline dönüştürülerek eğitim ve test verileri oluşturulur. Oluşturulan veri kümesi geliştirilen veri önişleme ve sınıflandırma süreçlerine tabi tutularak ısıl elementin sağlam veya hasarlı olduğu konusunda bir sonuca ulaşılır.

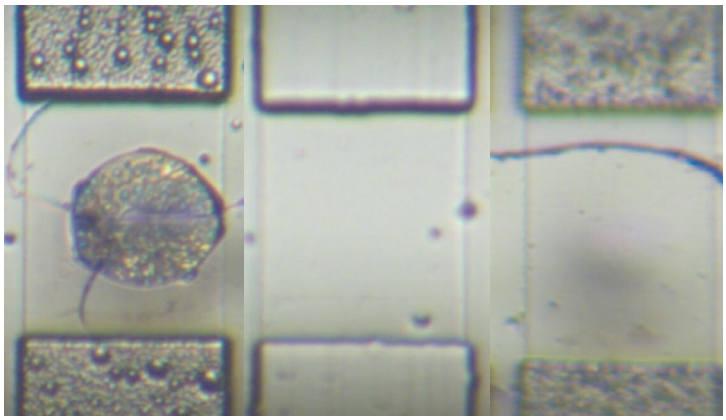


Şekil 7.15 Geliştirilen Sistemin Veri İşleme Blok Şeması

Isıtma hücre bölgelerinin bir program tarafından temiz görüntüleri alınarak karşılaştırma için kullanılır. Temiz görüntüler üzerinden şişmemiş dış baskı kafaları seçilir. Desenleri, şablon eşleme algoritmalarıyla karşılaştırma için hazırlanır. Şekil 7.16'da pixel kıyaslandıktan sonraki son pixel görüntüleri elde edilmiştir.

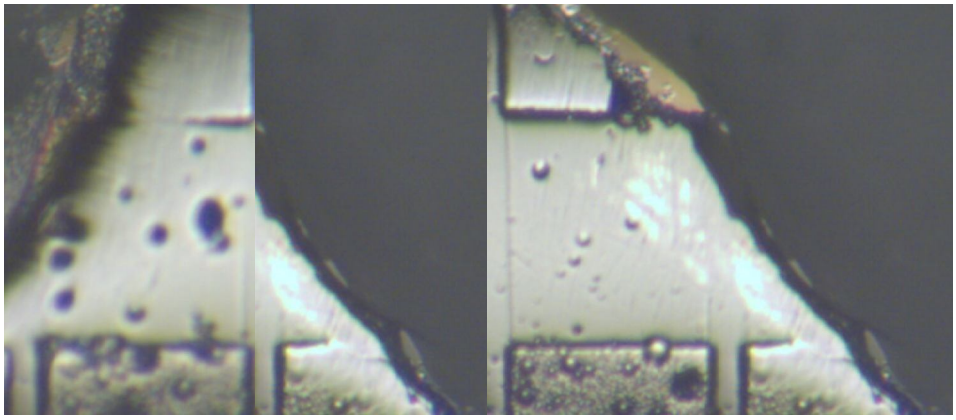


Şekil 7.16 Son pixel görüntüleri



Şekil 7.17 Termal baskı kafalarından elde edilen örnek görüntüler

Şekil 7.17'de elde edilen termal baskı kafalarının görüntülerinden sağlam, hasar görmüş örnek resimler bulunmaktadır. Görüntünün solundan itibaren ilk resim hasarlıdır. İkinci resim sağlam olduğu görünmektedir. Üçüncü resim çizikler ve boya kalıntıları olmasına rağmen hasarsızdır.



Şekil 7.18 Baskı kafalarının sisteme alınamadığı görüntülerinden örnekler

Şekil 7.18'de göründüğü gibi termal baskı kafa ısıl elementlerin ayrımı sırasında parçalanmış veya baskı kafasının en baş ve en son hücreleri olduğundan alınamayan görüntüler sınıflama verileri içine konmamıştır.

Tasarlanan sistemden elde edilen termal yazıcı kafaları görüntüleri bir uzman tarafından bozuk olup olmadığına göre sınıflandırılmıştır. Elde edilen görüntüler RGB formatında MATLAB'da yazılan bir program tarafından okunup gri formata

dönüştürülüp tek boyutlu bir matris elde edilmiştir. Bu matris daha sonra 80'e 48'lik boyuta indirgenmiştir. Matris tek satıra dönüştürülüp sınıflamanın giriş değerlerini temsil edecek şekilde program tarafından ayarlanmıştır. Giriş değerleri DVM sınıflandırıcısında eğitilerek bir uzman sistem tasarlanmıştır. Sınıflandırma işlemi sırasında 10-kat çapraz doğrulama işlemi uygulanarak sınıflandırıcının güvenliği artırılmıştır.

Tasarlanan uzman sistemin başarı oranını artırmak için Bölüm 7.2'de uyguladığımız yanlış sınıfa atanan değerleri ve sınıflamanın başarısını düşüren gürültülü değerleri eleyen uygulama sistem üzerinde denenmiştir. Sistemin başarısının artığı görünmüştür.

Giriş değeri olarak termal yazıcı kafaları görüntülerinin eleme işlemi aşağıdaki Şekil 7.19'daki gibi yapılmıştır;

```
Adım 1. Veriler organize edilir.( kolonlarda özellikler, satırlarda örnekler olmak üzere)
Adım 2. (kümelenmiş veri seti, dataindex)= k_means_algoritması(veri setinin girişleri)
Adım 3.
for i=1 to küme sayısı
  Adım 3.1. dağılım_değeri(i)=Frekans_dağılımı(kümelenmiş_veriseti(i)'deki veriseti(dataindex)'in
çıkışı)
  Adım 3.2.
        if dağılım değeri(sınıf-1) > dağılım değeri(sınıf-2)
        then kümelenmiş_veriseti(i) deki sınıf-2 veriyi ele
        else kümelenmiş_veriseti(i) deki sınıf-1 veriyi ele)
end for
```

Şekil 7.19 Önerilen veri eleme yönteminin basitleştirilmiş kodunun sunumu

Tek satırda temsil edilecek şekilde elde edilen 3466 adet termal yazıcı kafası görüntüsü, sütunlarda özellikler, satırda ise örnekler olacak şekilde düzenlenmiştir. Düzenlenen veriler üzerinde önerilen veri eliminasyonu metodu uygulanmıştır. Bu metot da N (9) adet kümeye ayrılan eğitim verileri, aynı kümede yer alan örneklerin aynı sınıfa ait olması gerektiği prensibi ile elenir. Buna göre hangi sınıf bir küme içerisinde baskın ise o veriler kalır. Diğerleri elenir. Verilerin N adet kümeye ayırılma işlemi için k-ortalamalar algoritması kullanılmıştır. Eleme işlemi sırasında istenirse bir oran tespit edilip bu oranın altında olan verilerin atılması sağlanabilir. Bu sayede tutarsızlığın yüksek olduğu kümelerde veri eliminasyonu yapılmazken yüksek tutarlı kümelerde bozucu veriler elemine edilip, elimine edilecek verilerin sayısı da

ayarlanabilmektedir. Elde edilen veri seti DVM sınıflandırıcısının girişine verilmiştir. Sonuçta sistemin başarısının arttığı görünmüştür. Elde edilen sonuçlar Çizelge 7.29'da gösterilmiştir.

7.4.2 Sınıflandırma Parametreleri

Bu çalışmada, gürültülü ve hatalı sınıflamış verileri DVM sınıflandırıcısının eğitim algoritmaları kısmında tespit edip, sınıflandırma başarısını düşürmesindeki etkilerini azaltmak için tasarlanan sistemimizi, termal yazıcı kafalarından elde ettiğimiz görüntülerden oluşturulmuş veri setleri üzerinde test edilmiştir. Sınıflandırıcının performans değerinin güvenirliliği 10-kat çapraz doğrulama yöntemi ile sağlanmıştır. Bu sınıflandırıcı kullanırken, eğitim Çizelge 7.29'da verilen parametrelere göre yapılmıştır.

Çizelge 7.29 Sınıflamada kullanılan parametre listesi

Parametreler	Değer
Metot	DVM
Optimizasyon algoritması	SMO
Doğrulama metot	k- kat çapraz doğrulama (10-kat ÇD)
Kernel_Fonksiyon	Lineer
TolKKT	1.0000e-003
MaxIter	15000
KernelCacheLimit	5000
Başlangıç değeri	Rastgele

7.4.3 Deneysel Sonuçlar ve Tartışmalar

Çizelge 7.30 Termal yazıcı kafalarından elde ettiğimiz görüntülerden oluşturulmuş veri seti için sınıflandırma performansı

Performans Kriterleri	DVM	k-ortalamalar +DVM	M. k-ortalamalar +DVM
Doğruluk (%)	96.45	99.77	99.88
Hassasiyet (%)	89.72	99.03	99.30

Özgünlük (%)	98.37	99.92	100
Pozitif öngörü değeri (%)	97.11	99.81	99.85
Negatif öngörü değeri (%)	94.08	99.61	100

Çizelge 7.30'da göründüğü gibi termal yazıcı kafalarından elde edilen görüntü verilerinin sınıflama başarısı uyguladığımız sistem sayesinde artmıştır. Sistem yanlış sınıflanan veya gürültü olarak tabir edilen sınıflama başarısını bozan veriler elenmiştir. Çizelge 7.31'de termal yazıcı kafalarından elde edilen görüntülerin kaç adet verisinin kullanıldığı ve çıkan sonuçların karşılaştırılmaları verilmiştir.

Çizelge 7.31 Termal yazıcı kafalarından elde ettiğimiz görüntülerden oluşturulmuş veri seti için saf hali ve ön işlemeden sonraki sonuçları

Veri Seti	Sınıf Sayısı	Özellik Sayısı	Küme Sayısı	Örnekler	Temizlenmiş Örnek sayısı	10-Kat ÇD ile ortalama doğruluk	KN=500 ve 10-Kat ÇD için Ortalama Doğruluk
Normal veri	2	3840	-	3466	3466	96.45	96.22
Ön işleme	2	3840	9	3466	3138	99.77	99.66
Modifiye Ön İşleme yapılmış	2	3840	9	3466	3207	99.88	99.75

Kirli, sıcak ortamlarda çalışma gibi dış etkilerin sonucu olarak termal transfer yazıcı baskı kafalarının bozulmaları görüntü işleme ve sınıflama yöntemiyle belirlenmiştir. Bu çalışma hızlı bir şekilde yanmış ve bozulmuş parçaları tespit etmek için yararlı olmuştur. Sistem geliştirilerek ilerde daha iyi termal transfer yazıcılar geliştirmek için veriler elde edilmesini sağlayacaktır. Ayrıca bu sistemle termal transfer yazıcıların bazı sorunlarının belirlenmesi ve bunların çözümü ile ilgili çalışmalarda yardımcı olabilecek bilgilerin elde edilebilmesi içinde geliştirilebilir.

Tasarlanan sistem ile bir uzmana gerek kalmadan termal yazıcı kafalarının bozuk olup olmadığını anlayan otomatik bir sistem geliştirilmiştir. Bu sistem uzman kişinin bile karar vermekte zorlanacağı bilgilerle hızlı, otomatik bir şekilde ve %99.88 civarında doğruluk oranı ile termal yazıcı kafasının bozuk olup olmadığını tespit edebilmiştir.

8 SONUÇ VE ÖNERİLER

Bilişim alanındaki teknolojinin gelişmesi ile oluşan veri yığınları çözümü karmaşık problemleri beraberinde getirmiştir. Veri yığınlarındaki veriler değerlendirilip anlamlı bilgiye dönüştürülmediği sürece değersizdir. Bu verilerin bilgiye dönüştürülme işlemleri için genellikle veri madenciliği yöntemleri kullanılmıştır. Bu tezde de ayrıntılı bir şekilde veri madenciliği yöntemleri açıklanmıştır. Veri madenciliği yöntemleri kullanılarak oluşturulan dört tane yeni sistem önerilmiştir.

Örneklerden yararlanarak bir olayın öğrenilmesi işlemi yapay zekâ ve istatistik ile uğraşan bilim adamlarının ve araştırmacıların odak noktası olmuştur. İçinde örneklerden faydalanılarak eğitim yapılan algoritmalardan en yaygın olarak kullanılanları YSA ve DVM'dir. YSA ve DVM'ler sınıflandırılma işlemleri için çözümü zor birçok problemlerde hızlı ve etkin çözüm tekniği olarak başarı ile uygulanmaktadır. Bu iki algoritma klasik yöntemlere göre günümüz problemlerini çözmede başarılı olsa da bu algoritmaların uygulanması sırasında verilerin elde edilmesi ve hazırlanması esnasında oluşan sorunlar nedeniyle istenen başarılı sonuçlar elde edilemeyebilmektedir. Verilerin büyümesiyle gürültü diye tabir edilen yanlış, eksik girilmiş değerler, örneklerden sonuç çıkaran sistemlerin sonuçlarını etkileyebilen sorunlar çıkarmaktadır. Tez çalışmasında çeşitli yöntemlerle bu gürültülü verileri tespit ederek sınıflamaya etkilerini azaltılacak 3 farklı sistem geliştirilmiştir. Geliştirilen sistemler literatürde sıklıkla kullanılan veri tabanlarında ve kendi elde ettiğimiz termal baskı kafalarının görüntülerinden oluşan veri tabanlarında test edilmiş ve sınıflama başarısını artırdığı görülmüştür.

Geliştirilen ilk sistemde, sınıflandırıcıların hızlı ve etkili bir şekilde çalışmasını engelleyen gürültülü ve tutarsız verilerin tespiti için veri madenciliği metotlarından k-ortalamalar algoritması kullanılarak sınıflandırıcının önünde bir ön işleme yöntemi tasarlanmıştır. Tasarlanan sistemde k-ortalamalar algoritmasında kullanılan Öklid uzaklık ölçütüne bir ağırlık eklentisi geliştirilmiş ve daha hassas gürültü tespiti sağlanmıştır. Geliştirilen yöntem UCI veri tabanından alınan kalp ve diyabet hastalıkların tespiti için kullanılan veri setleri ile test edilmiştir. Kalp veri setlerinden birinde elde edilen bilgiler, kabin sol kapacığı üzerinden alınan görüntü verilerinden elde edilmiştir. Geliştirilen sistemlerin sonuçlarının güvenliği açısından 10-kat çapraz

doğrulama kullanılmıştır. Ayrıca 10-kat çapraz doğrulama sonucunda elde edilen ortalama sonuçlarla yetinilmeyip sistem 10-kat çapraz doğrulama yöntemi 500 kere çalıştırılarak ortaya çıkan sonuçların da ortalamaları ve standart sapmaları ilgili bölümde verilmiştir. Çıkan sonuçlar literatürde bulunan sonuçlarla karşılaştırılmıştır. Statlog kalp hastalığı veri seti için elde edilen 10-kat çapraz doğrulamalı sonuç, literatürdeki 10-kat çapraz doğrulama kullanılan veya kullanılmayan tüm sistemlerden daha iyi sonuç vermiştir. K-ortalamalar algoritması geliştirilerek elde edilen sistemin sonucu ile karşılaştırdığımız diğer sistemlerin sonuçları arasında %6.67'lik bir fark ortaya çıkmıştır. SPECT görüntüleri veri setinde de elde edilen 10-kat çapraz doğrulamalı sonuç literatürdeki katlı ve kat kullanılmadan geliştirilen tüm sistemlerden daha iyi sonuç vermiştir. Literatürde taranan sistemler arasında bizim dışımızda tasarlanan sistemlerden en yakın sonuç ile %13.88'lik bir fark ortaya çıkmıştır. Diyabet veri seti için elde edilen 10-kat çapraz doğrulamalı sonuç, literatürdeki 10-kat çapraz doğrulama kullanılan veya kullanılmayan tüm sistemlerden daha iyi sonuç vermiştir. Literatürde taranan sistemler arasında bizim dışımızda tasarlanan sistemlerden en yakın sonuçla %4.33'lik bir fark ortaya çıkmıştır. Test edilen sonuçlardan da görüleceği gibi geliştirilen sistem ile veri yığınlarının içindeki sınıflamayı bozan veriler çok başarılı bir şekilde tespit edilmiştir.

Geliştirdiğimiz gürültülü ve yanlış sınıflanmış verileri tespit eden sistemimizin gerçek problemlerdeki performansını ölçmek amacıyla termal baskı kafalarının bozuk olup olamadığını tespit etmek için tasarladığımız sistem üzerinde testler yapılmıştır. Bu sistemde de %3.35'lik bir başarı artışı tespit edilmiştir. Sistemde bir mikroskop ile kurduğumuz sistemde termal yazıcı kafalarının görüntüleri alınmıştır. Elde edilen görüntüler her bir termal hücre ayrı ayrı resimlere ayrılmıştır. Elde edilen her bir hücrenin görüntüsü RGB formatına çevrilerek 80'e 48'lik boyutta bir matris ile ifade edilmiştir. Her bir termal hücre görüntüleri için tanımlanan matrisler, sıralı tek bir satır matrise dönüştürülüp oluşan bu veriler sınıflandırıcının girişine verilmiştir. Elde edilen sınıflama başarı değerini yükseltmek için k-ortalamalar algoritmasıyla oluşturduğumuz ön işleme yöntemi, sisteme ilave edilerek sistem tekrar çalıştığında sistemin başarısını artıran bir etki görülmüştür. Elde edilen sistem ile örneklemlerde yanlış sınıflanan veya gürültü adını verdiğimiz elemanlar tespit edilerek sınıflama başarısı yüksek veri setleri elde edilmiştir. Bu uygulama tıp, mühendislik, uzay bilimleri gibi birçok alanda

görüntülerden elde edilen veri setlerini düzenlemek için kullanılıp daha başarılı sınıflama başarıları elde edilebileceği düşünülmektedir.

Geliştirdiğimiz gürültü ve tutarsız verileri temizleme sistemimizi k-kat çapraz doğrulama sisteminin içine entegre edilerek(bütünleştirilerek) sadece eğitim verileri üzerinde eliminasyon yapıp test verileri orijinal verilerle korunarak karşılaştırılan başka bir sistem geliştirilmiştir. Bu sistem sayesinde katlı kısmında eğitilen verinin başarısını ölçmek için kullanılan test verisi üzerinde hiçbir işlem yapılmayarak ham halde başarı testi yapılmıştır. Burada gerçek uygulamalarda veri setlerine hala gürültülü veri gelmesi halinde sistemin nasıl çalışacağının testinin yapılması sağlanmıştır. Bu yöntemin performansını ölçmek için veri madenciliği çalışmalarında performans karşılaştırma için sıklıkla kullanılan veri tabanlarından alınan dört adet veri setinde denenmiştir. Bu veri setleri UCI veri tabanından alınan Hepatitis, Liver Disorders, SPECT görüntüleri ve Statlog (Heart) verileridir. Bu veri setlerinin seçilmesinin sebebi literatürde raporlanmış diğer araştırmalarla geliştirdiğimiz yöntemi karşılaştırmaktır. Uyguladığımız yöntemde sınıflama sonucunun güvenirliliğini artırmak için 10-kat çapraz doğrulama kullanılmıştır. Yapılan testler göstermiştir ki, k-kat çapraz doğrulama kullanılmadığı takdirde, elde edilen sonuçlar daha yüksek çıkmaktadır. Buna rağmen sistemimizi çapraz doğrulama yapılmayan sistemlerle de karşılaştırılmıştır. Hepatitis veri seti için uygulanan yöntem ulaşabildiğimiz literatürdeki kat kullanılarak ve kat kullanılmadan performans hesabı yapılan sistemler arasında en yüksek sonucu vermiştir. Bu veri setinde karşılaştırılan en yakın yöntemlerde doğrulama tekniğinin kullanıldığından bahsedilmemiştir. Buna rağmen daha yüksek sonuç elde edilmiştir. Liver Disorders veri seti için uygulanan yöntem ulaşabildiğimiz literatürdeki k-kat ÇD doğrulama kullanılarak ve k-kat ÇD doğrulama kullanılmadan performans hesabı yapılan sistemler arasında da en yüksek sonucu vermiştir. En yakın rakibi katlı doğrulama kullanmadan veri setinin %75 eğitim için ayırmış, %25'ini de test için ayırarak başarısını ölçmüştür. Buna rağmen çalışmamız ondan daha iyi sonuç vermiştir. Spect veri setini kullanan çalışmalar içerisinde en yüksek değer, önişleme için uyguladığımız yöntemi saymazsak Polat ve Güneş'e aittir. Bu sonuç ile bizim katlı çapraz doğrulama içinde uyguladığımız çalışmamız arasında %2.29 fark vardır. Polat ve Güneş veri kümesini %50 eğitme %50 test olarak ayırmış ve herhangi bir doğrulama işlemi kullandığını belirtmemiştir. Statlog kalp hastalığı veri seti üzerinde yapılan çalışmalar arasında ulaşabildiğimiz literatürdeki

katlı çapraz doğrulama kullanılarak ve katlı çapraz doğrulama kullanılmadan performans hesabı yapılan sistemler arasında da en yüksek sonucu vermiştir. Bu çalışmada da en yakın çalışma ile arasında %3.2'lik fark oluşmuştur. Elde edilen sonuçlara göre önerilen yöntemin sınıflama doğrulukları diğer sonuçlarına göre genellikle başarılı ve örüntü tanıma uygulamaları için çok umut verici olduğu görülmüştür.

Çağımızda teknolojinin gelişmesiyle elde edilmek istenen bilgiye çabuk ulaşılma eğilimi vardır. Elde edilen veri yığınlarında istenen bilgiye ulaşmak için kullanılması gerekmeyen veri özellikleri vardır. Tez çalışmasında aynı zamanda bu özelikleri tespit eden hibrid bir sistem geliştirilmiştir. Bu hibrid sistem ile örneklemlerden sonuç çıkarmak için kullanılan YSA ve DVM sınıflandırıcılarının daha az nitelik ile çalıştırılarak daha başarılı sonuçlar bulunmuştur. Yapılan bu hibrid nitelik azaltımı ile sistemin daha hızlı çalışması sağlanmıştır. Aynı zamanda tespit edilen niteliklerin bu sınıflama işlemleri için gerekmediğinden dolayı bundan sonra bu alanda yapılan çalışmalarda o niteliklerin elde edilmesi için kaybedilen zamandan kazanılmıştır. Gerekmeyen nitelikler için ekstradan kullanılan iş gücü varsa araç gereçlerden de tasarruf edilmesi sağlanmıştır. Sistem büyük boyutlu veri yığınlarının içindeki öznitelikler arasındaki ilişkileri tespit eden birliktelik kuralarından apriori algoritması ile ilişkili nitelikler içeren veri yığınlarının boyutlarını daha az boyuta indirgenmesini sağlayan bir dönüşüm tekniği olan TBA (PCA) birleştirilerek bir hibrid öznitelik seçme yöntemi oluşturulmuştur. Sistemi, YSA ve DVM sınıflandırıcıların önüne ön işleme yöntemi olarak koyarak sınıflamayı artıran başarı elde edilmiştir. Yöntem UCI veri tabanında sıklıkla kullanılan Wisconsin göğüs kanseri veri setinde test edilmiştir. Literatürde bu veri setinde yapılan araştırmalarla karşılaştırıldığında k- katlı çapraz doğrulama kullanan araştırmalar arasında Polat ve Güneş tarafından sunulan LS-SVM sisteminden sonra en iyi sonucu veren sistem olmuştur. Bu çalışma ile de bizim sistemimiz arasında %0.24'lük bir fark çıkmıştır. Literatürde bulunan diğer sistemlerde sınıflama sonucunun güvenirliliği için kullanılan k-katlı çapraz doğrulama kullanıldığı belirtilmemiştir. Muhtemelen tek kat için bulunan en yüksek değer yazılmış olmasına rağmen aradaki fark %1'den az çıkmıştır. Kullandığımız sistemin sağladığı avantajlar düşünüldüğünde özellik seçiminde ilerde başka tıbbi veri setlerinde kullanılarak diğer hastalıkların tespitinde de gerekli olamayan niteliklerin tespit edilip hastalardan bu

niteliklerin bulunması için istenen testler alınmadan daha kısa sürede ve daha az test ile işlemler gerçekleştirilmiş olacaktır. Tıp alanı dışında da kullanılarak gerekmeyen nitelikler tespit edilip örneklemden sınıflama işlemi kullanan sistemlerde gereksiz niteliklerin tespit edilmesi için harcanan zaman ve emekten kazançlar elde edilecektir.

Tez çalışmasında geliştirilen sistemler ile hem uygulamalarda avantajlar sağlanmış hem de literatürdeki görüntü ve metin tabanlı verilerde oluşan gürültülü ve gereksiz özniteliklerin kaldırılması konusunda mevcut olan yöntemlere yeni anlayışlar ortaya konulmuştur. Bu anlayışlar oluşturulurken veri madenciliği algoritmalarının geliştirilmesi için yeni eklentiler yapılmıştır. Sınıflandırıcıların önünde ve k-katlı çapraz doğrulamanın içine koyduğumuz ön işleme yöntemiyle sınıflamalarda ciddi bir performans artışı sağlanmıştır. Yapay zekâ sistemlerindeki gelişmeler sayesinde verilerin daha etkin kullanımına yönelik çalışmaların hızlandığı günümüzde artan ve karmaşıklaşan veri yığınlarından ihtiyaç duyulan veri ön işleme sistemlerine duyulan ihtiyacın arttığı da bir gerçektir. Birçok uygulamada özellikle YSA ve DVM gibi sınıflandırıcıların başarılı ve pratik bir şekilde kullanılan bir yöntem olsa da, bu sınıflandırıcıları bozan gürültü tabir edilen veriler sınıflandırıcıların başarılarını düşürerek bazı problemlerde kullanımların da sorun oluşturmuşlardır. Ayrıca bu sınıflandırıcılarda sınıflama için gerekmeyen nitelikler hem başarıların düşmesini hem de daha uzun sürelerde sonuç verilmesini sağlarlar. Bu ihtiyaçlar göz önüne alınarak geliştirilen sistemlerimizin sergilediği başarılı sonuçlar ile literatürdeki yapay zekâ ve sınıflama çözümleri arasında hak ettiği yeri almaya başlamışlardır. Sınıflamaların daha başarılı bir şekilde yapılması için geliştirdiğimiz sistemler, bu alanda önemli bir yer edinecektir.

KAYNAKÇA

Abonyi J, Szeifert F, 2003. Supervised fuzzy clustering for the identification of fuzzy classifiers. Pattern Recogn Lett, 24, 14, 2195-207.

Agrawal R, Imielinski T, Swami A, (1993). Mining Association Rules Between Sets of Items in Large Databases. ACM SIGMOD Conference on Management of Data. Washington.

Agrawal R, Srikant R, (1994). Fast algorithms for mining association rules. Proceeding of the 20th Int'l Conference on Very Large Databases. Santiago.

Ahmad F, Isa NAM, Hussain Z, Osman MK, 2013. Intelligent Medical Disease Diagnosis Using Improved Hybrid Genetic Algorithm - Multilayer Perceptron Network. J Med Syst, 37, 2.

Akay MF, 2009. Support vector machines combined with feature selection for breast cancer diagnosis. Expert Syst Appl, 36, 2, 3240-7.

Akbulut S, 2006. Veri Madenciliği Teknikleri Đle Bir Kozmetik Markanın Ayrılan Müsteri Analizi Ve Müsteri Segmentasyonu.

Akdemir B, 2009. Tahmin Uygulamalarında Performans Geliştirmek İçin Kullanılan Normalizasyon Metotlarına Yeni Bir Yaklaşım, Selçuk Üniversitesi, Konya.

Alataş B, 2002. Veri madenciliği, Fırat Ünv. .

Almuallim H, Dietterich T, (1991). Learning with many irrelevant features. Proceeding of AAAI 91, AAAI Press**:** 547-52.

Alpaydın E. Zeki Veri Madenciliği: Ham Veriden Altın Bilgiye Ulasma Yöntemleri, 1-10, Bilisim 2000 Eğitim Semineri.

Anderberg MR, 1973. Cluster Analysis for applications, New York, Academic Press, p.

Ayhan S, 2013. Kaba Küme ve Destek Vektör Makineleri Kullanılarak Nitelik İndirgeme ve Sınıflandırma Problemlerinin Çözümü İçin Bütünleşik Bir Yaklaşım, Eskişehir Osmangazi Üniversitesi.

B. Y, 2006. Döviz Kuru Tahmininde Yapay Sinir Ağlarının Kullanımı, Osmangazi üniversitesi, Eskişehir.

Bascil MS, Temurtas F, 2011. A Study on Hepatitis Disease Diagnosis Using Multilayer Neural Network with Levenberg Marquardt Training Algorithm. J Med Syst, 35, 3, 433-6.

Baş N, 2006. Yapay Sinir Ağları Yaklaşımı ve Bir Uygulama, Mimar Sinan Güzel Sanatlar Üniversitesi, İstanbul.

Bishop CM, 1995. Neural Networks for Pattern Recognition, Oxford, Clarendon Press, p.

Blake CL, M.C.J., (1998). UCI repository of machine learning databases.

University of California at Irvine Repository of Machine Learning Databases, 1998. http://www.ics.uci.edu/~mlearn/MLRepository.html, Erişim. Erişim adresi, http://www.ics.uci.edu/~mlearn/MLRepository.html.

Blashfield RK, Aldenferder MS, 1978. The Literature on Cluster Analysis. Multivariate Behavioral Research.

Bozik O, 2011. Ana Bileşenler Analizine Dayalı Yapay Sinir Ağı Yaklaşımı İle Radarla Uçak Sınıflandırma, Hava Harp Okulu Komutanlığı.

Breiman L, Friedman J, Olshen R, Stone C, 1984. Classification and regression trees, Monterey,CA, Wadsworth & Boks / Cole Advanced Boks & Software, p.

Burges CJC, 1998. A tutorial on support vector machines for pattern recognition,data mining and knowledge discovery, Kluwer Academic Publishers, p.

Busuttil S, 2003. Support vector machines, University of Malta, In Proceedings of the Computer Science Annual Research Workshop.

Chan KCC, Wong AKC, 1991. A statistical technique for extracting classificatory knowledge from databases, p. 107-123.

Chang PC, Lin JJ, Liu CH, 2012. An attribute weight assignment and particle swarm optimization algorithm for medical database classifications. Comput Meth Prog Bio, 107, 3, 382-92.

Chapelle O, 2007. Training a support vector machine in the primal. Neural Comput.

Chen HL, Liu DY, Yang B, Liu J, Wang G, 2011. A new hybrid method based on local fisher discriminant analysis and support vector machines for hepatitis disease diagnosis. Expert Syst Appl, 38, 9, 11796-803.

Chen HL, Yang B, Liu J, Liu DY, 2011. A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis. Expert Syst Appl, 38, 7, 9014-22.

Chen LF, Su CT, Chen KH, Wang PC, 2012. Particle swarm optimization for feature selection with application in obstructive sleep apnea diagnosis. Neural Comput Appl, 21, 8, 2087-96.

Chen MC, Lin CP, 2007. A data mining approach to product assortment and shelf space allocation. Expert Syst Appl, 32, 4, 976-86.

Chen Y-L, Chen J-M, Tung C-W, 2006. A Data Mining Approach For Retail Knowledge Discovery With Consideration Of The Effect Of Shelf-Space Adjacency On Sales. 42, 3.

Cherkassky V, Mulier F, 2007. Learning from Data: Concepts, Theory, and Methods, ABD, Wiley-IEEE Press, p.

Chikh MA, Saidi M, Settouti N, 2012. Diagnosis of Diabetes Diseases Using an Artificial Immune Recognition System2 (AIRS2) with Fuzzy K-nearest Neighbor. J Med Syst, 36, 5, 2721-9.

Yapay Sinir Ağları, 2015. Erişim. Erişim adresi, http://ube.ege.edu.tr/~cinsdiki/UBI521/Chapter-1/cinsdikici-neural-net-giris.pdf.

Cortes C, Vapnik V, (1995). Support-vector Networks. Machine Learning**:** 273–97.

Cristianini N, Shawe-Taylor J, 2000. An introduction to support vector machines, Cambridge University Press, p.

Crone SF, (2004). Business Forecasting with Artificial Neural Networks. Institute of Business Forecasting. Boston, Presentation of IBF Tutorial.

Cura T, 2008. Modern Sezgisel Teknikler ve Uygulamaları, İstanbul, Papatya Yayıncılık, p.

Çakar Ö, 2007. Fonksiyonel analize giriş I, A.Ü. Fen Fakültesi Döner Sermaye İşletmesi Yayınları no:13, p.

Çatalbaş MC, 2014. Temel Bileşenler Analizi Ve Kanonik Korelasyon Analizi İle İmge Tanıma Ve Sınıflandırma, Hacettepe Üniversitesi.

Çorumluoğlu Ö, Özbay Y, Kalaycı İ, Şanlıoğlu İ, (2005). GPS yüksekliklerinden ortometrik yüksekliklerin elde edilmesinde yapay sinir ağı (YSA) tekniğinin kullanımı ". 2.Mühendislik Ölçmeleri Sempozyumu. İstanbul, İTU**:** 552-60.

Diday E, Simon JC, 1976. Clustering analysis in Digital Pattern Recognition. Springer-Verlag.

Dietterich TG, 2003. Machine Learning, London, Nature Encyclopedia of Cognitive Science, p.

Dinçer E, 2006. Veri Madenciliğinde K-Means Algoritması Ve Tıp Alanında Uygulanması, KOCAELİ ÜNİVERSİTESİ.

Ding Q, 2001. Association rule mining survey.

Dogantekin E, Dogantekin A, Avci D, 2009. Automatic hepatitis diagnosis system based on Linear Discriminant Analysis and Adaptive Network based on Fuzzy Inference System. Expert Syst Appl, 36, 8, 11282-6.

Döslü A, 2008. Veri Madenciliğinde Market Sepet Analizi Ve Birliktelik Kurallarının Belirlenmesi, YILDIZ TEKNİK ÜNİVERSİTESİ.

Duch W, Adamczak R, Grabczewski K, 2001. A new methodology of extraction, optimization and application of crisp and fuzzy logical rules. IEEE Transactions on Neural Networks, 12, 2, 277-306.

Efe MÖ, Kaynak O, 2004. Yapay Sinir Ağları ve Uygulamaları, İstanbul, Boğaziçi Üniversitesi, p.

Eick CF, Rouhana A, Bagherjeiran A, Vilalta R, 2006. Using clustering to learn distance functions for supervised similarity assessment. Eng Appl Artif Intel, 19, 4, 395-401.

Elmas Ç, 2003. Yapay Sinir Ağları Kuram, Mimari, Eğitim, Uygulama, Ankara, Seçkin Yayıncılık, p. 27-28.

Elmas M, 2012. Destek Vektör Makineleri İle Fiyat Tahminleri Ve Kuyumculuk Sektöründe Bir Uygulama, İstanbul Üniversitesi.

Emir Ş, 2013. Yapay Sinir Ağları ve Destek Vektör Makineleri Yöntemlerinin Sınıflandırma Performanslarının Karşılaştırılması: Borsa Endeks Yönünün Tahmini Üzerine Bir Uygulama, İstanbul Üniversitesi, İstanbul.

Eray O, 2008. Destek vektör makineleriyle ses tanıma uygulaması, Pamukkale Üniversitesi.

Erdoğan ŞZ, 2004. Veri Madenciliği ve Veri Madenciliğinde Kullanılan K-means Algoritmasının Öğrenci Veri Tabanında Uygulanması, İstanbul Üniversitesi.

Everitt B, Landau S, Leese M, 2001. Cluster Analysis, London, Oxford University Pres, p.

Fausett L, 1994. Fundamentals of neural networks: architectures, algorithms, and applications, p.

Fausett L, 1994. Fundementals of Neural Networks - Architectures, Algorithms, and Applications, Prentice Hall, p.

Fayyad UM, Piatesky-Shapiro G, Smyth P, Uthurusamy R, 1994. Advances in data mining and Knowledge Discovery, USA, AAAI Pres, p.

Support vector machines explained, 2009. Erişim. Erişim adresi, www.cs.ucl.ac.uk/sta_/T.Fletcher/.

Freeman JA, Skapura DM, 1991. Neural Networks – Algorithms, Applications and Programming Techniques, USA, Addison-Wesley Publishing Company, p.

Goncalves LB, Vellasco MMBR, Pacheco MAC, de Souza FJ, 2006. Inverted hierarchical neuro-fuzzy BSP system: A novel neuro-fuzzy model for pattern classification and rule extraction in databases. Ieee T Syst Man Cy C, 36, 2, 236-48.

Goodman DEB, L. ; Watkins, A. , (2002). Artificial immune system classification of multiple-class problems. Proceedings of the Artificial Neural Networks in Engineering ANNIE 02**:** 179-83.

Griffith DA, 2005. An Examination Of The Influences Of Store Layout In Online Retailing. Journal of Business Research, 58, 10.

Grossberg S, 1988. Nonlinear Neural Networks: Principles, Mechanisms, and Architectures, Neural Networks, p. 17-61.

Grzymala-Busse JW, On the unknown attribute values in learning from examples, Proceeding of Methodologies for Intelligent Systerrzs. Lecture Notes in AI, Springer-Verlag**:** 368-77.

Gunn SR, 1998. Support vector machines for classification and regression, School of Electronics and Computer Science.

Gülce AC, 2010. Veri Madenciliğinde Apriori Algoritması Ve Apriori Algoritmasının Farklı Veri Kümelerinde Uygulanması, Trakya Üniversitesi

Günal EŞ, 2010. Örüntü Tanıma Uygulamalarında Fraktal Boyut Yardımıyla Öznitelik Çıkarımı, Osmangazi Universitesi, Eskişehir.

Günal S, 2008. Örüntü Tanıma Uygulamalarında Altuzay Analiziyle Öznitelik Seçimi ve Sınıflandırma, Eskişehir Osmangazi Üniversitesi.

Günay Atbaş A, Celile., 2008. Kümeleme Analizinde Küme Sayısının Belirlenmesi Üzerine Bir Çalışma, Ankara Üniversitesi.

Hamel LH, 2009. Knowledge Discovery with Support Vector Machines, Wiley-Interscience, p.

Han J, Fu Y, (1995). Discovery Of Multiple-Level Association Rules From Large Databases. Proceedings of the 21st VLDB Conference. Zurich**:** 420-31.

Han J, Kamber M, 2001. Data Mining Concepts and Techniques, Morgan Kauffmann Publishers, p.

Haykin S, 1999. Neural Networks. Prentice-Hall, New Jersey, p. 842.

He J, Lan M, Tan C-L, Sung S-Y, Low H-B, (2004). Initialization of Cluster Refinement Algorithms: AReview and Comparative Study. Neural Networks IEEE International Joint Conference**:** 297 – 302.

Hertz JA, Krogh AS, Palmer RG, 1991. Introduction to the Theory of Neural Computation. Redwood City, Addison-Wesley Publishing Company, p.

http://en.wikipedia.org/wiki/Standard_score, 2015. Erişim. Erişim adresi, http://en.wikipedia.org/wiki/Standard_score.

Diabet, 2015. Erişim. Erişim adresi, http://tr.wikipedia.org/wiki/Diyabet.

Meme Kanseri, 2015. Erişim. Erişim adresi, http://tr.wikipedia.org/wiki/Meme_kanseri.

Hubert L, 1974. Approximate Evaluation Techniques for the Single-Link and Complete-Link Hierarcihal Clustering Procedures. Journal of the American Statistical Association.

Isa NAM, Mamat WMFW, 2011. Clustered-Hybrid Multilayer Perceptron network for pattern recognition application. Appl Soft Comput, 11, 1, 1457-66.

Ishizuka M, (1989). Application of the thermal network method to the thermal analysis of thermal printer heads used in high speed thermal printers. IEEUCHMT '89 Japan IEMT Symposium**:** 311-4.

Jain AK, Dubes RC, 1988. Algorithms for Clustering Data, New Jersey, Prentice Hall, p.

Johnson AR, Wichern DW, 1988. Applied Multivariate Statistical Analysis, New Jersey, Prentice-Hall International Editions, p. 554.

Kahramanli H, 2008. Hibrit Bulanık Sinir Ağını Kullanarak Bir Sınıflandırma Ve Kural Çıkarma Sisteminin Geliştirilmesi, Selçuk Üniversitesi.

Kahramanli H, Allahverdi N, 2008. Design of a hybrid system for the diabetes and heart diseases. Expert Syst Appl, 35, 1-2, 82-9.

Kahramanli H, Allahverdi N, 2009. Extracting rules for classification problems: AIS based approach. Expert Syst Appl, 36, 7, 10494-502.

Kantardzic M, 2011. Data Mining: Concepts, Models, Methods, and Algorithms, New Jersey, Wiley-Blackwell, p.

Karabatak M, 2008. Özellik Seçimi, Sınıflama Ve Öngörü Uygulamalarına Yönelik Birliktelik Kuralı Çıkarımı Ve Yazılım Geliştirilmesi, FIRAT ÜNİVERSİTESİ.

Karabatak M, Ince MC, 2009. An expert system for detection of breast cancer based on association rules and neural network. Expert Syst Appl, 36, 2, 3465-9.

Karabulut E, 2009. Mamografi Görüntülerinin Değerlendirilmesinde Örüntü Tanıma Temelli Bir Karar Destek Sistemi, Fırat Üniversitesi

Karabulut EM, Ibrikci T, 2012. Effective Diagnosis of Coronary Artery Disease Using The Rotation Forest Ensemble Method. J Med Syst, 36, 5, 3011-8.

Kaufman L, Rousseeuw PJ, 1990. Finding Groups in Data: An Introduction to Cluster Analysis, John Wiley and Sons, p.

Kecman V, 2001. Learning and Soft Computing: Support Vector Machines, Neural Networks and Fuzzy Logic Models, Cambridge, A Bradford Book, p.

Kira KR, L., (1992). The feature selection problem: Traditional methods and a new algorithm. Proc. of AAAI 92, AAAI Press**:** 129-34.

Kohavi G, John R, 1997. Wrappers for feature subset selection. Artificial Intelligence, 97, 273–324.

Kohavi R, (1995). A study of cross validation and bootstrap for accuracy estimation and model selection. The Fourteenth International Joint Conference on Artificial Intelligence. San Francisco, Morgan Kaufman**:** 1137-45.

Kohei A, Barakbah AR, (2007). Hierarchical K-means: an algorithm for centroids initialization for K-means. Reports of the Faculty of Science and Engineering, Sga University.

Koyuncu F, 2004. Veri madenciliğinde apriori temelli ilişkilendirme kurallarının uygulama ve karşılaştırması, Yıldız Teknik Ünv., İstanbul.

A Brief Introduction to Neural Networks, 2007. Erişim. Erişim adresi, http://www.dkriesel.com.

Kuan TW, Wang JF, Wang JC, Lin PC, Gu GH, 2012. VLSI Design of an SVM Learning Core on Sequential Minimal Optimization Algorithm. Ieee Transactions on Very Large Scale Integration (Vlsi) Systems, 20, 4, 673-83.

Kurgan LA, Cios KJ, Tadeusiewicz R, Ogiela M, Goodenday LS, 2001. Knowledge discovery approach to automated cardiac SPECT diagnosis. Artif Intell Med, 23, 2, 149-69.

Küçükönder H, 2011. Yapay Sinir Ağları Ve Tarımda Bir Uygulama, Kahramanmaraş Sütçü İmam Üniversitesi, Fen Bilimleri Enstitüsü.

Ladha L, Deepa T, 2011. Feature Selection Methods And Algorithms. International Journal on Computer Science and Engineering (IJCSE) 3 (5).

Larose DT, 2005. Discovering Knowledge in Data, Wiley, p.

Larose DT, 2005. Discovering Knowledge in Data: An Introduction to Data Mining, Wiley-Interscience, p.

Lay D, 2000. Linear Algebra and It's Applications, Newyork, Addison-Wesley, p.

Lee SK, (1992). An extended relational database model for uncertain and imprecise information. Proceeding of the ISth VLDB conference**:** 211-8.

Lee YJ, Mangasarian OL, 2001. SSVM: A smooth support vector machine for classification. Comput Optim Appl, 20, 1, 5-22.

Li DC, Liu CW, Hu SC, 2011. A fuzzy-based data transformation for feature extraction to increase classification performance with small medical data sets. Artif Intell Med, 52, 1, 45-52.

Li S, Li H, Li M, Shyr Y, Xie L, Li Y, 2009. Improved prediction of lysine acetylation by support vector machines. Protein and peptide letters, 16.

Luba T, Lasocki R, (1994). On unknown attribute values in functional dependencies. Proc. of the International Workshop on Rough Sets and Soft Computing. San Jose, CA**:** 490-7.

MacQueen J, (1967). Some methods for classification and analysis of multi-variate observations. In: Proc. of the Fifth Berkeley Symp. on Math., Statistics and Probability, Berkeley: U. California Press**:** 281-97.

Marcano-Cedeno A, Quintanilla-Dominguez J, Andina D, 2011. WBCD breast cancer database classification applying artificial metaplasticity neural network. Expert Syst Appl, 38, 8, 9573-9.

Marill T, Green DM, 1963. On the effectiveness of receptors in recognition system. IEEE Trans. Inform. Theory, 9, 11–7.

McSherry D, 2011. Conversational case-based reasoning in medical decision making. Artif Intell Med, 52, 2, 59-66.

Mehrotra KM, Chilukuri K.; Ranka, Sanjay., 1997. Elements of Artificial Neural Networks. Complex Adaptive Systems Series, The MIT Press, p.

Michalski R, Stepp RE, Diday E, 1983. Automated construction of classifications:conceptual clustering versus numerical taxonomy. IEEE Trans.Pattern Anal Mach. Intell.

Mitchell T, 1997. Machine Learning, ABD, McGraw-Hill, p.

Munakata T, 2008. Fundamentals of the New Artificial Intelligence: Neural,Evolutionary, Fuzzy and More, London, Springer, p.

Narendra PM, Fukunaga K, 1977. A branch and bound algorithm for feature subset selection. IEEE Trans. on Computers, 917–22.

Nolde K, Morari M, 2010. Modeling and Control of Thermal Printing. Ieee T Contr Syst T, 18, 2, 405-13.

Osowski S, Siwekand K, Markiewicz T, (2004). MLP and SVM Networks – a Comparative Study. Proceedings of the 6th Nordic Signal Processing

Ozcift A, 2012. SVM Feature Selection Based Rotation Forest Ensemble Classifiers to Improve Computer-Aided Diagnosis of Parkinson Disease. J Med Syst, 36, 4, 2141-7.

Ozsen S, Gunes S, 2008. Effect of feature-type in selecting distance measure for an artificial immune system as a pattern recognizer. Digit Signal Process, 18, 4, 635-45.

Ozsen S, Gunes S, Kara S, Latifoglu F, 2009. Use of Kernel Functions in Artificial Immune Systems for the Nonlinear Classification Problems. Ieee T Inf Technol B, 13, 4, 621-8.

Özkan Y, 2008. Veri Madenciliği Yöntemleri, İstanbul, Papatya Yayınları, p.

Özşahin M, 2009. Türkiye Otomotiv Sektörünün Kendini Örgütleyen Haritalar İle Finansal Analizi, Çukurova Ünv, Adana.

Öztemel E, 2006. Yapay Sinir Ağları, İstanbul, Papatya Yayıncılık, p.

Patil BM, Joshi RC, Toshniwal D, 2010. Hybrid prediction model for Type-2 diabetic patients. Expert Syst Appl, 37, 12, 8102-8.

Patil BM, Joshi RC, Toshniwal D, 2010. Impact of K-Means on the Performance of Classifiers for Labeled Data. Comm Com Inf Sc, 94, 423-34.

Pennington KS, Crooks W, 1985. Resistive Ribbon Thermal Transfer Printing - a Historical Review and Introduction to a New Printing Technology. Ibm J Res Dev, 29, 5, 449-58.

Platt J, 1998. Sequential Minimal Optimization:A Fast Algorithm for Training Support Vector Machines, Microsoft Research, p.

Polat K, Guenes S, 2009. A new feature selection method on classification of medical datasets: Kernel F-score feature selection. Expert Syst Appl, 36, 7, 10367-73.

Polat K, Gunes S, 2006. Hepatitis disease diagnosis using a new hybrid system based on feature selection (FS) and artificial immune recognition system with fuzzy resource allocation. Digit Signal Process, 16, 6, 889-901.

Polat K, Gunes S, 2007. Breast cancer diagnosis using least square support vector machine. Digit Signal Process, 17, 4, 694-701.

Polat K, Gunes S, 2007. An expert system approach based on principal component analysis and adaptive neuro-fuzzy inference system to diagnosis of diabetes disease. Digit Signal Process, 17, 4, 702-10.

Polat K, Gunes S, 2007. Medical decision support system based on artificial immune recognition immune system (AIRS), fuzzy weighted pre-processing and feature selection. Expert Syst Appl, 33, 2, 484-90.

Polat K, Gunes S, 2007. Prediction of hepatitis disease based on principal component analysis and artificial immune recognition system. Appl Math Comput, 189, 2, 1282-91.

Polat K, Gunes S, Arslan A, 2008. A cascade learning system for classification of diabetes disease: Generalized discriminant analysis and least square support vector machine. Expert Syst Appl, 34, 1, 482-7.

Pudil P, Novovicova J, Kittler J, 1994. Floating search methods in feature selection. Pattern Recogn Lett, 15, 1119–25.

Quinlan JR, 1986. The effect of noise on concept learning. In Machine Learning: In An Artificial Intelligence Aproach, SanMateo, Morgan Kauffmann Inc, p. 149-166.

Quinlan JR, 1986. Induction of decision tree, p. 81-106.

Reed RD, Marks RJ, 1999. Neural Smithing: Supervised Learning, The MIT Press, p.

Rumelhart DEH, E.Geoffrey ; Williams, Ronald J., 1986. Learning Representations by Back-propagating Errors, Nature,, p.

Sagıroğlu Ş, Beşdok E, Erler M, 2004. Mühendislikte Yapay Zeka uygulamaları I :Yapay Sinir ağları, Kayseri, Ufuk Kitap Kırtasiye, p.

Sahan S, Polat K, Kodaz H, Gunes S, 2005. The medical applications of attribute weighted artificial immune system (AWAIS): Diagnosis of Heart and Diabetes Diseases. Artificial Immune Systems, Proceedings, 3627, 456-68.

Saracoglu R, 2012. Hidden Markov model-based classification of heart valve disease with PCA for dimension reduction. Eng Appl Artif Intel, 25, 7, 1523-8.

Saraç T, 2004. Yapay Sinir Ağları, Gazi Üniversitesi, Ankara.

Sartakhti JS, Zangooei MH, Mozafari K, 2012. Hepatitis disease diagnosis using a novel hybrid method based on support vector machine and simulated annealing (SVM-SA). Comput Meth Prog Bio, 108, 2, 570-9.

Schölkopf B, Smola AJ, 2002. Learning with Kernels, MIT Press, p.

Sever H, Oğuz B, 2002. Veritabanlarında Bilgi Keşfine Formal Bir Yaklaşım, Bilgi Dünyası, p.

Sevinç İ, 2009. Küme Merkezlerinin Belirlenmesinde Yeni Bir Yöntem Kocaeli Üniversitesi.

Siedlecki W, Sklansky J, 1989. A note on genetic algorithms for large-scale feature selection. Pattern Recogn Lett, 335–47.

Soman KP, Loganathan R, Ajay V, 2011. Machine learning with SVM and other kernel methods, PHI Learning Pvt. Ltd, p.

Stearns SD, (1976). On selecting features for pattern classifiers. 3rd lnternational Conference on Pattern Recognition. Coronado**:** 71-5.

Stehman SV, 1997. Selecting and interpreting measures of thematic classification accuracy. Remote Sens Environ, 62, 1, 77-89.

Sub bulakshmi CV, Deepa SN, Malathi N, (2012). Extreme Learning Machine for Two Category Data Classification. IEEE International Conference on Advanced Communication Control and Computing Technologies (ICACCCT), IEEE.

Suykens JAK, 2002. Least squares support vector machines, NJ : World Scientific, p. 294.

Tan PN, Steinbach M, Kumar V, 2006. Inroduction to Data Mining, Pearson Education, p.

Tang W, Khoshgoftaar TM, 2004. Noise identification with the k-means algorithm. Proc Int C Tools Art, 373-8.

Tarassenko L, 1998. A Guide to Neural Computing Applications, London: Arnold, p.

Tatlıdil H, 1996. Uygulamalı Çok Değişkenli Analiz, Ankara, p. 329-343.

Tebelkıs J, 1995. Speech Recognition Using Neural Networks, School of Computer Science Carnegie Mellon University, Pennsylvania.

Theodoridis S, Koutroumbas K, 2003. Pattern recognition, USA, Academic Press, p.

Tian J, Li MQ, Chen FZ, 2009. A hybrid classification algorithm based on coevolutionary EBFNN and domain covering method. Neural Comput Appl, 18, 3, 293-308.

Tokunaga Y, Yoshida M, Kizawa M, 1981. Life Estimation of Thermal Print Heads Consisting of Ta2n Thin-Film Resistors. Ieee T Compon Hybr, 4, 1, 148-53.

Turkoğlu İ, 2002. Durağan Olmayan ĠÇaretler Ġçin Zaman-Frekans Entropilerine Dayalı Akıllı Örüntü Tanıma, Fırat Üniversitesi, Elazığ.

2008. Erişim. Erişim adresi, http://www.tdk.org.tr.

Uyhan R, King-Hele JA, 2008. Modelling of thermal printers. Appl Math Model, 32, 4, 405-16.

Uzer MS, Inan O, Yilmaz N, 2013. A hybrid breast cancer detection system via neural network and feature selection based on SBS, SFS and PCA. Neural Comput Appl, 23, 3-4, 719-28.

Uzer MS, Yilmaz N, Inan O, 2013. Feature Selection Method Based on Artificial Bee Colony Algorithm and Support Vector Machines for Medical Datasets Classification. Sci World J.

Ünlükaplan Y, 2008. Çok Değişkenli İstatistiksel Yöntemlerin Peyzaj Ekolojisi Araştırmalarında Kullanımı,Yayınlanmamış, Çukurova Üniversitesi, Adana.

Van Gestel T, Suykens JAK, Lanckriet G, Lambrechts A, De Moor B, Vandewalle J, 2002. Bayesian framework for least-squares support vector machine classifiers,

Gaussian processes, and kernel Fisher discriminant analysis. Neural Comput, 14, 5, 1115-47.

Vanschoenwinkel B, Manderick B, 2004. Appropriate kernel functions for support vector machine learning with sequences of symbolic data. Deterministic and Statistical Methods in Machine Learning.

Vong CM, Wong PK, Li YP, Ho CM, 2005. Modelling of Modern Automotive Petrol Engine Performance Using Support Vector Machines. Journal of Zhejiang University SCIENCE, 6A(1).

Webb A, 2002. Statistical pattern recognition, England, John Wiley & Sons Ltd, p. 496.

WH. O, 2012. World Health Organization.

Whitney AW, 1971. A direct method of nonparametric measurement selection. IEEE Trans. Comput, 1100–3.

Widrow B, A. LM, 1990. 30 years of Adaptive Neural Networks: Perceptron,Madaline, and Backpropagation. IEEE, 78(9).

2015. Erişim. Erişim adresi, http://tr.wikipedia.org/wiki/Yapay_sinir_a%C4%9Flar%C4%B1.

Lecture Notes on the Kernel Trick, 2006. Erişim. Erişim adresi, http://www.che.iitm.ac.in/~naras/ch544/KernelTrick.pdf.

Xu Y, Zhu Q, Wang JH, 2012. Breast cancer diagnosis based on a kernel orthogonal transform. Neural Comput Appl, 21, 8, 1865-70.

Yalçın T, 2010. Temel Bileşenler Yardımıyla Gerçek Zamanlı Motor Koruma Algoritması, Ondokuz Mayıs Üniversitesi.

Yang J, Honavar V, 1998. Feature subset selection using a genetic algorithm. IEEE Intelligent Systems, 44–9.

Yang TC, Lai H, 2006. Comparison of product bundling strategies on different online shopping behaviors. Electron Commer R A, 5, 4, 295-304.

Yeh WC, Chang WW, Chung YY, 2009. A new hybrid approach for mining breast cancer pattern using discrete particle swarm optimization and statistical method. Expert Syst Appl, 36, 4, 8204-11.

Yilmaz N, Erdogan K, Boyali ON, (2012). Determining Damaged Regions that Occur on the Printhead of High Speed Thermal Transfer Printers. 3rd Int. Conf. on Computer and Computational Intelligence-ICCCI 2012. Bali, Indonesia, on proceedings of the 5th Int. Conf. on Computer Research and Development Asme Press.

Yilmaz N, Inan O, Uzer MS, 2014. A New Data Preparation Method Based on Clustering Algorithms for Diagnosis Systems of Heart and Diabetes Diseases. J Med Syst, 38, 5.

Zaman S, Karray F, 2009. Features Selection using Fuzzy ESVDF for Data Dimensionality Reduction. 2009 International Conference on Computer Engineering and Technology, Vol I, Proceedings, 81-7.

Zhang B, Li SS, Wu CS, Gao LR, Zhang WJ, Peng M, 2013. A neighbourhood-constrained k-means approach to classify very high spatial resolution hyperspectral imagery. Remote Sens Lett, 4, 2, 161-70.

Zurada JM, 1992. Introduction to Artificial Neural Systems, West Group, p.

ÖZGEÇMİŞ

KİŞİSEL BİLGİLER

Adı Soyadı : Onur İNAN
Uyruğu : T.C.
Doğum Yeri ve Tarihi : İzmir 18.02.1977
Telefon : 05425477779
Faks : 03322410064
e-mail : oinan@selcuk.edu.tr

EĞİTİM

Derece		Adı, İlçe, İl	Bitirme Yılı
Lise	:	**Selma Yiğitalp Lisesi,Merkez,İzmir**	**1994**
Üniversite	:	**Selçuk Üniversitesi,Selçuklu,Konya**	**1999**
Yüksek Lisans	:	**Selçuk Üniversitesi,Selçuklu,Konya**	**2003**
Doktora	:		

İŞ DENEYİMLERİ

Yıl	Kurum	Görevi
2000	**Selçuk Üniversitesi**	**Uzman**

UZMANLIK ALANI
Uzman sistemler

YABANCI DİLLER
İngilizce

BELİRTMEK İSTEĞİNİZ DİĞER ÖZELLİKLER

Uluslararası hakemli dergilerde yayımlanan makaleler :

(SSCI,SCI,SCI-EXPANDED,AHCI)

YILMAZ NİHAT,İNAN ONUR,UZER MUSTAFA SERTER (2014). A New Data Preparation Method Based on Clustering Algorithms for Diagnosis Systems of Heart and Diabetes Diseases. Journal of Medical Systems, 38(5) (Tezden Üretilmiştir.)

Uzer M S, Inan O, Yilmaz N (2013). Hybrid breast cancer detection tem via neural network and feature ion based on SBS, SFS and PCA. NEURAL COMPUTING & APPLICATIONS, 23(34), 719-728.

İnan O, Uzer M S, Yılmaz N (2013). A New Hybrid Feature Selection Method Based on Association Rules and PCA For Detection of Breast Cancer. International Journal of Innovative Computing, Information and Control, 9(2), 1-15. (Tezden Üretilmiştir.)

Uzer M S, Yilmaz N, Inan O (2013). Feature Selection Method Based on Artificial Bee Colony Algorithm and Support Vector Machines for Medical Datasets Classification. Scientific World Journal

Uluslararası bilimsel toplantılarda sunulan ve bildiri kitabında (Proceedings) basılan bildiriler :

Tam metin bildiri

İnan. O, İşcan. H, Yıldız. F (2013). The Analysis of the impact osf socio-Economic status of students with decision tree and cluster Algorithms. Advanced Computer Systems and Networks: Desing and Application, 211-215.

İNAN. O, YILMAZ. N, UZER. M S (2013). A New Data Elimination Method Based on Custering Algorithms for Diagnosis of Diabetes Diseases. 4th WORLD CONFERENCE on INFORMATION. (Tezden Üretilmiştir.)

Uzer. M S, Yılmaz. N, İnan. O (2013). A Case Study: Effect of ABC-based Feature election Algorithms on Breast Cancer Diagnosis. 4th World Conference on Information Technology