T.C

SELÇUK ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

GENETİK ALGORİTMALAR İLE TIBBİ VERİ MADENCİLİĞİ

EMİNE TUĞ

YÜKSEK LİSANS TEZİ

BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

Konya, 2005

T.C.

SELÇUK ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

GENETİK ALGORİTMALAR İLE TIBBİ VERİ MADENCİLİĞİ

EMİNE TUĞ

YÜKSEK LİSANS TEZİ

BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

Bu tez 20/12/2005 tarihinde aşağıdaki jüri tarafından oybirliği / oyçokluğu ile kabul edilmiştir

Prof.DrAhmet ARSLAN (Danışman)	Doç.Dr. Şirzad KAHRAMANLI (Üye)	Doç.Dr.Saadettin HERDEM (Üye)

ÖZET

Yüksek Lisans Tezi

GENETİK ALGORİTMALAR İLE TIBBİ VERİ MADENCİLİĞİ

EMİNE TUĞ

Selçuk Üniversitesi Fen Bilimleri Enstitüsü

Bilgisayar Mühendisliği Anabilim Dalı

Danışman : Prof. Dr. Ahmet ARSLAN

2005, 74 Sayfa

Jüri : Prof..Dr.Ahmet ARSLAN

Doç. Dr. Şirzad KAHRAMANLI

Doç.Dr.Saadettin HERDEM

Hazırlanan bu tezde, öncelikle kısaca Genetik Algoritma yöntemi açıklanmıştır. Genetik Algoritma evrim sürecini modelleyen ve en iyi çözümü bulmayı hedefleyen bir makine öğrenmesi tekniğidir. Daha sonra veri madenciliğinden bahsedilmiştir. Veri madenciliği depolanan verilerden belli teknikler kullanarak faydalı bilgiler elde etme işlemidir. Çalışmada Doğal Bağışıklık Sisteminden esinlenerek oluşturulmuş bir metot olan Yapay Bağışıklık Sistemi önerilmiştir. Burada SAND olarak bilinen bir yaklaşım sunulmuştur. Bu yaklaşım hatayı minimize eden bir arama ile veri kümesinde en iyi başlangıç populsyonunu üretmeyi hedefler. Ayrıca daha hassas sonuçlar üretmek için bir yöntem olarak Bulanık Mantıktan yararlanılmıştır. Bu çalışmada genetik algoritma kullanılarak bir veri madenciliği uygulaması geliştirilmiştir. Bu uygulamada tıbbi veriler kullanılmıştır. Kanser hastalarına ait veriler içinde hastalığa sebep olan etken grupları araştırılmaktadır. Amaç sık karşılaşılan hastalık sebeplerini bulmaktır.

Anahtar Kelimeler : Veri Madenciliği, Genetik Algoritma, Bulanık Mantık, Yapay Bağışıklık Sistemleri Sınıflandırma

ABSTRACT

Master of Science Thesis

MEDICAL DATA MINING via GENETIC ALGORITHMS

EMİNE TUĞ

Selçuk University

Graduate School of Natural and Applied Sciences

Department of Computer Engineering

Supervisor : Prof. Dr. Ahmet ARSLAN

2005, 74 Pages

Jury : Prof.Dr.Ahmet ARSLAN

Assoc.Prof.Dr. Şirzat KAHRAMANLI

Assoc.Prof.Dr. Saadettin HERDEM

In this thesis, at first Genetic Algorithms explained. Genetic Algorithm is a machine learning tecnique that aims to modelling the evolution and tries to find out the best solution. And then this thesis is about Data Mining. Data mining is an operation that uses some techinques to verify inferences from stored data. Artificial Immune System which is generated by inspiration of Natural İmmune System was suggested in this study. In this thesis, approach that is known as SAND was offered , that aims at generating a dedicated initialize a population of ahromosomes that best covers the search space, to be searched in order to minimize the error surface..Additionally to generate more delicate results Fuzzy Logic was used as a technique. In this study by using genetic algorithm we have developed data mining application. In the application medical data was used. we serarcehed effective groups that causes cancer on data about cancer people. Object is to find mostly encountered cancer causes.

Key Words : Data Mining, Genetic Algorithm, Fuzzy Logic, Artificial Immune Systems, Classification

İÇİNDEKİLER

ÖZET……..……..i

ABSTRACT……..……..iii

1. GİRİŞ 3
2. GENETİK ALGORİTMALAR 4
   2.2. Genetik Algoritmaların Temelleri 4
   2.3. Genetik Algoritma Terminolojisi 9
   2.4. Bireylerin Kodlanması 9
   2.5. Genetik Operatörler 10
      2.5.1. Seçim 11
         2.5.1.2. Kararlı Hal (Steady-State)Yöntemi 12
         2.5.1.3. Elitizm 12
      2.5.2. Çaprazlama 12
      2.5.3. Mutasyon 14
   2.6. Uygunluk Fonksiyonu 14
   2.7. Şema Teorisi 15
   2.8. Genetik Algoritmaların Performansını Etkileyen Faktörler 16
   2.9. Genetik Algoritmaların Uygulama Alanları 17
3. VERİ MADENCİLİĞİ 19
   3.1. Veri Madenciliğinde Aramanın Önemi 21
   3.2. Veri Madenciliğinde Arama İçin Geçerli Yaklaşımlar 21
   3.3. Veritabanlarında Bilgi Keşfi Süreci 24
   3.4. Veri Madenciliği Modelleri 26
      3.4.1. Birliktelik Kuralları 26
      3.4.2. Ardışıl Örüntülerin Madenciliği 28
      3.4.3. Sınıflandırma 29
      3.4.4. Kümeleme 30
   3.5. Veri Madenciliği Uygulama Alanları 31
4. YAPAY BAĞIŞIKLIK SİSTEMLERİ 35
   4.1 İmmünolojik Yaklaşım (Doğal Bağışıklık Sistemi) 36
   4.2 Yapay Bağışıklık Sistemi (Artificial Immune System) 40
   4.3 Bağışıklık Mühendisliği 41
   4.4 Yapay Bağışıklık Sistemi Uygulamaları ve Örnek Algoritmalar 41
      4.4.1 Takviyeli Öğrenme (Reinforcement Learning) 41
      4.4.2 Örüntü Tanıma (Pattern Recognition) 43
      4.4.3 Yapay Bağışıklık Ağ Teorisi (Artificial İmmune Network-aiNet) 44
      4.4.4 Klonal Seçim Algoritması (Clonal Selection Algorithm) 47
      4.4.5 Simulated Annealing Algoritması (SAND) 49
5. BULANIK MANTIK 53
   5.1. Üyelik Fonksiyonları ve Fonksiyon Etiketleri (Dilsel Değerler) 54
   5.2. Bulanık Mantık Kavramları 55
6. ÖRNEK ÇALIŞMA 57
   6.1. Veri Kümesi ve Veritabanı 58
   6.2. Genetik Algoritma 62
      6.2.1. Başlangıç Populasyonu 64

6.2.2. Uygunluk Fonksiyonu 65
6.2.3. Verilerin Bulanıklaştırılması 66
6.3. Uygulama Sonuçları 66
7. SONUÇ VE GELECEĞE YÖNELİK ÇALIŞMALAR 69
REFERANSLAR 71

1. GİRİŞ

Elektronik bir devrimin yaşandığı günümüzde her alanda potansiyel olarak depolanan veri hacmi hızla artmaktadır. Bununla birlikte bu verilerden faydalı bilgiler elde etmek giderek zorlaşmaktadır. Ancak bu konu göz ardı edilemeyecek kadar önemlidir. Özellikle sağlık sektöründe gizli bilgilerin keşfedilmesi işlemi hayati önem taşımaktadır. Son zamanlarda tıbbi veritabanlarında depolanan verilerin hızla artması ile veri madenciliği özellikle karar alma ve yönetim aşamalarında "bilgi" elde etmek için stratejik bir yol olarak benimsenmiştir.

Çok büyük miktarda veri toplayan işletmeler ellerindeki veriden geleceğe yönelik tahminler yapmanın önemini kavramışlardır. Bu bağlamda 1990'lı yıllara kadar daha çok , verinin toplanması ve depolanması ile ilgilenilmiş 1990'lı yıllardan itibaren ise veri ambarlarının kullanımının yaygınlaşması ve verinin analizi ile birlikte veri madenciliği ön plana çıkmıştır.

Veri madenciliği bir çok disiplindeki teknikleri kullanır. Makine öğrenmesi, Yapay Zeka, İstatistik vs. alanlarıyla beraber çalışarak ilginç ve yararlı bilgi elde etmeyi amaçlar. Bu çalışmada Genetik Algoritmaların veri madenciliği üzerinde uygulaması anlatılmıştır. Çalışma için T.C. Sağlık Bakanlığı Kanser Savaş Dairesi Başkanlığı tarafından toplanan kanser hastalarına ait veriler kullanılmıştır.

Genetik Algoritmalar (GA) öğrenme işleminde evrime dayalı bir yaklaşım kullanan Evrimsel Hesaplamaların (Evolutionary Computing) bir bölümüdür. Evrimsel Hesaplamalar, Yapay Zekanın (YZ) hızla gelişen bir alt dalıdır.

GA doğadaki evrim sürecinin modellenmesini kullanarak problem uzayında çok yönlü ve adaptif bir arama ile en iyi çözümü bulmaktadır. Evrim, amacı belli olan ya da kontrollü bir süreç değildir. Kaynakların kısıtlı olduğu değişken bir ortamda, farklı genetik bilgilere sahip canlılar yarışırlar ve kabaca güçlü olanlar hayatta kalıp çoğalarak/üreyerek genetik bilgilerini bir sonraki kuşağa geçirme şansını yakalarlar. Zayıf olanlar ise elenirler.

2. GENETİK ALGORİTMALAR

2.1. Genetik Algoritmaların Tarihçesi

Genetik Algoritmalar (GA) öğrenme işleminde evrime dayalı bir yaklaşım kullanan Evrimsel Hesaplamaların (Evolutionary Computing) bir bölümüdür. Evrimsel Hesaplamalar, Yapay Zekanın (YZ) hızla gelişen bir alt dalıdır ve I. Rechenberg'in "Evrim Stratejileri (Evolution Strategies)" adlı çalışmasıyla 1960'lı yıllarda ortaya çıkmıştır. Bu görüş araştırmacılar tarafından geliştirilmiştir. Amaç en iyi olanın hayatta kalması ilkesine dayanan evrim sürecinin modellenerek gerçek dünya problemlerine uygulanmasını sağlamaktır. "Genetik Algoritmalar" terimi bilimsel literatürde ilk defa J. D. Bagley'in bir çalışmasında ortaya atılmıştır (Bagley 1967). Daha sonra 1975'te yayınlanan J. Holland'ın "Doğal ve Yapay Sistemlerde Uyum (Adaptation in Natural and Artificial Systems)" kitabında GA'ın teorik esasları gösterilmiştir (Holland 1975). Michigan Üniversitesinde Makine Öğrenmesi (Machine Learning) üzerinde çalışan J. Holland , Darwin'in evrim kuramından etkilenerek geliştirdiği GA yöntemi ile ilgili teorik bilgileri bu kitapta toplamıştır. Deneysel çalışmalar ve uygulamalar öğrencileri ve çalışma arkadaşları tarafından geliştirilmiştir. J. Holland'ın öğrencisi olan D. E. Goldberg, çok sayıda kollara ayrılan gaz borularındaki gaz akışını düzenlemek için GA yöntemini kullanmıştır (Goldberg 1983). Böylece GA'ın uygulanabilirliği de göstermiştir. Son 20 yılda GA, özellikle optimizasyon ve makine öğrenmesi alanlarında oldukça yaygın bir şekilde kullanılmıştır.

2.2. Genetik Algoritmaların Temelleri

GA doğadaki evrim sürecinin modellenmesini kullanarak problem uzayında çok yönlü ve adaptif bir arama ile en iyi çözümü bulmaktadır. Evrim, amacı belli olan ya da kontrollü bir süreç değildir. Kaynakların kısıtlı olduğu değişken bir ortamda, farklı genetik bilgilere sahip canlılar yarışırlar ve kabaca güçlü olanlar hayatta kalıp çoğalarak/üreyerek genetik bilgilerini bir sonraki kuşağa geçirme şansını yakalarlar. Zayıf olanlar ise elenirler. Evrimin temeli farklılaşmadan oluşan

çeşitliliktir. Çeşitlilik ise üreme ve mutasyon ile sağlanır. Bu üreme iki şekilde olur: Basit canlılarda eşeysiz üreme gerçekleşir yani yeni oluşan canlı bir öncekinin aynısıdır ve farklılık üreme sırasında genlerdeki istemsiz hatalardan kaynaklanır. Bu olaya *mutasyon* denir. Gelişmiş canlılarda ise eşeyli üreme gerçekleşir yani değişik genetik bilgiye sahip iki canlı birleşerek tamamen farklı yeni bir canlı oluşturur. Farklılık, çoğalma sırasında hem mutasyonlar hem de yeniden birleşimler *(rekombinasyon)* sonucu oluşur. Her iki canlının genetik olarak farklı ama işlevsel olarak eş kromozomları üst üste gelerek bilgi değiş tokuşu yaparlar. Bu işlem çaprazlama olarak ifade edilir. Her iki durumda da oluşan farklı canlılardan bir kısmı, bir önceki kuşağa göre değişen çevreye daha iyi uyum gösterip (*adaptasyon*) daha güçlü olabilir. Bunların yaşaması ve üreyerek soyunu devam ettirmesine ise *doğal seleksiyon* denir.

GA sistemin matematiksel model bilgisine gerek duymaz. Karmaşık matematiksel hesaplamalar yerine yalnızca giriş -çıkış bilgilerine ihtiyaç duyar. Bu avantajlardan dolayı karmaşık problemlerin optimizasyonu için tercih edilebilir bir yöntemdir. GA büyük parametre ve yapısal belirsizlikleri tolere edebilir. Özellikle parametre değerleri önceden bilinmeyen kontrol sistemlerinde de çok iyi sonuçlar verebilmektedir. Ayrıca uygun performans indeksi seçimi ile sistemin arzu edilen dinamik davranışları ve kararlılığı elde edilebilir. Bu şekilde, performans indekslerinin seçimi, sistemin istenilen dinamik davranışların elde edilmesi için çok önemlidir (Zua 1995, Alli ve Kaya 2001).

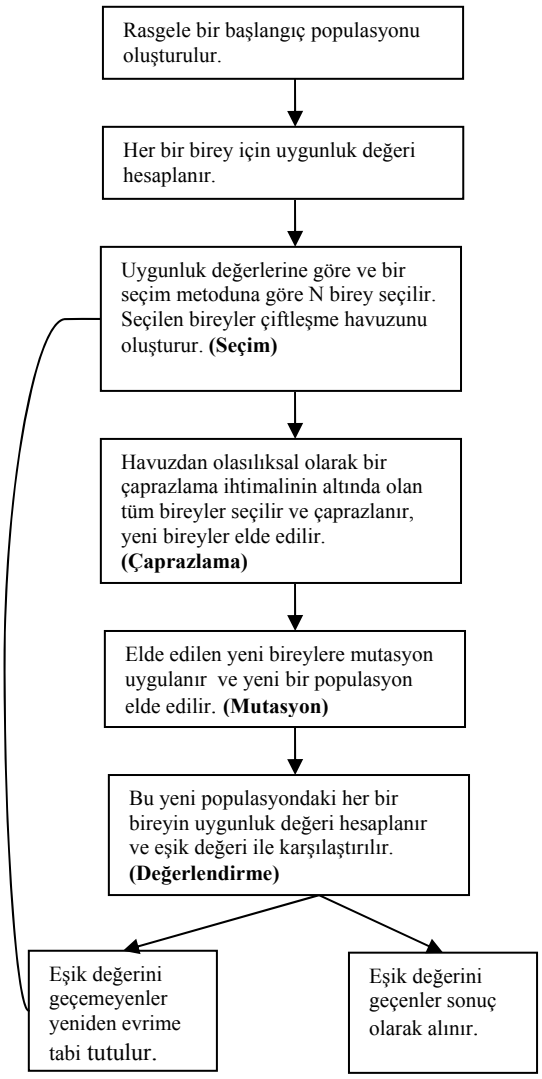
GA'yı diğer metotlardan ayıran ve avantaj sağlayan belli başlı farklar aşağıdaki gibi sıralanabilir :

- GA, sadece bir arama noktası değil bir çok arama noktası üzerinde çalışır. Arama uzayında yerel değil global bir arama yapar.
- GA'da işlemler olasılıksal kurallara göre yapılır. Bireyler yaşama ihtimallerine göre algoritma içinde yer alırlar. Programın üreteceği sonuçlar önceden kesin olarak tahmin edilemez.

- GA, arama uzayında bireylerin uygunlukluk değerlerini hesaplamak için bir uygunluk fonksiyonu kullanır. Sistemin matematiksel modeliyle ilgilenmez , kompleks işlemler yapmayı gerektirmez. Türevsel bilgilere gerek duymaz.
- Sürekli ve ayrık parametreleri optimize edebilir.
- GA, adaptif bir yöntem kullanır. Her çalışma sonunda kendini iyi yönde geliştirir.
- Çok sayıda parametre ile çalışma imkanı vardır.
- Paralel PC'ler kullanılarak çalıştırılabilir. Bu şekilde daha iyi sonuçlar elde edilir.
- Karmaşık uygunluk fonksiyon parametrelerini, lokal minimuma veya maksimuma takılmadan optimize edebilir.
- Sadece tek çözüm değil birden fazla parametrelerin optimum çözümlerini elde edebilir.

GA problem uzayındaki aday çözümlerin bir populasyon içerinse bulunduğunu kabul eder. Her bir aday çözüm ise populasyondaki bir bireyi temsil eder. Birey her yeni nesilde kendi kendini geliştirerek bir öncekinden daha iyi hale gelecektir. Bu adaptasyon boyunca rekombinasyon, mutasyon ve çaprazlamalar yeni üretilecek bireylerin daha uyumlu olmasını sağlayacak ve çeşitliliği oluşturacak mekanizmalar olacaktır.

GA'da en önemli problemlerden biri bireylerin *kodlanmasıdır*. Çünkü bireyler bizim için aday çözümlerdir. İyi bir kodlama işlemi yapılmadığı takdirde verimli sonuçlar elde etmekte güçleşecektir. Bireyler kodlandıktan sonra üzerinde arama içleminin gerçekleştirileceği başlangıç populasyonu üretilir. Daha sonra bu populasyon üzerinde seçim, mutasyon ve çaprazlama işlemleri ile bireylerin adaptasyonu sağlanır. Son olarak ise bir değerlendirme işlemi ile en iyi bireyler tahmin edilmeye çalışılır. GA'ın blok diyagramı Şekil 2.1'de verilmiştir. Her bir adım ise ilerleyen bölümlerde açıklanacaktır. GA'ın temel çalışma algoritması ise Tablo 2.1'de gösterilmiştir.



Şekil 2.1 Genetik Algoritma Blok Diyagramı

Terimler :

B :	Bir aday çözümü temsil eden birey
Uygunluk :	Her bir bireye bir evrim puanı atayan fonksiyon
Uygunluk Eşiği :	Algoritmanın sonlandırma kriteri
P :	Populasyon
r :	Her adımda çaprazlama ile değişecek populasyondaki bireylerin oranı
m :	Mutasyon oranı
n:	Populasyondaki birey sayısı

Yöntem :

Başlangıç Populasyonu : Rasgele olarak **n** bireyden oluşan bir başlangıç populasyonu oluşturulur. **(P)**

Evrim : P Populasyonundaki her bir **B** bireyinin uygunluk değeri hesaplanır.

Uygunluk(B)

While [max $_n$ Uygunluk(B)] < Uygunluk Eşiği do

Yeni bir nesil oluştur : $\mathbf{P_s}$

Seçim : Olasılıksal olarak P'den (1-r)n eleman yeni P_S nesline eklenir. Herhangi bir B bireyinin seçilme olasılığı :

$$\mathbf{P_r\,(B) = Uygunluk\,(B) / \sum Uygunluk}$$

Çaprazlama : Olasılıksal olarak P'den (r-n)/2 çift seçilir. Bu seçim yukarıda verilen $P_r(B)$ olasılığına göre yapılır.Her çift için çaprazlama yöntemine göre çaprazlama işlemi uygulanarak iki yeni yavru birey üretilir. Tüm yavrular P_S'ye eklenir.

Mutasyon : P_S'nin m yüzdelik elemanı eşit olasılıkla seçilir. Her birisi için seçilmiş bir genin değeri değiştirilir.

Güncelleme : P_S'deki bireyler P ile değiştirilir.

Değerlendirme : P'deki her bir birey için uygunluk değerleri hesaplanır.

Sonuç : P'den en yüksek uygunluğa sahip birey çözüm olarak geri dönderilir

Tablo 2.1 Genetik Algoritmanın Çalışma Modeli

Şekil 2.1'de de görüldüğü gibi GA'ın temel aşamaları *seçim*, *çaprazlama* ve *mutasyondur*. Bu aşamalar GA'ın *operatörleri* olarak kabul edilir. Bu operatörlerin uygun performans indeksi seçimleri sayesinde sistem dinamik ve kararlı bir yapı kazanır.

2.3. Genetik Algoritma Terminolojisi

Bu bölümde GA'da kullanılan terminolojiden bahsedilecektir. GA, doğal genetikteki terimleri kendine göre uyarlayarak kullanmaktadır. Bu terimler ve onların doğal genetikdeki karşılıkları arasında da bir uyum vardır. Tablo 2.2. doğal genetik ve GA arasındaki terminoloji karşılaştırmasını göstermektedir (Goldberg 1989).

Doğal Genetik	**Genetik Algoritmalar**
kromozom	string
gen	nitelik, karakter veya detektör
allele	nitelik değeri
locus	string pozisyonu
Genotype	yapı
ghenotype	parametre kümesi, alternatif çözüm, çözümlenmiş yapı
Epistasis	doğrusal olmama

Tablo 2.2. Genetik Algoritma Terminolojisi

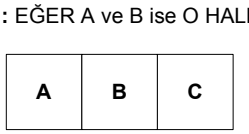
2.4. Bireylerin Kodlanması

GA'nın geleneksel metotlarla arasındaki temel farklardan biriside parametreleri kodlayarak çalışmasıdır. Bu sayede çok sayıda parametre ile çalışabilmektedir. Biyolojiden de bilindiği üzere genlerden her biri farklı bir karakteristiği ifade etmektedir. Kromozom ise genlerin oluşturduğu topluluktur ve birey hakkındaki bilgilere kromozoma bakarak karar veririz. Benzer şekilde GA'da genler bir parametreyi, parametrelerin toplu kümesi ise bir kromozomu temsil eder. Kromozom'a bakarak bir birey yani bir aday çözüm hakkında bilgi sahibi oluruz.

Kromozomların kodlama modeli GA'nın performansını etkileyen faktörlerden biridir. Kodlama için genelde iki yöntem kullanılır: Parametrelerin ikili(binary) sayılarla veya ondalık(decimal) sayılarla kodlanması yöntemleri. Algoritmadan iyi bir performans alabilmek için probleme uygun kodlama yöntemi seçilmelidir. İkili kodlama yönteminde her bir kromozomun kodu, bir genin alabileceği değer aralığına göre belirlenen bit sayılarına göre ve genin alacağı değerin 1 ve 0'lardan oluşan bit

stringleri şeklinde gösterilmesine göre oluşturulur. İkili kodlama yöntemi mutasyon ve çaprazlama gibi genetik operatörleri uygulamakta kolaylık sağlar. Ancak parametre sayısı ve parametrelerin değer aralığı arttığında kromozomların boyu oldukça büyüyecektir. Algoritma da ikili bir kodlama yöntemi tercih edilmiş olsa bile uygunluk fonksiyonu genlerin ondalık karşılıklarını kullanır. Dolayısıyla algoritmadan daha iyi bir performans sağlamak için çok sayıda parametre ile çalışan ve parametrelerin değer aralıklarının geniş olduğu problemlerde ondalık kodlama yöntemini kullanmak daha iyi bir performans sağlayacaktır.

EĞER – O HALDE kurallarının da kodlanması oldukça kolaydır. Her kuralın ön durumu (precondition) ve son durumu (postcondition) için ayrı ayrı genler tanımlamak suretiyle bir kural bir kromozom şeklinde kullanılabilir. Örnek bir kuralın kodlaması Şekil 2.2'de gösterildiği gibidir.



Şekil 2.2 Bir kuralı ifade eden kromozomun gösterimi

2.5. Genetik Operatörler

Genetik operatörler varolan populasyonda rekombinasyon ve mutasyon işlemleri ile yeni nesiller üretmeye çalışır. Amaç çeşitliliği artırmak ve daha iyi bireyler oluşturmaktır. Bu operatörler biyolojik evrimden bilinen genetik operasyonların bir yapay modelidir. GA'da en çok kullanılan operatörler *seçim (selection), çaprazlama (crossover)* ve *mutasyon (mutation)* operatörleridir.

2.5.1. Seçim

Biyolojik evrimde olduğu gibi GA'da da bireyler gelecek nesil içerisinde yer alma mücadelesi verirler. Bireyler uygunluk değerlerine göre algoritma içinde yer alırlar. Dolayısıyla seçim işlemi uygunluk fonksiyonuna da bağlıdır. Bu yüzden seçim için kullanılacak yöntem belirlenirken uygunluk fonksiyonu göz önünde tutulmalıdır. Seçim için kullanılan yöntemlerden bazıları alt bölümlerde verilmiştir.

2.5.1.1. Rulet Çarkı (Roulette Wheel) Yöntemi

Bu yöntemde populasyondaki her birey uygunluk oranı boyutunda rulet çarkında yer almaktadır. Yöntem çarkın N kez döndürülmesinden sonra N adet yeni birey seçmesi şeklinde modellenebilir. Uygunluğu yüksek olan bireyler çarkta daha fazla yer kaplayacağından seçilme ihtimalleri yüksek olacaktır. İyi bireylerin yeni nesildeki sayıca oranları da fazladır.Ayrıca bu yöntemle düşük uygunluğa sahip bireylerin elenme ihtimalleri de yüksek olacaktır. Formül 2.1 bir bireyin seçilme olasılığını aynı zamanda rulet çarkında kaplayacağı alanı gösterir.

$$P_{seçim}\,(\text{Birey}) = \text{Uygunluk}\,(\text{Birey}) \,/\, \sum \text{Uygunluk} \qquad \textbf{Formül 2.1}$$

Tablo 2.2 Rulet çarkı ile seçim metodununun algoritmasını göstermektedir.

[Sum] Populasyondaki tüm kromozomların uygunlukları toplamı S hesaplanır.
[Select] (0,S) aralığında rasgele bir r sayısı üretilir
[Loop] Populasyon boyunca uygunluklar toplamı hesaplanır ve toplam r'den büyük olduğunda bulunan birey seçilir.

Tablo 2.2 Rulet Çarkı Yöntemi

2.5.1.2. Kararlı Hal (Steady-State)Yöntemi

Bu yöntem bireylerin çoğunun bir sonraki nesildede olması gerektiği ilkesine dayanır. Uygunluğu düşük olan zayıf bireyler bu yöntemde yerlerini daha iyi yeni yavru bireylere bırakırlar. Kararlı hal yönteminde öncelikle populasyondaki en iyi bireyler uygunluk değerlerine göre seçilir, bu bireylerden yeni yavrular elde edilir ve bu yavrular uygunluğu düşük zayıf bireylerle değiştirilir.

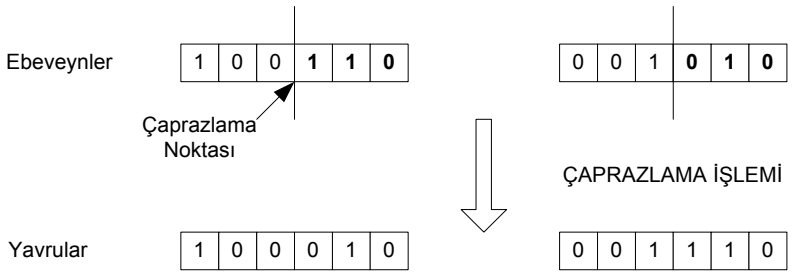
2.5.1.3. Elitizm

Elitizm yöntemine kararlı hal yöntemi gibi bir çeşit yer değiştirme metodu uygular. Bu yöntemde elde edilen iyi bireylerinde kaybolması önlenir. Ayrıca bu yöntem seçim metodu yanında artı bir operatör gibide kullanılabilir. Yöntem uygunluk değerleri çok yüksek olan bireylerin populasyondaki zayıf bireylerle yer değiştirmesi şeklinde çalışır. Bu doğal genetikte bir populasyona göçmenlerin dahil olması gibide yorumlanabilir. Elitizm yöntemi algoritmanın kısa sürede kararlı duruma geçmesini sağlar.

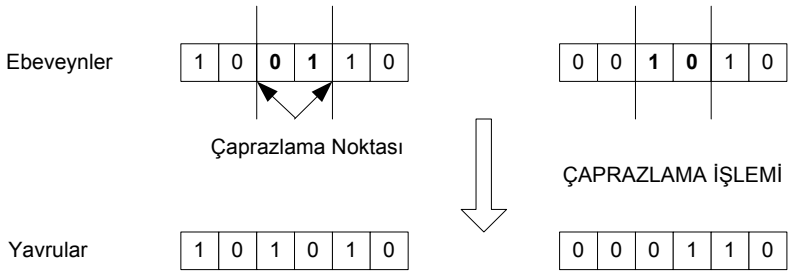
2.5.2. Çaprazlama

Çaprazlama operatörü her bir ebeveynin seçilen bitlerinin kopyalanması yoluyla , iki ebeveynden iki yeni yavru birey üretme şeklinde tanımlanabilir. Elde elden yeni yavrular ebeveynlerinden daha iyi olacaklardır yani daha yüksek uygunluk değerine sahip olacaklardır.

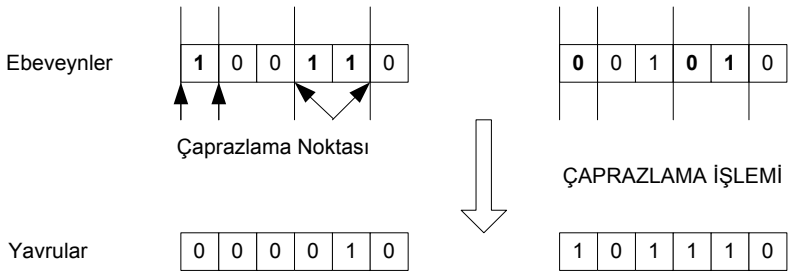
Çaprazlama operatörü seçilen çaprazlama yöntemine göre yapılmalıdır. Bazen seçilen yöntem uygunlukları iyileştirmeyip daha kötü bireyler elde etmeye bile sebep olabilir. Dolayısıyla probleme uygun çaprazlama yöntemi belirlenmelidir. Şekil 2.3 bazı çaprazlama yöntemlerini göstermektedir.



Şekil 2.3 (a) Tek noktalı çaprazlama işlemi



Şekil 2.3 (b) İki noktalı çaprazlama işlemi

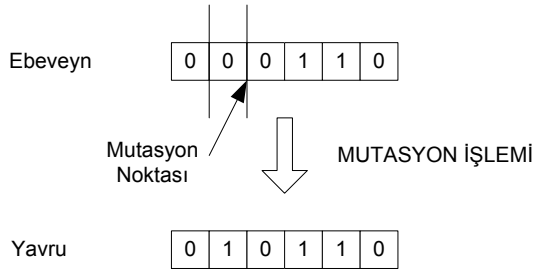


Şekil 2.3 (b) Uniform çaprazlama işlemi

Çaprazlama işlemi uygunluk değerine göre seçilmiş iki ebeveyn bireyden iyi özellikte yeni bireyler elde etmeye yarar. Belirlenen çaprazlama noktasına göre karşılıklı bireylerde genlerin yer değiştirmesiyle yavru bireyler elde edilir. Çaprazlama işleminde çaprazlama noktası rasgele olarak seçilir. Tüm işlem boyunca rasgele belirlenen bir çaprazlama noktası tüm bireylerde aynı olabileceği gibi her ikili birey içinde ayrı ayrı çaprazlama noktaları belirlenebilir.

2.5.3. Mutasyon

Mutasyon işlemi çaprazlama gibi populasyonda çeşitliliği sağlayan yöntemlerden biridir. Ancak bireylerde çaprazlamadaki kadar belirgin farklar ortaya çıkmaz. Değişiklikler daha küçüktür. Mutasyon işlemi belirlenen mutasyon oranına göre seçilen bireyin rasgele belirlenen geninin değer değiştirmesi şeklinde özetlenebilir. Şekil 4 mutasyon işleminin bir kromozom üzerinde uygulamasını göstermektedir. Mutasyon noktası da , çaprazlama noktası gibi rasgele belirlenir.



Şekil 2.4 Tek nokta mutasyon işlemi

2.6. Uygunluk Fonksiyonu

Başlangıç populasyonu bir kez oluşturulduktan sonra evrim başlar. GA bireylerin uygunluk değerlerini bularak iyiliklerini belirlemeye çalışır. Uygunluk, populasyondaki bir kısım bireyin problemi nasıl çözeceği için iyi bir

ölçüdür.Uygunluk ölçüsüne göre iyi olarak değerlendirilen bireyler tekrar üreme, çaprazlama ve mutasyon operatörleriyle seçilirler.

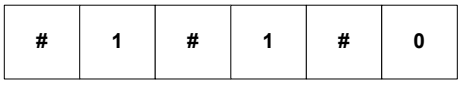
Bazı problemler için bireyin uygunluğu, bireyden elde edilen sonuç ile tahmin edilen sonuç arasındaki hatadan bulunabilir. Daha iyi bireylerde bu hata sıfıra yakın olacaktır.

Uygunluk fonksiyonu, her bir kromozomu değerlendirmek için iyi bir kaynaktır. Bu GA ve sistem arasında önemli bir bağlantıdır. Fonksiyon giriş olarak kodu çözülmüş şekilde kromozom (phenotype) alır ve kromozomun performansına ölçü olarak objektif bir değer üretir. Bu işlem diğer kromozomlar içinde yapıldıktan sonra bu değerler belirli bir düzende planlanırlar. Bu planlamayı sağlayan uygunluk tekniği probleme uygun olarak belirlenmelidir.

2.7. Şema Teorisi

GA'da üretilen çözüm kümesindeki bireyler incelenirse bireyler arasında bazı benzerlikler bulunabilir. Bu benzerliklerden yola çıkılarak şemalar oluşturulabilir. İkilik dizi kodlaması için aşağıdaki yöntem önerilebilir:

Genler 0,1 ve # ile kodlanır. # o konumda 0 veya 1 olmasının önemsiz olduğunu gösterir. Şekil 2.5 örnek bir şemayı göstermektedir. Şekil ikinci ve dördüncü bitleri 1, altıncı biti 0 olan çözümlerin başarılı olduğu bir populasyonda oluşturulan bir şemayı ifade eder.



Şekil 2.5 Örnek bir Şema

Bu şemaya uygun olarak aşağıdaki ikilik diziler yazılabilir:

010100, 010110,011100,011110,110100,110110,111100,111110

Görüldüğü gibi şemaların katılması ikilik dizilerle gösterilen arama aralığını büyütmektedir. Arama aralığının büyümesinin sonucun bulunmasını zorlaştırması beklenir ancak durum böyle değildir. Seçim ve yeniden kopyalama ile iyi özellikler daha çok bir araya gelerek daha iyi değerlere sahip şemalarla uygun çözümler elde edilir.

GA kendi içinde sanal olarak şemalar oluşturur. Populasyonun bireyleri incelenerek bu şemalar ortaya çıkarılabilir. GA şemaları oluşturmak için populasyon üyelerinin kodları dışında bir bilgi tutmaz. GA'ın bu özelliğine *içsel paralellik (implicit parallelism)* denir. Her nesilde, iyiyi belirleyen şemalardaki belirsiz yada önemsiz elemanlar azalır. Böylece GA sonuca doğru belli kalıplar içinde ilerler.

2.8. Genetik Algoritmaların Performansını Etkileyen Faktörler

Kromozom sayısı : Kromozom sayısını artırmak çalışma zamanını artırırken azaltmak da populasyondaki çeşitliliği azaltır.

Mutasyon oranı : Kromozomlar birbirine benzemeye başladığında hala çözüm noktalarının uzağında bulunuluyorsa mutasyon işlemi bu sorunu çözmek için en iyi yöntemdir. Ancak mutasyon oranı yüksek seçilecek olursa da GA kararlılığını kaybedecektir.

Çaprazlama yöntemi : En fazla kullanılan çaprazlama yöntemi tek noktalı çaprazlama olsa da yapılan araştırmalar bazı problemlerde farklı çaprazlama yöntemleri kullanmanın çok yararlı olduğunu göstermiştir.

Parametrelerin kodlanması : Nasıl bir kodlama yöntemi seçileceği GA'da önemli bir problemdir. Kimi zaman bir parametrenin doğrusal yada logaritmik kodlanması GA'ın performansında önemli farklar ortaya çıkarabilir.

Uygunluk fonksiyonunun belirlenmesi : Probleme uygun bir uygunluk fonksiyonu belirlenmediği takdirde GA'ın yaptığı değerlendirme işlemleri de yanlış

olacaktır. Ayrıca seçilen uygunluk fonksiyonuna göre yapılacak değerlendirme işlemi çalışma zamanını uzatacağı gibi çözüme hiçbir zaman ulaşılamamasına da sebep olabilir.

2.9. Genetik Algoritmaların Uygulama Alanları

GA bir çok alanda uygulamaları yapılmaktadır. En yaygın olarak kullanıldığı alanlar aşağıda belirtilmiştir. (Mitchell 1996)

Optimizasyon : GA sayısal ve kombinasyonel optimizasyon problemlerini içeren optimizasyon uygulamalarında kullanılabilir. GA en iyi yani en optimal çözümü bulma ilkesiyle kullanıldığı için optimizasyon en çok kullanılan alanlardan biri olarak kabul edilir. Devre tasarımı, doğrusal olmayan denklem sistemlerinin çözümü ve fabrika-üretim planlamalarında kullanılabilir.

Otomatik Programlama : GA spesifik uygulamalar için evrimsel bilgisayar programları geliştirmeye olanak sağlar. Belirli bir görev için tasarlanmış ve çevresinde aldığı bilgilere göre sürekli öğrenen etmenler (agent) geliştirmede kullanılabilir.

Hücresel Programlama (Cellular-Programming) : Doğadaki benzerleri gibi işleyen canlıların temel özellikleri olan uyum ve öğrenme becerilerini kullanan adeta yaşayan makinalar olarak tabir edebileceğimiz hücresel programlama (Cellular-Programming) için uygulanabilir.

Makine Öğrenmesi : GA hava tahmini veya protein yapılarının analizi gibi sınıflandırma ve tahmin işlemleri içeren bir çok makine öğrenmesi uygulamasında kullanılmaktadır. GA, yapay sinir ağları için ağırlıklar , öğrenen sınıflandırıcı sistemleri için kurallar gibi parçalı makine öğrenmesi sistemleri veya sembolik üretim sistemleri ve robotik sistemleri gibi alanlarda uygulamaları mevcuttur.

Ekonomi : GA pazarlama stratejileri belirleme, ekomomik modeller geliştirme, marketçilik gibi alanlarda kullanılır.

İmmün Sistemler : GA somatik hipermutasyon gibi yapıları içeren doğal immün sistemlere benzer modeller geliştirmede kullanılabilir. Doğal immün sistemler GA için uygun bir hayatta kalma mücadelesini çok iyi bir şekilde uygular.

Ekoloji : GA immün sistemler gibi biyolojik bir yarışı içeren ekolojide de kullanılmaktadır.

Populasyon Genetiği : GA genetikle ilgili ”Evrimsel olarak hangi koşullar alında bir gen hayatta kalabilir” gibi sorulara cevap vermek için kullanılabilir.

Evrim ve Öğrenme : GA bireyleri öğrenmesi ve türlerin evrimleştirilmesinde kullanılır.

Sosyal Sistemler : GA böcek kolonilerinin evrimi ve genel olarak çoklu-etmenli sistemlerde (multi-agent systems) iletişim ve çalışmanın evrimi gibi sosyal sistemlerin evrimleştirilmesi çalışmalarında kullanılır.

Biyo-Bilişim (Bio-Computing) : Genetik kodların analizi, DNA ve RNA kıvrılmalarının analizi ve tahmini uygulamalarında kullanılır.

3. VERİ MADENCİLİĞİ

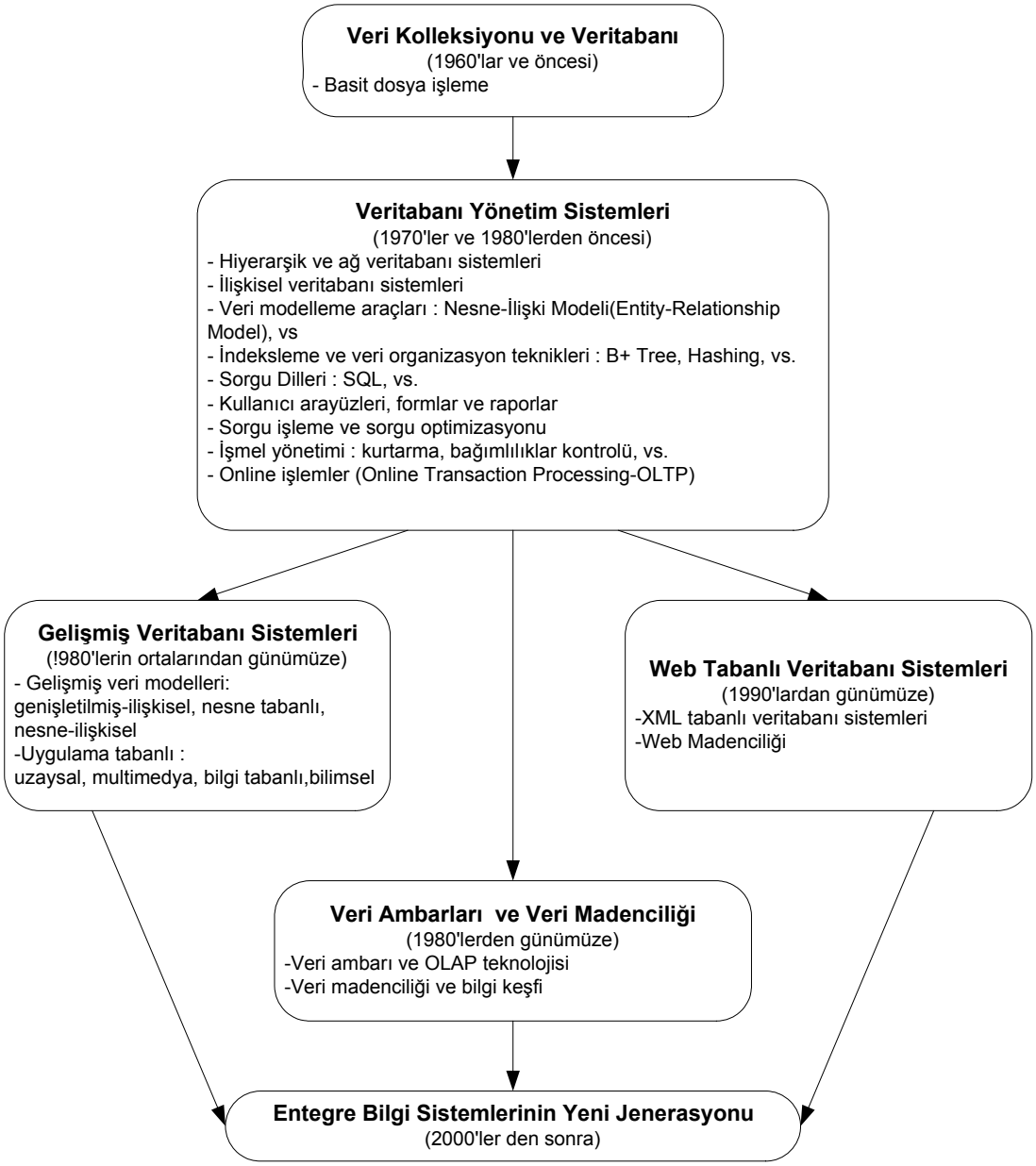
Günümüzde modern veri tabanları çok büyük miktarlarda veri içermektedir. Bununla birlikte işlenmemiş ham veriden fayda sağlamak oldukça zordur. Enformasyon ve bilgiye ulaşmak verinin iyi bir analizi ile olur. Bir takım enformasyonun araştırılmasında verinin analizini otomatikleştirmek için ise veri madenciliğine başvururuz. Veri madenciliği çeşitli araştırma sahalarından farklı teknik ve metodları içermektedir.

Yeni teknolojiler ve otomatikleşen sistemler tam verimli ve ucuz bir yolla verinin büyük miktarlarının toplanması ve depolanmasını sağlarlar. Bu şekilde her büyük organizasyon müşterilerinin, çalışanlarının, faaliyetlerinin verilerini oluşturur ve depolar. Elektronik verinin bu koleksiyonu veritabanlarında tutulur. Çok büyük koleksiyonlar için ise veri ambarları kullanılır. Veri tabanları ve veri ambarları sürekli olarak geliştirilir ve güncelleştirilir. Çünkü organizasyon hakkındaki enformasyon ve bilginin birçok çeşidini onların kodladığına inanılır. Bilgi gelecekte beklenen akımları tespit etmeye veya analizler yapmaya imkan sağlayacaktır. Bununla birlikte ham veri bazı işlemlerin dışında çok az faydalıdır.

"Veritabanında Bilgi Keşfi (Knowledge Discovery in Database-VTBK)" ifadesi verinin çok büyük bir koleksiyonundan enformasyon ve bilgi çıkarma teknikleri ve araçlarının geniş bir sahasını belirtir. Veriden daha önceden bilinmeyen ilginç ve yararlı bilgi elde etmeyi amaçlar. Bu amaçla Veri Madenciliği (Data Mining) teknikleri kullanılır. Veri ambarlarının büyüyen boyutlarında yüksek seviyeli otomatik işlemlerle bilgi arama insan analistler için kolaylaşmaktadır. Bu amaçla bir takım algoritmalar genel problemlerin çözümüne çalışırlar ve sürekli gelişirler.

Çok büyük miktarda veri toplayan işletmeler ellerindeki veriden geleceğe yönelik tahminler yapmanın önemini kavramışlardır. Bu bağlamda 1990'lı yıllara kadar daha çok , verinin toplanması ve depolanması ile ilgilenilmiş 1990'lı yıllardan itibaren ise veri ambarlarının kullanımının yaygınlaşması ve verinin analizi ile

birlikte veri madenciliği ön plana çıkmıştır. Şekil 3.1 veritabanı teknolojisinin evrimini özetlemektedir.



Şekil 3.1 Veritabanı Teknolojisinin Evrimi

Veri madenciliği ile büyük veri kümeleri üzerinde ilginç ve gizli kalmış örüntüleri tespit etmek mümkündür. Bu noktada veri madenciliği istatistikten ayrılır. İstatistikçiler bilinen faktörler arasındaki ilişkilerin güçlülüğünü araştırır. Veri madenciliğinde ise bilinmeyen faktörler arasında tahmin edilemeyecek ilişkilerin güçlülüğü araştırılır. Örneğin "Çocuk bezi alan müşterilerin %30'u birada satın alır" gibi bir ilişkiyi istatistik yaklaşımları ile bulmak zordur. Çünkü; "Çocuk bezi" ve "Bira" bizim için arasında ilişki olduğu düşünülecek ürünler değildir. Ancak veri madenciliği yaklaşımından baktığımızda "Çocuk bezi" ve "Bira" arasındaki ilişkiyi önceden tahmin edemediğimiz halde, bu teknikler keşfedilebilecek bir ilişkidir.

3.1. Veri Madenciliğinde Aramanın Önemi

Veri madenciliğinin önemli özelliklerinden biri verimli ve etkili bir arama mekanizması kullanmasıdır. Arama mekanizması önemlidir ve geniş veri kümeleri içerisinde potansiyel ilişkileri ortaya çıkarma prosesinde temel bir bileşendir.

Elektronik veri toplama yeteneği geliştikçe verilerin kullanma ve yorumlanma yeteneği azalır. Bu veritabanı analizi için geliştirilecek teknikler ve araçların yeni jenerasyonları için bir ihtiyaçtır. Veri madenciliği veri özü hakkında genelleştirme üretmek ve eğilimleri bulmak için örnekleri aramada büyük veri tabanları içinde eleme yapmak için metodlar kullanır.

Desenleri bulmak, eğilimleri tespit etmek veya veri özü hakkında genelleştirme yapmak için yoğun bir arama gerekir. Veri madenciliğinde arama alternatifler sınanarak bir çözüme konumlanma işlemidir.

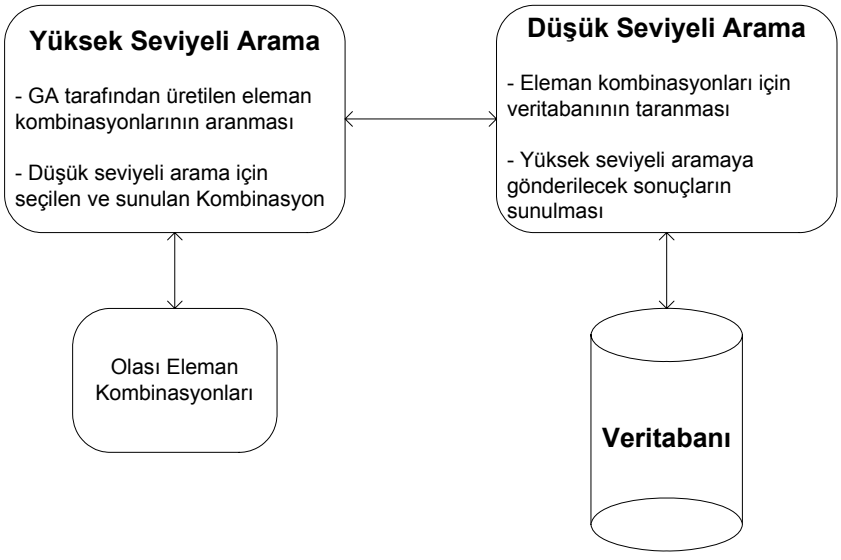
3.2. Veri Madenciliğinde Arama İçin Geçerli Yaklaşımlar

Veri madenciliği çerçevesinde arama mekanizmasını düşündüğümüz zaman arama tipleri arasında farklılıklar olmak zorundadır. Veritabanında "düşük seviyeli (low level)" arama ve "yüksek seviyeli (high level)" arama olarak temelde iki farklı arama yaklaşımı söz konusudur.

Veri madenciliğinde aramanın farklı seviyeleri bir veritabanında bilgiyi elde etme konusunda ayrılır. En düşük seviyede özel dosyalara yada kayıtlara erişilmesi gerektiği zaman erişim yapıları indeksler olarak adlandırılır. İndeksler belirli arama şartlarına yanıt vererek kayıtlara ulaşmayı hızlandırmak için kullanılır. Bir indeks bir anahtar terim ile birlikte sayfa numarasını veya anahtar terimin bulunduğu sayfa numaralarını içerir. Biz adres veya adreslerin listesini bulmak için indeksi arayabilir ve sonra veritabanında terimin yerleştiği adreslere ulaşabiliriz. Alternatif olarak ilgilenilen terimi bulmak için tüm veritabanı taranmalıdır.

Yüksek seviyeli bir arama, yukarıda açıklanmış olan daha düşük seviye arama mekanizmasını işleyen bir arama mekanizmasıdır. Yüksek seviyeli bir arama bir veritabanı içerisinde gerçek aramadan ziyade daha düşük seviyede aramanın ne olduğuna karar verir. Şekil 3.2 veri madenciliği çerçevesinde arama seviyelerinin nasıl olduğunu göstermektedir.

Veri madenciliği çerçevesinde yüksek seviyeli aramayı uygulamak için bir çok olası yaklaşım vardır. Yüksek seviyeli arama stratejileri tartışmasında kullanılan bazı tanımlamalar ve mevcut uygulama için popüler alternatifleri belirten çeşitli yüksek seviyeli arama stratejileri aşağıda belirtilmiştir.



Şekil 3.2 Veri madenciliğinde arama seviyeleri

Tanımlamalar :

Arama Uzayı : Bir arama uzayı durumların bir kümesini ve komşu durumların hareketlerinin kümesini içerir. Arama uzayı, arama boyunca ziyaret edilebilir durumları gösteren ve arama uzayında dolaşmak için rehberlik ederek hareket edebilir veri yapılarının önceden bulunmasını içerebilir.

Arama Stratejisi : Bir arama uzayını işlemek ve inşa etmek için genel bir yaklaşımdır. Örneğin, heuristik arama stratejisi, aranmamış arama uzayı parçasından ayrılmış bir ayrıntılı arama stratejisinden farklıdır.

Heuristik Arama : Bir heuristik arama toplam arama uzayının sadece bir parçasını işler veya oluşturur. Hill climbing ve beam arama aşağıda tanımlandığı gibi heuristik arama yaklaşımlarıdır. Verilen bu tanımlamaya göre bir genetik algoritma heuristik arama olarak düşünülmektedir.

Exhaustive (veya Enumarative) Arama : Optimal çözümü sağlamak için arama uzayında her duruma ulaşabilen bir aramadır. Bu yaklaşım veri madenciliği uygulamalarında olduğu gibi arama uzayı oldukça geniş olduğu zaman genellikle esnek değildir.

Veri Madenciliği Arama Stratejileri :

Hill Climbing : Aramanın her adımından sonra bir çözüm tutan bir heuristik arama stratejisidir. Her adımda geçerli durumun optimizasyon kriterine uygun en iyi komşusunu seçer. Eğer her komşu geçerli durumdan kötü ise arama durur.

Yayılarak (Beam) Arama : Hill climbing'e benzer bir heuristik arama stratejisidir. Her adımda en iyi N kısmi çözümleri optimizasyon kriterine uygun arama için tutulur.

Exhaustive (veya Enumerative) Arama : Yukarıda açıklanmıştır.

Hiyerarşik Ağaç Araması: Bir hiyerarşik seviyedeki olası alternetiflerin veri kümesi, sonraki daha düşük hiyerarşik seviyede alt kavramlar içerisine tekrarlı (recursive) olarak bölünmektedir. Hiyerarşik ağaç arama algoritması "HierarchyScan" tipik olarak bir ağaçta arama için kullanılan geleneksel sıralı tarama için bir alternatif önerir.

3.3. Veritabanlarında Bilgi Keşfi Süreci

Veri madenciliği büyük veri kümelerinden "bilgi çıkarma" işlemini yerine getirir. Bunu yaparken de belirli aşamalar takip edilir. Veritabanında bilgi keşfi sürecinin aşamalarını aşağıdaki gibi sıralayabiliriz.

Problemin Tanımlanması : Veri madenciliği çalışmalarında başarılı olmanın ilk şartı, uygulamanın hangi işletme amacı için yapılacağının açık bir şekilde tanımlanmasıdır. İlgili işletme amacı işletme problemi üzerine odaklanmış ve açık bir dille ifade edilmiş olmalı, elde edilecek sonuçların başarı düzeylerinin nasıl ölçüleceği tanımlanmalıdır. Ayrıca yanlış tahminlerde katlanılacak olan maliyetlere ve doğru tahminlerde kazanılacak faydalara ilişkin tahminlere de bu aşamada yer verilmelidir.

Verinin Temizlenmesi : Veri madenciliğinde kullanılacak verilerin farklı kaynaklardan toplanması, doğal olarak veri uyumsuzluklarına neden olacaktır. Bu uyumsuzlukların başlıcaları farklı zamanlara ait olmaları, kodlama farklılıkları (örneğin bir veri tabanında cinsiyet özelliğinin e/k, diğer bir veri tabanında 0/1 olarak kodlanması), farklı ölçü birimleridir. Ayrıca verilerin nasıl, nerede ve hangi koşullar altında toplandığı da önem taşımaktadır. Bu nedenlerle, iyi sonuç alınacak modeller ancak iyi verilerin üzerine kurulabileceği için, toplanan verilerin ne ölçüde uyumlu oldukları bu adımda incelenerek değerlendirilmelidir. Bu adımda farklı kaynaklardan toplanan verilerde bulunan sorunlar mümkün olduğu ölçüde giderilerek veriler tek bir veri tabanında toplanır. Ancak basit yöntemlerle ve baştan savma olarak

yapılacak sorun giderme işlemlerinin, ileriki aşamalarda daha büyük sorunların sorunların kaynağı olacağı unutulmamalıdır. Genellikle yanlış veri girişinden veya bir kereye özgü bir olayın gerçekleşmesinden kaynaklanan verilerin *(Outlier),* önemli bir uyarıcı enformasyon içerip içermediği kontrol edildikten sonra veri kümesinden atılması tercih edilir.

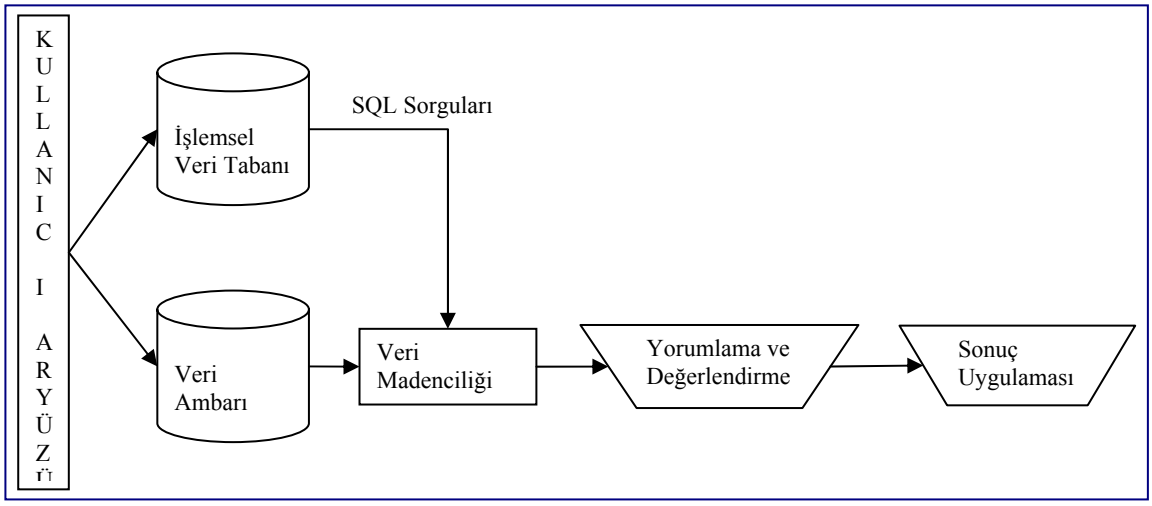
Verinin Seçimi : Bu adımda kurulacak modele bağlı olarak veri seçimi yapılır. Örneğin tahmin edici bir model için, bu adım bağımlı ve bağımsız değişkenlerin ve modelin eğitiminde kullanılacak veri kümesinin seçilmesi anlamını taşımaktadır. Sıra numarası, kimlik numarası gibi anlamlı olmayan ve diğer değişkenlerin modeldeki ağırlığının azalmasına da neden olabilecek değişkenlerin modele girmemesi gerekmektedir. Bazı veri madenciliği algoritmaları konu ile ilgisi olmayan bu tip değişkenleri otomatik olarak elese de, pratikte bu işlemin kullanılan yazılıma bırakılmaması daha akılcı olacaktır.Verilerin görselleştirilmesine olanak sağlayan grafik araçlar ve bunların sunduğu ilişkiler, bağımsız değişkenlerin seçilmesinde önemli yararlar sağlayabilir.

Veri Dönüştürme : Kredi riskinin tahmini için geliştirilen bir modelde, borç/gelir gibi önceden hesaplanmış bir oran yerine, ayrı ayrı borç ve gelir verilerinin kullanılması tercih edilebilir. Ayrıca modelde kullanılan algoritma, verilerin gösteriminde önemli rol oynayacaktır. Örneğin bir uygulamada bir yapay sinir ağı algoritmasının kullanılması durumunda kategorik değişken değerlerinin evet/hayır olması; bir karar ağacı algoritmasının kullanılması durumunda ise örneğin gelir değişken değerlerinin yüksek/orta/düşük olarak gruplanmış olması modelin etkinliğini artıracaktır.

Veri Madenciliği : Tanımlanan problem için en uygun modelin bulunabilmesi, olabildiğince çok sayıda modelin kurularak denenmesi ile mümkündür. Bu nedenle veri hazırlama ve model kurma aşamaları, en iyi olduğu düşünülen modele varılıncaya kadar yinelenen bir süreçtir. Veri madenciliği modelleri ilerleyen bölümlerde anlatılacaktır. Bu aşamada seçilecek modele göre en iyi olan yöntem belirlenir ve uygulanır. Veri madenciliğinde kullanılan çeşitli yöntemler vardır.

Makine öğrenmesi, yapay zeka algoritmaları, bulanık mantık vs. bir çok yöntem veri madenciliğinde model oluşturmada kullanılabilir.

Şekil 3.3'de veri madenciliği ile bilgi çıkarım işleminin işleyişi görülmektedir (Tuğ ve Ark. 2003, Roiger ve Geatz 2003).



Şekil 3.3 Veri Madenciliği İşleyiş Modeli

3.4. Veri Madenciliği Modelleri

Veri madenciliğinde kullanılan bir çok model vardır. Bu modellerin alt bölümlerde açıklanmıştır.

3.4.1. Birliktelik Kuralları

Birliktelik kuralları veri madenciliğinin en iyi örneklerinden biridir. Bu işlem iş hayatı veri tabanlarında sıklıkla uygulanır. Bu tür bir veritabanında örneğin tipik bir satış işlemine ait tüm veriler tutulur.

Formal olarak bir işlem şu şekilde tanımlanır: Her işlem ürünlerin bir alt kümesini içerir. Ürünler stoktaki olası ürünler ise işlem belirli bir müşteriye bu ürünlerin satışını göstersin. Birliktelik kuralı A ➔ B ilişkisi gibi tanımlanır. A ve B'nin her ikisi de ürünlerin bir alt kümesidir. Eğer bu kural geçerli bir kural ise veritabanında hem A hem de B'yi içeren tüm kayıtlarda gösterilir.

Bir örnek üzerinde bir kuralı şu şekilde ifade edebiliriz:
"#10 ve #14 kodlu ürünleri alan müşteriler #19 ve #75 kodlu ürünleri de almıştır."

Birliktelik kuralı madenciliğinin amacı tüm olası kurallar içinde kullanıcının belirlediği ilginçlik ölçüsüne göre keşif yapmaktır. Bu durumları formülize edersek iki yeni kavramla karşılaşırız : Bir kuralın *desteği (support)* ve *güvenilirliği (confidence)*. Destek tüm kayıtlar içerisinde A ve B'nin beraber olma olasılığıdır. Bu A ve B'yi içeren tüm kayıtların sayısının veritabanındaki toplam kayıt sayısına oranıdır. Aşağıdaki gibi formülize edilir.

Destek (A➔B) = P(AUB) = {A ve B'yi içeren kayıt sayısı}/{tüm kayıtlar}

Güvenilirlik ise A ürününü içeren bir kayıt gelirse B ile karşılaşma olasılığıdır. Aşağıdaki gibi formülize edilir.

Güvenilirlik (A➔B) = P (B\A) = {A ve B'yi içeren kayıt sayısı}/{ A'yı içeren kayıt sayısı }

Her iki değer için bir alt eşik tanımlanabilir (min_des ve min_güv). Her kural için min_des ve min_güv değerleri güçlü bir kuralı belirleme kriteridir. Bir veritabanında güçlü kurallar için keşif algoritmalarına giriş verisi gibi verilen min_des ve min_güv değerleri kurallar için bir eşik değeri olarak kullanılır.

Bizim stok örneğimize geri dönersek tipik bir güçlü kuralı "Süt alan müşteriler genellikle bisküvide alır" şeklinde ifade edebiliriz. Keşfedilen kompleks ve anlaşılması zor ilişkiler müşteri davranışlarından bağımsızdır.

Genellikle satış uygulamalarında kullanılan birliktelik kurallarının madenciliği market sepet analizi olarak ifade edilir. Fakat bu ifade birliktelik kurallarının sadece satış uygulamalarında kullanıldığı anlamına gelmez. Tıp gibi diğer uygulama alanları da vardır. Örneğin büyük bir tıbbi veritabanımız varsa semptomlar (bulgular) ve hastalıklar arasındaki yeni ilişkileri bulabilir, teşhis işlemlerinde yardımcı olabiliriz.

3.4.2. Ardışıl Örüntülerin Madenciliği

Ardışıl örüntülerin madenciliği birliktelik kurallarının çalışmasından sonra gelen doğal bir adımdır. Bir önceki adımdaki gibi bu teknik işlemsel veritabanlarına uygulanır. Burada işlemin bir sahibi ve bir zaman göstergesi olmalıdır. Örneğin bir banka veritabanında her bir işlem belirli bir müşteriye uygulanır ve belirli bir zamanda gerçekleşmiştir.

Ardışıl örüntülerin madenciliğinin amacı yeteri kadar sık karşılaşılan ürün alt kümelerinin sıralarının keşfedilmesidir. Ayrıca burada kastedilen "yeteri kadar" ifadesi kullanıcının vereceği bir giriş parametresidir. Belirli bir s sırasının desteği , veritabanının içerdiği s sıralarının sıklığı gibi tanımlanır.

Ardışıl örüntüleri bulma, belirli bir eşik değerinden yüksek bir desteğe sahip tüm s sıralarını keşfetme anlamına gelir. Banka örneğinde olası durumları gösteren veritabanından elde edilen ardışıl örüntülere bakıldığında üyelerin beklenen davranışları gösterdikleri gözlemlenir. Örneğin müşterilerin hesabı kapatılmadan önce benzer davranışlar gösterdiği keşfedilebilir. Bu banka için bir müşterinin servis ve önlemlerden memnun olup olmadığını anlamak için yararlı olabilir. Uygulamadaki diğer örnekler tıp alanında verilebilir. Ardışıl örüntülerin madenciliği ile belirli bir hastanın hastalığının gidişatı ile tedavilerin bağlantısını bulmak mümkün olabilmektedir.

3.4.3. Sınıflandırma

Sınıflandırma bir veritabanındaki nesnelerin bir kümesinin genel özelliklerini bulma işlemidir. Sınıflandırmanın amacı, veritabanındaki nesneleri bir gurup veya sınıfa dahil ederek belirli bir sınıflandırma modeli kurmaktır. Böyle bir model veritabanının eğitimli küme gibi bir örneği şeklinde düşünülebilir.

Eğitim algoritmaları sayesinde eğitimli küme, sınıfları ve onların özelliklerini öğrenmeye izin verir. Böylece bir sınıflandırma modeli doğru olarak oluşturulabilir ve veritabanındaki tüm kayıtların sınıflandırılmasında kullanılabilir.

Sınıflandırma işlemi 3 aşamadan oluşur :

1-) Öğrenme : Algoritma eğitim kümesinin tüm kayıtlarını inceleyerek sınıflandırma modelinin bir tanımlamasını oluşturur.

2-) 1. adımda oluşturulan model veritabanının yeni bir test kümesi ile karşılıklı test edilir. Eğitim kümesi için test kümesinin tüm kayıtları daha önceden sınıflandırılmalıdır. Test kümesi eğitim kümesinden farklı olmalıdır. Sınıflandırma modeli eğitim kümesinde daima çok iyi çalışır.

3-) Sınıflandırma : Sınıflandırma modeli veritabanının kayıtları sınıflandırır.

En genel sınıflandırma modeli karar ağaçlarıdır. Bu modelde her bir düğüm kayıtların özelliklerinin bir test fonksiyonuna tabi tutulmasıyla oluşturulur. Bunun yanında Yapay Sinir Ağları(Artificial Neural Networks), Genetik Algoritmalar(Genetic Algorithms), k-en yakın komşu, Bayes sınıflandırıcısı kullanılan diğer modellere örnek verilebilir.

Problem karmaşıklaştığında sınıflandırma modeli her zaman doğru sonucu vermeyebilir. Bunun için yeterli güvenilirliğe ulaşana kadar model test edilmelidir.

Sınıflandırma veri madenciliğinin genel bir uygulamasıdır ve veritabanını küçük homojen gruplara ayırmaya yardımcı olur. Performans tahmini, marketçilik ve diğer alanlarda sıklıkla kullanılmaktadır.

3.4.4. Kümeleme

Kümeleme işlemi sınıflandırmaya benzer çalışır. Yine amacı benzer olarak veritabanındaki kayıtları homojen gruplara ayırmaktır. Fakat analizden önce kullanıcı bu sınıfları bilemez. Kümeleme algoritması kayıtların oluşturduğu grubu daha doğal bir yolla keşfedecek ve sonra gruplamayı devam ettirecektir.

Kümelemenin en iyi uygulaması uzaysal veritabanlarında gerçekleştirilir. Bu tür veritabanlarında her bir kayıt belirli bir uzayda bir nokta olarak sunulur. Kümeleme algoritması aynı kümelere uyan tüm noktaları bulur. Bunun yanında her hangi bir kümeye dahil olmayan noktalarda olabilir. Böyle noktalar gürültü olarak ifade edilir. Gürültüler kümeleme algoritmasının gücü açısından önemlidir. Özel durumlarda gürültüler araştırmamızın nesneleri olabilir. Örneğin veritabanımızda bir sigorta şirketinin müşteri bilgileri tutulsun ve benzer davranışlara göre bu müşteriler kümelenecek olsun. Bir gürültü alışılmamış davranışlar gösteren bir müşteriyi gösterecektir. Bu gibi bir durumda örneğin şirkete yapılabilecek olası bir dolandırıcılık girişimi gizlenebilirdi ve daha ilerde araştırılmaya gerek duyulabilirdi. Burada kümeleme dolandırıcılık tespiti yapmak için kullanılabilir.

Kümeleme problemine basit olarak 4 farklı yaklaşım vardır:

1-) Parçalı Kümeleme : Algoritmanın bu sınıfı keşfedilecek k ile gösterilen kümelerin sayısı gibi bir giriş parametresine ihtiyaç duyar. Algoritma daha sonra k noktalarını izole eder. Bu noktaları küme merkezleri gibi düşünür. Veritabanındaki diğer noktalar ise bu merkezlere göre kümelenir.

2-) Izgara(Grid) Tabanlı Kümele : Bu metotta veritabanı uzayı farklı hücrelerden oluşan bir ızgaraya bölünür. Bir hücre eğer yeteri kadar sayıda nokta içeriyorsa yoğun olarak düşünülür. Kümeler yoğun bitişik hücrelerin gruplarına karşılık gelir.

3-) Hiyerarşik Kümeleme : Bu metot hiyerarşik yapıya sahip veritabanlarında geliştirilebilir. Bu yapıda ağaçtaki her bir düğüm bir kümeye dahildir. Babanın dahil olduğu kümeye çocuklarda dahildir. Bu yapı tepeden-aşağıya veya en alttan-yukarıya bir stratejiyle oluşturulmuştur. Veritabanı ilk durumda basit büyük bir küme gibi düşünülebilir. Daha sonra veritabanındaki her bir nokta bir kümeye dahil gibi düşünülür ve her bir küme beraber gelişir.

4-) Yoğunluk Tabanlı Kümeleme : Bu metod veritabanı gruplamasını yoğunluk ilişkili noktalara göre yapar. Her bir nokta yerel yoğunluk eşiğine göre gruplanır.

Kümelemenin en genel uygulamaları görüntü işleme alanındadır. Kümeleme algoritmaları dünya yüzeyinin uydu görüntülerini, astronomik görüntüleri ve medical görüntüleri analiz etmede kullanışlıdır.

3.5. Veri Madenciliği Uygulama Alanları

Veri madenciliğinin geniş bir uygulama alanı mevcuttur. Müşterilerin satın alma eğilimlerinin ve alışkanlıklarının belirlenmesi, web sayfalarına erişimlerin analizi, ilaçların yan etkilerinin ortaya çıkartılması, hastalıklar ve semptomlar arasında bağlantıların kurulması, bankacılıkta sahteciliklerin tespit edilmesi, kredi borcu ödeme tahminleri gibi bir çok alanda karar verme işlemlerine yardımcı olacak önemli sonuçlar elde edilebilir (Tuğ, Şakiroğlu, Bulun 2003).

Günümüzde ver madenciliğinin sıklıkla kullanıldığı bazı temel alanlardaki uygulama örnekleri aşağıda verilmiştir.

Tıp : Aşağıda literatürlerde geçen tıp alanında yapılmış veri madenciliği uygulamalarından bazı örnekler görülmektedir. (Tuğ ve Ark. 2003, Roiger ve Geatz 2003)

- Mitchell(1997) , kadınlarda acil C-section gerekliliğinin yüksek riskinin tespit edilebilmesi ile ilgili bir örnek üzerinde veri madenciliği işlemi yapmıştır.

- Merck-Medco Managed Care , aynı tip hastalıklar için daha ucuz ama eşit derecede etkili ilaç tedavilerinin bulunmasına yardımcı olması için veri madenciliğinden faydalanmıştır.

- Gerçek zamanlı veri kümeleri ve halka açık domainler üzerinde göğüs kanseri riski ile ilgili bir çalışmada veri madenciliği işlemi gerçekleştirilmiştir. Göğüs kanseri ameliyathanesine gelen hastaların %30'undan fazlasında göğüs kanseri tekrar oluşmaktadır. Bu veri kümesi her biri 9 nitelik bulunduran 286 kayıt içermektedir. Amaç hastanın tekrar kanser olup olmayacağına karar vermektir. Bu yüzden burada sadece iki sınıf söz konusudur. Bunlar tekrar oluşmama olayları (no-recurrence events) ve tekrar oluşma olaylarıdır (recurrence events). Bütün nitelikler kesindir (kategorikal). Meme kanseri domaini Yugoslavya Ljubljana dan Medical Centre Üniversitesi, Onkoloji enstitüsünden elde edilmiştir.

- Antipsikotik ilaçların kalp kası hastalıkları üzerine etkisi , solunum fonksiyon testlerinin analizi, genetik bozuklukların tespiti gibi bir çok uygulamada veri madenciliği tekniklerinden yararlanılmıştır.

Pazar Analizleri ve Yönetimi : Günümüz serbest rekabet ortamında zaman ve piyasa verilerinin karar destek amaçlı bilgi haline dönüşmüş şekli, satış ve pazarlama faaliyetleri açısından kritik önemi haiz iki unsurdur. Doğru kararlar alma ve bu kararların gecikmeden hayata geçirilmesi işletmelerin varlığını devam ettirebilmeleri açısından çok önemlidir (SPSS).

Pazar analizlerinde kullanılacak verinin kaynağı kredi kartı işlemleri, indirim kuponları, müşteri şikayet aramaları, yaşam stili çalışmaları, müşteri kayıtları, finans kayıtları olabilir. Bu kapsamda yapılan çalışmalar; ilgi alanı, gelir düzeyi, harcama alışkanlıkları, vb. özellikler açısından benzer niteliklerdeki müşteriler için bir model belirlenmesi ve hedef pazarın saptanması; müşterinin fiyat artışı ile değişen satın alma alışkanlıklarının belirlenmesi; çapraz pazar analizleri ile ürün satışları arasındaki birlikteliklerin ve ilişkilerin belirlenmesi ve bu bilgilere dayanılarak ürün satış tahminleri yapılması; müşteri profili belirleme çalışmaları kapsamında hangi özelliklerdeki müşterilerin hangi ürünleri satın aldıklarının belirlenmesi (kümeleme veya sınıflama); müşteri ihtiyaçlarının belirlenmesi kapsamında farklı müşteri tipleri için en iyi ürünlerin neler olduğunun belirlenmesi ve yeni müşterileri çekmede hangi faktörlerin etkili olacağının tahmini; çok boyutlu özetleme raporları ve istatistiksel özetleme bilgileri şeklinde özetlenebilir(SPSS).

Şirket Analizleri ve Risk Yönetimi : Veri madenciliği'nin şirkete ait analizlerde ve risk yönetimi konusundaki kullanımı oldukça yaygındır. Bu yönde yapılan bazı çalışmalar finansal planlama ve aktif varlıkların değerlendirilmesi kapsamında nakit akışlarının analizi ve tahmini, aktif varlıkların değerlendirilmesi için şüpheli alacakların analizi, zaman serileri ve cross-sectional analizi (finansal oranlar, trend analizi, vb.); kaynakların planlanması, işletme performansı değerlendirme ve izleme, işletmenin performansına yönelik geleceğe yönelik tahminler, kaynakların ve harcamaların karşılaştırılması ve özetlenmesi; rekabetsel incelemeler kapsamında rakiplerin ve pazarlama yönelimlerinin incelenmesi (CI: competitive intelligence), müşterilerin sınıflara ayrılması ve sınıflara göre fiyat politikası tayini; rekabetin yüksek olduğu bir pazarda fiyat politikalarının belirlenmesidir (SPSS).

Hilekarlıkların Tespiti ve Yönetimi (Fraud Detection) : Geçmişe ait veriler kullanılarak, geçmişte hilekarlık yapmış kişilere ait veriler incelenebilir ve bunlara ait bir model kurulabilir. Geliştirilen bu model kullanılarak hilekarlığa meyilli olanlar tespit edilebilir. Hilekarlık belirlemenin en yaygın kullanım alanları sigortacılık sektörü, finans sektöründe kredi kartı servisleri, parekendecilik sektörü ve

telekomünikasyon sektörüdür. Örneğin sigorta poliçelerinde yapılan hilekarlıkların tespiti (Fraud Detection) amaçlı bir model ile hilekarlık yapmaya meyilli gruplar belirlenebilir, farklı müşteri gruplarına uygun poliçe türü tespit edilebilir, maliyeti yüksek poliçeler belirlenebilir ve mevcut poliçeler için poliçelerdeki risk tayini yapılabilir(SPSS).

Müşteri İlişkileri Yönetimi (CRM) : Her müşteri aynı zamanda potansiyel bir müşteri adayıdır ve günümüzün yoğun rekabet ortamında şirketlerin başarıya ulaşabilmesi müşterilerini iyi tanıması, var olan müşterilerine yeni satışlar yapabilmesi ve onları memnun edebilmesi, müşterilerinin aynı sektörde hizmet veren diğer şirketlerle çalışmaya başlamasını (churn) engelleme amaçlı pazarlama politikaları geliştirebilmesi, yaptığı promosyonlardan hangilerinin hangi özellikli müşterileri tarafından ilgi göreceğini önceden bilmesi, şirkete kalıcı müşteriler kazandıracak promosyon paketlerinin hangileri olduğunu belirleyebilmesi gerekir. Şirket bünyesinde geçmiş dönemlerde toplanan verilerden veri ayrıştırma yöntemleri kullanılarak müşteri profili belirleme, kampanya yönetimi, müşteri sadakati belirleme, vb. amaçlı modeller geliştirilebilir ve geliştirilen modeller ile istenilen karar destek amaçlı bilgilere çok hızlı bir şekilde ulaşılabilir. Örneğin parekende sektöründe hangi müşterilerin hangi ürün kombinasyonlarını satın aldıklarının belirlenmesi (market sepet analizi) oldukça önemli bir veri ayrıştırma uygulamasıdır ve elde edilen sonuçlar promosyonlarda hedef kitlenin daha doğru belirlenmesi ve ürünlerin yerleştirimi ile ilgili kararlarda önemli karar desteği sağlar(SPSS).

4. YAPAY BAĞIŞIKLIK SİSTEMLERİ

Son yıllarda bilim adamları kompleks problemlerin çözümünde tabiattan esinlenmişlerdir. İnsan sinir sisteminin çalışması modellenerek Yapay Sinir Ağları geliştirilmiştir. Darwin'in evrim prosesi evrimsel algoritmaların oluşturulmasına olanak sağlamıştır. Çok yakın bir geçmişte ise yeni sayısal zeki paradigmalar geliştirmek için güçlü bir metafor olan bağışıklık sisteminin kullanımına önem verilmiştir. Bağışıklık sistemi oldukça güçlü, adaptif, dağıtık, bellek bulunduran, kendi kendine organize olan, güçlü örüntü tanıma yeteneğine sahip ve yabancı etkenlere karşı tepki vermek için evrimsel bir yapısı olan bir sistemdir (de Castro L. & Von Zuben F., 2001). Bu sistemin yetenekleri bilim adamları, mühendisler, matematikçiler, filozoflar ve diğer araştırmacıların ilgisini çekmiştir. Ancak bu konuda yapılan çalışmalar halen başlangıç seviyesindedir. Bağışıklık prensiplerini uygulayan araştırma alanları hızla gelişmektedir ve bu alan Yapay Bağışıklık Sistemleri – YBS (Artificial Immune Systems) veya Bağışıksal Hesaplama (Immunological Computation) olarak bilinmektedir.

İnsan bağışıklık sistemi neredeyse sınırsız sayıda virüs, bakteri, mantar, parazit gibi hastalık yapıcı mikroorganizmalara karşı vücudu koruyan kompleks bir ağ yapısı gibi görülebilir. Bu yapı potansiyel olarak çok zeki hesaplama uygulamalarına sahip , paralel ve dağıtılmış bir adaptif (kendi kendine öğrenebilen) sistemdir. Böyle bir yapının modellenmesi problemlerin çözümünde yeni bir yaklaşım olarak kullanılabilir. Bu yaklaşım YBS'dir. YBS tanıma (recognition), özellik çıkarma (feature extraction), çeşitlilik (diversity), öğrenme (learning) , bellek (memory), dağıtılmış algılama (distributed detection) ve kendi kendine düzenleme (self-regulation) (Nasaroui O., Gonzalez F. ve Dasgupta D.; Dasgupta D., 1999; Dasgupta D., 1997) gibi çalışma alanlarında kullanılabilir.

Bağışıklık sisteminin aşağıdaki özellikleri bilim adamlarının ve araştırmacıların ilgisini çekmiştir (de Castro L., Von Zuben F., 1999) :

- Eşsizlik(Uniqueness) : Her birey kendine özgü bir bağışıklık sistemine sahiptir.
- Yabancı etkenin tanınması : Bağışıklık sistemi ile vücudun doğasında olmayan moleküller tanınabilir ve yok edilebilir.
- Anormal olanı bulma : Bağışıklık sistemi vücutta daha önceden hiç rastlanmamış patojenleri keşfedebilir ve tepki verebilir.
- Dağıtılmış Algılama (Distributed Detection): Bağışıklık sisteminin hücreleri vücudun her tarafına dağıtılmıştır ve daha önemlisi herhangi bir merkezi kontrole maruz değildir.
- Gürültü Toleransı (Noise Tolerance , Imperfect Detection) : Patojenlerin kesin olarak tanınmasına gerek yoktur. Bu yüzden sistem esnektir.
- Takviyeli Öğrenme ve Hafıza (Reinforcement Learning and Memory) : Sistem patojenlerin yapılarını öğrenebilir. Bu yüzden gelecekte aynı patojenlere verilecek yanıtlar daha hızlı ve güçlüdür.

Zeki hesaplama tekniği sayesinde örüntü tanıma, veri analizi, sınıflandırma, öğrenme, hata algılama, optimizasyon, bellek kazanımı, robotik, bilgisayar güvenliği gibi bir çok alanda yapay bağışıklık sistemleri uygulanmıştır. YBS verimli olarak örüntü (pattern) algılayıcıları oluşturmak için kullanılabilir. Bağışıklık prensiplerine dayalı tekniklerle gözlemlenmiş veriden elde edilen bilgiler sayesinde gözlemlenmemiş veri hakkında tahminler yapılabilir. Böylece elde bulunan veri kümeleri analiz edilebilir.

4.1 İmmünolojik Yaklaşım (Doğal Bağışıklık Sistemi)

Bağışıklık sistemi temel görevi bakteri, virüs, mantar, parazit gibi hastalık yapıcı mikroorganizmalara karşı vücudu korumak olan hücreler, moleküller ve organlardan oluşan oldukça kompleks bir sistemdir. Bu sistem vücudun her yerindeki stratejik yerlere yerleştirilmiş fonksiyonel bileşenlere sahip bir örüntü algılama sistemi olarak düşünülebilir (Nasaroui O. , Gonzalez F. ve Dasgupta D.). Bağışıklık sisteminin vücuttaki bütün hücreleri ve molekülleri tanımak, onları sınıflandırmak, yabancı etkenlerle karşılaştığında uygun savunma mekanizmasını uyarmak, yabancı

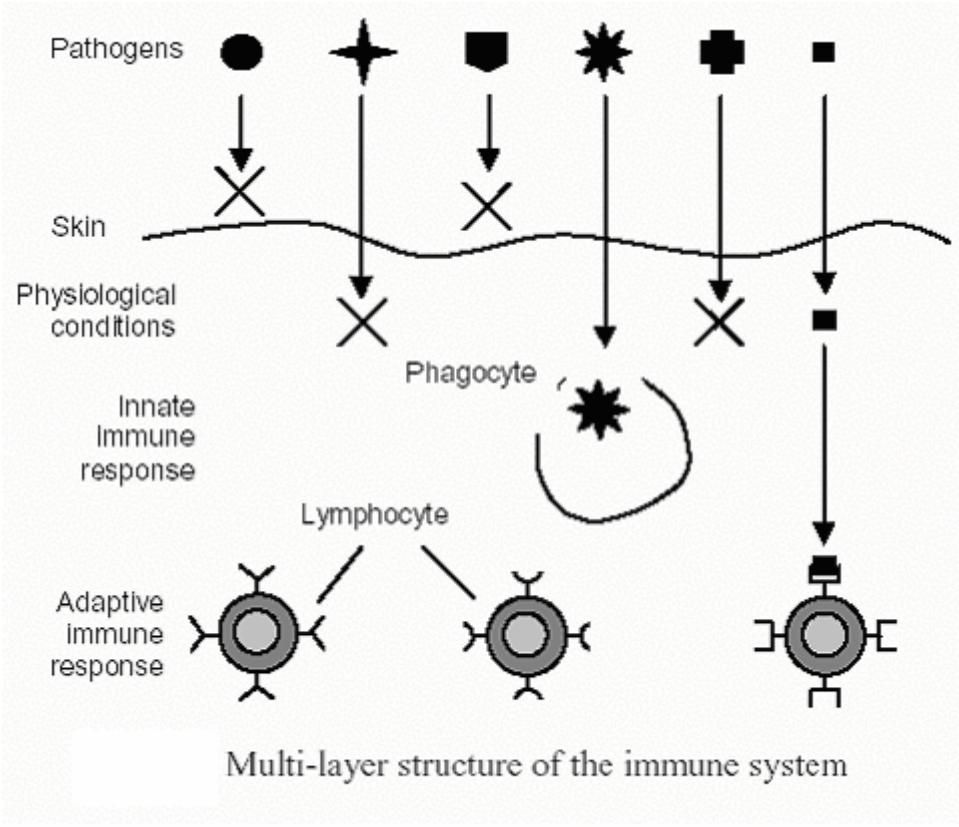
etkenler ve vücudun sahip olduğu hücre ve molekülleri ayırt etmek gibi çeşitli görevleri vardır.

Bağışıklık sistemi sayesinde hastalık yapıcı yabancı hücreler tanınabilir ve hastalık yapıcı olmayan doğal hücreler de ayırt edilebilir. Yabancı bir mikroorganizma organizma ile ilişki kurduğu zaman organizmanın direnci ile karşılaşır. Bu direnç çeşitli mekanizmalara ve temellere bağlı olarak oluşur. Yabancı madde tanınır ve yok edilmeye çalışılır. Sistemin hatırlama yeteneği sayesinde aynı yabancı madde ile karşılaşıldığında tanıma işlemi kısa süreceğinden mikroorganizma daha etkili bir şekilde yok edilir. İki türlü direnç vardır : Doğal direnç ve kazanılmış(edinsel) direnç.

Doğal Direnç : Organizmanın yağısal ve genetik özelliklerine bağlı olan ve vücuda giren tüm yabancı maddelere karşı ilk verilen tepki doğal direnç olarak adlandırılır. Organizmanın bir antijene o antijenle temas etmeden önce gösterdiği tepkidir. Doğal bağışıklık , nonspesifik bağışıklık, yapısal direnç de denilen bu savunma hattı, canlının genetik, anatomik(deri, tüy, mikroflora, vs...), hücresel (makrofajlar, nötrofiller) ve sıvısal (komplement, interferonlar, mukozal salgı ve sıvılar, inflamatuar/yangısal reaksiyonlar) yapısı ile savunma sistemlerinden oluşur.

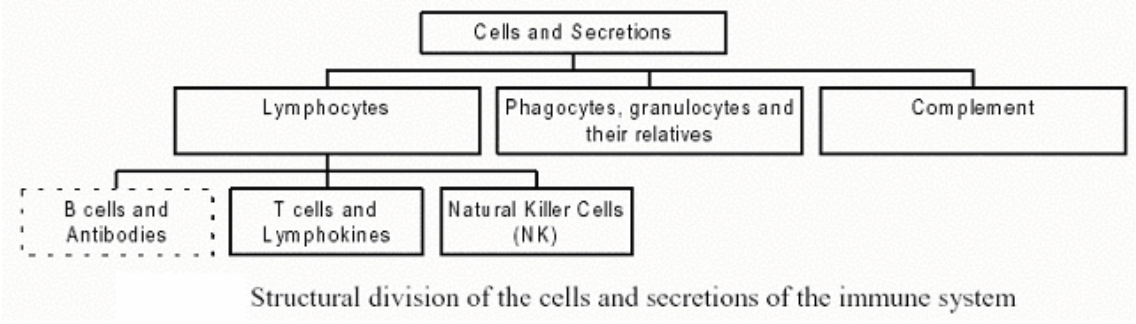
Kazanılmış(Edinsel) Direnç : Doğal bağışıklık sisteminin yetersiz kaldığı yada başaramadığı durumlarda kazanılmış bağışıklık devreye girer ve organizmayı korumaya çalışır. Kazanılmış yada edinsel bağışıklık organizmanın hastalık yapıcı etken veya onun alt birimleri ile ilişki kurmasından sonra oluşur ve oluştuğu canlıda etkilidir. Bu sistem lenfositler(T ve B lenfositler) ve bunların yabancı etkene özgü ürettikleri maddelerden(antikorlar,lenfokinler) oluşur.

Bağışıklık sistemi çok katmanlı bir mimariye sahiptir. Savunma çeşitli katmanlara yayılmıştır. (Janeway Jr. & Travers, 1997; Rensberger, 1996; Hofmeyr, 1997, 2000). Çok katmanlı mimari tek katmanla çözülemeyecek problemlerin çözümünde yardımcı olur. Bağışıklık sistemi mimarisi Şekil 4.1'de gösterildiği gibidir.(de Castro L. ,Von Zuben F., 1999)



Şekil 4.1 Bağışıklık Sisteminin Çok Katmanlı Mimarisi

Bağışıklık sistemi kemik iliğinden meydana gelen hücrelerin büyük bir çeşitliliğinden meydana gelmiştir. Bu hücrelerin bir kısmı genel savunmadan sorumludur diğerleri ise özel patojenlerle savaşmayla görevlidir. Ayrıca etkili bir savunma için hücreler arasında devamlı bir işbirliği vardır. Şekil 4.2'de bağışıklık sistemi tarafından üretilen hücreler ve salgılar gösterilmektedir (de Castro L. ,Von Zuben F., 1999).



Şekil 4.2 Bağışıklık Sistemi Hücreleri ve Salgıları

Lenfositler bağışıklık sisteminde önemli bir sorumluluğa sahip küçük lökositlerdir. İmmun cevaptan sorumludurlar. T ve B lenfosit alt grupları vardır.

Kemik iliğinde üretilen lenfositlerin bazıları, iyice olgunlaşıp fonksiyonel hale geldikten sonra buradan ayrılarak, kan yoluyla lenf dokularına taşınırlar. Bu tip lenfositlere B hücreleri denir. B hücreleri bakteri, virüs, parazit gibi yabancı maddelere karşı tepki vermek için antikor üretmek ve salgılamakla sorumludur. Her B hücresi yalnızca belirli bir antikora özel olarak üretilir.

T hücreleri timüs içerisinde olgunlaştıkları için bu ismi almışlardır (Dreher, 1995). Diğer hücrelerin hareketlerini düzenlemek ve organizmaya giren yabancı hücrelere atak yapmakla sorumludur. Lenfokinler hücresel gelişme, aktivasyon ve regulasyonu sağlarlar. Ayrıca lenfokinler öldürme hedefli hücrelerdir ve makrofajları uyarabilirler.

Doğal killer hücreleri (NK) öldürücü lenfositlerin bir başka türünü meydana getirir. Öldüren T hücreleri gibi güçlü kimyasal maddelerle doldurulmuş tanecikler içerirler. Öldüren T hücrelerinin aksine doğal killer hücrelerinin harekete başlamadan önce spesifik bir antijeni tanımalarına gerek yoktur. Onlar başlıca tümörlere saldırırlar ve bulaşıcı mikroplara karşı vücudu korurlar. Bu hücreler aynı zamanda bağışıklık düzenlemesine, limpokinlerin geniş miktarda salgılanmasına katkıda bulunurlar.

Fagositler ("hücre yiyenler") mikroorganizmaları ve antijenik tanecikleri yeme ve sindirme yeteneği olan beyaz kan hücreleridir. Bazı fagositlerin aynı zamanda lenfositler için antijen sunma yeteneği vardır. Bu fagositler antijen sunan hücreler (APCs) olarak adlandırılmaktadır. Bağışıklık tepkisinin başlamasında önemli bir rol oynarlar.

Tamamlayıcı sistem (The Complement System) bileşenlerinin her biri bir zincir reaksiyonla organizmaya giren yabancı maddelerle savaşır. Ayrıca fagositlerin operasyonlarını kolaylaştırmakla sorumludurlar.

4.2 Yapay Bağışıklık Sistemi (Artificial Immune System)

Doğal bağışıklık sisteminin bahsedilen özellikleri bilim adamlarının ilgisini çekmiştir. Kendi kendine öğrenebilmesi, güçlü ve çok zeki bir tanıma yeteneğinin olması, adaptif olması, paralel, dağıtılmış ve çok katmanlı bir ağ yapısı sunması zeki makinelerin alternatif bir yapay modeli olabileceğini göstermiştir. Yeni bir sayısal araştırma alanı olarak 1990'lı yıllarda Yapay Bağışıklık Sistemleri ortaya çıkmıştır. İlk pratik uygulamaları Dasgupta ve Forrest tarafından yapılmıştır (Dasgupta D. 1999,1997; Dasgupta D. & Forrest S., 1997; Dasgupta D. & Attoh-Okine N., 1997; Forrest S., Perelson A. S., Allen L. ve Cherukuri R., 1994; Forrest S., Javornik B., Smith R. E. ve Perelson A. S., 1993). Günümüzde ise bir çok alanda uygulamaları mevcuttur. Özellikle örüntü tanıma olarak bilinen mühendislik alanında , bağışıklık sisteminin yabancı maddeyi kolayca tanıması eğer tanıyamıyorsa güçlü bir alternatif çözümü bulabilmesi özelliği sayesinde daha iyi sonuçlar elde etmeyi sağlamıştır. Örüntü tanıma dışında , büyük veri kümeleri üzerinde bilgi keşfetme olarak bilinen veri madenciliği sahasında verilerin analizinde, zararlı bir antijenle organizmanın kendi sahip olduğu hücreleri ayırt edebilme yeteneğinden esinlenerek bilgisayar güvenliği ve hata algılama problemlerinde, sürekli öğrenen sistemler geliştirmede, sınıflandırma problemlerinin çözümünde, en iyi çözümü bulma amacını güden optimizasyon alanında, yine yapay zeka sahalarıyla ortak çalışmalar sonucu insan gibi davranabilen zeki sistemler geliştirmeye yönelik olarak robotik alanında veya zeki makineler sayılabilecek robotlar geliştirmede avantajlar sağlayacaktır. Yeni bir araştırma sahası olmasına rağmen bu konuda yapılan çalışmalar hızla artmaktadır. Bu alanda Ishida(1996), Hunt & Cook (1996), Dasgupta (1999) ve Holmeyr & Forrest(1999) çalışmalarından yararlanabiliriz.

Yapay bağışıklık sisteminin amacı , doğal bağışıklık sisteminin çalışma yapısını referans alarak gerçek dünya problemlerini çözmeye yönelik olarak algoritmalar geliştirmektir. Bunun için bağışıklık sisteminin çalışmasının iyi bir şekilde yorumlanması gerekir. Bu sebeple yeni fikirler üretebilmek için farklı alanlarda çalışan bilim adamlarının ortak çalışması çok önemlidir.

4.3 Bağışıklık Mühendisliği

Mühendislik belirli bir görevi uygulamak için temel bir sistem tasarlamakla ilgilenir. Geleneksel yaklaşımlar problemlerin çözümünü adım adım işleyen prosedürler kullanarak bulmaya çalışırlar. Ancak bu yöntem kompleks mühendislik problemlerinin çözümünde verimli sonuçlar elde etmeyi zorlaştırır. Bu yüzden yeni yaklaşımlar geliştirmek gerekir. Doğal sistemlerin yetenekleri geleneksel teknoloji araçlarına göre çok üstündür. Doğal sistemlerin davranışlarını yöneten genel prensiplerin açıklanması ve uygulanması mühendisliğin yeni biçimlerini oluşturacaktır (Wolframi, 1986). Bu düşünceyle bağışıklık bilimi (immunology) ve sinir biyolojisi (neurobiology) gibi doğal sistemlerden toplanan düşünceler ve mühendislik sistemleri kompleks davranışları kontrol edebilir ve belirli görevlere uygulanabilir.

Bağışıklık mühendisliği (Immune Engineering) doğal bağışıklık sistemine dayanan sayısal bir sistem olan yapay bağışıklık alanında belirli bir problem için çözüm araçlarını tanımlayan ve elde edilen çözümleri uygulayan bir prosestir. Bu amaçla kompleks problemleri çözmek için tamamen bağışıklıktan esinlenilen algoritmalar ve yöntemler geliştirmekle görevlidir.

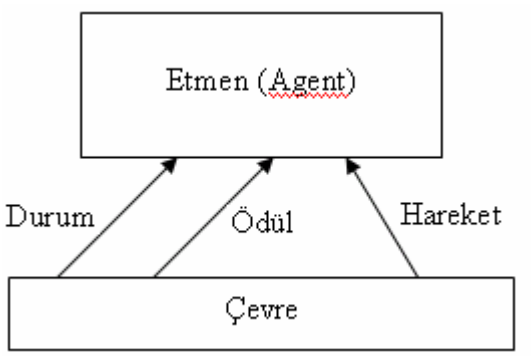
4.4 Yapay Bağışıklık Sistemi Uygulamaları ve Örnek Algoritmalar

Bu bölümde bağışıklık sistemindeki fikirler kullanılarak geliştirilmiş olan bazı algoritmalar ve uygulamalar sunulacaktır. Takviyeli öğrenme, örüntü tanıma, yapay bağışıklık ağ teorisi ve evrimsel bir sistem olarak klonal seçim algoritması konularından bahsedilecektir.

4.4.1 Takviyeli Öğrenme (Reinforcement Learning)

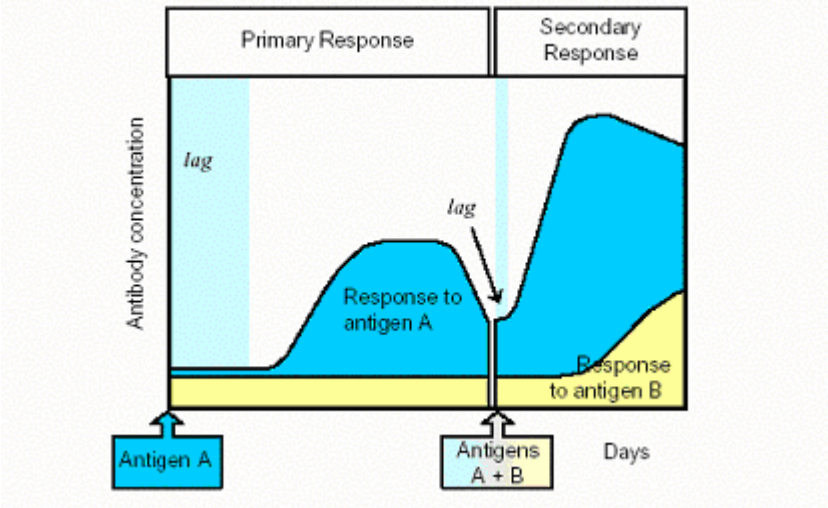
Takviyeli öğrenme bir ortamı algılama ve hareket etme yeteneklerine sahip otonomos bir etmenin(agent), amacını başarmak için en uygun hareketleri nasıl öğreneceği sorusuna cevap verir (Sutton R. S. , Barto A. G. , 1998; Kaelbling L. P.,

Littman M. L. ve Moore A. W. , 1996). Etmenin ortam hakkında ön bilgiye sahip olmaması bir problemdir. Standart bir takviyeli öğrenme prosesi Şekil 4.3' de gösterildiği gibidir (Mitchell T. M., 1997). Bir etmen kendi çevresiyle etkileşimdedir. Çevreden aldığı olası durumlara karşılık en uygun hareketi belirleyip uygulamakla sorumludur. Seçmiş olduğu hareket ne kadar iyi ise buna göre ödül alır. Böylece daha sonraki durumlar karşısında hedefine en uygun olarak seçeceği hareket daha iyi olacaktır.



Şekil 4.3 Takviyeli Öğrenme Prosesi

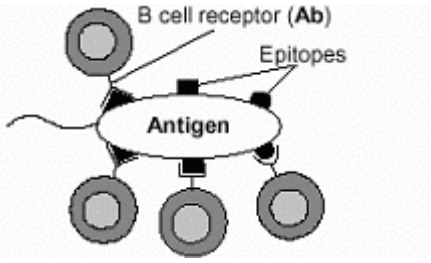
Bağışıklık sisteminin gelişiminde bir organizmanın hayatı boyunca defalarca belirli bir antijenle karşılaşması beklenebilir. Adaptif bağışıklık tepkisini uyaran bir antijen ilk açığa çıktığında, her bir farklı eğilim için antikor üreten B hücrelerinin küçük klonlarının bir spektrumu tarafından ele alınır. İkinci kez karşılaşınca bağışıklık tepkisinin geçerliliği, ilk bulaşmadaki yüksek eğilimli antikor üreten hücreler (bellek hücreleri) depolanarak oldukça geliştirilmiş olacaktır (Ada & Nossal, 1987). Her zaman sıfırdan başlamaktan ziyade bu strateji uygulanarak her bulaşmadan sonra ard arda antijenle karşılaştıkça bağışıklık tepkisinin hem hızı hem de doğruluğunun artması sağlanır (Perelson 1978, Farmer 1986). Bu şema takviyeli öğrenme stratejisine özgüdür (Sutton & Barto, 1998). Sistem sürekli olarak görevin uygulanma yeteneğini geliştirir. Şekil 4.4'de bu durum gösterilmiştir.



Şekil 4.4 Birinci ve İkinci İmmün Yanıt Etkileri

4.4.2 Örüntü Tanıma (Pattern Recognition)

Bağışıklık sistemi potansiyel olarak örüntü tanıma alanında güçlü çözümler üretir. Bir örüntü(patojen) formunda küçük bir değişiklik olsa dahi sistem bir tepki üretebilir (çapraz reaktivite). Bağışıklık sisteminde örüntü tanımanın bakış açısından hem B hem de T hücrelerinin en önemli özelliği antijenleri tanıyabilen arayüzlerindeki reseptör moleküllerine sahip olmasıdır (ya serbest ya da bir MHC molekülü sınırında). Şekil 4.5'de bu durum gösterilmiştir.



Şekil 4.5 Antikor Molekülünün Antijeni Tanıması

T hücreleri durumunda reseptör basitçe T hücre reseptörü (TCR) olarak adlandırılır. Bağışıklık sisteminde tanıma, moleküler seviyede meydana gelir ve reseptörün bağlandığı yer ile epitop olarak adlandırılan antijen parçasının birbirini

tamamladığı kabul edilir. Antikorlar tek bir reseptör türüne sahip iken antijenler birden fazla epitop'a sahip olabilir. Bunun anlamı tek bir antijenin farklı antikor molekülleri tarafından tanınabildiğidir.

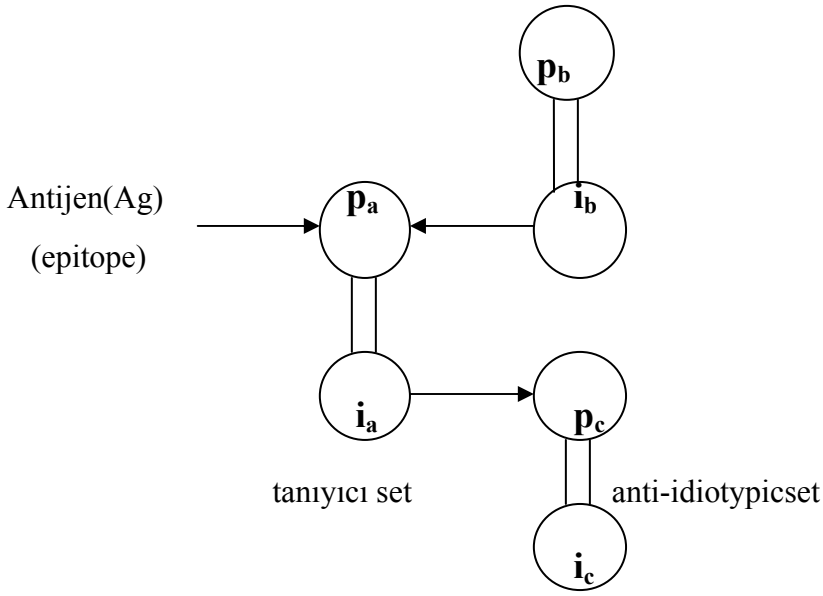
B ve T hücre reseptörleri bir antijenin farklı özelliklerini tanır. B hücre reseptörleri bozulmamış antijen moleküllerinde bulunan epitoplarla birlikte birbirlerini etkiler. T hücre reseptörleri sadece hücre arayüz molekülleri ile etkileşimdedir. T hücreleri bağışıklık tepkisinin düzenlenmesinde çok önemli bir rol oynayan diğer hücreleri öldürebilen veya onların gelişmesini ilerleten kimyasal maddeler salgılar. Bir hücre arayüz molekülü tanınarak T hücresi eritilebilen bir molekülden ziyade bir başka hücre ile etkileşimde olup olmadığını belirlenir. T hücresi reseptörü bir major histocompatibility complex (MHC) olarak adlandırılan hücre arayüz molekülü ile sınırlanan antijeni tanır.

4.4.3 Yapay Bağışıklık Ağ Teorisi (Artificial İmmune Network-aiNet)

De Castro ve Von Zuben tarafından oluşturulan aiNet yani Yapay Bağışıklık Ağ Teorisi**(**Artificial Immune Network) modeli Yapay Sinir Ağlarına (Artificial Neural Network) alternatif olarak bir öğrenme modeli sunar.

Bağışıklık Ağ Teorisi, antijenlerin yokluğunda bile birbiri ile etkileşim içerisinde olan, bir başka deyişle birbirini tanıyan hücre ve moleküllerden oluşan bir yapıyı öngörür. Bu teori ortaya atıldığı sıralarda , bağışıklık sistemindeki klonların ya da hücrelerin sadece antijenler tarafından uyarıldıkları zaman aktive olduklarına , diğer zamanlarda resting-state (uyku modu) denilen bir durumda kaldıklarına inanılıyordu. Biliyoruz ki B hücreleri tarafından salgılanan Antikorlar (Ab) sadece yabancı proteinleri değil aynı zamanda birbirlerini de tanırlar. Burada iki kavram karşımıza çıkmaktadır: **idiotype** ve **idiotope.** İdiotype, bir Ab setinin tanıdığı epitopların oluşturduğu bir settir. İdiotope ise bu setteki her bir epitopa verilen addır.

Bağışıklıksal Etkileşimler:



Şekil 4.6 Bağışıklık Ağ Modeli

Yukarıdaki şekilde bağışıklıksal etkileşimlerin bir kısmını görmekteyiz. Bağışıklık sistemi bir antijen(Ag) tarafından uyarıldığında, bu antijenin epitopu farklı paratoplar tarafından farklı derecelerde tanınır. Bu paratopların oluşturduğu sete pa adını veriyoruz. Bu pa paratopları, Ab ve reseptör moleküllerinin paratoplarıdır. Bu setteki her paratopun tanıdığı idiotoplar vardır. Böylelikle pa paratop setinin tanıdığı bir idiotope seti yani ia seti karşımıza çıkar. Paia sembolü , Ag'i tanıyan Ab molekülleri ile , potansiyel olarak ona yanıt oluşturabilecek lenfositlerin oluşturduğu toplam bir settir. Bağışıklık ağında pa setindeki her paratop, , bir idiotope setini tanır. Böylece toplam pa seti oldukça geniş bir idiotope setini tanımış olur. Buradaki ib idiotope seti, Ag antijenini tanıyan pa seti tarafından tanınan (eğer varsa) diğer epitopların dahili görüntülerini temsil eder. Bu idiotope setini yani ib'yi tanıyan diğer paratopların oluşturduğu set ise pb setidir. Bunun yanında , paia setindeki her idiotope, bir paratop seti tarafından tanınır. Bu da toplam ia setinin daha büyük bir pc paratop seti tarafından tanınması anlamına gelir. Yine benzer şekilde, bu pc setindeki her bir paratop bir ic idiotope seti tarafından tanınır ve toplam pc seti daha geniş bir ic seti ile birlikte ifade edilir-pcic. Fakat bu sefer bu sete tanıyıcı set yerine anti-idiotypic set adı verilir çünkü bu set asıl Ag'i değil , idiotope'ları tanımaktadır. Bu

yöntemi bu şekilde devam ettirirsek, ağ içerisinde bir önce tanımlanan setleri tanıyan ve yine bu setler tarafından tanınan oldukça büyük setlere ulaşırız.

Bu ağ yaklaşımı özellikle hesapsal yöntemlerin oluşturulması adına oldukça dikkat çekicidir. Çünkü öğrenme, hafıza, boyut kontrolü ve çeşitlilik gibi özellikleri barındırmaktadır. Genel olarak birçok network modelinin yapısını şu şekilde formülüze edebiliriz:

RPV(rate of population variation)=uyarılmayan hücrelerin ölümü+uyarılan hücrelerin çoğalması

Buradaki son terim, hem Ab-Ab hem de Ag-Ab tanımasını içermektedir.

Bağışılıksal modellemeyi tanımlayabilmek için tüm bağışıklıksal olayların bir S şekil uzayında gerçekleştiği farzedilmiştir. Bu S şekil-uzayı, çok boyutlu bir metric uzayıdır ve her eksen bir moleküler şeklin fiziko-kimyasal ölçütünü karakterize eder. Ayrıca şekil uzayının L boyutlu olduğu kabulü yapılmıştır. Bu şekil-uzayındaki her bir s Є S noktası bize Ab-Ab ya da Ag-Ab etkileşimlerini belirlemek için gerekli olan özelliklerin bir setini vermektedir. Matemetiksel olarak ise s noktasını L boyutlu bir dizi şeklinde ifade edebiliriz.

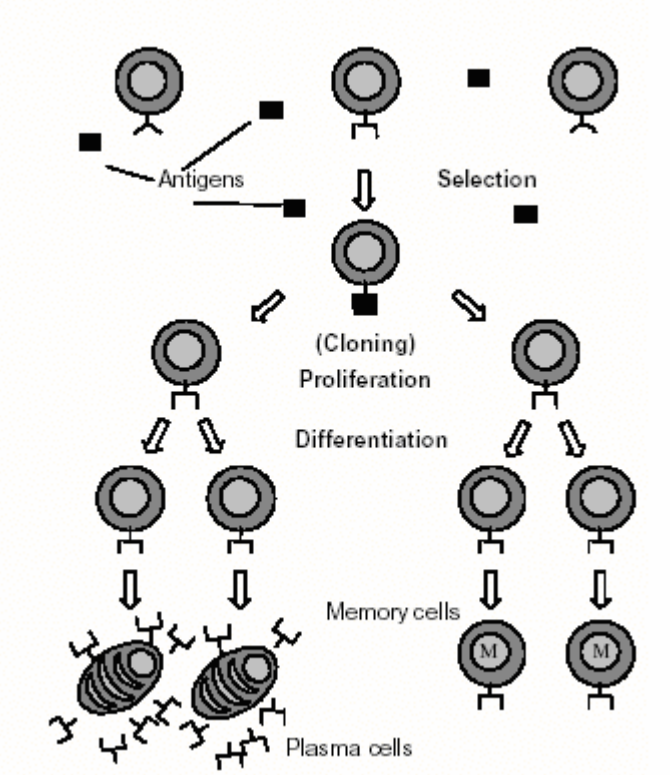
Oluşturulacak ağın giriş örüntüleri Ag'lerdir. Ag-Ab ve Ab-Ab etkileşimleri benzerlik ölçütlerine dayalı olarak belirlenir. Bu benzerlik ölçütleri de öklid uzayı ile ilişkilidir. Buradaki amacımız, tanınması gereken antijenlerin ağdaki görüntülerini temsil eden bir Ab repertuarı oluşturmak için bir öklid ölçütü- metric- kullanmaktır. Ag-Ab duyarlılığı, aralarındaki uzaklığın tersi ile orantılı olacaktır yani aralarındaki uzaklık az ise duyarlılık fazla, çok ise duyarlılık az olacaktır. Hesaplanan bu duyarlılıklara bağlı olarak , birbirine karşı duyarlılıkları önceden belirlenmiş bir eşik seviyesinden büyük olanlar atılır. Elemeden sonra kalan klonları tüm ağ hücreleri arasına eklenir. Ağdaki her bir hücre çifti arasındaki benzerliği belirlenir. aiNet öğrenme algoritmasının amacı, veriyi tanıyan ve onun yapısal organizasyonunu temsil eden bir hafıza seti oluşturmaktır. Ab'ler daha spesifik olurlarsa, ağdaki eleman sayısı daha fazla olur. Bu nedenle diyebiliriz ki baskılama eşiği , Ab'lerin spesifiklik seviyesini, kümeleme doğruluğunu ve ağdaki esnekliği kontrol eder.

Ağ çıkışı olarak ise hafıza antikorlarının koordinatlarının yer aldığı $\mathbf{Ab}_{\{m\}}$ matrisi ile , bu antikorlar arsındaki benzerlikleri barındıran bir **S** matrisi elde edilir. $\mathbf{Ab}_{\{m\}}$ matrisi, aiNet'e sunulan Ag'lerin bir görüntüsünü temsil ederken, **S** matrisi ise hangi network Ab'lerinin birbirleri ile bağlantılı olduklarını ve bu bağlantıların derecelerini temsil eder, başka bir deyişle genel ağ yapısını temsil etmiş olur.(de Castro L. & Von Zuben F., 2001)

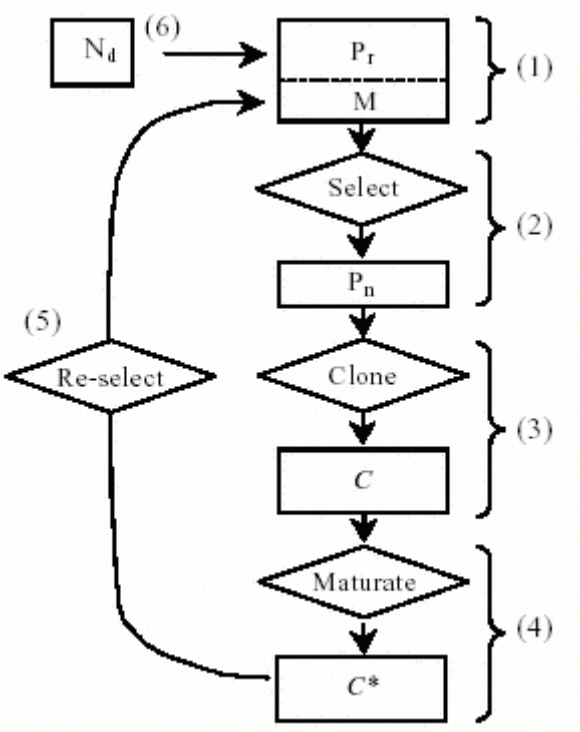
4.4.4 Klonal Seçim Algoritması (Clonal Selection Algorithm)

Klonal Seçim Algoritması (Clonal Selection Algorithm), doğal immune sistemden yararlanarak bir antijenik uyarıcıya bir immün yanıtın temel niteliklerini kullanır. Vücuda bir antijen girdiği zaman kemik iliğinden elde edilen B lenfosit hücrelerinin alt populasyonu antikorlar(Ab) üretir. Her bir hücre yalnızca bir çeşit antikor salgılar ve bu antikor da antijene özeldir. Bu antikorların bağlanmasıyla ve T-Helper hücresi gibi aksesuar hücrelerinden ikinci bir sinyal alınmasıyla antijen B hücrelerini uyarır. Birçok hücre bölünerek (mitoz) bir klon oluşturur. (de Castro L. & Von Zuben F., 2001)

İmmune Sistemin Temel Yönleri : repertuardan bağımsız olarak hafıza hücrelerinin bakımı ve uyarılan hücrelerin klonlanması, uyarılmayan hücrelerin yok edilmesi, affinity maturation ve yüksek affinity'li klonların tekrar seçimi ve hipermutasyondur.



Şekil 4.7 Klonal Seçim Modeli



Şekil 4.8 Klonal Seçim Algoritması

Algoritmanın Adımları

- (1) aday çözümlerin bir kümesi (P) üretilir, hafıza hücrelerinin (M) alt kümesi populasyona eklenir.(P=PR+M)
- (2) Bir affinity ölçüsü kullanılarak populasyonun (Pn) en iyi n bireyine karar verilir(Seçim)
- (3) Populasyonun bu n bireyi çoğaltılır (klonlama), klonların geçici bir populasyonu verilir. Klonun boyu antijenin affinity'sinin artan bir fonksiyonudur.
- (4) Bir hipermutasyon şemasına klonların populasyonu uygulanır. Burada hipermutasyon antijenle antikorun affinitysine orantılıdır. Olgunlaşan antikor polulasyonu üretilir(C*)
- (5) C*'den hafıza kümesine(M) birleştirilen gelişmiş bireyler tekrar seçilir. P'nin bazı üyeleri C*'nin gelişmiş diğer üyeleriyle değiştirilebilir.
- (6) Yeni bireylerle d antikorları değiştirilir. Düşük affinityli hücrelerin değişitirilmesi ihtimali daha yüksektir.

Bu algoritma kompleks makine öğrenmesi araçlarına uygun bir evrimsel strateji sunar. Öğrenme ve hafıza kazancı gibi alanlarda performans sağlar. Algoritmanın binary karakter tanıma, çok-modlu ve kombinasyonel optimizasyon , bağışıklık ağ teorisi gibi alanlarda uygulaması mevcuttur. Ayrıca algoritma GA' ya alternatif bir algoritmadır.

4.4.5 Simulated Annealing Algoritması (SAND)

Doğal bağışıklık sistemi, ana amacı vücudumuzu antijen adı verilen zararlı dış istilacılardan korumak olan kompleks bir örüntü tanıma elemanıdır. Bu elemandaki temel elementler antikorlardır (Ab). Ab'ler antijenlere (Ag) bağlanarak bu kompleksin (Ab-Ag kompleksi) diğer hücreler tarafından imha edilmesini sağlarlar. Bağışıklık sistemimizdeki Ab sayısının, olası Ag sayısından çok daha az olduğu bilinmektedir. Bu da Ab repertuarının sağlaması gereken en önemli özelliklerin çeşitlilik ve bireysel bağlanabilme yeteneği olduğunu göstermektedir. SAND

yaklaşımı araştırma alanına en iyi yanıtı verecek aday çözümlerle dolu bir havuz üretmeyi amaçlar. Bu yaklaşım problem hakkında hiçbir ön bilginin olmadığını farz ederek, sabit uzunlukta bir potansiyel çözüm adayları kümesi üretir. Daha sonra binary dizinler arasındaki Hamming uzaklığını hesaba katan bir enerji fonksiyonunu maksimize ederek populasyondaki çeşitliliği sağlar. Modelimiz başlangıçtaki uygun bir şekilde üretilmiş Ab populasyonu ile rasgele üretilmiş Ab populasyonunu karşılaştırır.

Genetik Algoritmaların tipik uygulamalarında evrimsel strateji, oldukça belirsiz ortamlarda, çok büyük bir aday populasyonu içerisinden giderek sayısı fazlalaşan uygun organizmalar üretir. Bu populasyonun tek bir üyesi, en iyi çözümü belirler. Kullanılan genetik operatörler populasyonu daha iyiye doğru sevk ederler. GA'daki rasgelelik öğrenme hızında güçlü bir etkiye ve çözüm kalitesine sahiptir. Yetersiz bir başlangıç populasyonu seçimi lokal minimuma takılmaya veya sayısal problemlerle karşılaşmaya neden olabilir. Kolen ve Pollack bir yapay sinir ağında başlangıç ağırlık vektörünün seçiminin önemini vurgulamışlardır (Kolen J.F. & Pollack J.B., 1990). Optimal ağırlık kümesinin elde edilmesinde global bir arama yapmanın mantıklı olmadığını göstermişlerdir. Lokal arama ile daha iyi çözümler elde edilebilir. Benzer şekilde GA'da başlangıç populasyanının iyi bir seçimi algoritmanın performansını iyileştirecektir.

Başlangıç populasyonu ve çözüm kalitesi arasındaki korelasyon başlangıç Ab repertuarı ve doğal immüne sistemin yanıt kalitesi arasındaki var olan korelasyona benzer. Bu kompleks bir örüntü tanıma aracı gibi görülebilir.

SAND algoritması istatistiksel mekanikler ve kombinasyonel optimizasyon arasındaki bağlantıyı sağlar.Metodun kaynağı katı veya sıvı madde örneklerinde bulunan çok sayıdaki atom özellikleri ile ilgilidir. Verilen bir sıcaklıkta, ısıl dengedeki bir sistemin davranışı deneysel olarak ortalama bir davranışın etrafındaki küçük sapmalar olarak karakterize edilebilir. Her atomik pozisyon bir olasılık faktörü ile çarpılır. Bu olasılık faktörü :

$$P(\Delta E) = \exp(-\Delta E/T)$$

Burada E, konfigürasyonun enerjisi; T, sıcaklık ve ΔE, ölçülen enerjideki küçük bir sapma miktarıdır.

Algoritmanın her adımında, bir atoma rasgele bir yer değiştirme verilir ve sonuçta sistemin enerjisinde oluşan ΔE değişimi hesaplanır. Eğer $\Delta E \leq 0$ olursa, yer değiştirme kabul edilir ve yer değiştiren atomun bulunduğu konfigürasyon bir sonraki adım için başlangıç noktasını oluşturur. $\Delta E > 0$ olma durumu da olasılıkla değerlendirilir : konfigürasyonun kabul edilme olasılığı P(ΔE) olasılığıdır.

Sıcaklık, enerji fonksiyonu gibi, enerji fonksiyonuyla aynı birimli bir kontrol parametresidir.SAND yaklaşımında önce optimize edilecek sistem yüksek bir sıcaklıkta çözünür, daha sonra sıcaklık sistem donuncaya kadar ve artık değişiklik olmayıncaya kadar küçük miktarlarda düşürülür. Her sıcaklıkta; simülasyon sistemin kararlı hale ulaşmasını sağlayacak kadar uzun bir zamanda gerçekleştirilmelidir.Dikkat edilirse, sıfır haricindeki sıcaklıklarda her zaman lokal optimumdan geçişler mümkündür. Sıcaklık sonuçları ve her sıcaklıkta dengeye ulaşmak için gerçekleşen ΔE değişimlerinin büyüklüğü bir annealing programı olarak düşünülebilir.

Başlangıç populasyonu belirlemede bu populasyon Ab repertuarı olarak belirlenmelidir. Bizim problemimizde Ag populasyonu önemsenmeyecektir. Ab populasyonunun enerji ölçüsü ise populasyondaki bireylerden elde edilecektir.

Ag-Ab bağlantısı bir antikor ve bir antijen arasındaki tamamlayıcı bir ölçü ile sunulur. Hedef (Ab – Ab) antikorlar arasındaki uzaklığı maksimize etmektir.Bu populasyondaki benzerliklerin azalmasına dolayısıyla çeşitliliğe imkan sağlar.

Ag –Ab etkileşimlerinin tanımlanması modelinde L parametrelerinin bir toplamıyla tanımlanan bir antijen ve bir antikor reseptörünün bağlandığı varsayılır. Antikorlar ve Antijen belirleyicileri ($Ag^{(-1)}$) S olarak kabul edilen L boyutlu öklit

şekil uzayında sunulur. Ab ve $Ag^{(-1)}$ arasındaki yakınlık onların affinity ölçüsü ile belirlenir. Antikorlar ve antijenlerin komponenti gerçek değerli koordinatların bir kümesiyle sunulur. Matematiksel olarak her bir molekül L boyutlu gerçek değerli uzayda bir nokta gibi gösterilir ve Ab – Ag affinity'si onlar arasındaki öklit uzaklığının(ε) tersiyle ilişkilidir. Bu uzaklığı bazı yazarlar bir uyarım topluluğu olarak ifade ederler. Çünkü bu bir antijen grubunu gösterir.

Öklit uzaklığının tersi moleküller arasındaki affinity ölçüsü olarak kullanılabilir :

$$Aff(xi , xj) = 1 / ED (xi, xj) + \varepsilon$$

Burada xi ve xj L boyunda bağımsız vektörlerdir. ε küçük bir pozitif ölçüdür.

$$ED(xi , xj) = \sqrt{\sum}^{1}_{k=1}(xik - xjk)^2$$

Öklit arama uzayında , enerji ölçüsü antikor populasyonundaki tüm vektörler için öklit uzaklıklarının toplamı şeklinde optimize edilebilir :

$$E = \sum^{N}_{i=1}\sum^{N}_{j=i+1} ED(xi , xj)$$

5. BULANIK MANTIK

1965 yılında California Üniversitesinden Lotfi Zadeh tarafından ortaya atılan Bulanık Mantık (BM) teorisi 80'li yıllardan itibaren kullanılmaya başlanmıştır. İlk uygulaması İbrahim Mandani tarafından bir buhar makinasında gerçekleştirilmiştir. Öncesinde Aristo mantığı ya da diğer adıyla ikili(Boole) mantık kullanılmaktaydı. İkili sınıflandırma mantığı keskin bir sınıflandırma gerektirir. Ancak BM ikili mantıkdan farklı olarak ara kavramlarlada ilgilenir ve daha gerçekçi sınıflandırmalar yapar. İkili mantık ya siyah ya da beyaz der. Ama BM siyah ve beyaz arasındaki değerleride biraz siyah, koyu beyaz, gri vs. kabul eder. Böylece gerçek dünya problemleri için daha hassas çıkarımlar sağlanır. Tam ve kesin olmayan bilgilere dayanarak tutarlı ve doğru kararlar vermeyi sağlayan düşünme ve karar verme mekanizması Bulanık Mantık olarak adlandırılır(Allahverdi 2002).

Bir sistemin veya bir olayın niceliğini bilmek ve onu analiz edebilmek için onun miktarsal karakteristiklerini bilmemiz gerekir. Bilgisayar sistemleri ile birlikte özellikle kontrol sistemlerinde bu karakteristik veriler kullanılmakta ve verimli sonuçlar elde edilmektedir. İnsancıl sistemlerde ise bu tür modellemeler yapamayız. Çünkü insancıl sistemlerde çok büyük bir hassasiyet istenir. Bilgisayar sistemlerinde ise bu yoktur. Bir sistemin karmaşıklığı arttıkça hassasiyeti azalır. Bunun için kesin matematiksel formüller ve veriler yerine dilsel veriler kullanırız.

Ahmet Uzun Boyludur. (Dilsel Değer Boy = Uzun)
Ahmet'in Boyu 1,90'dır. (Kesin Değer Boy = 1,90)

Ahmet Gençtir. (Dilsel Değer Yaş = Genç)
Ahmet 25 Yaşındadır. (Kesin Değer Yaş = 25)

Değişkenlerin dilsel değerlerini kullanarak ise BM yardımı ile hesaplamalar yapabiliriz. Örneğin bir çamaşır makinesinde çamaşırın ağırlığına göre deterjan kullanılması işlemi gerçekleştirilebilir.

Bulanık Teorinin Olumlu Yönleri

- Bulanık teori insan düşünce mantığına uygundur.
- Bulanık teorinin uygulanmasında karmaşık matematiksel formüllere gerek yoktur.
- Yazılım geliştirmek çok kolaydır.

Bulanık Teorinin Olumsuz Yönleri

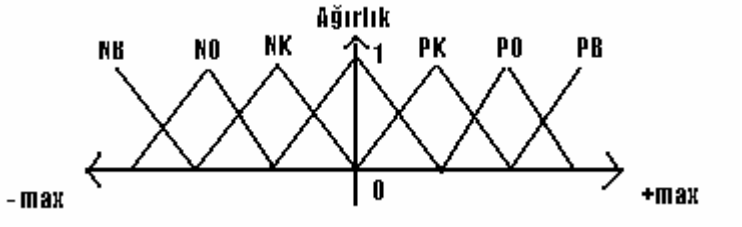
- Bir uzman sisteme ve uzmana ihtiyaç duyar.
- Bazı sonuçlar (üyelik fonksiyonları) dene yanılmayla bulunur.
- Matematiksel model olmadığından sistem kararlılığını anlamak zordur.

5.1. Üyelik Fonksiyonları ve Fonksiyon Etiketleri (Dilsel Değerler)



Şekil 5.1 Üyelik Fonksiyon Örnekleri

Şekil 5.1'de de görüldüğü gibi sistemin karakteristiğine uygun olarak herhangi bir üyelik fonksiyonu seçilir.



Şekil 5.2 Üyelik Fonksiyon Etiketleri

PK = Pozitif Küçük PO = Pozitif Orta PB = Pozitif Büyük
NK = Negatif Küçük NO = Negatif Orta NB = Negatif Büyük

Şekil 5.2'de örnek bir üyelik fonksiyonu ve bu fonksiyona bağlı dilsel değerler görülmektedir. Bu değerler sistemin karakteristiğine göre bir uzman tarafından belirlenir.

5.2. Bulanık Mantık Kavramları

Bulanık kümelerde, söz konusu evrenin elemanlarının bir A bulanık kümesine ait olma derecelerini temsil etmek amacıyla üyelik fonksiyonları kullanılır. Bulanık kümeler, üyelik fonksiyonları ile tanımlanır. Bu fonksiyonlar, elemanlara [0,1] kapalı aralığında gerçel değerler atayarak elemanların A bulanık kümesi ile temsil edilen kavrama ne derece uygun olduklarını veya A bulanık kümesi ile temsil edilen özellikleri ne derece taşıdıklarını gösterir. Bir A bulanık kümesinin üyelik fonksiyonu şu şekilde formulize edilir (Allahverdi 2002).

$$m_{A(x)} \rightarrow [0,1] \text{ , } x \in X$$

Burada x, X evrensel kümesinden alınan elemanları, $m_{A(x)}$ ise x elemanının A bulanık kümesine üyelik derecesini göstermektedir. $m_{A(x)}$ 'in değeri aşağıdaki gibi hesaplanır.

$$m_{A(x)} = \begin{cases} 1, \text{Eğer } x \in A \\ 0, \text{Eğer } x \in A \end{cases}$$

X = { x1, x2, … , xn } sonlu kümesi için A bulanık kümesi şöyle hesaplanır :

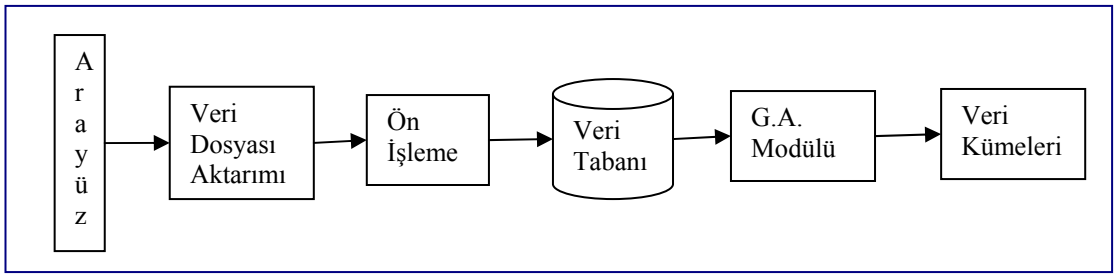
$$A = m_{A(x1)} / x1 + m_{A(x2)} / x2 + \ldots + m_{A(xn)} / xn = \sum_{i=1}^{n} m_{A(xi)} / xi$$

6. ÖRNEK ÇALIŞMA

Tıp alanında bulunan mevcut veri oldukça fazla ve hayati öneme sahiptir. Hastane bilgi sistemleri sayesinde bu veriler düzenli olarak tutulmaktadır. Bu verilerden daha fazla yaralanmak mümkündür. Hastane bilgi sistemlerinden veya diğer tıbbi veri toplayan sistemlerden alınan veriler üzerinde yapılan veri madenciliği çalışmaları hem uzmanlar için hem hastane yönetimi için hem de hastaların daha kaliteli bir hizmet almalarında etkin rol alabilir (Kaya, Bulun ve Arslan 2003).

Bu çalışma, GA kullanılarak kanser hastalarına ait veriler içinden sıklıkla karşılaşılan veri kümelerinin bulunması ile ilgilidir. T.C. Sağlık Bakanlığı Kanser Savaş Daire Başkanlığına bağlı olarak her ilden belirli dönemlerde kanser hastalarına ait veriler toplanmaktadır. Bu birim tarafından toplanan veriler daha sonra istatistiksel bilgiler üretmede kullanılmaktadır. Bu istatistiksel bilgiler coğrafi bölgelere,cinsiyete,yaşa ait kanser oranları gibi istatistiklerdir.Ancak bu birimde herhangi bir veri madenciliği uygulaması yapılmamaktadır. Bununla birlikte toplanan verilerin her ilden farklı yapıda gönderilmesi sebebiyle bu istatistikler uzun zamanda elde edilmektedir.

Çalışmada öncelikle verilere uygun bir veritabanı şeması oluşturulmuştur. Her ilden gelen farklı yapıdaki veri dosyalarından bilgiler, oluşturulan bu "kanser" veritabanı yapısına göre düzenlenmiş ve gereksiz verilerden (gürültülerden) arındırılmıştır. Daha sonra elde edilen yararlı bilgiler kanser veritabanına kaydedilmiştir. Bunu takiben Yapay Bağışıklık Sistemi tabanlı SAND algoritması yardımıyla hazırlanan başlangıç populasyonunu girdi alan Genetik Algoritma kullanılarak anlamlı bilgi kümelerinin bulunması amaçlanmıştır. Ayrıca GA modülünde Bulanık Mantıktan da yararlanılmıştır. Şekil 6.1 çalışmanın işleyiş modelini göstermektedir.



Şekil 6.1 Örnek Çalışmanın İşleyiş Modeli

6.1. Veri Kümesi ve Veritabanı

Çalışmada T.C. Sağlık Bakanlığı Kanser Savaş Daire Başkanlığı tarafından toplanan veriler kullanılmıştır. Her ilden kanser hastalarına ait veriler disket ortamında ve farklı yapılarda gelmektedir. Bu birim tarafından veriler ortak bir yapıya dönüştürülerek üzerinde çalışılmaktadır. Şekil 6.2 örnek bir veri dosyası bilgisini göstermektedir.

```
27 1
 ANTALYA SA   14   2  30  38   2   0   0 112 ANTALYA SAGLIK MUDURLUGU
 SIRANO       6   3  30  23   3  12   7 112 SIRANO       :
 ADI          6   4  30  23   4   1  15 112 ADI         :
 SOYADI       6   5  30  23   5   1  15 112 SOYADI      :
 DOGUMYERI    6   6  30  23   6   0   2 112 DOGUMYERI    :
 YASI         6   7  30  23   7   0   2 112 YASI         :
 CINSIYETI    6   8  30  23   8   0   1 112 CINSIYETI    :
 EGITIMDURU   6   9  30  23   9   0   1 112 EGITIMDURUMU  :
 MESLEGI      6  10  30  23  10   0   2 112 MESLEGI      :
 ADRESI       6  11  30  23  11   0   2 112 ADRESI       :
 ILCE        28  11  30  38  11   0   2 112 ILCE   :
 KOY         43  11  30  53  11   1  20 112 KOY    :
 SIGARA       6  12  30  23  12   0   1 112 SIGARA       :
 TERKYIL     28  12  30  38  12   0   2 112 TERK/YIL:
 ICMEYIL     43  12  30  53  12   0   2 112 ICME/YIL:
 TANEGUN     61  12  30  71  12   0   2 112 TANE/GUN:
 TANITAR      6  13  30  23  13  11  10 112 TANI/TAR     :
 TANIYONTEM   6  14  30  23  14   0   1 112 TANI/YONTEMI  :
 TUMYERI      6  15  30  23  15 101   4 112 TUMYERI      :
 TM          43  15  30  53  15   1  20 112 TM     :
 HISTTANI     6  16  30  23  16 101   6 112 HISTTANI     :
 HT          43  16  30  53  16   1  20 112 HT     :
 EVRE         6  17  30  23  17   0   1 112 EVRE         :
 TEDAVI      43  17  30  53  17   0   2 112 TEDAVI  :
 KONTROLTAR   6  18  30  23  18  11  10 112 KONTROLTARIHI  :
 HASTKODU    43  18  30  53  18 103   6 112 HASTKODU:
 KANBAGIDER   6  19  30  23  19   0   1 112 KANBAGIDER    :
    1RAMAZAN       ALAGOZ        483713 348                2  202017!
/04/2001416.9MˇDE              8145.3DˇFFUZADENOCA      2 512/06/2001 7.200!
0!
```

Şekil 6.2 Veri Dosyasından Bir Örnek

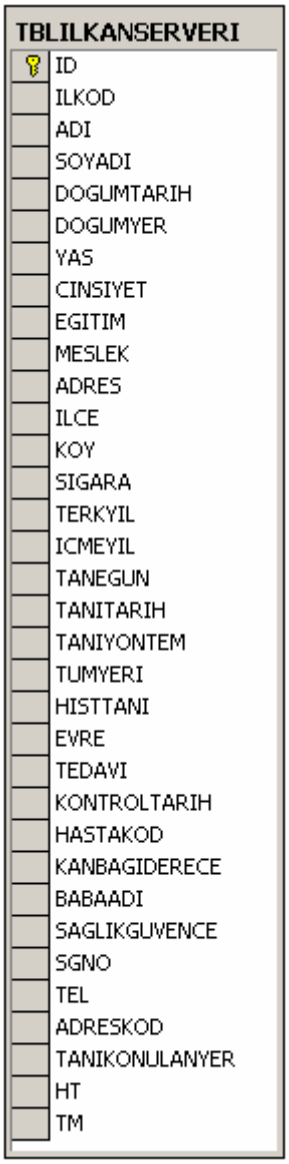
Veri Satırının İlgili Kısmı	Anlamı
IL	Verilerin alındığı il plaka kodu

YAS	Üyelik Fonksiyon Değerleri (Bölüm 6.2.3)	
	1	Bebek
	2	Çocuk
	3	Çok Genç
	4	Genç
	5	Orta Yaşlı
	6	Yaşlı
	7	Çok Yaşlı
TEDAVI	Hastaya uygulanan tedavi yöntemi	
	0	Tedavi
	1	Cerrahi
	2	Radyoterapi
	3	Kemoterapi
	4	Hormonterapi
	5	C+R
	6	C+K
	7	C+H
	8	R+K
	9	R+H
	10	K+H
	11	C+R+K
	12	C+R+H
	13	C+K+H
	14	R+K+H
	15	Hepsi
	99	Bilinmiyor
SIGARA	Hastanın sigara alışkanlığı	
	0	Hiç içmemiş
	1	İçmiş bırakmış
	2	Halen içiyor
	3	Bilinmiyor
EGITIM	Hastanın eğitim durumu	
	1	Eğitimsiz
	2	İlk Orta
	3	Lise
	4	Yüksek Okul
	9	Bilinmiyor
EVRE	Kanser hastalığının evresi	
	1	İnsute
	2	Lokalize
	3	Bölgesel Yaylım
	4	Metaztas
	9	Bilinmiyor
TANIYONTEM	Kanser hastalığı tespit edilme yöntemi	
	1	Radyoloji, Sitoloji,Hemotoloji

	2	Radyoloji
	3	Klinik, Cerrahi, Diğer
	4	Pataloji, Hematoloji, Sitoloji, Radyoloji
	5	Pataloji, Hematoloji, Sitoloji, Radyoloji, Diğer
	6	Radyoloji, Klinik, Cerrahi, Diğer
	7	Hepsi
	9	Bilinmiyor
MESLEK	Hastanın Mesleği	
	0	İşsiz
	1	İlmi, Teknik Eleman, Serbest Meslek
	2	Müteşebbisler (İşadamları, Direktörler,Yöneticiler)
	3	İdari Personel ve Benzeri Çalışanlar
	4	Ticaret ve Satış Personeli
	5	Hizmet İşlerinde Çalışanlar
	6	Tarım, Hayvancılık, Ormancılık, Avcılık Çalışanı
	7	Tarım Dışı Üretim Faaliyetleri Çalışanları
	8	Ev Hanımı
	9	Emekli
	10	Ağır İşlerde Çalışanlar
	99	Bilinmiyor

Tablo 6.1 Veri Dosyasının Çözümlenmesi

Şekil 6.2'de de görüldüğü gibi veri dosyalarında bulunan ham veri bir ön işleme tabii tutulup gürültülü verilerden arındırılır. Düzenlenen veriler ise Tablo 6.1'de olduğu gibi çözümlenerek veritabanına kaydedilir. Veritabanı şeması Şekil 6.3'de verilmiştir. Veritabanı tüm bilgileri tutan tek bir tablodan oluşmaktadır.



Şekil 6.3 Veritabanı Şeması

Veri kümesi yaklaşık 80 bin kayıttan oluşmaktadır. Veriler Ankara, İstanbul, Bursa, Elazığ ve Antalya illerine ait 2001,2002,2003 yılları kanserli hasta verileridir.

6.2. Genetik Algoritma

Bu çalışmada da genetik algoritmanın en iyiyi bulma özelliğinden faydalanılarak sık görülen veri gruplarının bulunması amaçlanmıştır. Bulunan veri grupları ise kanser hastalığı sebeplerine ışık tutmaktadır. Aranan veri grubuna göre

tasarlanan kromozomlar bizim için aday çözümleri temsil etmektedir. Geliştirilen uygunluk fonksiyonuna tabi tutulan kromozomlardan eşik seviyesini aşanlar çözüm uzayını belirtirler.

Genetik Algoritmalarda parametreler sıklıkla rasgele olarak belirlenmektedir. İnsan bağışıklık sistemini model alan Yapay Bağışıklık sisteminde ise bu rasgeleliğin önüne geçilebilmektedir. Öyleki her insan doğduğunda bir bağışıklık sistemine sahip olarak dünyaya gelir. Bu noktada Genetik Algoritmaların parametrelerini belirlemek için Yapay Bağışıklık Sistemlerinden yararlanılabilir. Bu çalışmada başlangıç populasyonu seçimi bir YBS modeli olan SAND yaklaşımı ile yapılmıştır.

Şekil 6.4 örnek bir kromozomu göstermektedir. Kromozomdaki her bir gen bir veri alanını ifade eder. Belirlenen genler guruplandırılabilir veri tiplerinden seçilmiştir (Örneğin cinsiyet gibi. Cinsiyet = Kadın veya Erkek). YAS geni kromozomda bireyin yaşının sayısal değerini değil dilsel yaş değerini tutmaktadır. Bu dilsel değerler Bölüm 6.2.3'de gösterilen üyelik fonksiyonunu göre üretilmektedir.

İL	DOĞUMYER	YAŞ	CİNSİYET	EĞİTİM	MESLEK	ADRES	SİGARA	TEDAVİ	TANIYÖNTEM	EVRE

Şekil 6.4 Bir Kromozomun Gösterimi

Ön işleme adımında veri dosyasından okunan her bir anlamlı kayıt veri tabanına kaydedilir. Veri alanları her bir dosyada farklı yapıda tutulmaktadır. Bu dosyalar çözümlenerek veriler "TBLILKANSERVERI" tablosuna aktarılır. Daha sonra veri madenciliği aşamasında bu tabloya başvurulur. Veritabanına kaydedilen her bir kayda ise kendisine özel olan bir "ID" bilgisi ile erişilir.

Çalışma için tasarlanan genetik algoritma aşağıdaki adımlardan oluşmaktadır:

Adım 1 : Başlangıç populasyonu seçimi için SAND yaklaşımı kullanılmıştır. Populasyon 100 bireyden oluşur.

Adım 2 : Her bir bireyin uygunluk değeri uygunluk fonksiyonuna göre hesaplanır.

Adım 3 (Seçim) : Bireyler rulet çarkı seçim metoduna göre seçilir. Seçilen 100 birey çiftleşme havuzuna atılır.

Adım 4 (Çaprazlama) : Çiftleşme havuzundan çaprazlama ihtimaline göre rasgele seçilen bireyler tek nokta çaprazlama işlemine tabi tutulurlar. Uygulamada çaprazlama noktası rasgele belirlenmiştir. Çaprazlama ihtimali ise %30 seçilmiştir.

Adım 5 (Mutasyon) : Yeni populasyonda çeşitliliği artırmak için nokta mutasyon işlemi uygulanmıştır. Mutasyon olasılığı uygulama %30 seçilmiştir.

Adım 6 (Değerlendirme) : Populasyondaki bireylerin uygunluk değerleri yeniden hesaplanır ve uygunluk eşiği geçen bireyler tespit edilmeye çalışılır. Eşik değerini aşamayan bireyler yeniden evrim sürecine katılmak üzere Adım 3'e gönderilir.

6.2.1. Başlangıç Populasyonu

Genetik Algoritmalarda kullanılan başlangıç populasyonu öğrenme hızında güçlü bir etkiye ve çözüm kalitesine sahiptir. Yetersiz bir başlangıç seçimi eğitim işleminin lokal minimuma takılmasına veya sayısal problemlerle karşılaşmaya neden olur. Bu çalışmada doğal bağışıklık sisteminden esinlenilerek oluşturulmuş bir metot olan yapay bağışıklık sistemi başlangıç populasyonu seçiminde önerilmiştir.

Başlangıç populasyonu ve çözüm kalitesi arasındaki korelasyon başlangıç antikor repertuarı ve doğal immün sistemin yanıtının kalitesi arasındaki korelasyona benzer. Bu kompleks bir örüntü tanıma aracı gibi görülebilir.

Bu çalışmada SAND olarak bilinen bir yapay bağışıklık modelinden yararlanılmıştır. Modelin ayrıntıları Bölüm 4.4.6'da anlatılmıştır. Bu yaklaşım hatayı minimize eden bir arama ile veri kümesi içinden en iyi başlangıç populasyon kümesini üretmeyi hedefler.

6.2.2. Uygunluk Fonksiyonu

GA'nın en önemli özelliklerinden biri uygunluk fonksiyonudur. Çözümlerin değerlendirme fonksiyonu olan uygunluk fonksiyonu GA'nın ana unsurudur. Bu uygulamada probleme uygun bir uygunluk fonksiyonu tasarlanmıştır. Uygunluk fonksiyonu Tablo 6.2'de verilmiştir.

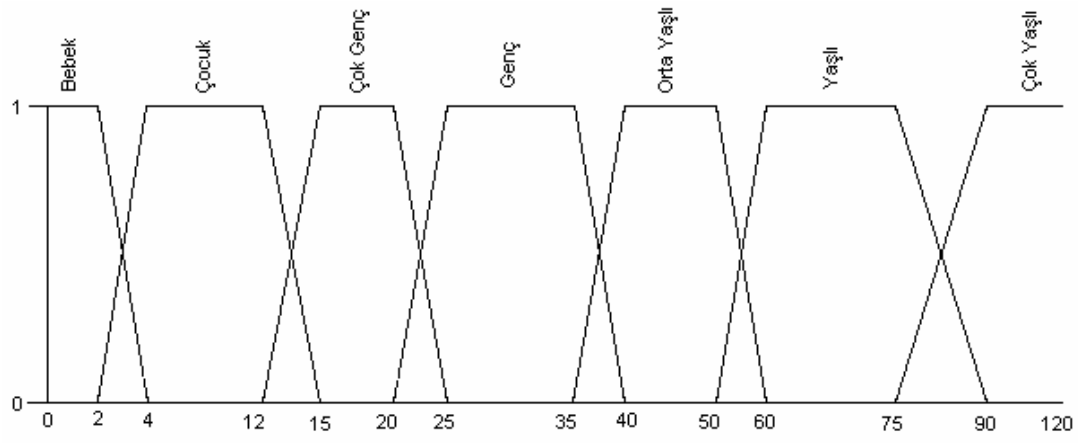
Destek = ∑ Bireyin veri kümesindeki tekrarlanma sayısı/ ∑ Veri kümesindeki toplam kayıt sayısı
Benzerlik Oranı : $\left(\sum_{n=1}^{11} 2^n \right.$ / n'li uygun gen sayısı)
m_{YAS} = YAS geni için üyelik değeri (Bkz. Bölüm 6.2.3)
Uygunluk Fonksiyonu = Destek * Benzerlik Oranı * m_{YAS} şeklindedir.

Tablo 6.2 Uygunluk Fonksiyonu

Uygulamada eşik değeri olarak %20 seçilir.

6.2.3. Verilerin Bulanıklaştırılması

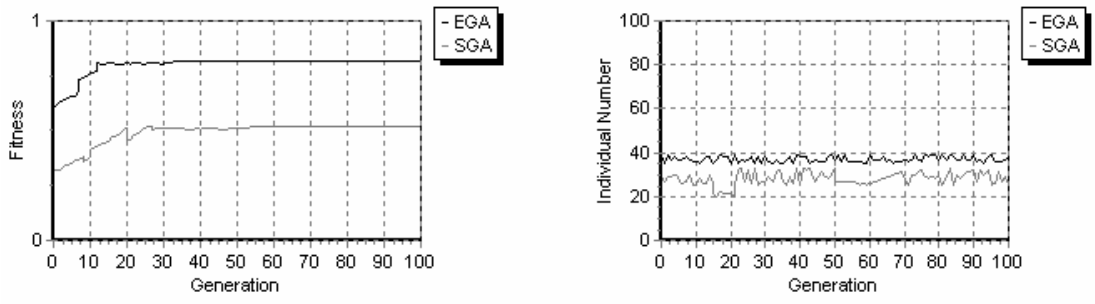
Çalışmada BM teorisinden de yararlanılmıştır. Bu sayede daha hassas sonuçlar üretilebileceği gösterilmeye çalışılmıştır. GA modülünde kullanılan populasyondaki bir bireyin yaşını ifade eden YAS geni Şekil 6.5'de görüldüğü gibi bulanıklaştırılmıştır (Delgado ve Ark. 2001). YAS geni için GA modülünde gerçek değerler değil bulanık değerler kullanılmıştır.



Şekil 6.5 YAS Geni İçin Üyelik Fonksiyonu

6.3. Uygulama Sonuçları

Uygulamada iki farklı uygunluk fonksiyonu kullanılmıştır. İlk uygunluk fonksiyonunun ayrıntıları Bölüm 6.2.1'de verilmiştir. İkinci uygunluk fonksiyonu ise karşılaştırma amaçlı kullanılan genel bir fonksiyondur. Bu fonksiyon veri madenciliğinden bilinen "Destek" değerini kullanmaktadır.



Şekil 6.6 Uygunluk Fonksiyonların Performans Karşılaştırmaları

(a) %20 eşik seviyesine göre her bir jenerasyonda elde edilen birey sayısı

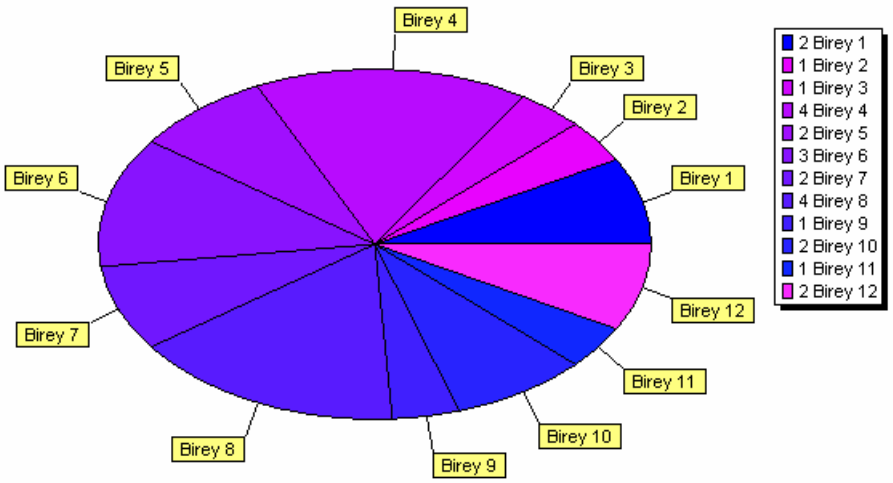
(b) %20 eşik seviyesine göre her bir jenerasyonda elde edilen en iyi bireyin uygunluğu

EGA : Uygulama için geliştirilen GA

SGA : Destek değerini kullanan GA

Şekil 6.6 uygunluk fonksiyonlarının performans karşılaştırmalarını göstermektedir. Şekil 6.6 (a) %20 eşik seviyesine göre her bir jenerasyonda elde edilen en iyi bireylerin sayısını göstermektedir. Bunun için GA 100 defa çalıştırılmıştır. Grafikte de görüldüğü gibi geliştirilen uygunluk fonksiyonu ile daha iyi ve tutarlı sonuçlar elde edilmiştir.

Şekil 6.6 (b) %20 eşik seviyesine göre her bir jenerasyonda elde edilen en iyi bireyin uygunluk değerini göstermektedir. Bunun için GA 100 defa çalıştırılmıştır. Grafikte de görüldüğü gibi geliştirilen uygunluk fonksiyonu , standart destek değerini kullanan GA'ya göre daha kısa sürede kararlı duruma geçmektedir.



Şekil 6.7 Uygunluk Fonksiyonunun 100 çalıştırma sonunda bulduğu veri grupları

	İL	D.YER	YAŞ	CİNSİYET	EĞİTİM	MESLEK	ADRES	SİGARA	TEDAVİ	T.YÖNTEM	EVRE
BİREY1	06	35	7	2	1	7	16	3	12	5	3
BİREY2	23	01	5	1	4	5	23	3	3	7	3
BİREY3	07	15	7	1	3	2	07	1	3	3	2
BİREY4	23	23	6	1	1	1	23	1	10	2	1
BİREY5	06	35	6	2	2	1	35	0	1	4	1
BİREY6	23	01	4	2	1	0	23	0	5	6	5
BİREY7	06	34	6	1	3	3	06	3	7	5	2
BİREY8	07	42	5	2	4	4	34	1	2	5	4
BİREY9	34	07	7	2	2	6	34	1	8	5	3
BİREY10	07	16	3	2	3	4	07	2	8	2	6
BİREY11	23	23	5	1	1	8	23	2	8	2	7
BİREY12	34	40	6	1	1	9	34	1	2	5	3

Tablo 6.3 Veri Gruplarının Açıklaması

Şekil 6.7 %20 eşik seviyesine göre uygulama için geliştirilen GA'nın 100 çalıştırma sonunda bulduğu veri gruplarını ve bu verilerle karşılaşma oranlarını göstermektedir. Burada daha sık görülen durumlar grafikte daha büyük paya sahiptir. Tablo 6.3'de ise bu veri gruplarının açıklaması görülmektedir.

7. SONUÇ VE GELECEĞE YÖNELİK ÇALIŞMALAR

Büyük boyutlardaki veri ile çalışan bir çok sektörde "veri" değil "enformasyon" ve "bilgi" ön plandadır. Veri madenciliği , veri depolama birimlerinde tutulan veriden anlamlı model ve kuralların oluşturulması işlemidir. Bu amaçla pek çok teknik mevcuttur ve daha verimli yeni tekniklerin geliştirilmesi halen araştırma konusudur.

Veri madenciliği bir çok disiplindeki teknikleri kullanır. Makine öğrenmesi, Yapay Zeka, İstatistik vs. alanlarıyla beraber çalışarak ilginç ve yararlı bilgi elde etmeyi amaçlar. Bu çalışmada Genetik Algoritmaların veri madenciliği üzerinde uygulaması anlatılmıştır.

GA doğadaki evrim sürecinin modellenmesini kullanarak problem uzayında çok yönlü ve adaptif bir arama ile en iyi çözümü bulmaktadır. Evrim, amacı belli olan ya da kontrollü bir süreç değildir. Kaynakların kısıtlı olduğu değişken bir ortamda, farklı genetik bilgilere sahip canlılar yarışırlar ve kabaca güçlü olanlar hayatta kalıp çoğalarak/üreyerek genetik bilgilerini bir sonraki kuşağa geçirme şansını yakalarlar. Zayıf olanlar ise elenirler.

Tıbbi veri tabanlarında yeni bilgilerin keşfedilmesi veritabanlarında depolanan verinin etkili bir şekilde kullanılması açısından çok önemlidir. Son zamanlarda tıbbi veritabanlarında depolanan veri miktarı büyük bir hızla artmaktadır. Bu büyümeye karşılık geleneksel veri analiz araçları yetersiz kalmaktadır. Verilerin analizi karar alma ve yönetim aşamalarında çok önemlidir. Sağlık sektöründe gizli bilgilerin keşfedilmesi hayati önem taşımaktadır. Örneğin semptomlar ve hastalıklar arasındaki gizli kalmış ilişkilerin bulunması, tecrübe ile elde edilebilecek bilgilerin önceden keşfedilmesi, çok faktörlü hastalıklarda önemli faktörler arasındaki ilişkilerin ortaya çıkarılması gibi bir çok alanda tıbbi veri madenciliğinin önemi tartışılmazdır.Tıp alanında bulunan mevcut veri oldukça fazla ve hayati öneme sahiptir. Hastane bilgi sistemleri sayesinde bu veriler düzenli olarak tutulmaktadır. Bu verilerden daha fazla yaralanmak mümkündür. Hastane bilgi sistemlerinden veya diğer tıbbi veri toplayan

sistemlerden alınan veriler üzerinde yapılan veri madenciliği çalışmaları hem uzmanlar için hem hastane yönetimi için hem de hastaların daha kaliteli bir hizmet almalarında etkin rol alabilir. Bu sebeplerden dolayı gelecekteki çalışmalarımı tıbbi veri madenciliği konusunda devam ettirmeyi düşünüyorum.

REFERANSLAR

Allahverdi N. , *Uzman Sistemler , Bir Yapay Zeka Uygulaması,* Atlas Yayın Dağıtım, Mart 2001.

Alli H. , Kaya M. , *Genetik Algoritma Kullanılarak PID Kontrol Parametrelerinin Bulunması*, Fırat Üniversitesi Fen ve Mühendislik Bilimleri Dergisi 13, 2, 1-8, 2001

Bagley J. D., *The Behavior of Adaptive Systems Which Emply Genetic and Correlation Algorithms,* PhD Thesis, University of Michigan, 1967.

Baysal B. , *Temel İmmünoloji Ders Notları,* Selçuk Üniversitesi, Eylül 2001.

Dasgupta D. (Editor). *Artificial Immune Systems and Their Applications,* A book published by Springer Verlag Inc., January 1999.

Dasgupta D., *Artificial Neural Networks and Artificial Immune Systems: Similarities and Differences,* In Proceedings of the IEEE International Conference on Systems, Man and Cybernetics (SMC '97), Orlando, October 12-15 1997.

Dasgupta D. (Editor), *Artificial Immune Systems and Their Applications,* A book published by Springer Verlag Inc., January 1999.

Dasgupta D., *Artificial Neural Networks and Artificial Immune Systems: Similarities and Differences,* In Proceedings of the IEEE International Conference on Systems, Man and Cybernetics (SMC '97), Orlando, October 12-15 1997.

Dasgupta D., Forrest S., *Novelty Detection in Time Series Data using Ideas from Immunology,* In ISCA 5th International Conference on Intelligent Systems, Reno, Nevada, June 19-21 1996.

Dasgupta D., Attoh-Okine N., *Immunity-based systems: A survey,* In Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics, pages 363-374, Orlando, Florida, October 12-15 1997.

de Castro L. ,Von Zuben F., *An Immunological Approach to Initialize Feedforward Neural Network Weights,* Proc. of ICANNGA, pp. 126-129, Prague/C.R., 22-25 April 2001.

de Castro L. ,Von Zuben F., *Artificial Immune Systems : Part I – Basic Theory and Applications,Technical Report TR – DCA 01/99* December, 1999.

de Castro L., Von Zuben F. , *Immune and Neural Network Models: Theoretical and Empirical Comparisons,* International Journal of Computational Intelligence and Applications (IJCIA), 1(3), pp. 239-257, 2001.

de Castro L., Von Zuben F. , *The Clonal Selection Algorithm with Engineering Applications,* Workshop on AIS and Their Applications, Las Vegas, USA

Delgado M., Sanchez D., Martin – Bautista M. J., Vila M. A., *Mining association rules with improved semantics in medical databases,* Artificial Intelligence in Medicine, 21, 241-245, 2001.

Forrest S., Perelson A. S., Allen L., Cherukuri R., *Self-nonself discrimination in a computer,* In Proceedings of the IEEE Symposium on Research in Security and Privacy, IEEE Computer Society Press, Los Alamitos, CA, pp. 202-212, 1994.

Forrest S., Javornik B., Smith R. E., Perelson A. S., *Using genetic algorithms to explore pattern recognition in the immune system,* In Evolutionary Computation 1:3, pp. 191-211, 1993.

Freeman K., Karr C., *Industrial Applications of Genetic Algorithms, Content: Data Mining Using Genetic Algorithms - Darrin J Marshall*, CRC Press, 1999.

Goldberg D. E., *Computer-Aided Gas Pipeline Operation Using Genetic Algorithms and Rule Learning*, PhD Thesis, The University of Michigan, 1983.

Goldberg D. E., *Genetic Algorithms in Search, Optimization, and Machine Learning,* Addison Wesley Longman, 1989

Groth R. , *Data Mining Building Competitive Advantage*, Prentice Hall PTR, 1999.

Holland J.H., *Adaptation in Natural and Artificial Syste*ms, MIT Press, 1975.

Kaya E., Bulun M., Arslan A., *Tıpta Veri Ambarları Oluşturma ve Veri Madenciliği Uygulamaları,* Akademik Bilişim 2003, Çukurova Üniversitesi,Adana, 3-5 Şubat 2003,

Kaelbling L. P., Littman M. L., Moore A. W. , *Reinforcement Learning : A survey,* Artificial Intelligence Research, Vol. 4, pp.237-285, 1996.

Kolen J.F., Pollack J.B., *Back Propagation is Sensitive to Initial Conditions,* Technical Report TR 90-JK-BPSIC, 1990.

Lee K. C., Kim J. S., Chung N. H., Kwon S. J., *Fuzzy cognitive map approach to web-mining inference amplification,* Expert Systems with Applications, 22, 197-211, 2002.

Mitchell M. , *An Introduction to Genetic Algorithms*, MIT Press, 1996.

Mitchell T. M., *Machine Learning,* The McGraw-Hill Companies, Inc. , 414 pages,1997.

Nasaroui O. , Gonzalez F. , Dasgupta D. , *The Fuzzy Artificial Immune System: Motivations, Basic Concepts, and Application to Clustering and Web Profiling*

Roiger R. J., Geatz M. W., *Data Mining, A Tutorial–Based Primer,* Addison Wesley, 2003.

SPSS, http://www.spss.com.tr/clementine.asp

Sutton R. S. , Barto A. G. , *Reinforcement Learning : An Introduction,* Cambridge, MA: MIT Pres., 1998.

Tuğ E., Şakiroğlu M., Bulun M., *Tıbbi Veri Tabanlarında Gizli Bilgilerin Keşfedilmesi,* 20. Bilişim Kurultayı, İstanbul, 2003.

Fritz A., Percy C., Jack A., Sobin L., Parkin D. M., Whelan S. Çeviri : Canda M. Ş., *UHS-0(Uluslararası Hastalık Sınıflandırması: Onkoloji) / ICD-O(International Classification of Diseases for Oncology),* DSÖ / WHO Kütüphanesi, 2000.

Zuo W. , *Control Theory Appication,* IEE Proc. Vol. 142, No. 2, March 1995.