SAKARYA UNIVERSITY JOURNAL OF COMPUTER AND INFORMATION SCIENCES
VOL. 2, NO. 3 DECEMBER 2019
DOI: 10.35377/saucis.02.03.648342

Detection of Crime Regions with Biclustering Approach and Comparison of Methods

Ibrahim CIL[1], Sümeyye Gizem ÇAKAR[2], Nazan SARI[3], Olcay EYDEMİR[4]
[1]Corresponding Author; Industrial Engineering Department, Sakarya University, Serdivan, Sakarya; icil@sakarya.edu.tr; https://orcid.org/0000-0002-1290-3704; +90264295671
[2]Industrial Engineering Department, Sakarya University, Serdivan, Sakarya; s.gizemcakar@gmail.com; https://orcid.org/0000-0003-0431-0236;
[3]Industrial Engineering Department, Sakarya University, Serdivan, Sakarya; nazansarii42@gmail.com; https://orcid.org/0000-0001-9667-8665;
[4]Industrial Engineering Department, Sakarya University, Serdivan, Sakarya; olcayeydemir@gmail.com; https://orcid.org/0000-0002-7119-344X

Received 18 November 2019; Revised 16 December 2019; Accepted 22 December 2019; Published online 31 December 2019

Abstract

In terms of safety of the social life, it is very important to foresee the crimes and take the necessary precautions before the crime is committed. For this purpose, crime analysis should be carried out in order for security units to take necessary measures. In this regard, the data mining approach makes a significant contribution to the security units in the analysis of large data. In this context, different data analysis methods are used to estimate and identify potential crime areas. By using dual clustering methods in the detection of crime zones, clustering of crime areas and crime types at the same time provides more comprehensive results than traditional clustering methods. In this study, CC and Xmotif algorithms were used on the data set of “Crimes in Boston” to determine the crime sites by using data mining approach. The results were obtained by using R-project 3.5.3 software. It was found that CC algorithm gives better results for this data set than Xmotif algorithm.

Keywords: Biclustering Method, CC Algorithm, Xmotif Algorithm, Crime Data, Data Mining

İkili Kümeleme Yaklaşımıyla Suç Bölgelerinin Tespiti ve İkili Kümeleme Yöntemlerinin Karşılaştırılması

Öz

Sosyal yaşamın güvenliği açısından, suç işlenmeden önce, suçların önceden öngörülmesi ve gerekli önlemlerin alınması oldukça önemli bir konudur. Bu amaçla güvenlik birimlerinin gerekli önlemleri alması için suç analizi yapılması gerekmektedir. Bu konuda veri madenciliği yaklaşımı güvenlik birimlerine büyük verilerin analizinde önemli bir katkı sağlamaktadır. Bu kapsamda potansiyel suç bölgelerinin tahmin edilerek belirlenmesinde farklı veri analiz yöntemleri uygulanmaktadır. Suç bölgelerinin tespitinde ikili kümeleme yöntemlerini kullanarak suçun işlendiği bölgeler ile suç türlerini aynı anda kümelemek, geleneksel kümeleme yöntemlerine göre daha kapsamlı sonuçlar sağlamaktadır. Bu çalışmada veri madenciliği yaklaşımı ile suç bölgelerini belirlemek için "Boston'daki Suçlar” veri seti üzerinde CC ve Xmotif algoritmaları kullanılmıştır. Elde edilen ikili kümelerin etkinliğini ölçmek amacıyla algoritmaların performansı Chia ve Karuturi ikili küme skoruna (CKSB) bakılarak karşılaştırılmıştır. Elde edilen sonuçlar R-project 3.5.3 yazılımı kullanılarak sağlanmıştır. Kullanılan bu veri seti için CC algoritmasının Xmotif algoritmasına göre daha iyi sonuçlar verdiği ortaya çıkmıştır.

Anahtar Kelimeler: İkili Kümeleme Yöntemi, CC Algoritması, Xmotif Algoritması, Suç Verisi, Veri Madenciliği

1. Giriş

Suç, siyasal, psikolojik, fiziki, ekonomik koşullar ile bölgesel faktörler arasındaki etkileşimin bir sonucu olarak ortaya çıkan, yasalara, ahlaka veya törelere aykırı davranışlar olarak tanımlanmaktadır [1]. Suç ve suçu meydana getiren unsurların birlikte incelenerek, aralarındaki ilişkinin tespit edilmesi suç analizi terimini meydana getirmiştir. Suç analizi yapılarak suç ve suçlu arasındaki ilişkinin ve suç bölgelerinin

tespiti, işlenme ihtimali yüksek olan suç türlerinin belirlenmesi ve suçların işlenmeden önce tahmin edilmesi, emniyet güçlerinin gerekli tedbirleri alması açısından önem arz etmektedir [2]. Bu bağlamda suç analizini, suç ve suça ait özelliklerin bilindiği verilerde gerçekleştirmekte önem kazanmaktadır.

Veri madenciliği analizi pek çok alanda önemli katkılar sağlamaktadır [3]. Diğer alanlarda olduğu gibi suçla ilgili verilerin analizinde de önemli sonuçlar ortaya çıkarılmaktadır [4]. Suç verileri, büyük boyutlu verilerdir. Bu verilerin ayrıntılı bir şekilde incelenerek analiz edilmesi yararlı bilgiler sunmakta ve yapılan tespitler ile alınan tedbirleri etkin kılmaktadır. Suç verilerinin analizi, yapılış amacına göre pek çok farklı yöntem ile gerçekleştirilebilmektedir. Büyük boyutlu verilerden anlamlı sonuçlar çıkardığı ve gelecekteki durumlara yönelik karar vermeyi sağladığı kümeleme tabanlı veri madenciliği yöntemleri, suç verilerinin analizinde sıkça kullanılmaktadır. Klasik kümeleme yöntemleri ile bir veri seti sadece satır ya da sadece sütunları dikkate alınarak gruplandırılabilmektedir. Ancak verileri, veri setinin satır ve sütunu aynı anda dikkate alarak gruplamak daha kapsamlı sonuçlar elde etmeyi sağlamaktadır. Bu tür kümeleme işlemine ikili kümeleme yöntemi denmektedir. Öyle ki bir suç verisinde, suçun işlendiği yer ile işlenen suç arasında bağlantı olabilmekte ve bunlar birleşerek suç kümelerini oluşturabilmektedir [1]. Bu sebeple ikili kümeleme yöntemlerini kullanarak, suçun işlendiği yerler ile suç türlerini aynı anda kümelemek suç bölgelerinin oluşturulmasına ve kümeler arasındaki benzerlik ve farklılıkların analiz edilmesine olanak sağlamaktadır [5, 6]. Bu yönü itibariyle güvenlik birimlerine suçları önleyici nitelikte, veriye dayalı etkin bir karar desteği sağlanabilecektir [7].

Literatürde ikili kümeleme yöntemi kullanılarak pek çok alanda çalışma yapılmıştır. Pontes vd., gen açıklama verilerini araştırma yönteminde, değerlendirme ölçütlerini kullanıp kullanmamalarına göre iki kategoriye ayırarak kapsamlı bir ikili kümeleme yaklaşımı sunmuştur [8]. Zhu vd., ikili kümeleme yöntemlerinin performansını geliştirmek amacıyla bulanık üye matrisi ve bulanık matematiği, çok amaçlı optimizasyon algoritmaları ile birleştirerek yeni bir algoritma önermiştir. Önerdikleri algoritmanın etkinliğini doğrulamak için üç gen/protein açıklama verisini ikili kümeleme yöntemlerinden NSGA2B, MOP- SOB ve PHMM algoritmaları ile önerdikleri algoritmaya uygulayarak karşılaştırmışlardır. Yeni algoritmanın daha iyi sonuçlar ürettiğini göstermişlerdir [9]. Arslan vd. ikili kümeleme yöntemlerinden Bimax algoritmasını kullanarak, kaçakçılıkta yakalanan suçluları yakalandıkları malzemenin türüne göre kümelenirmiştir. Çalışmada ikili kümeleme yöntemlerinden Bimax algoritmasını kullanarak suçluların ortak özellikleri belirlemişlerdir [10]. Kocatürk, ikili kümeleme algoritmalarından CC, Bimax, Plaid, Spectral, Quest ve Xmotif'i gen açıklama verileri üzerinde uygulamış, algoritmaların ürettiği sonuçları görsel ve sayısal açıdan karşılaştırmıştır [11]. Altunkaynak ve Örkcü, şehirlerde yaşanan suçların, türlerine göre gruplanmasında ikili kümele algoritmalarından Bimax'ı kullanarak suç bölgeleri arasındaki farklılıkları analiz etmiştir [1]. Singh ve Mehrotra, işbirlikçi filtreleme sistemlerinde kullanılan kümeleme yaklaşımlarının kapsamlı bir incelemesini yapmış, küme sayısının ile örtüşen derecenin öngörü ve öneri kalitesi üzerindeki etkilerini ikili kümeleme yöntemi kullanılarak belirlemiştir [12]. Dominguez ve Exposito, veri madenciliği yöntemlerinden ikili kümeleme yöntemini kullanarak, işletim sistemi araçlarını baz alarak, söz konusu ikili aramanın hızlandırılmasını amaçlamışlardır [13]. Gu ve Vaselkov, Nükleer Manyetik Spektroskopi (NPS) ve Kütle Spektrometresi (MS) teknolojileri kullanarak, biyolojik numunelerden elde edilen veriler üzerinde numunelerin alt kümelerinde benzer davranış sergileyen metabolitleri belirlemek amacıyla ikili kümeleme yöntemini uygulamıştır [14]. Kocatürk ve Altunkaynak, ikili kümeleme algoritmalarının performanslarını karşılaştırmak amacıyla gerçek veri setlerini kullanarak çalışma yapmıştır. Gerçek veri setlerinden maya, insan ve fare verilerini kullanarak, ikili kümeleme yöntemlerinden CC, Bimax, Plaid, Quest, Spectral ve Xmotif algoritmalarına uygulamışlardır. Elde ettikleri sonuçları farklı değerlendirme ölçütleri ve ısı grafiklerini kullanarak karşılaştırmışlardır [15].

Suç ve suçu meydana getiren faktörleri bir arada incelenip, aralarındaki ilişkiyi tespit etmek ve çıkarımda bulunmak için pek çok amaçla suç analizi ile ilgili çalışma yapılmıştır. Subhashini ve Milani, insanların suç işleme oranlarının artması nedeniyle; teknolojiden istifade ederek, Coğrafi Bilgi Sistemi (GIS) tabanlı bir uygulama ile suç oranının aşağıya düşürülmesine katkı sağlamıştır [16]. Jalil vd., suç analizi için bilgi sunum modelini oluşturmuştur [17]. Oguntunde vd., Nijerya'da toplanan verilere dayanarak, suç modelini ve oranını göstermek amacıyla çeşitli suç türleri arasında var olan ilişkileri göstermiştir [18]. Alves vd., suç analizleri sonucunda korelasyon ve regresyon analizlerini kullanarak kentlerdeki suç tahmin oranını belirlemiştir [19]. Quick vd., çok değişkenli mekânsal modeller ile suçlar

arasındaki korelasyon yapılarını doğrudan ölçmek ve çoklu suç tipleri arasında paylaşılan temel suç modellerini ortaya çıkarmak için birçok model sunmuştur [20].

Bu çalışma, suçun işlendiği bölgeler ile suç türlerini aynı anda kümeleyerek, suç bölgelerini tespit etmek amacıyla yapılmıştır. Çalışmada, cinayet, hırsızlık, dolandırıcılık, insan kaçakçılığı, haneye tecavüz ve taciz gibi toplam 52 suç türü ele alınmıştır. Suç bölgelerinin tespit edilmesi için ikili kümeleme yöntemlerinden CC ve Xmotif algoritmaları gerçek veri setine uygulanmıştır. Gerçek veri seti uygulamasında "Crimes in Boston" veri seti kullanılmıştır. Elde edilen ikili kümelerin etkinliğini ölçmek amacıyla algoritmaların performansı CKSB' ye bakılarak karşılaştırılmış ve sonuçlar elde edilmiştir.

Makalenin kalan kısmı şöyledir. İkinci bölümde çalışmada kullanılan yöntem detaylıca anlatılmış, üçüncü bölümde gerçek veri seti üzerinde yapılan uygulama anlatılmış ve dördüncü bölümde çalışmadan elde edilen sonuçlar paylaşılmıştır.

2. Materyal ve Metot

Veri seti, gözlem birimlerini ifade eden satırlar ile değişkenleri ifade eden sütunlardan oluşan iki boyutlu matristir[21]. Çalışılan bir veri setinin sadece satır ya da sadece sütuna göre gruplandırılması kümeleme yöntemi olarak tanımlanmaktadır. Klasik kümeleme yöntemlerinin aksine satır ve sütunların aynı anda kümelenmesi işlemine ikili kümeleme yöntemi denmektedir.

İkili kümeleme yöntemi, gen açıklama verilerinin analizi ile ortaya çıkmış bir kavramdır. Bir gen açıklama verisi genlere karşılık gelen satırlar ile koşullara karşılık gelen sütunlardan oluşan çok boyutlu matrisi ifade etmektedir [10]. Gen açıklama verilerinin analizinde gen ve koşulların aynı anda kümelenmesi ile birlikte çok sayıda genin hücre ve dokulardaki ifade düzeyleri eş zamanlı olarak incelenebilmektedir [22]. Bir ikili kümeleme matrisinin yapısı Eş.1'deki gibi gösterilmektedir.

$$B = \begin{pmatrix} b_{11} & \cdots & b_{1|J|} \\ \vdots & \ddots & \vdots \\ b_{|I|1} & \cdots & b_{|I||J|} \end{pmatrix} \tag{1}$$

$g_i = \{b_{i1}, b_{i2}, \dots, b_{i|J|}\}$ olmak üzere g_i, i. satırdaki genleri, $c_j = \{b_{1j}, b_{2j}, \dots, b_{|I|j}\}$ olmak üzere c_j, j. sütundaki koşulları ifade etmektedir. Eş.1'de yer alan ikili küme elemanlarının hesabına yönelik formüller Eş.2,3 ve 4'te verildiği gibidir.

$$b_{Ij} = \frac{1}{|I|}\sum_{i=1}^{|I|} b_{ij} \tag{2}$$

$$b_{iJ} = \frac{1}{|J|}\sum_{j=1}^{|J|} b_{ij} \tag{3}$$

$$b_{IJ} = \frac{1}{|I||J|}\sum_{i=1}^{|I|}\sum_{j=1}^{|J|} b_{ij} \tag{4}$$

Yukarıdaki eşitliklerde $|I|$ satır, $|J|$ sütun değeri olmak üzere, b_{Ij} j. sütunun ortalamasını, b_{iJ} i. satırın ortalamasını ve b_{IJ} B matrisinin genel ortalamasını ifade etmektedir.

Literatürde, gen verilerinin sergilediği desene bağlı olarak, ikili kümeleme türlerinin çeşitli şekillerde tanımlanması ve sınıflandırılması yapılmıştır. Bu türlerin bazıları için ikili kümelerin değerlerini denklem ile göstermek mümkündür. π, B matrisi için herhangi bir sabit değeri ifade etmek üzere, $\beta_i (1 \leq i \leq |I|)$ ve $\beta_j (1 \leq j \leq |J|)$ her bir i geni ya da j koşulu toplamsal modellerde kullanılan sabit değerlere, $\alpha_i (1 \leq i \leq |I|)$ ve $\alpha_j (1 \leq j \leq |J|)$ ise bir i geni ya da j koşulu çarpımsal modellerde kullanılan sabit değerlere karşılık gelmektedir. Buna göre ikili kümeleme türleri sabit değerli ikili küme, satır ya da

sütunlarda sabit değerli ikili küme, toplamsal ve çarpımsal model ikili küme başlıkları ile kategorize edilebilir.

Sabit değerli ikili kümeler, bir koşul kümesinde benzer ifade değerleri olan gen alt kümelerine karşılık gelmektedir. Bu durum $b_{ij} = \pi$ eşitliği ile gösterilmektedir. Satır ya da sütunlarda sabit değerli ikili küme, koşullar ya da genler kümesinde benzer ifadelere sahip gen ya da koşul alt kümesine karşılık gelmektedir. Toplamsal ve çarpımsal olarak ifade edilebilirler. Toplamsal modele göre ikili küme elemanları $b_{ij} = \pi + \beta_i$ veya $b_{ij} = \pi + \beta_j$ eşitliklerine göre değer alırken, çarpımsal modele göre $b_{ij} = \pi.\alpha_i$ veya $b_{ij} = \pi.\alpha_j$ eşitliklerine göre değer almaktadır. Toplamsal ve çarpımsal modeli bir arada bulunduran ikili kümenin elemanları $b_{ij} = \pi_i.\alpha_i + \beta_i$ eşitliğine göre değer almakla birlikte toplamsal ve çarpımsal ikili kümeleme modeli olarak ifade edilmektedir.

İkili küme türlerinin farklılık göstermesi sebebiyle geliştirilen birçok ikili kümeleme algoritması vardır. Bu çalışmada suç bölgelerini tespit etmek amacıyla CC ve Xmotif algoritmaları kullanılmıştır.

3.1 CC Algoritması

CC algoritması, Cheng ve Church tarafından literatüre kazandırılmıştır [19]. Genlerin koşullar üzerinde tutarlı dalgalanmalar gösterdiği veri matrisinden alt matrisler elde etmek için oluşturulmuştur. Bu algoritmanın matematiksel modeli ikili kümelerin Ortalama Karesel Artık Skoru' nu (MSR, H) en küçük yapmak üzere kurulmuştur. MSR, ikili kümenin tüm elemanlarının varyansını ifade etmektedir. MSR'nin değeri küçüldükçe satır ve sütun değerlerinin tutarlı, ikili kümelerin daha anlamlı olduğu sonucuna varılır [15]. Bu değerin hesabı Eş.5'te verildiği gibidir.

$$H(I,J) = \frac{1}{|I||J|} \sum_{i \in I, j \in J} (b_{ij} - b_{iJ} - b_{Ij} + b_{IJ})^2 \tag{5}$$

CC algoritması, sezgisel yaklaşımla satır ve sütunları yinelemeli olarak güncelleyerek ilerleyen, iteratif bir algoritmadır. Algoritmanın ilk adımında veri matrisinden MSR H değeri hesaplanarak MSR H '> αH ile birkaç satır ve sütun silme işlemi yapılmaktadır. Burada α çoklu düğüm silme için belirlenen eşik değeridir. İkinci adımda MSR H'nin değerini azaltabilmek için satır veya sütunlar tek tek silinmeye devam edilmektedir. Üçüncü adımda ise daha iyi sonuçlar elde etmek ve ikili kümenin boyutunu artırmak amacıyla MSR H değeri δ değerini geçmeyecek şekilde satır veya sütun ekleme işlemi yapılmaktadır. $\delta \geq 0$ olmak üzere en büyük kabul edilebilir ortalama kare artığı skoruna karşılık gelmektedir. İkili kümenin boyutu sabit kaldığında algoritma sona ermekte ve veri matrisi bir sonraki iterasyon için güncellenmektedir.

2.2 Xmotif algoritması

Murali ve Kasif tarafından önerilen Xmotif algoritması, bir veri matrisinde benzer satır ve sütunlardan oluşan en büyük boyutlu alt kümeleri bulmayı amaçlamaktadır [23]. Tekrarlı yapıda olan Xmotif algoritmasının çalışma süreci, veri matrisindeki en büyük boyutlu alt kümenin bulunması, bulunan alt kümenin veri matrisinden çıkarılması, kalan veri matrisindeki en büyük boyutlu alt kümenin bulunması şeklinde devam etmektedir. Tüm alt kümelerin bulunmasıyla birlikte algoritma sonlanır ancak tüm satır ve sütunlar alt kümelere dahil olmak zorunda değildir [15].

2.3 MSR Skoru

MSR sık kullanılan bir yaklaşım olmakla birlikte ikili kümelerin anlamlılığını, satır ve sütun değerlerinin tutarlılığına bakarak ölçmeye yarayan bir değerlendirme ölçütüdür [24]. MSR'nin değeri küçüldükçe ikili kümenin anlamlılığı artmaktadır. Bu ölçütün hesaplandığı denklem Eş.6'da verildiği gibidir.

$$MSR(B) = \frac{1}{|I||J|}\sum_{i=1}^{I}\sum_{j=1}^{J}(b_{ij} - b_{iJ} - b_{Ij} + b_{IJ})^2 \quad (6)$$

2.4 Chia ve Karuturi İkili Kümeleme Skoru

İkili küme içinde bulunan ve bulunmayan sütun elemanları arasındaki benzerliğe göre ikili kümenin anlamlılığını ölçen bu değerlendirme ölçütü Chia ve Karuturi tarafından önerilmiştir [25]. MSR skoruna göre ikili kümelerin satır ve sütun etkileri bulunarak CKSB hesaplanır. CKSB'nin değeri büyüdükçe ikili kümenin anlamlılığı artmaktadır [15]. Satır $(T_k(b))$ ve sütun $(B_k(b))$ etkileri ile CKSB'nin hesaplanma denklemi sırasıyla Eş.7, 8 ve 9'da verildiği gibidir.

$$T_k(b) = \frac{1}{|I(b)|}\sum_{i=1, i\in b}^{I(b)} b_{iJk}^2 - \frac{MSR_k(b)}{J_k(b)} \quad (7)$$

$$B_k(b) = \frac{1}{|J(b)|}\sum_{j=1, j\in b}^{J(b)} b_{Ijk}^2 - \frac{MSR_k(b)}{I(b)} \quad (8)$$

$$CKSB(b) = LOG\left(\frac{\max(T_1(b) + \alpha,\ B_1(b) + \alpha)}{\max(T_2(b) + \alpha,\ B_2(b) + \alpha)}\right) \quad (9)$$

Eşitliklerde görülen k indisinin değeri, ikili kümede bulunan eleman için 1, bulunmayan eleman için 2 olmaktadır. $I(b)$ ikili kümedeki satır sayısına karşılık gelen değer olmakla birlikte $J_k(b)$ ikili küme içindeki ve dışındaki sütun sayısına karşılık gelmektedir. α değeri 0 ile 1 arasında değişkenlik göstermektedir [15].

3. CC ve Xmotif Algoritmaları İle Suç Bölgelerinin Tespit Edilmesi

Bu çalışmada, suçun işlendiği yerler suç türlerine göre gruplandırılarak, suç bölgelerinin tespit edilmesi problemi çözülmüştür. Çalışmada ikili kümeleme yöntemlerinden CC ve Xmotif algoritmaları kullanılarak gerçek veri setine uygulanmıştır. Gerçek veri seti uygulamasında "Crimes in Boston" veri seti kullanılmıştır. Verilerin ikili kümeleme işlemi R-project 3.5.3 yazılımı kullanılarak gerçekleştirilmiştir. Elde edilen ikili kümelerin etkinliğini ölçmek amacıyla algoritmaların performansı Chia ve Karuturi ikili küme skoruna (CKSB) bakılarak karşılaştırılmıştır.

"Crime in Boston", 2015-2018 yılları arasında Boston'da meydana gelen 140,140 adet suç kaydından oluşan bir veri setidir. Veri setinde suç türleri, suçun meydana geldiği sokak adları ve işlendiği zaman ile ilgili bilgiler yer almaktadır. Bu veri setiyle ikili kümeleme yapmak amacıyla veri matrisi oluşturulmuştur. Oluşturulan veri matrisinin satırları suçların meydana geldiği sokak adlarına, sütunları ise suç türlerine karşılık gelmektedir ve 2,695 sokak ile 52 suç türü mevcuttur. Yani, ikili kümeleme analizi için 2695x52 boyutunda bir veri matrisi kullanılmıştır. Matris elemanları i. sokakta meydana gelen j. suçun adetini göstermektedir. Tablo 1'de 2015-2018 yılları arasında Boston'da meydana gelen başlıca suç türleri kod ve tanımlarıyla birlikte verilmiştir.

Tablo 1 Suç Kod ve Tanımları

Suç Kodu	Suç Tanımı	Suç Kodu	Suç Tanımı
X111	Cinayet	X2612	Yangına İlişkin Kayıtlar
X112	Adam Öldürme	X2618	Patlayıcılar
X301	Hırsızlık	X2628	Telefon Görüşmesi Şikâyetleri
X402	Ağır Ceza Gerektiren Saldırı	X2629	Taciz
X511	Konut Hırsızlığı	X2642	Kaçmak (Okul vb.)
X530	Ticari Hırsızlık	X2648	Asılsız Bomba İhbarı
X670	Plaka İle İlgili Olaylar	X2662	Balistik
X706	Oto Hırsızlığı	X2670	Cezai Taciz

X801	Basit Saldırı	X2672	Biyolojik Tehdit
X900	Yangın Çıkarma	X2905	İhlaller
X1001	Dolandırıcılık	X3001	Tıbbi Yardım
X1201	Yolsuzluk	X3106	Mala Yönelik Zarar
X1402	Vandallık	X3109	Polis Teşkilatı Olayları
X1501	Ateşli Silah İhlalleri	X3111	Ruhsat İhlali
X1504	Diğer	X3112	Ev Sahibi/Kiracı Anlaşmazlıkları
X1601	Fuhuş	X3116	Limanla İlgili Olaylar
X1610	İnsan Kaçakçılığı	X3119	Ateşli Silah Keşfi
X1805	İlaç İhlalleri	X3125	Tutuklama Emri
X1901	Kumar	X3130	Arama İzni
X2003	Çocuğa/Aileye Karşı Suçlar	X3201	Kayıp Eşya
X2006	Yasak İhlalleri	X3301	Sözel Anlaşmazlıklar
X2010	Haneye Tecavüz	X3302	Toplanma veya Toplama İhlalleri
X2101	Baskı Altında Çalışmak	X3402	Hayvan Olayları
X2204	İçki İhlali	X3410	Motorlu Araçların Çekilmesi
X2401	Ahlaka Aykırı Davranış	X3502	Bulunan Kayıp Kişi
X2606	Mahkumla İlgili Olaylar	X3801	Trafik Kazaları

Tablo 2'de kümeleme analizi yapılacak olan veri setinin yapısı gösterilmiştir.

Tablo 2 İkili Kümeleme Veri Matrisi Yapısı

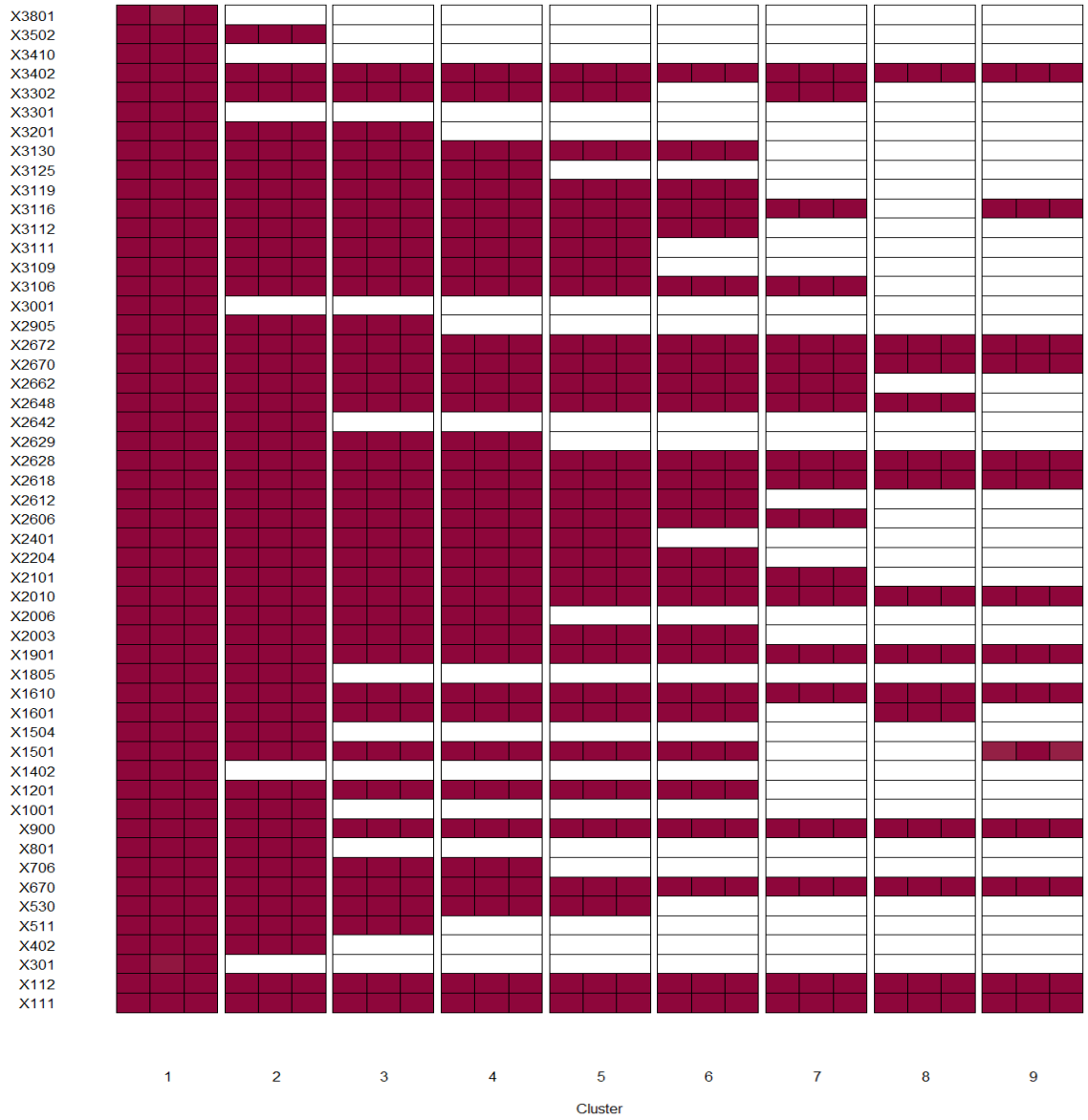
		Suç Türü				
		X111	X112	…	X3502	X3801
Sokak Adı	54	0	0	…	0	0
	15 1	0	0	…	0	0
	.	.	.		.	.
	.	.	.	.	.	.
	.	.	.		.	.
	Zeigler St.	0	0	…	0	5
	Zeller St.	0	0	…	0	1

Oluşturulan veri matrisi CC algoritmasına uygulanarak algoritmanın ürettiği sonuçlar kaydedilmiştir. Elde edilen sonuçlara göre 9 adet ikili küme bulunmuştur. Her bir ikili kümenin boyutu ve içerdiği sokak sayısının oranı Tablo 3' te verilmiştir.

Tablo 3 İkili Kümelerin Boyutları

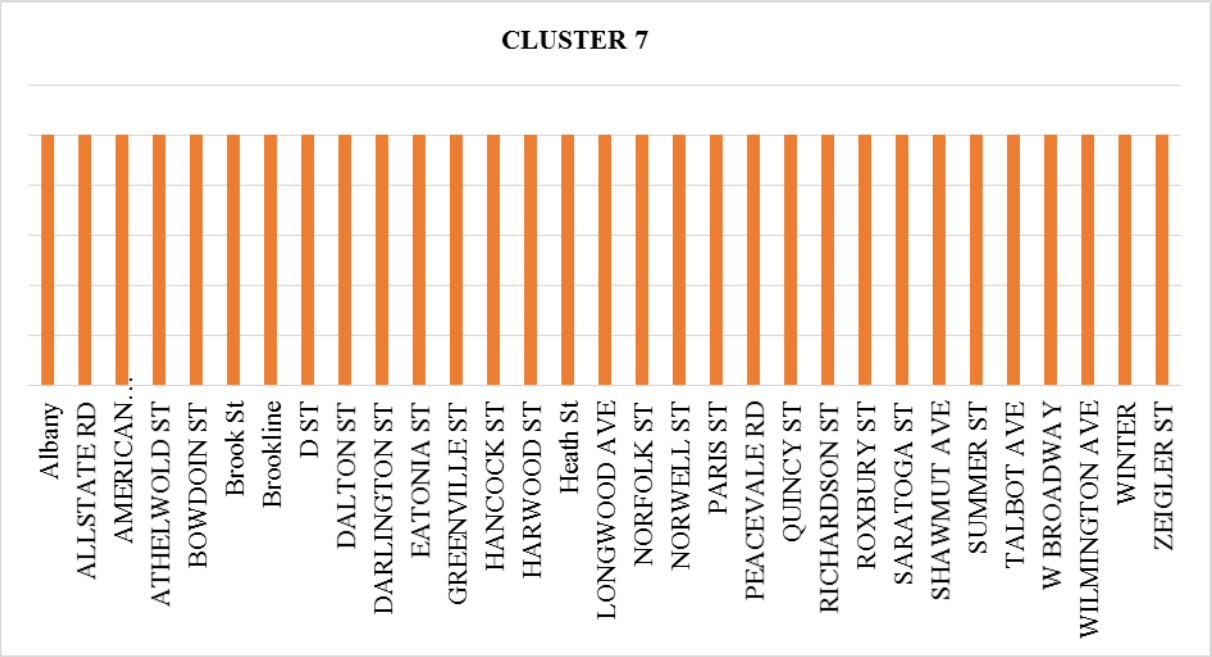
İkili Küme	Boyut	Oran (%)
1	1966x52	72.95
2	209x46	7.75
3	194x39	7.2
4	115x36	4.27
5	90x32	3.34
6	72x27	2.67
7	31x19	1.15
8	16x14	0.59
9	2x14	0.12

Tablo 3'te de görüldüğü gibi en büyük ikili küme %72.95 oranla 1.küme olurken, en küçük ikili küme %0.12 oranla 9. küme olmuştur. Her bir ikili kümenin içerdiği suç türüne ait üyelik grafiği Şekil 1'de verilmiştir. İkili kümelerin satır boyutları büyük olduğundan örnek olması amacıyla 7. ikili kümenin içerdiği sokaklar adlarıyla birlikte Şekil 2'deki gibi verilmiştir.



Şekil 1 CC Algoritmasına Göre İkili Küme Üyelik Grafiği

Algoritma sonucunda elde edilen ikili kümelerin boyutlarına ve üyelik grafiğine bakıldığında 1.ikili kümenin tüm suç türlerini kapsadığı görülmektedir. 1.ikili küme tüm suç türlerini kapsadığı için diğer kümeler bu kümenin alt kümesi olmaktadır. Bu sebeple 1.ikili kümede yer alan suç türlerinin önlenmesine yönelik tedbirler diğer kümeler için de alınabilir. Diğer ikili kümeler incelendiğinde X301, X1402, X3001, X3301, X3410 ve X3801 kodlarına karşılık gelen, "Hırsızlık", "Vandallık","Tıbbi Yardım", "Sözel Anlaşmazlıklar", "Motorlu Araçların Çekilmesi" ve "Trafik Kazaları" suçlarının yalnızca 1.ikili kümeye ait olduğu sonucuna varılmıştır. İkili kümeler kendi aralarında karşılaştırıldığında 7. ikili kümenin 6. ikili kümeden farklı olarak X3302 kodlu "Toplanma ve Toplama İhlalleri" suçunu içerdiği görülmektedir. X1601 koduna karşılık gelen "Fuhuş" suçu 7. ikili kümede bulunmazken 8. ikili kümenin elemanı olduğu görülmüştür. 9. ikili küme en küçük ikili küme olmasına rağmen 7. ve 8. ikili kümede bulunmayan X1501 kodlu "Ateşli Silah İhlalleri" suçunu içermektedir.



Şekil 2 CC Algoritmasına Göre 7. İkili Kümeye Ait Sokak Adları

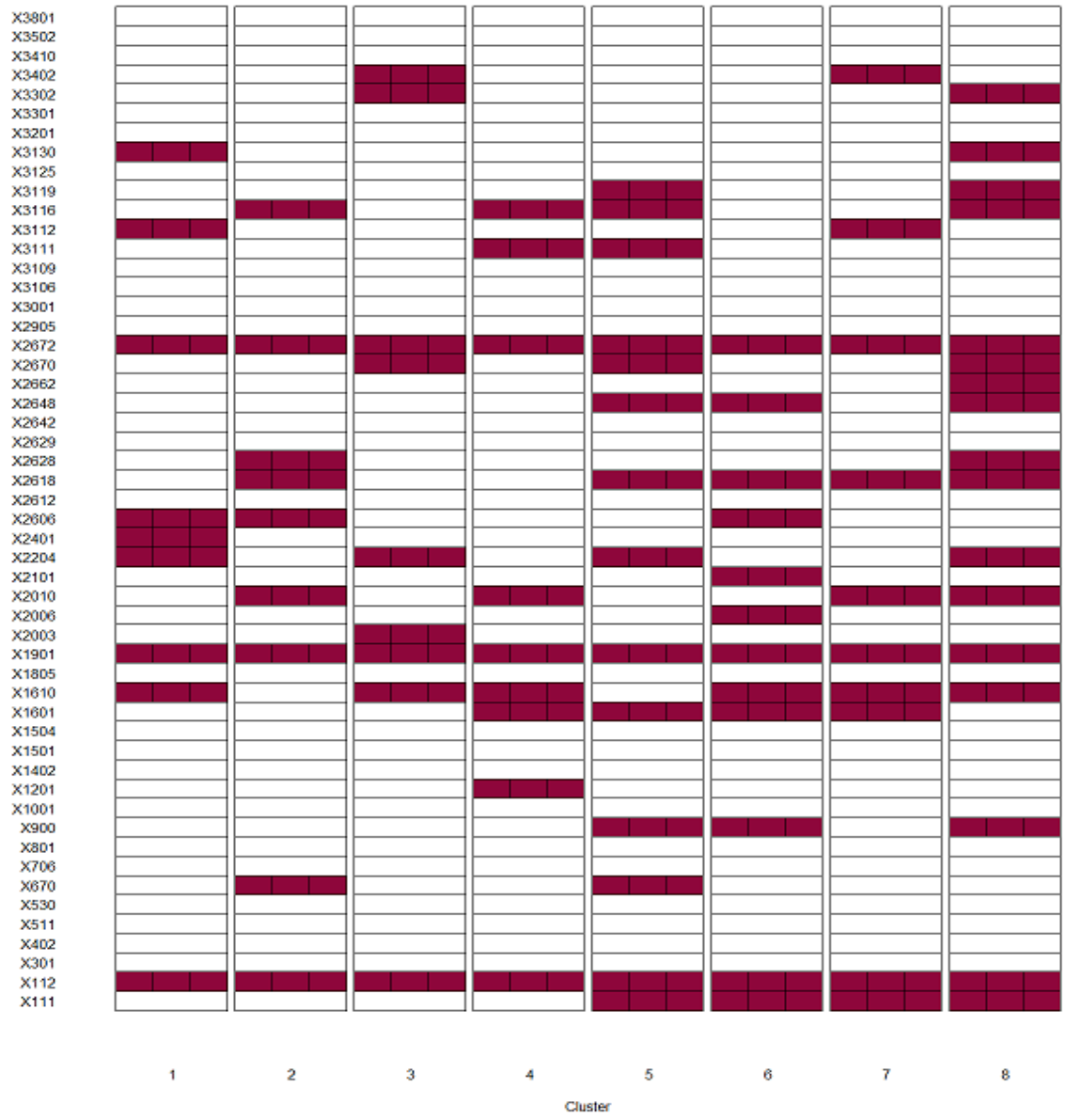
Şekil 1 ve Şekil 2'ye bakılarak, CC algoritmasına göre elde edilen 7. ikili kümeye dahil olan sokaklarda, 7. ikili kümenin içerdiği cinayet, adam öldürme, plaka ile ilgili olaylar, yangın çıkarma, insan kaçakçılığı, kumar, haneye tecavüz, baskı altında çalışmak, mahkumla ilgili olaylar, patlayıcılar, telefon görüşmesi şikayetleri, asılsız bomba ihbarı, balistik, cezai taciz, biyolojik tehdit, mala yönelik zarar, limanla ilgili olaylar, toplanma veya toplama ihlalleri ve hayvan olayları suçlarına karşı önlem alınması gerektiği sonucuna varılmıştır.

Oluşturulan veri matrisi Xmotif algoritmasına uygulanarak algoritmanın ürettiği sonuçlar kaydedilmiştir. Elde edilen sonuçlara göre 8 adet ikili küme bulunmuştur. Her bir ikili kümenin boyutu ve içerdiği sokak sayısının oranı Tablo 4' te verilmiştir.

Tablo 4 İkili Kümelerin Boyutları

İkili Küme	Boyut	Oran (%)
1	1935x9	71.8
2	504x9	18.7
3	103x9	3.82
4	78x9	2.89
5	8x14	0.29
6	8x12	0.29
7	13x10	0.48
8	2x17	0.074

Tablo 4'te de görüldüğü gibi en büyük ikili küme %71.8 oranla 1.küme olurken, en küçük ikili küme %0.074 oranla 8. küme olmuştur. Her bir ikili kümenin içerdiği suç türüne ait üyelik grafiği Şekil 3'teki gibidir. İkili kümelerin satır boyutları büyük olduğundan örnek olması amacıyla 7. ikili kümenin içerdiği sokaklar adlarıyla birlikte Şekil 4'teki gibi verilmiştir.



Şekil 3 Xmotif Algoritmasına Göre İkili Kümeleme Üyelik Grafiği

Algoritma sonucunda elde edilen ikili kümelerin boyutlarına ve üyelik grafiğine bakıldığında en fazla suç türünün 8.ikili kümeye ait olduğu görülmektedir. Diğer ikili kümeler birbirleriyle karşılaştırıldığında X2401 kodlu "Ahlaka Aykırı Davranış" suçunun yalnızca 1. ikili kümeye, X2003 kodlu "Çocuğa/Aileye Karşı Suçlar" adlı suçun yalnızca 3. ikili kümeye ait olduğu ve X1201 kodlu "Yolsuzluk" suçunun yalnızca 4. İkili kümeye ait olduğu sonucuna varılmıştır.

Şekil 3 ve Şekil 4'e bakılarak, Xmotif algoritmasına göre elde edilen 7. ikili kümeye dahil olan sokaklarda, 7. ikili kümenin içerdiği cinayet, adam öldürme, fuhuş, insan kaçakçılığı, kumar, haneye tecavüz, patlayıcılar, biyolojik tehdit, ev sahibi/kiracı anlaşmazlıkları ve hayvan olayları suçlarına karşı önlem alınması gerektiği sonucuna varılmıştır.



Şekil 4 Xmotif Algoritmasına Göre 7. İkili Kümeye Ait Sokak Adları

Veri seti her iki algoritmaya da uygulanarak maksimum büyüklükteki ikili kümeler elde edilmiştir. Algoritmaların sonucuna göre elde edilen ikili kümelerin etkinliğini ölçmek amacıyla Chia ve Karuturi ikili küme skoru (CKSB) hesaplanmıştır. CC ve Xmotif algoritmaları için CKSB değerleri Tablo 5'te verilmiştir.

Tablo 5 CC ve Xmotif Algoritmalarının CKSB Değerlerine Göre Karşılaştırılması

Algoritma	Tscore1	Bscore1	Tcore2	Bscore2	SBscore	TScore
CC	54.72253	352.022	74581.18	207823.1	**-6.38072**	-1.024799
Xmotif	0	0	634.6469	2320.307	-12.35463	-1.296375

Tablo 5'ten de görüldüğü gibi CC algoritmasının SBscore değeri Xmotif algoritmasına göre daha yüksek çıkmıştır. Bu da "Crime in Boston" veri seti için CC algoritmasının performansının daha iyi olduğunu göstermektedir. CC algoritmasının ürettiği ikili kümeler Xmotif algoritmasının ürettiği ikili kümelere göre daha anlamlıdır.

4. Sonuçlar

Suçların önlenmesi konu olunca, öncelikle suçluların yakalanıp cezalandırılarak, potansiyel suçların engellenmesi akla gelmektedir. Oysaki günümüzde suça sebep olacak potansiyel suç bölgelerinin belirlenip suçların işlenmeden önce analiz edilerek önlenmesine yönelik çalışmalar, suçların önlenmesinde daha etkili bir yoldur. Hangi analiz yöntemlerinin ve algoritmaların hangi veri türlerini analiz etmede daha etkin olduğunun bilinmesi suçların önlenmesinde etkili müdahale yapma fırsatı sağlamaktadır. Suç ve suçu meydana getiren etmenlerin birlikte incelenmesiyle farklı amaçlara yönelik suç analizi yapılmaktadır. Suç analizi yapılarak suç ve suçlu arasındaki ilişkinin ve suç bölgelerinin tespit edilmesi, işlenme ihtimali yüksek olan suç türlerinin belirlenmesi ve suçların işlenmeden önce tahmin edilmesi, emniyet güçlerinin gerekli önlemleri alması açısından önemlidir. Bu açıdan suç bölgelerinin belirlenmesinde ikili kümeleme yöntemlerini kullanarak suçun işlendiği bölgeler ile suç türlerini aynı anda kümelemek, klasik kümeleme yöntemlerine göre daha ayrıntılı sonuçlar elde etmeyi sağlamaktadır. Her geçen gün giderek daha çok önem kazanan ikili kümeleme yaklaşımı, veri setlerindeki satır ve sütunlardaki alt grupları arayıp bulan bir veri madenciliği yaklaşımıdır. Buradaki önemli nokta, satır ve sütunları aynı anda değerlendirerek birden farklı örüntü yapısını tespit etme imkânı sağlamasıdır.

Çalışmadaki tüm analizler, hesaplama ve grafiksel sonuçlar için kapsamlı bir yazılım ortamı olan R paket programı kullanılarak gerçekleştirildi. Burada belirtmek gerekirse, R paket programı, ikili kümeleme algoritmalarını, ön işleme yöntemlerini, sonuçları doğrulama ve görselleştirme tekniklerini kapsayan zengin içerikli bir yazılımdır. Çalışmada suç bölgelerinin tespit edilmesi için ikili kümeleme yöntemlerinden CC ve Xmotif algoritmaları gerçek veri setlerinden "Crimes in Boston" veri setine

uygulanmıştır. İkili kümeleme analizi için 2695x52 boyutunda bir veri matrisi kullanılmıştır. Son 3 yıldaki Boston'da 2695 sokakta meydana gelen başlıca suç türleri ele alınmıştır. Elde edilen ikili kümelerin etkinliğini ölçmek amacıyla CKSB'ye bakılarak algoritmaların performansları karşılaştırılmıştır. Bu veri seti için CC algoritmasının Xmotif algoritmasına göre daha anlamlı kümeler bulduğu sonucuna varılmıştır. Hem suç türlerinin hem de suç bölgelerinin geleneksel kümeleme yöntemlerinin ötesinde daha ileri kümele yaklaşımlarından biri olan ikili kümeleme yöntemiyle analiz edilmesi bu çalışmanın odak noktasını oluşturmaktadır. Bu sayede ikili kümeleme yaklaşımı yalnızca suçları ve suç bölgelerini temsil eden satırların yada sütunların kümelendiği tek boyutlu algoritmalara nazaran hem suç türünün hem de suç bölgesinin aynı anda kümelenmesine imkan sağlamaktadır. İkili kümeleme yöntemlerinin bir diğer avantajı, bu çalışmada da görüldüğü üzere elde edilen sonuçlar görsel ve grafiksel olarak kolay bir biçimde anlaşılıp yorumlanabilmektedirler. Algoritma sonucunda elde edilen ikili kümelerin boyutlarına ve üyelik grafiğine bakılarak her bir kümenin hangi suçları kapsadığı ve bu suçların hangi sokaklarda oldukları açık olarak belirlenebilmektedir. Bu sonuçlara bakarak suçun türüne göre, pratik bir biçimde suç önleme stratejileri geliştirilebilmesine ve güvenlik birimlerinin daha etkin kullanılmasına katkı sağlayabilmektedir. Çalışmada takip edilen bu yaklaşımla elimizde geçmiş suç kayıtlarının mevcut olduğu takdirde ülkemizdeki hemen hemen her yerleşim mahallinde etkin bir biçimde kullanılabilir olmasıdır. Daha ileri uygulamalarda verilerin güncellenmesi çevrim içi hale dönüştürülerek güvenlik birimlerine anlık karar desteği sağlanabilir.

Çalışmayı yapmanın amaçlarından biri, var olan ikili kümeleme uygulamalarının suç verileri ile ilgili büyük veri kümelerini işleyip işleyemeyeceğini ve nasıl işleyebileceğini doğrulamaktı. Bu bakımdan elde edilen sonuçlara bakıldığında, anlamlı sonuçlar elde edilmesi açısından bu amaç sağlanış oldu. Bu çalışmanın bir kısıtı olarak özgün veri seti kullanılması yerine, internet üzerinde var olan hazır gerçek veri setlerinden oluşan "Crimes in Boston" veri seti kullanılmıştır. Bu veri seti tarihsel veri niteliğindeki geçmiş kayıtların oluşturduğu hazır bir veri setidir. "Crimes in Boston" veri seti algoritmaların denenmesi ve test edilmesi açısından oldukça yararlıdır. Bu kapsamda kullanılan veri setinin ikili kümeleme yönteminin bu alanda kullanılabilirliği açısından oldukça uygun olduğu anlaşılmıştır. Bu çerçevede kıyaslanan CC algoritmasının Xmotif algoritmasına göre daha anlamlı kümeler oluşturduğu sonucuna da varılmıştır. Tüm bu araştırma sonuçlarından elde edilen bulgulardan ve statik veri analizinden hareketle, "Akan Veri" yaklaşımıyla ve "Büyük Veri" analizi kapsamında çalışmaların yapılmasına da yol açarak bir temel teşkil edeceği anlaşılmıştır. Gelecekte yapılması planlanan diğer çalışmalar ile çevrim içi gerçek veriler ile anlık analizler yapılarak, bu analizlere göre suçların oluşumundan önce önlenmesine yönelik daha etkin müdahale ve önlemler alınarak veya suçları önlemede etkin stratejiler geliştirerek güvenlik birimlerine suçları önleyici nitelikte, veriye dayalı etkin bir karar desteği sağlanacaktır. Ayrıca bu çalışmadaki veri setinde kullanılan veriler yalnızca suç türlerine dayalı verilerden oluşmaktadır. Bunun yanında işlenen suçların etkileri, oluşan kayıplar ve suçların sonuçları ile ilgili daha kapsamlı verilerin analize dâhil edilmesiyle yeni çalışmalar yapılabilecektir.

Referanslar

[1] B. Altunkaynak, H. H. Örkcü and A. Ramazan, "Şehirlerin suç türlerine göre ikili kümeleme yöntemi ile gruplandırılması: Türkiye örneği"*Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Dergisi*, vol. 22, pp. 110–120, Apr. 2018.

[2] Orakcı, M., Ciylan, B., Kök, İ. and Sevri, M., "Suç Analizinde Veri Madenciliği Teknikleri Ve Makine Öğrenmesi Algoritmalarının Kullanılması", 2019 [Online] Available:https://ab.org.tr/ab16/bildiri/111.pdf, Accessed:22-May-2019 .

[3] I. Cil, "Consumption universes based supermarket layout through association rule mining and multidimensional scaling". *Expert Systems with Application* 39.10, 8611-8625. 2012.

[4] M.F. Adak, C. Sen, I. Cil, "İçme Suyu Parametrelerinin Ölçülen Değerleri Üzerine Veri Madenciliği Uygulayarak Parametreler Arası İlişkiyi Belirlemek". *Akademik Platform Mühendislik ve Fen Bilimleri Dergisi* 1, 18-25. 2013

[5] I. Cil, D. Ay, Y.S. Turkan, "Data driven decision support to supermarket layout", *In Proceeding AIKED'09 Proceedings of the 8th WSEAS International Conference on Artificial Intelligence*, Knowledge Engineering and Data Bases 2009 Feb 21.

[6] D. Bozdağ, A. S. Kumar and U.V. Catalyurek. "Comparative analysis of biclustering algorithms." In Proceedings of the First ACM *International Conference on Bioinformatics and Computational Biology*, pp. 265-274. ACM, 2010.

[7] İ. Çil, "Bilgi Tabanlı İmalat Karar Destek Sistemleri ve Bir Uygulama." *Endüstri Mühendisliği*.1.1, 15-27. 2002

[8] B. Pontes, R. Giraldez and J.S. Aguilar-Ruiz, "Biclustering on expression data: A review", *Journal of Biomedical Informatics*, vol.57, pp.163‑180, Jul.2015.

[9] X. Zhu, J. Qui, M. Xie and J. Wang, "A multi-objective biclustering algorithm based on fuzzy mathematics", *Neurocomputing*, vol.253, pp.177–182, Mar. 2017.

[10] R. Arslan, H.H. Örkcü and B. Altunkaynak, "Kaçakçılıkta yakalanan malzeme türlerine göre suçluların kümelenmesi: İkili kümeleme yöntemi", *Uluslararası İktisadi ve İdari İncelemeler Dergisi*, 18. EYİ Özel Sayısı, pp.883-896, Jan. 2018.

[11] A. Kocatürk, "İkili kümeleme algoritmalarının görsel ve sayısal açıdan karşılaştırılması", Gazi Üniversitesi, Ankara, 2018.

[12] M. Singh, M. Mehrotra, "Impact of biclustering on the performance of biclustering based collaborative filtering", *Expert Systems With Applications*, vol. 113, pp.443–456, Jun. 2018.

[13] J. G. Dominguez, R. R. Exposito, "Accelerating binary biclustering on platforms with CUDA-enabled GPUs", *Information Sciences*, vol. 000, pp.1–9, May. 2018.

[14] Q. Gu, K. Veselkov, "Bi-clustering of metabolic data using matrix factorization tools", *Methods*, vol.151, pp.12‑20, Feb. 2018.

[15] A. Kocatürk, B. Altunkaynak, "Gen açıklama verileri için ikili kümeleme algoritmalarının karşılaştırılması ve uygulanması", *Türkiye Klinikleri J Biostat*, vol.10, pp.137-52, Sep.2018.

[16] R. Subhanishi, V. Milani, "Implementing geographical information system to provide evident support for crime analysis", *Procedia Computer Science*, vol.48, pp.537 – 540, May. 2015.

[17] M. A. Jalil, C. P. Ling, N. M. M. Noor and F. Mohd, "Knowledge representation model for crime analysis", *Procedia Computer Science*, vol.116, pp.484–491, Oct. 2017.

[18] P. E. Oguntunde, O. O. Ojo, H. I. Okagbue and O. A. Oguntunde, "Analysis of selected crime data in Nigeria", *Data in Brief*, vol.19, pp.1242–1249, Jun.2018.

[19] L. G. A. Alves, H. V. Riberio and F. A. Rodrigues, "Crime prediction through urban metrics and statistical learning", *Physica*, vol.505, pp.435–443, Apr.2018.

[20] M. Quick, G. Li and I. B. Smith, "Crime-general and crime-specific spatial patterns: A multivariate spatial analysis of four crime types at the small-area scale", *Journal of Criminal Justice*, vol. 58, pp.22–32, Jul. 2018.

[21] "Veri setinin hazırlanması ve düzenlenmesi", [Online] Available: https://acikders.ankara.edu.tr/pluginfile.php/1375/mod_resource/content/2/B2_Veri%20Seti.pdf, Accessed: 22-May-2019.

[22] S. H. Bal, F. Budak, "Mikroarray Teknolojisi", *Uludağ Üniversitesi Tıp Fakültesi Dergisi*, vol.38, pp.227-233, Oct.2012.

[23] Y. Cheng, G. M. Church, "Biclustering of expression data", *In Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology*, vol.8, pp.93–103. 2000,

[24] T. M. Muralı, K. Simon, "Extracting conserved gene expression motifs from gene expression data", *Pacific Symposium on Biocomptuing*, vol.8, pp.77-88, 2003.

[25] B. K. H. Chia, R. K. M. Karuturi, R., "Differential co-expression framework to quantify goodness of biclusters and compare biclustering algorithms", *Algorithms for Molecular Biology*, vol.5, pp.23, May. 2010.