Düzce Üniversitesi Bilim ve Teknoloji Dergisi

Araştırma Makalesi

Gözetimsiz Makine Öğrenme Teknikleri ile Miktara Dayalı Negatif Birliktelik Kural Madenciliği

Zahraa Mohammed Malik MALIK [a], Shadi AL-SHEHABI [a], Tansel Dökeroğlu [a,*]

[a] *Bilgisayar Mühendisliği Bölümü, Mühendislik Fakültesi, THK Üniversitesi, Ankara, TÜRKİYE*
** Sorumlu yazarın e-posta adresi: tansel@ceng.metu.edu.tr*

ÖZET

Birliktelik kuralları, veri kümesindeki nesnelerin varlığının diğer nesnelerin varlığını nasıl etkilediğini tanımlanmaktadır. Bu kurallar, alış veriş sepeti analizinde, bir ürünün aynı işlemdeki diğer ürünler üzerindeki etkisini incelemek için kullanılmaktadır. Pozitif ve negatif birliktelik kuralları olarak iki şekilde ifade edilebilirler. Pozitif birliktelik kuralı, bir ürün varlığının aynı işlemde diğer ürünü bulma olanağını arttırdığını, negatif birliktelik kuralı, bir çeşidin bulunmasının, diğer ürünün aynı işlemde olabilme ihtimalini azalttığını göstermektedir. Daha önceki işlemlerdeki sıklığı araştırdığı için pozitif birliktelik kuralı madenciliği, negatif birliktelik kuralları madenciliğine göre daha kolaydır. Negatif birlikteki kuralı madenciliğinde daha önceki işlemler araştırıldığında, ilgisiz ürünler arasındaki kurallarının madenciliği ile karşılaşılır. Bu kuralların çıkarımından kaçınmak için madencilik tekniklerine sağlanan önceden tanımlı alan bilgisi kullanılmaktadır. Dolayısıyla bu bilgi, bulunan kuralların ilgili ürünlere ait olmasını gerektirir. Bu çalışmada, satın alınan miktarlara dayalı veri kümesinden otomatik olarak bilgi alınması ile veri kümesindeki ürünler arasındaki negatif birliktelik kurallarını bulma kabiliyetine sahip yeni bir teknik önerilmektedir. Birliktelik kuralı madenciliği, gözetimsiz veri madencilik tekniği olduğundan, sağlanan veri kümesi etiketsiz verilerden oluşmaktadır. DBSCAN kümeleme yönteminin kullanımı, gerçek yaşam işlem veri tabanında test edildiğinde %0,21 destek ve %91,84 ortalama güven değerleri ile elde edilen 4.086 kural ile daha iyi sonuçlar göstermektedir. K-ortalama kümeleme yönteminin kullanımı ile çıkarılan alan bilgisine dayalı negatif birliktelik kuralları madenciliği sonucunda, %0,19 destek ve %85,84 ortalama güven değerine sahip 1.780 kural bulunurken, önerilen alan bilgisiz negatif birliktelik kuralı sonucu %0,12 destek ve %99,37 güven ortalama değerli 9.066 kural ile bu sonucu vermiştir.

Anahtar Kelimeler: *Veri madenciliği, Birliktelik kuralları, Negatif birliktelik kuralları, Kümeleme, Gözetimsiz makine öğrenmesi.*

Quantity-Based Negative Association Rule Mining Using Unsupervised Machine Learning Techniques

ABSTRACT

Association rules are defined as the relationships between objects in the dataset where the existence of one object in a certain condition affects the probability of the existence of the other object. These rules are widely investigated

Geliş: 12/04/2018, Düzeltme: 08/06/2018, Kabul: 20/07/2018

in the analysis of shopping baskets, to examine the effect of one item on the other in the same transaction. These rules may appear in two terms, positive and negative association rules. The negative association rule indicated that the existence of an item decreases the chance that the other item may appear in the same transaction. Mining positive association rules is relatively easy by simply investigating frequent patterns in earlier transactions. Mining negative association rule faces the main challenge of mining uninteresting rules between unrelated items, when earlier transactions are investigated. To avoid the extraction of such rules, existing negative association rule mining techniques rely on a predefined domain knowledge provided to the mining techniques. So that, this knowledge is used to ensure that the extracted rules are for related items. In this study, a novel technique is proposed that has the ability to mine interesting negative association rules between items in the transactions dataset, by automatically extracting knowledge from that dataset based on the purchased quantities. As miningassociation rules is an unsupervised data mining technique, the provided dataset is unlabeled data. The use of DBSCAN clustering method has shown better negative association rule mining results of 4,086 rules, with an average of 0.21% support and 91.84% confidence, when tested on a real-life transactions dataset. Mining negative association rules based on the domain knowledge extracted using the K-means clustering method has 1,780 rules with an average of 0.19% support and 85.84% confidence, while mining negative rules without any domain knowledge results in 9,066 rules with an average support of 0.12% and average confidence of 99.37%, using the same dataset.

Keywords: *Data mining; Association rules; Negative association rules; Clustering; Unsupervised machine learning.*

I. GİRİŞ

Bilgisayarların sunduğu yüksek performans ve doğruluk ile farklı alanlardaki kullanımının hızla artması, uygulamaların büyük miktarda veri ile çalışması zorunluluğunu da beraberinde getirmiştir. Faydalı bilgilerin bulunması, makine öğrenmesine dayanan veri madencilik teknikleri kullanılarak gerçekleştirilebilmektedir. Veri madenciliği, en yaygın makine öğrenme alanlarından birisidir. Yaygın olarak kullanılan madencilik tekniklerinden birisi, birliktelik kural madenciliğidir [1,2]. Bir veri kümesindeki nesneleri birleştiren veya bağlayan ilişkileri araştırarak bilgiyi büyük bir veri kümesinden bulup çıkarır. Diğer makine öğrenme teknikleri gibi, veri madencilik, gözetimli ya da gözetimsiz olabilir. Birliktelik kuralları, gözetimsiz bir veri madenciliği tekniğidir. Bu teknikte veri etiketlemeye gerek olmaksızın nesne ile özellikler arasındaki ilişkiler incelenir. Birliktelik kuralı madencilik tekniği, nesnelerin işlem veri setlerinin analizi için yaygın olarak kullanılmaktadır. Birliktelik kuralı, aynı alış-veriş sepetindeki iki nesne arasındaki ilişki olarak tanımlanabilir. Bir nesnenin sepette bulunması, diğer nesnenin aynı sepette bulunma olasılığı üzerinde doğrudan etkiye sahiptir. Bu ilişkiler, pozitif ve negatif birliktelik kuralı olmak üzere iki kategoriye ayrılabilir [3,4].

Bir sepette, öncül kalem olarak bilinen bir nesnenin varlığı, ardıl (bağlı) kalem olarak bilinen diğer nesnenin olma ihtimalini arttırdığında, iki nesne arasında pozitif birliktelik kuralı vardır denir. Ayrıca, bir sepette öncül nesnenin varlığı, ardıl nesnenin aynı sepette veya işlemde olmaması olasılığını arttırdığında ise n-öncül nesnenin ardıl nesnenin ile negatif birliktelik kuralından söz edilir. Birliktelik kuralları, simetrik değildir, dolayısıyla, *X* nesne ile *Y* nesne arasında belli bir birliktelik kuralının olması, ters yönde de yani *Y*'nin *X* ile bir birliktelik kuralı olmasını gerektirmez.

Negatif birliktelik kurallarının çıkarımı ile karşılaştırıldığında, pozitif birliktelik kuralının çıkarımının daha kolay olduğu görülür. Pozitif birliktelik kuralları, söz konusu iki nesnenin önceden belirlenen bir

eşik değer ile karşılaştırıldığında veri kümesinde birlikte görülme sıklığına göre bulunur. Öte yandan, negatif birliktelik kuralında, büyük bir zorluk vardır: ilgisiz negatif kural çıkarımı. Bu çıkarım, veri seti, alan bilgisi olmaksızın araştırıldığında, elde edilen negatif birliktelik kuralları bağlantısız nesneleri bir araya getirdiğinde görülür. Yani ardıl nesnenin yokluğuna neden olur. Çünkü bu nesneler birbirleri ile bağlantılı ve öncül nesnenin etkisine göre değildir. Bu kuralların çıkarımından kaçınmak için, bulunan negatif birliktelik kurallarının ilgili ve yararlı kural olduğunun kesinleştirilmesi alan bilgisine ihtiyaç duymaktadır.

Bir birliktelik kuralının gücünü belirlemek için destek ve güven olmak üzere iki güçlülük ölçümü kullanılır. Destek, çıkarımı yapılan birliktelik kuralının ne kadar güvenilir olduğunu gösterirken, güven, bu kuralın olma ihtimalini niteler. Destek, bu kural kalemlerinin veri kümesinde görülmesine göre hesaplanırken, güven, işlem kısmını temsil eder. Yani bu kural, bu kalemleri içeren işlem toplam sayısı için doğrudur. Zayıf kuralların önlenmesi için destek ve güven eşik değerleri belirlenir.

Veri madenciliği teknikleri, özel bir alan performansının iyileştirilmesi için kullanılabilecek faydalı bilgiyi arayıp bulmak amacıyla farklı uygulama alanları için toplanan verilerin analizi için yaygın olarak kullanılır. Veri setinden değerli bilgiyi bulup çıkarmak için, nesneler arasındaki güçlü ilişkileri ve veri seti özelliklerini bulmak önemlidir. Birliktelik/ilişkilendirme kuralı madenciliği, pozitif ve negatif birliktelik kuralı madenciliği olarak ikiye ayrılabilen yaygın veri madenciliği uygulamalarından birisidir. Pozitif birliktelik kurallarının bulunması, negatif olanların bulunmasından çok daha kolaydır.

Negatif birliktelik kuralı madenciliğinde alan bilgisinin kullanılması önemli olduğundan, bu araştırmada herhangi bir dış bilgiye gerek olmaksızın gerekli bilgiyi otomatik olarak bulan yeni bir yöntem önerilmiştir [18]. Önerilen yöntem, her bir kalemin satın alındığı miktar ile ilgili bilginin bulunmasına dayanır. Daha sonra, negatif birliktelik kuralları, satın alınan miktarların dağılımından elde edilen bilgiden çıkarılır. Bu bilgi, işlem veri setindeki her bir nesneye ait işlem başına satın alınan miktarların homojen gruplarını bulmak için uygulanan kümeleme yöntemleri kullanılarak çıkarılır. Bu grupları bulmak için, bu nesnenin mevcut satın alınan miktarları ve veri seti ile miktar sıklığını kullanarak bu veriler için histogramlar oluşturulur. Sonrasında, nesne kümelerinin dağılımı arasındaki ilişkileri araştırmak için bu kümeleme teknikleri kullanılır. Kümeleme sonuçları, ürünlerin satın alındığı sıklık örüntülerini ortaya çıkarır. Bu örüntüler daha sonra, negatif birliktelik kurallarını bulmak için kullanılır.

İlgili olmayan negatif birliktelik kuralları madenciliği sorununun önüne geçmek için ilk olarak pozitif birliktelik kuralları aranır. Pozitif birliktelik kuralları, birliktelik kuralındaki nesneler arasında ilişki olduğunu doğrular. Sonrasında bu nesneler, satın alınan miktar histogramlarına göre gruplara ayrılır. Ardından birbirlerine birleştiren pozitif birliktelik kuralına sahip kalem grupları arasındaki negatif birliktelik kuralları araştırılır. Bu işlem, önceden alan bilgisine ihtiyaç olmaksızın ilginç negatif birliktelik kurallarını oluşturur.

İkinci bölümde, araştırmada kullanılan veri madencilik teknikleri ile ilgili literatür incelemesi verilmiştir. Üçüncü bölümde, önerilen yöntem açıklanmaktadır. Dördüncü bölümde yapılan deneylerin sonuçları gösterilip tartışılmıştır. Beşinci bölümde araştırmanın sonuçları değerlendirilmiştir.

II. ÖNCEKİ ÇALIŞMALAR

Veri madenciliği, veri setindeki özellik (feature) değerleri arasındaki ilişki ve örüntüleri tespit etmek için veri kümelerini analiz eden makine öğrenme alanıdır [5]. Birliktelik kuralları, bu kuralların destek ve güven değerlerinin hesaplanması ile bir veri kümesindeki sıklık örüntülerini belirten şartlar kümesidir. Bir işlemler veri setinden birliktelik kurallarının tespiti, her bir işlemdeki kalemler arasındaki ilişkiyi bulmak için çok yaygındır [6]. Destek, veri setindeki toplam işlemler sayısına göre kalemi veya kalemler kümesini içeren işlemler sayısını temsil eder. Dolayısıyla, bir T işlem veri setindeki bir X kalemi için Eşt. 1'deki gibi hesaplanır [15]:

$$Support(X) = \frac{|X|}{|T|} \tag{1}$$

Ayrıca, $X \Rightarrow Y$ eşitliğinde destek değeri, hem X öğesi hem de Y öğesine sahip işlemler sayısının veri setindeki işlem toplam sayısına oranına eşit bir destek değerine sahiptir. Dolayısıyla $X \Rightarrow Y$ destek Eşt. 2'deki gibi hesaplanır [16].

$$Support(X \Rightarrow Y) = \frac{|X \cup Y|}{|T|} \tag{2}$$

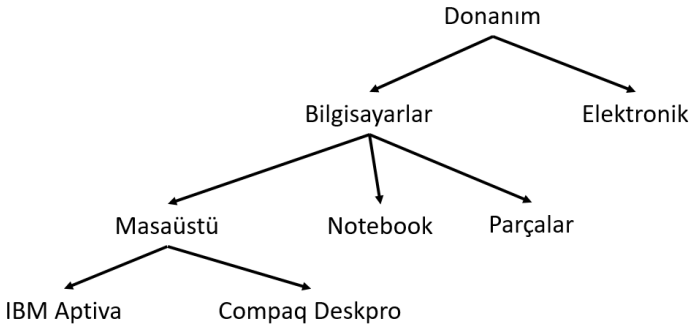
$X \Rightarrow Y$ birliktelik kuralında, sol taraftaki X, öncül, sağ taraftaki Y ise ardıl olarak bilinir. Öte yandan güven, bu kuralın görülme ihtimalini gösterir. İşlemlerdeki bu kural sıklığının öncül öğeyi içeren işlemler toplam sayısına oranı olarak hesaplanır. Kural güven değeri $X \Rightarrow Y$ Eşt. 3 kullanılarak hesaplanır [17].

$$Confidence(X \Rightarrow Y) = \frac{|X \cup Y|}{|X|} \tag{3}$$

Birliktelik kuralları, pozitif ve negatif birliktelik kuralları olmak üzere iki kategoriye ayrılabilir [7]. Negatif birliktelik kuralları, ilk olarak [8]'de tartışılmış, değişkenler arasındaki bağımsızlığı tespit etmek için istatistik testleri uygulanmıştır. Önerilen model, diğerine göre bir nesnenin varlığı ve yokluğuna dayanmaktadır. Negatif birliktelik kurallarının bulunmasında ilgisiz negatif birlikteliklerin önlenmesi zorluğu ile karşılaşılır. Örneğin, araştırma konusu veri setinde yer alan faturalarda hiç satın alınmamış veya nadiren satın alınmış bir kalem, diğer her sıklıkla satın alınan kalemle negatif birliktelik kuralına sahip değildir. Bu kuralların çıkarılmasının önlenmesi için, [9]'da önerilen yöntem, güçlü anlamlı negatif birliktelik kuralı bulmak için pozitif birliktelik kuralları çıkarımını birleştirir. Bu yöntemde kullanılan alan bilgisi, kalemleri gruplara sınıflandırır. Dolayısıyla bir kalem ile bir kategori arasında bir pozitif birliktelik kuralı keşfedildiğinde bu kategorideki bu öncül ile her kalemin negatif kuralları araştırılır. Örneğin, yoğurt ile su arasında bir pozitif birliktelik kuralı keşfedilir ve algoritmada sağlanan taksonomi (sınıflandırma), yoğurdun sağlıklı ve normal yoğurt şeklinde iki kategoriye ayrılabildiğini, suyun ise arıtma ve maden suyu gibi iki kategoriye ayrıldığını göstermektedir. Sonrasında bu kalemlerdeki kategorilerin her olası kombinasyonu arasında negatif birliktelik kuralları araştırılır. Bu yöntemde kullanılan taksonomi veri kümesinden otomatik olarak keşfedildiğinden, negatif birliktelik kurallarını bulmak için algoritmaya el ile veri sağlanması gerekmektedir.

Ayrıca, [10]'da önerilen negatif birliktelik kuralı araştırma tekniği, ilk olarak pozitif birliktelik kurallarını çıkarmaya bağlı olarak bu kuralları araştırır. Sonrasında, işlem veri setindeki kalemlerin

benzerliğine dayalı olarak negatif birliktelik kuralları çıkarılır. Satın alınan nesneler, algoritmaya sağlanan nesnelerin taksonomisine dayalı benzerliklerine göre gruplandırılır. Şekil 1'de gösterilen örnek taksonomide, araştırma taksonomi ağacında üst seviyeden alt seviyeye indikçe bu gruplar daha spesifik hale gelir. İki nesne grubu arasında bir pozitif birliktelik kuralı keşfedildiğinde, taksonominin aşağı seviyesindeki gruplar negatif birliktelik kuralı için araştırılır. İlgisiz negatif birliktelik kuralının bulunmasından kaçınılması ve bu bilginin kullanılması için, ayrıca alan bilgi olarak işlem veri setindeki nesneler taksonomisinin bir tanımı da gerekir.



Şekil 1. *Örnek tasonomi.*

Çalışma [11]'de önerilen yöntem, nesnelerin markalarına göre negatif birliktelik kurallarını araştırır. Dolayısıyla belli bir markadan bir ürün satın alan müşteri, başka bir ürün satın alma eğilimindedir. Ancak özel bir markayı satın almaktan kaçınır. Örneğin, atıştırmalıklar ile meşrubatlar arasında pozitif birliktelik kuralı bulunur ve bu ürünlerin taksonomisi, her bir ürünün farklı markaları olduğunu göstermektedir. Atıştırmalıkların markaları ile meşrubat markası arasında negatif birliktelik kuralları araştırılır. Bu araştırmanın sonuçları, algoritmaya sağlanan ve taksonomi ile gösterilen alan bilgisinin, işlem veri setinden çıkarılan negatif birliktelik kuralları sayısı ve veri setini işlemek için gerekli yapım süresi üzerinde güçlü bir etkiye sahip olduğunu göstermektedir.

Daha önce tartışılan negatif birliktelik kuralı araştırma teknikleri ve bazı diğer teknikler, ilk olarak pozitif birliktelik kurallarının keşfine dayalı bu kuralların çıkarılmasına dayanmaktadır. Bu işlem, birbirleri ile ilgili olmayan ürünlere ilişkin bu kuralları çıkarmak yerine ilgili negatif birliktelik kurallarının bulunmasını sağlar. Pozitif birliktelik kuralları da adaylar arasında bir ilişki olduğunu gösterir. Daha sonra, öncül ve ardıl kalemlerin spesifik özelliklerine göre bir araştırma yapılır. Bu yöntemleri negatif birliktelik kuralları araştırılmasında kullanmak için nesnelerin özelliklerinin aralarındaki ilişkilerin önceden tanımlanması gerektirir. Diğer taraftan pozitif birliktelik kuralı araştırması için bu bilgiye gerek yoktur.

Bir başka önemli makine öğrenme alanı, veri kümeleme olup bu da aynı zamanda veri kümesine herhangi bir etiketleme gerektirmeyen gözetimsiz makine öğrenme veri madenciliğidir. Veri kümeleme, nesnelerin gruplara ayrılması olarak tanımlanabilir. Bir gruptaki bir nesne, başka gruptaki başka bir nesneye göre bu gruptaki diğer nesnelere daha çok benzerdir [12]. Dolayısıyla her bir küme, bu nesneleri karakterize eden özelliklere ait değerler ile ilgili olarak daha çok homojen nesneler içerir. Veri kümeleme gözetimsiz makine öğrenme tekniği olduğundan, bu nesneleri belli bir kümeye gruplayan ilişkiler, kümeleme yöntemi olarak bilinmemekte ve daha büyük veri setlerinde insanların dikkatini çekmemektedir.

K-ortalama, veri kümelerini sayısal öznitelikler ile kümeleme için kullanılan en etkin veri kümeleme yöntemlerinden birisidir. Bu yöntem, n-boyutlu uzayda K rastgele sentroidlerden başlatarak önceden tanımlı "K" kümesini oluşturur. Burada *n,* veri setindeki öznitelikler sayısıdır. Veri setinin nesneleri, gruplara dağıtılır. Her bir nesne, en yakın önerilen sentroide ait grubun bir üyesi olarak değerlendirilir. Bu sentroidlerin konumu, her bir sentroid ile grubundaki nesne arasındaki karesi alınmış mesafelerin toplamının hesaplanması ile optimize edilir. Bu optimizasyon işlemi, karesi alınmış mesafelerin minimum toplamına ulaşılıncaya kadar tekrar edilir. Bu noktada, sentroid konumu ve aynı zamanda her bir kümedeki nesne tanımlanır [13].

Daha homojen kümeler oluşturabilmek için gerekli küme sayısı K'yı sınıflandırılan nesnelerin optimal sayısına ayarlamak önemlidir. Bu sayı, dirsek yöntemi kullanılarak hesaplanabilir. Bu yöntem, veriler iki küme veya daha fazla kümeye ayrıldığında, her bir tekrar için bir ilave küme ekleyerek karesi alınmış mesafelerin toplamını maliyet fonksiyonu olarak hesaplar. Maliyet fonksiyonu değerleri, kümelerin alt sayısında keskin düşüş gösterebilir. Sonrasında bu düşüş, düzleşmeye başlar, bunun anlamı, ilave kümeleme ile eklenen bilginin önceki kümeler sayısı kadar çok olmadığıdır. Dolayısıyla dirsek noktası olarak bilinen bu noktadaki kümelerin sayısı, kümelerin optimal sayısı olarak seçilir [14].

Bir veri seti için küme sayısını otomatik olarak tespit etme kabiliyetine sahip başka bir veri kümeleme yöntemi, Gürültülü Uygulamaların Yoğunluk Esaslı Uzaysal Kümelenmesi (DBSCAN) yöntemidir, Ester et al. tarafından 1996 yılında önerilmiştir [19,20]. DBSCAN, nesnelerin komşuları ile olan mesafelerini hesaplayarak belirli bir bölgede önceden belirlenmiş eşik değerden daha fazla nesne bulunan alanları gruplandırarak kümeleme işlemini gerçekleştirir. DBSCAN algoritması veri madenciliğine birçok yeni terim ve yaklaşım getirmiştir.

Histogramlar, belli bir veri setinde bir değerin ne sıklıkta görüldüğünü gösterir. Her bir değerin veri setinde görülme sayısı ile birlikte toplanması ile oluşturulur. Bu frekansların dağılımı, makine öğrenme teknikleri kullanılarak saklanabilen bazı bilgileri içerir. Bu bilgi bilinmediğinden, kümeleme yöntemleri kullanılarak çağrılabilir zira bu yöntemler gözetimsiz öğrenme yöntemleri olup veri etiketleme gerektirmez. Bilgiyi bulup çıkarmak için kümeleme histogramları farklı alanlarda uygulanmıştır. Dolayısıyla elde edilen bilgi, aralarındaki spesifik ilişkileri test etmek için veri setinden bulunup alınan başka bilgilerle farklı teknikler kullanılarak karşılaştırılır.

III. ÖNERİLEN YÖNTEM

Önerilen negatif birliktelik kuralı madencilik yöntemi, makine öğrenme tekniklerinden yararlanarak veri kümesindeki negatif birliktelik kural madenciliği için gerekli alan bilgisini elde eder. Bu yöntem, satın alınan nesnelerin miktarına bağlı olduğundan, negatif birliktelik kurallarını bulup çıkarmak için histogramlar kullanılır. Bu işlem için veri kümeleme yöntemleri kullanılır. Çünkü işlem veri setleri ve miktarlar histogramı etiketli veri değildir. Bu kümeler, pozitif birliktelik kurallarını paylaşan öge grupları arasındaki negatif birliktelik kurallarını bulmak için kullanılır.

Önerilen yöntem, gözetimsiz makine öğrenme tekniği olup tüm değişkenlerin algoritma ile hesaplanması önemlidir. Kümelerin optimal sayısını hesaplayabilen bir tekniğe ihtiyaç bulunmaktadır. Dirsek yöntemi, istenen kabiliyete sahip yaygın olarak kullanılan yöntemlerden birisidir. Bu yöntem, önceden belirlenen maksimum sayısına kadar kümeleme sonuçları için her bir nesne ve sentoroidin ait olduğu kümenin sentroidi arasındaki mesafenin toplamını hesaplar. Daha sonra, toplamda değişikliğin

keskin olarak düştüğü değer, kümelerin optimal sayısı olarak değerlendirilir. Bu yöntem kullanılarak oluşturulan grafik, dirsek birleşim yeri noktasının kümelerin optimal sayısı olarak seçildiği nokta, insan dirseğine benzediği için bu yöntem dirsek yöntemi olarak adlandırılmıştır. Daha az sayıdaki küme alınan alan bilgisinde büyük kayba neden olurken, kümelerin yüksek sayısı veri setinin kümelenmesi ile bulunan alan bilgisine herhangi bir önemli katkı olmaksızın daha fazla işlem gerektirir.

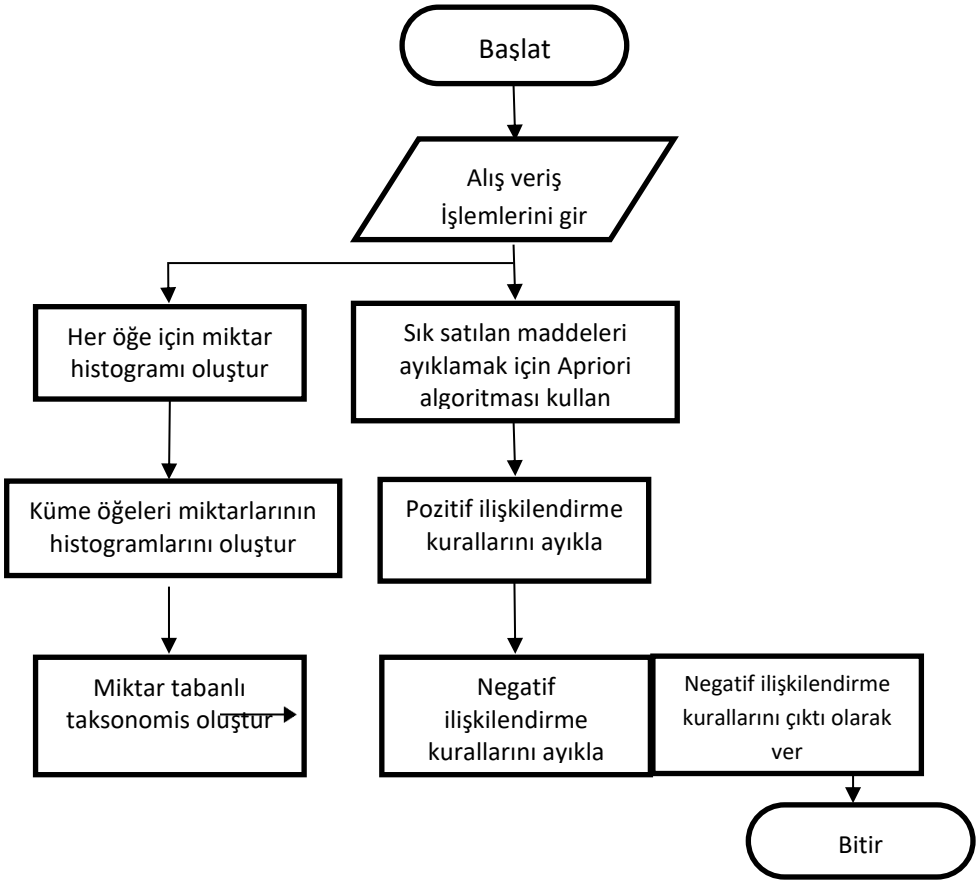
Diğer kümeleme yöntemi, DBSCAN çok boyutlu uzaydaki bir veri kümesini, bir tek kümedeki nesne minimum izin verilebilir sayısını ve aynı kümedeki iki bitişik nesne arasında maksimum mesafe sağlayarak, otomatik olarak optimal küme sayısına kümeleme kabiliyetine sahiptir. Kümeleme işlemi, veri setinden rastgele bir nesne seçilerek başlatılır, maksimum izin verilebilir mesafeden az mesafeli en yakın nesne, aynı küme içinde olmak üzere seçilir. İzin verilebilir mesafe limitleri içinde başka nesne dahil edilmeyinceye kadar devam edilir. Diğer rastgele kümelendirilmemiş nesne, sonraki küme kriteri için seçilir. Bu işlemler, tüm nesneler kümelendirilinceye kadar devam eder.

Kümelendirilen miktarlar sistem taksonomisini oluşturmak amacıyla kullanılır. Taksonomi, kümelendirilen miktarlara göre her bir ögenin, ögelerin yeni kümesine azaltılması ile oluşturulur. Böylece öge A ve öge B'nin her bir ögesi için, öge A için A1 ve A2 ve Öge B için B1 ve B2 ile gösterilen iki yeni öge oluşturulur. A1, 1, 3 veya 4 miktarlarına satın alınan A ögesi veri setindeki tüm girdileri gösterirken, A2, 7, 9 veya 10 miktarlarında satın alındığında A'nın girdilerini ve aynı şekilde B1 ve B2 B'nin girdilerini gösterir. Satın alınan A ve B ögeleri miktarlarının kümeleme sonuçlarına göre bulunan taksonomi Şekil 2'de gösterilmiştir.

Şekil 2. *Satın alınan miktarlara dayalı öge taksonomisi.*

Kümeleme kullanılarak bulunan alan bilgisi, negatif birliktelik kuralları için araştırılan ögelerdeki izafiyeti temin için pozitif birliktelik kurallarına sahip öge setlerinden negatif birliktelik kuralı keşfi için kullanılır. İşlem veri kümesinde bulunan ve minimum izin verilebilir destek değerinden büyük destek değere sahip her bir öge, kümeleme sonuçlarına göre ögelerin alt kümesi ile değiştirilir. Kendilerini birleştiren bir pozitif birliktelik kuralına sahip olan ögeler veya öge setleri arasındaki negatif birliktelik kuralları, Şekil 3'te gösterilen akış şemasında gösterildiği gibi öncüldeki ögelerin her altkümesi ile sonuç ögesindeki ögeler alt kümeleri arasında araştırılır.

Böylece X ögesinin satın alınan miktarları iki kümeye kümelenirse ve Y ögesinin miktarları da aynı şekilde kümelenirse, bulunan pozitif birliktelik kurallarında pozitif birliktelik kuralı $X \Rightarrow Y$ vardır. X ve Y'den tüm olası kombinasyonların destek değerleri hesaplanır, böylece bu aday negatif birliktelik kuralları destek değerleri $X_1 \Rightarrow \neg Y_1$, $X_1 \Rightarrow \neg Y_2$, $X_2 \Rightarrow \neg Y_1$, $X_2 \Rightarrow \neg Y_2$, $X \Rightarrow \neg Y_1$, $X \Rightarrow \neg Y_2$, $X_1 \Rightarrow \neg Y$, ve $X_2 \Rightarrow \neg Y$ hesaplanır ve minimum izin verilebilir destek değerinden büyük destek değerine sahip olanlar, sonuç olarak elde edilen negatif birliktelik kuralları olarak bulunur. Dolayısıyla, önceki örnekte yer alan A ve B ögeleri için, pozitif birliktelik kuralı $A \Rightarrow B$ için, bir tarafta A1'den öte tarafta B1 ve B2'ye ve aynı zamanda A1'den B1 ve B2'ye negatif birliktelik kuralları araştırılır. Sonra aynı işlem, varsa, pozitif birliktelik kuralı $B \Rightarrow A$ için tekrarlanır. Miktara dayalı negatif birliktelik kural madenciliği için kullanılan algoritma, Algoritma 1'de sunulmuştur.



Şekil 3. *Önerilen negatif birliktelik kuraları madencilik yöntemi akış şeması.*

Algoritma 1. *Miktar esaslı Negatif Birliktelik Kuralları Madenciliği.*

1: Girdi minimum destek(minSup)
Girdi minimum negatif destek (minNSup)
Girdi minimum güven (minConf)
Girdi minimum negatif güven (minNConf)
2: F = Apriori (I, minSup)
3: F'deki her bir x için
4: F'deki her bir y için
5: Eğer x≠y ise
6: $C_{xy} = \frac{|X \cup Y|}{|X|}$
7: Sonlandır eğer
8: için sonlandır
9: için sonlandır
10: $P = \{X \Rightarrow Y \mid C_{xy} \geq \text{minConf}\}$.
11: P'deki her bir p için
12: S = cluster(F)

13: NF = Apriori($S_x \cup S_y$, minNSup)
14: NF'deki her bir m için
15: NF'deki her bir n için
16: Eğer m≠n ise
17: $NC_{mn} = \frac{|m \cup \neg n|}{|m|}$
18: Sonlandır eğer
19: için sonlandır
20: için sonlandır
21: Her biri için son
22: $N = \{M \Rightarrow \neg N \mid NC_{mn} \geq minNConf\}$.
23: N Dön

IV. DENEYSEL SONUÇLAR

Önerilen negatif birliktelik kuralı madencilik yöntemleri performansını test etmek için, California Üniversitesi, Irvine (UCI) bilgi havuzundan gerçek hayat veri kümesi değerleri kullanılarak üç deney yapılmıştır. Veri seti, 3.925 ürün için 25.295 işlem bilgisi içeren, Fatura Kimlik, Stok Kodu, Adı, Miktarı, Fatura Tarihi, Birim Fiyatı, Müşteri Kimliği ve Ülkesi olmak üzere sekiz özellikle karakterize olan 541.909 veri grubundan oluşmaktadır. Deneylerde kullanılan veri setinin bir örneği Tablo 1'de verilmiştir.

Birliktelik kural madenciliği için sadece İşlem Kimliği, Stok Kodu ve Miktar kullanılmıştır. Tüm deneyler, Intel® Core™ i7-7700HQ @ 2.8 GHz işlemci ve a 16 GB belleğe sahip bir bilgisayarda Windows 10 işletim sisteminde Python programlama dili kullanılarak yapılmıştır. Deneyler, satın alınan farklı gruplar kullanılarak yapıldı, ilk deneyde her bir satın alınan miktar tek başına kullanılırken ikinci ve üçüncü deneylerde farklı kümeleme teknikleri kullanıldı.

Deney A: Bu deneyde, negatif birliktelik kuralları, ögelerin sonuç ögesindeki her miktarı ile öncül ögenin her bir satın alınan miktarı araştırılarak tespit edildi ve veri setinden pozitif birliktelik kuralları bulundu. Deneyin bir özeti Tablo 2'de verilmiştir.

Tablo 1. *Deneylerde kullanılan veri seti örneği.*

Fatura No	Stok Kodu	Adı	Miktar	Fatura Tarihi	Birim Fiyat	Müşteri Kimliği	Ülke
536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	12/1/2010 8:26	2,55	17850	İngiltere
536365	71053	WHITE METAL LANTERN	6	12/1/2010 8:26	3,39	17850	İngiltere
536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	12/1/2010 8:26	2,75	17850	İngiltere
536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	12/1/2010 8:26	3,39	17850	İngiltere
536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	12/1/2010 8:26	3,39	17850	İngiltere
536365	22752	SET 7 BABUSHKA NESTING BOXES	2	12/1/2010 8:26	7,65	17850	İngiltere
536365	21730	GLASS STAR FROSTED T-LIGHT HOLDER	6	12/1/2010 8:26	4,25	17850	İngiltere
536366	22633	HAND WARMER UNION JACK	6	12/1/2010 8:28	1,85	17850	İngiltere
536366	22632	HAND WARMER RED POLKA DOT	6	12/1/2010 8:28	1,85	17850	İngiltere
536367	84879	ASSORTED COLOUR BIRD ORNAMENT	32	12/1/2010 8:34	1,69	13047	İngiltere
536367	22745	POPPY'S PLAYHOUSE BEDROOM	6	12/1/2010 8:34	2,1	13047	İngiltere
536367	22748	POPPY'S PLAYHOUSE KITCHEN	6	12/1/2010 8:34	2,1	13047	İngiltere
536367	22749	FELTCRAFT PRINCESS CHARLOTTE DOLL	8	12/1/2010 8:34	3,75	13047	İngiltere
536370	21731	RED TOADSTOOL LED NIGHT LIGHT	24	12/1/2010 8:45	1,65	12583	Fransa
536370	22900	SET 2 TEA TOWELS I LOVE LONDON	24	12/1/2010 8:45	2,95	12583	Fransa
536370	21913	VINTAGE SEASIDE JIGSAW PUZZLES	12	12/1/2010 8:45	3,75	12583	Fransa
536370	22540	MINI JIGSAW CIRCUS PARADE	24	12/1/2010 8:45	0,42	12583	Fransa
536370	22544	MINI JIGSAW SPACEBOY	24	12/1/2010 8:45	0,42	12583	Fransa
536370	22492	MINI PAINT SET VINTAGE	36	12/1/2010 8:45	0,65	12583	Fransa

Tablo 2. *Deney A negatif birliktelik kuralı madenciliği sonuçların bir özeti.*

Kural sayısı	Yapım. zamanı	Destek			Güven		
		Minimum	Maksimum	Ortalama	Minimum	Maksimum	Ortalama
9066	1932,74	%0,10	%0,13	%0,12	%66,67	%100	%99,37

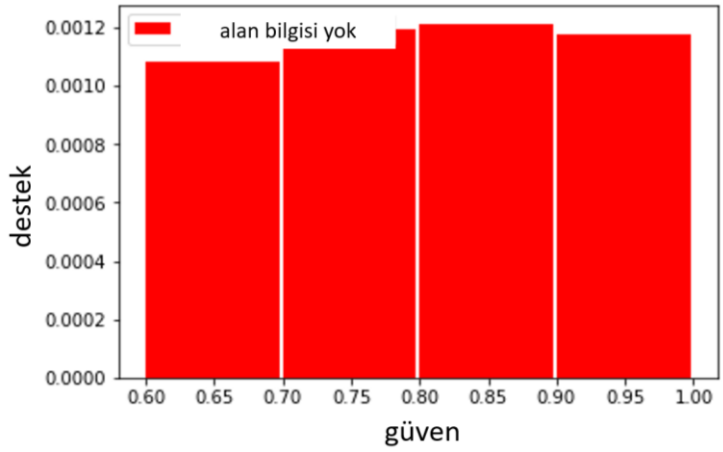
Bu deneyde çıkartılan negatif ilişki kurallarının bir örneği Tablo 3'te da gösterilmiştir.

Tablo 3. *Deney A'dan örnek negatif ilişki kuralları.*

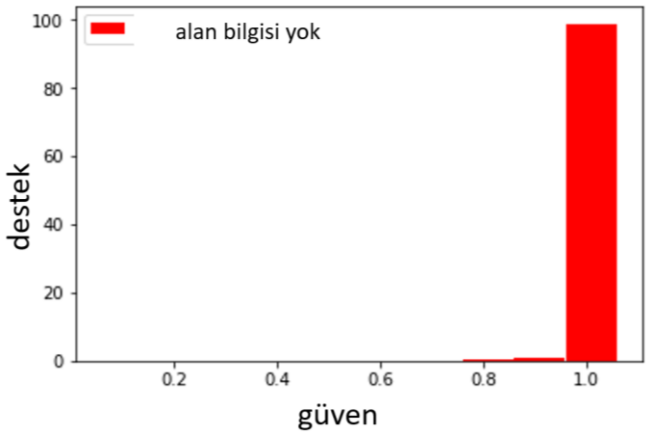
Önceki öge	Sonuç ögesi	Destek (%)	Güven (%)
JUMBO BAG RED RETROSPOT	JUMBO BAG TOYS	0,13	100,00
JUMBO BAG RED RETROSPOT	JUMBO BAG WOODLAND ANIMALS	0,13	100,00
JUMBO BAG RED RETROSPOT	JUMBO BAG OWLS	0,13	100,00
JUMBO BAG RED RETROSPOT	RED RETROSPOT SHOPPER BAG	0,13	100,00

WHITE HANGING HEART T-LIGHT HOLDER	REGENCY CAKESTAND 3 TIER	0,10	100,00
WHITE HANGING HEART T-LIGHT HOLDER	JUMBO SHOPPER VINTAGE RED PAISLEY	0,10	100,00
WHITE HANGING HEART T-LIGHT HOLDER	SMALL POPCORN HOLDER	0,10	100,00
WHITE HANGING HEART T-LIGHT HOLDER	HOME BUILDING BLOCK WORD	0,10	100,00
SET OF 3 CAKE TINS PANTRY DESIGN	JUMBO BAG RED RETROSPOT	0,11	99,65
SET OF 3 CAKE TINS PANTRY DESIGN	JAM MAKING SET PRINTED	0,11	99,83

Ayrıca, ortalama destek – güven karşılaştırması Şekil 4'te verilmiş olup, bulunan negatif birliktelik kuralları yüzdesi bu kuralların güven değerleri Şekil 5'te gösterilmiştir.

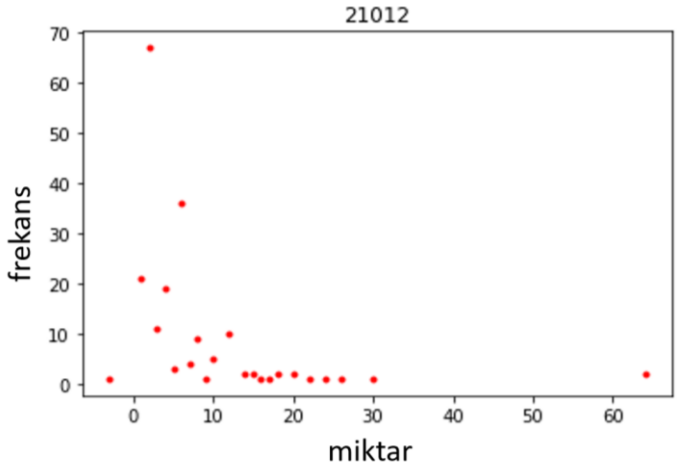


Şekil 4. *Deney A ortalama destek – güven değerleri.*

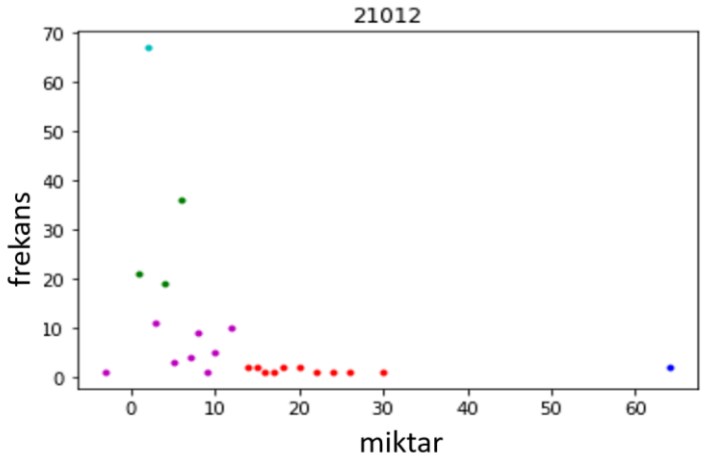


Şekil 5. *Deney A kurallar yüzdesi – güven.*

Deney B: Bu deneyde, satın alınan miktarlar, K-ortalama kümeleme yöntemi kullanılarak kümelendi. Bu yöntem, kümelerin sayısını otomatik olarak belirleme kabiliyetinde olmadığından, optimal küme sayısının seçimi için dirsek yöntemi kullanıldı. Pozitif birliktelik kurallarındaki öncül ve sonuç ögeleri daha sonra, bu kümeler kullanılarak araştırıldı. Satın alınan nitelik histogramına göre tamamlandı. Veri setinden bir tek nesnenin satın alınan miktarlarının bir örnek histogramı Şekil 6'da verilmiştir. Bu öge için dirsek yöntemi ile belirlenen kümelerin optimal sayısı beş kümedir. K-ortalama yöntemleri kullanılarak bu miktarlar Şekil 7'de gösterildiği gibi kümelendirildi.



Şekil 6. *Örnek satın alınan miktarlar histogramı*



Şekil 7. *Örnek kümelendirilen miktarlar histogramı*

Deney B'da elde edilen aynı pozitif birliktelik kurallarına göre bulunan negatif birliktelik kuralları özeti Tablo 4'te verilmiştir.

Tablo 4. *Deney B negatif birliktelik kuralı madenciliği sonuçlarının bir özeti.*

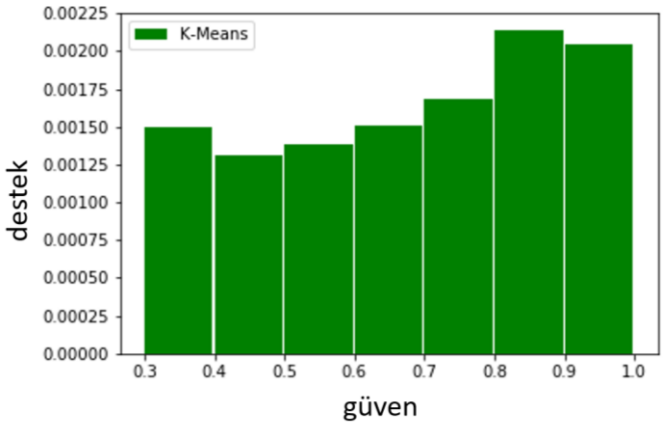
Kural sayısı	Yapım Süresi	Destek			Güven		
		Minimum	Maksimum	Ortalama	Minimum	Maksimum	Ortalama
1.780	353.57	%0,10	%0,44	%0,19	%31,69	%100	%85,84

K-aracı yöntemi kullanılarak çıkarılan etki alanı bilgisine dayalı olarak çıkarılan negatif ilişki kurallarının bir örneği Tablo 5'te gösterilmiştir.

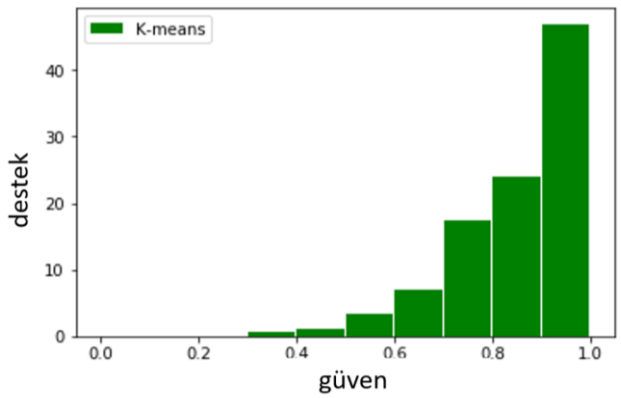
Tablo 5. *Deney B'de çıkarılan negatif ilişki kurallarının örneği.*

Önceki öge	Sonuç ögesi	Destek (%)	Güven (%)
WHITE HANGING HEART T-LIGHT HOLDER	VICTORIAN GLASS HANGING T-LIGHT	0,44	97,18
WHITE HANGING HEART T-LIGHT HOLDER	CANDLEHOLDER PINK HANGING HEART	0,44	97,18
JUMBO BAG RED RETROSPOT	NATURAL SLATE HEART CHALKBOARD	0,40	98,89
PACK OF 72 RETROSPOT CAKE CASES	PACK OF 60 PINK PAISLEY CAKE CASES	0,26	96,30
JUMBO BAG PINK POLKADOT	JUMBO BAG OWLS	0,23	98,40
JUMBO BAG PINK POLKADOT	JUMBO BAG SCANDINAVIAN PAISLEY	0,23	98,40
LUNCH BAG CARS BLUE	LUNCH BAG ALPHABET DESIGN	0,22	86,80
LUNCH BAG CARS BLUE	STRAWBERRY CHARLOTTE BAG	0,22	86,72
JAM MAKING SET PRINTED	SET OF 4 PANTRY JELLY MOULDS	0,22	92,95
JAM MAKING SET PRINTED	GREEN REGENCY TEACUP AND SAUCER	0,22	92,11

Bu deneyde güvene göre ortalama destek dağılımı Şekil 8'de gösterilmiş olup bulunan kural yüzdesi ve bu kuralların güven değerleri Şekil 9'da verilmiştir.



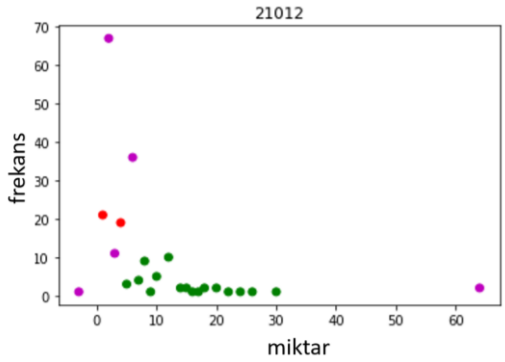
Şekil 8. *Deney B güven – destek ortalaması.*



Şekil 9. *Deney B için güven aralığı başına bulunan kural yüzdesi.*

Deney C: Bu deneyde, satın alınan miktarları veri setinde kümelemek için DBSCAN yöntemi kullanıldı. Bu yöntem, histogramdaki değerleri önceden tanımlı değerlere göre ve satın alınan ögelerin her bir

histogramı için küme sayısı vermeye gerek olmaksızın otomatik olarak kümeleme kabiliyetine sahiptir. Şekil 4'te gösterilen aynı örnek öge için, DBSCAN bu ögeyi Şekil 10'da gösterildiği gibi üç kümeye gruplandırdı.



Şekil 10. *DBSCAN kullanılarak örnek öge miktarları histogramının kümeleme sonuçları.*

DBSCAN yöntemi kullanılarak oluşturulan kümelere göre bulunan negatif birliktelik kurallarının özeti Tablo 6'da verilmiştir.

Table 6. *Deney C negatif birliktelik kuralı madenciliği sonuçlarının bir özeti.*

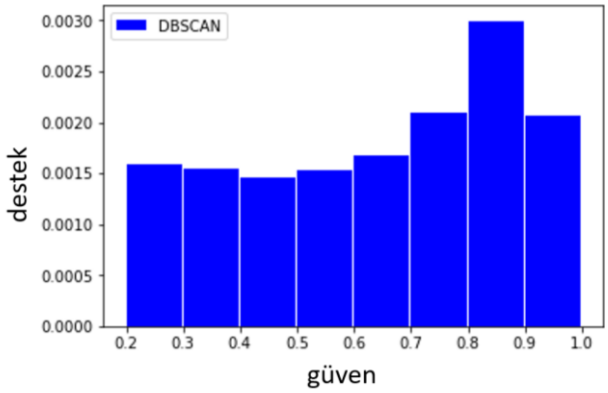
Kural sayısı	Yapım Süresi	Destek			Güven		
		Minimum	Maksimum	Ortalama	Minimum	Maksimum	Ortalama
4086	783,3	%0,10	%0,40	%0,21	%22,54	%100	%91,84

Bu deneyde çıkarılan örnek negatif ilişki kuralları Tablo 7'de gösterilmiştir.

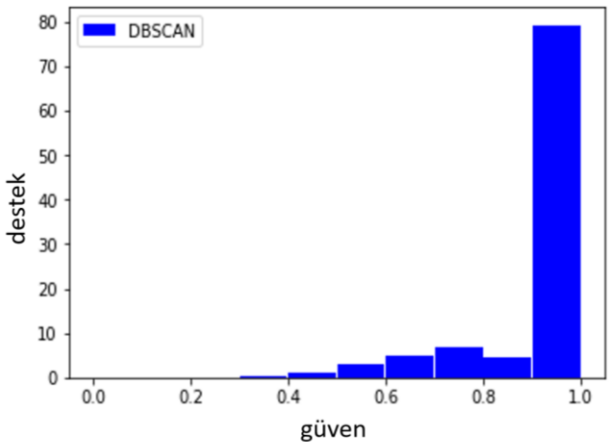
Tablo 7. *Deney C'nin örnek negatif ilişki kuralları.*

Önceki öge	Sonuç ögesi	Destek (%)	Güven (%)
WHITE HANGING HEART T-LIGHT HOLDER	NATURAL SLATE HEART CHALKBOARD	0,41	97,18
WHITE HANGING HEART T-LIGHT HOLDER	RECIPE BOX PANTRY YELLOW DESIGN	0,41	97,18
JUMBO BAG BAROQUE BLACK WHITE	WOODLAND CHARLOTTE BAG	0,38	98,92
JUMBO BAG BAROQUE BLACK WHITE	LUNCH BAG SUKI DESIGN	0,38	98,92
RED RETROSPOT CHARLOTTE BAG	LUNCH BAG BLACK SKULL.	0,29	98,14
RED RETROSPOT CHARLOTTE BAG	JUMBO STORAGE BAG SKULLS	0,29	78,09
GUMBALL MONOCHROME COAT RACK	RECIPE BOX PANTRY YELLOW DESIGN	0,27	98,96
ASSORTED COLOUR BIRD ORNAMENT	WHITE HANGING HEART T-LIGHT HOLDER	0,26	98,48
LUNCH BAG SPACEBOY DESIGN	SMALL POPCORN HOLDER	0,24	96,86
LUNCH BAG SPACEBOY DESIGN	TEA TIME PARTY BUNTING	0,24	96,78

Bu deney için ortalama destek – güven değerleri Şekil 11'de ve her bir güven aralığı için bulunan kurallar yüzdesi ise Şekil 12'de gösterilmiştir.



Şekil 11. *Deney C için ortalama destek – güven değerleri.*



Şekil 12. *Deney C için bulunan kuralların güven karşısında dağılımı.*

Ayrıca, önerilen yöntemlerin karmaşıklığını karşılaştırmak amacıyla, Yapım Süresi, 10,000, 500,000 ve 2,000,000 kayıttan oluşan üç farklı veri seti kullanılarak her bir algoritma için bir ölçümdür. Ölçüm Yapım Süresi Tablo 8'de gösterilmiştir.

Tablo 5.8: *Farklı büyüklükteki veri setleri için her bir algoritma başına yapım süresi.*

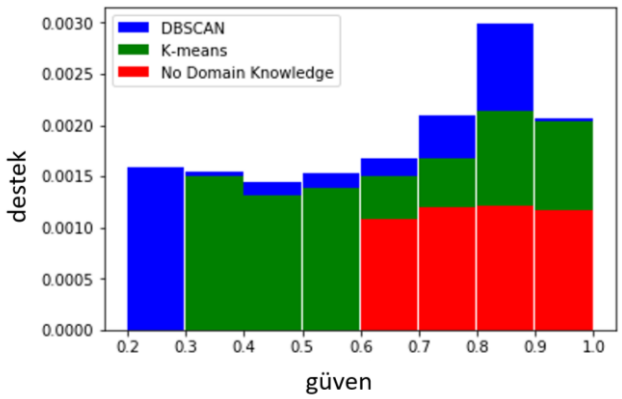
	Süre (saniye)		
Kayıt sayısı	Kümeleme yok	K-ortalama	DBSCAN
10.000	1,27	0,41	0,33
500.000	1232,92	1429,90	395,41
2.000.000	15326,43	10410,22	7721,35
Ortalama	5520,21	3946,84	2705,70

Deney A sonuçları, herhangi bir alan bilgisi olmaksızın negatif birliktelik kuralları madenciliğini göstermektedir. Bu negatif birliktelik kural madenciliğinin, ilgisiz kurallar oluşturması beklenir. Tablo 9'da gösterildiği gibi bu deney sonuçlarını bu araştırmada yapılan diğer deney ile karşılaştırılmış olup, alan bilgisi olmaksızın bulunan negatif birliktelik kuralları ortalama destek değeri, alan bilgisi kullanılarak bulunan diğer kurallar destek değerinden küçüktür.

Tablo 5.9: *Farklı teknikler kullanılarak bulunan negatif birliktelik kuralı özeti.*

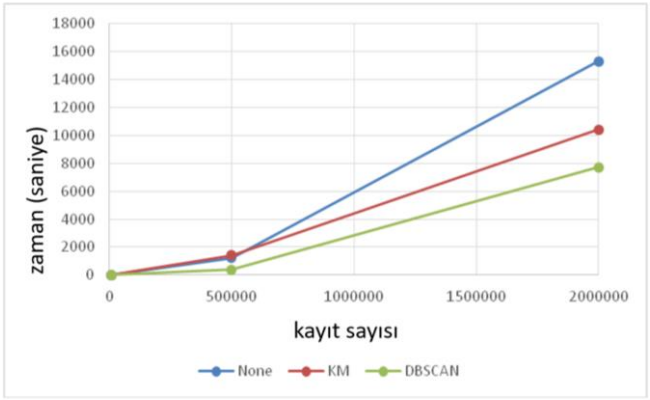
Deney (Kümeleme)	Kural sayısı	Yapım Süresi	Destek (%)			Güven (%)		
			Minimum	Maksimum	Ortalama	Minimum	Maksimum	Ortalama
A (none)	9066	1932,74	0,10	0,13	0,12	66,67	100	99,37
B (K-Ortalama)	1780	353,57	0,10	0,44	0,19	31,69	100	85,84
C DBSCAN)	4086	783,30	0,10	0,40	0,21	22,54	100	91,84

İlgili kuralların bulunmasını teminen pozitif kurallardan negatif birliktelik kurallarının bulunup alınması için bu bilgi teknolojisinin kullanılması amacıyla, önceki çalışmalarda alan bilgisi, el ile girilmiştir. Bu çalışmada, alan bilgisi, gözetimsiz makine öğrenme tekniği olan veri kümeleme kullanılarak otomatik olarak alınmıştır. Bu bilginin kullanımı, Deney B ve C'de gösterildiği gibi bulunan negatif birliktelik kuralının genel gücünü iyileştirmiştir. Bu iyileştirme Şekil 13'de çok iyi gösterilmiştir. Alan bilgisi olmadan negatif birliktelik kurallarının güven değeri yüksek seviyelerde yoğunlaşmasına rağmen, bu kuralın destek değeri, alan bilgisi keşfine göre bulunan değerlerden daha küçüktür.



Şekil 13. *Tüm yapılan deneyler için ortalama destek – güven değerleri.*

K-ortalama kümeleme yöntemi, çok yaygın olmasına ve bir çok uygulamada kullanılmasına rağmen, belli bir veri seti için kullanılacak optimal küme sayısını otomatik olarak belirleme kabiliyetine sahip değildir. Bu eksiklik, veri setini kümelemek için gereken algoritmanın karmaşıklığını arttırmaktadır. Aslında veri setini kümeler sayısı yelpazesine kümeleyerek, kümelerdeki bozulmayı ölçmek ve her kümeleme işlemindeki bozulmayı ölçmek önemlidir. Bu bozulmaya göre, kümelerin optimal sayısını tespit etmek ve sonrasında kümelerin bu sayısını kullanarak K-ortalama yönteminden yararlanarak kümelemek için dirsek yöntemi kullanılır. Bu yöntemin yapım süresi, DBSCAN kümelemeye göre madencilik yönteminin gerektirdiği yapım süresinden daha azdır. Ama bulunan kuralların sayısı, DBSCAN kümeleme yöntemine göre bulunan kuralların sayısından azdır. Dolayısıyla, DBSCAN kümeleme yöntemi kümelerine göre bulunan her bir negatif birliktelik kuralı için harcanan süre, 191.70 milisaniye iken, K-ortalama kümelemeye göre bulunan her bir kural için harcanan ortalama süre 198.63 mili saniyedir. Ayrıca, Her bir algoritma için Tablo 4'te ölçülen yapım süresi Şekil 14'de gösterilmiştir. Bu sonuçlar, DBSCAN yöntemi ifasının daha hızlı olduğunu, dolayısıyla da daha az karmaşık olduğunu göstermektedir. Bu davranışa, K-ortalama yöntemi kullanılarak kümelendirilen veri seti üzerinden çoklu geçiş ihtiyacı neden olur. İlk geçişte, veri seti Dirsek yöntemi yardımı ile kümelerin optimal sayısını seçmek için kümeleri farklı sayılara kümelendirilir. Öte yandan DBSCAN yönteminde kümelerin optimal sayısı, veri seti nesneleri kümelere dağıtıldığından bu yöntem ile belirlenir.

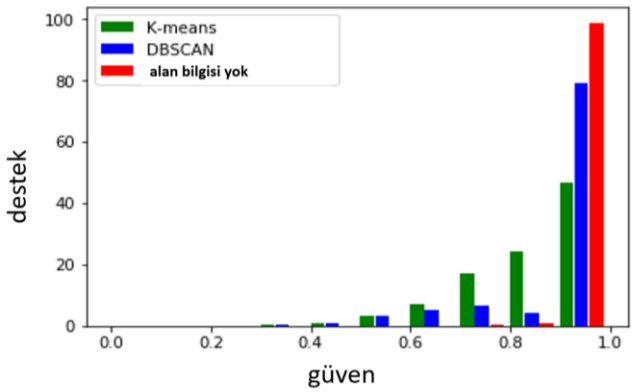


Şekil 14. *Farklı büyüklükte veri setleri kullanan her bir algoritmanın harcadığı yapım süresinin grafiksel gösterimi.*

DBSCAN kümeleme yöntemi, gereken kümenin önceden belirlenen özelliklerine göre kümenin optimal sayısını otomatik olarak belirleme kabiliyetindedir. Bu özellik, bir veri setinin kümeleme karmaşıklığını azaltır ve aynı zamanda dirsek yöntemi ile tespit edilen küme sayısına göre K ortalama ile kümeleme yöntemi ile karşılaştırıldığında iyileştirilmiş alan bilgisi de oluşturur. Bu da Şekil 10'da gösterilmiştir. Bulunan negatif birliktelik kuralının destek değerinin daha yüksek seviyelerde olduğu dikkat çekmektedir ve K ortalama kümeleme yöntemi ile sağlanan alan bilgisinde bulunanlara göre daha fazla kural bulma kabiliyetini de gösterir.

K ortalama yöntemine göre bu yöntemin yüksek üstünlüğünü de gösteren DBSCAN yöntemi için güven seviyesinde kuralların yüzde sayısının dağılımı, bulunan kuralların daha büyük olan kısmının yüksek seviyede güven değerinde olduğu Şekil 15'te gösterilmiştir. En yüksek seviyede bulunan kuralların en yüksek kısmının Deney A'dan çıkarılmış olmasına rağmen, Şekil 4 ve Şekil 5'de gösterilen kurallar güven değeri, daha önceki çalışmalarda önerilen hipotezlerle uyumlu olarak düşüktür. Herhangi bir önceki alan bilgisi olmayan negatif birliktelik kuralları madenciliği ilgisiz kurallar ile sonuçlanır.

K-ortalama yöntemi kullanılarak dirsek yöntemi yardımı ile hesaplanan küme sayısına göre bulunan negatif birliktelik kural desteği, DBSCAN kümeleme sonuçlarına göre bulunan kuralların ortalama destek değeri ile karşılaştırıldığında bulunan negatif birliktelik kurallarının yakın ortalama desteği olduğunu göstermektedir. Ancak bu kuralların ortalama güven değeri, DBSCAN kümeleme yöntemi ile sağlanan alan bilgisi kullanılarak bulunan güven değerinden önemli derecede küçüktür. Bu, DBSCAN kümeleme yöntemi kullanılarak bulunan alan bilgisinin, özellikle küme sayısı otomatik olarak bulunduğunda olmak üzere K-ortalama kümeleme yönteminden elde edilen değerden çok değerli olduğunu göstermiştir.



Şekil 12. *Tüm deneyler için güven seviyelerinde negatif birliktelik kuralının dağılımı.*

V. SONUÇ

Bu çalışmamızda, gözetimsiz veri madenciliği tekniği kümeleme kullanılarak, veri setinden otomatik olarak bulunup alınan alan bilgisine dayanan yeni bir negatif birliktelik kuralı madencilik tekniği önerilmektedir. K-ortalama ve DBSCAN kümeleme yöntemleri olmak üzere iki kümeleme yöntemi ile karşılaştırmalar yapıldı. Alan bilgisi olmaksızın, tek başına satın alınan her bir miktar kullanılarak bulunup çıkarılan negatif birliktelik kurallarının, bulunan kurallarda çok yüksek güven değerine sahip olduğu tespit edildi. Ancak bu kurallara ilişkin ortalama destek değer, diğer yöntemlerdeki değerden daha küçük olarak tespit edilmektedir. Bu durumda alan bilgisi kullanılmadığında ilgisiz kuralların bulunup çıkarılmasını göstermektedir. K-ortalama yöntemi ve küme optimal sayısını tespit etmek için dirsek yönteminin kullanılması ile oluşturulan miktar kümeye dayalı keşfedilen negatif birliktelik kurallarının destek değeri, DBSCAN kümeleme yöntemi ile oluşturulan kümelerin kullanılması ile bulunup çıkarılan kuralların ortalama destek değerine yakın olmak üzere bulunup çıkarılan kuralların ortalama destek değerinde önemli iyileşme göstermektedir. DBSCAN kümelerine göre bulunup çıkarılan negatif birliktelik kurallarındaki ortalama güven değeri, K-ortalama yöntemine göre bulunan güven değerlerinden çok daha yüksektir. Bu durumda negatif birliktelik kural madenciliğinde alan bilgisi kullanmanın önemini ve bu uygulama alanında DBSCAN kümeleme yöntemi ile bulunup çıkarılan alan bilgisinin, K-ortalama kümeleme yöntemi kullanılarak bulunup çıkarılan alan bilgisinden çok değerli olduğunu göstermektedir. Alan bilgisi olmaksızın bulunup çıkarılan negatif birliktelik kuralları ortalama değeri %0,12 iken K-ortalama ve DBSCAN alan bilgi çıkarımı durumunda ise sırasıyla %0,19 ve %0,21'dir. Ortalama güven değeri, alan bilgisiz madencilikte %99,37 iken, K-ortalama ve DBSCAN küme yöntemleri ile bulunup çıkarılan alan bilgisi kullanıldığında bu değer sırasıyla %85,84 ve %91,84'tür.

Gelecekteki araştırmalar için, işlem veri setinden bilgi bulup çıkarmak ve bu bilgiyi negatif birliktelik kuralları madenciliğinde kullanmak için, veri kümeleme tekniklerinin dışında farklı gözetimsiz makine öğrenme tekniklerinin kullanılması düşünülebilir.

VI. KAYNAKLAR

[1] A. L. Buczak ve E. Guven, "A survey of data mining and machine learning methods for cyber security intrusion detection," IEEE Communications Surveys & Tutorials, vol. 18, pp. 1153-1176, 2016.

[2] A. Holzinger ve I. Jurisica, "Knowledge discovery and data mining in biomedical informatics: The future is in integrative, interactive machine learning solutions," in Interactive knowledge discovery and data mining in biomedical informatics, ed: Springer, 2014, pp. 1-18.

[3] M. Hahsler ve R. Karpienko, "Visualizing association rules in hierarchical groups," *Journal of Business Economics*, vol. 87, pp. 317-335, 2017.

[4] Y. Zhao ve S.S. Bhowmick, "Association Rule Mining with R," A Survey Nanyang Technological University, Singapore, 2015.

[5] P. Kazienko, "Associations: discovery, analysis and applications", Oficyna Wydawnicza Politechniki Wrocławskiej, 2008.

[6] G. Suchacka ve G. Chodak, "Using association rules to assess purchase probability in online stores," *Information Systems and e-Business Management*, vol. 15, pp. 751-780, 2017.

[7] S. Datta ve S. Bose, "Discovering association rules partially devoid of dissociation by weighted confidence," Recent Trends in Information Systems (ReTIS), 2015 IEEE 2nd International Conference, 2015, ss. 138-143.

[8] S. Brin, R. Motwani, ve C. Silverstein, "Beyond market baskets: Generalizing association rules to correlations," in Acm Sigmod Record, pp. 265-276, 1997.

[9] A. Savasere, E. Omiecinski, ve S. Navathe, "Mining for strong negative associations in a large database of customer transactions," Data Engineering, 1998. Proceedings, 14th International Conference, 1998, pp. 494-502.

[10] X. Yuan, B. P. Buckles, Z. Yuan, ve J. Zhang, "Mining negative association rules," Computers and Communications, 2002. Proceedings. ISCC 2002. Seventh International Symposium, 2002, pp. 623-628.

[11] L.-M. Tsai, S.-J. Lin, ve D.-L. Yang, "Efficient mining of generalized negative association rules," in Granular Computing (GrC), 2010 IEEE International Conference, 2010, pp. 471-476.

[12] L. Aliahmadipour, V. Torra, ve E. Eslami, "On hesitant fuzzy clustering and clustering of hesitant fuzzy data," in Fuzzy Sets, Rough Sets, Multisets and Clustering, ed: Springer, 2017, pp. 157-168.

[13] M. B. Cohen, S. Elder, C. Musco, C. Musco, ve M. Persu, "Dimensionality reduction for k-means clustering and low rank approximation," in Proceedings of the forty-seventh annual ACM symposium on Theory of computing, 2015, pp. 163-172.

[14] T. M. Kodinariya ve P. R. Makwana, "Review on determining number of Cluster in K-Means Clustering," *International Journal*, vol. 1, pp. 90-95, 2013.

[15] R. Agarwal, ve R. Srikant, Fast algorithms for mining association rules. In Proc. of the 20th VLDB Conference, 1994, pp. 487-499.

[16] G. Sheng, H., Hou, X, Jiang, ve Y. Chen (). A novel association rule mining method of big data for power transformers state parameters based on probabilistic graph model. IEEE Transactions on Smart Grid, vol. 9, no. 2, pp. 695-702, 2018.

[17] T.D.B., Le, ve D. Lo, "Beyond support and confidence: Exploring interestingness measures for rule-based specification mining", In Software Analysis, Evolution and Reengineering (SANER), 22nd International Conference, *IEEE*, 2015, pp. 331-340.

[18] Artamonova, I. I., Frishman, G., and Frishman, D.. "Applying negative rule mining to improve genome annotation," *BMC bioinformatics*, vol. 8, no.1, ss. 261, 2007.

[19] M. Ester, H.P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," *In Kdd*, vol. 96, no. 34, pp. 226-231, 1996.

[20] T.T., Bilgin, ve Y. Çamurcu, "DBSCAN, OPTICS ve K-Means Kümeleme Algoritmalarının Uygulamalı Karşılaştırılması,", *Politeknik Dergisi*, vol. 8, no. 2, 2005.