TÜRK HAVA KURUMU ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

GÖZETİMSİZ MAKİNE ÖĞRENİM TEKNİKLERİ İLE MİKTARA DAYALI NEGATİF BİRLİKTELİK KURAL MADENCİLİĞİ

YÜKSEK LİSANS TEZİ

Zahraa Mohammed Malik MALIK

1406010005

Elektrik ve Bilgisayar Anabilim Dalı

Elektrik ve Bilgisayar Mühendisliği Programı

Ocak 2018

TÜRK HAVA KURUMU ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

GÖZETİMSİZ MAKİNE ÖĞRENİM TEKNİKLERİ İLE MİKTARA DAYALI NEGATİF BİRLİKTELİK KURAL MADENCİLİĞİ

YÜKSEK LİSANS TEZİ

Zahraa Mohammed Malik MALIK

1406010005

Elektrik ve Bilgisayar Anabilim Dalı

Elektrik ve Bilgisayar Mühendisliği Programı

Tez Danışmanı: Doç. Dr. Tansel Dökeroğlu

Türk Hava Kurumu Üniversitesi Fen Bilimler, Enstitüsü'nün 1406010005 numaralı Yüksek Lisans öğrencisi, "Zahraa mohammed malik", ilgili yönetmeliklerin belirlediği gerekli tüm şartları yerine getirdikten sonra hazırladığı "Gözetimsiz Makine Öğrenim Teknikleri ile Miktara Dayalı Negatif Birliktelik Kural Madenciliği" başlıklı tezini, aşağıda imzaları olan jüri önünde başarı ile sunmuştur.

Tez Danışmanı : **Doç. Dr. Tansel Dökeroğlu** .....
Türk Hava Kurumu Üniversitesi

Eş Danışman : **Yrd. Doç. Dr. Shadi AL-SHEHABI** .....
Türk Hava Kurumu Üniversitesi

Jüri Üyeleri : **Prof. Dr. Ahmet Coşar**
Orta Doğu Teknik Üniversitesi .....
Bilgisayar Mühendisliği

: **Doç. Dr. Tansel DÖKEROĞLU**
Türk Hava Kurumu Üniversitesi .....
Bilgisayar Mühendisliği

: **Yrd. Doç.Dr. Meltem Yıldırım İMAMOĞLU**
Türk Hava Kurumu Üniversitesi .....
Bilgisayar Mühendisliği

Tez Savunma Tarihi: 18/1/2018

TÜRK HAVA KURUMU ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ MÜDÜRLÜĞÜ'NE

Yüksek Lisans Tezi olarak sunduğum, "Gözetimsiz Makine Öğrenim Teknikleri ile Miktara Dayalı Negatif Birliktelik Kural Madenciliği" adlı çalışmamın, tarafımdan akademik etik ve kurallara aykırı düşecek bir yardıma başvurmaksızın yazıldığını ve yararlandığım kaynakların kaynakçada gösterilenlerden oluştuğunu, bunlara atıf yapılarak yararlanılmış olduğunu belirtir ve bunu onurumla doğrularım.

Zahraa Mohammed Malik MALIK
Ocak 2018

ÖNSÖZ

Bu vesile ile, bu çalışmamın tüm aşamalarında gösterdiği sabır ve sunduğu destek için Doç. Dr. Tansel Dökeroğlu'na teşekkürlerimi sunmak isterim. Zorluklarla mücadelemde sunduğu rehberlik, deneyim ve cesaretlendirmeleri için Yrd. Doç. Dr. Shadi AL-SHEHABI'ye en samimi şükranlarımı sunarım. Akademik ve yaşam yolculuğum boyunca elimi hiç bırakmayan dostum, sevgili eşime de teşekkür ederim. Aileme ve güzel gülümsemesi ile hayatımı aydınlatan küçük yıldızım kızım Tala'ya sevgi ve minnetle…

İÇİNDEKİLER

TABLOLAR LİSTESİ

ŞEKİLLER LİSTESİ

ALGORİTMA LİSTESİ

KISALTMALAR LİSTESİ

AR	: Birliktelik Kuralları (Association Rules)
DBSCAN	:Gürültülü Uygulamaların Yoğunluk Tabanlı Uzaysal Kümelenmesi (Density-Based Spatial Clustering of Applications with Noise)
DM	: Veri Madenciliği (Data Mining)
ML	: Makine Öğrenimi (Machine Learning)
NAR	: Negatif Birliktelik Kuralları (Negative Association Rules)
PAR	: Pozitif Birliktelik Kuralları (Positive Association Rules)

ÖZET

Gözetimsiz Makine Öğrenim Teknikleri ile Miktara Dayalı Negatif Birliktelik Kural Madenciliği

MALIK, Zahraa Mohammed Malik
Yüksek Lisans, Elektrik ve Bilgisayar Mühendisliği Bölümü
Danışman: Prof. Dr. Tansel Dökeroğlu
Eş Danışman: Yrd. Prof. Dr. Shadi AL-SHEHABI
Ocak 2018, 49 Sayfa

Birliktelik kuralları, veri kümesindeki nesnelerin varlığının diğer nesnelerin varlığını nasıl etkilediğini tanımlanmaktadır. Bu kurallar, alış veriş sepetleri analizinde, bir ürünün aynı işlemdeki diğer ürün üzerindeki etkisini incelemek için yaygın olarak kullanılmaktadır. Pozitif ve negatif birliktelik kuralları olarak iki şekilde ifade edilebilirler. Bir çeşitten diğerine pozitif birliktelik kuralı, bir ürün varlığının aynı işlemde diğer ürünü bulma olanağını arttırdığını gösterirken, negatif birliktelik kuralı, bir çeşidin bulunmasının, diğer ürünün aynı işlemde olabilme ihtimalini düşürdüğünü göstermektedir.

Daha önceki işlemlerdeki sıklıkların araştırdığı için pozitif birliktelik kuralı madenciliği, negatif birliktelik kuralları madenciliğine göre nispeten daha kolaydır. Negatif birlikteki kuralı madenciliğinde daha önceki işlemler araştırıldığında, ilgisiz ürünler arasındaki ilgisizlik kurallarının madenciliği ile karşılaşılır. Bu kuralların çıkarımından kaçınmak için, mevcut negatif birliktelik kuralı, madencilik tekniklerine sağlanan önceden tanımlı alan bilgisine dayanır. Dolayısıyla bu bilgi, bulunan kuralların ilgili ürünlere ait olması için kullanılır.

Bu çalışmada, satın alınan miktarlara dayalı veri kümesinden otomatik olarak bilgi alınması ile veri kümesindeki ürünler arasındaki ilginç negatif birliktelik kurallarını bulma kabiliyetine sahip yeni bir teknik önerilmektedir. Birliktelik kuralı

madenciliği, gözetimsiz veri madencilik tekniği olduğundan, sağlanan veri kümesi etiketsiz verilerden oluşmaktadır.

Bu çalışmada, *K-Ortalama(k means)* ve *Gürültülü Uygulamaların Yoğunluk Tabanlı Uzaysal Kümelenmesi* (*DBSCAN*) olmak üzere iki yöntem test edilmiştir. Bu tekniklerin sonuçları, herhangi bir alan bilgisi olmaksızın bulunan negatif birliktelik kuralı sonuçları ile karşılaştırılmıştır. *DBSCAN* kümeleme yönteminin kullanımı, gerçek yaşam işlem veri tabanında test edildiğinde %0.21 destek ve %91.84 güven ortalama değerleri ile 4,086 şeklinde daha iyi negatif birliktelik kuralı sonucu göstermiştir. *K-Ortalama* kümeleme yönteminin kullanımı ile çıkarılan alan bilgisine dayalı negatif birliktelik kuralları madenciliği sonucu, %0.19 destek ve %85.84 güven ortalama değerine sahip 1,780 iken, alan bilgisiz negatif birliktelik kuralı sonucu %0.12 destek ve %99.37 güven ortalama değerli 9,066 sonucunu vermiştir.

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

ABSTRACT

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

MALIK, Zahraa Mohammed Malik

Master, Department of Electrical and Computer Engineering

Supervisor: Prof. Dr. Tansel Dökeroğlu

Co-Supervisor:Asst. Prof. Dr. Shadi AL-SHEHABI

January 2018, 49 page

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 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. A positive association rule from one item to another indicated that the existence of that item increases the chance to find the other in the same transaction, while 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, compared to mining negative association rules, 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 mining

association rules is an unsupervised data mining technique, the provided dataset is unlabeled data.

Two clustering methods are tested in this study, which are the K-means and Density-Based Spatial Clustering of Applications with Noise (DBSCAN) methods. The results of these techniques are compared to the results of extracting negative association rules without any domain knowledge. 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.

BİRİNCİ BÖLÜM

GİRİŞ

Bilgisayarların sunduğu yüksek performans ve doğruluk ile farklı alanlarda bilgisayar kullanımının hızla büyümesi, bu uygulamalar için büyük miktarda veri elde edilmesini de beraberinde getirmiştir. Bu verilerin bir kısmı, bu uygulama alanlarında sağlanan hizmetlerin iyileştirilmesinde destek olabilecek çok değerli bilgiler içerir. Bu bilgilerin bulunup alınması, makine öğrenmeye dayanan veri madencilik teknikleri kullanılarak gerçekleştirilir. Makine öğrenme teknikleri, veri kümesinden istenen bilgiye bağlı olarak gözetimli ve gözetimsiz olmak üzere iki ana kategoriye ayrılabilir. Gözetimli makine öğrenme, ön veri etiketlemeyi gerektirir. Dolayısıyla bu teknikler veri kümesindeki nesnelerin özellik değerleri arasındaki örüntü ve ilişkiler ile bunlara verilen etiketi tespit eder. Gözetimsiz makine öğrenme ise, veri etiketlemeyi gerektirmez. Çünkü nesneler arasındaki ilişkiyi ve veri kümesindeki değerlerin sıklık örüntüsünü araştırarak bilgiyi bulup getirir. Yaygın olarak kullanılan madencilik tekniklerinden birisi, birliktelik kural madenciliğidir [1, 2].

Veri madenciliği, en yaygın makine öğrenme alanlarından birisidir. Bir veri kümesindeki nesneleri birleştiren veya bağlayan ilişkileri araştırarak bilgiyi büyük veri kümelerinden bulup çıkarır. Diğer makine öğrenme tekniği gibi, veri madencilik tekniği, gözetimli veya gözetimsiz teknik olabilir. Gözetimli veri madenciliği teknikleri, etiketli veri gerektirir. Etiketli veri, veri kümesindeki her nesne için bir etiket belirlenmesi gerekliliğidir. Veri madencilik tekniği sonrasında bir nesneyi karakterize eden her bir özellik değeri ile bu nesneye verilen etiket arasındaki ilişkileri araştırır. Böylece çıkarılan bilgi, nesne üzerindeki özellik değerlerine bağlı olarak yeni bir nesne için bir etiket belirlenmesi amacıyla kullanılabilir. Bu da, öngörülen etiket kullanılarak yeni nesnelerin gelecekteki davranışının kestirilebilmesine imkân tanır. Bu makine öğrenme teknikleri sınıflandırıcılar olarak bilinmektedir.

Öte yandan, kümeleme ise, gözetimsiz bir makine öğrenme tekniğidir. Kümeleme, veri kümesindeki nesnelerin gruplar halinde dağıtılması işlemi olup, bu dağıtımda her bir gruptaki nesneler, diğer gruplardaki nesnelere göre daha çok birbirlerine benzerdir. Bu gruplar kümeler olarak bilinmektedir. Kümeleme, önceden veri etiketlemesi gerektirmez ve kümeleme işlemi, sadece veri kümesindeki nesnelerin özel değerlerine bağlıdır. Veri kümeleme, çok önemli veri madencilik alanıdır çünkü veri etiketlemeye gerek olmaksızın veri kümesinden bilgi çıkarımı yapma kabiliyetindedir.

Birliktelik kuralları, bir diğer gözetimsiz veri madencilik tekniğidir. Bu teknikte veri etiketlemeye gerek olmaksızın nesne ile özellikler arasındaki ilişkiler araştırılır. Birliktelik kuralı madencilik teknikleri, kalemlerin aynı işlemde bir araya getirildiği işlem veri setlerini analiz için yaygın olarak kullanılmaktadır. Birliktelik kuralı, aynı alış-veriş sepetindeki iki kalem (çeşit) arasındaki ilişki olarak tanımlanır. Bir kalemin sepette bulunması, diğer kalemin 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 kalemin varlığı, ardıl (bağlı) kalem olarak bilinen diğer kalemin olma ihtimalini arttırdığında, iki kalem arasında pozitif birliktelik kuralı olduğu söylenir. Ayrıca, bir sepette öncül kalemin varlığı, ardıl kalemin aynı sepette veya işlemde olmaması olasılığını arttırdığında ise n-öncül kalemin ardıl kalem ile negatif birliktelik kuralından söz edilir. Birliktelik kuralları, simetrik değildir, dolayısıyla, X kalemi ile Y kalemi 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 çıkarımı kolaydır. Pozitif birliktelik kuralları, söz konusu iki kalemin 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 kalemleri bir araya getirdiğinde görülür. Yani ardıl kalemin yokluğuna neden olur. Çünkü bu kalemler birbirleri ile bağlantılı ve öncül kalemin etkisine göre değildir. Bu kuralların çıkarımından kaçınmak için, bulunan negatif birliktelik kurallarının ilgili ve yararlı kural olduğunun temin edilmesi, alan bilgisi gerektirmektedir.

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.

Bir veri setindeki nesneleri daha homojen gruplara dağıtmak için kullanılan birçok kümeleme algoritması bulunmaktadır. *K-Ortalama*, veri setinin kümelenmesinde kullanılan en yaygın algoritmalardan birisidir. Bu teknik, veri seti nesnelerini bölmek için kullanılan küme sayısının önceden belirlenmesini gerektirir. Kümeleme, gerekli küme sayısına eşit bir rastgele sentroid kümesi ile başlayarak gerçekleştirilir. Veri setindeki nesneler, kümelere dağıtılır. Her bir nesne, en yakın sentroid kümesinde yer alır. Sonrasında her bir kümeye ait mesafelerin toplamı hesaplanır ve sentroid yerinin optimize edilmesi için kullanılır. Tüm sentroidler için en iyi yer bulununcaya kadar mesafe toplamı minimize edilerek devam edilir. Bu noktada, nesneler, bu sentroidlere göre kümelendirilir.

Bu teknik, optimal küme sayısını otomatik olarak belirleme kabiliyetine sahip olmadığından, (K-Ortalama kümeleme yöntemi kullanarak) bu veri setini kümelemeden önce veri setine ait küme optimal sayısını bulmak için bu yöntem ile birlikte dirsek yöntemi yaygın olarak kullanılmaktadır. Dirsek yöntemi, veri setinin bir kümeye kümelenmesi ile başlar, sonrasında mesafelerin toplamı hesaplanır. Küme sayısını arttırdığından her defasında bir küme için yapılır. Daha sonra, mesafelerin önceki adımdakinden çok daha az olması durumundaki noktayı bulmak için hesaplanan toplamdaki varyasyon ölçülür. Bu nokta, ölçülen değer değişiklik oranının birden düştüğü bir dirsek benzeri noktadır. Dirsek yöntemi denmesinin nedeni budur [5, 6].

Diğer bir yaygın kümeleme yöntemi, Gürültülü Uygulamaların Yoğunluk Esaslı Uzaysal Kümelenmesidir (DBSCAN). Bu yöntem, aynı kümedeki nesneler arasındaki maksimum mesafe ve bir kümedeki nesnelerin minimum sayısı olan, önceden tanımlı bir ayarlamayı kullanarak her farklı veri seti için optimal küme sayısını otomatik olarak tespit etme kabiliyetine sahiptir. Bu ayarları kullanan algoritma, rastgele bir nesneden kümeleme yapmaya başlar ve sonrasında bu nesneye algoritma için belirlenen maksimum mesafeden daha yakın olan tüm nesneleri birleştirir. Sonra, yeni bir küme/demet oluşturmaya başlamak için henüz kümelendirilmemiş olan başka bir

nesne seçilir. Sonunda, her bir veri seti için farklı küme sayısı kullanılarak tüm nesneler kümelerde/demetlerde dağıtılır.

1.1 Araştırmanın Amacı

Veri madenciliği teknikleri, spesifik bir alan performansının iyileştirilmesi için kullanılabilecek faydalı bilgiyi arayıp bulmak amacıyla farklı uygulama alanları için toplanan verileri analiz için yaygın olarak kullanılır. Veri setinden değerli bilgiyi bulup çıkarmak için, nesneler arasındaki güçlü ilişkileri ve veri seti özelliklerini bulmak önemlidir. Birliktelik/ilişkilendirme kuralları 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 madenciliğinin karşılaştığı zorluğa göre olup, ilgili olmayan birliktelik kurallarının bulunmasıdır ve ilgili olmayan nesneleri veri setinde birleştiren kurallardır. Bu kurallardan kaçınmak için, alanın önceki bilgisine gerek vardır. Mevcut negatif birliktelik kuralı bilgisinin bulunması, negatif birliktelik kuralları bilgisinin bulunması için modele sağlanan bilgiye bağlıdır.

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 önermek amaçlanmaktadır. Önerilen yöntem, her bir kalemin satın alındığı miktar ile ilgili bilginin bulunmasına dayanır. Sonrasında negatif birliktelik kuralları, satın alınan miktarların dağılımından elde edilen bilgiden çıkarılır.

Bu bilgi, işlem veri setindeki her bir kaleme ait işlem başına satın alınan miktarların homojen gruplarını bulmak için uygulanan kümeleme yöntemleri kullanılarak çıkarılır. Bu grupları bulmak için, bu kalemin mevcut satın alınan miktarları ve veri seti ile miktar sıklığını kullanarak bu veriler için histogramlar oluşturulur. Sonrasında, kalem kümelerinin dağılımı arasındaki ilişkileri araştırmak için bu histogramları kümelemek için 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 sorunun önüne geçmek için ilk olarak pozitif birliktelik kuralları aranır. Bu pozitif birliktelik kuralları,

birliktelik kuralındaki kalemler arasında ilişki olduğunu doğrular. Sonrasında bu kalemler, satın alınan miktarlar histogramları kümeleme sonuçlarına göre gruplara ayrılır. Ardından birbirlerine birleştiren pozitif birliktelik kuralına sahip kalem grupları arasındaki negatif birliktelik kuralları araştırılır. Bu işlem, önceden alan bilgisine gerek olmaksızın ilginç negatif birliktelik kurallarını oluşturur.

Bu araştırmanın geri kalanı, aşağıdaki sunulduğu şekilde yapılandırılmıştır: İkinci bölümde, araştırmada kullanılan veri madencilik teknikleri ile ilgili literatür incelemesi verilmiştir. Üçüncü bölümde birliktelik kuralları ile nesneleri nasıl veri setinde birleştirdikleri ve satın alınan miktarlar histogramlarında kullanılan kümeleme teknikleri incelenmiştir. Dördüncü bölümde, önerilen yöntem açıklanmaktadır. Beşinci bölümde yapılan deneylerin sonuçları gösterilip tartışılmıştır. Altıncı bölümde bu araştırmanın sonuçları ve gelecekte yapılacak çalışmalar sunulmuştur.

İKİNCİ BÖLÜM

LİTERATÜR İNCELEMESİ

Makine öğrenmesi, programa ve çıktılardaki sonuçlara göre spesifik bir yolda girdilerle etkileşim içinde olan, önceden konfigüre edilen programlara gerek olmaksızın bilgisayarlara verileri analiz etme, karar verme ve/veya önlem alma kabiliyeti sağlayan teknikleri inceleyen bir bilim alanıdır. Makine öğrenme yöntemleri, programa dahil edilmeyen durumlarla etkileşime girme kabiliyetine sahiptir [8]. Veri madenciliği, veri setindeki özellik (feature) değerleri arasındaki ilişki ve örüntüleri tespit etmek için veri setlerini analiz eden makine öğrenme alanıdır. Veri madenciliği teknikleri, diğer makine öğrenme teknikleri gibi, gözetimli ve gözetimsiz olmak üzere iki ana kategoriye ayrılabilir. Gözetimli veri madenciliği, etiketlenmiş veri gerektirir. Yani, veri madencilik tekniklerinin, veri gruplarının ilgili sınıfa girmelerine neden olan örüntüleri bulup çıkarabilmek için veri gruplarının sınıflara tasnif edilmesidir. Bu örüntü ve ilişkiler, veri madencilik teknikleri ile bulunup çıkarılan bilgi olarak bilinir ve herhangi bir yeni veri grubunu sınıflandırmak ve özellik değerlerinin ait olduğu sınıfın kestirilmesinde kullanılabilir. Bu sınıflandırma, girmesi kestirilen sınıfa göre yeni veri grubunun ilerideki davranışının kestirilmesinde kullanılabilir [9-11]. Gözetimsiz makine öğrenme teknikleri, etiketsiz veri setinde öznitelik değerleri arasındaki örüntü ve ilişkileri bulur. Yani bu teknikler, toplanan veri setine herhangi bir ilave gerektirmez [11, 12].

En yaygın gözetimsiz makine öğrenme tekniklerinden birisi, birliktelik kuralı madenciliğidir. Birliktelik/ilişkilendirme, birlikteliğin önemi ve niteliğine bağlı olarak ilişkinin gücünü belirleyen belli ölçümlerin kullanımını temsil eden, münferit nesneler veya nesne grupları arasındaki bir ilişkidir [13]. Birliktelik kuralları, bu kuralların *destek* ve *güven* değerlerinin hesaplanması ile bir veri setindeki sıklık örüntülerini belirten şartlar kümesidir. Bir işlemler veri setinden birliktelik kurallarının tespiti, her bir işlemdeki kalemler arasındaki ilişkiyi bulmak için çok yaygındır [14].

Destek, veri setindeki toplam işlemler sayısına göre kalemi veya kalemler kümesini içeren işlemler sayısını temsil eder [15]. Dolayısıyla, bir T işlem veri setindeki bir X kalemi için aşağıdaki gibi hesaplanır:

$$support(X) = \frac{|X|}{|T|} \qquad \text{denklem(2.1)}$$

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 aşağıdaki gibi hesaplanır

$$support(X \Rightarrow Y) = \frac{|X \cup Y|}{|T|} \qquad \text{denklem(2.2)}$$

$X \Rightarrow Y$ 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 [15, 16]. Kural güven değeri $X \Rightarrow Y$ aşağıdaki eşitlik kullanılarak hesaplanır.

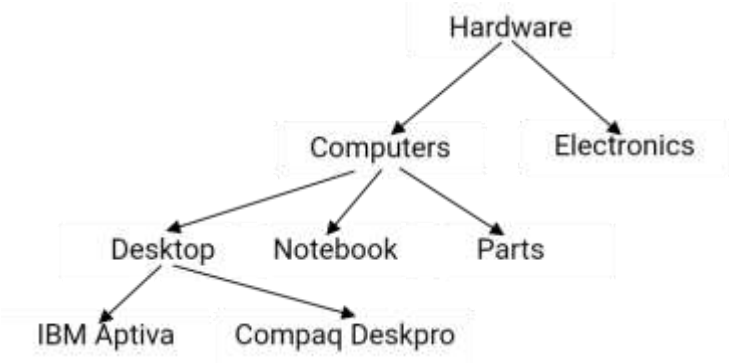
$$confidence(X \Rightarrow Y) = \frac{|X \cup Y|}{|X|} \qquad \text{denklem(2.3)}$$

Birliktelik kuralları, pozitif ve negatif birliktelik kuralları olmak üzere iki kategoriye ayrılabilir. Bir öğeden diğer öğeye pozitif birliktelik kuralı, öncül kalemin varlığının, aynı işlemdeki ardıl kaleme sahip olma ihtimalini arttırmasını önerir. Öte taraftan bir negatif birliktelik kuralı tersini belirtir. Bir öncülün varlığı, aynı işlemde ardıl öğenin olma ihtimalini düşürür veya öncül öğenin yokluğu, diğer öğenin aynı işlemde olma ihtimalini arttırır [17-19]. Negatif birliktelik kuralları, ilk olarak [20]'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, [21]'de 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 (sınıflandırma) veri setinden otomatik olarak keşfedildiğinden, negatif birliktelik kurallarını bulmak için algoritmaya el ile sağlanması gerekmektedir.

Ayrıca, [22]'de 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 kalemler, algoritmaya sağlanan kalemlerin taksonomisine dayalı benzerliklerine göre gruplandırılır. Şekil 2.1'de gösterilen örnek taksonomide gösterildiği gibi, araştırma taksonomi ağacında üst seviyeden alt seviyeye indikçe, bu gruplar daha spesifik hale gelir. İki kalem 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 kalemler taksonomisinin bir tanımı da gerekir.



Şekil 2.1: [22]'de önerilen yöntemin gerektirdiği örnek tasonomi.

[23]'te önerilen yöntem, kalemlerin markalarına göre negatif birliktelik kurallarını araştırır. Dolayısıyla belli bir markadan bir kalem satın alan müşteri, başka bir kalemi satın alma eğilimindedir. Ancak spesifik bir markayı satın almaktan kaçınır. Örneğin, atıştırmalıklar ile meşrubatlar arasında pozitif birliktelik kuralı bulunur ve bu ürünlerin taksonomisi, her bir kalemin 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 aynı zamanda [24, 25] gibi bazı diğer teknikler, ilk olarak pozitif birliktelik kurallarının keşfine dayalı bu kuralların çıkarılmasına dayanmaktadır. Bu işlem, birbirleri ile ilgili olmayan kalemlere ilişkin bu kuralları çıkarmak yerine ilgili negatif birliktelik kurallarının bulunmasını temin ederken, pozitif birliktelik kuralları da adaylar arasında bir ilişki olduğunu temin eder. Daha sonra, öncül ve ardıl kalemlerin spesifik özelliklerine göre daha belirgin bir araştırma yapılır. Bu yöntemler, negatif birliktelik kuralları araştırılmasında kullanmak için bu kalemlerin önceden özelliklerinin ve aralarındaki ilişkilerin tanımlanmasını gerektirir. Öte yandan pozitif birliktelik kuralı araştırması için bu bilgiye gerek yoktur.

Bir başka önemli makine öğrenme alanı, veri kümeleme olup bu da aynı zamanda veri setine herhangi bir etiketleme gerektirmeyen gözetimsiz makine öğrenme veri madenciliğidir. Veri kümeleme, nesnelerin gruplara ayrılması olarak tanımlanır. Bir gruptaki bir nesne, başka gruptaki başka bir nesneye göre bu gruptaki diğer nesnelere daha çok benzerdir [26, 27]. Dolayısıyla, her bir küme, bu nesneleri karakterize eden özelliklere ait değerler ile ilgili olarak daha çok homojen nesne içerir.

Veri kümeleme gözetimsiz makine öğrenme tekniği olduğundan, bu nesneleri belli bir kümeye gruplayan ilişkiler, kümeleme yöntemi olarak bilinmemektedir ve özellikle daha büyük veri setlerinde insanların dikkatini çekmesi zordur [28, 29].

K-Ortalama, veri setlerini sayısal öznitelikler ile kümeleme için kullanılan en etkin veri kümeleme yöntemlerinden birisidir. Bu yöntem, n-boyutlu uzayda K rastgele sentroidlerini başlatarak önceden tanımlı "K" kümesini oluşturur. Burada n, veri setindeki öznitelikler sayısıdır. Sonrasında, 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 [30, 31].

Mesafe toplamının hesaplanması için kullanılan eşitlik:

$$W(X,C) = \sum_{k=1}^{K} \sum_{j \in X_k} \|Y_j - C_k\|^2 \qquad \text{denklem(2.4)}$$

Burada X, veri setinin çok boyutlu uzamında Y vektörü ve sentroid C'nin K sayısı kullanılarak belirlenen nesneler grubudur.

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 kümelendirildiğinde, 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 sayısının yaptığı kadar çok olmadığıdır. Dolayısıyla dirsek noktası olarak bilinen bu noktadaki kümelerin sayısı, kümelerin optimal sayısı olarak seçilir [5, 32].

Bir veri seti için küme sayısını otomatik olarak tespit etme kabiliyetine sahip bir başka veri kümeleme yöntemi, Gürültülü Uygulamaların Yoğunluk Esaslı Uzaysal Kümelenmesi (*DBSCAN*) yöntemidir, Ester et al. tarafından önerilen 1996'da [33]. Bir kümede minimum sayıda nesne ve bir kümedeki iki nesne arasında maksimum mesafe

sağlanması ile algoritma için ayarlanan maksimum mesafe içindeki tüm nesneler bu kümeye toplanıncaya kadar, bu kümedeki nesneleri toplamaya başlamak için algoritma bir nesne seçer. Sonra, aynı işlem için bir başka kümelenmemiş nesne seçilir. Bu işlem, tüm nesneler kümelere gruplandırılıncaya kadar tekrar edilir [34-35].

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 almak 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 [37-39].

ÜÇÜNCÜ BÖLÜM

BİRLİKTELİK KURALLARI VE VERİ KÜMELEME

3.1 Birliktelik Kuralları

Birliktelik kuralları madenciliği, büyük işlem veri setinden bilgi tespiti ve keşfedilen bilginin alınması için kullanılan en yaygın tekniklerden biridir [40]. $X \Rightarrow Y$, form gösterimi olarak tanımlanabilir. Burada X ve Y, $T = \{t_1, t_2, t_3, \ldots, t_n\}$ işlem veri setinde görülen $I = \{i_1, i_2, i_3, \ldots, i_n\}$ öge setinden; her bir işlemin kendisine özgü bir kimlik numarasının olduğu ve her bir alım başına bir veya daha fazla öge içerdiği öge veya öge grupları vardır. Kuralın sol tarafındaki öge veya öge seti, öncül olarak bilinirken sağ taraftaki ise ardıl olarak bilinir. Daha iyi anlaşılması için Tablo 3.1'de gösterilen veri seti, bu bölüm boyunca bir örnek olarak kullanılmaktadır.

Tablo 3.1: İşlem veri seti örneği.

İşlem Kimliği	Öge Kodu	Miktar
1	A	1
1	B	4
1	C	2
2	A	5
2	D	3
3	B	1
3	C	3
3	D	1
4	A	2
4	B	3
4	C	1
5	A	8
5	D	4

Birliktelik kuralı madenciliği genel olarak iki adımlı bir işlem ile elde edilir. İlk adım, işlemler veri setindeki her bir ögenin destek değerinin hesaplanmasını içerir. Burada destek değeri, bu öge veya öge setinin görüldüğü işlemin, işlem veri setindeki işlem toplam sayısına oranıdır. Hesaplanan destek değeri, önceden tanımlanan eşik değerden büyükse, bu öge sık öge olarak değerlendirilir ve birliktelik kurallarını almak için sonraki adımda işlenir. Büyük olmayan ögeler, bu kural olsa bile bunlardan bulunup çıkarılan kural için yeterli desteğe sahip olmadığından ihmal edilir. Daha sonra aday birliktelik kuralları, daha önce bulunup çıkarılan ögelerin tüm olası kombinasyonlarının hesaplanması ile çıkarılır. Birliktelik kural destek değeri $X \Rightarrow Y$ aşağıdaki eşitlik kullanılarak hesaplanır:

$$s(X \Rightarrow Y) = \frac{|X \cup Y|}{|T|} \qquad \text{denklem(3.1)}$$

İkinci adımda, yeterli güven değerine sahip adayları geçerli kılmak için önceki işlemden adayları araştırır ve $X \Rightarrow Y$ kuralı için, aynı işlemdeki X ögesi ile birlikte görülen Y ögesinin sıklığını gösterir. Güven, aşağıdaki denklem kullanılarak hesaplanır:

$$c(X \Rightarrow Y) = \frac{|X \cup Y|}{|X|} \qquad \text{denklem(3.2)}$$

3.1.1 Pozitif Birliktelik Kuralları

Bir öge veya öge seti, X kabul edilebilir destek ve güven değerine sahip olduğu tespit edildiğinde Y ögesi ile pozitif birliktelik kuralına sahip olduğu söylenir. Bu kural, işlemde X ögesinin varlığının, aynı işlemde Y ögesinin varolma ihtimalini arttırdığını belirtir. Öte yandan, Y ögesinin varlığı, aynı işlemde X ögesinin varolma ihtimali artışını gösterebilir veya göstermeyebilir. Birliktelik kuralları simetrik değildir, yani belli yöndeki bir kural, aynı kuralın ters yönde varolma zorunluluğunu getirmez.

İşlemler veri setinde sıklık öge setlerinin araştırılmasında, birliktelik kuralları madenciliğinde en popüler algoritmalardan birisi *Apriori* algoritmasıdır. Önceden tanımlı destek değerinden daha yüksek destek değerine sahip ögeler, kabul edilebilir

destek değerine sahip tüm diğer ögelerle bu öge için destek değeri ölçülerek sonraki öge kümelerine atılır. Önceden belirlenen seviyenin altında destek değere sahip ögeler, kabul edilemez destek değere sahip ögeye göre kabul edilebilir destek seviyesi ile sık öge seti oluşturmanın imkânsız olması durumuna göre atılır. Bu, sık öge kümesindeki tüm ögelerin sık ögeler olduğunu da gösterir ve bir öge seti destek değeri her zaman bu öge setindeki en az destekli öge destek değerinden daha küçüktür. Bu da anti monoton özellik olarak bilinir [44].

Bu özelliklere göre Apriori algoritması işlem veri setindeki her bir öge için destek değerini hesaplayarak aday öge seti aramasını başlatır. Belirlenen eşik değerin altında destek değere sahip ögeler sonraki adımda azaltma işleminden çıkarılır. Çünkü istenen kritere uymayan destek değerli bir ögeyi dahil ederek kabul edilebilir destek değerli bir öge seti oluşturmak imkansızdır. Bu işlem, Apriori algoritmasının tüm sık öge setlerini daha hızlı oluşturmasına imkân tanır. Dolayısıyla daha büyük veri setlerini işlemek mümkün hale gelir. Apriori azaltma algoritması Algoritma 3.1'de gösterilmiştir.

Algoritma 3.1: Apriori algoritması [21].

1: Girdi ögeleri (I), minimum destek (minSup)
2: J=1.
3: Sık ögeleri al. $\mathrm{F_j} = \{\mathrm{i} | \mathrm{i} \in \mathrm{I} \;\&\; \frac{|\mathrm{i}|}{|\mathrm{T}|} \geq \mathrm{minSup}\}$.
4: $F_j \neq \emptyset$ iken
5: J=J+1
6: Önceki sık öge setlerinden aday oluştur $\mathrm{C_j} = \mathrm{apriori.gen}(\mathrm{F_j}\text{-}1)$
7: i = 1 ila |T| için
8: $C_i = subset(C_j, t_i)$
9: Her bir aday için $c \in C_t$
10: $\mathrm{S_c} = \mathrm{S_c} + 1$
11: İçin son
12: İçin son
13: $F_j = \{c | c \in C_j \;\&\; \frac{S_c}{|T|} \geq minSup\}$
14: İken son
15: F Dön

Tablo 3.1'de gösterilen örnek veri seti için, her bir münferit ögenin destek değeri Tablo 3.2'de gösterilmiştir. Örneğin, öge A, beş işlemin dördünde görünmektedir. Dolayısıyla destek değeri %80'dir ve kalan ögeler için de aynıdır.

Tablo 3.2: Örnek veri setindeki her bir münferit öge destek değeri

Öge Kodu	Destek
A	0.80
B	0.60
C	0.60
D	0.60

Bu örnek için minimum %50 destek değeri olduğunu kabul edelim, Tablo 3.2'deki tüm ögeler, öge seti oluşturma adaylarıdır. Dolayısıyla, her bir olası öge set desteği Tablo 3.1'den hesaplanır. Hesaplanan destek değerleri Tablo 3.3'te gösterilmiştir.

Tablo 3.3: Örnek veri setindeki öge setleri destek değeri.

Öge seti	Destek
AB	0.40
AC	0.40
AD	0.40
BC	0.60
BD	0.20
CD	0.20

Minimum izin verilen destek değerinden yüksek destek değerli tek öge seti BC öge setidir. Dolayısıyla, birliktelik kural madenciliğinin tüm adaylarını içeren bir tablo oluşturmak için bu öge, Tablo 3.2'deki münferit ögelere eklenir.

Tüm sık öge setleri çıkarıldıktan sonra, algoritma her olası öge çifti veya öge setleri arasındaki birliktelik kurallarını araştırmaya başlar. Her çıkarılan kural için güven değeri hesaplanır ve minimum kabul edilebilir gerekli güven değeri ile karşılaştırılır. Son olarak birliktelik kuralları, bu kuralın güçlülüğünü belirlemek için her bir kuralın destek ve güven değerleri ile birlikte alınır. Destek, bir birliktelik kuralının istatistiksel önemini gösterir, burada düşük destek seviyeli kurallar genellikle göz ardı edilir. Çünkü tesadüfen görüldükleri kabul edilir ve bu kurallarda

önemli olacak yeteri karlılıkta değildirler. Ayrıca, bir kuralın güven değeri, bulunan kuralın güvenilirliğini yansıtır. Yani güven değeri ne kadar yüksekse bu ögelerin aynı işlemlerde birlikte görülme ihtimali o kadar yüksektir.

Tablo 3.1'de gösterilen örnek veri setindeki ögeler arasında birden fazla pozitif birliktelik kuralı olabilmesine rağmen, tartışma için sadece bir birliktelik kuralı seçilir; $A \Rightarrow B$ ilişkisi. Bu kural güven değeri %50'dir. Çünkü A ögesini içeren dört işlemden ikisi B ögesini içerir. Dolayısıyla, A ögesine sahip sepetin B ögesine de sahip olma şansı %50'dir.

3.1.2 Negatif Birliktelik Kuralları

Negatif birliktelik kuralları, öncül öge varlığının, sonuç ögenin satın alımda olmaması ihtimalini arttırdığı öge çiftleri veya öge setleri arasındaki ilişkiyi belirtir. Dolayısıyla, X'den Y'ye negatif birliktelik kuralı $X \Rightarrow \neg Y$ olarak gösterilir. Bu kurallar, doğrudan veri setinden aranamaz. Çünkü bu tür ilişkiler ilgili olmayan ögeler arasında görülür ve doğrudan madencilik ilgisiz kurallar ile sonuçlanır. Dolayısıyla, bu problemin önlenmesi için pozitif birliktelik kuralları önce çıkarılır, sonra alan bilgisi kullanılarak pozitif birliktelik kurallarına sahip ögelerden gruplar oluşturulur. Sonrasında öncüllerin grupları ile sonuç ögelerden gruplar arasında negatif birliktelik kuralları araştırılır. Bu işlem, ilgili negatif birliktelik kurallarının bulunmasını temin eder. Önceki negatif birliktelik kural madencilik yöntemlerinde [19, 21, 22, 24, 25, 45-49], farklı ögeler arasındaki ilişkiyi temsil eden, marka, tip ve model gibi alan bilgisi. Bu yöntemlerle çıkarılan negatif birliktelik kurallar, ögeler arasında var olan pozitif birliktelik kurallarına ve modele sağlanan bilgiye dayanır.

Örneğin, Tablo 3.1'de gösterilen örnek veri setinden elde edilen pozitif birliktelik kuralı, $A \Rightarrow B$ bu ögeler arasında varolan ilişki olduğunu gösteriyor. Öte yandan, satın alınan A ögesi miktarı dörtten büyük olduğunda da negatif birliktelik kuralı olduğu dikkat çekebilir. Bu kuralın güven değeri %50'dir. Bu işlem, bulunan negatif birliktelik kurallarının değerli ve ilgili olmasını temin eder. Böyle bir örnekte, bu kurallara dikkat etmek kolaydır. Çünkü daha büyük veri setinde bunları bulmak zordur. Alan bilgisi olmaksızın ilgisiz negatif birliktelik kurallarını bulmanın riski, daha büyük veri setlerine arttıkça, bulunup çıkarılan kurallara daha fazla değer katmak için bu kuralları araştırmak ve bazı alan bilgilerini kullanmak da daha iyidir.

3.2 Veri kümeleme (Data Clustering)

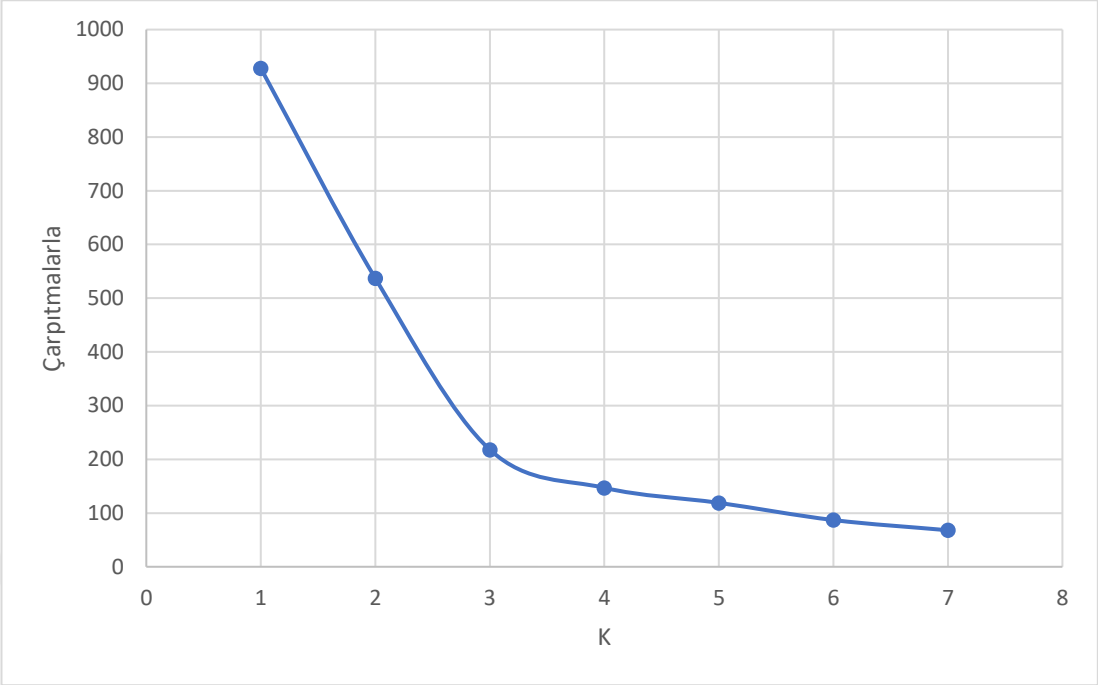
Veri nesnelerinin, her bir gruptaki nesnelerin, veri setinin diğer nesnelerine göre birbirlerine daha fazla benzer oldukları gruplara dağıtılması, veri kümeleme olarak adlandırılır. Kümeleme teknikleri, değerler arasındaki mesafe kolayca hesaplanabildiğinden sayısal özelliklerle kolayca ve daha doğru şekilde bulunabilen benzer nesneleri bulmak için her bir nesnenin öz değerlerine dayanır. Kümeleme teknikleri, gözetimsiz makine öğrenme teknikleri olup, bir tarafta veri nesneleri diğer tarafta veri etiketleri arasındaki ilişkinin araştırıldığı gözetimli makine öğrenmenin tersine, veri nesneleri arasındaki ilişkileri araştırır. Dolayısıyla, kümeleme teknikleri herhangi bir veri etiketlemesi gerektirmez. En yaygın veri kümeleme yöntemlerinden ikisi K-Ortalama ve DBSCAN yöntemleridir [50, 51].

K-Ortalama (K-means clustering method), adından da anlaşılacağı gibi her bir küme için rastgele K oluşturarak, veri setini K kümelerine kümeler. Sonra her bir nesne, kendisine en yakın sentroide dahil edilir. Nesneleri bu kümelere dağıttıktan sonra, bu kümedeki nesneler arasındaki mesafe ortalaması alınarak her bir sentroid için yeni konum hesaplanır. Sonra nesneler, sentroidlerin yeni konumuna göre yeni kümelere yeniden dağıtılır. Tüm sentroidler optimal konumlarına yaklaşıncaya kadar bu işlem tekrarlanır ve daha fazla tekrar, küme sentroidlerinin daha fazla yer değişimi getirmez. K-Ortalama kümeleme yöntemi algoritması Algoritma 3.2'de gösterilmiştir. Bu yöntem, önceden tanımlı kümeler sayısı gerektirdiğinden, dirsek yöntemi ile küme optimal sayısını hesaplamak için bu yöntemle beraber Dirsek yöntemi yaygın olarak kullanılmaktadır. Daha iyi kümeleme sonuçları elde etmek için veri setindeki nesnenin dağıtımına bağlı olarak, veri setini farklı küme sayısına kümeleme kabiliyetine sahip K-Ortalama yöntemine geçilir [5, 32, 52].

Algoritma 3.2: K-Ortalama kümeleme algoritması.

```
1:   Girdi Veri seti (D)
     Girdi gereken küme sayısı (K)
2:   Başlangıç rastgele küme sentroidleri C = C1 … Ck
3:   Önceki sentroid yerleri farklı iken Cp ≠ C
4:       Veri setindeki her bir nesne için D
5:           Küme C'ye küme sentroidi ata
6:       Her biri için son
7:       Her bir küme sentroidi için C'de c
8:           c = kümedeki nesne ortalaması
9:       Her biri için son
10:  İken son
23:  C'ye dön
```

Dirsek yönteminde (Elbow method), her bir tekrara bir küme ekleyerek veri setini kümelemek için K-Ortalama kümeleme tekniği kullanılır. Sonrasında, bir kümedeki nesneler ve bu kümenin sentroidi arasındaki ortalama mesafe toplamı hesaplanır. Bu hesaplanan değer, küme sayısı arttıkça düşer. Spesifik küme sayısında, daha fazla küme eklenmesi, hesaplanan mesafe toplamını düşürür. Ancak bu azalma değişim oranı, son kümeleme değişim oranından oldukça küçüktür. Değişim oranındaki bu fark, yönteme de adını veren, dirseğe benzer bir nokta oluşturur. Dirsek noktasındaki küme sayısı, küme optimal sayısı olarak kümeleme yöntemine geçer. Çünkü küme sayısının daha da arttırılması veri setinden alınan bilgiye çok şey katmaz. Bozulma olarak da bilinen ortalama mesafenin hesaplanan toplamına ait örnek grafik Şekil 3.1'de verilmiştir.



Şekil 3.1: Dirsek yöntemi bozulmalar sonucu örnek grafiği.

Diğer bir yaygın veri kümeleme tekniği, DBSCAN (Density-Based Spatial Clustering of Applications with Noise) yöntemidir. Bu yöntem, iki önceden tanımlı konfigürasyon değerine göre her bir veri seti için küme sayısını otomatik olarak oluşturma kabiliyetine sahiptir. Bu değerler, aynı kümedeki nesneler arasında izin verilebilir maksimum mesafeyi ve bir kümenin içerebileceği minimum nesne sayısını gösterir. Bu iki değer kullanılarak, kümeleme veri setinden rastgele bir nesne seçimi ile başlatılır ve buna göre yeni bir küme oluşturulur. Sonra, nesne ile mesafesi maksimum izin verilen mesafenin altında olan bitişik nesneler aynı kümeye dahil edilir. Bu işlem, veri setinde bir kümedeki bir nesneye daha yakın olan tüm nesneler algoritma için konfigüre edilinceye kadar tekrarlanır. Ayrıca, başka kümelenmiş nesne seçilir ve aynı işlem tüm nesneler kümelendirilinceye kadar tekrarlanır. DBSCAN kümeleme yönteminin algoritması Algoritma'da gösterilmiştir 3.3.

Algoritma 3.3: DBSCAN kümeleme yönteminin algoritması

```
Input Dataset (D)
Input parameters (esp, min_samples)
While unclustered object exists
        Select random unclustered object
        clusterCount=0
        While object (o) with distance <esp exists
                Include object (o) in cluster
                clusterCount+=1
        If clusterCount ≥ min_samples
                Create new cluster
        Else
                Objects are noise
Return (Clusters)
```

DÖRDÜNCÜ BÖLÜM

ÖNERİLEN YÖNTEM

Bu çalışmada önerilen negatif birliktelik kuralı madencilik yöntemi, makine öğrenme tekniklerinden yararlanarak veri setinden negatif birliktelik kural madenciliği için gerekli alan bilgisini bulup alır. Bu yöntem, satın alınan ögelerin satın alınan miktarına dayandığından, satın alınan miktarların histogramı sonrasında, negatif birliktelik kurallarını bulup çıkarmak için bu bilgi bu histogramlardan alınır. Bu bilgiyi bulup çıkarmak için, veri kümeleme yöntemleri kullanılır. Çünkü işlem veri setleri ve aynı zamanda miktarlar histogramı etiketli veri değildir. Bu kümeler, sonrasında, pozitif birliktelik kurallarını paylaşan öge grupları arasındaki negatif birliktelik kurallarını bulmak için kullanılır.

4.1 Kümeleme Miktarları

Veri setinden negatif birliktelik kuralları çıkarımında kullanılabilecek faydalı bilgi elde etmek ve her bir ögenin miktarları için histogramlar oluşturulur. Miktar değerlerinin x-eksenine ve bu miktarda satın alınan ögenin yer aldığı fatura sayısının da y eksenine dağıtılması ile bir histogram oluşturulur. Sonrasında, bu miktarların histogramda dağıtımı gruplar içinde toplanır. Kümeleme, bir veri seti nesnesinin gruplar içine dağıtılması işlemidir. Bu işlemde gruptaki bir nesne, diğer gruplardaki diğer bir nesneye göre aynı gruptaki diğer nesnelere daha çok benzerdir. Kümeleme, gözetimsiz makine öğrenme tekniğidir yani, etiketlenmemiş verilerden bilgi bulabilme kabiliyetine sahiptir. Bu kümeler, negatif birliktelik kuralları tespitinde kullanılır. Bu çalışmada, bu amaçla daha uygun kümeleme yöntemini bulmak için iki kümeleme yöntemi kullanılmıştır. Bu yöntemler, K-Ortalama ve Uygulamaların Yoğunluk esaslı Uzaysal Kümelenmesi (DBSCAN) kümeleme yöntemleridir.

K-Ortalama kümeleme yöntemi, veri seti nesnelerinin dağıtıldığı grupların sayısına karşılık gelen önceden belirlenen küme sayısı gerektirir. Bu yöntem, veri setini temsil eden çok boyutlu uzayda K rastgele noktalı kümeleme yöntemini başlatır. Daha sonra her bir sentroidden ve bu kümedeki nesneden mesafelerin toplamını minimize etmek için bu noktaların konumları ayarlanır. Bu mesafeleri hesaplamak için kullanılan eşitlik:

$$W(X, C) = \sum_{k=1}^{K} \sum_{j \in X_k} \left\| Y_j - C_k \right\|^2 \qquad \text{denklem(4.1)}$$

Burada X, veri setinin çok boyutlu uzamında Y vektörü ve sentroid C'nin K sayısı kullanılarak belirlenen nesneler grubudur.

Önerilen yöntem, gözetimsiz makine öğrenme tekniği olduğundan, tüm değişkenlerin algoritma ile hesaplanması önemlidir. Dolayısıyla, kümelerin optimal sayısını hesaplayabilen bir tekniğin kullanılması önemlidir. Dirsek yöntemi, istenen kabiliyete sahip yaygın olarak kullanılan yöntemlerden biridir. Bu yöntem, kümelerin önceden belirlenen maksimum sayısına kadar bir kümenin, kümeleme sonuçları için veri tabanının her bir nesnesi ve sentoridin ait olduğu kümenin sentroidi arasındaki mesafenin toplamını hesaplar. 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 bundan sonra dirsek yöntemi olarak adlandırılır. 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 yöntemi olup, bu yöntem çok boyutlu uzaydaki bir veri setini, 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, sonrasında, 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 tekrarlanır.

4.2 Bilgi Madenciliği

Kümeleme tekniği kullanılarak çıkarılan bilgi, her bir öge başına satın alınan miktarların örüntüsüne bağlı olarak, işlem veri setindeki ögeler için bir taksonomi oluşturmak için kullanılır. Örneğin, Tablo 4.1'de gösterildiği gibi işlem başına her bir ögenin satın alınan miktarında iki öge örnek işlemi olduğunu düşünün.

Tablo 4.1: İki öge için işlem örnek veri seti.

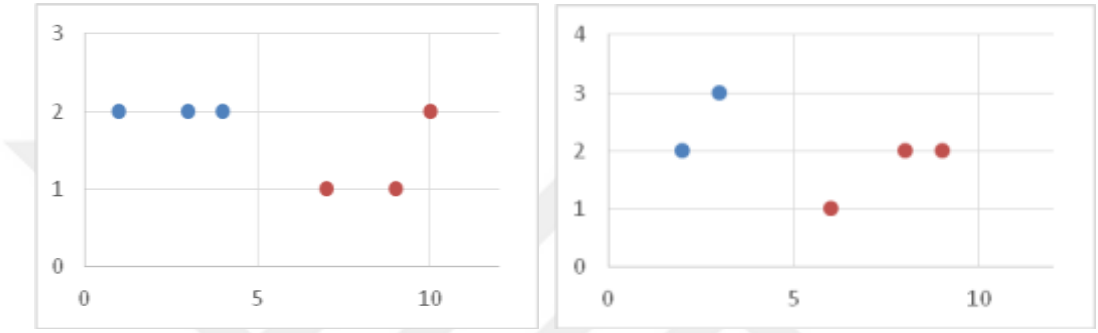
İşlem Kimliği	Ürün Kimliği	Miktar
1	A	1
1	B	8
2	A	7
2	B	2
3	A	9
3	B	3
4	A	4
4	B	9
5	A	3
5	B	3
6	A	10
6	B	3
7	A	10
7	B	9
8	A	1
8	B	8
9	A	4
9	B	2
10	A	3
10	B	6

Satın alınan miktarların frekansları, bu miktarları frekanslarına göre homojen gruplara kümelemek amacıyla her bir öge için, Tablo 4.2'de belirtildiği gibi

hesaplanır. Tablo 4.2'deki verilere göre oluşturulan histogram ve kümelerde nesnelerin dağıtımı Şekil 4.1'de gösterilmiştir.

Tablo 4.2: Her bir öge için satın alınan miktarların frekansı.

	Miktar frekansı									
Öge	1	2	3	4	5	6	7	8	9	10
A	2	0	2	2	0	0	1	0	1	2
B	0	2	3	0	0	1	0	2	2	0



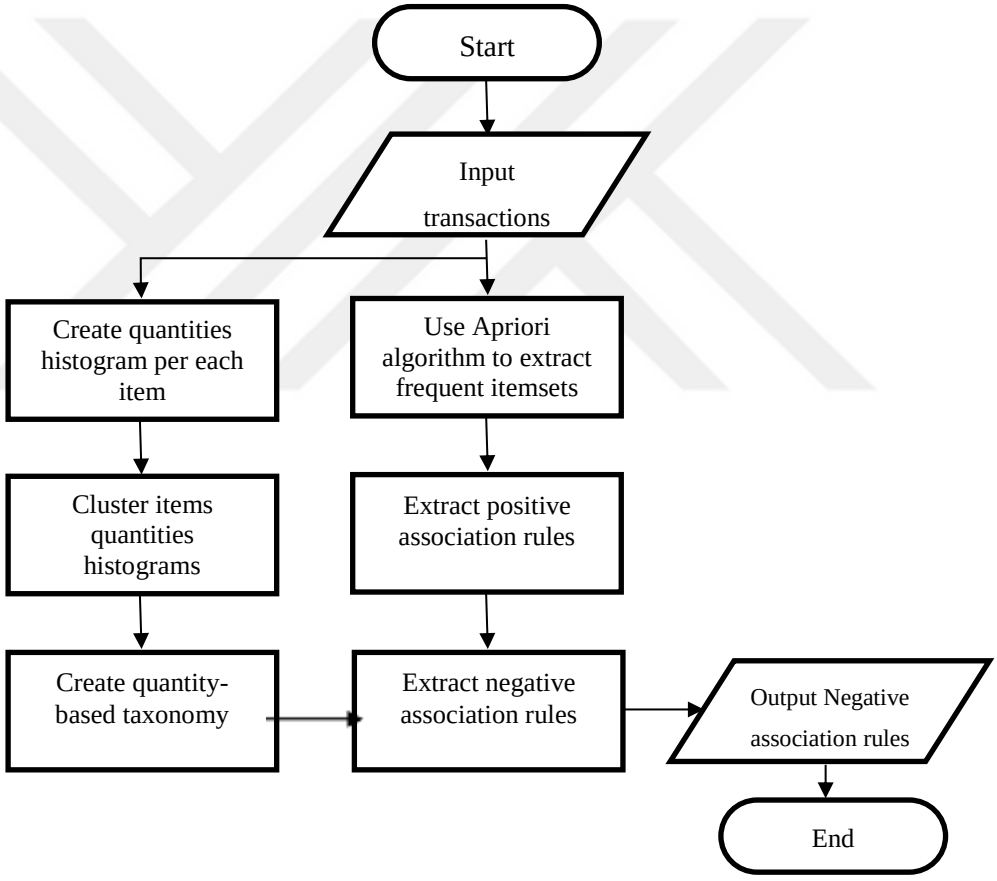
Şekil 4.1: Satın alınan miktarların kümeli histogramları Sol: Öge A Sağ: Öge B.

Kümelendirilen miktarlar daha sonra, 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 Ü1 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 4.2'de gösterilmiştir.

Şekil 4.2: Satın alınan miktarlara dayalı öge taksonomisi.

4.3 Miktar esaslı Negatif Birliktelik Kuralları Madenciliği

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 setinde 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 4.3'te gösterilen akış şemasında gösterildiği gibi öncüldeki ögelerin her altkümesi ile sonuç ögesindeki ögeler alt kümeleri arasında araştırılır.



Şekil 4.3: Önerilen negatif birliktelik kuraları madencilik yöntemi akış şeması.

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 4.1'de gösterilmiştir.

Algoritma 4.1: 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: İçin sonlandır
9: İçin sonlandır
10: $P = \{X \Rightarrow Y | C_{xy} \geq minConf\}$.
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: İçin sonlandır
20: İçin sonlandır
21: Her biri için son
22: N = $\{M \Rightarrow \neg N | NC_{mn} \geq minNConf\}$.
23: N Dön

BEŞİNCİ BÖLÜM

DENEYSEL SONUÇLAR

5.1 Deneyler

Önerilen negatif birliktelik kuralı madencilik yöntemleri performansını test etmek için, California Üniversitesi, Irvine (UCI) bilgi havuzundan gerçek hayat çevrimiçi depo işlemleri veri seti değerleri kullanılarak üç deney yapılmıştır [53]. 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 5.1'de verilmiştir.

Tablo 5.1: 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

Ögeleri ve satın alınan miktarları içeren işlemlere göre ögeler için birliktelik kuralları araştırıldığından, birliktelik kural madenciliği için sadece İşlem Kimliği, Stok Kodu ve Miktar kullanılmıştır. Tüm deneyler, an Intel® Core™ i7-7700HQ @ 2.8 GHz işlemci ve a 16 GB belleğe sahip bir bilgisayarda Windows 10 işletim sisteminde Python programlama dili [54] kullanılarak yapılmıştır. *Apriori* algoritma uygulaması için python kütüphane apyori 1.0.0 kütüphanesi kullanıldı. Deneyler, satın alınan farklı gruplar kullanılarak yapıldı, ilk deneyde her bir satın alınan miktar tek başına kullanılırken ikinci ve üçüncü deneylerde farklı kümeleme teknikleri kullanılmıştır.

5.1.1 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 5.2'de verilmiştir.

Tablo 5.2: Deney A negatif birliktelik kuralı madenciliği sonuçların bir özeti.

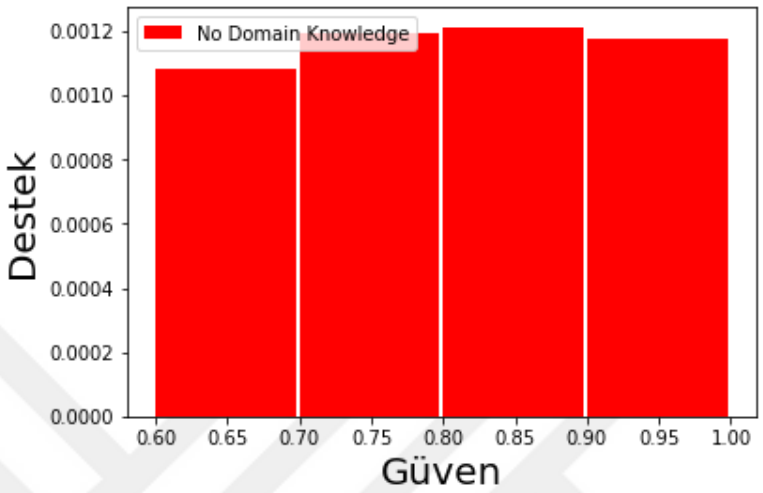
Kural sayısı	Yapım. zamanı	Destek			Güven		
		Minimum	Maksimum	Ortalama	Minimum	Maksimum	Ortalama
9066	1932.74	0.10%	0.13%	0.12%	66.67%	100%	99.37%

Bu deneyde çıkartılan negatif ilişki kurallarının bir örneği Tablo5.3te' gösterilmiştir.

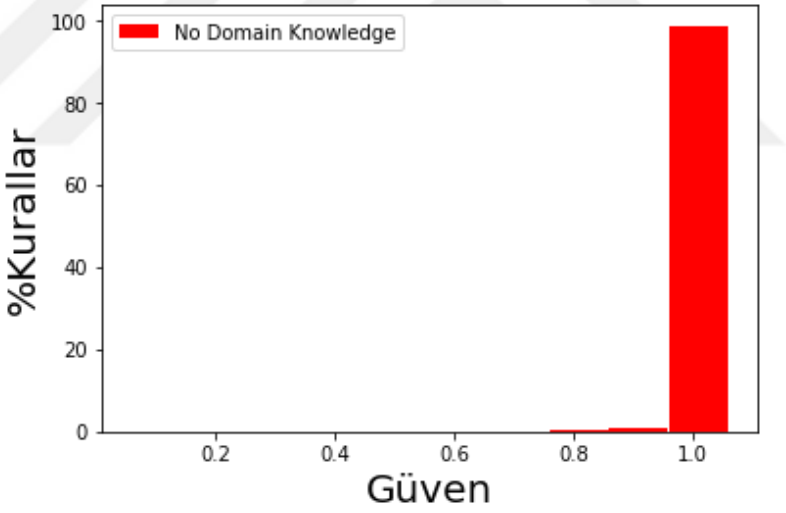
Tablo 5.3: Deney A'dan örnek negatif ilişki kuralları.

Antecedent Item	Consequent Item	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 5.1'de verilmiş olup, bulunan negatif birliktelik kuralları yüzdesi – bu kuralların güven değerleri Şekil 5.2'de gösterilmiştir.



Şekil 5.1: Deney A ortalama destek – güven değerleri.

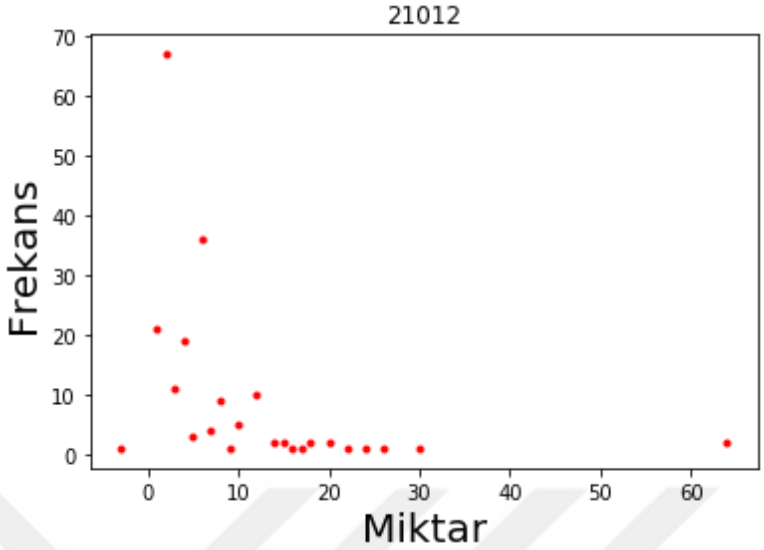


Şekil 5.2: Deney A kurallar yüzdesi – güven.

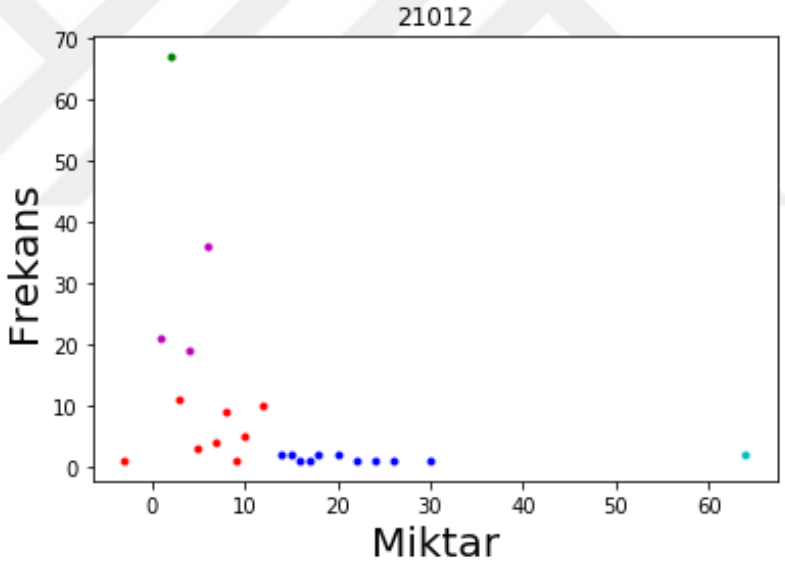
5.1.2 Deney B

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 5.3'te verilmiştir. Bu öge

için dirsek yöntemi ile belirlenen kümelerin optimal sayısı beş kümedir. K ortalama yöntemleri kullanılarak bu miktarlar Şekil 5.4'te gösterildiği gibi kümelendirildi.



Şekil 5.3 : Örnek satın alınan miktarlar histogramı.



Şekil 5.4 : Örnek kümelendirilen miktarlar histogramı.

Deney B'da elde edilen aynı pozitif birliktelik kurallarına göre bulunan negatif birliktelik kuralları özeti Tablo 5.4'te verilmiştir.

Tablo 5.4: Deney B negatif birliktelik kuralı madenciliği sonuçlarının bir özeti.

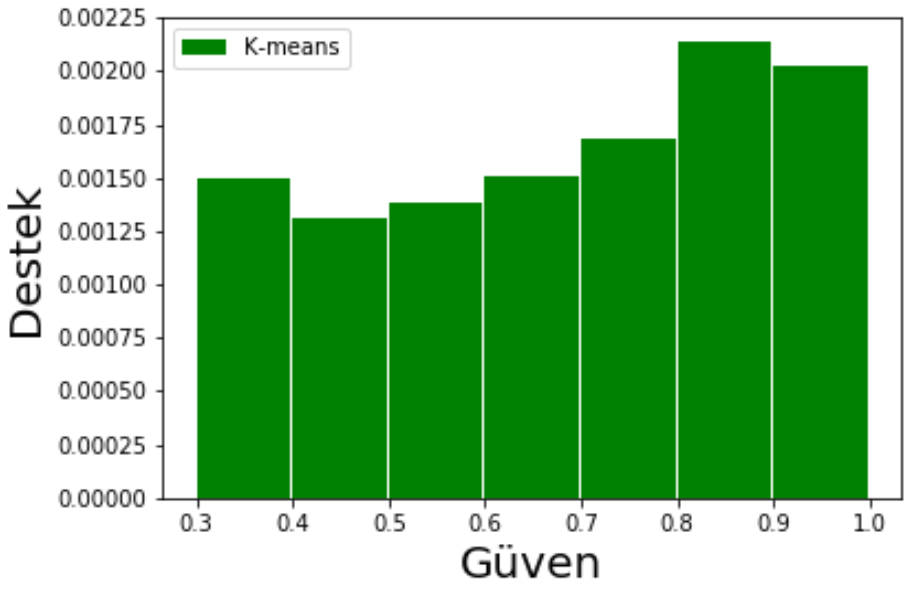
Kural sayısı	Yapım Süresi	Destek			Güven		
		Minimum	Maksimum	Ortalama	Minimum	Maksimum	Ortalama
1780	353.57	0.10%	0.44%	0.19%	31.69%	100%	85.84%

K-aracı yöntemi kullanılarak çıkarılan etki alanı bilgisine dayalı olarak çıkarılan negatif ilişki kurallarının bir örneği Tablo'da gösterilmiştir 5.5.

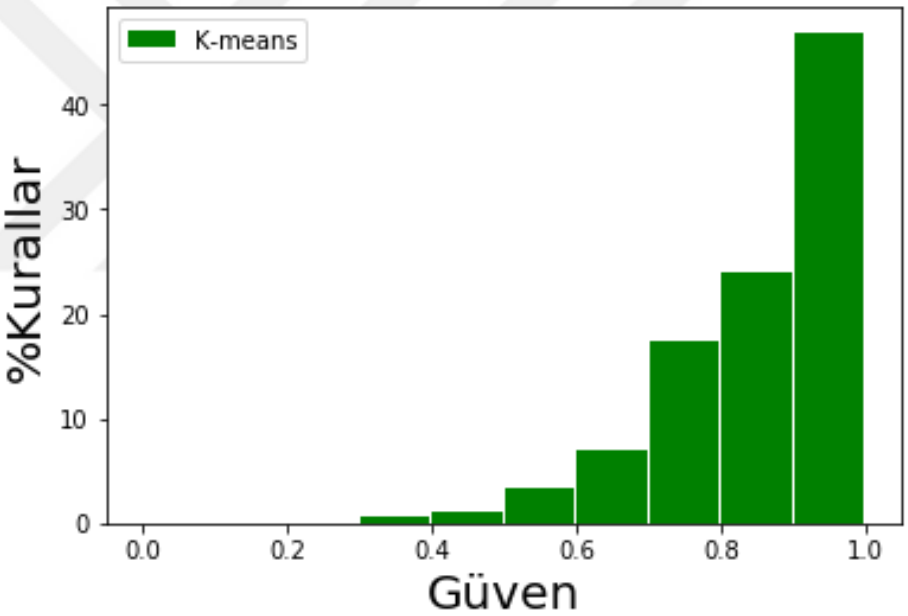
Tablo 6.5: Deney B'de çıkarılan negatif ilişki kurallarının örneği.

Antecedent Item	Consequent Item	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 5.5'de gösterilmiş olup bulunan kural yüzdesi ve bu kuralların güven değerleri Şekil 5.6'da verilmiştir.



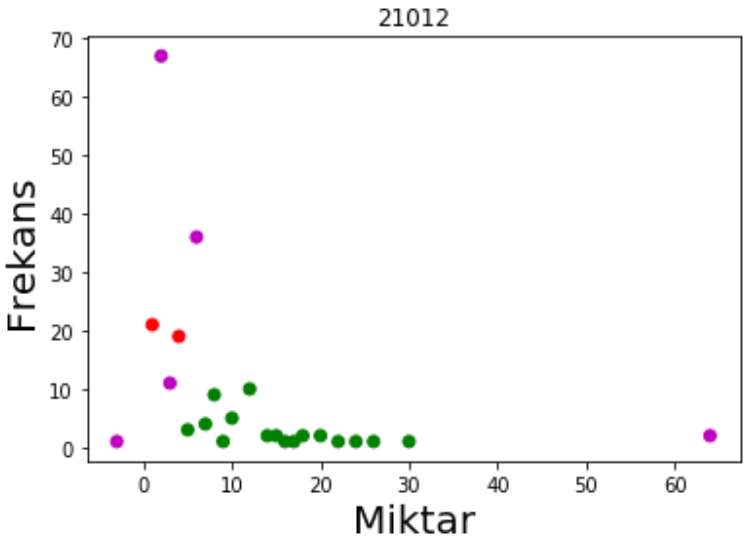
Şekil 5.5: Deney B güven – destek ortalaması.



Şekil 5.6: Deney B için güven aralığı başına bulunan kural yüzdesi.

5.1.3 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 5.2'de gösterilen aynı örnek öge için, DBSCAN bu ögeyi Şekil 5.7'te gösterildiği gibi üç kümeye gruplandırdı.



Şekil 5.7: DBSCAN kullanılarak örnek öge miktarları histogramının kümeleme sonuçları.

DBSCAN yöntemi kullanılarak oluşturulan kümelere göre bulunan negatif birliktelik kurallarının özeti Tablo 5.6'te verilmiştir.

Table 5.6: Deney C negatif birliktelik kuralı madenciliği sonuçlarının bir özeti.

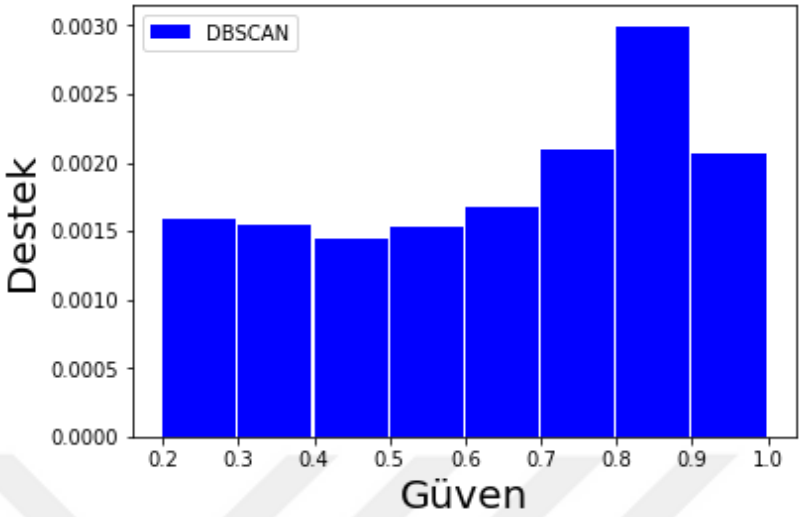
Kural sayısı	Yapım Süresi	Destek			Güven		
		Minimum	Maksimum	Ortalama	Minimum	Maksimum	Ortalama
4086	783.30	0.10%	0.40%	0.21%	22.54%	100%	91.84%

Bu deneyde çıkarılan örnek negatif ilişki kuralları Tablo'da gösterilmektedir 5.7.

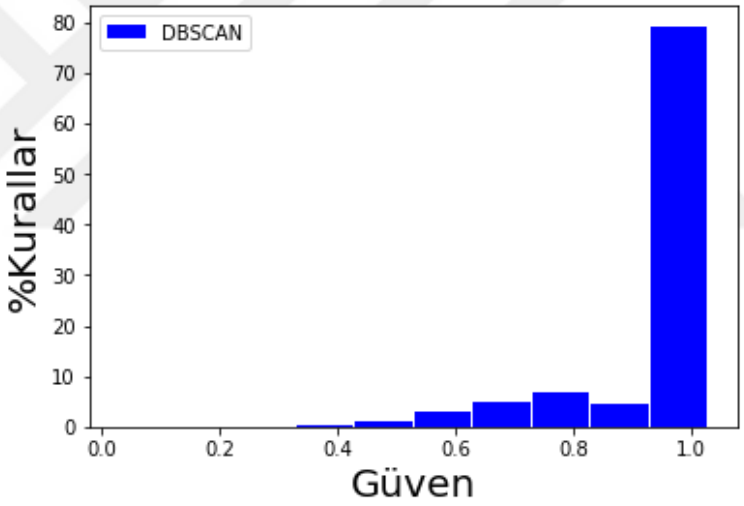
Tablo 5.7: Deney C'nin örnek negatif ilişki kuralları.

Antecedent Item	**Consequent Item**	**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 5.8'de ve her bir güven aralığı için bulunan kurallar yüzdesi ise Şekil 5.9'da gösterilmiştir.



Şekil 5.8: Deney C için ortalama destek – güven değerleri.



Şekil 5.9: Deney C için bulunan kuralların güven karşısında dağılımı.

Ayrıca, önerilen yöntemlerin karmaşıklığını karşılaştırmak amacıyla, Yapım Süresi, 10000, 500000 ve 2000000 kayıttan oluşan üç farklı veri seti kullanılarak her bir algoritma için bir ölçümdür. Ölçüm Yapım Süresi Tablo 5.5'te gösterilmiştir.

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

	Süre (saniye)		
Kayıt sayısı	Kümeleme yok	K-Ortalama	DBSCAN
10000	1.27	0.41	0.33
500000	1232.92	1429.90	395.41
2000000	15326.43	10410.22	7721.35
Ortalama	5520.21	3946.84	2705.70

5.2 değerlendirme

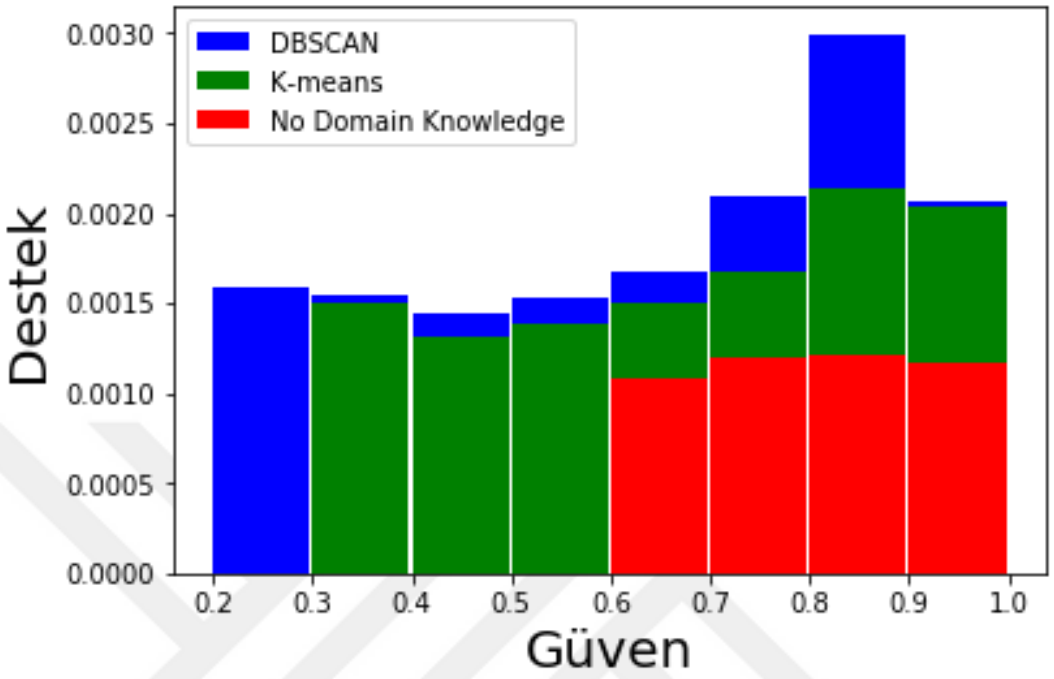
Deney A sonuçları, herhangi bir alan bilgisi olmaksızın negatif birliktelik kuralları madenciliğini göstermektedir. Bu negatif birliktelik kural madenciliğinin, ilgisiz kurallar oluşturması beklenir. Tablo 5.9'da gösterildiği gibi bu deney sonuçlarını bu araştırmada yapılan diğer deney ile karşılaştırılmış 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 [19, 21, 22, 24, 25, 45-49] 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 5.10'de çok iyi gösterilmiştir. Alan bilgisi olmadan negatif birliktelik

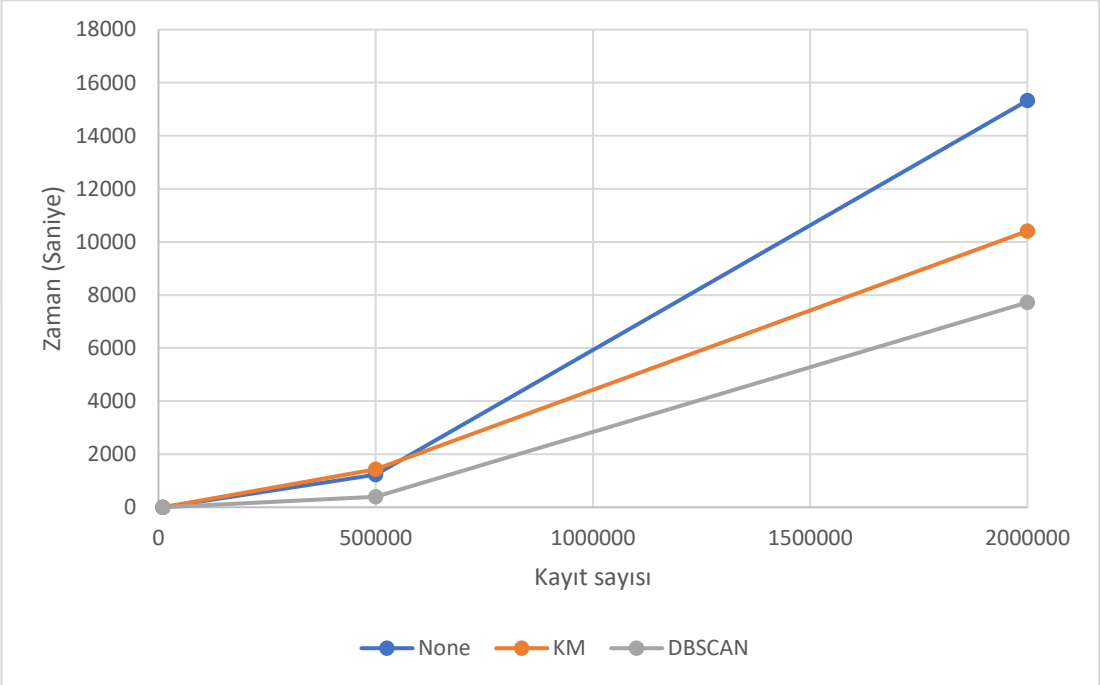
kurallarının güven değeri yüksek seviyelerde yoğunlaşmasına rağmen, bu kuralın destek değeri, alan bilgisi keşfine göre bulunan değerlerden daha küçüktür.



Şekil 5.10: Tüm yapılan deneyler için ortalama destek – güven değerleri.

K ortalama kümeleme yöntemi, çok yaygın olmasına ve bir çok uygulamada kullanılmasına rağmen [5, 30-32, 55, 56], 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 5.4'te ölçülen yapım süresi Şekil 5.11'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.



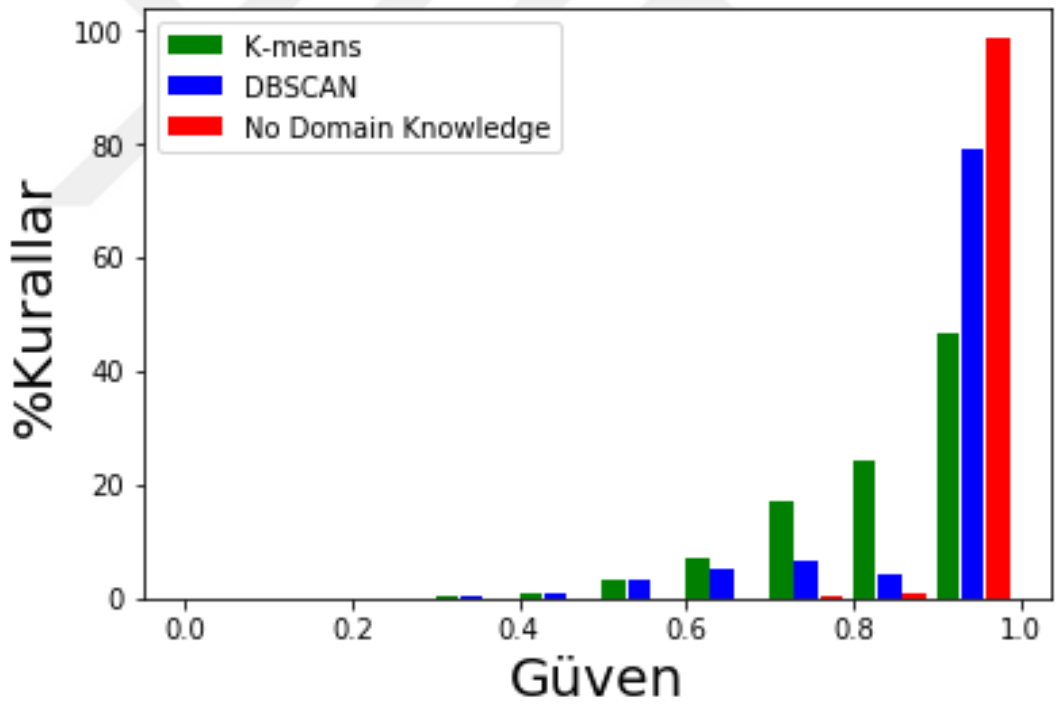
Şekil 5.11: Farklı büyüklükte veri setleri kullanan her bir algoritmanın harcadığı yapım süresinin grafiksel gösterimi.

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 5.10'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 5.12'de gösterilmiştir. En yüksek seviyede bulunan kuralların en yüksek kısmının Deney A'dan çıkarılmış olmasına rağmen, Şekil 5.1 ve Şekil 5.10'da

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 5.12: Tüm deneyler için güven seviyelerinde negatif birliktelik kuralının dağılımı.

ALTINCI BÖLÜM

SONUÇ

Makine öğrenmesi bilgisayarların, spesifik şartlarla etkileşmeye gerek olmaksızın etraflarındaki çevreyi analiz edip etkileşim içine girebilme kabiliyetidir. Bu algoritma, bilgisayarlara daha önce hiç olmadığı şartlarla etkileşmesine imkân tanır. Veri madenciliği, ilgi alanı veri setleri için faydalı bilgileri bulup çıkarmak olan makine öğrenme alanlarından birisidir. Bu bilgi, özellikle büyük veri setlerinde, bazen insanların dikkatini çekmeyebilir. Veri madenciliği, bulunan alan bilgileri olarak sunmak üzere, veri özellikleri arasındaki gizli ilişkileri ve sık örüntüleri keşfetmeye çalışır.

Veri madenciliği, diğer herhangi bir makine öğrenme gibi, gözetimli ve gözetimsiz olmak üzere iki kategoriye ayrılır. Gözetimli makine öğrenmede, veri, gözetimli veri madencilik teknikleri uygulanmadan önce etiketlenmelidir. Bu durumda, veri madenciliği teknikleri, veri nesnelerini birbirinden ayıran özelliklerin değerleri arasındaki ilişkileri bulmaya ve her bir nesnenin bir diğeri ile etiketlemeye çalışır. Sınıflandırma, nesneler için etiketlerin kestiriminde bulunarak yeni nesnelerin gelecekteki davranışlarını kestirmek için veri setinden bilgi bulup çıkarmak için kullanılan gözetimli veri madenciliği tekniklerinden biridir. Öte yandan, kümeleme gibi gözetimsiz veri madenciliği teknikleri, veri nesnesinin öz değerlerindeki sık örüntüyü bulmaya çalışır. Böylece bu teknikler, bir gruptaki bir nesnenin diğer gruplardaki nesnelere göre bu gruptaki nesnelere daha fazla benzer olduğu daha homojen gruplara dağıtır. Gözetimsiz veri madencilik teknikleri, veri etiketleme gerektirmez ve alan bilgisini ham veri setinden bulup çıkarma kabiliyetindedir.

Birliktelik kuralları, aynı veri setindeki bir nesnenin diğer nesneye göre davranışını tanımlayan kurallardır. Bir veri setinde bir nesnenin görülmesinin diğer bir nesnenin görülmesi üzerindeki etkisini bulmak için, birçok birliktelik kuralı madencilik tekniği önerilmektedir. Bir ögenin belli bir pozisyonda varlığı, başka bir

nesnenin bulunma olasılığını arttırıyorsa, birinci nesnenin diğerine bir pozitif birliktelik kuralı vardır. Eğer bu ögenin varlığı, ikinci ögenin belli durumda bulunma ihtimalini azaltıyorsa, negatif birliktelik kuralından söz edilir.

Negatif birliktelik kuralı madenciliğinde, ilgisiz nesneleri birleştiren kurallar olan ilgisiz kuralların bulunup çıkarılması zorluğu vardır. Bu davranışı önlemek ve negatif birliktelik kuralları genellikle ilgili ögeler için; ilk olarak, pozitif birliktelik kurallarını arayıp çıkararak, sonra da bunları birleştiren pozitif birliktelik kuralına sahip olduğu tespit edilen nesnelerin alt kategorileri arasında negatif birliktelik kuralları arayıp bulmak için işlenir. Bu işlem, güçlü ilgili negatif birliktelik kurallarının bulunmasını temin eder. Nesnelerin alt kategorilerini belirlemek için, alt kategoriler ile ana nesne arasındaki ilişkileri niteleyen ana bilginin kullanılması önemlidir. Dolayısıyla, mevcut yöntemlerin çoğu, negatif birliktelik kuralı madencilik teknikleri el ile sağlanan veri bilgisine dayanır.

Bu araştırmada, gözetimsiz veri madenciliği tekniği, kümeleme, kullanılarak veri setinden otomatik olarak bulunup alınan alan bilgisine dayanan yeni bir negatif birliktelik kuralı madencilik tekniği önerilmektedir. Bu çalışmada, K-Ortalama ve DBSCAN kümeleme yöntemleri olmak üzere iki kümeleme yöntemi test edildi. Bu kuralları sınamak ve kümeleme yöntemleri ile bulunup çıkarılan alan bilgisine dayanarak çıkarılanlarla karşılaştırmak için alan bilgisi olmaksızın negatif birliktelik kuraları madenciliği de yapıldı. Bu yöntemleri sınamak için UCI deposundan elde edilen gerçek yaşam işlemler veri seti kullanıldı. Negatif birliktelik kuralı madenciliği, satın alınan miktarlara dayanmaktadır. Bu araştırmada üç senaryo denendi: ilk senaryoda, sadece alan bilgisi olmaksızın madencilik negatif birliktelik kurallarını temsil eden her satın alınan miktar kullanılmış olup, diğer iki deneyde ise K ortalama ve DBSCAN kümeleme yöntemleri kullanılarak, bulunup çıkarılan alan bilgisi kullanıldı.

Satın alınan miktarlar histogramları oluşturuldu. Sonrasında bu histogramlara göre, satın alınan miktarlar, farklı sayılarda kümelere kümelendirildi. K-Ortalama kümeleme yöntemi, belli bir veri seti için kümelerin optimal sayısını hesaplayabilme kabiliyetine sahip olmadığından, bu sayıyı bulmak için dirsek yöntemi kullanıldı. Ayrıca, DBSCAN yöntemi, herhangi bir veri setini otomatik olarak, önceden tanımlı konfigürasyona göre optimal sayıda kümeye kümeleme kabiliyetine sahiptir. Bu kümeler, daha sonra negatif birliktelik kurallarının araştırılması için kullanılan, her bir ögeye ait alt gruplar oluşturmak için kullanıldı.

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 olmuştur. Bu durumda alan bilgisi kullanılmadığında ilgisiz kuralların bulunup çıkarılmasını göstermektedir. Ayrıca, 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. Ancak 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 %0.19 ve %0.21'dir. Öte yandan, ortalama güven değeri, alan bilgisiz madencilikte %99.37 iken sonuç olarak, K-Ortalama ve DBSCAN küme yöntemleri ile bulunup çıkarılan alan bilgisi kullanıldığında bu değer %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 tavsiye edilir.

KAYNAKLAR

[1] A. L. Buczak and 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 and 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 and R. Karpienko, "Visualizing association rules in hierarchical groups," *Journal of Business Economics*, vol. 87, pp. 317-335, 2017.

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

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

[6] T. S. Krishna, A. Y. Babu, and R. K. Kumar, "Determination of Optimal Clusters for a Non-hierarchical Clustering Paradigm K-Means Algorithm," in *Proceedings of International Conference on Computational Intelligence and Data Engineering*, 2017, pp. 301-316.

[7] A. Rodriguez and A. Laio, "Clustering by fast search and find of density peaks," *Science*, vol. 344, pp. 1492-1496, 2014.

[8] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical machine learning tools and techniques*: Morgan Kaufmann, 2016.

[9] C. C. Aggarwal, *Data classification: algorithms and applications*: CRC Press, 2014.

[10] G. Shmueli, P. C. Bruce, I. Yahav, N. R. Patel, and K. C. Lichtendahl Jr, *Data Mining for Business Analytics: Concepts, Techniques, and Applications in R*: John Wiley & Sons, 2017.

[11] D. T. Larose, *Discovering knowledge in data: an introduction to data mining*: John Wiley & Sons, 2014.

[12] R. J. Roiger, *Data mining: a tutorial-based primer*: CRC Press, 2017.

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

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

[15] K. Lai and N. Cerpa, "Support vs. confidence in association rule algorithms," in *Proceedings of the OPTIMA Conference, Curicó*, 2001.

[16] B. Minaei-Bidgoli, R. Barmaki, and M. Nasiri, "Mining numerical association rules via multi-objective genetic algorithms," *Information Sciences,* vol. 233, pp. 15-24, 2013.

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

[18] A. S. Sadh and N. Shukla, "Association rules optimization: A survey," *International Journal of Advanced Computer Research (IJACR),* vol. 3, 2013.

[19] S. Mahmood, M. Shahbaz, and A. Guergachi, "Negative and positive association rules mining from text using frequent and infrequent itemsets," *The Scientific World Journal,* vol. 2014, 2014.

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

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

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

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

[24] C. Cornelis, P. Yan, X. Zhang, and G. Chen, "Mining positive and negative association rules from large databases," in *Cybernetics and Intelligent Systems, 2006 IEEE Conference on*, 2006, pp. 1-6.

[25] J. Tsiligaridis, "Mining Positive and Negative Association Rules," in *The International Conference on E-Technologies and Business on the Web (EBW2013)*, 2013, pp. 110-114.

[26] L. Aliahmadipour, V. Torra, and 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.

[27] S. Guha and N. Mishra, "Clustering data streams," in *Data Stream Management*, ed: Springer, 2016, pp. 169-187.

[28] A. Fahad, N. Alshatri, Z. Tari, A. Alamri, I. Khalil, A. Y. Zomaya, *et al.*, "A survey of clustering algorithms for big data: Taxonomy and empirical analysis," *IEEE transactions on emerging topics in computing*, vol. 2, pp. 267-279, 2014.

[29] A. S. Shirkhorshidi, S. Aghabozorgi, T. Y. Wah, and T. Herawan, "Big data clustering: a review," in *International Conference on Computational Science and Its Applications*, 2014, pp. 707-720.

[30] M. B. Cohen, S. Elder, C. Musco, C. Musco, and 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.

[31] M. E. Celebi, H. A. Kingravi, and P. A. Vela, "A comparative study of efficient initialization methods for the k-means clustering algorithm," *Expert systems with applications*, vol. 40, pp. 200-210, 2013.

[32] P. Bholowalia and A. Kumar, "EBK-means: A clustering technique based on elbow method and k-means in WSN," *International Journal of Computer Applications*, vol. 105, 2014.

[33] Ester, Martin, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. "A density-based algorithm for discovering clusters in large spatial databases with noise." In *Kdd*, vol. 96, no. 34, pp. 226-231. 1996.

[34] K. M. Kumar and A. R. M. Reddy, "A fast DBSCAN clustering algorithm by accelerating neighbor searching using Groups method," *Pattern Recognition,* vol. 58, pp. 39-48, 2016.

[35] T. N. Tran, K. Drab, and M. Daszykowski, "Revised DBSCAN algorithm to cluster data with dense adjacent clusters," *Chemometrics and Intelligent Laboratory Systems,* vol. 120, pp. 92-96, 2013.

[36] X. Meng, H. Li, and J. Cui, "Different strategies for differentially private histogram publication," *Journal of Communications and Information Networks,* vol. 2, pp. 68-77, 2017.

[37] E.-E. M. Azhari, M. M. M. Hatta, Z. Z. Htike, and S. L. Win, "Brain tumor detection and localization in magnetic resonance imaging," *International Journal of Information Technology Convergence and services,* vol. 4, p. 1, 2014.

[38] M. H. Mozaffari and S. H. Zahiri, "Unsupervised Data and Histogram Clustering Using Inclined Planes System Optimization Algorithm," *Image Analysis & Stereology,* vol. 33, pp. 65-74, 2014.

[39] B. Li, Y. Chen, J. Ren, and L. Cheng, "A Fast Video Stabilization Method Based on Feature Matching and Histogram Clustering," in *Information Technology and Intelligent Transportation Systems*, ed: Springer, 2017, pp. 315-325.

[40] R. Agrawal, T. Imieliński, and A. Swami, "Mining association rules between sets of items in large databases," in *Acm sigmod record*, 1993, pp. 207-216.

[41] J. Dongre, G. L. Prajapati, and S. Tokekar, "The role of Apriori algorithm for finding the association rules in Data mining," in *Issues and Challenges in Intelligent Computing Techniques (ICICT), 2014 International Conference on*, 2014, pp. 657-660.

[42] P. Fournier-Viger, A. Gomariz, T. Gueniche, A. Soltani, C.-W. Wu, and V. S. Tseng, "SPMF: a Java open-source pattern mining library," *The Journal of Machine Learning Research,* vol. 15, pp. 3389-3393, 2014.

[43] M. Al-Maolegi and B. Arkok, "An improved apriori algorithm for association rules," *arXiv preprint arXiv:1403.3948*, 2014.

[44] P.-N. Tan and V. Kumar, "Chapter 6. association analysis: Basic concepts and algorithms," *Introduction to Data Mining. Addison-Wesley. ISBN*, vol. 321321367, 2005.

[45] B. Ramasubbareddy, A. Govardhan, and A. Ramamohanreddy, "Mining positive and negative association rules," in *Computer Science and Education (ICCSE), 2010 5th International Conference on*, 2010, pp. 1403-1406.

[46] X. Wu, C. Zhang, and S. Zhang, "Efficient mining of both positive and negative association rules," *ACM Transactions on Information Systems (TOIS)*, vol. 22, pp. 381-405, 2004.

[47] D. Martin, A. Rosete, J. Alcala-Fdez, and F. Herrera, "A new multiobjective evolutionary algorithm for mining a reduced set of interesting positive and negative quantitative association rules," *IEEE Transactions on Evolutionary Computation*, vol. 18, pp. 54-69, 2014.

[48] B. K. Rani, K. Srinivas, B. R. Reddy, and A. Govardhan, "Mining negative association rules," *International Journal of Engineering and Technology*, vol. 3, pp. 100-105, 2011.

[49] C. Chen and D. Wang, "Research on Association Rules Mining Base on Positive and Negative Items of FP-tree," 2016.

[50] S. Chakraborty, N. Nagwani, and L. Dey, "Performance comparison of incremental k-means and incremental dbscan algorithms," *arXiv preprint arXiv:1406.4751*, 2014.

[51] S. Chakraborty and N. K. Nagwani, "Analysis and study of Incremental DBSCAN clustering algorithm," *arXiv preprint arXiv:1406.4754*, 2014.

[52] J. Štrobl, M. Piorecký, and V. Krajča, "METHODS FOR AUTOMATIC ESTIMATION OF THE NUMBER OF CLUSTERS FOR K-MEANS ALGORITHM USED ON EEG SIGNAL: FEASIBILITY STUDY," *Lékař a technika-Clinician and Technology*, vol. 47, pp. 88-95, 2017.

[53] D. Chen, S. L. Sain, and K. Guo, "Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining," *Journal of Database Marketing & Customer Strategy Management*, vol. 19, pp. 197-208, 2012.

[54] M. F. Sanner, "Python: a programming language for software integration and development," *J Mol Graph Model*, vol. 17, pp. 57-61, 1999.

[55] P. Chévez, D. Barbero, I. Martini, and C. Discoli, "Application of the k-means clustering method for the detection and analysis of areas of homogeneous residential electricity consumption at the Great La Plata region, Buenos Aires, Argentina," *Sustainable Cities and Society*, vol. 32, pp. 115-129, 2017.

[56] Y. Zhang, K. Tangwongsan, and S. Tirthapura, "Streaming Algorithms for k-Means Clustering with Fast Queries," *arXiv preprint arXiv:1701.03826*, 2017.

ÖZGEÇMİŞ

KİŞİSEL BİLGİLER

Adı Soyadı	:	ZAHRAA MOHAMMAD MALIK
Uyruğu	:	Irak
Doğum Yeri ve Tarihi	:	irak 18.7.1989
Medeni Hali	:	Evli
Adres	:	Ahi Mesut Mh.1776 Cad 11/45 Etmesgut-Ankara
E-Posta	:	zahraa_mm17@yahoo.com
İletişim	:	05347276920

EĞİTİM

Lise	:	Baghdad High School / Irak-Baghdad
Lisans	:	Bilgisayar Mühendisliği
Yükse Lisans	:	Türk Hava Kurumu Üniversitesi

MESLEKİ DENEYİM

20011- : Bilgisayar Mühendisliği

YABANCI DİL

İngilizce, Arapça, Türkçe