Öneri C.6.S.22
Haziran 2004 ss.315-321

VERİ MADENCİLİĞİNDE BİRLİKTELİK KURALLARI

Ayşe OĞUZLAR

Uludağ Üniversitesi, İİBF, Ekonometri Bölümü, Yardımcı Doçent. Dr.

ASSOCİATİON RULES İN DATA MİNİNG

Abstract: *Data mining is an interdisciplinary exercise and statistics, database technology, machine learning, pattern recognition, artificial intelligence and visualization, all play a role. Modelling is a very important phase of data mining. One of the modelling techniques is association rules. Association rules associate a particular conclusion with a set of conditions. At the end of the processing, a table of the best rules is presented. The algorithms use a generate and test method for finding rules. Association rule algorithms automatically find the associations that you could find manually using visualization techniques. Apriori algorithm is one of the major algorithm for association rules. In this study DİE 2002 III. term household manpower survey was used. For this data Apriori algorithm was used and the results were explained.*

Keywords: Data Mining, Association Rules, Apriori Algorithm.

VERİ MADENCİLİĞİNDE BİRLİKTELİK KURALLARI

Özet: *Veri madenciliği; istatistik, veri tabanı teknolojisi, makine öğrenimi, örüntü tanıma, yapay zeka ve görselleştirmenin rol oynadığı disiplinler arası bir yaklaşımdır. Modelleme, veri madenciliğinin çok önemli bir aşamasıdır. Modelleme tekniklerinden biri de birliktelik kurallarıdır. Birliktelik kuralları, bir koşul kümesi ile kısmi bir sonuca ulaşma ile ilgilidir. Sürecin sonunda en iyi kuralların bir tablosu oluşmaktadır. Algoritmalar, kuralların bulunması için bir türetim ve test yöntemi kullanmaktadırlar. Birliktelik kural algoritmaları, görselleştirme teknikleri yolu ile ortaya çıkarılan birlikteliklerin, otomatik olarak bulunmasını sağlarlar. Apriori algoritması, birliktelik kuralları türetmek için kullanılan temel algoritmalardan biridir. Bu çalışmada DİE 2002 III. dönem hanehalkı işgücü anketi sonuçları kullanılmıştır.Bu verilere Apriori algoritması uygulanmasıyla elde edilen sonuçlar açıklanmaya çalışılmıştır.*

Anahtar Kelimeler: Veri Madenciliği, Birliktelik Kuralları, Apriori Algoritması

I. GİRİŞ

Günümüzde veri tabanları artık tera byte'larla ölçülmektedir. Bu ölçekte büyük veriler, stratejik öneme sahip bilgileri gizlemektedir [1]. Veri madenciliği, büyük veritabanlarındaki gizli bilgi ve yapıyı açığa çıkarmak için kullanılan bir süreçtir. Veri madenciliği çoğu kişilerce özbilgi keşfi (knowledge discovery) ile aynı anlamda kullanılmaktadır. Bunun tersi olarak bir takım kişiler de veri madenciliğini özbilgi keşif sürecinin basit ve temel bir adımı olarak görmektedir. Özbilgi keşif süreci aşağıdaki adımları içermektedir. [2]

1. Verilerin Temizlenmesi (Gürültülü ve tutarsız verilerin temizlenmesi)
2. Verilerin Birleştirilmesi (Çok sayıdaki veri kaynağının birleştirilmesi)
3. Verilerin Seçilmesi (Analize uygun verilerin veri tabanından seçilmesi)
4. Verilerin Dönüşümü (Verilerin özetleme ve birleştirme süreçleri için uygun formlara dönüştürülmesi)
5. Veri Madenciliği (Sırasıyla veri örüntülerinin ortaya çıkarılması için uygulanan akıllı yöntemlere içeren esas adım)
6. Örüntülerin Değerlendirilmesi (Özbilgiyi temsil eden ilginç örüntülerin teşhis edilmesi) Özbilginin sunumu (Kullanıcı tarafından ortaya çıkarılan özbilginin gösterimi için
7. Görselleştirme ve özbilgi sunum tekniklerinin kullanımı)

Bu sürece göre veri madenciliği özbilgi keşif sürecinin bir adımıdır. Fakat yalnızca basit bir adım değil, değerlemeye alınacak gizli örüntüleri ortaya çıkaran temel bir basamaktır. Veri madenciliği astronomi, biyoloji, finans, sigorta, tıp ve bir çok başka dalda uygulanmaktadır. Son 20 yıldır Amerika Birleşik Devletleri'nde çeşitli veri madenciliği algoritmalarının gizli dinlemeden, vergi kaçakçılıklarının ortaya çıkartılmasına kadar çeşitli uygulamalarda kullanıldığı bilinmektedir [3].

II. VERİ MADENCİLİĞİ İÇİN ÇAPRAZ ENDÜSTRİ STANDART SÜRECİ (CRISP-DM)

Veri madenciliği için izlenecek adımları açıklayan Çapraz Endüstri Standart Prosedürü (CRISP-DM) geliştirilmiştir. Bu prosedür izlenecek adımlar için bir standart sağlamaktadır [4].

Veri madenciliği için genel Çapraz Endüstri Standart Süreci modeli altı aşamadan oluşmaktadır. Bu altı aşama bir dairesel süreçte birbiri ile ilişkilidir. Bu altı aşama tam bir veri madenciliği sürecini içermektedir. Bu altı aşamanın isimleri ve açıklamaları aşağıdaki gibidir:

1. **Business understanding (İşin kavranılması):** Veri madenciliğinin belki de en önemli aşamasıdır. İşin anlaşılması iş hedeflerinin belirlenmesi, durumun değerlendirilmesi, veri madenciliği hedeflerinin belirlenmesi ve bir proje planının üretilmesi işlemlerini kapsamaktadır.
2. **Data understanding (Verinin kavranılması):** Veriler, veri madenciliği için 'hammaddeler' dir. Bu aşama veri kaynaklarının neler olduğunun belirlenmesi ve bu kaynakların karekteristiklerinin anlaşılmasını içerir. Daha açık bir ifade ile başlangıç verilerinin toparlanması, verilerin tanımlanması, verilerin açıklanması ve veri kalitesinin doğrulanması gibi işlemleri içerir.
3. **Data preparation (Verinin hazırlanması):** Verilerin belirlenmesinin ardından, verilerin veri madenciliği için hazırlanması gerekmektedir. Hazırlama işlemi; seçme, temizleme, kurma, birleştirme ve uygun forma dönüştürme gibi işlemleri kapsamaktadır.
4. **Modeling (Modelleme):** Bu aşama veri madenciliğinin en çarpıcı kısmını oluşturmaktadır Bu aşamada analiz yöntemleri verilerden bilginin çıkartılması için kullanılmaktadır. Bu aşama modelleme tekniklerinin seçimi, test tasarımlarının türetilmesi ve modellerin kurulması ve uygulanması gibi adımları içermektedir.
5. **Evaluation (Değerlendirme):** Model seçiminin ardından, veri madenciliği sonuçlarının iş hedefine ulaşmada nasıl yardımcı olacağının değerlendirilmesi gerekmektedir. Bu aşamanın elemanları sonuçların değerlendirilmesi, veri madenciliği sürecinin gözden geçirilmesi ve sonraki adımların belirlenmesi gibi işlemleri içermektedir.
6. **Deployment (Yayılım):** Bu aşama, orijinal problemin çözülmesi için yeni özbilginin gündelik iş süreçleriyle birleştirilmesi işlemine odaklanmıştır. Plan yayılımı, izleme ve bakım, final raporunun üretilmesi ve projenin gözden geçirilmesi gibi adımları içermektedir.

III. VERİ MADENCİLİĞİNDE MODELLEME TEKNİKLERİ

Veri madenciliği bir süreç gerektirdiğinden bu süreç için en önemli adım, veri madenciliğine ihtiyaç duyulup duyulmadığıdır. Bu adım problemin doğru bir biçimde anlaşılmasını gerektirmektedir. Problemin etkin bir şekilde çözümü için aşağıdakiler de sağlanmalıdır:

Problemin perspektifi anlaşılmalı, hedefler ve sınırlamalar bilinmelidir,

Çıktıyı etkileyen önemli faktörler açıklanmalıdır.

Problemin belirlenmesinin ardından veri madenciliğinin başarılı olduğunun belirlenebilmesi için başarı kriteri de tanımlanmalıdır. Başarı kriteri sayısal olabileceği gibi subjektif veya doğası gereği kalitatif de olabilir.

Veri madenciliğinin başarılı olması için problem ve başarı kriterleri belirlendikten sonra, projenin hedefleri veri madenciliği terimleri ile ifade edilmelidir. Aşağıdaki tabloda veri madenciliği hedeflerinin problemin çözüm hedeflerinden farklılık gösterdiği görülebilmektedir [5].

Tablo 1. Problemin ve Veri madenciliğinin hedefi arasındaki farklılık

Problemin hedefi	Veri madenciliği hedefi
Satışlardaki artış	Tüketici özelliklerinin satın alma gücüne bağlı olarak belirlenmesi
Kredi kartı sahtekarlıklarının önlenmesi	Sahte kart kullanımı için kritik örüntülerin bulunması veya otomatik sahtekarlık araştırması için doğru bir algoritmanın kurulması

Problemin tanımı ve veri madenciliği hedefleri veri madenciliği problem tiplerinin temel ayırımıyla doğrudan ilintilidir:

Verilerin tanımlanması ve özetlenmesi
Sınıflandırma
Tahmin
Birliktelik araştırılması
Bağımlılık analizi
Segmentasyon

Kullanılan tekniğe bağlı olarak veri madenciliği çıktıları farklılık göstereceğinden, önce problem tiplerinin tanımlanması uygun olacaktır. Bütün bu aşamaların ardından projenin bir planı yapılmalı tüm adımlar ayrıntılarıyla bu planda belirtilmelidir.

Veri madenciliği, veri tabanı teknolojisi, istatistik, makine öğrenim, örüntü tanımı, yapay sinir ağları, verilerin görselleştirilmesi ve uzaysal veri analizi gibi farklı disiplinlerde yer alan tekniklerin bir birleşimini içerir.

Genel olarak veri madenciliği iki temel kategoriye ayrılabilir: Betimsel (descriptive) ve çıkarımsal (predictive).

Betimsel veri madenciliği, veri tabanındaki verilerin genel özelliklerinin ortaya çıkarılması ile ilgilidir. Çıkarımsal veri madenciliği ise, tahmin amaçlı olarak verilerden çıkarım yapma görevini içerir. Genel

olarak betimsel ve çıkarımsal veri madenciliğinin hedefleri, temel veri madenciliği yöntemlerinden birinin uygulanmasıyla başarılır. Tablo 2' de veri madenciliği problem tipleri ve problemlere uygun modelleme teknikleri gösterilmiştir [6].

Tablo 2.Veri Madenciliği Problemleri ve Uygun Modelleme Teknikleri

Problem	Modelleme Tekniği
Sınıflandırma (Classification)	karar çıkarma yöntemleri, karar ağaçları, sinir ağları, K-en yakın komşuluk, duruma dayalı nedensellik
Çıkarım (Prediction)	regresyon analizi, regresyon ağaçları, sinir ağları, K-en yakın komşuluk.
Bağımlılık Analizi (Dependency Analysis)	korelasyon analizi, regresyon analizi, birliktelik kuralları, Bayesian ağlar, inductive logic programlama.
Verilerin Tanımlanması Ve Özetlenmesi (Data description and summarization)	İstatistiksel teknikler, OLAP.
Bölme Veya Kümeleme (Segmentation Or Clustering)	Kümeleme teknikleri, sinir ağları, görselleştirme yöntemleri.

Modelleme teknikleri veri madenciliğinde önemlidir çünkü geleneksel istatistiksel analizlere ek olarak, önemli değişkenler arasındaki ilişkiler hakkında otomatik olarak yeni modeller türetirler. Modelleme tekniklerinden biri olan birliktelik kuralına aşağıda genel hatlarıyla değinilmiştir.

IV. BİRLİKTELİK KURALLARI (ASSOCİATİON RULES)

Bir kural, sol kısım (öncül veya koşul) ile sağ kısımdan (bağlı kısım) oluşur. Örneğin 'Yağmur yağarsa, yer ıslanır.'. Sol ve sağ kısmın her ikisi de Boolean (doğru veya yanlış) durumlarından oluşur. Kural, sol taraf (antecedent(s)) doğru olduğunda sağ tarafın da (consequent(s)) doğru olduğuna işaret eder. Bir olasılıksal kural ise, sol tarafın doğru olduğu bilindiğinde, sağ tarafın da p olasılığıyla doğru olmasını gösterir. p olasılığı, sol tarafın doğru olduğu bilindiğinde, sağ tarafın da doğru olmasının basit koşullu olasılığıdır [7].

Kurallar doğası gereği kesiklidir. Bu nedenle kurallar kesikli ve kategorik değişkenlerin modellenmesinde iyi uyum gösterir.

Birliktelik Analizi, verilen bir veri kümesinde sıkça tekrarlanan nitelik-değerli koşulları gösteren birliktelik kurallarının araştırılmasıdır. Birliktelik analizi pazar sepet (market basket) analizinde geniş bir kullanıma sahiptir. Pazar sepet analizinde, tüketicinin alışverişi sürecince beraber aldığı mallar belirlenir. Pazar sepet analizinin çıktısı, tüketici satın alma davranışına ilişkin birliktelikler kümesidir. Birliktelikler, birliktelik kuralları (association rules) olarak bilinen kuralların özel bir kümesi formunda verilir. Birliktelik kuralları, uygun ürün pazarlama stratejisinin belirlenmesinde yardımcı olacaktır [8].

Birliktelik Kuralları $X \Rightarrow Y$ formundadır. Bu birliktelik kuralı, X kuralını sağlayan veritabanındaki kayıtların aynı zamanda Y' nin koşullarını da sağladığı anlamındadır.

IV.1 Güven ve Destek (Confidence and Support)

Birliktelik kurallarında, geleneksel sınıflama kurallarının aksine bir nitelik bir kuralda ön koşul olarak görünürken, diğerinde bağlı koşul olarak görünebilmektedir. Bunun yanında geleneksel sınıflama kuralları bir kuralın bağlı kısmı tek nitelik olacak biçimde sınırlandırmıştır. Birliktelik kuralında ise bir kuralın bağlı kısmı, bir veya birden fazla nitelik değeri içerebilir. Birliktelik kuralları, tek bağlı değişken seçimi sınırlaması olmadan büyük veritabanlarındaki ilişkilerin bulunmasını sağladığından popülerdirler.

Tüketicinin satın alma trendindeki ilginç ilişkiler aşağıdaki dört ürün için belirlenmeye çalışılacaktır:

- Süt
- Peynir
- Ekmek
- Yumurta

Mümkün birliktelikler ise aşağıdaki biçimde verilmiştir:

1. Tüketici süt aldığında aynı zamanda ekmek de alır.
2. ,Tüketici ekmek aldığında aynı zamanda süt de alır.
3. Tüketici süt ve yumurta aldığında aynı zamanda peynir ve ekmek de alır.
4. Tüketici süt, peynir ve yumurta aldığında ekmek de alır.

İlk birliktelik, süt alan tüketicilerin aynı zamanda ekmek de satın aldığını söylemektedir. Bu kural için güven, süt alındığında ekmek almanın koşullu olasılığıdır. Eğer 10000 tüketici süt almışsa ve bunların 5000 aynı zamanda ekmek almışlarsa, süt alındığında ekmek de almanın güveni 5000/10000=%50 olarak bulunur.

İkinci kurala bakıldığında ilk kural ile aynı bilgiyi vermediği anlaşılabilmektedir. Bu birliktelikteki ilk kural ekmek satın alınmış olmasıdır. Bir örnek olarak 20000 tüketicinin ekmek aldığı biliniyorsa ve bunların herhangi 5000 tanesi aynı zamanda süt almış olsalar, ekmek alındığında süt de almanın güven değeri %25 olacaktır. Üçüncü ve dördüncü birlikteliklerin, birinci ve ikinci ile benzer biçimde güven değerleri hesaplanabilir.

Bir birliktelik kuralından bulunan nitelik değerlerini içeren tüm örneklerin yüzdesi güven

değerinden farklı olacaktır. Bu istatistik, kural için destek (support) olarak bilinir. Destek, en basit anlamıyla, belirli bir birliktelik kuralındaki tüm nitelikleri içeren, veri tabanındaki örneklerin minimum yüzdesidir. Destek ve güven değerleri öncül kısım A ile ve bağlı kısım da B ile gösterilmek üzere aşağıdaki biçimde gösterilebilir:

Destek $(A \Rightarrow B)=P(A \cup B)$
Güven $(A \Rightarrow B)=P(B \backslash A)$

Birliktelik kural madenciliği iki adımlı bir süreçtir:

1. Tüm sık tekrarlanan kümelerin bulunması: Tüm kümelerin her biri en az önceden belirlenen minimum destek sayısı kadar tekrarlanacaktır.
2. Sık tekrarlanan kümelerden güçlü birliktelik kurallarının türetilmesi: Bu kurallar minimum destek ve minimum güveni sağlamak zorundadır.

Etkin birliktelik kuralları üretebilmek için özel algoritmalar geliştirilmiştir. Bu tip algoritmalardan biri Apriori algoritmasıdır (Agrawal ve diğerleri). Bu algoritma madde kümeleri (itemsets) üretir. Madde kümeleri, belirli bir kapsam gereksinimini sağlayan nitelik-değer kombinasyonlarıdır. Bu nitelik-değer kombinasyonları, kapsamın içerdiği gereksinimlerin dışındakileri eler. Bu sayede kural türetim süreci makul bir zaman diliminde tamamlanır.

Apriori birliktelik kural türetimi iki adımdan oluşan bir süreçtir. İlk adım da madde kümeleri türetilir. İkinci adım, birliktelik kural kümesi yaratacak şekilde, üretilmiş madde kümelerinin kullanılmasını içerir.

Uygulama bölümünde Apriori algoritmasından yararlanılarak 2002 hanehalkı işgücü anketi verilerinden birliktelik kuralları türetilmeye çalışılmıştır. Kullanılan değişkenler ve elde edilen sonuçlar uygulama bölümünde özetlenmiştir.

V. UYGULAMA

Apriori algoritması için uygulama bölümünde 2002 hanehalkı işgücü anketi veri kümesinden yararlanılmıştır. Bu veri kümesinin 300689 gözlem değeri için kullanılan değişkenler aşağıdaki şekildedir:

1. Cinsiyet
    1. Erkek
    2. Kadın
2. Bitirilen Yaş
3. Hanehalkı Reisine Yakınlık
    1. Hanehalkı reisi
    2. Eşi
    3. Çocuğu
    4. Gelini veya damadı
    5. Torunu
    6. Ebeveyni
    7. Diğer Akrabası
    8. Akraba Olmayan
4. En Son Bitirilen Okul
    1. Bir okul bitirmedi
    2. İlkokul
    3. İlköğretim
    4. Ortaokul
    5. Meslek ortaokulu
    6. Lise
    7. Mesleki Lise
    8. 2 yıllık ön lisans
    9. 2 yıllık lisans
    10. 4 yıllık lisans
    11. Mastır, doktora vb.
5. Medeni Durum
    1. Hiç evlenmedi
    2. Evli
    3. Boşandı
    4. Eşi öldü
6. İş arama
    1. Evet
    2. Hayır
7. Kır-Kent
    1. Kır
    2. Kent

değişkenleri ele alınmıştır.

SPSS formatındaki veriler öncelikle Clementine 7.0 veri madenciliği programına aktarılmıştır [9]. Ardından veri kümesinde yer alan değişkenlerin türleri programa tanıtılarak, kayıp gözlemler veri kümesinden dışlanılmıştır [10]. Programın Modeling menüsünden Apriori modülü seçilerek uygulanmıştır [11]. Türetilecek birliktelik kurallarının bağlı kısmı olarak ISSIZ olarak kısaltılmış iş arama değişkeni kullanılmıştır. Birliktelik kuralının öncül kısmında ise ilk olarak cinsiyet, okul ile belirtilen en son bitirilen okul, medeni hal değişkeni olan medhal ve hanehalkı reisine yakınlığı gösteren haresyak değişkenleri el alınmıştır. Elde edilen beş birliktelik kuralı ile destek ve güven değerleri Tablo 3' de görülmektedir. Tablodaki sıralama güven değerleri büyükten küçüğe doğru sıralanacak biçimde gerçekleştirilmiştir.

Birliktelik kuralları ile özet değerlerin yorumlanması sırası ile aşağıdaki biçimde yapılabilir:

1.Kadın, ilkokul mezunu ve hanehalkı reisinin eşi olan kişiler iş aramamaktadır. Kadın, ilkokul mezunu, hanehalkı reisinin eşi ve iş aramayan kişi sayısı 33780' dir. Toplam kişi sayısı 300689 olduğundan destek değeri % 11,2' dir. Ön koşulları sağlayan kişilerden yani kadın, ilkokul mezunu ve hanehalkı reisinin eşi olanların, %

82,7' si bağlı koşulu sağlamaktadır ve iş aramamaktadır. Bu değer Tablo 3'de güven değeri olarak belirlenmiştir.

2.Kadın, evli, ilkokul mezunu ve hanehalkı reisinin eşi olan kişiler iş aramamaktadır. Kadın, evli, ilkokul mezunu, hanehalkı reisinin eşi ve iş aramayan kişi sayısı 33780' dir. Toplam kişi sayısı 300689 olduğundan destek değeri yine % 11,2 bulunmuştur. Ön koşulları sağlayan kişilerden dolayısıyla kadın, evli, ilkokul mezunu ve hanehalkı reisinin eşi olanların, % 82,7' si bağlı koşulu sağlamaktadır ve iş aramamaktadır.

Tablo 3.

Instances	Support	Confidence	Consequent	Antecedent1	Antecedent2	Antecedent3	Antecedent4
33780	11.2	82.7	ıSSıZ = 2	CıNSıYET=2	OKUL=2	HARESYAK=2	
33780	11.2	82.7	ıSSıZ = 2	CıNSıYET=2	MEDHAL=2	OKUL=2	HARESYAK=2
33938	11.3	82.4	ıSSıZ = 2	OKUL=2	HARESYAK=2		
33938	11.3	82.4	ıSSıZ = 2	MEDHAL=2	OKUL=2	HARESYAK=2	
39213	13.0	80.8	ıSSıZ = 2	CıNSıYET=2	MEDHAL=2	OKUL=2	

3.Tablo 3' deki üçüncü birliktelik kuralından ise ilkokul mezunu ve hanehalkı reisinin eşi olanların iş aramadığını söylemek mümkündür. İlkokul mezunu, hanehalkı reisinin eşi ve iş aramayan kişi sayısı 33938' dir. Toplam kişi sayısı 300689 olduğundan destek değeri % 11,3 bulunmuştur. Ön koşulları sağlayan ilkokul mezunu ve hanehalkı reisinin eşi olanların, % 82,4' ü bağlı koşulu sağlamaktadır ve iş aramamaktadır.

4.Dördüncü birliktelik kuralına bakıldığında evli, ilkokul mezunu ve hanehalkı reisinin eşi olanların iş aramadığı anlaşılmaktadır. Evli, ilkokul mezunu, hanehalkı reisinin eşi ve iş aramayan kişi sayısı 33938 çıkmıştır. Dolayısıyla bu birliktelik kuralı için destek değeri %11,3' dür. Evli, ilkokul mezunu ve hanehalkı reisinin eşi olanların %82,4' ü iş aramamaktadır.

5.Son birliktelik kuralından ise kadın, evli ve ilkokul mezunlarının iş aramadığını söylemek doğru olacaktır. Kadın, evli, ilkokul mezunu ve iş aramayanlar 39213 kişidir. Dolayısıyla bu birliktelik kuralı için destek değeri %13' dür. Kadın, evli ve ilkokul mezunu olanların %80,8' i iş aramamaktadır.

Tablo 4' de ise ikinci bir grup olarak ele alınan ve kır-kent değişkeninin de analize dahil edildiği birliktelik kuralları ile elde edilen destek ve güven değerleri görülmektedir. Bu ikinci grup için de birliktelik kurallarının bağlı kısmı olarak ISSIZ olarak kısaltılmış iş arama değişkeni kullanılmıştır. Birliktelik kuralının öncül kısmında ise ilk olarak kır-kent, cinsiyet, okul, medeni hal değişkeni olan MEDHAL ve hanehalkı reisine yakınlığı gösteren HARESYAK değişkenleri el alınmıştır. Elde edilen 11 birliktelik kuralı ile destek ve güven değerleri Tablo 4' de gösterilmiştir. Tablodaki sıralama yine güven değerleri büyükten küçüğe doğru sıralanacak biçimde gerçekleştirilmiştir.

1.Birliktelik kuralı: Kentte oturan, kadın ve ilkokul mezunu olan kişiler iş aramamaktadır. Kentte oturan, kadın, ilkokul mezunu ve iş aramayan kişi sayısı 37594' dür. Toplam kişi sayısı 300689 olduğundan destek değeri % 12,5' dir. Ön koşulları sağlayan kişilerden yani kentte oturan, kadın ve ilkokul mezunu olanların, % 87,1' i bağlı koşulu sağlamaktadır ve iş aramamaktadır.

2.Birliktelik Kuralı: Kentte oturan, kadın ve hanehalkı reisinin eşi olan kişiler iş aramamaktadır. Kentte oturan, kadın, hanehalkı reisinin eşi ve iş aramayan kişi sayısı 49108' dir. Bu kural için destek değeri % 16,3' dür. Ön koşulları sağlayan kişilerden yani kentte oturan, kadın ve hanehalkı reisinin eşi olanların, % 85,3' ü bağlı koşulu sağlamaktadır ve iş aramamaktadır.

3.Birliktelik Kuralı: Kentte oturan, kadın, evli ve hanehalkı reisinin eşi olan kişiler iş aramamaktadır. Kentte oturan, kadın, evli, hanehalkı reisinin eşi ve iş aramayan kişi sayısı 49108' dir. Bu kural için destek değeri % 16,3' dür. Ön koşulları sağlayan kişilerden yani kentte oturan, kadın, evli ve hanehalkı reisinin eşi olanların, % 85,3' ü bağlı koşulu sağlamaktadır ve iş aramamaktadır.

4.Birliktelik Kuralı: Kentte oturan, kadın ve evli olan kişiler iş aramamaktadır. Kentte oturan, kadın, evli, ve iş aramayan kişi sayısı evli olanların, % 85,2' si 54467' dir. Bu kural için destek değeri % 18,1' dir. Ön koşulları sağlayan kişilerden yani kentte oturan, kadın ve bağlı koşulu sağlamaktadır ve iş aramamaktadır. Destek değeri en büyük olan birliktelik kuralı bu kuraldır.

Tablo 4

Instances	Support	Confidence	Consequent	Antecedent1	Antecedent2	Antecedent 3	Antecedent4
37594	12.5	87.1	ıSSıZ = 2	KIR_KENT=2	CıNSıYET=2	OKUL=2	
49108	16.3	85.3	ıSSıZ = 2	KIR_KENT=2	CıNSıYET=2	HARESYAK=2	
49108	16.3	85.3	ıSSıZ = 2	KIR_KENT=2	CıNSıYET=2	MEDHAL=2	HARESYAK=2
54467	18.1	85.2	ıSSıZ = 2	KIR_KENT=2	CıNSıYET=2	MEDHAL=2	
49340	16.4	85.0	ıSSıZ = 2	KIR_KENT=2	HARESYAK=2		
49340	16.4	85.0	ıSSıZ = 2	KIR_KENT=2	MEDHAL=2	HARESYAK=2	
33780	11.2	82.7	ıSSıZ = 2	CıNSıYET=2	OKUL=2	HARESYAK=2	
33780	11.2	82.7	ıSSıZ = 2	CıNSıYET=2	MEDHAL=2	OKUL=2	HARESYAK=2
33938	11.3	82.4	ıSSıZ = 2	OKUL=2	HARESYAK=2		
33938	11.3	82.4	ıSSıZ = 2	MEDHAL = 2	OKUL=2	HARESYAK=2	
39213	13.0	80.8	ıSSıZ = 2	CıNSıYET=2	MEDHAL=2	OKUL=2	

5. Birliktelik Kuralı: Kentte oturan ve hanehalkı reisinin eşi olanlar iş aramamaktadır. Kentte oturan, hanehalkı reisinin eşi ve iş aramayan kişi sayısı 49340' dır. Bu kural için destek değeri %16,4 olarak bulunmuştur. Kentte oturan ve hanehalkı reisinin eşi olup da aynı zamanda iş aramayan kişi yüzdesi %85' dir.

6. Birliktelik Kuralı: Kentte oturan, evli ve hanehalkı reisinin eşi olanlar iş aramamaktadır. Kentte oturan, evli, hanehalkı reisinin eşi ve iş aramayan kişi sayısı 49340' dır. Bu kural için destek değeri yine bir önceki birliktelik kuralında olduğu gibi %16,4 olarak bulunmuştur. Kentte oturan, evli ve hanehalkı reisinin eşi olan ve aynı zamanda da iş aramayan kişi yüzdesi %85' dir.

7. Birliktelik Kuralı: Kadın, ilkokul mezunu, hanehalkı reisinin eşi olanlar iş aramamaktadır. Kadın, İlkokul mezunu, hanehalkı reisinin eşi ve iş aramayan kişi sayısı 33780 ve bu kural için destek değeri %11,2' dir. Kadın, ilkokul mezunu ve hanehalkı reisinin eşi olanlardan, bağlı koşulu da sağlayan dolayısıyla iş aramayanların yüzdesi ise %82,7' dir.

8. Birliktelik Kuralı: Kadın, evli, ilkokul mezunu, hanehalkı reisinin eşi olan kişiler iş aramamaktadır. Kadın, evli, ilkokul mezunu, hanehalkı reisinin eşi ve iş aramayanlar toplam 33780 kişidir. Bu birliktelik kuralı için destek değeri %11,2, ve güven değeri ise %82,7' dir.

9. Birliktelik Kuralı: İlkokul mezunu, hanehalkı reisinin eşi olanlar iş aramamaktadır. İlkokul mezunu, hanehalkı reisinin eşi olan ve iş aramayan kişi sayısı 33938' dir. Bu birliktelik kuralı için destek değeri %11,3, güven değeri ise %82,4 bulunmuştur.

10. Birliktelik Kuralı: Evli, ilkokul mezunu, hanehalkı reisinin eşi olanlar iş aramamaktadır. Evli, ilkokul mezunu, hanehalkı reisinin eşi olan ve iş aramayan 33938 kişi vardır. Bu kural için destek değeri%11,3 ve güven değeri ise % 82,4 olarak bulunmuştur.

11. Birliktelik Kuralı: Kadın, evli ve ilkokul mezunu olanlar iş aramamaktadır. Kadın, evli, ilkokul mezunu olup iş aramayan 39213 kişi mevcuttur. Bu son birliktelik kuralı için destek değeri %13 ve güven değeri ise % 80,8 olarak bulunmuştur.

VI. SONUÇ

Veri madenciliği çeşitli açılardan geleneksel istatistiksel yöntemlerle önemli farklılıklar gösterir. Özellikle zaman içinde verinin azlığının değil, çokluğunun bir sorun olması ve bilgisayarların veri saklama ve işleme hızlarındaki inanılmaz artışların sonucunda veri madenciliğinin popülerliği her geçen gün artmaktadır. Veri madenciliğinde birliktelik kurallarının ayrı bir yeri vardır. Özellikle pazar sepet analizi analizi uygulamalarında yararlanılan birliktelik kurallarında önemli iki değer, güven ve destek değerleridir. Çalışmamızda 2002 hanehalkı işgücü anketi verilerinden 300689 veri ve belirtilen değişkenler için Apriori algoritması kullanılarak birliktelik kuralları ile bu kuralların güven ve destek değerleri bulunmuştur. İlk bulunan 5 birliktelik kuralı içerisinden destek değeri en yüksek olan 5. birliktelik kuralıdır. Bu kuralın işaret ettiği öncül ve bağlı koşulları sağlayan kişi sayısı 39213' dür. Bu kurala göre kadın, evli ve ilkokul mezunu olanlar iş aramamaktadır.

Kır-kent değişkenini de değişken kümesine dahil ettiğimizde daha yüksek destek değerlerine ulaşılmıştır. Bu ikinci grup değişken için bulunan birliktelik kuralları içerisinden en yüksek destek değerine sahip olan ise 4. birliktelik kuralıdır. Bu birliktelik kuralının işaret ettiği öncül ve bağlı koşulları sağlayan kişi sayısı 54467' dir. Bu birliktelik kuralına göre ise kentte oturan, kadın ve evli kişiler iş aramamaktadır

Bulunan tüm birliktelik kuralları iş aramayanlar için oluşmuştur. Bunun nedeni veri kümesinde iş aramayan kişilerin oldukça fazla oluşudur. Dolayısıyla algoritma, bağlı koşulun iş arama olduğu durum (ISSIZ=1) için bir kural üretememektedir. Çarpıcı bir başka durum ise kır-kent değişkeninin birliktelik kurallarında destek değerini yükseltmesidir.

YARARLANILAN KAYNAKLAR

[1] ZHOU Z. (2002), "Three perspectives of data mining" **Artificial Intelligence**, 143 (2003).

[2] HAN J. and KAMBER M. (2001), **Data Mining: Concepts and Techniques**, Morgan Kaufmann Publishers, USA.

[3] AKPINAR, H. (2000), "Veri Tabanlarında Bilgi Keşfi ve Veri Madenciliği" **İ.Ü. İşletme Fakültesi Dergisi**, C:29 S:1, s.1-22.

[4] LUAN J. (2002), "Data Mining and Knowledge Management in Higher Education-Potential Applications" **Presentation at AIR Forum**, Toronto-Canada, pp.1-18.

[5] DMS Tutorial - Problem Understanding, http://dms.irb.hr/tutorial/tut_prob_understand.php.

[6] DMS Tutorial - Modelling techniques, http://dms.irb.hr/tutorial/tut_modelling.php.

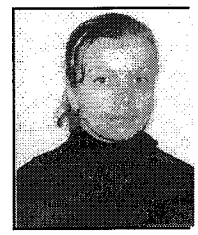
[7] HAND David, MANNİLA Heikki ve SMYTH Padhraic, **Principles of Data Mining,** MIT Press, USA, 2001.

[8] ROIGER Richard J. ve GEATZ Michael W., **Data Mining A Tutorial-Based Primer**, Addison Wesley, USA, 2003.

[9] GOEBEL M. and Gruenwald L. (1999), "A Survey Of Data Mining And Knowledge Discovery Software Tools", **SIGKDD Explorations**, Volume:1, Issue:1, USA.

[10] KIM W., CHOI B., HONG E., KIM S. And LEE D. (2003), "A Taxonomy of Dirty Data", **Data Mining and Knowledge Discovery**, 7.

[11] Clementine 7.0 User's Guide, SPSS Inc., USA, 2002.



Ayşe OĞUZLAR

Uludağ Üniversitesi, İ.İ.B.F.,
Ekonometri Bölümü
Görükle-BURSA

(224)442 89 41 – 41158

ayseog@uludag.edu.tr

Ayşe OĞUZLAR has Ph. D. Of Statistics at Uludag University Instute of Social Sciences. She is Assistant Professor since October 2000 in Uludag University, Department of Econometrics. Her scientific interest are multivariate statistical analysis, categorical data analysis, data mining and six sigma.