T.C
FIRAT ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

GİZLİLİĞİ KORUYAN BULANIK VERİ MADENCİLİĞİ YÖNTEMLERİNİN GELİŞTİRİLMESİ

Tolga BERBEROĞLU

Tez Yöntecisi
Yrd.Doç.Dr. Mehmet KAYA

YÜKSEK LİSANS TEZİ
BİLGİSAYAR MÜHENDİSLİĞİ ANA BİLİM DALI

ELAZIĞ,2008

T.C
FIRAT ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

GİZLİLİĞİ KORUYAN BULANIK VERİ MADENCİLİĞİ YÖNTEMLERİNİN GELİŞTİRİLMESİ

Tolga BERBEROĞLU

Yüksek Lisans Tezi
Bilgisayar Mühendisliği Anabilim Dalı

Bu tez, …… tarihinde aşağıda belirtilen jüri tarafından oybirliği / oyçokluğu ile başarılı / başarısız olarak değerlendirilmiştir.

Danışman : Yrd.Doç.Dr. Mehmet KAYA
Üye : Prof.Dr. Yakup DEMİR
Üye : Yrd.Doç.Dr. Burhan ERGEN

Bu tezin kabulü, Fen Bilimleri Enstitüsü Yönetim Kurulu'nun …../…../…….. tarihi ve …….. sayılı kararıyla onaylanmıştır.

TEŞEKKÜR

Bu tez çalışması boyunca bana yardımlarını esirgemeyen tez danışmanım sayın Yrd. Doç. Dr. Mehmet KAYA'ya ve bu tez çalışması için proje desteği sağlayan FÜBAP'a teşekkürlerimi sunarım.

İÇİNDEKİLER

Sayfa

ŞEKİLLER LİSTESİ

Sayfa

TABLOLAR LİSTESİ

KISALTMALAR

ISL : Kuralın sol tarafının destek değerinin arttırma

DSR : Kuralın sağ tarafının destek değerini azaltma

LHS : Kuralın sol tarafı

RHS : Kuralın sağ tarafı

ID : Kişiyi tanımlayıcı kimlik bilgisi

D : Orijinal veri tabanı.

F : Veri tabanını bulanıklaştırılmış hali .

X : Nesnelerin bir kümesi,

T_L : Sol taraftaki nesneye ait işlem verileri

T_R : Sağ taraftaki nesneye ait işlem verileri

U : Kural,

f : F veri tabanındaki nesneleri ifade etmektedir.

α : Kullanıcı tanımlı minimum destek değeri

λ : Kullanıcı tanımlı minimum güven değeri

I : İşlem verisi

z : Bulanık mantık işlemlerindeki az değeri

o : Bulanık mantık işlemlerindeki orta Değeri

b : Bulanık mantık işlemlerindeki çok değeri

ÖZET

Yüksek Lisans Tezi

GİZLİLİĞİ KORUYAN BULANIK VERİ MADENCİLİĞİ YÖNTEMLERİNİN GELİŞTİRİLMESİ

Tolga BERBEROĞLU

Fırat Üniversitesi

Fen Bilimleri Enstitüsü

Bilgisayar Mühendisliği Anabilim Dalı

2008, Sayfa: 51

Veri madenciliği veya veri tabanlarında bilginin keşfi, büyük miktardaki veriden önceden bilinmeyen ilişkilerin otomatik çıkarılması için yapılan araştırmalardır. Veri madenciliği üzerine yapılan çalışmalardaki artış ve internet ile diğer medya organları sayesinde bulunan bilgilerin çok fazla ortak kişilerle paylaşılması, kişisel gizlilik problemlerini ortaya çıkarmıştır. Kişisel gizliliği korumak için bir çok yöntem geliştirilmiştir, bu yöntemlerin çoğu keskin (1-0) değerleri üzerinde işlem yapmaktadır. Gerçek hayatta veri tabanları nicel değerlerden oluşmaktadır. Örneğin, insan kanında yapılan veri madenciliği araştırmalarında, kanda akyuvar olup olmadığı gibi keskin bir araştırmadan ziyade bunların oranı üzerinde çalışılmaktadır. Bu nicel değerler üzerinde yapılan veri madenciliği araştırmalarında, kişisel gizliliği korumaya yönelik araştırmalar çok azdır.

Bu tez çalışmasında, nicel değerli veri tabanlarında bulanık mantık yöntemi kullanılarak kişisel gizliğin sağlanması gerçekleştirilmiştir. Günümüzde nicel değerler en çok hastalık verilerinde bulunmaktadır. Bu tezde, birliktelik ve sınıflandırma işlemlerinde çıkabilecek kişisel gizlilik problemlerine karşı iki algoritma geliştirilmiştir. Bu algoritmalar için göğüs kanseri verileri üzerinde çalışan bir uygulama yazılımı oluşturulmuştur.

Anahtar Kelimeler: Veri madenciliği, Kişisel Gizlilik, Nicel değerli veriler,

ABSTRACT

Master Thesis

DEVELOPMENT OF PRIVACY PRESERVING FUZZY DATA MINING METHODS

Tolga BERBEROĞLU

Firat University

Institute of Natural an Applied Science

Department of Computer Enginering

2008, Page: 51

Data mining or discovering knowledge from a database are processes that analyzing the unknown relationships in the that database. Recently, the problem of privacy preserving occurs due to increasing of studies on this area and sharing the large information over internet and media environments. Some techniques have been developed fort his purpose, however many of these studies are working based on crisp (Bloolean) values. In real world, the databases of applications consist of quantitative values. For instance; in analyzing the leukocyte in one's blood, the main goal is to find the ratio of leukocyte instead of whether leukocyte is exists. The count of studies on the privacy preserving on such quantitative databases is very small.

In this thesis, we have implemented the privacy preserving on quantitative databases by using fuzzy logic techniques. Today, the most application area of such databases are about sickness knowledge. We have developed two algorithms fort he problems about the relations and clustering of these databases. The application software for a case study on the database of breast cancer have been developed.

Keywords : Data Mining, Privacy Preserving, Quantitative values

1. GİRİŞ

Veri madenciliği veya veri tabanlarında bilginin keşfi, büyük miktardaki veriden önceden bilinmeyen ilişkilerin otomatik çıkarılması araştırmalarıdır [1-4]. Veri depolama ve dağıtımları ile ilişkili teknolojilerindeki gelişmeler var olan veri madenciliği algoritmalarına farklı bir noktadan bakılmasına neden olmuşlardır. Veri madenciliği üzerine yapılan araştırmalardaki artış ve internet ile diğer medya organları sayesinde bulunan bilgilerin çok fazla ortak kişilerle paylaşılması, kişisel gizlilik problemlerine daha ciddiyetle yaklaşılmasına neden olmaktadır [2].

Gizliliği korumanın kesin ve ucuz yolu, yetkisiz kullanıcıların veriye erişmesini engellemekle kolayca halledilebilir. Fakat bu istenilmeyen bir yoldu, çünkü veri madenciliğinin birçok faydası vardır. Örneğin kolektif verinin analiz edilmesi, bir şirketin daha verimli çalışma yollarının bulmasında yardımcı olur. Veriye erişim engellendiğinde, bu faydalarında çıkarılması engellenmiş olur. Veri madenciliği; veri tabanı teknolojisi, istatistik, yapay zeka, makine öğrenmesi, örüntü tanımlama ve veri görselleştirmesi gibi pek çok teknik alan arasında köprü görevi yapan çok disiplinli bir alandır[1]. Veri madenciliği işlemleri gerçek amaçları doğrultusunda kullanıldığında özellikle astronomi, biyoloji, finans, pazarlama, sigorta, tıp gibi birçok alanda faydalı araştırmalar yapılmasına yardımcı olmuştur[3]. Veri madenciliğini kullanan kişilerin kötü niyetli olması ise faydadan çok zarar verilmesine neden olmaktadır. Peki, kişisel gizlilik ihlal edilmeden veri madenciliği çalışması yapılması mümkün müdür? Gerekli titizlik gösterilerek bu sorunun cevabının evet olduğuna inanılmaktadır.

Veri madenciliği ve bilgi güvenlik komitesi yakın zamanlarda toplanmış, kötü niyetli yapılan veri madenciliği sorunlarını adreslemişlerdir. Bu sorunları çözmek için bir çok teknik geliştirilmiştir. Bu tekniklerden bazıları Karıştırma (Perturbation) , Engelleme (Blocking), Veri Değiş Tokuşu (Swapping) ve Güvenli Çok Gruplu hesaplamadır. Yukarıdaki teknikler çeşitli gruplarda incelenmektedir.

Veri madenciliğinde önceden bilinmeyen birlikteliklerin bulunmasında sık kullanılan bağıntı kural madenciliği yönteminde hassas, çıkarılması tehlikeli olabilecek kuralların engellenmesi içinde bazı teknikler geliştirilmiştir: Hassas kuraların Destek değerlerinin keskin (1,0) değerleri ile değiştirilmesiyle bağıntı kurallarını gizleme veya hassas kuralların nesnelerinin, değerlerinin bilinmeyenler (?) ile değiştirilmesiyle bağıntı kurallarının çıkarılmasının engellenmesi [3]. Bu yöntemlerde keskin (1,0) değerleri üzerinde işlem yapılmaktadır. Fakat bağıntı kuralları, sadece keskin (1,0) değerlerinden ibaret değildir. Birçok birliktelik nicel değerler içermektedir. (10 tane “bira” veya 20 tane “çocuk bezi” gibi). Bu

birliktelik kurallarının çıkarılmasında kullanılan tekniklerden biri bulanık bağıntı kuralları yöntemidir. Bu yöntemin kötü niyetli kullanımlarından da çıkabilecek kişisel gizlilik sorunlarını engellemek için bir teknik henüz geliştirilmemiştir.

1.1. Bu Tezin Amacı

Bu tezde, nicel değerli verilerin bulunduğu veri tabanından hassas bağıntı ve sınıflandırma kurallarının çıkarılmasını önleme amaçlanmıştır. Veri madenciliği yeni ve her gün gelişen bir teknolojidir. Bu teknolojinin kötü niyetli kullanıcılar tarafından kullanılması, başta kişisel gizlilik olmak üzere birçok soruna neden olmaktadır. Bu hassas bağıntı ve sınıflandırma kurallarının engellenmesi için birçok teknik geliştirilmiş olmasına rağmen, bu tekniklerin çoğu hassas keskin değerler üzerinde işlem ve sınıflandırma yapmaktadır. Nicel değerli verilerin olduğu veri tabanlarında çıkarılabilecek hassas bağıntı kurallarını gizlemek için herhangi bir teknik geliştirilmemiştir. Bu tezde, bulanık mantık yöntemi kullanılarak bu hassas kuralların çıkarılmasını engellemek için geliştirilen yöntem ve uygulama sonuçları anlatılmaktadır.

1.2. Tezde Geliştirilenler

Bu tezde, nicel değerli verilerin bulunduğu bir veri tabanından biri hassas bağıntı kurallarının çıkarılmasını engelleyen diğeri de hassas sınıflandırma kurallarının çıkarılmasını engelleyen iki yöntem geliştirilmiştir. Bu algoritmalarda önce, olarak bulanık mantık yöntemi kullanılarak çıkabilecek birliktelikler bulunur. Daha sonra bu birlikteliklerden, kullanıcı tarafından hassas olabileceği tahmin edilen kurallar belirlenir. Bu kuralların gizlenmesi için önerilen algoritma ile birlikteliklerin sol tarafındaki nicelikler üzerinde değişiklik yapılarak, kuralın güven değeri güven eşik değerinden daha aşağı bir seviyeye çekilir. Bu tezde geliştirilen algoritmalar, göğüs kanseri hastalığına ait gerçek bir veri kümesi üzerinde denenmiş ve elde edilen sonuçlar tezin ilgili bölümlerinde verilmiştir.

1.3. Tez Taslağı

İkinci bölümde, veri madenciliği ve kişisel gizliliğin tanımları yapılmıştır. Veri madenciliğinin kişisel gizliliği nasıl ihlal ettiğine dair örnekler verilmiştir. Kişisel gizliliği korumak için önerilen yöntemler tanıtılmıştır. Üçüncü bölümde, geliştirilecek algoritmanın temel aldığı bağıntı kuralları ve bulanık küme kavramı tanıtılmıştır. Dördüncü bölümde, nicel değerli veriler üzerinde hassas bağıntı kurallarının çıkarılmasını engellemek için geliştirilen algoritmadan bahsedilmiştir. Beşinci bölümde, sınıflandırma ve hassas sınıflandırmaların engellenmesi için geliştirilen algoritma anlatılmıştır. Altıncı bölüm, geliştirilen yazılımın gerçek

bir veri kümesi üzerinde testini içerir. Yedinci bölümde ise elde edilen sonuçlara yer verilmektedir.

2. GİZLİLİĞİ KORUYAN VERİ MADENCİLİĞİ

2.1 Gizlilik Nedir ?

Standart bir sözlükte veri ile ilgili olarak mahremiyet "yetkisiz kişiler tarafından özgürlüğün zorlanması" olarak tanımlanır [2]. Gizlilik sağlanarak, veri madenciliği bu tanıma bir anlayış getirir. Eğer kullanıcılara veri madenciliğine özgü bir işlem yapmaları için veriyi kullanmalarına izin verilirse, mahremiyet sorunu kalmaz. Eğer kullanıcının izni yoksa bu durumda kişisel gizlilik problemi oluşur.

Çoğu gizlilik yasaları arasındaki yaygın bir standart (Avrupa Gizlilik yasaları [4] veya Amerikan Sağlık yasaları [5]) "mahremiyet yalnızca tek tek tanımlanabilir veriler için uygulanabilir" şeklindedir. İzinsiz girme ve tek tek tanımlanabilir olma birleştirildiğinde, kişisel gizlilik korunarak veri madenciliği yapma ile ilgili standart belirlenir. Bu standartta, kişisel gizliliği koruyan veri madenciliği tekniği herhangi bir bilginin açığa çıkmasını engellemelidir. Buna göre;

1. Bir birey araştırılamaz
2. Zorla giriş yapılamaz.

Her iki madde için formal tanımlar açık bir meydan okumadır. Bir uçta, bu kritere göre belirli bir birey hakkında tam doğru verinin verilmediği kabul edilirken, başka bir uç noktada ise ikinci madde veri madenciliği yöntemlerini olası problemlere sürükler. Bireylerin yoğun grupları dikkate alındığı halde, bir grup hakkında bilginin çoğalması o gruptaki bireyler hakkındaki bilgileri de çoğaltır. Bunun anlamı her iki bilginin kazancını hesaplamak ve bireye özgü ilişkilendirme yeteneğinin ve bunların eşik değerlerini geçip geçmediğini bulmaktır.

Kişisel gizlilikle ilgili endişeler her geçen gün artmakta ve toplumsal bir tepkiye neden olmaktadır. Bu toplumsal tepkinin büyümesinden dolayı bir çok ülkede kişisel gizlilikle ilgili toplantılar ve konferanslar düzenlenmekte ve devletler çeşitli isimler altında bu sorunla baş edebilmek için ofisler kurmaktadırlar. Bu yoğun çalışmaya rağmen, kişisel gizlilik ihlali henüz engellenebilmiş değildir.

Aslında bu kadar çalışmanın çok fazla sonuç vermemesi, kişisel gizliliği tanımlayan bir yapının henüz tam olarak oluşturulamamasından kaynaklanmaktadır. Yani neler kişisel gizliliğe girer veya neler girmez, kişinin hangi bilgilerinin başkaların eline geçmesi bireyin mahremiyetini tehdit eder veya neler tehdit etmez bilinmemektedir. Günümüzde, bilgi teknolojisindeki ilerlemeler ve internet ortamında bilgi paylaşımının yaygın olması verinin

kolay bir şekilde ele geçirilmesine neden olmaktadır. Kişinin verileri ne kadar korunmaya çalışılsa da, onları ele verecek kilit bilgiler ne derece gizlense de, kişinin serbest bırakılan verileri yardımıyla başka veri tabanlarından saklanılan kilit veriler ele geçirilebilecektir. Gerçek bir örnek [6]'da verilmiştir, burada medikal kayıtlardaki hasta verileri; isim ve adres bilgileri silinerek ifşa edilmiştir. Genel olarak erişilebilir seçmen kayıtlarından, hastaların doğum tarihi, cinsiyet ve posta kodu kayıtları kullanılarak isimsiz farz edilmiş medikal kayıtlardaki hasta isim ve adres bilgileri ortaya çıkarılabilir. Bu kilit bir noktaya neden olur. Çünkü veride birey yeterli derecede tanımlama bilgisi içermiyor. Veri diğer kaynaklarla birleştirilerek bu tanımlana bilirlik etkin olmaktadır.

2.2 Gizliliğin İhlal Edildiği Örnekler

Örnek 1. Alice kendisinde bazı genetik rahatsızlıkların olduğunu düşünüyor ve kendisi hakkında bir inceleme yapmak istiyor. Alice aynı zamanda, Bob'un çeşitli hastalıkların DNA modellerini kapsayan bir veri tabanına sahip olduğunu biliyor. Alice Bob'tan onun veri tabanında araştırma yapmak istediğini belirtiyor. Bob, Alice'e kendi veri tabanında veri madenciliği yapmasına izin veriyor. Buradaki problem ise Bob'un Alice'in veri madenciliği çalışmaları yaparken kişisel kayıtlardaki hassas bilgiye erişmesini nasıl engelleyebileceğidir [3].

Örnek 2. Süpermarketler zinciri olan Bigmart şirketinin satın alma müdürü, Dettrees Paper şirketi ile alışveriş görüşmesi yapıyor. Dettrees firması Bigmart firmasından, kendi mallarını takip etmeleri ve böylece ürünlerinin miktarına göre üretim yaparak stok maliyetlerini düşürmeyi amaçladıklarını söyleyerek, müşteri satın alma veri tabanına erişmelerine izin vermelerini istiyor. Bunun karşılığında da fiyatlarında indirime gideceklerini belirtiyorlar.

Bigmart firmasının haberi olmadan, Detrees şirketi onların veri tabanında veri madenciliği yapmaya başlıyor. Onlar bu veri madenciliği işlemiyle "soğuk algınlığı ilacı alan müşterilerin, kağıt mendilde aldıklarını" buluyor. Aynı zamanda bu ilişkiyi kullanarak "yağlı süt alan müşterilerin, Green Paper firmasının kağıt mendilini de aldıklarını" keşfediyorlar. Bu iki ürün kendi stoklarında olduğundan "Detrees ürünlerinden alanlara, 50 cent'lik yağlı süt" alış veriş kampanyası başlatıyorlar. Green Paper şirketinin satışlarının ani bir şekilde kesilmesi, Bigmart firmasına verdikleri ürünlerinin fiyatlarını yükseltmelerine neden oluyor. Daha sonra Bigmart firması yetkilileri Detrees şirketi ile görüşmeye gittiklerinde, düşük fiyat vermekte isteksiz olma durumuyla karşılaşıyorlar ve Bigmart firması ticari rakiplerine karşı iş kaybetmeye başlıyor [7].

Örnek 3. Amerikanın en önemli sağlık kuruluşlarından biri olan Kiser, yanlışlıkla 858 mail gönderiyor. Yanlış kişilere gönderilen bu maillerdeki mesajlar, üyelerin ID'leri ile birlikte hastaların hastalıkları için sordukları soruları ve cevapları içeriyor (Washington Post, 10 Ağustos 2000).

Örnek 4. Sağlık ürünlerini online satan Global Healthrax firması kazaran isim, ev telefon numaraları, banka hesap numaraları ve kredi kartı bilgilerini kendi web sitelerinde gösterdiler (MSNBC, 19 Ocak 2000).

Örnek 5. Bir medikal servis, bir veri tabanı reklamı yapıyor. Eczacılığa ait pazarlama bilgilerinin yanında, 4.3 milyon insanın isimlerini ve sahip oldukları alerjilerinin isimlerini de reklam yapıyor (www.mmslists.com).

Örnek 5. Boston Üniversitesi, Framingham kalp çalışmasının bölümü olarak 50 yılı aşkın bir süredir toplanmış olan verileri satmak için özel bir şirket kurdu. Tıbbi kayıtlar ve genetik numuneler içeren 12,000'den fazla insan üzerinde derlenen bu veriler satışa çıkarıldı (New York Times, 17 Haziran 2000).

2.3. Gizlilik ve Veri Madenciliği

Veri madenciliği, çok büyük miktardaki veriden bilginin kazanılması için anlamlı bir teknoloji olarak ortaya çıkmıştır. Ancak, bu teknolojinin kişilerin mahremiyetini ihlal ettiğine dair endişeler vardır ve bu endişeler toplumsal bir tepkiye neden olmaktadır. Örneğin, Amerika savunma departmanı tarafından Amerika Senatosuna sunulan bir bildiriden sonra araştırma ve geliştirmeleri de kapsayan tüm veri madenciliği programları yasaklandı.

Bir başka sorunda aslında kurumların kendi sonuçları ile ilgilidir. Örneğin nüfus dairesi, uzun bir zamandır kişisel gizliliği ihlal edecek risk taşıyan nüfus verilerini içerin özet tablolar yayınlamaktadır. Küçük bir bölge için bu özet tablosu, bir bireyin kim olduğunu tanımlamayabilir. Fakat çocuk sayısı ve eğitim seviyesi gibi bireyler hakkında bilinen bazı bilgilerin birleştirilmesi ile bireyin tecrit edilmesi ve özel bilginin tespit edilmesi mümkün olabilir. [8]'de yapılan çalışma, kişisel bilgiyi açığa çıkartmadan özet verinin nasıl serbest bırakılacağını gösteren bir araştırmadır.

Veri madenciliği alanında çok yeni bir dal olan kişisel gizlilik, tam olarak tanımlanamamıştır. Veri madenciliği ile yapılan bir çok faydalı araştırma, bu konunun popülaritesini devam ettirmektedir. Örneğin süper market alış veriş verileri üzerine yapılan

madenciliği düşünelim. Süper marketlerin birçoğu, müşteri eğilimlerini öğrenmek için kendi veri tabanlarında madencilik yapılmasını öneriyor. Bu veriler üzerinde yapılacak veri madenciliğinin üreteceği sonuçlara göre, müşteri eğilimini öğrenecek ve mağaza planını yeniden dizayn ederek daha etkili promosyonlar oluşturabilecektir [6].

Veri madenciliğinin yukarıdaki pazarlama alanının dışında, medikal alanda da bazı hastalıkların incelenmesinde kullanıldığı bilinmektedir. Sigorta firmaları da, kendi mali çıkarlarını korumak için kaza yapma eğilimi yüksek olan müşterileri tespit etmekte veri madenciliğini kullanmaktadır. Air bus ve Boeing firmaları daha kararlı uçuş planları belirlemek için veri madenciliği yöntemlerini kullanmışlardır.

2.4. Gizliliği Koruma Tekniklerinin Sınıflandırılması

Kişisel gizliliği korunmuş veri madenciliği kavramı, yakın zamanda veri madenciliği algoritmalarından kişisel gizliliğin keşfedilmesi endişelerine karşılık önerilmiştir. Kötü niyetli veri madenciliğine karşılık, çeşitli teknikler geliştirilmiştir. Bu teknikler çeşitli kategorilerde sınıflandırılmaktadır. Bu kategorilerden bazıları; bireysel gizlilik ve kurumsal gizlilik [4], sorgu kısıtlama ve veri karıştırma [5], çıkışta gizliliği sağlama ve girişte gizliliği sağlamadır [6]. Bu oluşturulan kategoriler, benzer teknikleri içermektedir. Bu kategorilerden günümüze kadar geliştirilmiş olan teknikleri en genel haliyle kapsayan, çıkışta kişisel gizliliği sağlama ve girişte kişisel gizliliği sağlama sınıflandırmasıdır. Bu sınıflandırmada, kişisel gizliliğin birinci tipi verinin değiştirilmesidir. Böylece veri madenciliği sonuçları belirli kişisel gizliliği koruyacaktır. Bu tip kişisel gizliliğe çıkışta kişisel gizlilik (output privacy) denir. Karıştırma, Engelleme, Değiş tokuş ve Örnekleme teknikleri çıkışta kişisel gizliliği koruma sınıfına giren tekniklerdir.

Kişisel gizliliğin ikinci tipi verinin manipüle edilmesine dayanır. Örneğin Güvenli Çok Parçalı Hesaplama gibi şifreleme tabanlı tekniklerde, global veri madenciliği sonuçları hala keşfedilebiliyor iken kullanıcılara verinin yalnızca bir alt kümesine erişilmeye izin verilir.

2.4.1. Çıkışta Gizliliği Koruma Yöntemleri

2.4.2. Karıştırma

Bu yöntem ilk olarak Agrawal ve diğ [27] tarafından önerildi. Orijinal verinin karıştırılmasına dayanarak kişisel gizliliği koruyan veri madenciliği yapma tekniğidir. Veri madenciliği yapmak isteyen kullanıcıya kaynak olarak bu karıştırılmış veri kümesi verilir. Burada bireysel varlıklar için veri değerleri çarpıtılmıştır. Bu yüzden ayrı ayrı tanımlanabilir değerler açığa çıkmaz. Bir anket örneği verilecek olursa; bir şirket özel veri değerlerinden oluşan bir anketten veri madenciliği yapmak istiyor. Bu veri değerlerine sahip olanlar bu verileri

direkt olarak vermekte pek gönüllü değil iken, çarpıtılmış (karıştırılmış) veriyi sunmakta gönüllü olabilirler.

Eğer bir özellik lineer bir değişim gösteriyor ise bu özelliği karıştırmak için veriye belirli bir olasılık dağılımından üretilen gürültü eklenir. Örneğin, X bir özellik olsun ve bir birey X = x değerine sahip olsun. Burada x gerçek bir değerdir. r'de ortalaması 0, varyansı 1 olan normal bir dağılımdan rasgele üretilen bir sayı olsun. x'in açığa çıkması yerine, onun x + r değeri açığa çıkar. Aslında çok daha karmaşık yöntemler tasarlanabilir. Warner [9] inceleme makalesinde gizlilik hassasiyeti ile baş edebilmek için rasgele tepki yöntemini önerdi. Örnek olarak; evet hayır değerlerini alan bir Y niteliği kabul edilsin. Örnekte açığa çıkması istenmeyen bu nitelik aynı zamanda özel ve bireyseldir. Y'yi inceleyen veya incelemeyen doğrudan sorular yerine aşağıdaki iki soru önerilmiştir :

1. Y niteliğine sahibim
2. Y niteliğine sahip değilim

Birey sonra hangi sorunun cevaplanacağının kararını vermek için bir rasgele hale getirme yöntemini kullanır. Birincisi θ olasılık ile seçilir ve ikincisi $1-\theta$ olasılık ile seçilir. Açıkçası Y değeri karıştırılmış bir değer olarak elde edilir ve gerçek değer veya gizli değer korunmuştur. [10]'da kişisel gizlilik korunarak karar ağaçlarının inşası için bu teknik kullanılmıştır. [11]'de market sepeti birliktelik kurallarının madenciliğin de, kişisel gizliliğin korunması için daha karmaşık bir yöntem önerilmiştir. Zhu and Liu [12]' de gürültü eklenmesi için yöntem önermişlerdir. Çünkü rasgele hale getirme, çoğu karıştırma metotlarının önemli bir parçasıdır.

Rasgele hale getirme ve gürültü verisi kişisel gizliliği korur fakat veri madenciliğine meydan okuyan bir duruş sergiler. Burada iki can alıcı sorun vardır. Bunlar; rasgele hale getirilmiş veride nasıl madencilik yapılacağı ve rasgele hale getirilmiş veriye dayanan sonuçların orijinal veriden alınan olası sonuçlar ile düzgün bir şekilde nasıl karşılaştırılacağıdır. Veri yeterli olduğunda, rasgele hale getirme yöntemi tam olarak bilinmese dahi birçok kümelenmiş veri yeterli doğrulukla hala çıkarılabilir. Rasgele hale getirme yöntemi bilindiğinde genellikle en uygun sonuçları elde edebilen olası veri madenciliği araçları geliştirilebilir. Burada rasgele hale getirilmeden dolayı bazı bilgi ve verimliliğin kaybedilebileceği veya ödün verilebileceği anlaşılır. Çoğu uygulamada veri madenciliğinin görevi genellikle sınırlı alanlar ile ilgilenmektir. Bu nedenle saklanmak istenen bilgi saklanırken, aynı zamanda rasgele hale getirmekle gereksiz bilgide saklanır.

Farklı veri madenciliği işlemleri ve uygulamaları, farklı rasgele hale getirme tasarımları gerektirir. Rasgele hale getirmenin derecesi, genellikle bir veri tabanında ne kadar bir mahremiyetin korunacağı veya ne kadarlık bilginin diğerlerinin öğrenmesine izin verileceği ile ölçülür. [13]'de Kargupta ve diğ. önemli bir noktaya dikkat çekti: Gelişi güzel yapılan rasgele hale getirmeler güvenli değildir. Her ne kadar rasgelendirilmiş veri orijinal veriden oldukça farklı görünse de kötü niyetli biri, orijinal veriden örnekleri ve ilişkileri kendi çıkarında kullanarak yaklaşık olarak onların değerlerini bulabilir. Örneğin, bir özellik içeren ve bu özellikteki değerler sabit olan bir veri kabul edilsin. Gizlilik ihlalini amaçlayan bir kişi rasgele hale getirilmiş verilere bakıp analiz ederek bu gerçeği kolaylıkla fark eder. Orijinal veri noktaları yüksek ardışık korelasyonlar gösteriyorsa ve hatta belirleyici örneklemeler içeriyorsa ve nitelikler çok yüksek derecede ilişkililerse benzer durumlar oluşur. [14]'de Hungel ve arkadaşları bu problemi de keşfettiler ve veri kolerasyonlarına dayanan iki farklı veri rekontrüksiyon metodu önerdiler. Bunlar; a Principal Component Analysis (PCA) tekniği ve a Bayes Estimate (BE) tekniğidir.

2.4.3 Engelleme

Bu yöntem ilk olarak Y.Saygın ve diğ. [30] tarafından önerildi. Bu tip gizlilik sağlama yöntemi, veri tabanından çıkarılabilecek bağıntı kuralları içerisinde hassas, çıkarılması tehlikeli olabilecek olan bağıntı kurallarının çıkarılmasının engellenmesidir. Bunun için bu kuralın çıkarılmasına neden olan nesnelerin destek değerleriyle oynanır. Bu yöntemde, destek değerleri iki şekilde değiştirilmektedir. Bunlardan birincisi, kuralın sol tarafındaki nesnenin destek değeri arttırılır. Bu yönteme ISL(**I**ncrease **S**upport of **L**HS) denilmektedir [3,30]. İkinci yöntem ise kuralın sağ tarafındaki nesnenin destek değerini düşürmektir. Bu yönteme de, DSR (**D**ecrease **S**upport of **R**HS) denilmektedir [3,30]. Her iki yöntemde, kuralın gizlenmesi için kuralı oluşturan nesnelerin değerleri (1-0) gibi keskin değerlerle veya değeri bilinmeyen (?) gibi bir değişkenle değiştirilir.

2.4.4. Veri Değiş Tokuşu

Bu yöntem ilk olarak Dalenius ve Reis [15] tarafından önerilmiştir. Veri değiş tokuşu bireysel veri kaydı seviyesindeki çalışmalarda açığa çıkarımları sınıflandırmak için öne sürülen bir tekniktir. Gizli kalması gereken bilginin muhafazası için veri tabanından seçilmiş kayıt çiftleri arasından yine seçilmiş alanlardaki değerlerin yer değiştirilmesi ile yapılmaktadır. Veri tabanındaki bir kaydın veya bir bireyin belirli bir özelliğini, veri tabanına giren kötü niyetli bir kullanıcı için onu bulmayı olanaksız kılar. Aşağıdaki örnek incelendiğinde bu yöntemin amacı ve şekli daha iyi anlaşılacaktır. Bu örnekte gelir verisi korunacak olan veridir.

Örnek: Aşağıda 6 kişinin yaş ve gelir bilgisini içeren bir mikro veri dosyası bulunmaktadır. Kayıtların kime ait olduğunu korumak için, gelir bilgisi kayıtlar arasında rasgele değiş-tokuş edilir. 1.Kayıt ile 6. Kayıt, 2.Kayıt ile 3. Kayıt ve 4. Kayıt ile 5. Kayıt gelir bilgileri kendi aralarında değiş tokuş edilir.

Tablo 2.1 Veri değiş tokuşunda kullanılan veri kümesi

	Orijinal Kayıtlar			**Veri Değiş Tokuşundan Sonra**	
#	**Yaş**	**Gelir**	**#**	**Yaş**	**Gelir**
1	21	20.000	**1**	21	15.000
2	24	30.000	**2**	24	30.000
3	35	30.000	**3**	35	30.000
4	36	25.000	**4**	36	55.000
5	45	55.000	**5**	45	25.000
6	50	15.000	**6**	50	20.000

2. ve 3. kayıt değerleri değiştirilmemiştir. Çünkü bu kayıt değerleri için değiş tokuş herhangi bir avantaj sağlamaz.

Veri – değiş tokuşun avantajları:

1- Veri değiş tokuşu her bir kayıt için doğru bilgiyi maskeler.
2- Bu yöntem oldukça basittir ve mikro veri dosyasından başka bir şey gerektirmez ve uygulama için rasgele bir sayı üretilir.
3- Veri değiş–tokuşu hassas olmayan veya tanımlayıcı özelliği olmayan alanları bozmaksızın bir veya daha fazla değişken üzerinde kullanılabilir.

Veri – değiş tokuşunun dezavantajları:

1- Büyük sayıda kayıtların olduğu bir mikro data dosyasında gelişi güzel yapılan kategorik değişiklikler nadir kombinasyon oluşumuna neden olur. Örneğin; bir tezgahtarın gelirinin bir beyin cerrahı ile yer değiştirmesi veya çok sayıda erkek sekreter ortaya çıkması.
2- Veri–değiş tokuşunun popülasyonun tamamı üzerinde yapılacak bir araştırmaya etkisi olmaz. Ama bir alt alan incelenmek istendiğinde bu incelemeyi etkileyebilir.

2.4.5 Modelleme

Liew ve diğ. [16] tarafından tanımlanan bu yöntem güvenli bir veri tabanında gizli kalması gereken bir özelliğin korunması için olasılık dağılımını kullanır. Bu yöntem sayısal ve kategorik özelliklerin her ikisine birden uygulanabilir. Ancak birbirine bağlı birçok özelliğin genellenmesine uygun değildir. Bu yöntem üç adımdan oluşur:

1. Adım: Nitelik değerlerinin temel olası yoğunluk fonksiyonunu belirler ve bu fonksiyonun parametreleri belirlenir.

2. Adım: Gizli kalması gereken bu niteliğin belirlenen olası yoğunluk fonksiyonundan verinin örnek bir modeli oluşturulur.

3. Adım: Orijinal veri için üretilen yedek veri, benzer aralıkta bir dizidir. Örneğin; bu yeni modeldeki en küçük veri orijinal verideki en küçük veri ile yer değiştirir. Veri tabanı küçük olduğu zaman bu yöntem ile tanıtılan gürültü büyük olur böylece daha iyi bir güvenlik sağlanmış olur. Veri tabanı büyüdükçe gizli kalması gereken özelliğin güvenliğinin sağlanması düşer.

2.4.6. Girişte Gizliliği Koruma Yöntemleri

2.4.7. Güvenli Çok Parçalı Hesaplama

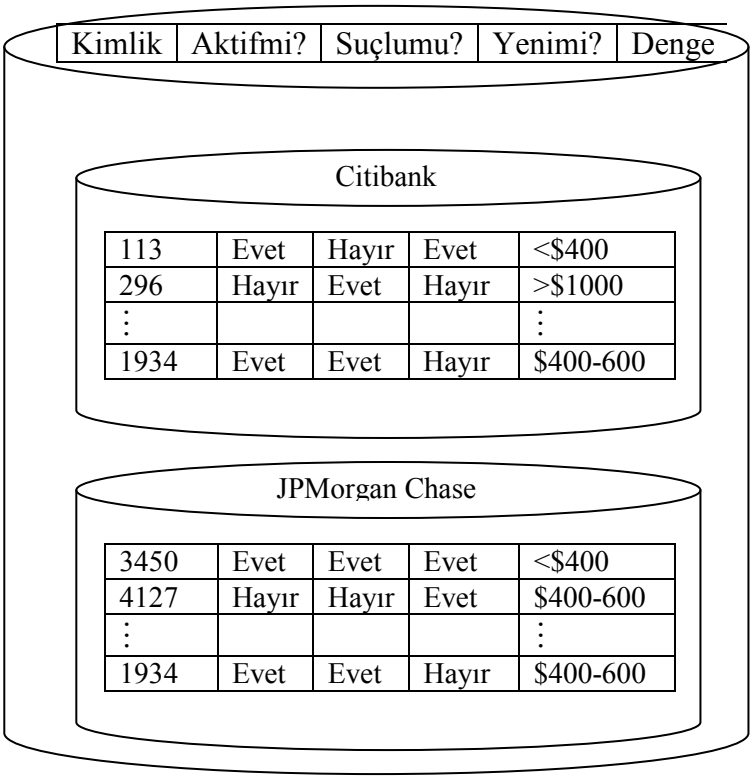
Farklı ilaç şirketleri veri madenciliği kullanarak yeni ilaçlar keşfetmek için medikal araştırmalarda kullandıkları veri tabanlarını birleştirmek istedikleri varsayılsın. Her bir şirket kendi veri tabanına erişebildiği halde kişisel gizlilik endişelerinden ve düzenlemelerinden dolayı diğer ilaç şirketlerinin verisine erişemiyor. Hatta daha fazlası ticari sebeplerden dolayı bazı kesin istatistiklerin açığa çıkmasını da istemeyebilirler. Kişisel gizliliği koruyarak veri madenciliği bu problemi adreslemeye çalışır. Amaç kendi sonuçlarından başka herhangi bir şey açığa çıkmadan veri madenciliği sonuçlarını öğrenmektir. Bu soruna yönelik çözümler dağıtık veri madenciliği olarak ele alınır ve sadece kişisel gizliliği korumayı amaçlamaz. Aynı zamanda amaçlanan sonuçtan başka her türlü bilginin sızdırılmasını da önler.

Dağıtık veri madenciliği olarak adlandırılan veri tabanlarından bilginin keşfi çok umut verici ve önemli bir araştırma alanıdır. Dağıtık veri madenciliği veri tabanı parçalara ayrılarak yapılır. Burada veri tabanı iki şekilde parçalanır: Yatay bölümleme, dikey bölümleme [17-19]. Bu modeller biçimsel olarak aşağıdaki şekilde tanımlanır.

Bireyler ve bu bireylere ait özelliklerin toplandığı bir veri kümesi tanımlansın. Bu durumda $D \equiv (E,I)$ dir. Burada E, bilgileri depolanan varlıkların kümesi; I, ise özellikler kümesi dir. k tanede farklı alanın olduğu kabul edilirse $P_1, \ldots, P_k$ sırasıyla $D_1 \equiv (E_1,I_1), \ldots D_k \equiv (E_k,I_k)$ veri kümelerini barındırır.

Verinin yatay bölümlendiği durumda farklı alanlar farklı bireylerin benzer özelliklerini toplar. Bu yüzden yatay bölümleme $E_G = \bigcup_i E_i = E_1 \cup \ldots \cup E_k$ ve $I_G = \bigcap_i = I_1 \cap \ldots \cap I_k$ ile temsil

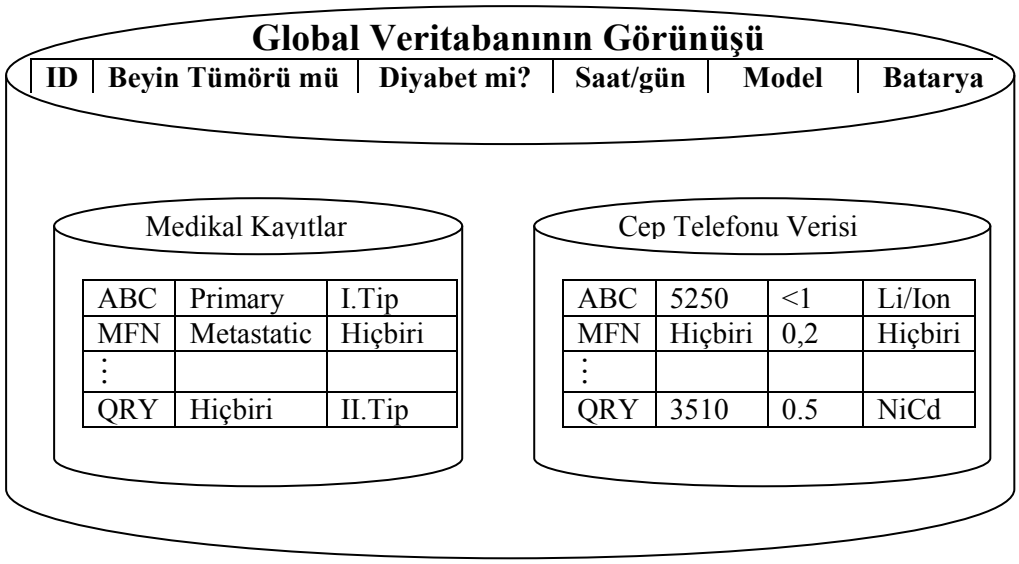
edilir. Gerçek hayatta böyle birçok durum mevcuttur. Örneğin bankalar çok benzer bilgiyi saklarlar. Bununla birlikte her bir banka için temel müşteri eğilimi oldukça farklıdır. Şekil 2.1 verinin yatay bölümlenmesini göstermektedir. Bu şekilde iki banka gösterilmektedir. Citibank ve JP Morgan Chase, her biri kendi müşterilerinin kredi kart bilgilerini topluyor. Her ikisi tarafından hesap bakiyesi ya da hesabın yeni olup olmaması, aktif olması ve borcu olup olmadığı gibi özellikler tutuluyor. Bu iki veri tabanın birleştirilmesi, dolandırıcılık gibi aktivelerin belirlenmesi için daha doğru kestirimci modellerin geliştirilmesine yol gösterir.



Şekil 2.1 Yatay bölümlenmiş veri tabanı

Diğer taraftan verinin dikey bölümlenmesi ile, benzer veri kümelerinin farklı niteliklerini bulunduran farklı alanlar anlaşılır. Örneğin; Ford binek araçlarının yapımı hakkında bilgi toplar. Fire Stone firması da tekerlek yapımı hakkında bilgi toplar. Binek araçları ile lastikler arasında bağıntı kurulabilir. Bu bağlantı bilgisi veri tabanlarına girilerek kullanılabilir. Genel veri tabanı daha sonra faydalı bilgilerin çıkarılmasında kullanılabilir. Şekil 2.2 verinin dikey bölümlenmesini göstermektedir. Birincisinde beyin tümörü ve diyabet çeşitleri gibi medikal kayıtları tutan hayali bir hastane/sigorta şirketi olduğu kabul edilmiştir. (şartları sağlamıyorsa çıkış boştur). Diğer taraftan, bir telsiz sağlayıcısı günlük kullanılan yaklaşık airtime miktarı, cep telefonu markaları ve kullanılan batarya tipi ile ilgili bilgiyi toplamaktadır. Ortak müşteriler için bu bilgiler birleştirildiğinde ve madencilik algoritmaları çalıştırıldığında

tam anlamıyla beklenmedik ilişkiler bulunabilir. Örneğin; her gün bir saatten fazla li/ion bataryalı cep telefonu kullanan birinci tip diyabet hastası bir kişi muhtemelen beyin tümörüne uğramaktadır. Buna karşılık her bir veri tabanı birbirinden izole edilmiş olsaydı, böyle bir bilginin elde edilmesi imkansız olurdu.



Şekil 2.2 Dikey Bölümlenmiş veri tabanı

3. BULANIK BAĞINTI KURAL MADENCİLİĞİ

3.1. Bağıntı Kuralları

Bir alış–veriş sırasında veya birbirini izleyen alışverişlerde müşterilerin hangi mal veya hizmetleri satın almaya eğilimli olduğunun belirlenmesi, müşteriye daha fazla ürün satın almasını sağlayan yollardan birisidir. Satın alma eğilimlerinin tanımlanmasını sağlayan bağıntı kuralları ve sıralı örüntüler, pazarlama amaçlı olarak pazar sepeti analizi adı altında veri madenciliğini kullanmaktadır. Bununla birlikte bu teknikler, tıp, finans ve farklı alanların birbiri ile ilişkili olduğunun belirlenmesi sonucunda değerli bilgi kazanımının söz konusu olduğu durumlarda da önem taşımaktadır.

Bağıntı kuralları aşağıda sunulan örneklerde görüldüğü gibi eş zamanlı olarak gerçekleşen ilişkilerin tanımlanmasında kullanılır [20].

- Müşteriler bira satın aldığında, %75 ihtimalle patates cipside alırlar.
- Düşük yağlı peynir ve yağsız yoğurt alan müşteriler, %85 ihtimalle diyet süt de alırlar

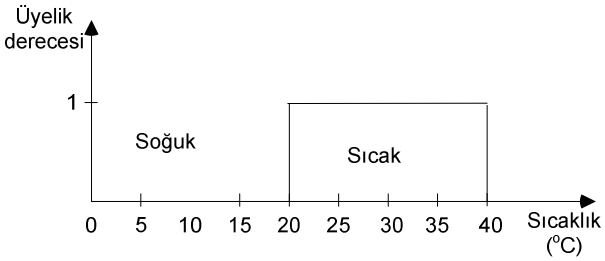
Sıralı örüntüler ise aşağıda sunulan örneklerde görüldüğü gibi birbirileri ile ilişkili olan ancak birbirlerini izleyen dönemlerde gerçekleşen ilişkilerin tanımlanmasında kullanılır [21].

- X ameliyatı yapıldığında, 15 gün içinde %45 ihtimalle Y enfeksiyonu oluşacaktır.
- İMKB endeksi düşerken A hisse senedinin değeri %15'ten fazla artacak olursa, üç iş günü içerisinde B hisse senedinin değeri %60 ihtimalle artacaktır.
- Çekiç satın alan bir müşteri, ilk üç gün içerisinde %15, bu dönemi izleyen üç ay içerisinde de %10 ihtimalle çivi satın alacaktır.

3.2. Bulanık Küme Kuramı

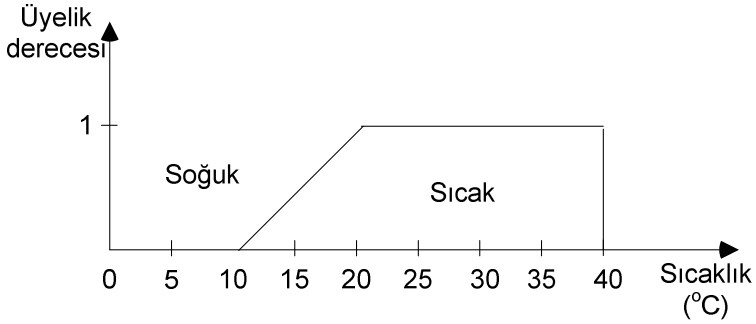
Bulanık küme teorisi ilk olarak Zadeh tarafından önerildi [22]. Klasik küme teorisinde bir eleman o kümenin ya elemanıdır ya da değildir. Hiçbir zaman kısmi üyelik olmaz. Nesnenin üyelik değeri 1 ise kümenin elemanı, 0 ise elemanı değildir. Başka bir deyişle, klasik veya yeni ürün kümelerinde elemanların üyelikleri {0,1} değerlerini alır. Bulanık mantık, insanın günlük yaşantısında nesnelere verdiği üyelik değerlerini, dolayısıyla insan davranışlarını taklit eder [23,24]. Örneğin elini suya sokan bir kişi hiçbir zaman tam olarak ısısını bilemez, onun yerine sıcak, az sıcak, soğuk, çok soğuk gibi dilsel niteleyiciler kullanır.

Klasik kümelere örnek, şekil 3.1'de verilmiştir. Eğer sıcaklık 20^{o}C'nin altına düşerse sıcak değildir. Yani klasik mantık kuramına göre 19,5 ^{o}C sıcak değildir. Doğal olarak bu mantığın hiçbir esnekliği yoktur. Gerçek dünyada ise sınırlar bu kadar keskin değildir.

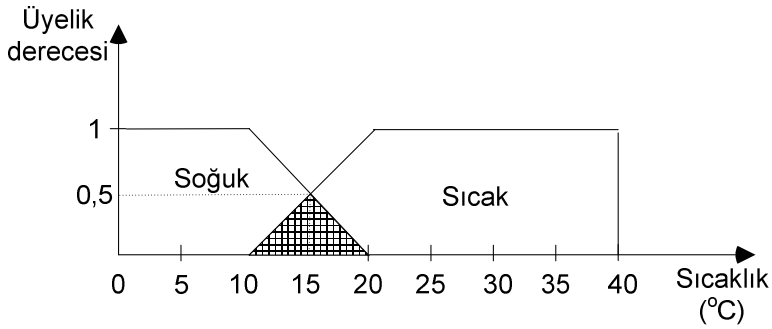


Şekil 3.1 Klasik küme grafiği

Klasik kümelerin aksine bulanık kümelerde elemanların üyelik dereceleri [0-1] aralığında sonsuz sayıda değişebilir. Bunlar üyeliğin derecelerinin devamlı ve aralıksız bütünüyle bir kümedir. Keskin kümelerdeki soğuk-sıcak, hızlı-yavaş, aydınlık-karanlık gibi ikili değişkenler, bulanık mantıkta biraz soğuk, biraz sıcak biraz karanlık gibi esnek niteleyicilerle yumuşatılarak gerçek dünyaya benzetilir. Bulanık kümeler için Şekil 3.2'de örnek verilmiştir. Burada 10–40 ^{o}C arasındaki değerler sıcak kümesine üyedirler. 20–40 ^{o}C arasındaki değerlerin üyelik dereceleri 1'dir. 10–20 ^{o}C derece arasındaki sıcaklıkların ise üyelik dereceleri 0 ile 1 değerleri arasında değişecektir. Başka bir deyişle örneğin 11 ^{o}C az sıcak, 15 ^{o}C biraz sıcak olarak değerlendirilecektir. 20 ^{o}C'yi oda sıcaklığı kabul ederek soğuk bulanık kümesi oluşturulduğunda şekil3.3 elde edilir.



Şekil 3.2 Sıcak bulanık küme grafiği



Şekil 3.3 Soğuk ve sıcak bulanık küme grafiği

3.3. Bulanık Bağıntı Kuralları

Veri madenciliği olarak adlandırılan veri tabanlarında bilginin keşfi çok umut verici ve önemli bir araştırma alanıdır. Yakın zamandaki araştırmaların çoğu ikili(1-0) değer içeren verileri ele almışlardır. Agrawal ve çalışma arkadaşları işlem verilerinden bağıntı kuralları bulmak için büyük yoğun nesne kümeleri kavramına dayanan birkaç madencilik algoritması önerdiler[25-27]. "nesne kümesi" kavramı bu veri madenciliği makalelerinde ilk olarak yer aldı. Bundan sonra bu kavram ortak bir kullanım haline geldi. Bu kelimenin anlamı parçalardan oluşmuş bir kümedir. [25]'de Agrawal ve diğerleri iki yeni yöntem Apriori ve AprioriTid algoritmalarını önerdiler. Bu iki algoritma birbirine çok benzemesine rağmen arasındaki farklılık şudur. Apriori algoritması önceki geçişte bulunan yoğun nesne kümelerini kullanarak yani aday nesneleri üretir. Burada önemli olan, bir yoğun nesne kümesinin tüm alt kümeleri de yoğun olmalıdır. Apriori TID algoritmasının, Apriori algoritmasından farkı veri tabanının ilk geçişten sonra aday nesnelerin desteğini hesaplamak için bir daha kullanılmamasıdır. Bunun yerine, önceki geçişte kullanılan aday nesnelerin kodlamış hali kullanılır. Daha sonraki geçişlerde, bu kodlamanın boyutu veri tabanından çok daha düşük olur. Böylece çok daha hızlı bir çalışma yöntemi elde edilir. [25]'de aynı yazarlar madencilik işlemlerini iki evreye ayırdılar. Birinci evrede aday nesne kümeleri üretilir ve işlem verileri taranarak sayılır. İşlem verilerinde bulunan nesne kümelerinin sayısı kullanıcının önceden tanımladığı eşik değerinden (minimum destek) yüksek ise bu nesne kümeleri yoğun nesne kümeleri olarak düşünülür. Daha sonra bu yoğun nesne kümelerinden ikili nesnelerden oluşan aday nesne kümeleri bulunur. Daha sonra tüm yoğun nesne kümeleri bulunana kadar bu işlemler devam eder. İkinci evrede, birinci evrede bulunan yoğun nesne kümelerinden bağıntı kurallarının çıkarılması işlemleri yapılır. Her bir yoğun nesne kümesi için bağıntı kurallarının tüm olası kombinasyon biçimleri alınır ve kullanıcının önceden tanımladığı eşik değerinden (minimum güven) büyük olan güven değerli yoğun nesne kümeleri istenen bağıntı kuralları olarak çıkışa verilir.

Yukarıda anlatılan yöntemler ikili [0-1] değerleri için önerilmiştir. Gerçek dünyadaki uygulamalarda işlem verileri genelde nicel ve kategorik değerlerden oluşmuştur. İkili değerli işlem verilerinden bağıntı kuralları üretmek için önerilen yöntemlere ek olarak. Srikant ve Agrawal [28] aynı zamanda nicel ve kategorik niteliklerden bağıntı kural madenciliği üretmek içinde bir yöntem önerdiler. Onların önerdiği yöntemde öncelikle her bir nitelik için kullanılacak küme sayısı belirlenir. Daha sonra her bir niteliğin alabileceği bütün değerler, elemanları ardışıl olan bir tamsayılar kümesine eşleştirilir. Kullanıcının önceden tanımladığı destek değerinden büyük destek değerine sahip olan yoğun nesne kümeleri bulunur. Bu yoğun nesne kümeleri daha sonra birliktelik kurallarını üretmek için kullanılır ve kullanıcılar için ilginç olan kurallar çıkarılır.

Veri madenciliğinde nicel değerli veriler üzerinde Srikant ve Agrawalın [28] önerdiği yöntemin yanında nicel değerli verilerden oluşan veri tabanlarında bağıntı kurallarının çıkarılmasında bulanık küme kavramı yöntemi de kullanılmaktadır. Bulanık küme teorisinin insan muhakemesine yakınlığı ve basitliği onun yoğun bir şekilde akıllı sistemlerde kullanılmasına neden olmaktadır. Eldeki veri kümesine uyan kuralları öğrenen birçok bulanık öğrenme algoritmaları tasarlanmış ve belirli alanlar için etkili katkılarda kullanılmıştır [22-24]. Veri madenciliğin de bulanık kümelerin kullanılması yakın zamanlarda geliştirilmiştir. [29]'da Hong ve diğerleri nicel değerli işlem verilerinden bulanık bağıntı kuralları ve ilginç nesne kümelerinin bulunması için [25]'deki Apriori madencilik algoritması ile bulanık küme kavramlarını birleştirerek bir madencilik yaklaşımı önerdiler. [23]'de yine [20]'de tanıtılan bir başka yaklaşım olan AprioriTid algoritmasına dayanan yeni bir bulanık madencilik algoritması önerdiler.

4. NİCEL DEĞERLİ VERİ TABANINDAN BULANIK BAĞINTI KURALLARININ GİZLENMESİ

Yakın zamandaki bağıntı kurallarının gizlenmesi çalışmalarının çoğu ikili (0-1) değerlerini taşıyan veriler üzerinde yapılmıştır. Gerçek hayatta ise veriler genellikle nicel değerler içermektedir. Örneğin bir hastadaki kan örneğinde kan içerisinde birçok nitelik bulunabilir. Çoğu hastalıkta bir niteliğin bulunmasından ziyade o niteliğin kandaki değerinin ne kadar olduğuna göre teşhis konmakta ve ona göre tedavi uygulanmaktadır. Örneğin her insanda kolesterol vardır, bu durum kişinin hasta olduğu veya olmadığı anlamına gelmez. Kolesterol değerinin çokluğuna göre bir teşhis konabilir.

Yukarıda anlatılan veri madenciliği yöntemlerinde ister ikili (1-0) değerlerini baz alan yöntemler olsun isterse de nicel değerleri göz önüne alan yöntemler olsun hepsinde kişisel gizliliği tehlikeye atabilecek bağıntı kurallarının çıkması muhtemeldir. Özellikle medikal alanda hastalarla ilgili çok geniş detayların bulunduğu veri tabanları bulunmaktadır. Bu veri tabanlarının kötü niyetli kullanımı hastaların kişisel gizliliğini tehlike altına sokmaktadır. Yakın zamanda bu veri tabanları ifşa edilmiştir. İfşa edilen veri tabanları ile ilgili örnekler bölüm 2.2'de verilmiştir.

Bu tezde [25]'de önerilen Apriori algoritması, [30]'da bağıntı kurallarının çıkarılmasını engellemek için önerilen ISL yöntemi ve [32]'te Tzung-Pei Hong ve diğerlerinin önerdiği nicel verilerin bulanıklaştırılması yöntemlerini birleştirip bunları bir adım daha öne götüren nicel verilerden hassas bağıntı kurallarının çıkarılmasını engellemek için yeni bir yöntem geliştirilmiştir.

4.1. Önerilen Algoritma

Nicel değerlere sahip bir veri tabanında $X \Rightarrow Y$ gibi bir kuralı gizlemek için o kuralın güven değeri minimum güven değerinin altına düşürülür. Güven değerini minimum güven değerinin altına düşürmek için kuraldaki nesnelerin destek değerleri ile oynanır. Kuralı gizlemek için ya kuralın sol tarafındaki X nesnenin destek değeri arttırılır yada kuralın sağ tarafındaki Y nesnenin destek değeri minimum destek değerinin altına düşürülür.

Bu tezde kuralın güven değerini düşürmek için kuralın sol tarafındaki nesnenin destek değeri arttırılmaktadır. Bunu yapmak içinde sol taraftaki nesnenin 0,5 ten ve sağ taraftaki nesnenin değerinden küçük olduğu işlem verilerinde sol tarafın değeri 1'den çıkarılmaktadır.

Algoritmada kullanılan kısaltmalar aşağıdaki gibidir.

D	: Orijinal veri tabanı.
F	: Veri tabanını bulanıklaştırılmış hali .
X	: Nesnelerin bir kümesi,
T_L	: Sol taraftaki nesneye ait işlem verileri
T_R	: Sağ taraftaki nesneye ait işlem verileri
U	: Kural,
f	: F veri tabanındaki nesneleri ifade etmektedir.
α	: Kullanıcı tanımlı minimum destek değeri
λ	: Kullanıcı tanımlı minimum güven değeri

Giriş Verileri

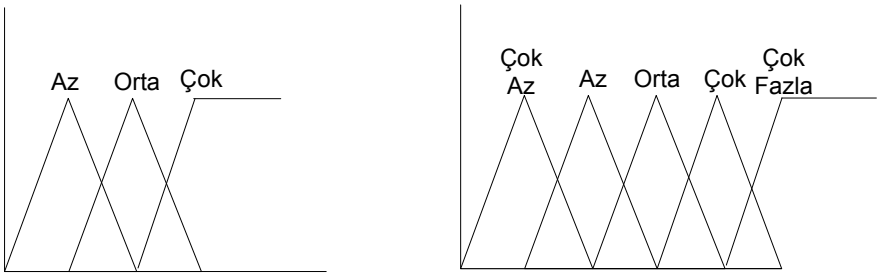
a) Kaynak veri tabanı
b) α minimum destek değeri
c) λ minimum güven değeri

Çıkış

1 Veri tabanı bulanıklaştır D → F
2. Bulanık F veri tabanındaki f ∈ F'deki her bir itemın destek değerini hesapla
3. Eğer tüm f(support)< min_support;EXIT:. // herhangibir kural yoktur.
4. F'deki Large 2_itemsetleri bul.
5. for each X'in içerdiği her bir Large 2 itemsetleri için
6. **U kuralın destek değerini hesapla → min(U_L , U_R)**
7. **eğer kural U(support) > min_support**
8. Kural U (confidence) hesapla
9. **if kural U(confidence) > min_confidence**
10. **for each T_L of rule**
11. **if T_L <0.5 and T_L < T_R**
12. **T_L = 1 - T_L**
13. **end // if**
14. **end // for each**
15. **end // for each**
16. **Kural U'nun confidencesini tekrar hesapla**
17. **if kural U(confidence) > min_confidence**
18. **remove all T_L and eski haline dön**
19. **end // if**
20. **end // for each**
21. **F → veri tabanını durulandır. F →D**
22. **end.**

4.2. Yöntemin Aşamaları

1.Adım Önce veri tabanı bulanıklaştırılarak dilsel terimler şekline getirilir. Burada bulanıklaştırma işlemi yapılırken değer aralıkları kullanıcı tarafından belirlenir. Örneğin;



Şekil 4.1 Bulanık Küme örnekleri

2.Adım Bulanıklaştırma işleminden sonra kullanıcı destek eşik değerini geçen nesneler alınır, geri kalanlar elenir.

3.Adım Nesnelerin 2'li birlikteliklerine bakılır. İkili birlikteliklere bakarken VE işlemi kullanılır. Buradaki VE işleminde destek değerini belirlemede küçük olanın değeri alınır. Örneğin;

	Ao	Bo
I_2	0,3	0,5

Burada destek değeri 0,3 olarak alınır.

4.Adım Varsa 3'lü birlikteliklere bakılır.

5.Adım Kullanıcı tanımlı güven eşik değerine göre kurallar çıkarılır. Güven formülü

$$\text{Güven} = \frac{\text{Destek ikili}}{\text{Destek tekil}}$$

$$\text{Güven(AoBo)} = \frac{\text{Destek(AoBo)}}{\text{Destek(Ao)}} \quad \text{ile belirlenir}$$

6.Adım Hassas kurallar belirlenir. Bu kurallar belirlendikten sonra kuralların gizlenmesi için güven değerinin düşürülmesine çalışılır. Bunu yaparken ikili nesnelerin destek değeri düşürülür veya soldaki nesnenin destek değeri yükseltilir.

$$\text{Güve(AoBo)} = \frac{\text{Destek(AoBo)} \downarrow}{\text{Destek(Ao)} \uparrow}$$

Bu destek değerlerinin düşürülmesi veya arttırılması işlemleri aşağıdaki örnekle açıklanmıştır.

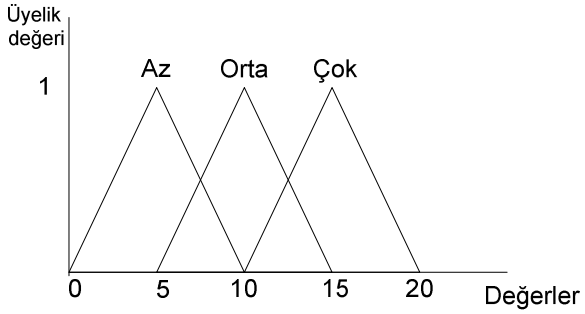
Örnek:

1.Adım. Eşik destek değeri 2,2 ve güven değeri %70 olarak alındı.

Tablo 4.1 Nesne Kümesi

	A	B	C	D
I_1	10	5	8	3
I_2	3	11	6	14
I_3	6	3	9	13
I_4	7	5	8	12
I_5	11	4	7	10

2.Adım. Tablo 4.1'deki veri tabanı bulanıklaştırma işlemine sokulur. Bu çalışmada üçgen üyelik fonksiyonu tercih edilmiştir.



Şekil 4.2. Örnekte kullanılan üyelik fonksiyonu

Tablo 4.2. Tablo 4.1'deki verilerin bulanık kümelere dönüştürülmesi

Kayıtlar	**A**			**B**			**C**			**D**		
	Az	**Ao**	**Ab**	**Bz**	**Bo**	**Bb**	**Cz**	**Co**	**Cb**	**Dz**	**Do**	**Db**
I_1	0	1	0	1	0	0	0,4	0,6	0	0,6	0	0
I_2	0,6	0	0	0	0,8	0,2	0,8	0,2	0	0	0,2	0,8
I_3	0,8	0,2	0	0,6	0	0	0,2	0,8	0	0	0,4	0,6
I_4	0,6	0,4	0	1	0	0	0,4	0,6	0	0	0,6	0,4
I_5	0	0,8	0,2	0,8	0	0	0,6	0,4	0	0	1	0
Toplam Destek	2,0	**2,4**	0,2	**3,4**	0,8	0,2	**2,4**	**2,6**	0	0,6	**2,2**	1,8

3.Adım. 2li birlikteliklerin ve bu birlikteliklerin destek değerlerinin belirlenmesi

Tablo4 3. Tablo 4.2'deki bulanık değerlerden ikili birlikteliklerin bulunması

Kayıtlar	Ao	Bz	Destek
I_1	1	1	**1**
I_2	0	0	**0**
I_3	0,2	0,6	**0,2**
I_4	0,4	1	**0,4**
I_5	0,8	0,8	**0,8**
Toplam			**2,4**

Ao	Cz	Destek
1	0,4	0,4
0	0,8	0
0,2	0,2	0,2
0,4	0,4	0,4
0,8	0,6	0,6
		1,4

Ao	Co	Destek
1	0,6	0,6
0	0,2	0
0,2	0,8	0,2
0,4	0,6	0,4
0,8	0,4	0,4
		1,6

Kayıtlar	Ao	Do	Destek
I_1	1	0	0
I_2	0	0,2	0,2
I_3	0,2	0,4	0,2
I_4	0,4	0,6	0,4
I_5	0,8	1	0,8
Toplam			1,6

Bz	Cz	Destek
1	0,4	**0,4**
0	0,8	**0**
0,6	0,2	**0,2**
1	0,4	**0,4**
0,8	0,6	**0,6**
		1,6

Bz	Co	Destek
1	0,6	0,6
0	0,2	0
0,6	0,8	0,6
1	0,6	0,6
0,8	0,4	0,4
		2,2

Kayıtlar	Bz	Do	Destek
I_1	1	0	0
I_2	0	0,2	0
I_3	0,6	0,4	0,4
I_4	1	0,6	0,6
I_5	0,8	1	0,8
Toplam			1,8

Cz	Do	Destek
0,4	0	**0**
0,8	0,2	**0,2**
0,2	0,4	**0,2**
0,4	0,6	**0,4**
0,6	1	**0,6**
		1,4

Co	Do	Destek
0,6	0	**0**
0,2	0,2	**0,2**
0,8	0,4	**0,4**
0,6	0,6	**0,6**
0,4	1	**0,4**
		1,6

Burada ($Ao \Rightarrow Bz$) kuralının hassas kural olduğu kabul edilirse, amaç Güven(AoBz) güven değerini düşürmektir. Bunun içinde ya Destek(AoBz)'yi düşürülür yada Destek(Ao)'nın değeri arttırılır. Destek(Ao)'nun değerini arttırmak için ;

4. Adım. Ao'nun Bz'den küçük olduğu kayıtlarda Ao 1'den çıkarılır.

$$\text{Güven(AoBz)}=\frac{\text{Destek(AoBz)}}{\text{Destek(Ao)}}=\frac{2,4}{2,4}=1=\%100$$

	Ao	Bz	Destek
I_1	1	1	**1**
I_2	0	0	**0**
I_3	0,2	0,6	**0,2**
I_4	0,4	1	**0,4**
I_5	0,8	0,8	**0,8**
Toplam			**2,4**

$I_3 = 1\text{-}0,2 = 0,8$

	Ao	Bz	Destek
I_1	1	1	**1**
I_2	0	0	**0**
I_3	0,8	0,6	**0,6**
I_4	0,4	1	**0,4**
I_5	0,8	0,8	**0,8**
Toplam	3.0		**2,8**

$$\text{Güven(AoBz)}=\frac{\text{Destek(AoBz)}}{\text{Destek(Ao)}}=\frac{2,8}{3}=1=\%0,93$$

Durum 2 : Ao'nun 0'a eşit olduğu ve Bz'nin 1'den küçük olduğu işlem verilerinde Ao'nun değeri 1'den çıkarılır.

	Ao	Bz	Destek
I_1	1	1	**1**
I_2	0	0	**0**
I_3	0,8	0,6	**0,6**
I_4	0,4	1	**0,4**
I_5	0,8	0,8	**0,8**
Toplam	3.0		**2,8**

$I_2 = 1 - 0 = 1$

	Ao	Bz	Destek
I_1	1	1	**1**
I_2	1	0	**0**
I_3	0,8	0,6	**0,6**
I_4	0,4	1	**0,4**
I_5	0,8	0,8	**0,8**
Toplam	4.0		**2,8**

$$\text{Güven(AoBz)}=\frac{\text{Destek(AoBz)}}{\text{Destek(Ao)}}=\frac{2,8}{4}=0,7=\%0,70$$

5.Adım. Değiştirilmiş değerlerin bulanıklaştırılmış veri tabanında yerlerine yazılmaları

Tablo 4.4 Değiştirilmiş değerli bulanık veri tabanı

Kayıtlar	**A**			**B**			**C**			**D**		
	Az	**Ao**	**Ab**	**Bz**	**Bo**	**Bb**	**Cz**	**Co**	**Cb**	**Dz**	**Do**	**Db**
I_1	0	1	0	1	0	0	0,4	0,6	0	0,6	0	0
I_2	0,6	**1**	0	0	0,8	0,2	0,8	0,2	0	0	0,2	0,8
I_3	0,8	**0.8**	0	0,6	0	0	0,2	0,8	0	0	0,4	0,6
I_4	0,6	0,4	0	1	0	0	0,4	0,6	0	0	0,6	0,4
I_5	0	0,8	0,2	0,8	0	0	0,6	0,4	0	0	1	0
Toplam Destek	2,0	**4**	0,2	**3,4**	0,8	0,2	**2,4**	**2,6**	0	0,6	**2,2**	1,8

Yukarıdaki veri tabanında da görüldüğü gibi I_2, I_3'ün değerine bakıldığında bulanık değerlerin toplamı 1'i geçmektedir. Bu sorunu giderebilmek için bulanık destek değeri değiştirilen öğenin değeri 1'den çıkarılır kalan değer diğer bulanık mantık öğesinin yerine yazılır. Örneğin I_3 değeri için ;

Az + Ao = 0,8 + 0,8 = 1,6 burada Ao 1'den çıkarılır. 1-0,8 = 0,2 çıkan değer Az'nin yerine yazılır. Az'nin yeni değeri 0,2 olur.

Tablo 4.5. Tablo4.4'un düzeltme işleminden sonraki durumu

Kayıtlar	**A**			**B**			**C**			**D**		
	Az	**Ao**	**Ab**	**Bz**	**Bo**	**Bb**	**Cz**	**Co**	**Cb**	**Dz**	**Do**	**Db**
I_1	0	1	0	1	0	0	0,4	0,6	0	0,6	0	0
I_2	0	**1**	0	0	0,8	0,2	0,8	0,2	0	0	0,2	0,8
I_3	0,2	**0.8**	0	0,6	0	0	0,2	0,8	0	0	0,4	0,6
I_4	0,6	0,4	0	1	0	0	0,4	0,6	0	0	0,6	0,4
I_5	0	0,8	0,2	0,8	0	0	0,6	0,4	0	0	1	0
Toplam Destek	2,0	**4**	0,2	**3,4**	0,8	0,2	**2,4**	**2,6**	0	0,6	**2,2**	1,8

6.Adım. I_2'nin değeri de aynı şekilde değiştirilir Az = 0 veri tabanının düzeltme işleminden sonraki hali tablo 4.6'da gösterilmiştir

Tablo 4.6. Veri tabanının düzeltme işleminden sonraki durumu

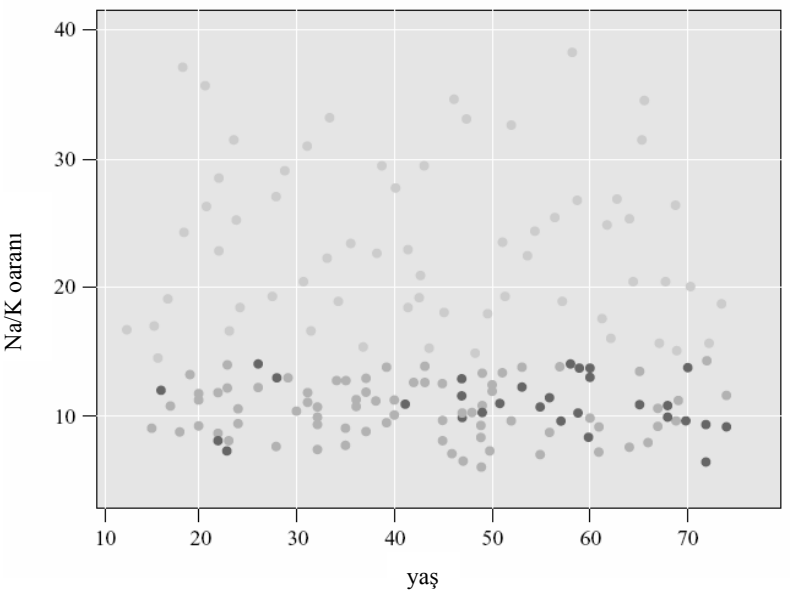
	A	B	C	D
I_1	10	5	8	3
I_2	**10**	11	6	14
I_3	**9**	3	9	13
I_4	7	5	8	12
I_5	11	4	7	10

5.SINIFLANDIRMA KURALLARININ GİZLENMESİ

Sınıflandırmada, hedef bir kategorik değer vardır. Örneğin yüksek gelir, orta gelir ve düşük gelir gibi değerler, gelir niteliğini üç sınıf veya kategori içerisinde böler. Veri madenciliği modeli kayıtların yoğun nesne kümelerini inceler. Her bir kayıt hedef verinin yanında giriş kümesi gibi bir bilgiyi tutar. Örneğin, Tablo 5.1'de gösterilen veri kümesinden bir alıntı üzerinde çalışılırsa, araştırmacı bu veri tabanındaki kişinin yaş, cinsiyet ve meslek ile ilgili karakteristiklerine dayanarak gelir düzeyinin sınıflandırılmasını yapmak isteyebilir. Bu işlem bir sınıflandırma işlemidir. Bu algoritma kabaca aşağıdaki gibi yürütülür. İlk olarak hedef değeri içeren (önceden sınıflandırılan) gelir düzey verisi ve diğer kestirimci veriler içeren veri tabanı incelenir. Bu şekilde algoritma hangi değer kombinasyonlarının hangi gelir düzeyi ile birliktelik kurduğunu öğrenir. Örneğin yaşlı kadınlar ile yüksek gelir düzeyi birlikteliği kurulabilir. Bu veri eğitim kümesi olarak adlandırılır. Daha sonra algoritma mevcut gelir düzeyi bilgisi kalmayana kadar her bir kaydı inceler. Algoritma yeni kayıtların sınıflandırmasını yapmak için eğitim verisindeki sınıflandırmayı baz alır. Örneğin 63 yaşındaki bayan profesör yüksek gelir düzeyinde sınıflandırılabilir.

Tablo 5.1 Gelir sınıflandırması için veri kümesinden seçilmiş kısım

ID	Yaş	Cinsiyet	Meslek	Gelir
001	47	F	Yazılım Müh.	Yüksek
002	28	M	Pazarlama Uzm	Orta
003	35	M	İşsiz	Düşük
⋮				



Şekil 5.1 Hangi hasta tipi için hangi ilaç yazılmalıdır.

Örneğin medikal bir alanda, hastanın yaşı, hastanın sodyum/potasyum oranı gibi belirli hasta karakteristiklerine dayanarak bir hastaya ilaç önerilebilmesi için ilaçların sınıflandırılması ile ilgili bir örnek şekil 5.1'de verilmiştir [32]. Şekilde 200 hastanın yaş, sodyum ve potasyum oranlarının dağılımı gösterilmektedir. Noktanın yoğunluğu bir ilacı sembolize etmektedir. Açık gri Y ilacının belirtisi, normal gri noktalar A veya X ilacının belirtisi, koyu gri noktalar B veya C ilacını belirtisini verir. Bu çizim SPSS tarafından yayınlanan Clementine veri madenciliği yazılımı tarafından üretilmiştir.

Bu dağıtık çizimde Na/K (sodyum/potasyum) oranı dikey eksende, yaş oranı ise yatay eksende gösterilmiştir. Hasta reçetesi bu veri kümesine dayanarak hazırlandığı kabul edilirse;

1. Yüksek sodyum potasyum oranına sahip genç bir hasta için hangi ilaç verilmelidir ?.

- Genç hastalar grafiğin sol tarafında yer almaktadır. Yüksek sodyum/potasyum oranın da grafikteki dikey eksenin üst tarafında yer almaktadır. Yüksek sodyum/potasyum oranına sahip genç hastalar için açık gri renge sahip olan Y ilacı tavsiye edilmektedir. Böyle hastalar için en çok tavsiye edilen sınıflandırma Y ilacını vermektir.

2. Düşük sodyum/potasyum oranına sahip daha yaşlı hastalar için hangi ilaç yazılabilir.

- Grafiğin sağ alt tarafında koyu gri (B ve C ilaçları) veya orta gri (A ve X ilaçları) hemen hemen aynı yoğunlukta olmasından dolayı farklı reçeteler hazırlanabilir. Daha belirleyici bilgi olmadan burada bir sınıf tanımlanamaz. Belki bu ilaçlar farklı etkileşime sahip olabilir. Örneğin; östorojen, astım veya kalp rahatsızlığı gibi.

Grafikler ve noktalar veride iki veya üç boyutlu birlikteliklerin anlaşılmasında yardımcı olur. Fakat bazen sınıflandırma farklı kestirilmelere, daha çok boyutsal çizimlere ihtiyaç duyabilir. Bu yüzden sınıflandırma işlemlerinin performansını arttırabilmek için daha karmaşık çalışmalara ihtiyaç vardır. Sınıflandırmada en çok bilinen veri madenciliği yöntemler; *k*-en yakın komşu, karar ağaçları ve yapay sinir ağlarıdır.

5.1. Bulanık Küme Kavramı ve Apriori Algoritması Kullanılarak Sınıflandırma

Nicel değerlere sahip bir veri tabanında $X\ Y \Rightarrow Z$ bir sınıflandırma kuralını gizlemek için, gizlenilmesi istenilen sınıflandırma kuralındaki nesnelerin güven değeri minimum güven değerinin altına düşürülür. Güven değerini minimum güven değerinin altına düşürmek için kuraldaki nesnelerin destek değerleri ile oynanır. Sınıflandırma kuralını gizlemek için tıpkı bağıntı kurallarında olduğu gibi ya kuralın sol tarafındaki X nesnesinin destek değerini arttırılır yada kuralın sağ tarafındaki Y nesnesinin destek değeri minimun destek değerinin altına düşürülür.

Bu tezde kuralın güven değerini düşürmek için kuralın sol tarafındaki nesnenin destek değeri arttırılmaktadır. Bunu yapmak içinde sol taraftaki nesnenin 0,5 ten ve sağ taraftaki nesnenin değerinden küçük olduğu işlem verilerinde sol tarafın değeri 1'den çıkarılır.

Algoritmada kullanılan kısaltmalar şu şekildedir.

D	: Orijinal veri tabanı.
C	: Sınıflar
F_i	: Veri tabanını bulanıklaştırılmış hali .
X	: Nesnelerin bir kümesi,
T_L	: Sol taraftaki nesneye ait işlem verileri
T_R	: Sağ taraftaki nesneye ait işlem verileri
U	: Kural,
f	: F veri tabanındaki nesneleri ifade etmektedir.
α	: Kullanıcı tanımlı minimum destek değeri
λ	: Kullanıcı tanımlı minimum güven değeri

Giriş Verileri

d) Kaynak veri tabanı
e) α minimum destek değeri
f) λ minimum güven değeri

Çıkış

1 Veri tabanı taranarak sınıflar ve sınıflara ait olan kayıtlar belirlenir.
2 Ayrılan sınıflara ait nesnelerdeki değerler bulanıklaştırılır $C_i \rightarrow F_i$
3. for bulanık F_i veri tabanındaki $f \in F_i$'deki her bir nesnenin destek değerini hesapla
4. Eğer tüm $F_i \in$ f(destek)$< \alpha$;EXIT: // herhangi bir kural yoktur.
5. F_i'deki 2'li nesne kümelerini bul.

6. for X'in içerdiği her bir 2'li nesne kümeleri için
7. **U kuralın destek değerini hesapla → min(U_L , U_R)**
8. **eğer kural U(destek) >** α
9. Kural U (güven) hesapla
10. **if kural U(güven) >** λ

5.2. Yöntemin Aşamaları

1.Adım Önce veri tabanı sınıf alanına göre parçalara ayrılır.

Tablo 5.2 Orijinal veri kümesinin sınıflara göre iki ayrı kümeye ayrılması

	A	B	C	D	Sınıf
I_1	10	5	8	3	1
I_2	3	11	6	14	1
I_3	6	3	9	13	2
I_4	7	5	8	12	2
I_5	11	4	7	10	2
I_6	8	2	4	6	1
I_7	6	7	9	15	2

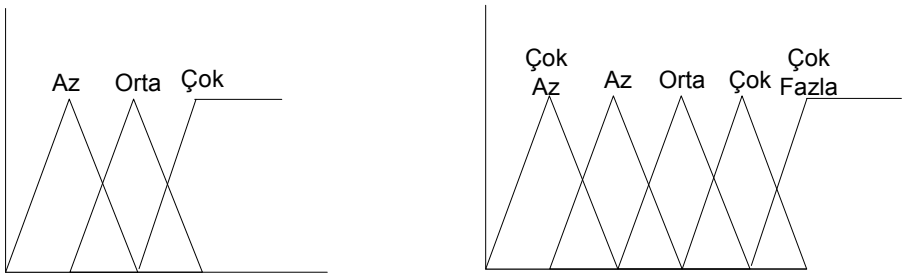
Sınıf 1 için veri tabanı değişimi

	A	B	C	D	Sınıf
I_1	10	5	8	3	1
I_2	3	11	6	14	1
I_3	6	3	9	13	0
I_4	7	5	8	12	0
I_5	11	4	7	10	0
I_6	8	2	4	6	1
I_7	6	7	9	15	0

Sınıf 2 için veri tabanı değişimi

	A	B	C	D	Sınıf
I_1	10	5	8	3	0
I_2	3	11	6	14	0
I_3	6	3	9	13	1
I_4	7	5	8	12	1
I_5	11	4	7	10	1
I_6	8	2	4	6	0
I_7	6	7	9	15	1

2.Adım Önce veri tabanı bulanıklaştırılarak dilsel terimler şekline getirilir. Burada bulanıklaştırma işlemi yapılırken değer aralıkları kullanıcı tarafından belirlenir.



Şekil 5.2 Yöntem için kullanılabilecek bulanık küme örnekleri

3.Adım. Bulanıklaştırma işleminden sonra her bir sınıf için kullanıcı destek eşik değerini geçen nesneler alınır, geri kalanlar elenir.

4.Adım. Nesnelerin 2'li birlikteliklerine bakılır. İkili birlikteliklere bakarken VE işlemi kullanılır buradaki VE işleminde destek değerini belirlemede küçük olanın değeri alınır. Örneğin;

	Ao	Bo
I_2	0,3	0,5

Burada destek değeri 0,3 olarak alınır.

5.Adım Varsa 3'lü birlikteliklere bakılır.

6.Adım. Kullanıcı güven eşik değerine göre kurallar çıkarılır. Güven formülü aşağıdaki şekilde verilmiştir.

$$\text{Güven} = \frac{\text{Destek ikili}}{\text{Destek tekil}}$$

$$\text{Güven(AoBo)} = \frac{\text{Destek(AoBo)}}{\text{Destek(Ao)}}$$

Her sınıf için hassas kurallar belirlenir. Bu kurallar belirlendikten sonra kuralların gizlenmesi için güven değerinin düşürülmesine çalışılır. Bunu yaparken ikili birlikteliklerin destek değeri düşürülür veya tekil destek değeri yükseltilir.

$$\text{Güven(AoBo)} = \frac{\text{Destek(AoBo)} \downarrow}{\text{Destek(Ao)} \uparrow}$$

Bu destek değerlerinin düşürülmesi veya arttırılması işlemleri aşağıdaki bir örnekle açıklanmıştır.

ÖRNEK

1.Adım Eşik destek değeri 2 ve güven değeri %80 olarak alındı.

Tablo 5.3 Orijinal Veri Kümesi

	A	B	C	D	Sınıf
I_1	10	5	8	3	1
I_2	3	11	6	14	1
I_3	6	3	9	13	2
I_4	7	5	8	12	2
I_5	11	4	7	13	2
I_6	8	2	4	6	1
I_7	5	7	9	15	2

2.Adım Önce veri tabanı sınıf alanına göre parçalara ayrılır.

Tablo 5.4 Örnekteki orijinal veri kümesinin sınıflara göre iki ayrı kümeye ayrılması

	A	B	C	D	Sınıf
I_1	10	5	8	3	1
I_2	3	11	6	14	1
I_3	6	3	9	13	2
I_4	7	5	8	12	2
I_5	11	4	7	10	2
I_6	8	2	4	6	1
I_7	6	7	9	15	2

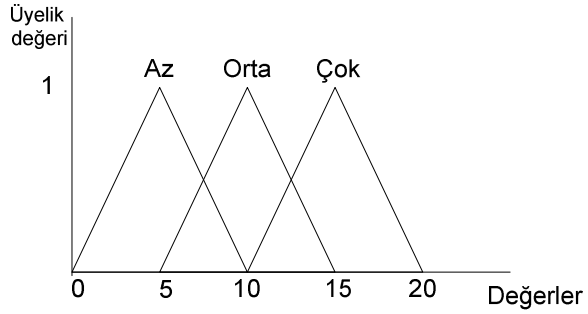
Sınıf 1 için veri tabanı değişimi

	A	B	C	D	Sınıf
I_1	10	5	8	3	1
I_2	3	11	6	14	1
I_3	6	3	9	13	0
I_4	7	5	8	12	0
I_5	11	4	7	10	0
I_6	8	2	4	6	1
I_7	6	7	9	15	0

Sınıf 2 için veri tabanı değişimi

	A	B	C	D	Sınıf
I_1	10	5	8	3	0
I_2	3	11	6	14	0
I_3	6	3	9	13	1
I_4	7	5	8	12	1
I_5	11	4	7	10	1
I_6	8	2	4	6	0
I_7	6	7	9	15	1

2.Adım Tablo 5.1'deki veri tabanının sınıflara ait işlem verileride bulanıklaştırma işlemine sokulur. Burada üçgen tipi bulanık üyelik fonksiyonları tercih edilmiştir.



Şekil 5.3. Örnekte kullanılan üyelik fonksiyonu

3.Adım Örnekte hesaplamalarda Sınıf bir için herhangi bir kural oluşmadığından işlemler Sınıf 2' ye göre yapılmıştır. Bu adımda aynı zamanda birlikteliklerin ve bu birlikteliklerin destek değerlerinin belirlenmesi işlemleri yapılmaktadır.

Sınıf 2 için

Tablo 5.5. Verilerin Bulanık kümelere dönüştürülmesi

İşlem verileri	**A**			**B**			**C**			**D**			**Snf**
	Az	**Ao**	**Ab**	**Bz**	**Bo**	**Bb**	**Cz**	**Co**	**Cb**	**Dz**	**Do**	**Db**	
I_1	0	1	0	1	0	0	0,4	0,6	0	0,6	0	0	**1**
I_2	0,6	0	0	0	0,8	0,2	0,8	0,2	0	0	0,2	0,8	**1**
I_3	0,8	0,2	0	0,6	0	0	0,2	0,8	0	0	0,4	0,6	**0**
I_4	0,6	0,4	0	1	0	0	0,4	0,6	0	0	0,6	0,4	**0**
I_5	0	0,8	0,2	0,8	0	0	0,6	0,4	0	0	0,4	0,6	**0**
I_6	0,4	0,6	0	0,4	0	0	0,8	0	0	0,8	0,2	0	**1**
I_7	1	0	0	0,6	0,4	0	0,2	0,8	0	0	0	1	**0**
Toplam Destek	**3,4**	3	0,2	**4,4**	1,2	0,2	**3,4**	**3,4**	0	1,4	1,8	**3,4**	

Tablo 5.6 Sınıf 2'ye ait ikili birlikteliklerin bulunması

İşlem verileri	**Az**	**Bz**	**Destek(AzBz)**	**Snf**	**Destek(AzBz $\Rightarrow$ Sınıf2)**
I_1	0	1	0	0	0
I_2	0,6	0	0	0	0
I_3	0,8	0,6	0,6	1	0,6
I_4	0,6	1	0,6	1	0,6
I_5	0	0,8	0	1	0
I_6	0,4	0,4	0,4	0	0
I_7	1	0,6	0,6	1	0,6
Toplam Destek	**3,4**	**4,4**	**2,2**		1,8

İşlem verileri	**Az**	**Cz**	**Destek(AzCz)**	**Snf**	**Destek(AzCz $\Rightarrow$ Sınıf2)**
I_1	0	0,4	0	0	0
I_2	0,6	0,8	0,6	0	0
I_3	0,8	0,2	0,2	1	0,2
I_4	0,6	0,4	0,4	1	0,4
I_5	0	0,6	0	1	0
I_6	0,4	0,8	0,4	0	0
I_7	1	0,2	0,2	1	0,2
Toplam Destek	**3,4**	**3,4**	1,8		1

İşlem verileri	Az	Co	Destek(AzCo)	Snf	Destek(AzCo ⇒ Sınıf2)
I_1	0	0,6	0	0	0
I_2	0,6	0,2	0,2	0	0
I_3	0,8	0,8	0,8	1	0,8
I_4	0,6	0,6	0,6	1	0,6
I_5	0	0,4	0	1	0
I_6	0,4	0	0	0	0
I_7	1	0,8	0,8	1	0,8
Toplam Destek	**3,4**	**3,4**	**2,4**		**2,2**

İşlem verileri	Az	Db	Destek(AzDb)	Snf	Destek(AzDb ⇒ Sınıf2)
I_1	0	0	0	0	0
I_2	0,6	0,8	0,6	0	0
I_3	0,8	0,6	0,6	1	0,6
I_4	0,6	0,4	0,4	1	0,4
I_5	0	0,6	0	1	0
I_6	0,4	0	0	0	0
I_7	1	1	1	1	1
Toplam Destek	**3,4**	**3,4**	**2,6**		**2**

İşlem verileri	Bz	Cz	Destek(BzCz)	Snf	Destek(BzCz ⇒ Sınıf2)
I_1	1	0,4	0,4	0	0
I_2	0	0,8	0	0	0
I_3	0,6	0,2	0,2	1	0,2
I_4	1	0,4	0,4	1	0,4
I_5	0,8	0,6	0,6	1	0,6
I_6	0,4	0,8	0,4	0	0
I_7	0,6	0,2	0,2	1	0,2
Toplam Destek	**4,4**	**3,4**	**2,4**		1,4

İşlem verileri	Bz	Co	Destek(BzCo)	Snf	Destek(BzCo ⇒ Sınıf2)
I_1	1	0,6	0,6	0	0
I_2	0	0,2	0	0	0
I_3	0,6	0,8	0,6	1	0,6
I_4	1	0,6	0,6	1	0,6
I_5	0,8	0,4	0,4	1	0,4
I_6	0,4	0	0	0	0
I_7	0,6	0,8	0,6	1	0,6
Toplam Destek	**4,4**	**3,4**	**2,8**		**2,2**

İşlem verileri	Bz	Db	Destek(BzDb)	Snf	Destek(BzDb ⇒ Sınıf2)
I_1	1	0	0	0	0
I_2	0	0,8	0	0	0
I_3	0,6	0,6	0,6	1	0,6
I_4	1	0,4	0,4	1	0,4
I_5	0,8	0,6	0,6	1	0,6
I_6	0,4	0	0	0	0
I_7	0,6	1	0,6	1	0,6
Toplam Destek	**4,4**	**3,4**	**2,2**		**2,2**

İşlem verileri	Cz	Db	Destek(CzDb)	Snf	Destek(CzDb ⇒ Sınıf2)
I_1	0,4	0	0	0	0
I_2	0,8	0,8	0,8	0	0
I_3	0,2	0,6	0,2	1	0,2
I_4	0,4	0,4	0,4	1	0,4
I_5	0,6	0,6	0,6	1	0,6
I_6	0,8	0	0	0	0
I_7	0,2	1	0,2	1	0,2
Toplam Destek	**3,4**	**3,4**	**2,2**		1,4

İşlem verileri	Co	Db	Destek(CoDb)	Snf	Destek(CoDb ⇒ Sınıf2)
I_1	0,6	0	0	0	0
I_2	0,2	0,8	0,2	0	0
I_3	0,8	0,6	0,6	1	0,6
I_4	0,6	0,4	0,4	1	0,4
I_5	0,4	0,6	0,4	1	0,4
I_6	0	0	0	0	0
I_7	0,8	1	0,8	1	0,8
Toplam Destek	**3,4**	**3,4**	**2,4**		**2,2**

Burada 1.sınıfta minimum güven değerini geçen herhangi bir birliktelik bulunmamaktadır. 2.sınıfta (AzCo), (AzDb), (BzDb), (Bz,Co), (Co,Db) birliktelikleri minimum destek değerini geçmişlerdir. Daha sonra bu birlikteliklerin güven değerleri hesaplanır.

4.Adım.

$$\text{Güven}(\text{AzCo} \Rightarrow \text{Sınıf2}) = \frac{\text{Destek}(\text{AzCoSinif2})}{\text{Destek}(\text{AzCo})} = \frac{2,2}{2,4} = 0,91 = \%91 > \text{min\_guv}(\%85)$$

$$\text{Güven}(\text{AzDb} \Rightarrow \text{Sınıf2}) = \frac{\text{Destek}(\text{AzDbSinif2})}{\text{Destek}(\text{AzDb})} = \frac{2}{2,6} = 0,76 = \%76 < \text{min\_guv}(\%85)$$

$$\text{Güven}(\text{BzDb} \Rightarrow \text{Sınıf2}) = \frac{\text{Destek}(\text{BzDbSinif2})}{\text{Destek}(\text{BzDb})} = \frac{2,2}{2,2} = 1 = \%100 > \text{min\_guv}(\%85)$$

$$\text{Güven}(\text{BzCo} \Rightarrow \text{Sınıf2}) = \frac{\text{Destek}(\text{BzCoSinif2})}{\text{Destek}(\text{BzCo})} = \frac{2,2}{2,8} = 0,78 = \%78 < \text{min\_guv}(\%85)$$

$$\text{Güven}(\text{CoDb} \Rightarrow \text{Sınıf2}) = \frac{\text{Destek}(\text{CoDbSinif2})}{\text{Destek}(\text{CoDb})} = \frac{2,2}{2,4} = 0,91 = \%91 > \text{min\_guv}(\%85)$$

Yukarıdaki güven hesaplama işlemlerine göre güven(AzCo $\Rightarrow$ Sınıf2), Güven(BzDb $\Rightarrow$ Sınıf2) ve Güven(CoDb $\Rightarrow$ Sınıf2) sınıflamalarının minimum güven değerinin üzerinde olduğu görülmüştür. Bu kurallardan güven(CoDb $\Rightarrow$ 2) kuralının gizlenmek istendiği farz edilmiştir.

5.Adım (CoDb $\Rightarrow$ 2) sınıflamasının gizlenmesi için sınıf değerinin 0 olduğu durumlarda (CoDb) ikilisinin destek değeri yükseltilir. Bu destek değerini yükseltmek için (CoDb) ikilisinden destek değeri 0,5 ten ve diğer nenenin destek değerinden küçük olan nesnenin değeri 1'den çıkarılır. Bu işlem veri kümesinin en başından başlanarak güven değeri, minimum güven değerinin altına düşüne kadar veri kümesinin sonuna kadar devam edilir.

Tab

İşlem verileri	**Co**	**Db**	**Destek(CoDb)**	**Snf**	**Destek(CoDb $\Rightarrow$ Sınıf2)**
I_1	0,6	0	0	0	0
I_2	0,2	0,8	0,2	0	0
I_3	0,8	0,6	0,6	1	0,6
I_4	0,6	0,4	0,4	1	0,4
I_5	0,4	0,6	0,4	1	0,4
I_6	0	0	0	0	0
I_7	0,8	1	0,8	1	0,8
Toplam Destek	**3,4**	**3,4**	**2,4**		**2,2**

I_2 = Co = 1-0,2=0,8

İşlem verileri	**Co**	**Db**	**Destek(CoDb)**	**Snf**	**Destek(CoDb $\Rightarrow$ Sınıf2)**
I_1	0,6	0	0	0	0
I_2	**0,8**	0,8	**0,8**	0	0
I_3	0,8	0,6	0,6	1	0,6
I_4	0,6	0,4	0,4	1	0,4
I_5	0,4	0,6	0,4	1	0,4
I_6	0	0	0	0	0
I_7	0,8	1	0,8	1	0,8
Toplam Destek	**3,4**	**3,4**	**3**		**2,2**

$$\text{Güven(CoDb} \Rightarrow \text{Sınıf2)} = \frac{\text{Destek(CoDbSinif2)}}{\text{Destek(CoDb)}} = \frac{2,2}{3} = 0,73 = \%73 > \text{min\_guv}(\%85)$$

Yukarıda görüldüğü gibi (CoDb) $\Rightarrow$ 2 kuralının güven değeri ilk başta %91'idi ve minimum güven değerinin üzerinde olduğu için bu kural geçerli bir kural olmaktaydı. Yöntem uyguladıktan sonra (CoDb) $\Rightarrow$ 2 kuralının güven değeri %73 oldu ve minimum güven değerinin altına düştü. Böylece kural geçerliliğini yitirmiş olur ve kuralın gizlenmesi gerçekleşti.

6.Adım. Değiştirilmiş değerlerin bulanıklaştırılmış veri tabanında yerlerine yazılmalır.

Tablo 5.7. Değiştirilmiş değerlerle bulanık veri tabanı

İşlem verileri	**A**			**B**			**C**			**D**		
	Az	**Ao**	**Ab**	**Bz**	**Bo**	**Bb**	**Cz**	**Co**	**Cb**	**Dz**	**Do**	**Db**
$\mathbf{I_1}$	0	1	0	1	0	0	0,4	0,6	0	0,6	0	0
$\mathbf{I_2}$	0,6	0	0	0	0,8	0,2	0,8	**0,8**	0	0	0,2	0,8
$\mathbf{I_3}$	0,8	0,2	0	0,6	0	0	0,2	0,8	0	0	0,4	0,6
$\mathbf{I_4}$	0,6	0,4	0	1	0	0	0,4	0,6	0	0	0,6	0,4
$\mathbf{I_5}$	0	0,8	0,2	0,8	0	0	0,6	0,4	0	0	0,4	0,6
$\mathbf{I_6}$	0,4	0,6	0	0,4	0	0	0,8	0	0	0,8	0,2	0
$\mathbf{I_7}$	1	0	0	0,6	0,4	0	0,2	0,8	0	0	0	1
Toplam Destek	**3,4**	3	0,2	**4,4**	1,2	0,2	**3,4**	**3,4**	0	1,4	1,8	**3,4**

Yukarıdaki veri tabanında, I_2'i işlem verisinin C nesnesinin değerine bakıldığında bulanık değerlerin toplamı 1'i geçmektedir. Bu sorunu giderebilmek için değiştirilmiş nesnenin değeri 1'den çıkarılıp geri kalan değer diğer bulanık mantık öğesinin yerine yazılır. Örneğin

Cz + Co+Cb = 0,8 + 0,8+0 = 1,6 Burada Co 1'den çıkarılır. 1-0,8 = 0,2. Çıkan değer Cz'nin yerine yazılır.Cz'nin yeni değeri 0,2 olur.

Tablo 5.8. Tablo 5.8'deki bulanık veri kümesinin optimizasyondan geçirildikten sonraki durumu

İşlem verileri	**A**			**B**			**C**			**D**		
	Az	**Ao**	**Ab**	**Bz**	**Bo**	**Bb**	**Cz**	**Co**	**Cb**	**Dz**	**Do**	**Db**
$\mathbf{I_1}$	0	1	0	1	0	0	0,4	0,6	0	0,6	0	0
$\mathbf{I_2}$	0,6	0	0	0	0,8	0,2	**0,2**	**0,8**	0	0	0,2	0,8
$\mathbf{I_3}$	0,8	0,2	0	0,6	0	0	0,2	0,8	0	0	0,4	0,6
$\mathbf{I_4}$	0,6	0,4	0	1	0	0	0,4	0,6	0	0	0,6	0,4
$\mathbf{I_5}$	0	0,8	0,2	0,8	0	0	0,6	0,4	0	0	0,4	0,6
$\mathbf{I_6}$	0,4	0,6	0	0,4	0	0	0,8	0	0	0,8	0,2	0
$\mathbf{I_7}$	1	0	0	0,6	0,4	0	0,2	0,8	0	0	0	1
Toplam Destek	**3,4**	3	0,2	**4,4**	1,2	0,2	**3,4**	**3,4**	0	1,4	1,8	**3,4**

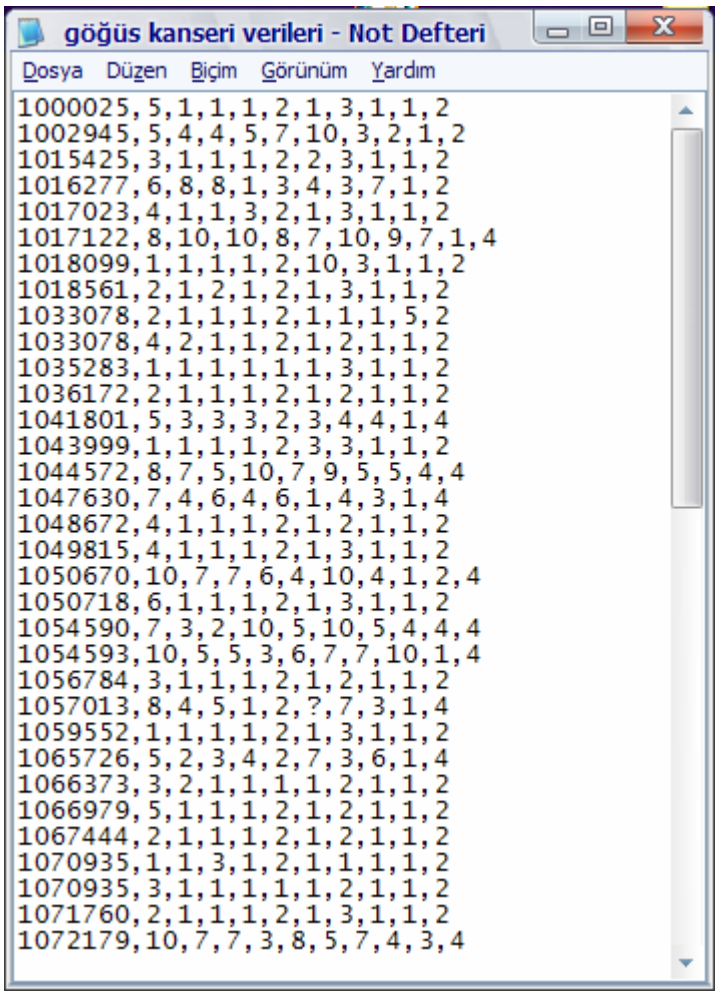
7.Adım Sınıflar birleştirilerek dönüştürme işlemi yapıldıktan sonra veri tabanı aşağıdaki duruma gelir.

Tablo 5.10. Veri tabanının dönüştürme işleminden sonraki durumu

	A	B	C	D	Sınıf
I_1	10	5	8	3	1
I_2	3	11	**9**	14	1
I_3	6	3	9	13	2
I_4	7	5	8	12	2
I_5	5	4	7	13	2
I_6	8	2	4	6	1
I_7	5	7	9	15	2

6. UYGULAMA YAZILIMI

Bu tezde, önerilen algoritmaların uygulanabilirliği ve etkisini test etmek için bir yazılım geliştirilmiştir. Bu yazılım, **http://mlearn.ics.uci.edu/databases/breast-cancer-wisconsin/** sitesinden alınan göğüs kanseri verilerini kullanmaktadır. Aşağıda uygulama yazılımının kullandığı göğüs kanseri verisi görülmektedir.



Şekil 6.1 Göğüs kanseri verileri

Uygulama yazılımında 200 adet göğüs kanseri hastalığı verisiyle çalışılmıştır. Bu veri kümesi 1sınıflandırma alanı ve 9 şart niteliğinden oluşmaktadır. Aşağıda her bir niteliğin ne anlama geldiği konusunda bilgi verilmiştir.

1: Hasta kimlik numarası
2: yoğunluk miktarı
3: Hücre büyüklüklerinin birbirine benzerliği
4: Hücre şekillerinin birbirine benzerliği

5: Marjinal yapışkanlık

6: Tek epitelyum hücre büyüklüğü

7: Çıplak DNA'lar

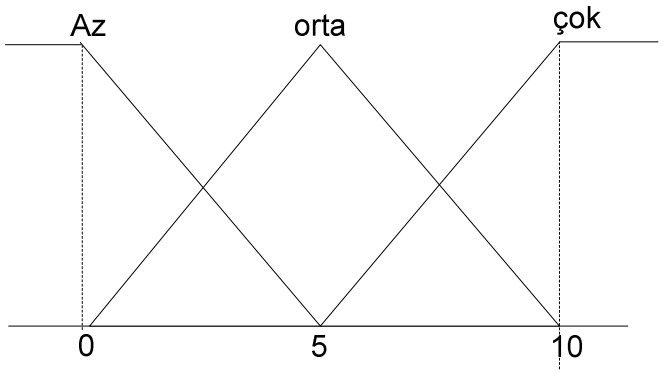
8: Bland Chromatin

9: Normal Nucleoli

10: Mitose

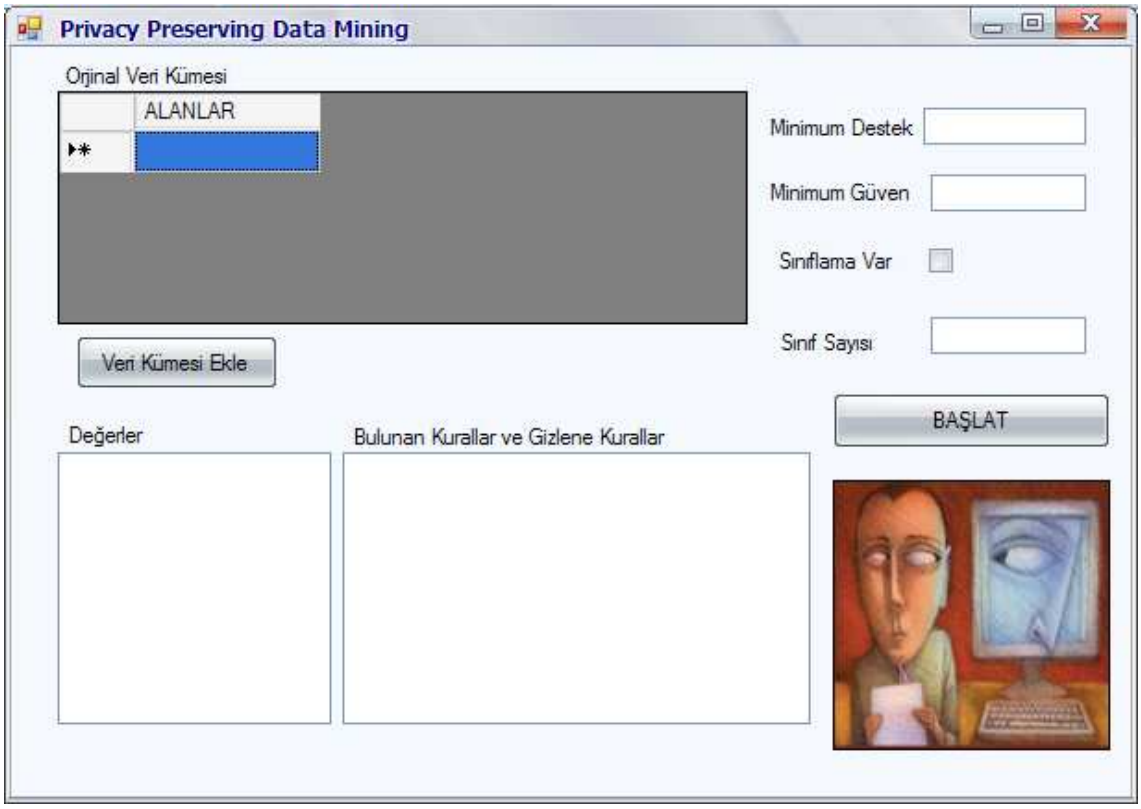
11: Sınıflar
a) 2 : İyi huylu b) 4: Kötü huylu

Uygulama yazılımında 2 algoritma kullanılmıştır. 4.1'teki algoritmada hassas birliktelik kurallarının gizlenmesi üzerine çalıştığından sınıflandırma niteliği dikkate alınmamıştır. 5.2'deki algoritma ise hassas sınıflandırma kurallarının gizlenmesi üzerine çalıştığından sınıflandırma verisi kullanılmıştır. Yukarıdaki göğüs kanseri verisinde 3 tane kayıp veri bulunmaktadır. Bu verilerin bulunduğu hastalar incelenmeye alınmamıştır. Her iki algoritmada verilerin aralıkları aynı (1-10) olduğundan, her iki algoritmada aynı üyelik fonksiyonu ile ifade edilmiştir.



Şekil 6.2 Uygulama yazılımında kullanılan üyelik fonksiyonu

Yukarıdaki üyelik fonksiyonu doğrultusunda program iki kısımdan oluşmaktadır. Birinci kısım hassas birliktelik kurallarının gizlenmesi ikinci kısım hassas sınıflandırma kurallarının gizlenmesinde kullanılmaktadır. Şekil 6.2'de gerçekleştirilen uygulama yazılımının ara yüzü görülmektedir.



Şekil 6.3 Uygulama yazılımının ara yüzü

6.1. Uygulama Ortamı

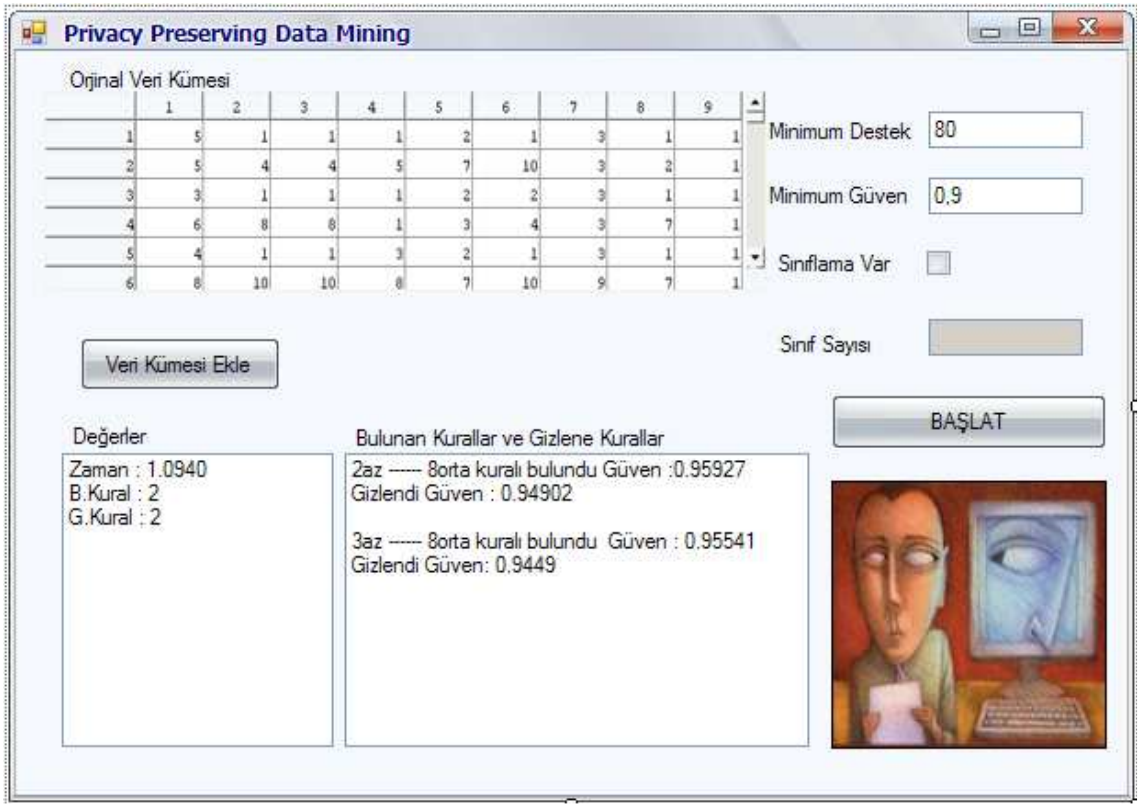
Uygulama yazılımı gerçekleştirilirken Matlab 6.5 programlama dili kullanılmıştır. Uygulama yazılımı için İntel Centrino (Sonoma) 1,7 GHz işlemciye, 512 MB RAM'e sahip bir bilgisayar kullanılmıştır.

6.2. Geliştirilen Yazılım İle Elde Edilen Sonuçlar

Geliştirilen uygulama yazılımında 2 farklı duruma göre elde edilen sonuçlara yer verilmiştir. Birinci durumda sınıflandırma işlemi olmaksızın birliktelik kurallarının gizlenmesi işlemi incelenmektedir. İkinci durumda ise sınıflandırma işlemi dikkate alınarak elde edilen sınıflandırmaların gizlenmesi işlemi incelenmektedir.

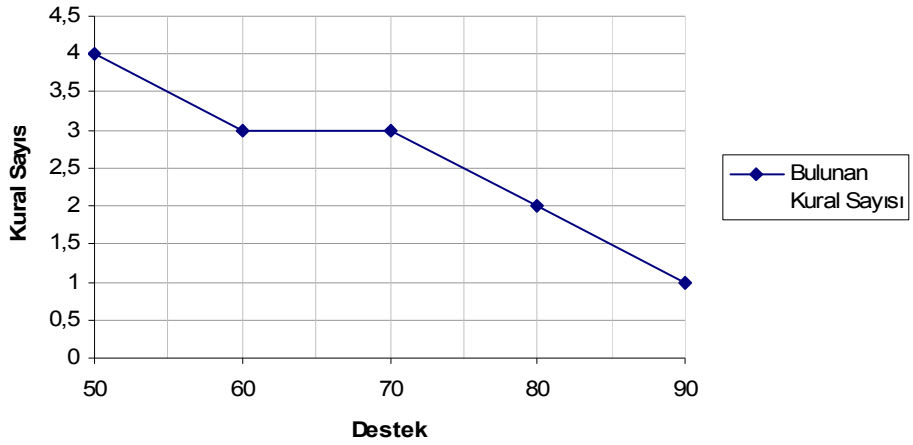
Birliktelik kurallarının kullanılması ile ilgili ekran görüntüsü Şekil 6.3'te verilmiştir. Bu kısımda sınıflandırma seçeneği iptal edilmiştir. Minimum destek değeri 80 ve minimum güven değeri 0,90 olarak alınmıştır.

6.2.1. Birliktelik Kuralları İçin Elde Edilen Sonuçlar



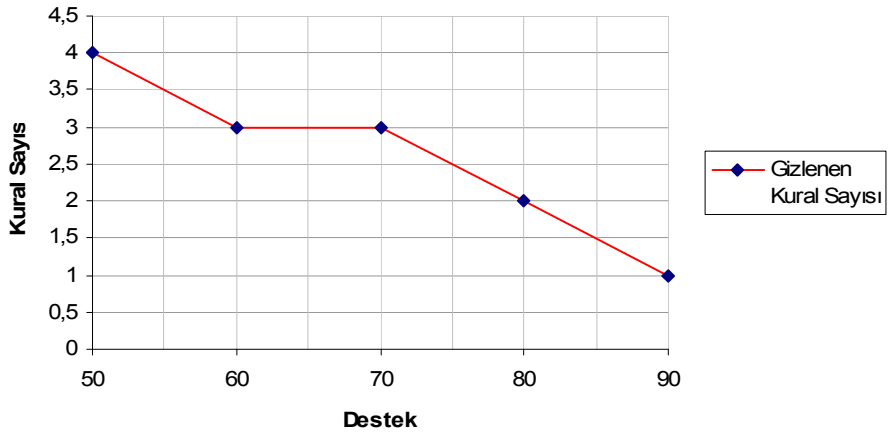
Şekil 6.4 1.Durum altında çalışan uygulama yazılımının ekran görüntüsü

Programın çalıştırılması sonucunda güven değeri minimum güven değerinin üzerinde olan 2 birliktelik kuralı bulunmuştur. Bu iki birliktelik kuralının ikisi de gizlenmiştir. Oluşan yeni veri kümesi tekrar işleme alındığında herhangi bir kural bulunmamıştır. Minimum güven değerinin ve minimum destek değerinin değiştirilmesi ile elde edilen birliktelik kuralları, gizlenen kurallar, bu gizleme neticesi değiştirilmiş veri tabanından elde edilen yeni kurallar aşağıdaki grafiklerde verilmiştir.



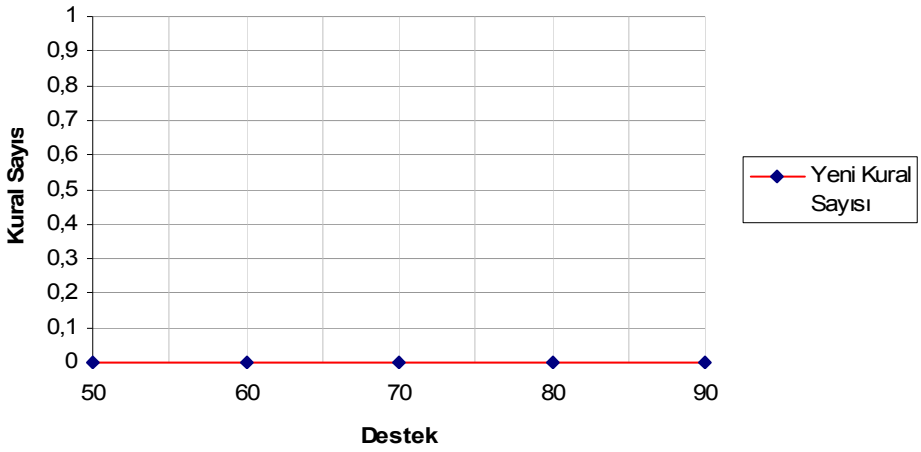
Şekil 6.5 Farklı destek değerlerinde bulunan kural sayısı

Şekil 6.5'teki grafikte değişen minimum destek değerine karşılık bulunan birliktelik kuralları gösterilmiştir. Minimum destek değeri arttıkça, destek değeri minimum destek değerini geçen nesnelerin sayısı düşmektedir. Destek değeri ile bulunan kural sayısı arasında ters bir orantı görülmektedir.



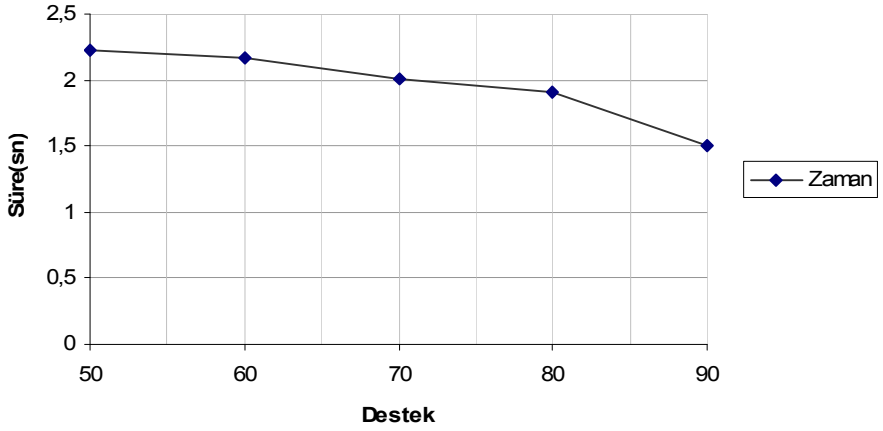
Şekil 6.6 Farklı destek değerlerinde gizlenen kural sayısı

Şekil 6.6'ta ki grafikte değişik destek değerleri altında bulunan birliktelik kurallarından gizlenebilenler gösterilmiştir. Grafikte görüldüğü gibi şekil 6.5'teki grafikte bulunan kuralların tamamı gizlenebilmiştir. Bu da yöntemin başarısını göstermektedir.



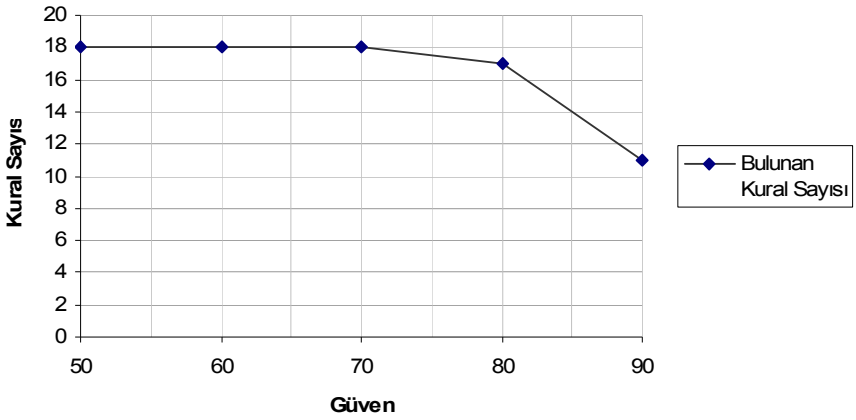
Şekil 6.7. Farklı destek değerlerinde değiştirilmiş veri tabanında elde edilen yeni kural sayısı

Şekil 6.6 'daki grafikte bulunan kurallar gizlendikten sonra veri değerleri değiştirilen veri tabanı tekrar incelenmeye alınmıştır. Bu incelemenin sonucunda Şekil 6.7'de gösterildiği gibi veri tabanından herhangi bir yeni kural çıkmamaktadır.



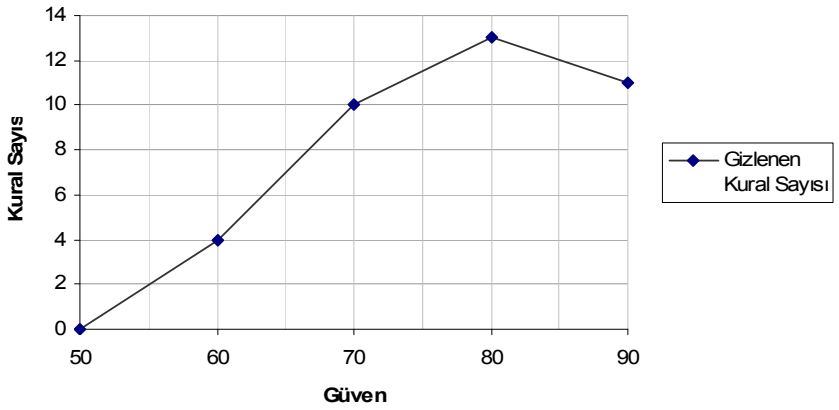
Şekil 6.8 Farklı destek değerlerinde programın çalışma zamanı

Şekil 6.8'deki grafikte algoritmanın çalışma zamanı gösterilmiştir. Bu grafikte destek değeri arttıkça algoritmanın çalışma zamanı azalmaktadır. Bunun nedeni minimum destek değeri arttıkça, minimum destek değerini geçen nesnelerin sayısı azalmaktadır. Bundan dolayı değiştirilecek birliktelik sayısı azalmakta ve algoritmanın çalışma zamanı kısalmaktadır.



Şekil 6.9 Farklı güven aralığında bulunan kural sayısı

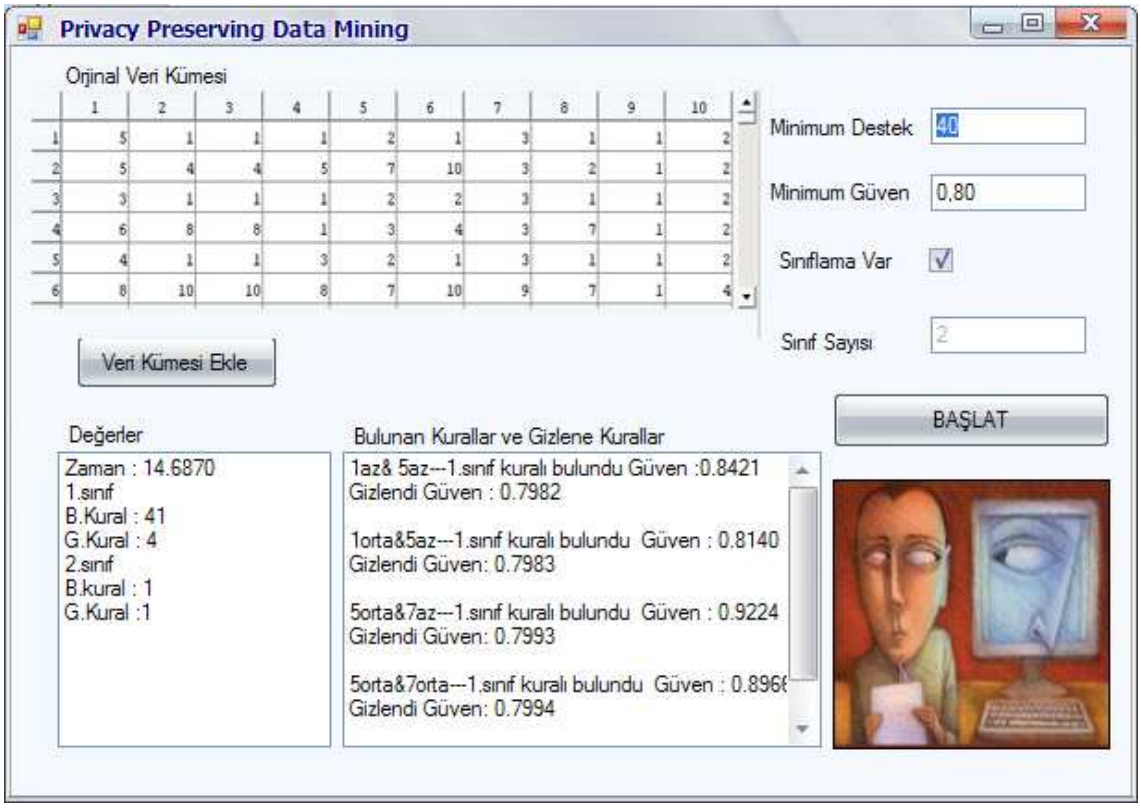
Şekil 6.9'daki grafikte değişen minimum güven değerlerine karşılık bulunan kural sayısı gösterilmektedir. Grafikte kuralların güven aralığı genelde %70 ve yukarısı olduğu için %70'e kadar hemen hemen tüm kurallar bulunabilmektedir.



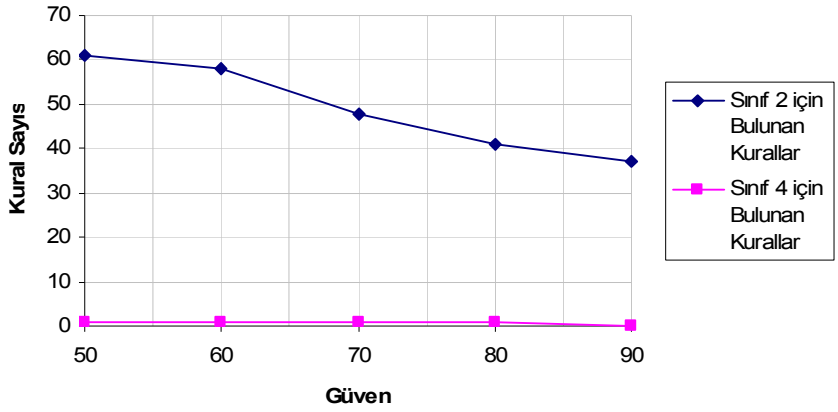
Şekil 6.10 Farklı güven aralığında gizlenen kural sayısı

Şekil 6.10'daki grafikte değişen güven değerleri altına bulunan kuralların gizlenme oranı gösterilmektedir. Bu grafikte, veri tabanındaki birlikteliklilerin güven değerleri genelde 70'in üzerinde olduğu için algoritma değerleri değiştirmesine rağmen, güven değerleri arasındaki fark çok büyüktür. Bu nedenle düşük güven değerlerinde kuralların gizlenme oranı düşüktür.

6.2.2. Sınıflandırma İçin Elde Edilen Sonuçlar



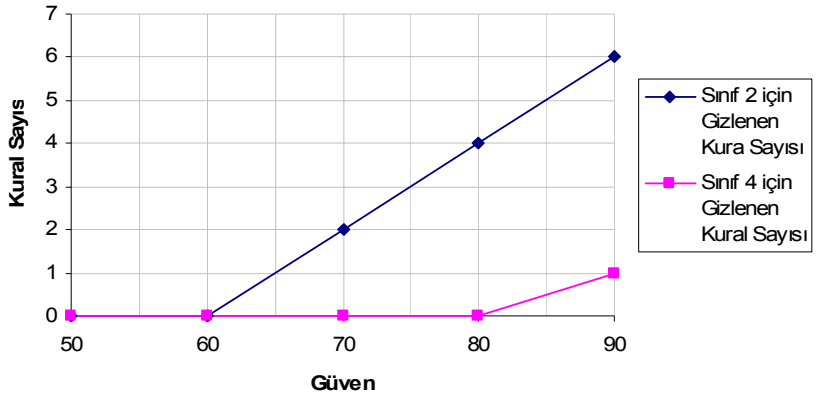
Şekil 6.11 2.durum altında çalışan uygulama yazılımının ekran görüntüsü



Şekil 6.12 Değişik Güven değerlerinde farklı sınıflar için elde edilen kural sayısı

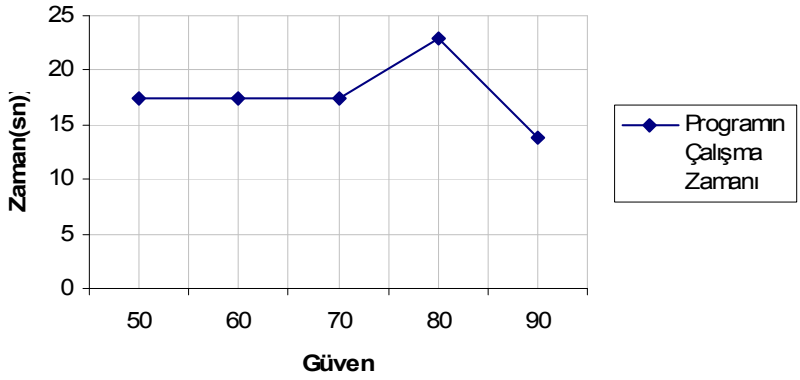
Şekil 6.12'deki grafikte farklı güven değerlerinde 2. ve 4.sınıflara ait bulunan kural sayısı gösterilmektedir. Grafikte 2. sınıfa ait dağılım normal bir şekilde dağılmaktadır. 4.sınıfa ait kural sayısının çok az olmasının nedeni 4.sınıfa ait verilerin az sayıda olması ve bu verilerin destek değerlerinin minimum destek değerinin altında kalmasıdır. Yukarıdaki grafikte ve güven

değerlerinin değiştiği grafiklerde destek değeri 40 alınmıştır. Destek değeri çok düşürüldüğü zaman ise 2. sınıfa ait çok fazla gereksiz kural bulunmaktadır. Buda algoritmanın verimliliğini düşürmektedir.



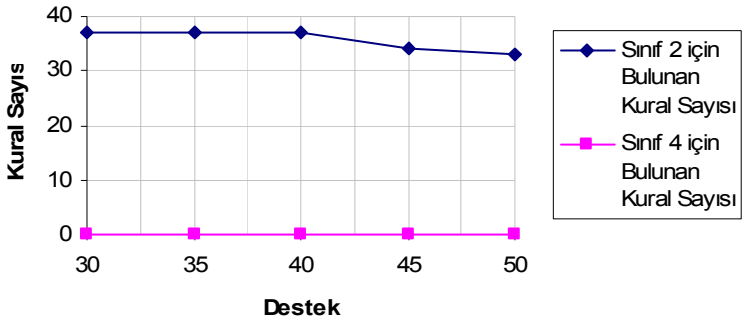
Şekil 6.13 Değişik Güven değerlerinde farklı sınıflar için gizlenen kural sayısı

Şekil 6.13'teki grafikte farklı güven değerlerine karşılık bulunan kurallar içerisinden gizlenebilen kuralların sayısı gösterilmiştir. Grafikte minimum güven değeri 60'ın altında olduğu zaman hiçbir kural gizlenememektedir. Bunun nedeni sınıf birlikteliklerindeki güven değerlerinin minimum güven değerlerinden çok büyük olmasındandır.



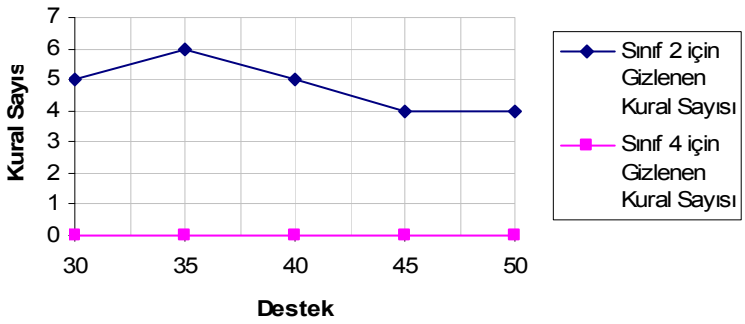
Şekil 6.14 Değişik Güven değerlerinde programın çalışma zamanı

Şekil 6.14'tek farklı güven değerleri altında grafik sınıflama işlemi dikkate alındığında algoritmanın çalışma zamanını göstermektedir. Grafikteki keskin değişimler birlikteliklerin yoğun olduğu güven değer aralıklarını göstermektedir.



Şekil 6.15 Değişik Destek değerlerinde farklı sınıflar için bulunan kural sayısı

Şekil 6.15'teki grafikte farklı destek değerleri altında sınıflara ait bulunan kural sayısı gösterilmiştir. Burada minimum destek aralığının 50'nin altında alınmasını nedeni 4. sınıfa ait nesnelerin destek değerlerinin hiç birinin 50'yi geçememesinden kaynaklanmaktadır.



Şekil 6.16 Değişik Destek değerlerinde farklı sınıflar için gizlenen kural sayısı

7. SONUÇLAR

Veri madenciliği, veri tabanlarında istenilen ilişkilerin otomatik çıkarılması için bir yöntemdir. Veri madenciliğinin kötü niyetli olarak kullanılması kişisel gizlikleri tehlikeye sokmaktadır. Bu durum veri madenciliği araştırmalarına izin verilmemesine kadar uzanmaktadır. Bu kötü niyetli veri madenciliği işlemlerini engellemek için bir çok yöntem geliştirilmiştir. Bu yöntemlerin çoğu veri tabanlarındaki nesnelerin bulunup bulunmamasına yani (1,0) mantığına göre işlem yapmaktadır. Bu tez çalışmasında nicel değerli veri tabanlarında yapılabilecek kötü niyetli veri madenciliğinin bulanık mantık yöntemi, apriori algoritması ve ISL yöntemi kullanılarak nasıl engelleneceğine dair bilgi verilmiştir. Birliktelik ve sınıflandırma olmak üzere iki farklı madenciliğin engellenmesi üzerinde durulmaktadır ve hassas bağıntı kurallarının gizlenmesi için iki farklı algoritma geliştirilmiştir.

Bu tez çalışması 4 aşamadan meydana gelmektedir. Birinci. aşama veri tabanının bulanıklaştırılmasıdır. Sınıflamada bu aşamaya ek olarak veri tabanındaki veriler sınıflara göre ayrılarak sınıf sayısı kadar veri kümelerine ayrılır ve işlemler bu kümeler üzerinde ayrı ayrı yapılır. İkinci aşama bulanıklaştırılan veri tabanından bağıntı kurallarının bulunması, üçüncü aşamada bulunan bağıntı kuraları içerisinde hassas bağıntı kurallarının gizlenmesi ve dördüncü aşamada da veri tabanın eski haline döndürülmesi işlemleri yapılır.

Tezde, geliştirilen algoritma, uygulama yazılımı haline getirilmiş ve elde edilen kurallar ve gizlenen kurallara yer verilmiştir. Ayrıca sonuçlar grafik üzerinde de gösterilmiştir. Uygulama yazılımında göğüs kanserine ait 200 adet kayıt ile çalışılmıştır. Bu çalışmada, hastalık verilerinin tercih edilmesindeki temel sebep, nicel değerlerin bu veri kümelerinde daha sık görülmeleri ve kötü niyetli madencilik işlemlerine çok fazla maruz kalmalarıdır.

Bu yöntemin dezavantajı, hassas olmayan bazı kuraların da gizlenebilmesidir. Fakat bu hassas olmayan kuralların gizlenme oranı düşük olmaktadır. Bu yüzden yöntemin verimliliğini etkilememektedir.

KAYNAKLAR

1. V. Verykios, E. Bertino, I.G. Fovino, L.P. Provenza, Y. Saygın and Y. Theodoridis, "State – of-the–art in Privacy Preserving Data Mining", *Sıgmod Record*, Vol. 33, No. 1, 50-57, March 2004.
2. http://www.merriam-webster.com/
3. Shyue-Liang Wang and Ayat Jarafi, "Using unknows for hiding sensitive Predective assosiaction rules", *IEEE IRI* 2005.
4. Directive 95/46/EC of the European ParHament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing ofpersonal data and on the free movement of such data. Official Journal of the European Communities, No I.(281):31-50, Oct. 24 1995.
5. Standard for privacy of individually identifiable health information. *Federal Register,* 67(157):53181-53273, Aug. 14 2002.
6. L. Sweeney, "K-anonymity: a model for protecting privacy". *International Journal on Uncertainty*, Fuzziness and Knowledge-based Systems, (5):557-570, 2002.
7. Chris Clifton and Don Marks. "Security and privacy implications of data mining", *ACM SIGMOD* pages 15–19, Montreal, Canada, June 2 1996.
8. M. Delgado and A. Gonzalez, "An inductive learning procedure to identify fuzzy systems", *Fuzzy Sets Syst* 55 (1993) 121–132.
9. S. L. Warner. "Randomized response: a survey technique for eliminating evasive answer bias" *Journal of the American Statistical Association*, 60(309):63-69, Mar. 1965.
10. W. Du and Z. Zhan. "Using randomized response techniques for privacypreserving data mining" *In The Ninth ACM SIGKDD International Conferenceon Knowledge Discovery and Data Mining*, pages 505-510, Washington, DC,Aug. 24-27 2003
11. A. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke. "Privacy preserving mining of association rules". *In The Eighth ACM SIGKDD International Conference on KnowledgeDiscovery and Data Mining,* pages 217-228, Edmonton, Alberta, Canada, July 23-26 2002.
12. Y. Zhu and L. Liu. "Optimal randomization for privacy preserving data mining" *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 761-766, New York, NY, USA, 2004.
13. H. Kargupta, S. Datta, Q. Wang, and K. Sivakumar. "On the privacy preserving properties of random data perturbation techniques". *In Proceedings of the Third IEEE International Conference on Data Mining (ICDM'03),* Melbourne, Florida, Nov. 19-22 2003.

14. Z. Huang, W. Du, and B. “Chen. Deriving private information from randomized data”. In *Proceedings of the 2005 ACM SICMOD International Conference on Management of Data*, Baltimore, MD, June 13-16 2005.
15. T.Dalenius and S.p.Reiss, “Data Swapping A Technique for Disclourse Control,” *J.Statistical Planning and Inference*, vol. 6, pp, 73-85, 1982
16. C.K. Liew, u.J.Choi, and C. J. Liew, “A data disortion by probability distribution.” *ACM Trabsection and Database systems (TODS),* 10(3):395-411, 1985
17. A.C. Yao, “How to Generate and Exchange Secrets,” *Proc. 27th IEEE Symp. Foundations of computer Science, IEEE CS Press*, 1986, pp. 162–167.
18. M. Kantarcıoğlu and C. Clifton, “Privacy-Preserving Distributed Mining of Association Rules on Horizontally Partitioned Data,” *IEEE Trans. Knowledge Data Eng.*,vol. 16, no. 9, pp. 1026–1037, 2004,
19. J. Vaidya and C. Clifton. “Privacy preserving association rule mining in vertically partitioned data”. In *The Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 639-644, Edmonton, Alberta, Canada, July 23-26 2002.
20. Agrawal R, Imielinki T. And Swami A, “Mining association rules between sets of items in large databases”, *ACM SIGMOD Conf. Managment of Data*, 207-216,1993.
21. Agraval R. And Srikant R,“Mining SequentialPatterns”, *Proc. Of the 11th International Conference on data Engineering*, pp.3-14, 995,
22. L.A. Zadeh, Fuzzy sets, Inform. Control 8 (3) 338–353. (1965)
23. A.F. Blishun, “Fuzzy learning models in expert systems”, *Fuzzy Sets Syst.* 22 57–70. (1987)
24. L.M. de Campos and S. Moral, “Learning rules for a fuzzy inference model”, *Fuzzy Sets Syst.* 59 247–257. (1993)
25. R. Agrawal and R. Srikant, “Fast algorithm for mining association rules”, *The International Conference on Very Large Data Bases*, pp. 487–499. 1994
26. R. Agrawal, R. Srikant and Q. Vu, “Mining association rules with item constraints,” *The Third International Conference on Knowledge Discovery in Databases and Data Mining*, Newport Beach, California, Ağustos 1997.
27. R. Agrawal, T. Imielinksi and A. Swami, “Database mining: a performance perspective”, *IEEE Trans. Knowledge Data Eng*. 5 (6) 914–25. (1993)
28. R. Srikant and R. Agrawal, “Mining quantitative association rules in large relational tables”, *ACM SIGMOD International Conference on Management of Data,* Monreal, Canada, June, pp. 1–12. 1996
29. Tzung-Pei Hong, Chan-Sheng Kuo and Shyue-Liang Wang, “A fuzzy AprioriTid mining algorithm with reduced computational time” *Applied Soft Computing*, pp. 1–10. (2004)

30. Y.Saygın. V.Verkios, and C.Clifton, "Using unknows to prevent Discovery of Associations Rules*", SIGMOD Record* 30(4): 45-54, December 2001.
31. T.P. Hong, C.S. Kuo and S.C. Chi," Mining association rules from quantitative data", *Intell. Data Anal*. 3 (5) 363–376. (1999)
32. Danıel T. Larose. "Dıscoverıng Knowledge In Data An Introduction To Data Mining," *John Wiley & Sons, Inc*. ISBN 0-471-66657-2
33. http://mlearn.ics.uci.edu/databases/breast-cancer-wisconsin/

ÖZGEÇMİŞ

Tolga BERBEROĞLU, 1979 yılında Elazığ'da doğdu. İlk ve orta öğretimini Elazığ'da tamamladı. 1997 yılında Elazığ Balakgazi Süper Lisesi'nden mezun oldu. Aynı yıl Gazi Üniversitesi Teknik eğitim Fakültesi Elektronik – Bilgisayar Anabilim dalı Bilgisayar Sistemleri Öğretmenliği bölümünde lisans eğitimi almaya hak kazandı. 2001 yılında ilgili bölümden mezun oldu. 2005 yılı güz döneminde Fırat Üniversitesi Bilgisayar Mühendisliği bölümünde yüksek lisans eğitimi almaya başladı.