BULANIK KABA KÜME YÖNTEMİ İLE NİTELİK İNDİRGEMEDE YENİ BİR ALGORİTMA

Bekir AĞIRGÜN

DOKTORA TEZİ
ENDÜSTRİ MÜHENDİSLİĞİ

GAZİ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

ŞUBAT 2009
ANKARA

Bekir AĞIRGÜN tarafından hazırlanan Bulanık Kaba Küme Yöntemi ile Nitelik İndirgemede Yeni Bir Algoritma adlı bu tezin Doktora tezi olarak uygun olduğunu onaylarım.

Prof. Dr. Cevriye GENCER …….
Tez Danışmanı, Endüstri Müh.A.B.D.

Bu çalışma, jürimiz tarafından oy birliği ile Endüstri mühendisliği Anabilim Dalında Doktora tezi olarak kabul edilmiştir.

Prof. Dr. Semra Oral ERBAŞ …….
İstatistik Bölümü Anabilim Dalı, G.Ü.
Prof. Dr. Cevriye GENCER …….
Endüstri Müh. Anabilim Dalı, G.Ü.
Doç.Dr. Aydın ULUCAN …….
Sayısal Yöntemler Anabilim Dalı, Hacettepe Üniversitesi
Yrd.Doç.Dr. Feyzan ARIKAN …….
Endüstri Müh. Anabilim Dalı, G.Ü.
Dr. Soner BİNAY …….
Harekat Araştırması Anabilim Dalı, Kara Harp Okulu

Tarih: 23/02/2009

Bu tez ile G.Ü. Fen Bilimleri Enstitüsü Yönetim Kurulu Doktora derecesini onamıştır.

Prof. Dr. Nail ÜNSAL …….
Fen Bilimleri Enstitüsü Müdürü

TEZ BİLDİRİMİ

Tez içindeki bütün bilgilerin etik davranış ve akademik kurallar çerçevesinde elde edilerek sunulduğunu, ayrıca tez yazım kurallarına uygun olarak hazırlanan bu çalışmada bana ait olmayan her türlü ifade ve bilginin kaynağına eksiksiz atıf yapıldığını bildiririm.

Bekir AĞIRGÜN

BULANIK KABA KÜME YÖNTEMİ İLE NİTELİK İNDİRGEMEDE YENİ BİR ALGORİTMA

(Doktora Tezi)

Bekir AĞIRGÜN

GAZİ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

Şubat 2009

ÖZET

Günümüzde bilgisayar teknolojilerinin gelişmesi ile yüksek boyutlu veritabanları ile karşılaşılmaktadır. Veritabanlarından gereksiz verilerin atılması, veri içerisindeki örüntülerin bulunması ve elde edilen bilginin kullanılması sürecine veri tabanlarında bilgi keşfi denilmektedir. Nitelik indirgeme, kısaca nitelik uzayının boyutunun belirli ölçütlere göre azaltılması olarak tanımlanmaktadır. Bu kapsamda literatürde birçok çalışma yapılmıştır. Bu çalışmaların içerisinde en ilginç olanlarından biri ise bulanık ve kaba kümelerin melezlenmesi ile yapılan nitelik indirgeme çalışmalarıdır.

Bu tezde, tip-2 bulanık kümeler kullanılarak bulanık-kaba kümeler için ayırtedilebilirlik matrisi tabanlı nitelik indirgemede yeni bir yaklaşım önerilmektedir. Önerilen algoritma, UCI makine öğrenimi veri setinden alınan veri kümeleri kullanılarak, literatürdeki nitelik indirgeme algoritmaları ile karşılaştırılmış olup, etkinliğini test etmek için sınıflandırma algoritmalarından faydalanılmıştır.

Bilim Kodu : **906.1.148**
Anahtar Kelimeler :**Nitelik indirgeme, Kaba Küme, Bulanık Küme, Bulanık-Kaba Küme.**
Sayfa Adedi : **136**
Tez Yöneticisi **Prof. Dr. Cevriye GENCER**

A NEW ALGORİTHM ON ATTRIBUTE REDUCTION WITH FUZZY ROUGH SET METHOD
(Ph.D. Thesis)

Bekir AĞIRGÜN

GAZI UNIVERSITY
INSTITUTE OF SCIENCE AND TECHNOLOGY
February 2009

ABSTRACT

Today, with the development of computer technologies, we encounter with high dimensional databases. The process of removing the unnecessary data form databases, finding the patterns inside the data and using the obtained knowledge is called knowledge discovery in databases. Attribute reduction is defined as reducing of attribute space dimension according to certain criteria. In this context, a lot of studies have been carried out in literature. In these studies, the most interesting one is the attribute reduction with hybridization of fuzzy and rough sets.

In this thesis, a new approach is proposed for attribute reduction based on discernibility matrix using the type-2 fuzzy sets. Proposed algorithm is compared with the other attribute reduction algorithms in the literature using the data sets taken from UCI machine learning repository and in order to test the effectiveness of the proposed algorithm, we made use of classification algorithms.

Science Code : **906.1.148**
Key Words : **Atrribute reduction, Rough sets, Fuzzy sets, Fuzzy-rough sets.**
Page Number : **136**
Adviser : **Prof. Dr. Cevriye GENCER**

TEŞEKKÜR

Çalışmalarım boyunca değerli yardım ve katkılarıyla beni yönlendiren Tez Danışmanım Prof. Dr. Cevriye GENCER'e, yine Tez İzleme Komitelerinde kıymetli fikir ve yönlendirmeleri ile beni destekleyen Prof.Dr.Semra Oral ERBAŞ , Dr. Soner BİNAY'a, savunma jürimde bulunan hocalarım Doç.Dr. Aydın ULUCAN ve Yrd.Doç.Dr. Feyzan ARIKAN'a ve değerli fikirlerinden faydalandığım Yrd.Doç.Dr. Emel KIZILKAYA AYDOĞAN'a, ayrıca bana desteklerinden ötürü aileme teşekkürü bir borç bilirim.

İÇİNDEKİLER

Sayfa

Sayfa

Sayfa

ÇİZELGELERİN LİSTESİ

Çizelge **Sayfa**

ŞEKİLLERİN LİSTESİ

Şekil **Sayfa**

SİMGELER VE KISALTMALAR

Bu çalışmada kullanılmış kısaltmalar, açıklamaları ile birlikte aşağıda sunulmuştur.

Kısaltmalar	**Açıklama**
VTBK	Veritabanlarında bilgi keşfi
T_L	Łukasiewicz t-norm
UCI	Kaliforniya Üniversitesi
YSA	Yapay sinir ağları
PCA	Temel bileşenler analizi
ÇBÖ	Çok boyutlu ölçekleme analizi
T	t-norm
S	t-conorm
N	Ters işlem

1. GİRİŞ

Yakın senelerde büyük miktardaki verilere ulaşma imkânının kolaylaşmış olması; ticari ürünlerin birçoğunda kullanılan barkodlar, birçok hizmet ve ürünün artık bilgisayar tabanlı yapılması, bilimsel ve kamusal hizmetlerde bilgisayar kullanımının yaygınlaşması, üretim ve satış işlemlerinin internet üzerinden yapılması, bu veri kütlesinin kullanışlı bilgiye dönüştürülmesini gerekli kılmıştır. Ayrıca, küresel bir bilgi sistemi olan internet ve dolayısı ile web sayfaları bizlere inanılmaz boyutlarda bilgi ve veri kaynağı sağlamaktadır. Zamanla beraber depolanmış verilerde görülen büyüme eğilimi, veri yığınlarının kullanışlı bilgilere dönüştürülmesi için yeni tekniklerin kullanılması ihtiyacını ortaya koymaktadır.

İşlenmemiş veriden doğrudan faydalanmak neredeyse imkânsızdır. Verinin asıl değeri, yeni bir bilginin keşfi, karar verme konusunda yardımcı olması ve veri kaynağını yönetim faaliyetlerinde kullanılması için kullanışlı bilgilerin çıkarılması kabiliyeti ile ölçülür. Çoğu iş alanlarında veri analizi geleneksel olarak elle yapılmaktadır. Veri kümesi ile aşina olan analizciler, istatistiksel teknikler yardımı ile çeşitli özetlemeler yapmaktadırlar. Aslında böyle bir durumda analizi yapan kişi/kişiler sorgu işlemcisi görevini yürütmektedir. Bununla beraber, böyle bir yaklaşımın, verinin sürekli artışı karşısında yapabileceği şeyler sınırlıdır. Dolayısı ile insan kapasitesini aşan bu olumsuzluklar karşısında otomatik veri işleme araçlarının kullanılması kaçınılmaz hale gelmiştir.

Veritabanlarında bilgi keşfi (VTBK) sürecinin amacı, yüksek miktardaki verilerden kullanışlı bilginin çıkarılmasıdır. Böyle bir süreç hem etkileşimli hem de tekrarlı bir işlemdir [Manila, 1996]. Ancak, VTBK süreci, verinin sihirli bir kutudan geçirilip bilginin çıkarılması gibi basite indirgenebilecek bir işlem de değildir. VTBK sisteminin kullanıcısı, doğru veri altkümelerinin seçimi, uygun örüntülerin bulunması ve bulunan örüntülerin ilginçliğinin tespiti için, alan bilgisine sahip olmalıdır. Dolayısı ile veriden bilginin çıkarılması süreci birkaç işlemden oluşmaktadır. Genel bir ifade ile bu adımlar; araştırma alanının anlaşılması, verinin hazırlanması, veri

madenciliği, keşfedilen örüntülerin ön işlemesi, sonuçların uygulamaya konulması olarak sıralanabilir.

Araştırma alanının analizci tarafından anlaşılması veya ön bilgiye sahip olunması, elde edilecek bilginin değerlendirilmesi için faydalı olacaktır.

Verinin hazırlanması aşaması ise, veri kaynaklarının seçimini, verinin heterojen olarak bütünleştirilmesini, verinin toplanmasından kaynaklanan çeşitli hataların giderilmesini, gürültülü verinin temizlenmesini ve eksik verilerle ilgilenilmesi faaliyetlerini içermektedir.

Veri madenciliği, veritabanlarında bilgi keşfi kavramı içerisinde, ilginç ve daha önceden keşfedilmemiş kullanışlı bilginin çıkarılması olarak ifade edilmekte olup, istatistik, makine öğrenimi ile değişik matematiksel tekniklerin kullanıldığı bir disiplindir.

Veri madenciliğinde başarılı olmanın ilk şartı, hedefin ya da istenen çıktının açıkça belirlenmesidir. Sorunun tam olarak belirlenememesi, veritabanlarında bilgi keşfi süreçleri içerisinde öngörülmemiş bazı sorunların ve maliyetlerin çıkmasına sebep olabilecektir.

Tanımlanan problem için en uygun veri madenciliği modelinin bulunması olabildiğince çok sayıda modelin denenmesi ile mümkündür. Bu nedenle veri hazırlama ve modelin kurulması aşamaları en uygun olduğuna karar verilen modelin oluşturulmasına kadar yinelenen bir süreci kapsamaktadır [Kızılkaya, 2008].

Kalıpların keşfi işleminden sonra VTBK işlemi sonlandırılmaz. Araştırmacı neyin keşfedildiğini anlamalı, veriyi ve kalıpları aynı anda analiz ederek alan bilgisi ile karşılaştırmalıdır.

Düşük kalitedeki veri, veri madenciliği aşamasında kullanılan tekniğin başarısının da düşmesine sebep olmaktadır.

Bu olumsuzluğun giderilmesi için veri kalitesini arttıracak ve sadece ilgili verilerin bir veri madenciliği algoritmasına girmesini sağlayacak değişik metotlar mevcuttur. Bunlardan biri de nitelik seçimidir. Nitelik seçimi (indirgemesi) belirli bir ölçüt ile orijinal veri kümesini daha az nitelikle ifade etmek amacı ile geliştirilmiş metotlar ve yaklaşımlardır.

Mantık ve yapay zekânın temel kavramı olan muğlâklık ve belirsizlik kavramları uzun yıllar felsefe ve mantıkçılar tarafından araştırma konusu olmuştur. Son yıllarda, özellikle yapay zekâ ile ilgilenen bilgisayar bilimcileri tarafından bu araştırma alanına temel katkılar sağlamışlardır. Muğlâklık kavramı; eldeki mevcut bilgiye bağlı olarak örneklerin bir kavrama ya da tümleyenine atanamadığı bir sınır durumunu ifade etmektedir. Diğer yandan belirsizlik ise, gözlem hakkında tam veya kesin bilginin olmayışından kaynaklanmakta, gözlenen değerin bir kümeye ne kadar ait olduğu ile ilgilenmektedir.

Belirsizliğin modellenmesi, son yıllarda bilginin temsili alanında önemli bir araştırma sahası olmuştur. Zadeh'in [1965] bulanık küme teorisi ve Pawlak'ın [1982] kaba küme teorisi gibi birçok yaklaşım belirsizlik problemini işaret etmektedir. Bulanık kümeler ve kaba kümeler, belirsizliğin modellenmesi için klasik küme teorisinin genelleştirilmesidir. Bu iki teori farklı, fakat ilişkili ve birbirini tamamlar niteliktedir [Tsang ve ark. 2005].

Tezde geliştirilen algoritma bir nitelik seçimi algoritmasıdır.

Tezin amacı, bulanık-kaba küme teorisi kullanılarak veri kümesi içerisinde nitelik indirgemesi yapacak bir algoritma tasarlamaktır. Tezde, nitelik indirgeme konusu, teorik temelde anlatılarak kullanılan teknikler gösterilmiş; kaba küme, bulanık küme ve bulanık-kaba küme melezleri ile yapılan nitelik indirgeme çalışmaları genel olarak tanıtılarak elde edilen bilgilerle bir veri kümesi içerisinde tüm indirgemeleri bulacak bir algoritma tasarlanmıştır. Algoritma, öncelikle sürekli değer alan nitelikleri bulanıklaştırmakta, elde edilen bulanık nitelikler ile kendi arasında benzerlik ilişkilerini bulmakta, elde edilen benzerlik ilişkileri belirlenen bir bütünleştirme

operatörüne göre birleştirmekte ve Łukasiewicz t-norm (T_L) çatısında ayırtedilebilirlik matrisi yardımı ile indirgemeler bulmaktadır.

Tezin birinci bölümü; bulanık kaba kümeler yardımıyla yapılan nitelik indirgeme çalışmasına esas olmak üzere genel bilgiler vermektedir. İkinci bölümde; veri madenciliği kavramı ve veri madenciliği metotları hakkında genel bilgi sunulmakta, veri madenciliği ve makine öğrenimi içerisinde bilinen boyutsal (nitelik) indirgeme problemi anlatılmakta, bu kapsamda kullanılan çeşitli nitelik indirgeme teknikleri açıklanmaktadır. Üçüncü bölümde; bulanık-kaba küme teorisinin daha iyi anlaşılabilmesi için kaba kümeler, bulanık kümeler ve nitelik indirgeme yöntemleri hakkında bilgi verilmekte, bulanık-kaba küme teorisi anlatılarak nitelik indirgemede kullanımı açıklanmaktadır. Dördüncü bölümde; önceki anlatılan konular ışığında tip–2 bulanık kümelerde nitelik indirgemede kullanılabilecek bir algoritma önerilmiş ve bir örnek çözülerek görselleştirilmiştir.

Önerilen algoritma UCI makine öğrenimi deposundan alınan on veri tabanı üzerinde denenmiş, müteakiben elde edilen sonuçların performansı diğer nitelik indirgeme yaklaşımları ile karşılaştırılmıştır.

Beşinci bölümde yapılan çalışmalar kısaca özetlenerek gelecek dönemde yapılabilecek çalışmalar sunulmuştur.

2. VERİ TABANLARINDA BİLGİ KEŞFİ ve ÖN İŞLEME

Veri, bir bilgisayar tarafından işlenebilen herhangi bir olgu, sayı veya metindir. Bu anlamda veri, enformasyonun da temelini oluşturmaktadır. Ayrıca veriler; satışlar, maliyetler, stoklar veya banka hesapları gibi işlemsel ya da tahmin verileri şeklinde ya da makroekonomik veriler gibi işlemsel olmayan şekilde karşımıza çıkabilmektedir. Meta-veri ise veri sözlüğü yapıları ya da mantıksal veri tabanı tasarımı gibi faaliyetlerde karşımıza çıkan verinin kendisi hakkındaki veridir.

Enformasyon, veriler arasındaki kalıplar, eşlemeler ya da ilişkilerdir. Diğer bir deyişle, ilgili verilerin bir araya getirilmesinden oluşur. Örneğin, satış noktaları (POS cihazları) işlemleri ne zaman hangi ürünlerin satıldığı hakkında enformasyon verebilir.

Bilgi ise, verinin işlemsel faaliyetlere dönüştürülmesi için kullanıcıya öz ama zengin altyapı sağlamaktadır. Tarihsel kalıplar ve gelecek eğilimleri hakkındaki enformasyon bilgiye çevrilir. Örneğin, bir süpermarket satışları hakkındaki enformasyon, indirimli ürünlerin zamanlaması ya da müşterilerin satın alma alışkanlıkları hakkında bilgiye dönüştürülebilir.

2.1. Veri Tabanlarında Bilgi Keşfi

Aktif araştırma alanlarından biri olan Veri Tabanlarından Bilgi Keşfi (VTBK) disiplini, çok büyük oylumlu verileri tam veya yarı otomatik bir biçimde analiz eden yeni kuşak araç ve tekniklerin üretilmesi ile ilgilenen son yılların gözde araştırma konularından biridir. Veri madenciliği ise, önceden bilinmeyen, veri içinde gizli, anlamlı ve yararlı örüntülerin (kalıpların) büyük ölçekli veri tabanlarından otomatik biçimde elde edilmesini sağlayan VTBK süreci içinde bir adımdır [Sever ve Oğuz, 2002].

Veri tabanlarında bilgi keşfi ihtiyacı;

1. Veri toplama ve depolama donanımlarının gelişimi sayesinde veri dağlarının oluşumu,

2. Veri depolamanın önemli bir maliyet kalemi haline gelmesi,

3. Veri tabanı yönetim sistemlerinin gelişmesi,

4. Bilgisayar teknolojisinde paralel mimari yapılarının da kullanılması ile hesaplama etkinliğinin fazlalaşması,

5. Otomatik öğrenme tekniklerinin sürekli gelişimi,

6. Veri toplamada karşılaşılabilecek muhtemel yanlışlıklar ve/veya belirsizliklerle başa çıkma gerekliliğinden doğmuştur.

VTBK'ni oluşturan süreçler [Sever ve Oğuz, 2002];

1. Veri temizleme : Kısaca verinin ayıklanması olarak tanımlanabilen işlem, tutarlılık kontrolü, veri hatalarının düzeltilmesi, boş değerlerin doldurulması veya silinmesi, kendini tekrar eden kayıtların silinmesi gibi işlemleri içerir. Bu işlem tamamen uygulama alanına bağımlıdır ve veritabanı kullanıcısının yardımına ihtiyaç duymaktadır [Freitas, 1997].

2. Veri bütünleştirme : Çoklu veri kaynaklarından ve veritabanlarından verinin bütünlük ve güvenilirlik özellikleri korunarak birleştirilmesini ifade eder [Lenzerini, 2002].

3. Veri seçimi : Analiz görevi ile ilgili verinin veri tabanından çıkarılmasıdır. Diğer bir ifade ile veri madenciliği araçlarının ve hedeflerinin seçilmesi ile analiz için uygun girdi niteliklerinin seçilmesi olarak tanımlanabilir.

4. Veri dönüşümü : Düzeltme, bütünleştirme, yüksek seviye kavramlarla açıklama, belirli bir aralığa indirgeme ve nitelik kurulumu gibi işlemlerle verinin veri madenciliği adımına hazırlanması işlemlerini ifade etmektedir.

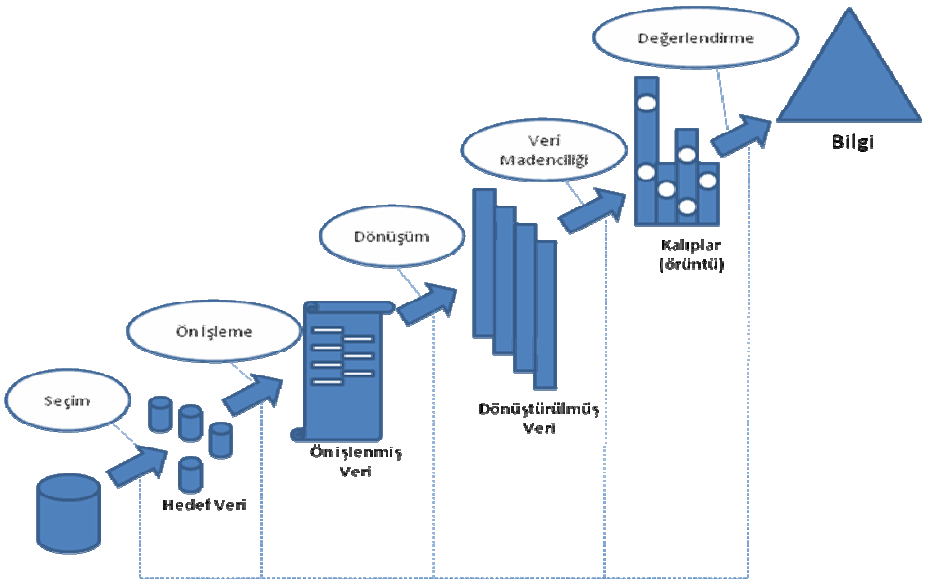
5. Veri madenciliği : Veri kalıplarının çıkarılması için zeki metotların kullanıldığı ana adımdır.

6. Kalıp değerlendirme : Belli bir ilginçlik ölçütü içinde bilgiyi sunan doğru ve ilginç kalıpların tanımlanmasıdır.

7. Bilginin sunumu : Görselleştirme ve bilgi gösterim teknikleri kullanılarak çıkarılan bilginin kullanıcıya tanıtılması olarak tanımlanabilir.

Genelde 1–4 ncü süreçlere "ön işleme" 6 ve 7'nci süreçlere ise "sonra işleme" denilmektedir [Bruha, 2000].

Veri tabanlarında bilgi keşfi süreçleri Şekil 2. 1 de verilmektedir.



Şekil 2.1. Veritabanlarında bilgi keşfi süreçleri [Han ve Kamber, 2001]

Veri madenciliği süreci kullanıcı veya bilgi tabanı ile etkileşim içerisinde bulunabilir. İlginç kalıplar kullanıcıya sunulur ve bilgi tabanına yeni bir bilgi olarak kaydedilebilir.

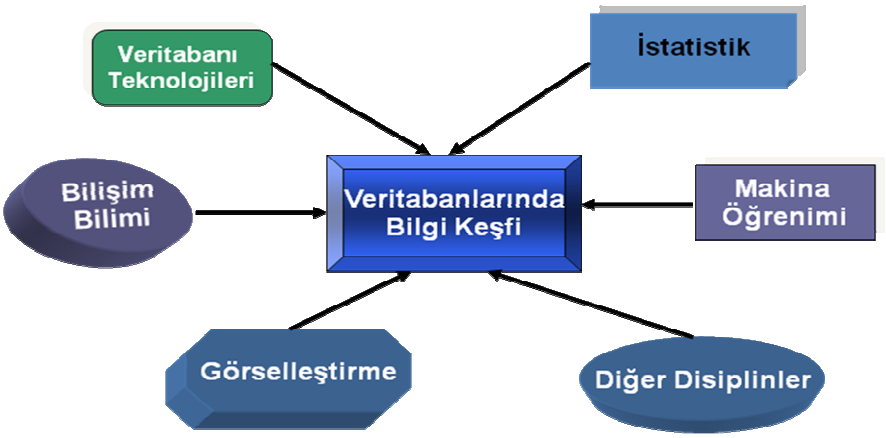
Bu bağlamda, veri madenciliği bütün işlemlerin arasında sadece bir adımdır. Tarihsel olarak veriden kullanışlı kalıplar bulmak işlemler bütünü için; veri madenciliği, enformasyon hasadı, veri arkeolojisi ve veri örüntü işleme gibi bir çok isim verilmiştir [Fayyad ve ark., 1996]. Veri madenciliği terimi ise en çok istatistikçiler, veri analizcileri ve yönetim bilişim sistemleri toplulukları tarafından kabul görmüştür. Veri tabanlarında bilgi keşfi kavramı ise ilk defa 1989 yılında VTBK çalıştayında bilginin, veri yönelimli keşfin son ürünü olduğunu ifade etmek için ortaya çıkmıştır [Piatetsky-Shapiro, 1991].

2.1.1. VTBK'nın diğer bilimlerle ilişkileri

VTBK açıkça disiplinler arası bir araştırma alanı olup, birçok bilimsel disiplinin ortak alanıdır. Bu disiplinler, VTBK teknikleri ile ilgili ve veri ile ilgili olmak üzere ikiye ayrılmaktadır [Han ve Kamber, 2001].

VTBK teknikleri ile ilgili olanlar, bir yapay zekâ alanı olan makine öğrenimi teknikleri [Langley,1992] ile özellikle istatistiksel örüntü tanıma ve keşif amaçlı veri analizi ile ilgilenen istatistik bilimidir [Lee ve ark.,1995]. Bunun haricinde VTBK ile ilgili diğer bir alan ise veri görselleştirmedir. Dolayısı ile VTBK işlemleri genel olarak; veri tabanı teknolojisi, istatistik, makine öğrenimi, akıllı bilgi sistemleri, veri ambarı ve uzman sistemlerde bilgi çıkarımı gibi birçok bilimsel disiplinle ortak çalışmak zorundadır. Şekil 2.2'de VTBK ve diğer disiplinlerin ilişkileri görsel olarak sunulmaktadır.

Veri ile ilgili olanlar ise, veri madenciliği uygulamasına göre görüntü analizi, sinyal işleme, biyo-enformasyon veya psikoloji gibi bilimlerde kullanılan teknikleri bütünleştirmektedir.



Şekil 2.2. Çoklu disiplinlerle kesişim noktası olarak VTBK.

Bununla beraber VTBK süreci kullanıldığı alana bağlı olarak, diğer disiplinlerde uygulama alanı bulabilmiştir. Bunlar; yapay sinir ağları, genetik algoritmalar, genetik programlama, karınca kolonisi yaklaşımı, bulanık kümeler, kaba kümeler ya da melezleri, bilgi sunumu, çıkarımsal mantık programlama veya yüksek performanslı hesaplama alanlarıdır.

2.1.2. Veri madenciliği sistemlerinin sınıflandırılması

Kullanılan veri, veri madenciliği sistemini yakından ilgilendirmektedir. Veri madenciliği sistemleri; uzaysal veri madenciliği, enformasyon çıkarımı, örüntü tanıma, görüntü analizi, sinyal işleme, web teknolojileri, ekonomi, iş yeri yönetimi, biyo-enformasyon ya da psikoloji alanlarında, uygulama alanının kendine has özelliklerini içeren verilerle de ilgilenmekte, dolayısı ile bu sistemlere uygun teknikler de geliştirilmektedir. Bu nedenle, veri madenciliği sistemlerin açık bir sınıflandırmasının yapılmasına ihtiyaç duyulmaktadır. Böyle bir sınıflandırma veri madenciliği tekniklerini kullanacak olanlara hem bu sistemlerin ayrımını yapmada hem de kendi ihtiyaçlarını en iyi karşılayan sistemlerin tanımlamasına yardımcı olabilecektir [Han ve Kamber, 2001].

Veritabanına bağlı olarak sınıflandırma

İlişkisel, işlemsel, veri ambarı gibi değişik veritabanı sistemleri farklı ölçütlere göre sınıflandırılabilir. Burada bahsedilenlerin tümü için kendi veri madenciliği modelleri oluşturulmaktadır.

Bilgi çeşidine göre sınıflandırma

Veri madenciliği işlevleri bakımından; sınıflandırma ve regresyon modelleri, kümeleme modelleri, birliktelik kuralları, ardışık zamanlı örüntüler, bellek tabanlı yöntemler gibi bilgi çeşitlerine göre de sınıflandırılır. Bununla beraber, veri madenciliği sistemleri çıkarılacak bilginin çoklu seviye bilgilerini ifade eden enformasyonun tanecikli yapısına veya soyutlama seviyelerine de bağlıdır.

Kullanılan tekniklere göre sınıflandırma

Veri madenciliği sistemleri kullanılan veri madenciliği tekniklerine göre de sınıflandırılabilir. Otonom sistemler, etkileşimli keşif sistemleri gibi kullanıcı ile etkileşim derecesine göre ve veri ambarı yönelimli teknikler, makine öğrenimi, istatistik, örüntü tanıma, yapay sinir ağları gibi kullanılan veri analizi metotlarına göre de sınıflandırılabilir.

Uygulama alanına göre sınıflandırma

Veri madenciliği sistemleri; maliye, iletişim, DNA, hisse senetleri, e-posta gibi uyarlanmış alanlara göre de sınıflandırılabilir. Değişik uygulamalar genelde ilgilenilen alana özgü değişik tekniklerin de beraber kullanılmasını gerektirmektedir.

2.1.3. Veri madenciliğinde kullanılan yöntemler

Veri madenciliğinde kullanılan modeller, tahmin edici ve tanımlayıcı olmak üzere iki ana başlık altında toplanabilir.

Tahmin edici modellerde; sonuçları bilinen verilerden hareket edilerek bir model geliştirilmesi ve kurulan bu modelden yararlanılarak sonuçları bilinmeyen veri kümeleri için sonuç değerlerinin tahmin edilmesi amaçlanmaktadır.

Tanımlayıcı modeller; karar vermeye rehberlik etmede kullanılabilecek mevcut verilerdeki örüntülerin tanımlanmasını sağlamaktadır.

Gerek tanımlayıcı gerekse tahmin edici modellerde yoğun olarak kullanılan belli başlı yöntemleri; sınıflandırma ve regresyon, kümeleme, birliktelik kuralları ve ardışık zamanlı örüntüler, bellek tabanlı yöntemler, yapay sinir ağları ve karar ağaçları olmak üzere altı ana başlık altında incelemek mümkündür [Akpınar, 2000].

Sınıflandırma ve Regresyon modelleri;

Mevcut verilerden hareket ederek geleceğin tahmin edilmesinde faydalanılan ve veri madenciliği teknikleri içerisinde en yaygın kullanıma sahip olan sınıflama ve regresyon modelleri arasındaki temel fark, tahmin edilen bağımlı değişkenin kategorik veya süreklilik gösteren bir değere sahip olmasıdır. Ancak çok terimli lojistik regresyon gibi kategorik değerlerin de tahmin edilmesine olanak sağlayan tekniklerle, her iki model giderek birbirine yaklaşmakta ve bunun bir sonucu olarak aynı tekniklerden yararlanılması mümkün olmaktadır. Sınıflama ve regresyon modellerinde kullanılan başlıca teknikler;

- Genetik Algoritmalar,
- K-En Yakın Komşu,
- Naïve-Bayes,
- Çoklu ve Lojistik Regresyon,
- Faktör ve Ayırma analizleridir.

Kümeleme modelleri;

Kümeleme modellerinde amaç, üyelerin birbirlerine çok benzediği, ancak özellikleri birbirlerinden çok farklı olan kümelerin bulunması ve veri tabanındaki kayıtların bu farklı kümelere bölünmesidir. Kümeleme analizinde; veri tabanındaki kayıtların hangi kümelere ayrılacağı veya kümelemenin hangi değişken özelliklerine göre yapılacağı hususu, konunun uzmanı olan bir kişi tarafından belirtilebileceği gibi, veri tabanındaki kayıtların kümelere ayrılması işlemi geliştirilen bilgisayar programları tarafından da yapabilmektedir.

Birliktelik kuralları ve Ardışık zamanlı örüntüler;

Bir alışveriş sırasında veya birbirini izleyen alışverişlerde müşterinin hangi mal veya hizmetleri satın almaya eğilimli olduğunun belirlenmesi, müşteriye daha fazla ürünün satılmasını sağlama yollarından biridir. Satın alma eğilimlerinin tanımlanmasını sağlayan birliktelik kuralları ve ardışık zamanlı örüntüler, pazarlama amaçlı olarak pazar sepeti analizi adı altında veri madenciliğinde yaygın olarak kullanılmaktadır. Bununla birlikte bu teknikler, tıp, finans ve farklı olayların birbirleri ile ilişkili olduğunun belirlenmesi gibi alanlara uygulanmakta ve sonucunda değerli bilgi kazanımları elde edilebilmektedir.

Bellek tabanlı yöntemler;

Bu yöntem, istatistikte 1950'li yıllarda önerilmiş olmasına rağmen o yıllarda gerektirdiği hesaplama ve bellek yüzünden kullanılamamış ancak günümüzde bilgisayarların ucuzlaması ve kapasitelerinin artmasıyla, özellikle de çok işlemcili sistemlerin yaygınlaşmasıyla, kullanılabilir olmuştur. Bu yöntemlere en iyi örnek en yakın k-komşu algoritmasıdır.

Yapay sinir ağları;

Yapay sinir ağlarında (YSA) kullanılan öğrenme algoritmaları, veriden üniteler arasındaki bağlantı ağırlıklarını hesaplar. YSA'lar istatistiksel yöntemler gibi veri hakkında parametrik bir model varsaymaz yani uygulama alanı daha geniştir ve bellek tabanlı yöntemler kadar yüksek işlem ve bellek gerektirmez.

Karar ağaçları;

İstatistiksel yöntemlerde veya YSA'nda veriden bir fonksiyon öğrenildikten sonra bu fonksiyonun insanlar tarafından anlaşılabilecek bir kural olarak yorumlanması zordur. Karar ağaçları ise ağaç oluşturulduktan sonra, kökten yaprağa doğru inilerek kurallar yazılabilir. Bu şekilde kural çıkarma veri madenciliği çalışmasının sonucunun doğrulanmasını sağlar. Bu kurallar uygulama konusunda uzman bir karar vericiye gösterilerek sonucun anlamlı olup olmadığı denetlenebilir. Sonradan başka bir teknik kullanılacak bile olsa karar ağacı ile önce bir kısa çalışma yapmak, önemli değişkenler ve yaklaşık kurallar konusunda karar vericiye bilgi verir.

2.2. VTBK Sürecinde Önişleme

Çoğu VTBK süreci pratik uygulamalarında, özellikle verilerin toplanması ve veri madenciliği aşamasında hangi verilerin önemli ve ilgili olduğu tam olarak belirlenemez. Böylece kullanışlı olacağı değerlendirilen veriler veritabanında toplanır. Dolayısı ile veritabanları aslında istenmeyen, ilgisiz ve keşif amacına yönelik olmayan veya belirli bir amaca yönelmemiş verilerden oluşur. Düşük kalitedeki veri, veri madenciliği aşamasında kullanılan tekniğin başarısının da düşmesine sebep olmaktadır [Han ve Kamber, 2001] .

Veri madenciliği sistemlerin karşı karşıya kaldığı en büyük sorunlardan biri de veritabanı boyutudur. Dolayısı ile veri madenciliği teknikleri ya sezgisel bir yaklaşımla arama uzayını taramalı, ya da veri tabanını yatay ve dikey olarak indirgemelidir [Sever ve Oğuz, 2002].

Veri önişleme aşaması için birçok teknik mevcuttur. Verinin temizlenmesi, veri içerisindeki gürültünün atılması ve tutarsızlıkların düzeltilmesi için kullanılır. Veri bütünleştirme ise çoklu kaynaktaki verileri, veri ambarı gibi tutarlı bir veri deposunda birleştirir. Bütünleştirme için normalleştirme gibi veri dönüşümü teknikleri de uygulanabilir. Veri indirgeme işlemi, verinin boyutunu; veri bütünleştirme, gereksiz nitelikleri ayıklama veya kümeleme gibi metotlarla azaltır. Bu teknikler tek başına kullanılabildiği gibi birbiri ile de kullanılabilir [Han ve Kamber, 2001].

Veri madenciliği aşamasında verinin toplanması ile ilgili bir takım problemler bulunabilir. Bunlar kısaca; gürültülü veri, boş değerler, eksik veri, artık veri, dinamik veri ve farklı tiplerdeki verilerdir.

2.2.1. VTBK'de verilerin temizlemesi

Gerçek dünyada veriler; gürültülü, boş ve/veya eksik olabilir.

Gürültülü veri;

Veri girişi veya veri toplanması sırasında oluşan sistem dışı hatalara gürültü adı verilir [Quinlan, 1986]. Büyük veri tabanlarında ölçülen değişken içerisindeki pek çok niteliğin değeri yanlış olabilir. Bu hata, veri girişi sırasında yapılan insan hataları veya girilen değerin yanlış ölçülmesinden kaynaklanabildiği gibi veritabanı tasarımcısının yanlış eşik değerler vermesinden de kaynaklanabilir. Örneğin boyu 1.65 m'den büyük olanların uzun kabul edilmesi gibi.

Quinlan [1986], gürültülü veri ile başa çıkabilmek için her niteliğe rassal seçilmiş belirli bir yüzde oranında değer ekleyerek gürültünün veri madenciliği algoritmasına etkisi üzerinde çalışmıştır. Benzer bir çalışma ise, Arditi ve Pulket [2007] tarafından yapılmıştır. Çalışmalarında, gürültülü ve gürültüsüz verilerin sınıflandırma doğruluğu karşılaştırılmış ve en iyi ihtimalle gürültülü bir veride gürültüsüz veriye oranla %12'lik bir fark olduğunu göstermişlerdir.

Gürültülü verinin ayıklanması için veri gruplandırma, kümeleme ve insan-bilgisayar etkileşimi gibi bazı teknikler kullanılabilir.

Veri gruplandırma tekniğinde; veriler kendi aralarında sıralanır. Eşit sayıda eleman içeren kümelere ayrılır ve kümelerin ortalaması o kümeye denk gelen değerlere atanır. Örneğin {4,8,15} kümesinde gözlenen değerlere 9 değeri verilir [Han ve Kamber, 2001]. Kümeleme tekniğinde; kümeleme yolu ile benzer değerler gruplara ayrılırken, bu kümelerin dışında kalan değerler aykırı değer olarak kabul edilir. İnsan bilgisayar etkileşimli teknikte ise, operatör bilgisayarın tespit ettiği aykırı gözlemlerin listesinden kendince işe yaramayan değerleri ayıklayabilir. Diğer bir metot ise veriyi bir regresyon fonksiyonuna uydurarak düzlemedir [Labib ve Malek, 2005].

Boş Değerler;

Boş değer tanımı gereği kendisi de dahil olmak üzere hiçbir değere eşit olmayan değerdir. Bir veri tabanında nitelik değeri boş ise o nitelik ya bilinmeyen ya da uygulanamaz bir değere sahiptir.

Boş değerlerle başa çıkmak için veri değerini elle doldurmak, o gözlemin tamamının silinmesi, boş değerlere sabit bir değer atamak, nitelik ortalamalarının atanması, aynı sınıfa ait örneklerin ortalamalarının atanması, olası bir değerin atanması teknikleri kullanılmaktadır [Scheffer, 2002].

Eksik veri;

Evrendeki her nesnenin ayrıntılı bir biçimde tanımlandığı ve bu nesnelerin alabileceği değerler kümesinin belirli olduğu varsayılsın. Verilen her bir bilgi sisteminde her bir nesnenin tanımı kesin ve yeterli olsa idi sınıflandırma işlemi basitçe nesnelerin alt kümelerinden faydalanarak yapılabilirdi. Ancak pratik uygulamalarda, veriler kurum ihtiyaçları doğrultusunda toplandığından, mevcut veri

bilgi keşfi açısından uygun olmayabilir [Sever ve Oğuz, 2002]. Eksik veri ile karşılaşıldığında veriyi içeren gözlemin silinmesi tekniği kullanılmaktadır.

2.2.2. Boyutsal (nitelik) indirgeme

Analiz edilecek veri kümeleri, birçoğu veri madenciliğinde kullanılmayacak gereksiz yüzlerce nitelikten oluşabilir. Gereksiz niteliklerin atılması ile düşük kaliteli kalıpların keşfi bir ölçüde engellenir. Bununla beraber, bu tip gereksiz nitelikler veri madenciliği algoritmasının yavaşlamasına sebep olabilmektedir.

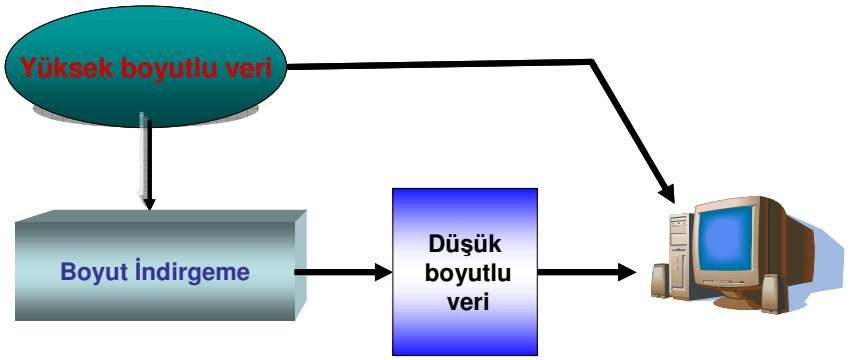
Müşteri ilişki yönetimi, borsa veri analizi gibi birçok bilgi keşfi işlemlerinde, ana amaç, odaklanılan konu ile ilgili ve en az onun kadar değerli olabilecek nitelik alt kümesinin bulunması işlemidir [Ortega, 2000]. Bu türlü bir işlem, gerçek dünyada verilerin, veri madenciliği için toplanmamasından kaynaklanmaktadır. Milyonlarca kaydı taşıyan niteliklerin çok azının karar kuralı olarak karşımıza çıkması, veri madenciliği uygulaması yapılmadan önce verinin hazırlanmasının gerekliliğini ortaya koymaktadır. Dolayısı ile veritabanındaki gereksiz verilerin indirgenmesi ve/veya veri kütlesi içerisindeki sadece kullanılacak niteliklerin ayırt edilmesi işlemi aslında veri madenciliği algoritmalarının etkinliğinin arttırılması için gerekli bir ön işlem olarak yapılmalıdır.

Genelde kullanışlı, bazen de zorunlu olarak veri boyutunun en az veri kaybıyla yönetilebilir hale getirilmesi gerekmektedir. Boyutsal (nitelik) indirgeme ile veri kümesi boyutu gereksiz nitelikler atılması suretiyle küçültülür. Bu işlem Şekil 2.3'de gösterilmiştir. Ayrıca, bu kısımdan itibaren boyutsal indirgeme yerine nitelik indirgeme ifadesi kullanılacaktır.

Sıklıkla, verinin orijinal hali iki sebepten dolayı gereksiz hale gelmektedir;

1. Birçok nitelik gürültünün varyansından daha küçük varyansa sahip olmakta ve gereksiz hale gelmektedir.

2. Birçok nitelik birbiri ile (lineer kombinasyonlar ya da diğer fonksiyona bağımlılık yolu ile) ilişkilidir. Birbiri ile ilişkisi olmayan yeni nitelikler bulunmalıdır [Miguel, 1997].



Şekil 2.3. Boyutsal (nitelik) indirgeme

Matematiksel olarak durum şu şekilde özetlenebilir; p-boyutlu rassal değişken $x = (x_1,....x_p)^T$ olsun. Bunun için orijinal verinin içeriğini koruyacak şekilde herhangi bir kriter yardımıyla $k < p$ sağlanarak $s = (s_1,....s_k)^T$ şekliyle ifade etmeye boyutsal indirgeme denilmektedir.

Burada, s'nin bileşenlerine bazen gizli bileşen de denilmektedir. p'nin ise birçok ismi olduğu gibi genelde "değişken" ya da "nitelik" de denilmektedir [Fodor, 2002].

Nitelik indirgeme problemlerine bakış açıları

Yüksek boyutlu problemlerin bütün bilimsel alanlarda görüldüğü değerlendirildiğinde birçok bilimsel yaklaşımının da boyutsal indirgeme yaklaşımlarını kullanması doğaldır. Bu anlamda nitelik indirgeme kavramına bazı açılardan bakmak doğru olacaktır.

1. Temel olarak nitelik indirgeme problemi $D > L$ olacak şekilde D-boyutlu bir uzayın L boyuta bir izdüşümü olarak nitelendirilebilir.

2. Nitelik indirgeme, istatistikte, çok değişkenli yoğunluk tahmini, regresyon ya da düzleme teknikleri ile ilgilidir.

3. Örüntü tanıma açısından nitelik çıkarma ile eşdeğerdir.

4. Enformasyon teorisi açısından, veri sıkıştırma ve dönüşümü ile ilgilidir.

5. Çoğu görselleştirme tekniklerinde, (Çok boyutlu ölçekleme analizi [Kruskal, 1964] ve Sammon eşlemesi [Sammon, 1969] gibi) nitelik indirgeme yapılmaktadır.

6. Karmaşıklık indirgemesi; eğer bir algoritmanın zamansal ya da hesapsal olarak karmaşıklığı kendi girdilerinin boyutuna bağımlı ise, boyutun azaltılması algoritmayı daha etkin yapacaktır.

Nitelik indirgeme problemi sınıfları

Nitelik indirgeme problemleri üç kategoride değerlendirilebilir [Jensen, 2005]:

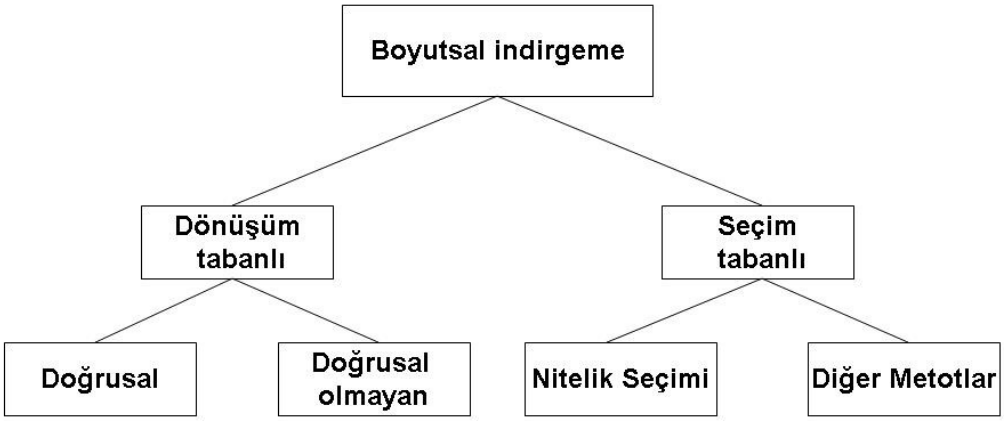
1. Fazla sayıdaki niteliklerin indirgenmesi için kullanılan metotlar; Binlerce bileşenden oluşan boyutsallığın azaltılması için kullanılır. Bu metotlar genelde yüz tanıma, karakter tanıma veya konuşmacı tanıma gibi örüntü tanıma problemlerinde sıklıkla kullanılmaktadır. Tipik metotlar temel bileşenler analizi ve kaba küme yaklaşımlarıdır [Dong ve ark., 1999].

2. Az sayıdaki niteliklerin indirgenmesi için kullanılan metotlar; Burada veri genelde 20–30 civarında bileşenden oluşur. Çoğu sosyal bilimler ve psikoloji gibi alanlardaki istatistiksel analiz metotları bu sınıfa düşmektedir. Sıklıkla kullanılan teknikler; PCA, faktör analizi ve çok boyutlu ölçekleme analizidir.

3. Görselleştirme problemleri; Burada, veri aslında çok boyutlu değildir ancak veriyi görselleştirmek için 2,3 veya 4 boyuta indirgenmelidir. Bu metotlar; "Projeksiyon izleme" ve "Çok boyutlu ölçekleme analizidir [Fodor, 2002].

Nitelik indirgeme probleminin tasnifi;

Nitelik indirgeme; verideki anlamsal bütünlüğü bozup bozmamasına göre dönüşüm ve seçim tabanlı olmak üzere iki şekilde incelenebilir. Nitelik indirgeme probleminin tasnifi şekil 2.4'de sunulmuştur.



Şekil 2.4. Boyutsal indirgeme probleminin tasnifi [Jensen, 2005]

Dönüşüm Tabanlı Nitelik İndirgeme

Bu indirgeme, orijinal veri kümesinin anlamına ileride yapılacak işlemlerde ihtiyaç olmadığında kullanılabilir. Doğrusal ve doğrusal olmayan şekilde ikiye ayrılırlar. Dönüşüm tabanlı metotlarda temel olarak iki yaklaşımdan bahsedilebilir. Bunlar nitelik çıkarma ve nitelik kurulumudur.

Dönüşüm Tabanlı İndirgemede kullanılan yaklaşımlar;

Nitelik Çıkarma Yaklaşımı

Nitelik çıkarma, yeni nitelik kümesinin orijinal nitelik kümesinden bir takım işlemler sonucu türetilmesi işlemidir [Wyse ve ark, 1980]. Nitelik çıkartmanın asıl amacı belirli performans ölçütlerine göre bazı dönüşüm metotları kullanarak minimum yeni nitelik kümesi üretmektir.

Burada bahsedilen performans ölçütleri, yeni çıkarılan niteliğin değerlendirilmesinde optimumu araştırır. Performans değerlendirilmesinin ana amacı ise, dönüşüm işleminden sonra orijinal niteliklerin korunmasını garanti etmektir. Sınıflandırma veya kümeleme gibi veri madenciliği uygulamaları, performans ölçütlerini belirlemekte olup, sınıflandırma için eğitim kümesinin tahmin doğruluğu, kümeleme için ise kümeler arası ya da küme içi benzerlikler, verilerdeki varyans gibi ölçütler kullanılmaktadır. Nitelik çıkartma için, ileri beslemeli sinir ağları, temel bileşenler analizi, gibi teknikler kullanılmaktadır.

Nitelik Kurulumu Yaklaşımı

Nitelik kurulumu, nitelikler arası bilinmeyen ilişkileri keşfeden, nitelik uzayını ek nitelikler yaratarak veya bunlardan sonuç çıkararak arttıran bir işlem olarak tanımlanır [Liu ve Motoda, 1998]. Buna basit bir örnek olarak n niteliğe sahip bir veri kümesi olsun. Burada, $A_1, A_2, ..., A_n$ dir. Nitelik kurulumu işleminden sonra, m adet ek nitelik çıkarılabilir. $n < k \leq n + m$ olacak şekilde yeni bir A_k niteliği, orijinal veri kümesindeki A_i ve A_j niteliklerinden bir işlemle kurulabilir. Örneğin, A_1 genişlik ve A_2 uzunluk olsun. Böyle bir problem B_1 alan olacak şekilde tek boyutlu hale dönüştürülebilir.

Bütün yeni kurulmuş nitelikler, orijinal nitelikler açısından tanımlanabilir. Nitelik kurulumu işleminde dışarıdan herhangi ek bir bilgi eklenmez, sadece orijinal niteliklerin anlamı arttırılmaya çalışılır. Genelde yeni nitelik kümesi orijinal nitelik kümesinden daha geniştir. Aslında bu durum nitelik kurulumunun bir dezavantajı olarak görülse de yeni nitelikler kurulduktan sonra birçok eski nitelik gereksiz hale gelmektedir. Orijinal niteliklerden yeni nitelikler çıkarma işlemi birçok nitelik kombinasyonunu gözden geçirmeyi gerektirdiğinden elle hesaplanması zordur. Nitelik kurulumu; iyileştirilmiş doğruluk, kolay anlaşılabilirlik, gerçekçi kümeleme, gizli kalıpların ortaya çıkarılması gibi veri madenciliği hedeflerine yardımcı olmaktadır.

Nitelik kurulumu konusunda dört ana yaklaşımdan bahsedilebilir. Bunlar; veri yönelimli, hipotez yönelimli bilgi tabanlı ve bu tekniklerin melezleridir. Veri yönelimli yaklaşım, çeşitli operatörler yardımıyla eldeki verinin analizine dayanarak yeni nitelikler oluşturur. Hipotez yönelimli yaklaşım ise, önceden oluşturulmuş hipotezlere dayanan yeni nitelikler oluşturmak için kullanılır. Oluşturulan hipotezlerden çıkarılan kullanışlı kavramlar yeni niteliklerin oluşturulması için kullanılır. Bilgi tabanlı yaklaşımda ise, yeni bileşik niteliklerin tipinin belirlenmesi ve uygun operatörlerin seçimini sağlayan alan bilgisinin kullanımı söz konusudur.

Bileşik niteliklerin oluşturulması için niteliklerin bileşkesi için birçok operatör bulunmaktadır. Veri tipleri her bir operatörün uygunluğu için önemli bir rol oynar. Nominal değerler için kesişim birleşim ve ters işlem operatörleri, sürekli değerler için ise toplama, çıkarma, çarpma, bölme, minimum, maksimum gibi operatörler kullanılmaktadır. Bütün operatörlerin kullanımı mümkün olmadığından, potansiyel olarak kullanışlı operatörlerin sezgisel olarak kullanımı sağlanmaktadır. Yeni oluşturulan niteliklerin tümü kullanışlı değildir. Eski ve yeni niteliklerin tümünün kullanımı boyutsal olarak çözümü zor problemler haline gelebilecektir. Bu kapsamda nitelik seçimi algoritmaları kullanılabilmektedir.

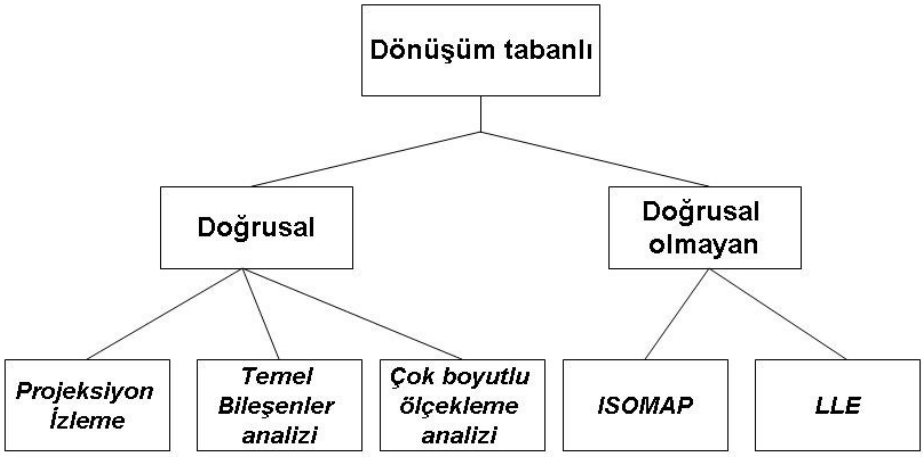
Dönüşüm Tabanlı Nitelik indirgemede kullanılan teknikler;

Dönüşüm tabanlı tekniklerin sınıflandırılması Şekil 2.5'de verilmiştir.

Doğrusal metotlar

Projeksiyon izleme, temel bileşenler analizi ve çok boyutlu ölçekleme analizi tekniklerini içeren ve literatürde birçok çalışmanın olduğu tekniklerdir.

Projeksiyon İzleme; Düşük boyutlu projeksiyonlar kullanarak yüksek boyutlu verinin analizi için tasarlanmıştır. Amacı yüksek boyutlu veri içerisinde olası lineer olmayan ve ilginç yapıları ortaya çıkarmaktır [Friedman ve Tukey, 1974].



Şekil 2.5. Dönüşüm tabanlı tekniklerin sınıflandırılması [Jensen, 2005]

Temel bileşenler analizi (PCA); değişkenler arası bağımlılık yapısının yok edilmesi ve (veya) boyut indirgenmesi ya da başka analizler için verinin hazırlanması amaçları ile kullanılırlar. PCA'da uygulanan adımlar basit olarak aşağıya çıkarılmıştır.

1. (n) ölçümdeki (p) değişkene ait veri matrisi standartlaştırılır,

2. Standartlaştırılmış veri matrisinin korelasyon matrisi bulunur,

3. Korelasyon matrisinin öz değerleri ve standartlaştırılmış öz vektörleri hesaplanır,

4. Öz değerlerden temel bileşenlerin toplam varyansı açıklama oranları elde edilir,

5. Her bir öz vektörün devriği ile standartlaştırılmış veri matrisinin devriği çarpılarak temel bileşen değerleri bulunur [Dinçer ve ark., 1996].

Çok boyutlu ölçekleme analizi (ÇBÖ) [Torgerson, 1958], nesne ya da birimler arasında gözlemlenen benzerlikler ya da farklılıklardan oluşan uzaklık değerlerine dayalı olarak bu nesnelerin tek ya da çok boyutlu uzaydaki gösterimini elde etmeyi amaçlayan, böylece nesneler arasındaki ilişkilerin belirlenmesini sağlayan çok değişkenli bir istatistiksel analiz yöntemidir. ÇBÖ analizinin kökleri psikofizik ve

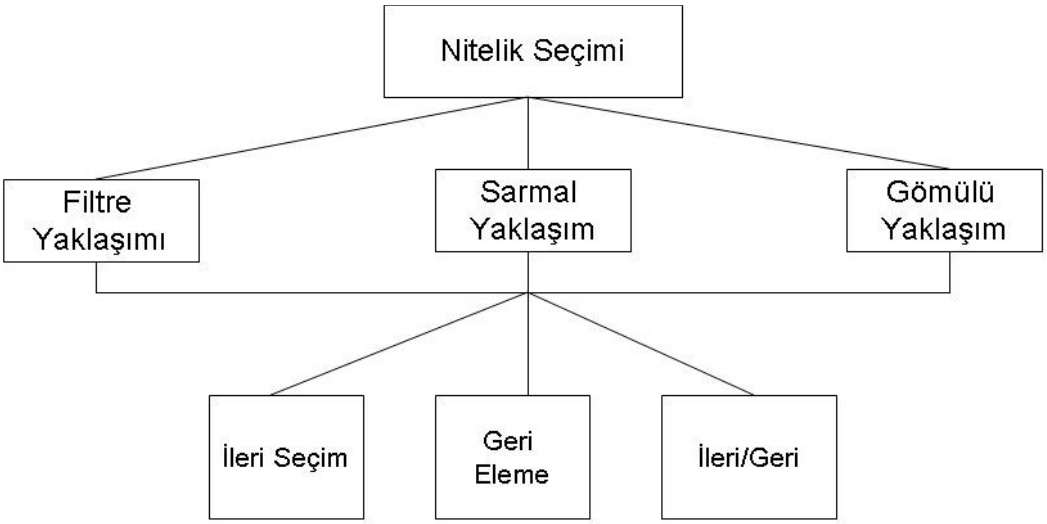
psikometri alanlarında yapılan çalışmalara dayanmakla birlikte pazarlama araştırmalarında yaygın olarak kullanılmaktadır [Yenidoğan, 2008].

Doğrusal olmayan metotlar;

Doğrusal metotların dezavantajı lineer olmayan veri ile boyutsal indirgeme yapamamalarıdır. Doğrusal olmayan ilişkilere sahip veri kümesinde bu metotlar sadece öklit yapısını bulabiliyorlardı. İlk çalışmalar PCA işleminin uzantılarıydı; ya veriyi başlangıçta kümeleme işlemine tabi tutuyor ve PCA'yı kümeler içerisinde kullanılıyordu ya da bir greedy optimizasyon işleminden geçiriyordu. Bu durum ise lineer olmama durumu ile başa çıkmaya çalışan tekniklerin gelişmesine yol açtı. Bu metotlar ise; ISOMAP [Tenenbaum ve ark., 2000] ve Locally Linear Embedding [Roweis ve Saul, 2000] metotlarıdır.

Seçim tabanlı indirgemede kullanılan yaklaşımlar;

Seçim tabanlı yaklaşımların bir tasnifi Şekil 2.6'de sunulmuştur.



Şekil 2.6. Seçim tabanlı yaklaşımların tasnifi [Mladenic,2006]

Seçilmiş bir alt kümenin optimal olup olmadığını belirlemek zor bir problemdir. Alt küme minimalitesi ve alt küme uygunluğu arasında karşılıklı ve ters yönlü bir değişim mevcuttur. Burada (özellikle birçok niteliğin gözlenmesinin masraflı oluşu ve pratik olmadığı) bazı uygulamalar için, daha küçük, daha az kesin nitelik altkümesi hem zaman hem de maliyet açılarından tercih edilebilir.

Nitelik seçimi yaklaşımı;

Nitelik seçimi kısaca belirli bir değerlendirme kriterine göre, bir araştırma problemi olarak tanımlanabilir. Nitelik seçimi işlemi, herhangi yeni bir niteliğin oluşturulmaması açısından nitelik çıkartmadan farklıdır.

Nitelik seçiminin amaçları;

1. Nitelik uzayı boyutunun azaltılması,

2. Öğrenme algoritması hızının arttırılması,

3. Sınıflandırma algoritması tahmin doğruluğunun arttırılması,

4. Öğrenme algoritması sonuçlarının daha iyi anlaşılması olarak sıralanabilir.

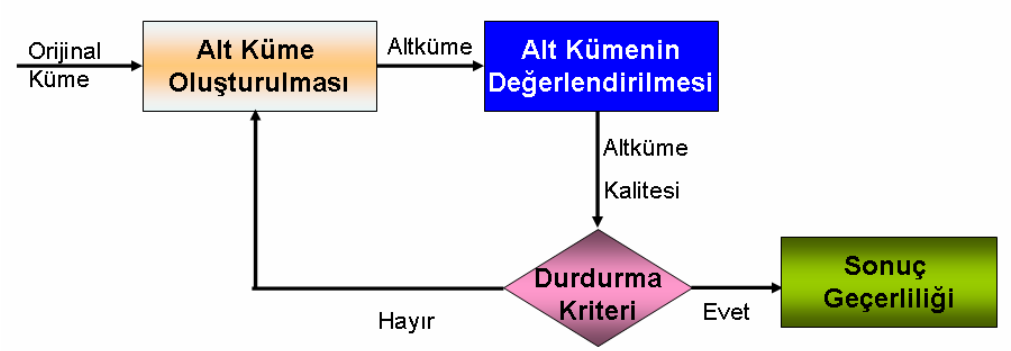
Şekil 2.7'de nitelik seçimi aşamaları sunulmuş olup, klasik bir nitelik seçimi problemi dört aşamadan oluşmaktadır [Dash ve Liu, 1997];

1. Nitelik alt kümesinin oluşturulması,

2. Seçilen niteliğinin belirli bir ölçüte göre değerlendirmesinin yapılması,

3. Seçilen alt kümenin uyumluluk kalitesine göre seçiminin yapılması,

4. Elde edilen sonucun geçerli olup olmadığına bakılması.



Şekil 2.7. Nitelik seçimi aşamaları

Nitelik alt kümesinin oluşturulması; Nitelik alt kümesinin oluşturulmasında, araştırma uzayında bir başlangıç noktasının belirlenmesi ilk aşamadır. Burada 3 değişik yöntemden bahsedilmektedir. Bunlardan birincisi, ileriye doğru üretim denilen, seçilen nitelikleri boş bir kümeye atayarak nitelik alt kümesinin bulunması işlemidir. Her aşamada seçilen nitelik boş kümeye eklenerek belirlenen durdurma ölçütü sağlanana kadar devam ettirilir. İkinci yöntem ise, geriye doğru eleme denilen, eldeki tüm niteliklerden başlamaktır. Burada, her bir işlemde belirlenen ölçüte göre, en düşük öneme sahip nitelik atılır. İşlem, durdurma kriteri sağlanana kadar ya da elde en az bir nitelik kalana kadar devam eder. Diğer bir yöntem ise, rassal olarak nitelik alt kümesinin oluşturulmasıdır. Bu yöntem yerel optimumdan kurtulmak için kullanılabilir olmasına rağmen nitelik sayısına bağlı olarak hesaplama zamanı açısından uygun bir yöntem olarak tercih edilmemektedir.

Değerlendirme ölçütü; Optimal alt küme kavramı belirli bir değerlendirme ölçütüne daima bağımlıdır. Diğer bir deyişle, başka bir değerlendirme ölçütü kullanıldığında sonuçlar değişecektir. Değerlendirme ölçütleri ise, bir öğrenme algoritmasına bağımlı olup olmadığına göre üç ana yaklaşımı benimsemektedirler. Bunlar filtre, sarmal ve gömülü yaklaşımlardır [Mladenic,2006]. Filtre yaklaşımında, uzaklık, enformasyon, bağımlılık ve tutarlılık gibi çeşitli ölçütler kullanılabilir. Sarmal yaklaşımda ise, seçilen alt küme üzerine öğrenme algoritmasının performansını ölçerek alt kümenin uygunluk kalitesini ölçer. Gözetimli öğrenmede sınıflandırma işleminin ilk hedefi tahmini doğruluğun ençoklanmasıdır. Bu yüzden araştırmacılar tarafından tahmini doğruluk genelde kabul edilen bir ölçüttür. Gözetimsiz öğrenmede

ise, kümeleme yoğunluğu, en çok olabilirlik kestirimi gibi teknikler kümeleme kalitesinin sonuçlarını ölçmek kullanılır.

Durdurma kriteri; her bir iterasyonda nitelik seçimi işleminin ilerleyip ilerlememesi kararını verir. Örneğin bir durdurma kriteri belirli bir nitelik sayısına ya da belirli bir işlem sayısına erişilmesi, bir optimal sonucun bulunması olabilir.

Sonuç Geçerliliği; bulunan nitelik alt kümesi ileride kullanılmak üzere doğrulanmalı ve diğer nitelik indirgeme algoritmaları ile karşılaştırılmalıdır.

Nitelik Seçimi Algoritmaları;

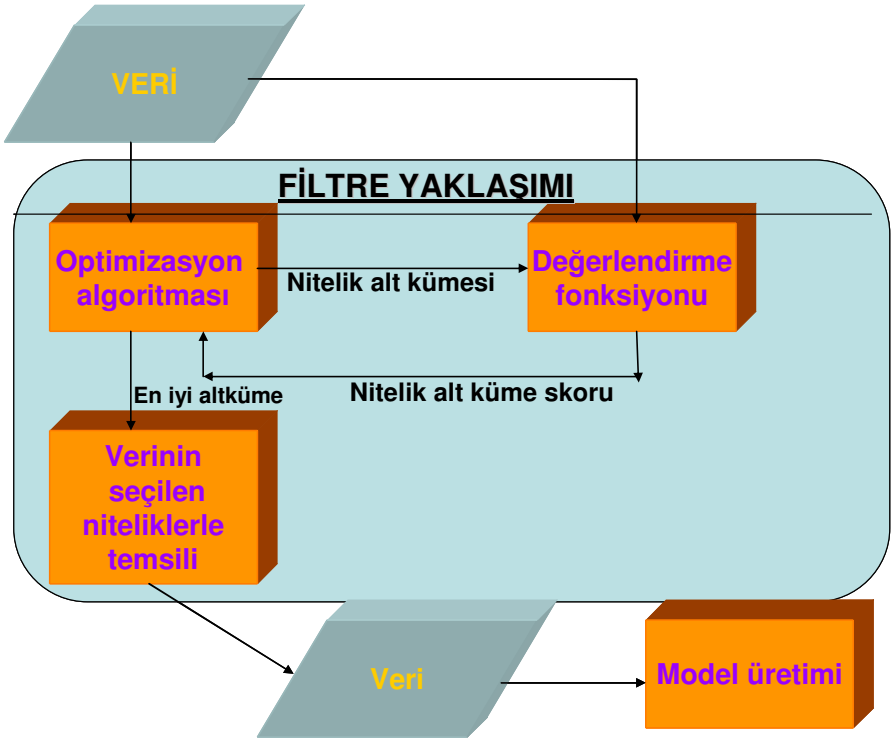
Nitelik seçim algoritmaları değerlendirme ölçütüne göre iki kategoriye ayrılabilirler. Eğer nitelik seçimini herhangi bir öğrenme algoritmasından bağımsız olarak yaparlarsa buna "Filtre yaklaşımı" denmektedir. Burada ilgisiz nitelikler kural çıkarımından önce ayıklanmaktadır. Filtreler özel bir kural çıkarım algoritmasının bir kısmı olmadıklarından birçok alanda kullanılabilmektedir [John ve Ark, 1994].

Şekil 2.8'de filtre yaklaşımının aşamaları sunulmaktadır.

Bu kapsamda kullanılan algoritmalar aşağıda kısaca tanıtılmıştır.

RELIEF [Kira ve Rendell; 1992]; filtre yaklaşımına dayanan ilk nitelik seçimi algoritmasıdır. RELIEF algoritmasında her bir niteliğe, karar sınıf etiketleri arasında ayırt edilebilme kabiliyetini gösteren bir "ilişki ağırlığı" verilir.

FOCUS [Almuallim ve Dietterich, 1991]; diğer bir filtreleme metodudur. Önce genişlik stratejisi kullanarak, eğitim verisinin tutarlı olarak etiketlenmesini sağlayan minimal nitelik kümesini araştırır.



Şekil 2.8. Filtre Yaklaşımı [Mladenic,2006]

LVF [Liu ve Setiono, 1996a]; bu metotta rassal nitelik alt kümeleri, Las Vegas algoritması kullanılarak üretilir.

SCRAP [Raman ve Ioerger, 2002]; Örnek uzayı içerisinde sırasal araştırma yaparak nitelik ilişkisini hesaplayan örnek tabanlı bir fitredir. SCRAP, diğer ileri ve geri araştırma tekniklerinin aksine nesneleri bir kerede ele alır. Buradaki ana fikir, veri tablosunda karar sınırlarını değiştiren niteliklerin tespit edilmesidir. Bu nitelikler en çok bilgi verici olarak kabul edilir.

EBR [Jensen ve Shen, 2001]; Diğer bir filtre tabanlı nitelik indirgeme tekniğidir. Bu yaklaşım C4.5 gibi makine öğrenimi teknikleri ile uygulanan entropi sezgiseline dayanır. Bir veri kümesi içerisinde en çok bilgi kazancını sağlayan niteliklerin bulunması temeline dayanır.

FDR [Traina ve Ark, 2000]; Değişik ölçeklerde veri tarafından sergilenen kendine benzerlik kavramına dayanan bir nitelik seçimi yaklaşımıdır.

Nitelik Gruplama (FG) [Yao, 2001]; genelde nitelik seçiminde üretim işlemi artan oranda tek tek nitelikleri ekler veya çıkarır. Son zamanlarda her aşamada niteliklerin gruplandırarak araştırmalar yapılmaktadır. Bu strateji aynı anda çeşitli nitelikleri seçerek optimal alt kümeleri bulmak suretiyle hesap zamanını azaltır.

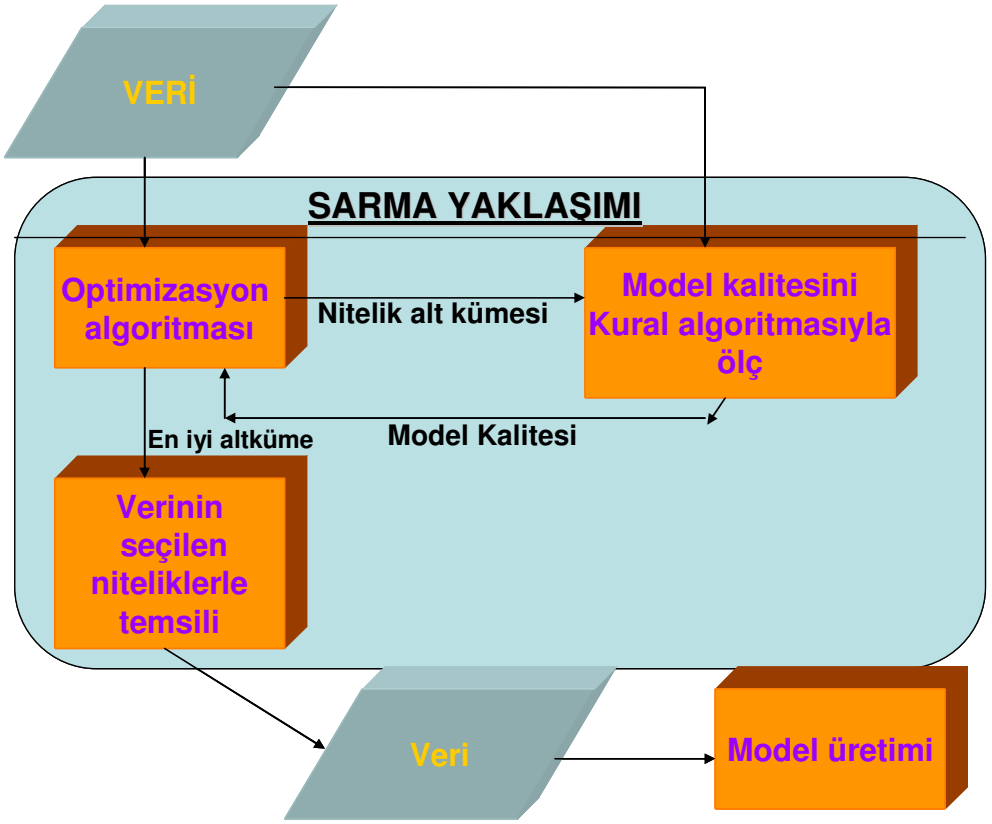
Eğer öğrenme algoritmasının değerlendirme işlemi bir uygulamaya (yani sınıflandırmaya) bağlı ise nitelik seçim algoritması sarma yaklaşımı kullanmaktadır. Bu metot bir kural çıkarım algoritmasından ölçülen kesinliği bir uygunluk ölçütü olarak kullanmak suretiyle nitelik altküme uzayını araştırır. Sarmal yaklaşımın daha iyi sonuçlar çıkarmalarına rağmen pahalı olmaları, çok fazla sayıda nitelikle başa çıkamamaları nedeniyle dezavantajlı oldukları değerlendirilmektedir.

Bu kapsamda kullanılan algoritmalar aşağıda sunulmuştur;

LVW algoritması [Liu ve Setiono, 1996b]; LVF algoritması temeline dayanan bir sarmal metottur. LVF ile olduğu gibi, LVW düşük sınıflandırma hatası ile sonuçlanacak şekilde çalışırken orta derecede çözümler üretir. Bu metot hesap zamanı açısından pek uygun sonuçlar elde edilemediğinden pek tercih edilmemektedir. Ancak durdurma kriterinin bir kural çıkarım mekanizmasına bağlı oluşu bakımından ise daha kesin sonuçlar vermesi kaçınılmazdır.

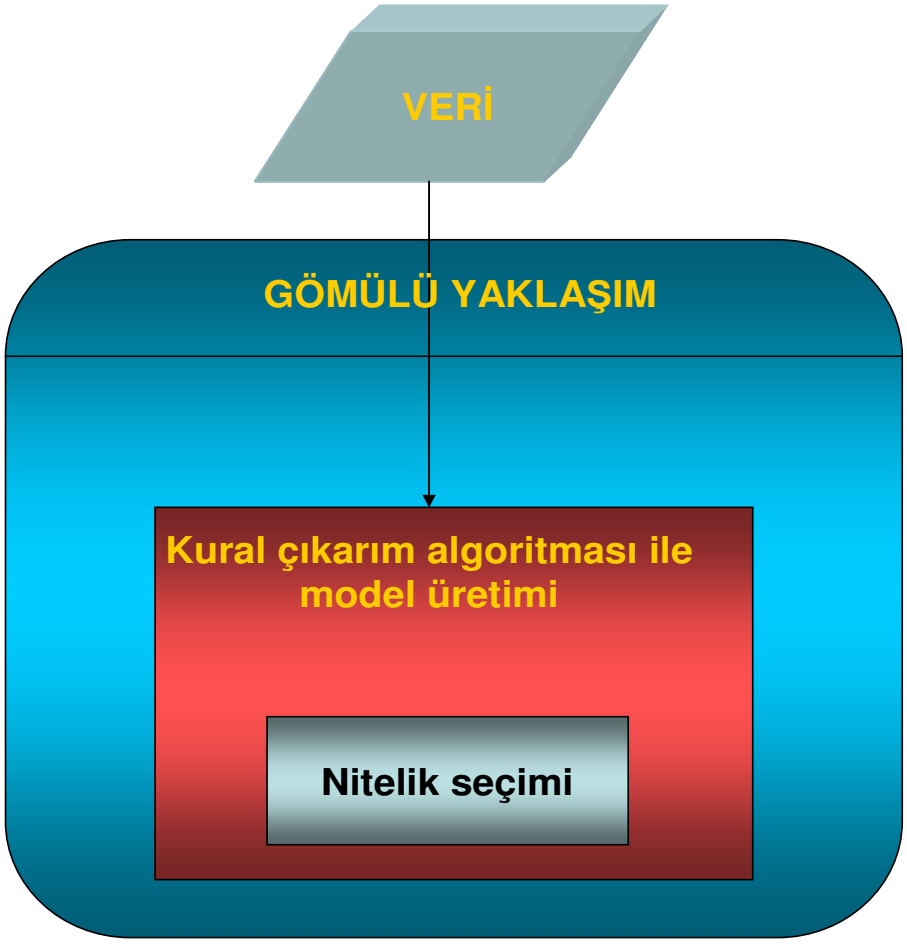
Setiono ve Lin [1997] tarafından, optimal alt kümeleri bulmak için geri eleme yöntemini kullanan yapay sinir ağları tabanlı nitelik seçici bir algoritma önerilmiştir.

Şekil 2.9'da sarmal yaklaşımın gösterimi sunulmaktadır.



Şekil 2.9. Sarmal Yaklaşım

Gömülü yaklaşımlarda nitelik seçimi ve öğrenme iç içe geçmiş durumdadır. Bunun en iyi bilinen örneği karar ağacı kural çıkarımıdır [Quinlan, 1993]. Burada nitelik bağımsızlığı, verinin altkümelere dağıtılması ve en iyi niteliğin bulunması için kullanılması öngörülür. Prosedür aşama aşama altkümeler üzerinde durdurma koşulu sağlanıncaya kadar devam eder. Çıktısı ise sadece bir nitelik altkümesinin kullanıldığı bir modeldir. Gömülü yaklaşım öğrenme algoritmasının öğrenme ile dönüşümlü çalışan nitelik seçimi mekanizmasına sahip olduğunu farz eder. Yani gömülü yaklaşım; şekil 2.10'da gösterildiği gibi kural çıkarım algoritmasının içerisinde nitelik seçim mekanizmasının olmasıdır.



Şekil 2.10. Gömülü yaklaşım

Diğer Metotlar;

Genetik algoritmalar [Holland, 1975]; Genelde geniş, doğrusal olmayan ve az anlaşılabilen uzayların hızlı bir şekilde araştırılması için oldukça etkili bir yöntemdir. Tek bir çözümün optimize edildiği klasik nitelik seçimi stratejilerinin tersine, çözüm topluluğu aynı anda değiştirilebilir. Bu çıktı olarak optimale yakın nitelik alt kümeleri üretir.

Tavlama benzetimi tabanlı nitelik seçimi [Kirkpatrick ve ark., 1983]; Tavlama maddenin kolay kırılganlığını azaltmak ve sertleştirmek için yavaşça ısıtıldığında ve soğutulma işlemidir. Örneğin bu işlem bir metalin minimum enerji ile belli bir

yapılanmaya ulaşabilmesi için yapılır. Eğer metal çok hızlı şekilde tavlanırsa başarıya ulaşılması mümkün değildir.

Bu çalışmada, nitelik indirgemede seçim tabanlı filtre yaklaşımlı yeni bir algoritma geliştirilmiştir.

3. BULANIK-KABA KÜMELER VE NİTELİK İNDİRGEME

Bu bölümde kısaca bulanık kaba küme kavramının anlaşılması için bulanık küme teorisi, kaba küme teorisi ve nitelik indirgemede kullanımı ve bulanık kaba kümelerin oluşturulması için temel özelliklerden bahsedilecektir.

3.1. Bulanık Küme Teorisi

Klasik matematikte nesnelerin veya olayların kesin ve tam tanımlarını araştırılmıştır. Fakat gerçek hayat senaryolarında bütün değerler kesin olarak elde edilemez. Mutlaka bu değerler içerisinde belirsizlik olacaktır. Bununla beraber, klasik matematikle modelleme esnasında bu muğlâklıktan dolayı problemlerle karşılaşacaktır. Bu çelişkiyi gidermenin bir yolu bulanık küme teorisidir.

Örneğin yağmurun yağması olayı; çiselemekten sağanak yağmura kadar değişik yoğunlukta rastlanabilen doğal bir olgudur. Fakat yağmur kelimesi kesin olarak yağmurun şiddetini kesin olarak tanımlamadığından bulanık bir olgudur.

Sıklıkla insan düşüncesinde oluşan olguların anlaşılması, hatırlanması ve kategorize edilmesi de bulanıktır. Bu kavramların sınıfları belli değildir. Bu yüzden bu kavramları yargılamak ve neden bulmak da bulanıktır. Örneğin yağmur hafif, orta ve ağır olarak sınıflandırılabilir. Maalesef sınırlar tanımlı olmadığından yağan yağmurun hafi, orta ve ağır derecede yağdığını söylemek zordur. Hafif, orta ve ağır kavramları bulanık kavramlardır.

Bulanık küme teorisi 1965'de Zadeh tarafından günlük yaşam içerisindeki bulanıklığın ve kesinlik içermeyen olguların gösterilmesi için ortaya atılmıştır. Bu teori çok karmaşık sistemlerin karakteristiklerin tanımlanması için etkin bir yaklaşım sağlar. Bulanık yaklaşım, insan düşüncesindeki ana öğenin sadece sayılardan oluşmadığını öngörür. Diğer bir ifade ile "kesin geçişten ziyade aşamalı geçiş öngören nesne sınıfları ile ifade edilebileceği" temeline dayanır. İnsan muhakemesi

arkasındaki mantık iki değerli ya da çok değerli mantık değil, bulanık gerçekler, bulanık ilişkiler ve bulanık kurallardır [Zadeh, 1965;1975a;1975b].

Örneğin, klasik A kümesinin gerçek sayılarının 6'dan büyük olması kesin bir sınırdır. Bu durum şu şekilde gösterilebilir. $A = \{x | x > 6\}$. Burada açık olarak 6 sınırı vardır. Eğer x, 6'dan büyükse bu kümeye girecek, değilse girmeyecektir. Fakat 6 km lik bir yol uzun saylıyorsa 5 km 900 metrenin durumu ne olacaktır? Klasik küme teorisinin tersine bir bulanık kümenin kesin sınırları yoktur. Bu *aittir* kümesinden *ait değildir* kümesine geçiş aşama aşamadır ve düzgün geçiş işlemi için "su sıcaktır" ya da "sıcaklık yüksektir" gibi dilsel ifade kullanılarak üyelik fonksiyonu ile kümelerin tanımlanmasında ve modellenmesinde esneklik kazandırılır.

Durum matematiksel olarak ifade edilirse, X evreni ve x ise $x \in X$ olacak şekilde belirli bir elemanı göstersin. Bulanık A kümesi [0,1] aralığında bir üyelik fonksiyonu ile tanımlanır.

$$\mu_A : X \rightarrow [0,1] \tag{3.1}$$

Bu yüzden, $x \in X$ için $\mu_A(x)$, x elemanının bulanık A kümesine aitlik derecesini gösterir. İki değerli (klasik) kümeler için $\mu_A(x)$, 0 veya 1 değerlerini alır.

Bulanık kümeler kesikli ve sürekli uzayda tanımlanabilir. X kesikli uzayında, bulanık kümeler ya n-boyutlu vektörlerle;

$$A = \{\mu_A(x_i)/x_i | x_i \in X, i = 1,...,n\} \tag{3.2}$$

ya da toplama işlemi ile gösterilir;

$$A = \mu_A(x_1)/x_1 + \mu_A(x_2)/x_2 + ... + \mu_A(x_n)/x_n = \sum_{i=1}^{n} \mu_A(x_i)/x_i \tag{3.3}$$

Burada toplama işlemini cebirsel toplama ile karıştırmamak gereklidir. Buradaki toplama işlemi sadece A kümesinin sıralı çiftlerden oluşturulduğunu gösterir.

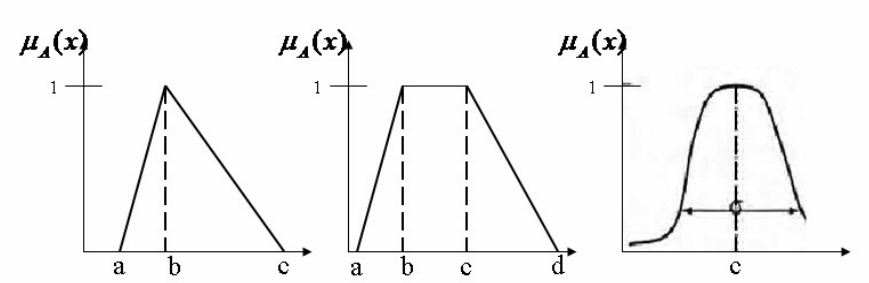
Bulanık üyelik fonksiyonu ile evren içerisinde her bir elemana 0–1 aralığında bir üyelik derecesi verilir.

$$A = \{x, \mu_A(x) | x \in X\}, \tag{3.4}$$

burada $\mu_A(x)$ üyelik fonksiyonudur.

Eğer üyelik değerleri sadece 0 veya 1 değerini alacak şekilde sınırlanırsa klasik A kümesi bulunur. Gerçek hayat problemlerinde üyelik fonksiyonları birkaç parametre ile ifade edilecek fonksiyon sınıfları ile (örneğin, üçgensel, yamuk ve gauss fonksiyonu şeklinde) sınırlandırılmıştır.

Şekil 3.1'de sık kullanılan üyelik fonksiyonu şekilleri sunulmaktadır.



Şekil 3.1. En çok kullanılan üyelik fonksiyonu şekilleri

Örneğin, üçgensel üyelik fonksiyonu {a,b,c} parametreleri ile ifade edilir.

$$\mu(x) = \text{üçgen}(x,a,b,c) = \begin{cases} 0 & x \le a \\ (x-a)/(b-a) & a \le x \le b \\ (c-x)/(c-b) & b \le x \le c \\ 0 & c \le x \end{cases} \tag{3.5}$$

Buradan da görülebileceği gibi bulanık kümeler, uygun evrenin tanımına ve uygun üyelik fonksiyonu seçimine bağlıdır.

3.1.1. Bulanık işlemler

Muğlak enformasyon altında belirli kararlara ulaşabilmek için kullanılan bulanık kümelerde ilişkiler ve işlemler tanımlanmalıdır [Klir ve Yuan, 1995].

Bulanık kümelerin eşitliği: Klasik kümeler için eğer, her iki küme de aynı elemana sahiplerse eşittir denilmektedir. Bulanık kümelerde ise durum biraz değişik olmakla beraber, bulanık üyelik dereceleri eşit olmasına bakılır. İki bulanık küme için, sadece ve sadece bütün $x \in X$ için $\mu_A(x) = \mu_B(x)$ ise, $A = B$ denir.

Bulanık kümelerde kapsama: Klasik kümelerde, eğer A'nın tüm elemanları B'de de görülüyorsa $A \subset B$ idi. Bulanık kümeler için ise elemanların o kümelere olan üyelik dereceleri de dikkate alınmalıdır. Bütün $x \in X$ için sadece ve sadece $\mu_A(x) \leq \mu_B(x)$ ise A bulanık kümesi B bulanık kümesinin alt kümesidir denilir ve $A \subset B$ ile gösterilir.

Normalite: Eğer, bir bulanık kümede herhangi bir elemanın aidiyeti 1 ise o bulanık küme normaldir denilir.

$$\exists x \in A \bullet \mu_A(x) = 1 \text{ veya diğer bir gösterimle } \sup_x A(x) = 1 \tag{3.6}$$

Normalleştirme: Bir bulanık küme üyelik fonksiyon değerlerini yüksekliğine bölünerek normalleştirilir.

$$Normalize(A) = \frac{\mu_A(x)}{yükseklik(x)} \tag{3.7}$$

Yükseklik; Bir bulanık kümenin yüksekliği üyelik fonksiyonlarının en küçük üst sınırı (supremum) olarak ifade edilir. Bir bulanık kümenin yüksekliği, o kümenin en küçük üst sınırıdır.

$$yükseklik(A) = \sup_x A(x) \tag{3.8}$$

Destek: A bulanık kümesinin desteği, X evreninde bulunan ve A'ya 0'dan farklı üyelik derecesine sahip olan bütün elemanların kümesidir.

$$destek(A)\{x \in X \,|\, \mu_A(x) > 0\} \tag{3.9}$$

Çekirdek (Öz): A bulanık kümesinin çekirdeği A kümesinde üyelik derecesi 1 olan bütün elemanların kümesidir.

$$Çekirdek(A) = \{x \in X \,|\, \mu_A(x) = 1\} \tag{3.10}$$

α-kesim: Üyelik derecesi belirli bir α değerinden yüksek A kümesinin bütün eleman kümesine A'nın α-kesimi denir.

$$A_\alpha = \{x \in X \,|\, \mu_A(x) \geq \alpha\} \tag{3.11}$$

Kardinalite: A bulanık kümesinin kardinalitesi;

$$Kardinalite(A) = \sum_{x \in X} \mu_A(x) \tag{3.12}$$

Bulanık sayı: Eğer bir bulanık küme dışbükey ve normalleştirilmiş ve parçalı (piecewise) sürekli ise, bulanık sayı denir.

3.1.2. Üçgensel Normlar

Üçgensel normlar (kısaca t-norm, T) bulanık mantık ve bulanık kümelerde kesişim özelliğinin anlaşılması için gerekli bir araçtır. t-normlar, ilk zamanlar olasılıklı metrik uzaylar kavramı içerisinde düşünülmüş, karar verme ve istatistik gibi bilim dallarında uygulama alanı bulmuştur. Cebirsel olarak t-normlar [0,1] aralığında kapalı, doğal elemanı 1 olan ikili bir işlemdir [Klementa ve ark., 2004].

Bir t-norm; [0,1] aralığında değişme, birleşme ve monotonluk özeliğini taşıyan ve nötr elemanı 1 olan ikili bir işlemdir. Bütün $x, y, z \in [0,1]$ için $T : [0,1]^2 \to [0,1]$ özelliğini taşırlar.

$$T(x, y) = T(y, x), \tag{3.13}$$

$$T(x, T(y, z)) = T(T(y, x), z), \tag{3.14}$$

$$T(x, y) \le T(x, z), \quad y \le z \tag{3.15}$$

$$T(x,1) = x. \tag{3.16}$$

Bazı yazarlar $T(x, y)$ gibi bir ifade yerine $x * y$ kullanmaktadırlar.

T-normlar Boole kesişimin bir uzantısı olduğundan [0,1] değerli ve bulanık mantıkta ($\wedge$) işareti ile de gösterilmektedir. Literatürde sayısız çeşitte t-norm bulmak mümkündür. Minimum T_M, çarpım T_P, Lukasiewicz t-norm T_L, güçlü çarpım T_D operatörleri bu anlamda önemli sayılabilir.

$$T_M(x, y) = \min(x, y), \tag{3.17}$$

$$T_P(x, y) = x.y, \tag{3.18}$$

$$T_L(x, y) = \max(x + y - 1, 0), \tag{3.19}$$

$$T_D(x, y) = \begin{cases} 0 & x, y \in [0,1[^2 \\ \min(x, y) & DH \end{cases} \tag{3.20}$$

T_D ve T_M en küçük ve en büyük t-normlardır. Minimum T_M her bir $x \in [0,1]$ bağımsız eşgüçlü eleman olduğundan ve T_P ve T_L t-normların önemli altsınıflarını ifade ettiğinden önemlidir. T_M, T_P, T_L ve T_D t-normları sırasıyla $M, \prod, W, Z$ olarak gösterilir.

Sınır koşulu (Eş. 3.16) ve monotonluk özelliği (Eş. 3.15) minimal gösterimde verilmiştir. Eş.(3.13)'de içine dahil edildiğinde $x \in [0,1]$ olacak şekilde her bir t-norm (T) şu koşulları karşılar:

$$T(0, x) = T(x, 0), \tag{3.21}$$

$$T(1, x) = x. \tag{3.22}$$

Bu yüzden bütün t-normlar $[0,1]^2$ birim alanının sınırları içerisindedir.

T'nin monotonluğu, Eş.(3.15) özelliği ile birleşince birleşik monotonluk kavramı ortaya çıkar.

$T(x_1, y_1) \le T(x_2, y_2)$ burada $x_1 \le x_2, y_1 \le y_2$ dir. (3.23)

t-normların Gücü: T_1 ve T_2 gibi iki t-norm için, $(x, y) \in [0,1]^2$ olmak üzere $T_1(x, y) \le T_2(x, y)$ ise, T_1 in T_2 'den daha zayıf olduğu söylenebilir. Buradan aşağıdaki sonuçlar çıkarılabilir:

Eğer $T_1 \le T_2$ ve $T_1 \ne T_2$ ise $T_1 < T_2$ dir.

Eğer $(x_0, y_0) \in [0,1]^2$ için $T_1(x_0, y_0) < T_2(x_0, y_0)$ dir.

Eş.(3.13), Eş.(3.15) ve Eş.(3.16)'in doğal bir sonucu olarak T_D en zayıf, T_M ise en güçlü t-norm olacaktır. Her bir t-norm için; $T_D \le T \le T_M$ denilebilir.

4 temel t-norm için ise $T_D < T_L < T_P < T_M$ yazılabilir.

t-altnorm: Bütün $(x, y, z) \in [0,1]$ için $F : [0,1]^2 \to [0,1]$ fonksiyonu, (Eş. 3.13-15) özellikleri ile beraber $F(x, y) \le \min(x, y)$ yazılabilirse buna t-altnorm denir.

Açıkça her bir t-norm bir t-altnorm'dur. Örneğin "0" fonksiyonu bir t-altnorm'dur aynı zamanda bir t-norm değildir. Her bir t-altnorm, t-norm'a birim karenin üst sağ sınır değeri tekrar belirlenerek dönüştürülebilir.

3.1.3. Üçgensel Conormlar

Üçgensel Conormlar t-normların duali gibi düşünülebilir.

Üçgensel conorm (S); [0,1] aralığında değişme, birleşme ve monotonluk özeliğini taşıyan ve etkisiz elemanı 0 olan ikili bir işlemdir. Bütün $x, y, z \in [0,1]$ için (Eş. 3.15-17) ve $S(x,0) = x.$ özelliğini taşırlar.

Aşağıdaki açıklanan ve isimleri sıra ile maksimum S_M, olasılıklı toplam S_P, Lukasiewics t-conorm S_L ve güçlü çarpım S_D ile gösterilmiştir.

$$S_M(x, y) = \max(x, y), \tag{3.24}$$

$$S_P(x, y) = x + y - x.y, \tag{3.25}$$

$$T_L(x, y) = \min(x + y, 1), \tag{3.26}$$

$$S_D(x,y)=\begin{cases}0 & x,y\in]0,1]^2\\ \max(x,y) & DH\end{cases} \tag{3.27}$$

S_M, S_P, S_L S_D t-conorm'ları $M^*,\prod^*,W^*,Z^*$ ile ifade edilirler. t-conorm'ların orijinal tanımı şu şekildedir; sadece ve sadece bütün $x,y\in[0,1]$ için aşağıdaki özeliklerin bir yada ikisini de sağlıyorsa $S:[0,1]^2\to[0,1]$ fonksiyonu bir t-norm'dur.

$$S(x,y)=1-T(1-x,1-y), \tag{3.28}$$

$$T(x,y)=1-S(1-x,1-y). \tag{3.29}$$

Eş. (3.28) durumuna T'nin dual t-conorm'u denir ve benzer olarak Eş. (3.29) ifadesine de S'nin dual t-norm'u denir. Açıkça (T_M,S_M), (T_P,S_P), (T_L,S_L) ve (T_D,S_D) çiftleri birbirinin dualidir.

Standart ters işlemi birim aralığın tümleyeni olarak düşünüldüğünde $N_s(x)=1-x$ olarak ortaya çıkacaktır. S-norm terimini kullanmamak için orijinal tanım üzerinden gidilmesi daha uygun olacaktır.

İkinci gereklilik durumuna açıklanan dualite kavramı t-conormlar için gerekli olan özelliklerin t-normlar'ın sahip olduğu özelliklerden çıkarılabildiği anlamına gelecektir.

Dualite kavramı bazı sıraları da değiştirir, diğer bir deyişle; bazı t-normlar için (T_1,T_2 için) $T_1\le T_2$ ve bunlara karşılık gelen (S_1,S_2) t-conormları için ise $S_1\ge S_2$ elde edilir.

Ters işlem;

(i) $N:[0,1]\to[0,1]$'da artmayan fonksiyona bir "*Ters işlem*" denir. Bu ters işlem N(0)=1 ve N(1)=0 koşulunu sağlamalıdır.

(ii) Normal koşullara ek olarak N'nin sürekli ve kesin olarak azalan olması koşulları sağlanırsa $N:[0,1]\to[0,1]$ ters işlem'e *sıkı ters işlem* denir.

Her bir güçlü ters işlem standart ters işlemin monoton bir dönüşümüdür. Aslında literatürde değişik ters işlem operatörleri de mevcuttur. $N_G[0,1]\to[0,1]$ için $N(x)=1-x^2$ fonksiyonu sıkı fakat güçlü bir ters işlem değildir. Bunlardan biri de Gödel ters işlemidir;

$$NG(x)=\begin{cases}1 & x=0,\\ 0 & x\in]0,1].\end{cases} \tag{3.30}$$

En önemli ve en çok kullanılan güçlü ters işlem ise $N_s:[0,1]\to[0,1]$ için standart ters işlemdir.

$$N_s=1-x\,. \tag{3.31}$$

Standart ters işlem (N); t-conorm'ları, t-norm'lar ile tanımlarken ve bir bulanık kümenin tümleyenini modellerken kullanabilir.

Bir t-norm (T) ve sıkı ters işlem (N) verildiğinde $S:[0,1]^2\to[0,1]$ için t-conorm T'ye N-dual olacak şekilde $S(x,y)=N^{-1}(T(N(x),N(y)))$ şeklinde ifade edilir. Bununla beraber eğer N sıkı olmayan bir ters işlemse yukarıdaki ifade kullanılmayacaktır.

Eğer N güçlü bir ters işlemse, t-conorm S'ye yukarıdaki yapı uygulanarak t-norm T elde edilebilir.

t-norm T ve t-conorm S arasında bir ilişki vardır. Bu De Morgan kanununun genişletilmiş halidir.

Eğer bütün $x, y \in [0,1]^2$ için aşağıdaki koşullar sağlanırsa, (T,S,N) üçlemesine De Morgan Duali denir.

$$T(x, y) = N(S(N(x), N(y))), \tag{3.32}$$

$$S(x, y) = N(T(N(x), N(y))). \tag{3.33}$$

Bu durum şu şekilde ifade edilebilir; bir t-norm T ve (T,S,N) verildiğinde sadece ve sadece güçlü ters işlem N ve S, T'nin N-duali olarak ifade edilebiliyorsa bir De Morgan Dual'dir denilir.

3.1.4. Bulanık kümelerde bütünleştirme işlemi

Bütünleştirme işlemi fonksiyonları içerisinde değişik özellikler barındıran sistemlerdir. [0,1] kapalı aralığında gerçek değerler alıp yine [0,1] aralığında değerler oluşturur. n bileşeni bulunan fonksiyonlar için $f : [0,1]^n \to [0,1]$ ile ifade edilirler.

Bütünleştirme operatörleri; bulanık kümelerde üyelik derecelerinin, Tercih derecelerinin, kanıt güçlülüğünün, hipotez desteği gibi argümanların birleştirilmesi için kullanılmaktadır [Beliakov ve ark.,2007].

Örneğin, bulanık mantıkta n bulanık kümedeki kısmi üyelik derecelerini içeren bir x nesnesi $\mu_1, \mu_2, ..., \mu_n$ ile ifade edilsin. Hedefimiz birleştirilmiş bulanık küme $\mu = f(\mu_1, \mu_2, ..., \mu_n).$ içinde bütün üyelik fonksiyonlarını temsil eden bir değer bulmaktır. Bu kombinasyon; kesişim, birleşim veya daha karmaşık bir işlem olabilir.

"0" girdi değeri "üyelik derecesinin" ya da buna dair herhangi bir tercih olmaması anlamlarına gelir. Benzer olarak "1" değeri ise tam üyelik derecesi veya tam tercih manasına gelmektedir. Bu kavram bize bütünleştirme fonksiyonları hakkında "sınırların korunması" gibi temel bir özelliğe işaret eder.

$$f(\underbrace{0,0,...,0}_{n-defa}) = 0 \text{ ve } f(\underbrace{1,1,...,1}_{n-defa}) = 1 \tag{3.34}$$

İkinci önemli özellik ise *Monotonluk* koşuludur. x, y gibi iki girdinin bütünleştirme işlemi yapıldığı düşünülsün. $x_1 < y_1$ ve $x_j = y_j$ olsun. f 'nin j'inci kriterine göre girdiler A ve B gibi iki alternatifi göstersin. Böylece B alternatifi A'ya göre tercih edilecektir. Elbette kriter sırası önemli değildir, böylece monotonluk sadece ilk kriter için değil bütün sıradaki girdiler için sağlanır.

Monotonluk ve sınırların korunması genel bütünleştirme fonksiyonlarının karakterize edilmesi için iki temel özelliktir. Eğer bu özelliklerden birisi sağlanmazsa f fonksiyonunu bir bütünleştirme fonksiyonu olarak ifade edilemez.

Bütünleştirme fonksiyonlarının sınıflandırılması

Yukarıda da açıklandığı gibi bütünleştirme işleminin değişik şekilleri mevcuttur. Bazı durumlarda birbirinden önemli ya da önemsiz nitelikleri bazı durumlarda ise mantıksal ilişkileri (birleşim ve kesişim gibi) modellemek için kullanılır. Bütünleştirme fonksiyonları için ortalama, kesişim, birleşim ve karışık olmak üzere dört değişik sınıftan bahsedilebilir.

Ortalama fonksiyonu: Bir f bütünleştirme fonksiyonunda eğer her bir x aşağıdaki koşulu sağlıyorsa ortalama davranışına sahiptir.

$$\min(x) \le f(x) \le \max(x). \tag{3.35}$$

Birleşim fonksiyonu: Eş. (3.39) sınırları içerisindeki her bir x bir birleşim davranışı sergiler.

$$f(x) \le \min(x) = \min(x_1, x_2, ..., x_n). \tag{3.36}$$

Kesişim fonksiyonu: Eş. (3.38) sınırları içerisindeki her bir x bir kesişim davranışı sergiler.

$$f(x) \ge \max(x) = \max(x_1, x_2, ..., x_n). \tag{3.37}$$

Karışık Fonksiyon; Eğer yukarıdaki sınırlara uymayan nitelikte değerler taşıyorsa karışık fonksiyon değerleri almaktadır.

Bütünleştirme fonksiyonlarının genel özellikleri

Eşgüçlülük: Eğer bir bütünleştirme fonksiyonu (f) bütün girdileri $x = (t, t, ..., t), t \in [0,1]$ ve çıktıları $x = (t, t, ..., t) = t$ ise eşgüçlü olarak adlandırılırlar. Monotonluk özelliği taşıdığından ortalama davranışı sergiler. Örneğin aritmetik ve geometrik ortalama da bir bütünleştirme fonksiyonudur.

Simetriklik: Eğer kendi değerleri girdilerin permütasyonuna bağımlı değilse (f) fonksiyonuna simetrik denir.

Sıkı Monotonluk: Eğer bir bütünleştirme fonksiyonu (f); $x \le y$ fakat $x \ne y$ şartlarını taşıyorsa $f(x) < f(y)$ yani fonksiyonun sıkı monoton olduğunu gösterir.

3.2. Kaba Küme Teorisi

Kaba küme teorisi 1980'lerin başında Pawlak [1982] tarafından önerilmiştir. Diğer yöntemlerin aksine, orijinal kaba küme yaklaşımı sadece veri içerisindeki bilgiyi

kullanır ve istatistiksel parametreler belirli oranlar veya belirli varsayımlara dayanmaz.

Kaba küme felsefesinin çıkış noktası, ilgilenilen her nesne ile bazı enformasyonun bağdaştırılması varsayımıdır. Diğer bir deyişle kaba küme teorisi, kümenin tek olarak elemanları ile tanımlandığı ve kümenin elemanları hakkında ilave hiçbir enformasyonun bulunmadığı klasik küme teorisinin aksine, bir kümenin tanımlanması için başlangıçta evrenin elemanları hakkında bazı bilgilere gereksinim olduğu varsayımına dayanır. Örneğin, nesneler; aynı hastalığa sahip olan hastalar ise, o hastalığın belirtileri hastalar hakkındaki enformasyonu oluşturur. Nesneler aynı enformasyon ile nitelendiriliyorlar ise, nesneler aynıdırlar veya ayırt edilemezler. Ortaya konulan ayırtedilemezlik ilişkisi, kaba küme teorisinin temelidir. Bütün aynı nesnelerin kümesine elemanter küme denir ve bilginin temel taşını (granülünü) oluşturur. Elemanter kümelerin herhangi bir birleşimine kesin küme denir, aksi takdirde küme kabadır. Bu tanımdan da anlaşılacağı gibi her bir kaba kümenin sınır alanı elemanları, yani kesinlikle kümenin kendisi ya da tümleyeni olarak sınıflandırılamayan elemanları vardır. Ancak kesin kümelerin sınır elemanları bulunmamaktadır [Binay, 2002].

Kaba küme teorisi, eksik ya da tam olmayan bilgi ile başa çıkabilen yeni bir matematiksel araçtır [Pawlak, 2004].

Eksik ya da tam olmayan bilgi birçok filozof, mantıkçı ve matematikçi tarafından uzun bir süre ele alınmıştır. Son yıllarda, bilgisayar bilimi, kısmen yapay zekâ alanlarında önemli bir konu haline gelmiştir. Eksik ya da tam olmayan bilginin nasıl anlaşılacağı ve işleneceği probleminin çözümü için birçok yaklaşım bulunmaktadır. Hiç şüphesiz ki en çok başarılı olanlardan biri de Zadeh [1965] tarafından önerilen bulanık küme teorisidir [Pawlak ve Skowron, 2007].

Kaba küme yaklaşımının ana avantajları şunlardır;

- Verideki gizli kalıpların bulunması için etkin metotlar, algoritmalar ve araçlar sağlar,
- İstatistikteki olasılık, bulanık kümelerdeki üyelik derecesi gibi veri hakkında ek bir enformasyona ihtiyaç duymaz,
- Orijinal veride olduğu kadar aynı bilgiyi sağlayacak verinin minimal altkümesinin (nitelik indirgeme) bulunmasını sağlar,
- Verinin önemliliğini değerlendirmeyi kolaylaştırır,
- Verilerden karar kuralları kümelerinin otomatik üretimini sağlar,
- Teori, kolayca anlaşılabilirdir,
- Elde edilen sonuçların açıkça anlaşılmasını sağlar,
- Paralel ve dağıtık işlemeye dönük algoritma yapılarına uygundur,
- Kaba küme teorisi, melezleri ve uygulamaları hakkında bilgi elde etmek için internet ortamında birçok kaynak bulunabilir [Suraj, 2004].

Kaba kümelerin uygulama alanı çok çeşitlidir. Özellikle kaba küme yaklaşımı, yapay zekâ, bilişsel bilimler, makine öğrenimi, bilgi keşfi, veri madenciliği, uzman sistemler, yaklaşık nedenleşme (approximate reasoning) ve örüntü tanıma konularında uygulama alanı bulmuştur.

3.2.1. Yaklaşım uzayları ve küme yaklaşımları

Kaba küme yaklaşımı veri içerisindeki belirsizlik ve muğlâklık ile ilgilenen diğer bir yaklaşımdır. Bulanık küme teorisine benzer olarak klasik küme teorisine bir alternatif değil onun içine gömülmüş ya da adapte edilmiş şekildedir [Pawlak, 2004] .

Kaba küme kavramı iç ve kapanış denilen topolojik işlemler yardımıyla genel olarak tanımlanır ve bu tanıma *yaklaşım* adı verilir.

Bu probleme daha dikkatli bakılacak olunursa; U evren, $R \subseteq UXU$ bir ikili ilişki olsun. Konunun kolay anlaşılması için R'nin bir eşdeğerlik ilişkisi olduğunu farz edilecektir. Burada, (U, R) çiftine *yaklaşım uzayı* denir.

X, U'nun bir alt kümesi olsun. Burada X kümesi R ilişkisi ile tanımlanmaya çalışılacaktır.

$R(x)$; x elemanı ile belirlenen R eşdeğerlik sınıfını ifade etsin. Burada R ayırtedilemezlik ilişkisi, evren (U) hakkındaki bilgi eksikliğini ifade eder. R ilişkisinin eşdeğerlik sınıflarına, R'ye göre idrak edebildiğimiz bilginin temel kısmını gösteren *granül* denir. Diğer bir ifade ile ayırtedilemezlik ilişkisi evren hakkındaki bilgi eksikliğini ifade eder. Diğer bir deyişle; beynimiz bir benzetme ilişkisi ile bir nesneyi bir başka nesneye benzetir ya da ayırt eder. Eğer iki nesne gerçekte farklı, fakat biz aynı olarak algılıyorsak, bu bizim evren hakkındaki bilgisizliğimizden kaynaklanır. Sadece ayırtedilemezlik ilişkisi kullanılarak sadece evrendeki nesneler teker teker gözlenmez, aynı zamanda bu ilişki ile tanımlanan bilginin granüllerine de erişilebilir.

Kaba kümeler; alt ve üst yaklaşımlar yardımıyla, kaba üyelik fonksiyonları ile veya benzerlik ilişkileri oluşturularak tanımlanırlar.

X kümesinin R'ye göre alt yaklaşımı; R'ye göre kesinlikle X olarak sınıflandırılan nesnelerin kümesine eşittir.

$$R_*(X) = \bigcup_{x \in U} \{R(x) : R(x) \subseteq X\} \tag{3.38}$$

X kümesinin R'ye göre üst yaklaşımı; R'ye göre muhtemelen X olarak sınıflandırılabilen nesnelerin kümesine eşittir.

$$R^*(X) = \bigcup_{x \in U} \{R(x) : R(x) \cap X \neq 0\} \tag{3.39}$$

X kümesinin R'ye göre sınır bölgesi; R'ye göre ne X ne de X değil olarak sınıflandırılabilen nesnelerin kümesine denir.

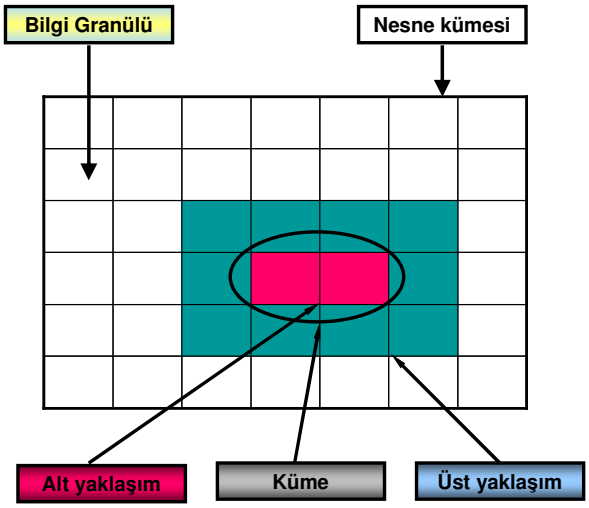
$$BN_R(X) = R^*(X) - R_*(X) \tag{3.40}$$

Eğer X'in sınır bölgesi boş ise, X kümesi R'ye göre kesindir.

Eğer X'in sınır bölgesi boş değilse X kümesi R'ye göre kabadır.

Görüleceği gibi, yaklaşımların tanımları bilgi granülü açısından ifade edilmiştir. Bir kümenin alt yaklaşımı tamamen küme içinde dahil olan bütün bilgi granüllerinin birleşimine, üst yaklaşım ise küme ile boş olmayan kesişim yapan granüllerin birleşimine eşittir. Sınır bölgesi ise üst ve alt yaklaşım arasındaki farktır. Diğer bir deyişle, bilginin granüler yapısından dolayı, eldeki bilgi kullanılarak nesnenin kesin bir tanımı yapılamaz. Bu yüzden, her bir kaba küme ile alt ve üst yaklaşım denilen iki klasik küme ile ilişkilendirilir. Böylece kaba küme, klasik kümenin tersi olarak boş olmayan sınır alanına sahiptir.

Kaba küme yaklaşımları Şekil 3.2'de gösterilmektedir.



Şekil 3.2. Kaba küme yaklaşımları

Alt ve Üst yaklaşımların özellikleri aşağıya çıkarılmıştır.

$$R_*(X) \subseteq X \subseteq R^*(X), \tag{3.41}$$

$$R_*(0) = R^*(0) = 0; R_*(U) = R^*(U) = U, \tag{3.42}$$

$$R^*(X \cup Y) = R^*(X) \cup R^*(Y), \tag{3.43}$$

$$R_*(X \cap Y) = R_*(X) \cap R_*(Y), \tag{3.44}$$

$$R_*(X \cup Y) \supseteq R_*(X) \cup R_*(Y), \tag{3.45}$$

$$R^*(X \cap Y) \subseteq R^*(X) \cup R^*(Y), \tag{3.46}$$

$$X \subseteq Y \rightarrow R_*(X) \subseteq R_*(Y) \,\&\, R^*(X) \subseteq R^*(Y), \tag{3.47}$$

$$R_*(-X) = -R^*(X), \tag{3.48}$$

$$R^*(-X) = -R_*(X), \tag{3.49}$$

$$R_*R_*(X) = R^*R_*(X) = R_*(X), \tag{3.50}$$

$$R^*R^*(X) = R_*R^*(X) = R^*(X), \tag{3.51}$$

Dolayısı ile bulanık küme teorisi ve kaba küme teorisi tamamen değişik matematiksel oluşum gerektirir.

Kaba kümeler yaklaşımların yerine kaba üyelik fonksiyonu ile de tanımlanabilir [Pawlak ve Skowron, 1994];

$$\mu_X^R : U \to \langle 0,1 \rangle, \tag{3.52}$$

Burada, μ_X^R ; x'in X kümesine R ilişkisi içerisinde aidiyetini ifade eder.

$$\mu_X^R(x) = \frac{card(X \cap R(x))}{card(R(x))}, \tag{3.53}$$

Card(x) X'in kardinalitesi yani küme içindeki eleman sayısıdır. Kaba üyelik fonksiyonu, R bilindiğinde, x'in X'e aidiyetinin koşullu olasılığını ifade eder.

Kaba üyelik fonksiyonu yaklaşımları ve sınır bölgeleri tanımlamak için de kullanılır.

$$R_*(X) = \{x \in U : \mu_X^R(x) = 1\}, \tag{3.54}$$

$$R^*(X) = \{x \in U : \mu_X^R(x) > 0\}, \tag{3.55}$$

$$BN_R(X) = \{x \in U : 0 < \mu_X^R(x) < 1\}. \tag{3.56}$$

Kaba üyelik fonksiyonu aşağıdaki özelliklere sahiptir;

Sadece ve sadece $x \in R_*(X)$ ise $\mu_X^R(x) = 1$, (3.57)

Sadece ve sadece $x \in U - R^*(X)$ ise $\mu_X^R(x) = 0$, (3.58)

Sadece ve sadece $x \in BN_R(X)$ ise $0 < \mu_X^R(x) < 1$. (3.59)

Üçüncü ve son olarak kaba kümeler benzerlik ilişkileri ile de tanımlanabilir. Klasik kaba küme yaklaşımları ayırt edilemezlik ilişkileri üzerine kuruludur. Bahsedilen eşdeğerlik ilişkisi yansıma, simetri ve geçişim özelliklerine sahiptir.

3.2.2. Enformasyon sistemleri

Yukarıda verilen özellikler kullanılarak temel kavramlar formüle edilebilir.

U ve A gibi iki sonlu, boş olmayan küme ele alalım, burada U evren, A ise nitelik kümesidir. $S = (U, A)$ çiftine enformasyon sistemi denir. Her bir $a \in A$ için a'nın yayılım alanı denilen V_a değer kümesi bulunur. A'nın herhangi bir B altkümesi U üzerinde ikili ilişkiyi , $I(B)$, belirler ve buna ayırtedilemezlik ilişkisi denir.

Her bir $a \in A$ için sadece ve sadece $a(x) = a(y)$ ise x $I(B)$ y ayırtedilmezlik ilişkisi sağlanır. Buradaki $a(x)$, x elemanı için a niteliğinin değerini ifade eder.Açıkça görüldüğü gibi $I(B)$ bir eşdeğerlik ilişkisidir. $I(B)$'nin bütün eşdeğerlik sınıfları ailesi yani B tarafından belirlenen bölüntü $U / I(B)$ ile ya da basitçe U / B ile ifade edilecektir. $I(B)$ eşdeğerlik sınıfı yani U / B bölüntü bloğu $B(x)$ ile gösterilecektir. Eğer (x,y) çifti $I(B)$'ye aitse buna x ve y, B-ayırtedilemez denecektir. $I(B)$ ilişkisinin eşdeğerlik ilişkisine de B-elemanter küme denecektir.

Yani kısaca, $I(B)$ ilişkisi ile birbirinden ayırt edilen sınıflara $B(x)$, eğer ayırtedilen nitelik yoksa da bu duruma B-ayırtedilemez denecektir.

3.2.3. Karar tabloları

İçindeki niteliklerin durum ve karar şeklinde iki sınıfa ayrıldığı bilgi sistemine karar tablosu denir. Durum ve karar nitelikleri ile oluşturulan karar tablosu evreninin bölüntülerini ifade eder. Çizelge 3.1'de örnek bir karar tablosu sunulmaktadır.

Çizelge 3.1'de Baş ağrısı, Kas ağrısı ve Ateş nitelikleri durum nitelikleri olarak değerlendirilirken Grip bir karar niteliği olarak ele alınır. Bir karar tablosunda durum nitelikleri C, karar nitelikleri D ve karar tablosu S=(U,C,D) olarak gösterilir. Karar tablosunun her bir satırı bir ilgili durumlar sağlandığında alınacak kararları belirleyen

bir karar kuralını ifade eder. Örneğin Çizelge 3.1'de (Baş ağrısı-Yok), (Kas ağrısı-Var), (Ateş-Yüksek) durumları (Grip-Evet) kararını belirler. Karar tablosundaki nesneler karar kurallarının etiketi gibi kullanılırlar.

Çizelge 3.1. Karar tablosu.

Hasta	**Baş ağrısı**	**Kas ağrısı**	**Ateş**	**Grip**
P1	Yok	Var	Yüksek	Evet
P2	Var	Yok	Yüksek	Evet
P3	Var	Var	Çok Yüksek	Evet
P4	Yok	Var	Normal	Hayır
P5	Var	Yok	Yüksek	Hayır
P6	Yok	Var	Çok Yüksek	Evet

Karar kuralları tutarlı ve tutarsız olabilir. Buna bazen kesin ve olası kurallar da denilebilmektedir. Karar tablosundaki tutarlı kuralların tüm kurallara oranı tutarlılık faktörü olarak tanımlanır ve $\gamma(C,D)$ olarak gösterilir. Eğer $\gamma(C,D)=1$ ise karar tablosu tutarlıdır. Bütün karar tablolarının tutarlı olması mümkün değildir.

3.2.4. Nitelik bağımlılığı

Veri analizinde önemli bir hususlardan biri de nitelikler arası bağımlılıkların bulunmasıdır. Eğer D'deki bütün nitelikler C'deki nitelik değerleri ile belirleniyorsa yani D nitelikleri tamamen C niteliklerine dayanıyorsa bu C=>D ifadesi ile gösterilir.

Örneğin Çizelge 3.1'de tam bağımlılık yoktur. Diğer bir deyişle, Çizelge 3.1'deki Grip niteliğinin P5 gözlemi için değeri "*hayır*" yerine "*evet*" olsaydı ateş ve grip arasında *tam bağımlılık var* denecekti. Dolayısı ile nitelik bağımlılık kavramı için daha genel bir kavrama ihtiyaç vardır. Buna "*Kısmi Bağımlılık*" denecektir.

Kolaylıkla görülebileceği gibi; eğer D, C ye tamamen bağlı ise $I(C) \subseteq I(D)$ dir. Bu C ile yaratılan bölüntünün D'den daha iyi olduğu manasına gelir.
Eğer D, C ye bir k derecesi ile bağımlı ise, 0<k<1;

$$\gamma(C,D) = \frac{Card(POS_C(D))}{Card(U)} \tag{3.60}$$

Burada; $POS_C(D) = \bigcup_{X \in U/I(D)} C_*(X)$ dir. (3.61)

$POS_C(D)$ ifadesine U/D bölüntüsünün C'ye göre pozitif alanı denir ve bu ifade C yardımıyla U/D bölüntüsünün bloklarına kesinlikle sınıflandırılabilecek evrenin elemanlarını içerir.

3.2.5. Nitelik indirgeme

Bir veri tablosunda bazı veriler veri tablosunun temel özellikleri kaybedilmeden atılması nitelik seçimi kavramını ifade eder.

Örneğin; Çizelge 3.1'deki baş ağrısı veya kas ağrısı niteliklerinden biri atılmış olsa yaklaşımlar ya da bağımlılıklar anlamında bir fark ifade etmeyecektir. Küçük nitelik kümeleri kullanarak aynı bağımlılık derecesini ve yaklaşım kesinliği bulunabilir. Bu fikri daha iyi ifade etmek için B, A kümesinin bir alt kümesi ve a niteliğinin B kümesine ait olduğu farz edilsin.

Eğer $I(B) = I(B - \{a\})$ ise a niteliği B kümesinde gereksizdir. Aksi halde a niteliği B kümesinde gereklidir.

B bütün nitelikleri gerekli ise bağımsızdır denir.

Eğer B' bağımsız ve $I(B') = I(B)$ ise B 'nin indirgenmiş altkümesidir.

Diğer önemli bir özellik ise çekirdek (öz, kor) özelliğidir.

$Çekirdek(B) = \bigcap \text{RED}(B)$ şeklinde ifade edilir.

$RED(B)$ ise B'nin tüm indirgenmiş kümeleridir.

3.2.6. Ayırtedilebilirlik matrisi ve fonksiyonlar

İndirgemeleri ve çekirdeği kolayca hesaplamak için Skowron ve Rauser [1992] tarafından ayırtedilebilirlik matrisi önerilmiştir.

$S = (U, A)$, n nesneli bir enformasyon sistemi olsun. S'nin ayırtedilebilirlik matrisi, $M(S)$ ile gösterilir. Matris simetriktir ve matrisin girdileri (c_{ij}) şu şekilde hesaplanır.

$$c_{ij} = \{a \in A : a(x_i) \neq a(x_j)\}, \text{ i,j=1,2,...n.} \tag{3.62}$$

Buradaki c_{ij} girdisi x_i ile x_j nesnelerini ayırt eden bütün niteliklerinden oluşur. $M(S)$ matrisi simetrik ve $c_{ii} = 0$ olduğundan bu matrisin sadece alt üçgen kısmının hesaplanması yeterlidir.

Her bir ayırtedilebilirlik matrisi ile aşağıda tanımlanan bir ayırtedilebilirlik fonksiyonu tanımlanabilir. Bir enformasyon sistemi S'nin ayırtedilebilirlik fonksiyonu f_S, $a_1,...,a_m$ niteliklerine denk gelen m adet $a_1^*,...,a_m^*$ boole değişkenin bir fonksiyonudur ve aşağıdaki şekilde tanımlanır;

$$f_S(a_1^*, a_2^*,..., a_m^*) = \forall\{\exists c_{ij}^* \,|\, 1 \leq j \leq i \leq n, c_{ij} \neq 0\} \tag{3.63}$$

Burada, $c_{ij}^{*} = \{a^{*} \mid a \in c_{ij}\}$ dir. f_S ayırtedilebilirlik fonksiyonunun temel içerenleri, A nitelik kümesindeki tüm indirgemeleri belirler.

Çizelge 3.1'de sunulan karar tablosunun ayırtedilebilirlik matrisi Çizelge 3.2 de gösterilmektedir.

Çizelge 3.2. Ayırtedilebilirlik matrisi

	P1	**P2**	**P3**	**P4**	**P5**	**P6**
P1						
P2	B,K					
P3	B,A	K,A				
P4	A,G	B,K,A,G	B,A,G			
P5	B,K,G	G	K,A,G	B,K,A		
P6	A	B,K,A	B	A,G	B,K,A,G	

Çizelge 3.2'de, Başağrısı (B), Kasağrısı (K), Ateş (A) ve Grip (G) olarak gösterilmiştir. Bu tablonun ayırtedilebilirlik fonksiyonu açısından ifadesi;

$f_S(B,K,A,G) = (B \vee K)(B \vee A)(A \vee G)(B \vee K \vee G)A(K \vee A)(B = (B \vee K \vee A \vee G)G$
$(B \vee K \vee A)(B \vee A \vee G)(K \vee A \vee G)B(B \vee K \vee A)(A \vee G)(B \vee K \vee A \vee G)$ dir.

Burada, $\vee$ işareti boole cebiri açısından ayrışmayı ifade eder. Ayrıca, birleşme bu formülasyonda gösterilmemiştir. Fonksiyonda her bir parantez birleşmeyi ifade eder. Boole cebiri kullanılarak yapılan sadeleştirmeden sonra Çizelge 3.2'de görülen tek elemanlı girdiler {B,A,G} nitelikleri indirgeme kümesi ve çekirdek olarak bulunur.

Ayırtedilebilirlik fonksiyonu kullanılarak göreli indirgemeler de bulunabilir. Bunun için formülasyonda ufak bir değişiklik yapmak gerekir.

$$c_{ij} = \{a \in C : a(x_i) \neq a(x_j) \text{ ve } w(x_i, x_j)\} \tag{3.64}$$

Burada, $w(x_i, x_j) \equiv x_i \in POS_C(D)$ ve $x_j \notin POS_C(D)$ veya, (3.65)

$$x_i \notin POS_C(D) \text{ ve } x_j \in POS_C(D) \text{ veya,} \quad (3.66)$$

$$x_i, x_j \in POS_C(D) \text{ ve } x_i, x_j \notin U/D,\ i, j = 1,2,...,n. \quad (3.67)$$

Eğer D ile tanımlanan bölüntü, C tarafından tanımlanabilir ise, yukarıdaki $w(x_i, x_j)$ koşulu $x_i, x_j \notin U/D$ olarak indirgenebilir.

Böylece, $IND(D)$ ilişkisinin eşdeğerlik sınıflarına ait olmayan, x_i ile x_j nesnelerini ayırt edebilen bütün nitelik kümelerinin birleşimi c_{ij} olarak ifade edilir. Dolayısı ile burada yukarıda bahsedilen iki kural yerine getirilmiş olur.

Eğer burada boole fonksiyonu yerine birleşme operatörünü sadece ayırtedilebilirlik matrisindeki k sütununda sınırlandırıldığında, k-göreli ayırtedilebilirlik fonksiyonu elde edilir. Bu fonksiyonun bütün temel içerenleri A'nın k-göreli indirgeme kümelerini belirler. Bütün bu indirgemeler $x_k \in U$ nesnesini diğer nesnelerden ayırt edecek minimum miktarda bilgiyi açığa çıkarır.

Çizelge 3.3. Karar göreli ayırtedilebilirlik matrisi

	P1	**P2**	**P3**	**P4**	**P5**	**P6**
P1						
P2						
P3						
P4	A	B,K,A				
P5	B,K		K,A			
P6				A	B,K,A	

Çizelge 3.3'den görülebileceği gibi tablonun ayırtedilebilirlik fonksiyonu; $A(B \vee K)(B \vee K \vee A)(K \vee A)$ olarak bulunur ve ayırtedilebilirlik fonksiyonu basitleştirildikten sonra $AB \vee AK$ indirgeme ve bu iki ifadenin kesişimleri olan A niteliği çekirdek olarak bulunur.

3.2.7. Niteliklerin önemi

Bir niteliğin önemi; nitelik bilgi tablosundan çıkarıldığındaki etkisi ile ölçülür. $\gamma(C, D)$ karar tablosunun tutarlılık derecesini veya C ile D arasındaki bağımlılık derecesini veya C tarafından U/D nin yaklaşımlarının kesinliğini ifade eder. "a" niteliğinin önemi $\gamma(C, D)$ ile $\gamma(C-a, D)$ arasındaki farkla ölçülür. Bir niteliğin önemi;

$$\sigma_{(C,D)}(a) = \frac{(\gamma(C,D) - \gamma(C-\{a\},D))}{\gamma(C,D)} = 1 - \frac{\gamma(C-\{a\},D)}{\gamma(C,D)} \text{ veya} \tag{3.68}$$

$$\sigma_{(C,D)}(B) = \frac{(\gamma(C,D) - \gamma(C-B,D))}{\gamma(C,D)} = 1 - \frac{\gamma(C-B,D)}{\gamma(C,D)} \text{ veya} \tag{3.69}$$

$$\varepsilon_{(C,D)}(B) = \frac{(\gamma(C,D) - \gamma(B,D))}{\gamma(C,D)} = 1 - \frac{\gamma(B,D)}{\gamma(C,D)} \tag{3.70}$$

ölçütlerinden biri ile ölçülebilir. [Geng ve Zhu, 2006]. Burada, 0< σ <1 dir.

Örnek için ;

σ(Baş ağrısı)=0, σ(Kas ağrısı) = 0, σ(Ateş)=0.75 olarak bulunur. Ateş niteliği öneminin 0.75 olması onun karar verici gücünü gösterir. Yani tutarlı karar kurallarının 3/4'ünün bu niteliğe bağlı olduğu manasına gelir.

3.2.8. Benzerlik ilişkileri ile kaba kümelerin tanımlanması

Kaba kümeler, alt ve üst yaklaşımların yanı sıra benzerlik ilişkileri ile de tanımlanabilir. Klasik kaba küme yaklaşımları ayırt edilemezlik ilişkileri üzerine kuruludur. Bahsedilen eşdeğerlik ilişkisi yansıyan, simetrik ve geçişkenlik özelliklerine sahiptir. Slowinski ve Vanderpooten, [1997] makalesinde, bu ilişkinin sadece geçişli olması için kısıtlanabileceği önerilerek benzerlik ilişkisi üzerinden kaba kümeleri tekrar tanımlanmıştır. Bu yöntemin diğer karmaşık işlemlere göre daha rahat anlaşılır olması bakımından avantajlı olabileceği değerlendirilmektedir.

U nesnesi (evren) bir tanımlama (nitelik değerleri) ile karakterize edilebilir. Nesneler arası benzerlikler eldeki tanımlama gücü ile aynı ya da birbirinden pek az farkı olan nesne sınıflarını oluşturan bir ikili ilişki (R) ile tanımlanabilir. Bu sınıflar evren hakkında *bilginin temel blokları* olarak da tanımlanabilir. Aslında bu ikili ilişki kaba küme yaklaşımında ayırtedilemezlik diye adlandırılan eşdeğerlik ilişkisi kavramıdır. Bunu oluştururken her bir nitelik için R_j olarak adlandırılan ilişki kümelerinin bütünleştirilmesi gerekir. Basitçe bu işlem bir kesişim operatörü ile yapılabilir. Sonuç olarak iki nesne R ilişkisine göre sadece ve sadece a_j üzerinde aynı değeri alıyorsa ayırtedilemez denilebilir.

Böyle bir tanımın çok kısıtlayıcı olabileceğinden hareketle, a_j değerleri arası küçük farklılıklara tolerans tanınabilir. Bununla beraber toplama operatörünün de kesişim kavramından daha az kısıtlayıcı şekilde düzenlenmesi gerekebilir. Bu düşünce sistemi ayırtedilemezlik ilişkisini daha genel bir şekil olan benzerlik ilişkilerinin kullanımı ile daha esnek çözümler üretebileceğine işaret eder.

$R(x) = \{ y \in U : yRx \}$ ile R(x) ile ayırtedilemeyen nesne kümesi ifade edilmektedir. R; yansıyan, simetriklik ve geçişkenlik özelliklerini taşıyan eşdeğerlik ilişkisi olarak tanımlanabilir. Böylece R, U'nun bir bölüntüsünün eşdeğerlik sınıfını ifade ettiği görülebilir.

R, U üzerinde bir ayırtedilemezlik (eşdeğerlik) ilişkisi ve $\tilde{R}$ ise R yi genişleten bir benzerlik ilişkisi olsun;

$$\forall x \in U, R(x) \subseteq \tilde{R}(x) \tag{3.71}$$

$$\forall x \in U, \forall y \in \tilde{R}(x),\ \ R(y) \subseteq \tilde{R}(x), \text{ burada } \tilde{R}(x) = \{ y \in U : y\tilde{R}x \} \tag{3.72}$$

Buradan 3 sonuç çıkarılabilir;

1. $\tilde{R}$ yansıyandır.

2. Herhangi bir benzerlik sınıfı ayırtedilmezlik sınıflarının birleşimi olarak görülebilir.

$$\tilde{R}(x) = \bigcup_{y \in \tilde{R}(x)} R(y) \tag{3.73}$$

3. $\tilde{R}$ U'nun bir örtüsünü sağlar. Bu örtü aynı zamanda R tarafından ortaya çıkarılan bölüntüyü içerir. Diğer bir deyişle $\tilde{R}$'nin yansıyan olması U'daki her nesnenin en az bir sınıf ile örtülenmesini sağlar.

Burada dikkat edilmesi gereken bir diğer nokta ise $\tilde{R}$'nin simetrik olmasının gerekmeyişidir. Diğer bir deyişle, benzerlik ilişkilerinin simetriklik özelliğini taşıyıp taşımadığı araştırılmalıdır [Slowinski ve Vanderpooten, 2000]. Örneğin "Berk babasına benzer" cümlesi "Babası Berk'e benzer" cümlesi ile eşdeğer değildir. Çünkü burada referans noktası olan "baba"dır. Dolayısı ile simetriklik kavramı tartışmaya açık bir konudur. Bu tartışmadan sonra geçişkenlik özelliği üzerine yorum yapmak gerekir. Örneğin, içinde şeker oranının hafifçe arttırıldığı 10 adet kahve fincanı düşünün. 1 ile 2'nci fincandaki kahvenin tatları birbirine benzer, hatta 1 ile 3 bile birbirine benzeyebilir, ancak 1 ve 10'uncu kahvelerin tadının birbirine benzediği

söylenemez. Dolayısı ile üzerinde çalışılan sistemin özelliği de dikkate alınmak suretiyle benzerlik ilişkilerinin kurulması önem arz etmektedir.

Klasik kaba küme içindeki eşdeğerlik ilişkisinin özellikleri;

Yansıma : $R(x,x)=1$ (3.74)

Simetriklik : $R(x,y)=R(y,x)$ (3.75)

T-geçişkenlik : $T(R(x,y),R(y,z))\le R(x,z)$ olarak gösterilebilir. (3.76)

3.2.9. Benzerlik ilişkileri

Burada, 2 küme arasındaki ilişki ile ilgilenileceği varsayımına dayanarak benzerlik ilişkileri anlatılmıştır. Literatürde, ikili ilişkiler kullanıldığı gibi daha fazla sayıda küme arasında ilişkilerin de kurulduğu çalışmalar mevcuttur.

Kati benzerlik ilişkileri;

Gündelik yaşamda ilişki, bir nesne kümesinin kendi veya diğer kümelerle arasındaki birlikteliği olarak tanımlanabilir. Örneğin Ahmet'in kırmızı Toyotası, Mehmet'in ise Renault arabası olsun. İlişkinin temeli bu birliktelikler olarak tanımlanabilir. Bu şekilde oluşturulan birliktelikler (yani insanlar ve arabalar arasındaki) bir ilişkidir. Bu tekil birliktelikleri göstermek için, Ahmet ve Toyota gibi ilişkili nesne kümeleri kullanılabilir. Bununla beraber, {Ahmet, Kırmızı Toyota} gibi basit ilişkiler yeterli bilgiyi vermeyebilir. Ahmet'in kırmızı Toyota'sının oluşu kırmızı Toyota'ların Ahmet'e ait olmasını gerektirmeyeceğinden nesnelerin sırası da ayrıca dikkate alınmalıdır. Böylece, bir ilişkiyi tanımlamak için kendi üyeleri arasındaki sıra korunmak suretiyle kurulacak kümelere ihtiyaç bulunmaktadır.

Bu kapsamda, X kümesi üzerindeki bazı önemli ikili ilişkilerin özelliklerini şu şekilde tanımlamak mümkündür;

Yansıma: X içerisindeki bütün x'ler için xRx sağlanır. Örneğin; büyük eşit operatörü yansıyandır, fakat büyük operatörü (>) yansıyan değildir.

Yansımazlık: X içerisindeki bütün x'ler için xRx sağlanmaz. Örneğin; büyük operatörü yansımazdır.

Tümleşik-yansıyan (co-reflexive): X içerisindeki bütün x'ler için xRx ise $x = y$ dir.

Simetriklik: X içerisindeki bütün x, y'ler için xRy ise yRx dir. Örneğin, biri ile kan bağının oluşu simetrik bir ilişkidir. A'nın B ile kan bağı varsa B'nin A ile kan bağı vardır.

Ters-simetriklik: X içerisindeki bütün x, y'ler için xRy ve yRx ise $x = y$ dir. Büyük eşit operatörü, x≥y ve y≥x sağlandığında x=y olduğundan ters-simetrik ilişkidir.

Asimetriklik: X içerisindeki bütün x, y'ler için xRy ise yRx olmaması durumudur. Büyük operatörü x>y ise y>x olmadığından asimetriktir.

Geçişkenlik: X içerisindeki bütün x, y, z'ler için eğer xRy ve yRz ise xRz olması sağlanır. Eğer x, y'nin öncülü, y de z nin öncülü ise x, z nin öncülü olması sonucu çıkarılabilineceğinden, bir şeyin öncülü olma durumu geçişlidir.

Doğrusallık: X içerisindeki bütün x, y'ler için xRy ve/veya yRx sağlanıyorsa doğrusaldır denir.

Kısaca, Çizelge 3.4'de benzerlik ilişkilerinin özellikleri verilmektedir.

Çizelge 3.4. Benzerlik ilişkilerinin özellikleri

Kati Benzerlik İlişkileri	Yansıma	Yansımazlık	Simetriklik	Ters simetriklik	Geçişkenlik
Eşdeğerlik	**X**		**X**		**X**
Yarı-eşdeğerlik			**X**		**X**
Tolerans	**X**		**X**		
Kısmi Sıra (Partial Order)	**X**			**X**	**X**
Kati Sıra (Strict Order)		**X**		**X**	**X**

3.2.10. Benzerlik ölçütlerinin kurulması için genel bir yaklaşım

Benzerlik ölçütlerini kurmanın en genel yolu bir uzaklık ölçütünü dikkate almaktır. Böyle bir metodun en büyük dezavantajı uzaklık ölçütünün benzerlik açısından negatif ya da pozitif bir kanıt olmamasıdır. Bununla beraber, uzaklık tabanlı benzerlik ölçütleri simetriktir.

Bu yüzden, daha genel bir kurulum olan pozitif ve negatif kanıtların uyumluluk ve uyumsuzluk denilen iki tip koşulun sağlanmasıdır [Slowinsky ve Vanderpooten, 1997].

1. Yeterince sayıdaki nitelik benzerlikle uyumlu olmalıdır.

2. Uyumlu olmayan nitelikler arasında hiçbirisi benzerlikle çok fazla (aykırı derecede) uyumsuz olmamalıdır.

Sonuç olarak, basit bir benzerlik ilişkisi;

$$C_j(x,y) = \begin{cases} 1 & |xj - yj| \le \varepsilon_j(y_j) \quad ise, \\ 0 & aksi\_hallerde. \end{cases} \tag{3.77}$$

olarak kurulabilir. Burada ε_j esneklik sağlamak için küçük bir hata oranıdır.

3.2.11. Kaba küme nitelik indirgeme algoritmalarına genel bakış

Kaba küme nitelik indirgeme ile ilgili literatürde sayısız örnek bulmak mümkündür. Fakat bunlardan en fazla bilinenleri genetik algoritma, yapay sinir ağları, tabu arama, karınca kolonisi yaklaşımı gibi arama yaklaşımları ile aşağıda sınıflandırılan algoritmik yapıların birleştirilmiş şeklidir. Bu algoritmaları şu üç kategoriye ayırmak mümkündür.

- Pozitif alanlı yaklaşımlar; [Shi ve ark., 2004], [Han ve ark., 2004], [Zhang ve Yao, 2004].

- Ayırt edilebilirlik matrisi ve nitelik sıklık ölçütlü yaklaşımlar; [Geng ve Zhu, 2006],[Wang ve Chen, 2004], [Hu ve ark., 2003], [Wang ve ark., 2006].

- Bilgi entropisi tabanlı yaklaşımlar; [Duoqian ve Jue, 1997], [Han ve Wu,2002].

Ayrıca optimal altı arama yaklaşımları ile klasik kaba küme nitelik indirgeme için Hedar ve ark, [2006] tarafından yayınlanan teknik rapor kaynak olarak alınabilir.

Bir karar tablosu birden fazla niteliğe sahip olabilir. Ancak bulunan nitelik altkümesi tüm karar tablosunun yerine kullanılabilir. Bir karar tablosundan bütün indirgemeleri bulmak NP-Zor yapıdadır [Lin, Cercone, 1997].

Pratik uygulamalarda bütün nitelik alt kümelerini bulmak gerekli değildir. Sadece bir nitelik alt kümesinin bulunması yeterli olacaktır [Hu ve Lin, 2004]. Hangi nitelik alt kümesinin seçileceği ise yararlanılan optimalite kriterine göre değişmektedir.

Kaba kümeler yaklaşımı kullanılarak literatürde birçok nitelik indirgeme çalışmaları yapılmıştır. Bu çalışmalardan dikkati çekenler ise;

Golan ve Ziarco [1995] makalesinde, veri analizi ve veri tabanlarında bilgi keşfi ile veriden muhakeme yapma gibi çoklu görevleri yerine getiren yapay zeka yazılımı olan DataLogic/R ile kaba küme indirgeme metodu önermiştir.

Duntsch ve Gediga [1997] makalesinde, kaba küme veri analizi yapabilen GROBIAN yazılımını gerçekleştirmiştir.

Hu ve ark., [2000], çalışmalarında ayırtedilebilirlik matrisi içerisinde nitelik sıklık ölçütü ve örnekleme teknikleri ile iki ayrı nitelik indirgeme algoritması önermişlerdir.

Bakar ve ark. [2000], makalesinde ikili tamsayılı programlama kullanarak, minimum indirgemeleri bulmak için bir algoritma önermişlerdir.

Beynon [2001], makalesinde değişken duyarlı kaba küme metodunda $\beta-$indirgemeleri bulmak için belli bir tolerans oranı ile nitelik indirgeme çalışmıştır.

Li ve Liu [2002], çalışmalarında niteliklerin ilgililik özelliğini bulmak için koşullu entropi kullanarak nitelik indirgemesi çalışmıştır.

Bu konuda özellik arz eden diğer bir çalışma ise pozitif alan tabanlı bir yaklaşım olan Shen ve Chouchoulas [2000] tarafından önerilen QuickReduct algoritmasıdır.

QuickReduct algoritması;

Nitelik seçimi işlemi için kullanılan algoritmaların belki de en çok bilineni QuickReduct algoritmasıdır.

QUICKREDUCT(C,D)
C; durum nitelik kümesi
D; karar nitelikleri kümesi
(1) $R \leftarrow \{\}$
(2) do
(3) $T \leftarrow R$
(4) $\forall x \in (C - R)$
(5) If $\gamma_{R \cup \{x\}}(D) > \gamma_T(D)$
(6) $T \leftarrow R \cup \{x\}$
(7) $R \leftarrow T$
(8) Until $\gamma_R(D) == \gamma_C(D)$
(9) Return R

Algoritma boş indirgeme kümesi ile başlar. Daha sonra nitelikleri teker teker indirgeme kümesine alarak bağımlılıktaki gelişmeye göre kabul eder. Bu işlem indirgeme kümesinin bağımlılık derecesinin tüm durum niteliklerinin bağımlılık derecesine eşit olana kadar devam eder. Ancak bu metot minimal altkümenin bulunmasını garanti etmeyebilir. Çünkü sadece 1 tane indirgeme bulabilir. Bu metot ekleme çıkarma stratejisi ile geliştirilebilir. Böylece daha minimale yakın sonuçlar bulunabilir.

Bu algoritma ile ilgili diğer bir strateji geri dönüş stratejisi olabilir. Bu durumda nitelik kümesi başlangıçta tüm nitelik kümesi ile başlar ve bağımlılık derecesinde fark yaratmayan nitelikler sırasıyla atılabilir [Yao ve ark., 2006].

Ayırtedilebilirlik matrisinden indirgeme bulmak, kaba küme literatürü içinde yoğunlukla uğraşılan konulardan biri olmuştur.

Bütün nitelik alt kümelerinden optimal indirgemeyi bulmak ise NP-zor yapıdadır [Geng ve Zhu, 2006]. Skowron ve Rauszer [1992] ayırtedilebilirlik matrisi ile ilgili özellikler ispatlayarak kor, indirgeme ve bağımlılıkların bulunması gibi konularda etkin algoritmalar tasarlamışlardır.

Bu kapsamda, Wang ve ark., [2006] tarafından bir algoritma önerilmiştir.

Burada, matris içerisinde en sıklıkla görülen nitelik indirgeme kümesine atılmakta ve bununla kesişim yapan diğer nitelik alt kümeleri ayırtedilebilirlik matrisinden silinmektedir. Bu işleme hiçbir nitelik altkümesi kalmayana kadar devam edilmektedir. Ayırtedilebilirlik matrisinin elemanları şu şekilde hesaplanmaktadır;

$$c_{ij} = \begin{cases} \phi & f_D(x_i) = f_D(x_j); \\ \{a \in A : a(x_i) = a(x_j)\}, & f_D(x_i) = f_D(x_j) \end{cases} \tag{3.78}$$

$M(IS)$ simetrik ve $c_{ii} = \phi \quad i = 1,2,..n$ olduğundan sadece alt üçgen $1 \le j < i \le n$ dikkate alınacaktır.

A'da bütün vazgeçilemez niteliklerin hepsine $Çekirdek(A)$ denir ve şu şekilde ifade edilir.

$$Çekirdek(A) = \{a \in A : c_{ij} = \{a\} bazı\ i, j\ için\} \tag{3.79}$$

Eğer B, A'da bağımsızsa ve $IND(B) = IND(A)$ ise $B \subseteq A$ ifadesini karşılayan tüm kümelere indirgeme denir. Dolayısı ile B içeren herhangi bir girdinin c_{ij} ile kesişiminin boş küme olmayacağı açıktır.

Anlatılanlar ışığında oluşturulan algoritma aşağıda sunulmaktadır:

Girdi: Enformasyon sisteminin ayırtedilebilirlik matrisi
Çıktı: Enformasyon sistemine ait bir indirgeme

(1) $R = \phi$
(2) M'de tekrarlanan tüm elemanları sil
(3) M'den Çekirdeği hesapla
(4) $R \Leftarrow C_0$
(5) M Matrisinden M_R'yi hesapla

(6) M_R deki elemanları kardinalitesine göre ayır. Kardinal sayısı i ; $L_i, i = 1,2..n$ ile ifade edilir.

(7) $a \in A - R$ için $i_0 = \min_{L_i \neq 0}\{i\}$ olsun. $SGF(a,R,A)\big|L_{i_0}$ ifadesini L_{i_0} dan hesapla. $SGF(a',R,A)\big|L_{i_0} = \underset{a \in A-R}{Max}\left\{SGF(a,R,A)\big|L_{i_0}\right\}$ olarak hesapla.

(8) $R \Leftarrow R \cup \{a'\}$

(9) $M_R = \phi_{nxn}$ elde edilene kadar devam et.

Örnek:

Bir bilgi sisteminden ac(1), ad(1), bc(1), bd(1), cd(1), abc(2), abd(2), abcd(144) olsun. Parantez içerisindeki sayılar M'de kaç defa gözlemlendiğini göstermektedir. $C_0 = 0$ olduğu açıktır. M'deki tekrarlanan elemanları silerek ve M_R'deki elemanları kardinalitesine göre sıralayalım. Böylece;

$L_1 = \{\phi\},$
$L_2 = \{ac, ad, bc, bd, cd\},$
$L_3 = \{abc, abd\},$
$L_4 = \{abcd\}.$
elde edilir.

Böylece kardinalitesi (eleman sayısı) 1 olan eleman bulunmadığına göre algoritma 2'li elemanlara bakacaktır.

$SGF(a,R,A)\big|L_2 = 2$, $SGF(b,R,A)\big|L_2 = 2$, $SGF(c,R,A)\big|L_2 = 3$, $SGF(d,R,A)\big|L_2 = 3$ elde edilir. $SGF(c,R,A)\big|L_2 = SGF(d,R,A)\big|L_2 = 3$ ve L_3 L_4 seviyelerinde de eşit olduklarından c veya d niteliklerinden biri seçilebilir. $R = \{c\}$ seçilir ise, $L_1 = \{\phi\}$, $L_2 = \{ad, bd\}$, $L_3 = \{abd\}$, $L_4 = \{\phi\}$ elde edilir. Diğer adımda ise hala 2'li elemanlar ele alınır ve d niteliği seçilirse $R = \{c, d\}$ bulunur. Dolayısı ile c,d nitelikleri bu bilgi sistemi için indirgeme olarak bulunur.

3.3. Bulanık Kaba Kümeler

Kesin olmayan verilerle ilgili problemlerle başa çıkabilmek için önerilen kaba ve bulanık kümelerin her ikisi de klasik küme metodunu genelleştirmektedir. Ancak, kaba küme ile bulanık küme yaklaşımı belirsizlik ve muğlaklık problemlerine değişik açılardan bakar. Bulanık küme bilgiyi yaklaşık olarak üyelik dereceleri kullanarak tanımlar. İnsanın konuşma diline has bulanık belirsizlikle ilgilenir. Kabalık kavramı ise enformasyonun granüler yapısının bir sonucudur. Eğer nesneler aynı tanıma sahipken değişik sınıflarla ifade ediliyorsa böyle bir tanımlamada kaba belirsizlik görülür. Bu aslında iki nesnenin aynı olduğu manasına gelmez, ancak bu durum yorumlayan kişinin bu nesneler hakkında kısıtlı kavrama derecesine sahip olduğundan ileri gelmektedir. Böyle bir kavrama derecesinde değişik nesneleri aynı olarak tanımlayabiliriz. Kaba belirsizlik, kavrama seviyesi arttıkça azalacaktır ve enformasyon granülü daha saf hale gelecektir.

Klasik kaba küme metodu kesikli değer içeren veritabanları ile etkin olarak çalışır. Ancak, gerçekte birçok veri tabanı sürekli, tamsayılı ya da kategorik değerler alabilmektedir. Kaba küme metodunda verilerin hazırlanması için kesikli hale getirme işlemi yapılmalıdır. Bu işlem esnasında birçok önemli bilgi kaybının ya da karışıklıkların yaşanması muhtemeldir.

Çizelge 3.5 belirli öğrencilere ait not karar tablosu olsun. Kesikli hale getirme işlemi için kesme noktaları 60 ve 85 olarak kullanılırsa Çizelge 3.6 elde edilmektedir.

Aslında Çizelge 3.5 deki 2 ve 3'ncü kayıtlar birbirine yakınken Çizelge 3.6'daki durum farklıdır. Böylece, orijinal karar tablosundaki bazı önemli enformasyon kaybolmuştur. Diğer bir deyişle, sürekli değer alan gözlemler kesikli hale getirme işlemine tabi tutulduktan sonra sınırların keskin olması sebebiyle tamamen değişik bir sınıfa geçebilmektedir.

Çizelge 3.5. Not karar tablosu

ÖĞRENCİ	FİZİK	MATEMATİK	DEĞERLENDİRME
1	60	95	Orta
2	85	97	İyi
3	84	67	Orta
4	53	62	Kötü
5	97	94	İyi
6	90	80	Orta
7	98	53	Orta
8	41	44	Kötü

Çizelge 3.6. Kesikli hale getirme işlemi sonrası karar tablosu

ÖĞRENCİ	FİZİK	MATEMATİK	DEĞERLENDİRME
1	Orta	İyi	Orta
2	İyi	İyi	İyi
3	Orta	Orta	Orta
4	Kötü	Orta	Kötü
5	İyi	İyi	İyi
6	İyi	Orta	Orta
7	İyi	Kötü	Orta
8	Kötü	Kötü	Kötü

Bulanık küme teorisi bu probleme etkin çözümler sağlamaktadır. Dolayısı ile bulanık küme ile kaba küme yaklaşımı beraber kullanılmasının bilginin en az kaybının sağlanması için gereklidir. Burada, evren bulanık olarak bölünebilir ve bulanık sınıflar birbirlerinin üstüne bindirilebilir. Her bir nitelik için uzmanlar tarafından bir üyelik fonksiyonu tanımlanır ve niceleyici değerler üyelik fonksiyonu yardımıyla bulanık kümelere çevrilebilir. Bu örnek için üyelik fonksiyonları aşağıdaki şekilde tanımlanabilir:

$$\dot{I}(x)=\begin{cases}0 & x\le 80\\ 1-\left(\dfrac{2x-190}{110-80}\right)^2 & 80<x<95\\ 1 & x\ge 95\end{cases} \tag{3.80}$$

$$K(x)=\begin{cases}1 & x\le 50\\ 1-\left(\dfrac{2x-100}{65-35}\right)^2 & 50<x<65\\ 1 & x\ge 65\end{cases} \tag{3.81}$$

$$O(x)=1-\dot{I}(x)-K(x) \tag{3.82}$$

Uzmanlardan elde edilen bu üyelik fonksiyonları ile elde edilen değerler Çizelge 3.7'de sunulmaktadır.

Çizelge 3.7. Bulanık karar tablosu

ÖĞRENCİ	**FİZİK**			**MATEMATİK**			**DEĞERLENDİRME**
	İ	**O**	**K**	**İ**	**O**	**K**	
1	0	0,44	0,56	1	0	0	Orta
2	0,56	0,44	0	1	0	0	İyi
3	0,46	0,54	0	0	1	0	Orta
4	0	0,04	0,96	0	0,64	0,36	Kötü
5	1	0	0	0,99	0,01	0	İyi
6	0,89	0,11	0	0	1	0	Orta
7	1	0	0	0	0,04	0,96	Orta
8	0	0	1	0	0	1	Kötü

Burada açıklanan konunun ışığında bulanık kümeler niteliklerinin isimlerine bağlı olarak (fizik, matematik) ya da olmadan ifade edilmektedir. Aşağıda tip–1 ve tip–2

bulanık kümeler açıklanmaktadır. Bulanık kaba kümelerle çalışırken tip–1 ve tip–2 şeklinde ayrılan bulanık değerler için ayrı matematiksel oluşumlar gerektirmektedir.

Tip-1 bulanık kümeler; N nesneli bir veri kümesinden her bir örnek x_i $(i \in [1, N])$ bulanık olan nitelikler veya N_F karakteristik ile tanımlanır. Bunlar μ_i^j, $j = 1,..,N_F$ kısmi üyelik dereceleri ile ölçülür.

$$x_i = [\mu_i^1, \mu_i^2, ..., \mu_i^{N_F}] \tag{3.83}$$

μ_i^j değerlerinin tutarlı olması için genellikle zayıf bulanık bölüntü oluşturmaları gerekir.

Örneğin, Çizelge 3.8'de Bodjanova [1997]'nin gösterimi kullanılmıştır. Burada, C1 düşük banka hesabı, C2 orta banka hesabı, C3 yüksek banka hesabı, C4 düşük aylık gelir, C5 orta aylık gelir ve C6 yüksek aylık gelir anlamına gelmektedir. D_1, D_2 ve D_3 ise karar niteliğini ifade etmektedir.

Çizelge 3.8 de görülebileceği gibi; $x_1 = [0;0,8;0,3;0,7;0,2;0]$ olduğu gözlemlenecektir.

Çizelge 3.8. Tip 1 bulanık bilgi sistemi

	C1	**C2**	**C3**	**C4**	**C5**	**C6**	**D1**	**D2**	**D3**
X1	0	0,8	0,3	0,7	0,2	0	0,5	0,7	0
X2	1	0	0	0	0,1	0,8	0	0	1
X3	0,6	0,3	0	0	1	0	0	0,5	0,7
X4	0	0,5	0,6	0	0	1	0,6	0,5	0
X5	0	1	0	0	0,7	0,5	0	1	0
X6	0	0,1	0,8	1	0	0	0,8	0,4	0
X7	0	0	1	0,4	0,6	0	1	0	0
X8	0	0	0	0,5	0,5	0	0	0	1

<u>*Tip-2 bulanık kümeler*</u>; N nesneli bir veri kümesi olsun. x_i $(i \in [1, N])$ örneği N_F adet karakteristik ile tanımlanırken her bir j niteliğinin değeri N_{LL} üyelik derecesi kümesi ya da dilsel derece ile μ_i^{jk} $(k = 1,..., N_c LL)$ olacak şekilde belirlenmektedir. Burada, her bir x_i değeri şu şekilde gösterilmektedir.

$$x_i = [(\mu_i^{11},...,\mu_i^{1k},...,\mu_i^{1N_{LL}}),(\mu_i^{21},...,\mu_i^{2k},...,\mu_i^{2N_{LL}}),....,(\mu_i^{j1},...,\mu_i^{jk},...,\mu_i^{jN_{LL}}),...,$$
$$(\mu_i^{N_F 1},...,\mu_i^{N_F k},...,\mu_i^{N_F N_{LL}})] \qquad (3.84)$$

Bu tip bir ifade ile Çizelge 3.8'deki gösterim; $x_1 = [(0,0.8,0.3),(0.7,0.2,0)]$ şeklinde olacaktır.

Burada, bu iki gösterim arasında herhangi bir fark olmayacağı iddia edilse bile, Eş.(3.87) şeklinde bir gösterim her bir niteliğe ayrı bir önem derecesi (ağırlık değeri) verilebilmesi bakımından kullanışlı olduğu değerlendirilmektedir.

3.3.1. Bulanık Eşdeğerlik Sınıfları

Klasik eşdeğerlik sınıfları nasıl kaba küme teorisinin merkezinde bulunuyorsa bulanık eşdeğerlik sınıfları da bulanık kaba kümelerin merkezindedir [Thiele, 1998].

Bulanık bir enformasyon sisteminde hem durum nitelikleri hem de karar nitelikleri bulanık değerler alabilir. Klasik eşdeğerlik sınıfları, bir bulanık benzerlik ilişkisi S ile genişletilebilir. Örneğin $\mu_S(x, y) = 0{,}9$ eşitliği x ve y nesnelerinin birbirine oldukça benzediğini ifade eder. Benzerlik ilişkisi aslında eşdeğerlik ilişkisinin genelleştirilmiş halidir. Daha açıkça bir benzerlik ilişkisi U evreninde S ile gösterilen bir bulanık ilişkidir. Bu ilişki yansıyan, simetrik ve T-geçişlidir [Dubois ve Prade, 1990;1992].

Yansıma ; $S(x, x) = 1,\ \forall x \in U$ (3.85)

Simetriklik ; $S(x, y) = S(y, x),\ \forall x, y \in U$ (3.86)

T-geçişkenlik ; $S(x,z) \geq T(S(x,y),S(y,z)),\ \forall x,y,z \in U$ (3.87)

Burada Pawlak'ın önerdiği klasik eşdeğerlik ilişkisi, R, bulanık bir ilişkiyi gösteren S'ye çevrilmiştir. Buradaki t-norm değişme, birleşme ve monotonluk özelliği olan bir bütünleştirme işlemidir. Bu işlem Zadeh'in orijinal benzerlik ilişkisindeki sınır koşullarına uyar.

R ilişkisinin direkt olarak bir bulanık ilişki ile yer değiştirmesinin yerine kendi eşdeğerlik sınıfları, U/R, Φ bulanık kümeler ailesi ile zayıf bulanık bölüntü oluşturmak için yer değiştirilebilir. Bu durumda şu koşullar da sağlanmalıdır.

$$\inf_x \max_{i=1} \mu_{F_i}(x) > 0 \tag{3.88}$$

$$\sup_x \min(\mu_{F_i}(x), \mu_{F_i}(x)) < 1 \tag{3.89}$$

İlk şart, Φ bulanık ailesinin evrenin (U) tamamını kapsamasını garanti eder. İkinci şart ise bulanık kümeler arasında ayrıklık koşuludur.

3.3.2 Bulanık kaba kümelerin genel özellikleri

$$R_*U = U, R*\phi = \phi; \tag{3.90}$$

$$R_*(A \cap B) = R_*A \cap R_*B,\ \ R*(A \cup B) = R*A \cup R*B; \tag{3.91}$$

$$R_*A^C = (R*A)^C\ ,\ R*A^C = (R_*A)^C; \tag{3.92}$$

$$R_*A \subseteq A\ ,\ A \subseteq R*A; \tag{3.93}$$

$$R_*(U - \{y\})(x) = R_*(U - \{x\})(y)\ ,\ R*x_1(y) = R*y_1(x); \tag{3.94}$$

$$R_* A \subseteq R_*(R_* A),\ R^*(R^* A) \subseteq R^* A \tag{3.95}$$

Burada $A, B \in F(U),\ x, y \in U,\quad x_\lambda(y) = \begin{cases} \lambda, y = x \\ 0, y \neq x \end{cases}$ (3.96)

[Morsi ve Yakout, 1998**].**

3.3.3 Bulanık kaba kümeler ile yapılan nitelik indirgeme çalışmaları

Bulanık kaba kümelerin oluşturulması ve beraberce veri madenciliği uygulamalarında kullanılması için birçok çalışmalar yapılmıştır. Bulanık kaba küme uygulamalarında nitelik indirgeme çalışmaları öğrenme algoritması çalışmalarına nazaran daha azdır. Bu kapsamda, bulanık eşdeğerlik özelliklerinin gevşetilmesi ve bulanık içerme ile bulanık kümelerin α-kesmelerinin kullanılması suretiyle değişik bulanık kaba küme tanımları yapılmıştır.

Bulanık ve kaba kümelerin melezlenmesinde, belitsel ve kurucu olmak üzere iki ana yaklaşım bulunmaktadır [Yeung ve ark., 2005].

Kurucu yaklaşımda, evrendeki bulanık ilişkiler birincil kavramdır ve alt ve üst yaklaşımlar bu kavramlar üzerine kurulur. Başlangıçta sadece bulanık eşdeğerlik ilişkileri kullanılırken daha sonraki genelleştirme çalışmalarında gelişigüzel ikili ilişkiler de değerlendirmeye alınmıştır.

Diğer yandan belitsel yaklaşımda ise, alt ve üst yaklaşım operatörleri birincil kavramdır. Bu yaklaşımda, yaklaşım operatörlerini tanımlamak için bir takım belitler kullanılmıştır [Wu ve Zhang, 2004].

Buradan da anlaşılacağı gibi kurucu yaklaşım daha çok pratik uygulamalar üzerinde, belitsel yaklaşım ise bulanık kaba kümelerin matematiksel karakteristikleri üzerinde çalışmaktadır.

Bulanık kaba küme yaklaşımını ilk öneren yazarlar Dubois ve Prade'dir [1990; 1992]. Çalışmalarında, t-norm min ve t-conorm max operatörleri kullanarak bir bulanık benzerlik ilişkisine göre alt ve üst yaklaşım çiftini kurmuştur.

Makalelerinde; U evreni, R bu evren üzerindeki ikili ilişkiyi tanımladığında, bulanık kaba kümenin alt ve üst yaklaşımları sırası ile;

$$R_*(F)(x) = \inf_{y \in U} \max(1 - (R(x, y), F(y)) \tag{3.97}$$

$$R^*(F)(x) = \sup_{y \in U} \min(1 - (R(x, y), F(y)) \text{ ile tanımlamışlardır.} \tag{3.98}$$

Morsi ve Yakout [1998], bulanık kaba kümeler üzerinde bir takım belitler üzerine çalışmıştır. Çalışmalarında bulanık kaba kümeleri bulanık t-benzerlik ilişkileri üzerine kurmuşlardır.

Thiele [2000a; 2000b; 2001], bulanık kaba yaklaşım operatörlerinin belitsel tanımlamaları ile değişik iki (diamond ve box) bulanık operatör yardımıyla modal mantık içerisinde kaba bulanık yaklaşım operatörlerini incelemiştir.

Radzikowska ve Kerre [2002] kaba kümelerin bulanıklaştırılması için daha genel bir metot önererek, bir t-norm ile belirlenen bulanık benzerlik ilişkisine göre bulanık kaba kümeleri tanımlamıştır.

Bulanık-kaba küme metodunu kullanarak ilk nitelik indirgeme çalışması [Kuncheva 1992] tarafından yapılmıştır. Çalışmasında bulanık örüntü tanıma ile uygulamalar yapmıştır.

Jensen ve Shen [2002a;b, 2004a;b;c, 2005a;b] tarafından önerilen yaklaşımda, bulanık pozitif alan ve bulanık alt yaklaşım kavramlarından faydalanarak bağımlılık

fonksiyonu tanımlamış ve bulanık-kaba QuickReduct algoritması önermişlerdir. Algoritmalarında, alt ve üst yaklaşımlar;

$$\mu_{\underline{P}X}(x) = \sup_{F \in U/P} \min(\mu_F(x), \inf_{y \in U} \max\{1 - \mu_F(y), \mu_X(x)\} \tag{3.99}$$

$$\mu_{\overline{P}X}(x) = \sup_{F \in U/P} \min(\mu_F(x), \sup_{y \in U} \min\{\mu_F(y), \mu_X(x)\} \tag{3.100}$$

olarak ifade edilmiştir.

Klasik kaba küme yaklaşımında pozitif alan alt yaklaşımların toplamı idi. Bir x niteliğinin bulanık pozitif alana aidiyeti ise;

$$\mu_{POS_P(Q)}(x) = \sup_{X \in U/Q} \mu_{\underline{P}X}(x) \tag{3.101}$$

Bulanık pozitif alan tanımlandıktan sonra bağımlılık fonksiyonu;

$$\gamma_P(Q) = \frac{\left|\mu_{POS_P(Q)}(x)\right|}{|U|} = \frac{\sum_{x \in U} \mu_{POS_P(Q)}(x)}{|U|} \quad \text{ile tanımlanır.} \tag{3.102}$$

Klasik kaba küme teorisi ile aynı şekilde eğer x niteliğinin pozitif alana ait olduğunda 1 değerini alsaydı aynı ifade elde edilebilirdi.

Herhangi bir a niteliği için {a} ile bölünen evren $U/IND\{(a)\}$ ile gösterilir ve o nitelik için bulanık eşdeğerlik sınıflarının kümesi olarak ele alınır.

$$U/P = \otimes\{a \in P : U/IND\{(a)\}\} \tag{3.103}$$

Burada dikkat çekilebilecek diğer önemli bir husus nitelik grupları ile çalışırken bulanık kaba yaklaşımının nasıl çalışacağıdır. Bunun için bulanık küme teorisindeki kartezyen çarpımı kullanmak yeterli olacaktır. Yani;

$$A \otimes B = \{X \cap Y : \forall X \in A, \forall Y \in B, X \cap Y \neq \phi\} \text{ dir.} \quad (3.104)$$

Eğer S kesin (kati) ise yansıyan, simetrik ve t-geçişkenlik gereksinimleri S'yi klasik eşdeğerlik ilişkisi haline getirir.

U/P deki her bir küme bir eşdeğerlik sınıfıdır. Örneğin eğer P={a,b} ise, $\text{U/IND}(\{\text{a}\}) = \{N_a, Z_a\}$ ve $\text{U/IND}(\{\text{b}\}) = \{N_b, Z_b\}$ dir.

Buradan; $U / P = \{N_a \cap N_b, N_a \cap Z_b, Z_a \cap N_b, Z_a \cap Z_b\}$ elde edilir. Bu yapının üzerine herhangi bir arama stratejisi kullanılarak algoritma üretmek de mümkündür.

Jensen ve Shen [2002] bu aşamaları kullanarak aşağıdaki algoritmayı geliştirmişlerdir.

FRQUICKREDUCT(C,D)

C; Durum nitelik kümesi

D; Karar nitelik kümesi

(1) $R \leftarrow \{\}; \gamma_{best} = 0; \gamma_{prev} = 0$

(2) do

(3) $T \leftarrow R$

(4) $\gamma_{prev} = \gamma_{best}$

(5) $\forall x \in (C - R)$

(6) If $\gamma_{R \cup \{x\}}(D) > \gamma_T(D)$

(7) $T \leftarrow R \cup \{x\}$

(8) $\gamma_{best} = \gamma_T(D)$

(9) $R \leftarrow T$

(10) Until $\gamma_{best}(D) == \gamma_{prev}$

(11) Return R

FRQR algoritmasının çalışma yapısı QuickReduct algoritması ile benzerlikler göstermekle beraber, algoritma eldeki en iyi nitelik alt kümesinin bağımlılık derecesi geliştirilemeyinceye kadar devam eder.

Bağımlılık ölçütü monoton olmadığından, QuickReduct stili diğer bir ifade ile tepe tırmanma türü araştırma stratejisi ile sadece yerel optimuma ulaşıldığında algoritmaya son verilir. Ancak optimum nokta araştırma uzayının başka bir yerinde olabilir. Dolayısı ile bu metot diğer araştırma metotları kullanılarak tüm araştırma uzayını da kapsayacak şekilde genişletilmelidir.

Jensen ve Shen, FRQR algoritmasını; web sitesi sınıflandırılma, karmaşık sistemlerin takibi ve adli olaylar gibi pratik uygulamalarda kullanmışlardır.

FRQR algoritmasının sadece tek indirgeme bulması, matematiksel temelden ve teorik analizden yoksun oluşu, sonlandırma kriterinin zayıf tasarımı ve indirgeme yapısının açık olmayışı nedeniyle Bhatt ve Gopal [2005a;b;c] tarafından eleştirilmiştir. Yazarlar, büyük bulanık veri tabanları için sıkı hesapsal alan önererek FRQR algoritmasının hesaplama performansını geliştirmişlerdir.

Parthalain ve ark. [2006] Shannon entropisi kullanarak daha küçük genişlikte nitelik indirgeme bulmaya çalışmışlardır. Makalelerinde, basit bir tepe tırmanma algoritması ile her aşamada her bir nitelik veya nitelik grubunun bulanık entropisini hesaplayarak indirgeme kümesine atamışlardır.

Bulanık yaklaşım uzaylarının, bulanık olasılıklı yaklaşım uzaylarına genelleştirilmesi ile yapılan indirgeme çalışmaları ise Hu ve ark., [2006] tarafından çalışılmıştır. Çalışmalarının, hem kategorik hem de sürekli nitelikleri içeren karar sistemlerine uygulanması bakımından önemli olduğu değerlendirilmektedir.

Salido ve Murakami, [2003] makalesinde kurucu bir yaklaşım önererek tip-2 bulanık kümelerin kullanılması suretiyle nitelik indirgeme ve sınıflandırma uygulamaları yapmış, en sık kullanılan t-norm'ların oluşturulması ve bütünleştirilmesi için sağlam bir alt yapı sağlamışlardır.

Tsang ve ark., [2005] ile Wang ve ark., [2005], tip-1 bulanık karar sistemleri için göreli indirgemeleri bulabilecek bir algoritma tasarlamışlardır. İndirgeme altkümelerini bulmak için bir ayırtedilebilirlik matrisi önermişlerdir. Ancak sonuçta elde edilen matrisin kati doğasından dolayı bilgi kaybı olabileceği konusunda eleştiriler almışlardır.

Chen ve ark. [2007], makalelerinde, Tsang ve ark. [2005] önerdiği nitelik indirgeme algoritmasını t-bulanık kaba kümelere uygulamış ve göreli indirgemeleri bulabilecek bir algoritma tasarlamışlardır.

4. NİTELİK İNDİRGEME İÇİN YENİ BİR ALGORİTMA

Bu bölümde, bulanık kaba kümeler yardımıyla ayırtedilebilirlik matrisi tabanlı bir nitelik indirgeme algoritması tasarlanmış; hazırlanan algoritma örnekle açıklanmış ve diğer nitelik indirgeme algoritmalarıyla karşılaştırılarak etkinliği test edilmiştir.

4.1. Önerilen Algoritma

4.1.1. Ön Hazırlık işlemleri

Bulanık kaba kümeler ile nitelik indirgeme yapabilmek için bazı ön hazırlıklar yapılmalıdır. Bunlardan birincisi; indirgemeden sonra neyin sabit kalacağı konusunda karar verilmesidir. Burada, klasik kaba küme teorisinde de olduğu gibi, karar niteliği pozitif alanının korunması fikri takip edilecektir. Bölüm 3'de anlatıldığı gibi, karar niteliğin pozitif alanı; karar sınıflarının alt yaklaşımlarının birleşimi olarak tanımlanır.

İkincisi, benzerlik ilişkilerinin oluşturulması için hangi t-normun kullanılacağıdır. Çalışmada tip-2 bulanık verisi kullanıldığından, hangi t-norm'un seçileceğine karar vermek için Valverde [1985]'nin t-norm dönüşüm çalışmasından faydalanılmıştır. Bu dönüşüm, tip-2 bulanık verilerin kullanımı için gereklidir.

Yakın-ters kavramına dayanarak bulanık veriden t-benzerlik ilişkisi kurma metodu aşağıda gösterilmektedir.

t-norm teorisinden de bilindiği gibi [Klementa ve ark., 2004].arşimedyan t-norm şu formüle göre elde edilmiştir;

$$T(x, y) = g^{[-1]}(g(x) + g(y)), \tag{4.1}$$

Burada $g(x)$ sürekli mutlak azalan $[0,1] \to [0,\infty]$ bir fonksiyondur. Burada $g(1) = 0$, $g^{[-1]}$; g 'nin sözde tersini gösterir ve $\boldsymbol{g^{[-1]}(x) = g^{-1}(\min(x, g(0)))}$ ile ifade edilir.

t-norm $T(x, y)$ 'nin yakın-ters işlemi $T\hat{}(x|y)$ ile gösterilir.

$$T\hat{}(x|y) = g^{[-1]}(g(y) - g(x)) \tag{4.2}$$

Eğer iki x_i ve x_j örneği ele alınırsa, her iki örnekteki k niteliği, N_{LL} dilsel değişken kümesine üyelik derecesi vektörü olarak tanımlanabilecektir. Diğer bir deyişle, $x_i^k = (\mu_i^{k1}, ..., \mu_i^{kl}, ..., \mu_i^{kN_{LL}})$ ve $x_j^k = (\mu_j^{k1}, ..., \mu_j^{kl}, ..., \mu_j^{kN_{LL}})$ olarak gösterilebilecektir. Valverde'ye göre k niteliğine bağlı olarak x_i ve x_j arasındaki t-benzerlik şu şekilde çıkarılabilir;

$$S_{ij}^k(x_i, y_j) = \underset{I \in [1,...N_{LL}]}{Inf} \; T\hat{}(Max(\mu_i^{kl}, \mu_j^{kl}) \,|\, Min(\mu_i^{kl}, \mu_j^{kl})). \tag{4.3}$$

Aşağıda Valverde'nin işlemi yardımı ile belirli t-normlar için yapılan dönüşümler gösterilmektedir [Salido ve Murakami, 2003].

Minimum operatörü $T(x, y) = Min(x, y)$;

$$S_{ij}^k(x_i, y_j) = \begin{cases} \inf_{l \in J_{ij}^k}(\min(\mu_i^{kl}, \mu_j^{kl})) & \text{Eğer} \quad J_{ij}^k \neq 0 \\ 1 & \text{Diğer hallerde.} \end{cases} \tag{4.4}$$

Çarpım operatörü $T(x, y) = xy$;

$$S_{ij}^k(x_i, y_j) = \begin{cases} \inf_{l \in J_{ij}^k}(\min(\dfrac{\mu_i^{kl}}{\mu_j^{kl}}, \dfrac{\mu_j^{kl}}{\mu_i^{kl}})) & \text{Eğer} \quad J_{ij}^k \neq 0 \\ 1 & \text{Diğer hallerde.} \end{cases} \tag{4.5}$$

Lukasiewicz operatörü $T(x, y) = Max(x + y - 1, 0)$

$$S_{ij}^{k}(x_i, y_j) = \inf_{I \in J_{ij}^{k}} (1 - \left| \mu_i^{kl} - \mu_j^{kl} \right|) \tag{4.6}$$

Yager operatörü ($T(x, y) = 1 - \min((1, ((1-x)^p + (1-y)^P)^{1/p}))$):

$$S_{ij}^{k}(x_i, y_j) = \inf_{I \in J_{ij}^{k}} (1 - \left| (\mu_i^{kl})^p - (\mu_j^{kl})^p \right|^{1/p}) \tag{4.7}$$

Örneğin, elimizdeki veri kümesinin iki elemanı, (0,7; 0,3) ve (0,71; 0,29) olsun. Eldeki verilere göre bu iki kümenin değişik operatörlere göre benzerlikleri sırasıyla; minimum operatörüne göre, 0,29; çarpım operatörüne göre, 0,966; Lukasiewicz operatörüne göre, 0,99 ve p=2 için Yager operatörüne göre 0,9859 değerleri elde edilmektedir. Burada birbirine çok yakın bulanık iki değer arasındaki benzerlik derecesi 1'e daha yakın olduğundan dolayı Lukasiewicz operatörünün kullanımı mantıklı bir yol olarak görülmektedir.

Üçüncüsü, t-bulanık ilişkileri için uygun bütünleştirme operatörünün seçilmesidir. Çünkü t-bulanık benzerlik ilişkilerinin birleştirilmesi sonucunda elde edilen ilişkinin de t-benzerlik ilişkilerinin özelliklerine sahip olması gerekmektedir. Aşağıda bütünleştirme operatörünün oluşturulması için önem arz eden t-geçişkenlik teoremi verilmektedir.

Teorem 4.1

$R_1, R_2, ...R_n$, T_L geçişli ilişki olsun. Burada T_L Lukasiewicz t-norm'dur. Sadece ve sadece $M(x)$ in De morgan duali $N(x) = 1 - M(1 - x_1, ..., 1 - x_n)$, $\forall x, y, z \in [0,1]^n \left| z = x + y, \right.$ olacak şekilde, $\boldsymbol{N(z) \le N(x) + N(y)}$ ifadesini sağlarsa $R = M(R_1, R_2, ...R_n)$ de T_L geçişlidir denir [Salido ve Murakami, 2003].

Aslında burada yatan problem uyumlu (M,T) çiftinin bulunmasıdır. Bahsedilen şekilde bir çiftin bulunması problemi bulanık küme teorisi içerisinde pek çalışılmamıştır. Ovchinnikov [1992], eğer t-geçişli benzerliklerin bütünleştirmesi sonucunda elde edilen ilişki de t-geçişli ise bütünleştirme operatörüne "*iyi tanımlanmış*" demektedir. Burada asıl problem iyi tanımlanmış $M(T)$ çiftinin bulunmasıdır. $M(x)$ üzerinde yapılabilecek tek kısıtlama monoton operatör olması ve sınır koşullarının $M(0)=0$ ve $M(1)=1$ sağlanmasıdır. Buna ilaveten, bütünleştirme operatörü belli bir uzaklık fonksiyonundan çıkarılabilmektedir. Bu tip operatörlerin özelliği bütünleştirme operatörünün minimum ile maximum arasında belirlenebilmesi ve kullanıcıya esneklik sağlamasıdır.

Bu yolla genelleştirilmiş ortalama operatörü Minskowsky uzaklığından direkt olarak çıkarılmıştır.

$$M(x)=(x_1^p+x_2^p+...+x_n^p)/n)^{1/p} \tag{4.8}$$

Teorem-4.1'den T_L çatısında geçişli bir benzerlik toplama operatörü olarak genelleştirilmiş ortalamaların De Morgan duali kullanarak;

$$M(x)=1-(((1-x_1)^p+(1-x_2)^p+....+(1-x_n)^p)/n)^{1/p} \quad \text{elde edilir.} \tag{4.9}$$

Burada p değerinin azalışı ve artışı ile minimum ve maximum operatörüne yaklaşılır.

Yukarıda tanımlanan ön hazırlıkların tamamlanması ile önerilen algoritma çözüme hazır hale gelmektedir.

4.1.2. Alt yaklaşım operatörüne göre ayırtedilebilirlik matrisinin oluşturulması

Bölüm 3'de ayırtedilebilirlik matrisinin kaba kümelerde kullanımı hakkında tanımlar verilmiştir. Bu bölümde ise, bulanık kaba kümeler için ayırtedilebilirlik matrisinin

oluşturulması maksadıyla Tsang ve ark. [2005]'nın tip-1 bulanık kümelerden indirgeme bulmak için geliştirdikleri algoritma temel alınarak tip-2 bulanık kümeler için yeni bir algoritma tasarlanmıştır.

U evreni, $\Re$, durumsal nitelik kümesi denilen t-benzerlik ilişkilerinin sonlu kümesi ile D, sembolik değerlerden oluşan karar niteliği denilen bir eşdeğerlik ilişkisini ifade etsin. $(U, \Re \cup D)$ ifadesine bulanık karar sistemi denir. $Sim(\Re) = \cap\{R : R \in \Re\}$ ile gösterilirse $Sim(\Re)$'de aynı zamanda bulanık t-benzerlik ilişkisidir.

Bütün $x \in U$ için $[x]_D$, D'ye göre eşdeğerlik sınıfı olduğunda, $Sim(\Re)$'ye göre D'nin pozitif alanı $POS_{Sim(\Re)}(D) = \bigcup_{x \in U} \underline{Sim(\Re)}([x]_D)$ ile ifade edilir. Eğer $POS_{Sim(\Re)}D = POS_{Sim(\Re - \{R\})}D$ ise, R'ye $\Re$ içerisinde D'ye göre *vazgeçilebilir* denir. Aksi halde "*vazgeçilemez*"dir. Eğer her bir $R \in \Re$, $\Re$ içerisinde D'ye göre vazgeçilemez ise $\Re$ ailesi D'ye göre *bağımsızdır* denir.

$\mathrm{P} \subset \Re$, D'ye göre niteliklerin bir indirgemesi olsun. Eğer P, D'ye göre bağımsız ve $POS_{Sim(\Re)}D = POS_{Sim(\mathrm{P})}D$ ise kısaca P'ye $\Re$'nin *göreli indirgemesi* denir. $\Re$ içinde D'ye göre bütün vazgeçilemez elemanların toplamına ise $Çekirdek_D(\Re)$ ile gösterilen *Çekirdek* denir.

Aşağıda hangi durumlarda $\mathrm{P} \subset \Re$'nin $\Re$'ye göre göreli indirgemesi olduğu tartışılacaktır.

Öncelikle pozitif alan alt yaklaşımla tanımlandığından A bulanık kümesi ve R bulanık benzerlik ilişkisi için alt yaklaşım R_*A yapısı incelenecektir.

$$(x_\lambda)_R(z) = \begin{cases} 0, & 1 - R(x,z) \geq \lambda \\ \lambda, & 1 - R(x,z) < \lambda \end{cases} \qquad z, x \in U \tag{4.10}$$

Burada $\lambda \in (0,1]$ dir. x_λ ise bir bulanık noktadır.

$$x_\lambda(z)=\begin{cases}\lambda, & z=x\\ 0, & z\neq x\end{cases} \quad x,z\in U \tag{4.11}$$

Burada, açıkça $x_\lambda \subseteq (x_\lambda)_R$ olduğu görülebilecektir.

A bulanık kümesinin alt yaklaşımı için, R_*A şu teorem önerilecektir;

Teorem 4.2

$R_*A=\cup\{(x_\lambda)_R:(x_\lambda)_R\subseteq A\}$, $[\lambda\in 0,1]$ [Tsang ve ark., 2005].

Öneri 4.1

$(x_\lambda)_R$ ve $(y_\lambda)_R$ gibi 2 bulanık nokta için eğer $(x_\lambda)_R\neq(y_\lambda)_R$ ise $(x_\lambda)_R\cap(y_\lambda)_R=\phi$ dır.

Öneri 4.2

$(x_\lambda)_{sim(\Re)}=\bigcap_{r\in\Re}(x_\lambda)_R$, $(x_\lambda)_R=(y_\lambda)_R$ ise sadece ve sadece $(x_\lambda)_{sim(\Re)}=(y_\lambda)_{sim(\Re)}$ dir.

Teorem 4.3

$(x_\lambda)_{sim(\Re)}\subseteq[z]_D$ [Tsang ve ark., 2005].

Teorem 4.3 ile $\Re$ 'nin göreceli indirgemesini karakterize edecek şu teorem oluşur.

Teorem 4.4

$\mathrm{P} \subset \Re$ olsun. $x \in U$ için, P sadece ve sadece $\lambda = sim(\Re)_*([x]_D)(x)$ olduğunda $(x_\lambda)_{sim(\Re)} \subseteq [x]_D$ ifadesi $\Re$'nin göreli indirgemesidir. [Tsang ve ark., 2005].

Teorem 4.5

Şu ifadeler eşdeğerdir;

1. $\mathrm{P}, \Re$'nin göreceli indirgemesini içerir.

2. Her bir $x \in U$ için, $\lambda = sim(\Re)_*([x]_D)(x)$ dir. Eğer $(y_\lambda)_{sim(\Re)} \not\subset [x]_D$ ise $sim(\mathrm{P})(x, y) \leq 1 - \lambda$ dır.

$U = \{x_1, x_2, ..., x_n\}$ olsun. $M_D(U, \Re)$ ile nxn (c_{ij}) matrisi tanımlansın. Buna ayırt edilebilirlik matrisi denir.

1. $c_{ij} = \{R : 1 - R(x_i, x_j) \geq \lambda_i\}$ (4.12)

$$\lambda_i = sim(\Re)_*([x_i]_D)(x_i) \quad (4.13)$$

$$\lambda_j = sim(\Re)_*([x_j]_D)(x_j) \text{ eğer } \lambda_j < \lambda_i \quad (4.14)$$

2. $c_{ij} = \phi$ aksi halde [Tsang ve ark., 2005].

4.1.3. Niteliklerin bulanıklaştırılması

Niteliklerin bulanıklaştırılması işlemi iki açıdan değerlendirilebilir. Birincisi, sürekli değerler alan niteliklerin bulanıklaştırılması; ikincisi ise, kategorik değerler alan niteliklerin bulanıklaştırılmasıdır. Konu ile ilgili literatürde birçok teknik bulmak

mümkündür. Ancak, konunun anlaşılması ve basitleştirilmesi için sadece veri içindeki enformasyon kullanılmak suretiyle bir yaklaşım Salido ve Murakami [2003] tarafından önerilmektedir.

Sürekli değer alan niteliklerin bulanıklaştırılması

Sürekli değer alan veri kümelerinin bulanıklaştırılması işlemi öncelikle verilerin normalleştirilmesi ile başlamalıdır. Verilerin normalleştirilmesi işlemi basitçe verinin değerine bağlı olarak 0 ve civarına dağıtılması işlemidir.

$$F_i^* = \frac{F_i - \mu_i}{\mu_i} \text{ yada} \tag{4.15}$$

$$F_i^* = \frac{F_i - F_{\min}}{F_{\max} - F_{\min}} \tag{4.16}$$

Burada, μ_i niteliğin ortalaması, $F_{\max}$ niteliğin en büyük değeri, $F_{\min}$ ise niteliğin en küçük değeridir.

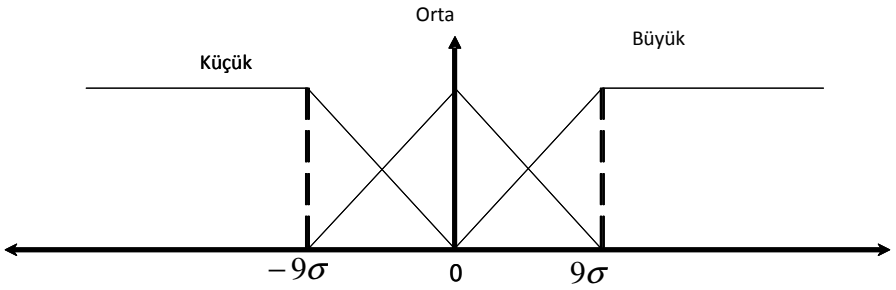
Kategorik değer alan niteliklerin bulanıklaştırılması

Veri tabanları her zaman sürekli değerler almazlar. Veri toplama sisteminin özelliğine göre bazı kategorik değerler de alabilirler. Örneğin, renk niteliği için sürekli bir değer oluşturulması imkânsızdır. Bu kapsamda veri toplarken, ya belirli renkler için kodlama yapılacak (kırmızı=1, mavi=2 gibi) ya da direk olarak renk yazılacaktır. Her iki durumda da, böyle bir değerin bulanıklaştırılması işlemi son derece basit olacaktır. Örneğin, elimizde renk niteliği ve bu niteliklerin alabileceği kırmızı, mavi ve sarı 3 farklı değer olsun. Veri kümesindeki gözlemler sırası ile kırmızı, sarı ve mavi olduğunda bulanık değerler, (1,0,0), (0,0,1) ve (0,1,0) olacaktır.

4.1.4. Bulanık üyelik fonksiyon değerlerinin bulunması

Salido ve Murakami [2003] çalışmalarında, pratik uygulamalarda bulanıklaştırma işlemi için uzman görüşünün olmadığı durumlarda kullanılabilecek basit bir yaklaşım önermişlerdir. Burada, bir bulanık bölüntü fonksiyonu yardımı ile küçük orta ve büyük olarak nesnelerin bulanık üyelik fonksiyonları bulunabilmektedir.

Şekil 4.1'de herhangi bir kesin evrenin bulanıklaştırması için çizilmiş fonksiyon grafiği görülmektedir. Burada küçük ile ifade edilen dilsel değişken yamuk şeklinde; orta dilsel değişken ile ifade edilen bulanık bölüntü üçgen şeklinde ve büyük dilsel değişken ile ifade edilen bulanık bölüntü ise yamuk şekline sahiptir. Ayrıca, burada σ ilgili nitelik değerinin standart sapmasını ifade etmektedir.



Şekil 4.1. Bulanık bölüntü fonksiyonu

4.1.5 Benzerlik ilişkilerinin oluşturulması

Burada ilgilenilen veri türü tip-2 bulanık verisi olduğundan, yukarıda açıklandığı gibi, benzerlik ilişkileri her bir nitelik için ayrı ayrı hesaplanacaktır. Bu maksatla Lukasiewicz operatörü'nün dönüştürülmüş şekli kullanılacaktır.

$$S_{ij}^{k}(x_i, y_j) = \inf_{I \in J_{ij}^{k}} (1 - \left| \mu_i^{kl} - \mu_j^{kl} \right|) \tag{4.8}$$

4.1.6 Benzerlik ilişkilerinin bütünleştirilmesi

Burada, amaç her bir bulanık benzerlik ilişkisini tanımlayacak tek bir bulanık benzerlik ilişkisi elde etmektir. T_L çatısında geçişli bir benzerlik toplama operatörü olarak genelleştirilmiş ortalamaların De Morgan duali kullanılacaktır.

$$M(x) = 1 - (((1-x_1)^p + (1-x_2)^p + ... + (1-x_n)^p)/n)^{1/p} \quad (4.9)$$

Algoritma boyunca minimuma yaklaştığı için p=2 indisi alınacaktır.

4.1.7. $M_D(U,\Re)$ ayırtedilebilirlik matrisini bulma

Bütünleştirme işlemi tamamlandıktan sonra $M_D(U,\Re)$ ayırtedilebilirlik matrisinin elde edilmesi için öncelikle λ_i değeri hesaplanmalıdır.

λ_i değeri basitçe, bütünleştirilmiş benzerlik ilişkileri matrisinden gözlem satırındaki en büyük değerin tersi ile bulunur. Bir sonraki işlem her bir bulanık benzerlik ilişkisi ile bu değeri eş.(4.12-14)'e göre karşılaştırmaktır. Burada, ilgili niteliğin bulanık değeri eğer buradaki karşılaştırmayı geçerse ayırtedilebilirlik matrisine alınır. Aksi halde alınmaz. Ayrıca, klasik kaba kümelerdeki ayırtedilebilirlik matrisinin aksine, $M_D(U,\Re)$'nin simetrik olması beklenmez.

4.2 Önerilen Algoritmanın Adımları

Bu bölümde kısaca, önerilen algoritma sunulacaktır.

ALGORİTMA

Girdi: $(U,\Re \cup D)$ bulanık karar sistemi

Çıktı : P, durum niteliklerinin indirgemeleri.

Adım 1. $Sim(R_1),...Sim(R_n)$ i hesapla,

Adım 2. $Sim(\Re) = M(Sim(R_1),...Sim(R_n))$ i hesapla,

Adım 3. $M_D(U,\Re)$ ayırtedilebilirlik matrisini $c_{ij} = \{R : 1 - R(x_i, x_j) \geq \lambda_i\}$ olacak şekilde hesapla.

Adım 4. Ayırtedilebilirlik matrisi içerisindeki $c_{ij} = 0$ girdilerini sil.

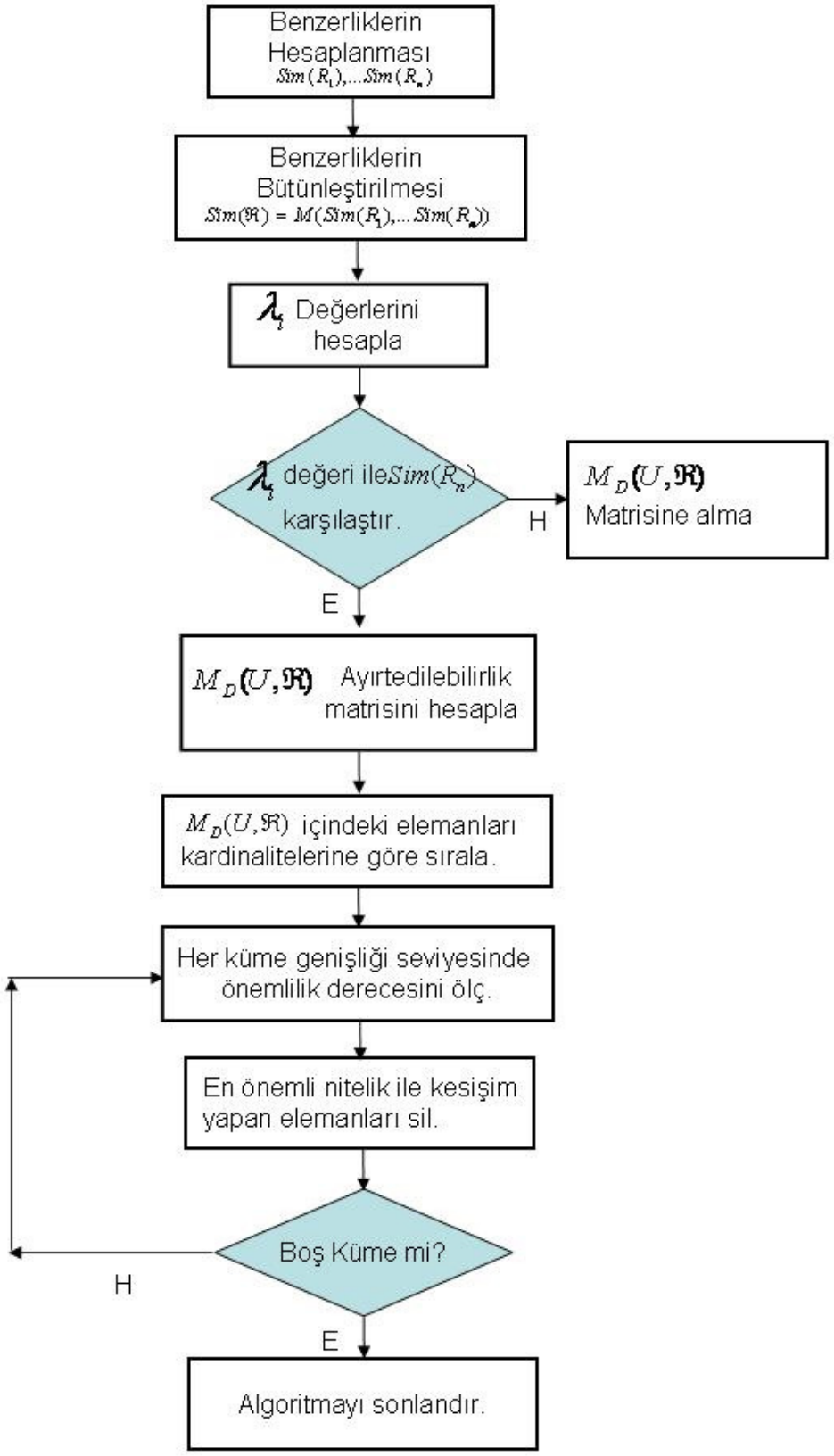
Adım 5. Ayırtedilebilirlik matrisi içerisindeki girdileri küme genişliklerine göre sırala. Kardinal sayı küme genişliği L_i, $i = 1,2,..,m$ ile ifade edilir.

Adım 6. $i_0 = \min_{L_i}\{i\}$, a durum niteliği ve P indirgeme kümesi olsun. Herhangi bir küme genişliği seviyesinde nitelik önemlilik derecesini $SGF(a', \Re, P|L_{i0}) = \max_{a \in P-\{R\}}\{SGF(a, \Re, P|L_{i0}\}$ hesapla.

Adım 7. $P <= P \bigcup \{a'\}$ ifadesini $\Re$ kümesinde hiçbir eleman kalmayana kadar hesapla.

Önerilen algoritmanın akış diyagramı şekil 4.2'de sunulmaktadır.

Önerilen algoritma, Java dilinde kodlanmıştır. Program 2 bölümden oluşmaktadır. Birinci bölüm eksik değerler temizlenerek uygun şekilde hazırlanmış verileri bulanıklaştırmakta, ikinci bölüm ise her nitelik genişlik seviyesinde nitelik alt kümelerini göstermektedir. Program görsel olarak EK-1'de sunulmuştur.



Şekil 4.2. Önerilen algoritmanın akış diyagramı

4.3 Algoritmayı açıklayıcı örnek

Çizelge 4.1'de verilen veri kümesi ele alınsın. Burada, $U = \{x_1, x_2, ..., x_{10}\}$, ve her bir nitelik küçük, orta ve büyük şeklinde 3 bulanık dilsel değişkenle tanımlanmaktadır.

Girdi :

Çizelge 4.1. Bulanık karar sistemi

	A			B			C			D			E			F			Q
	L	M	B	L	M	B	L	M	B	L	M	B	L	M	B	L	M	B	
1	0	0.05	0.95	0.87	0.13	0	0.77	0.23	0	1	0	0	0	0	1	0.52	0.48	0	0
2	0	0.13	0.87	0	0	1	0	0.75	0.25	0.63	0.37	0	0.69	0.31	0	0	0.51	0.49	1
3	0	0.27	0.73	0.39	0.61	0	0	0	1	0.21	0.79	0	0	0.34	0.66	0.8	0.2	0	1
4	1	0	0	0.57	0.43	0	0.77	0.23	0	0	0.02	0.98	0.69	0.31	0	0	0	1	0
5	1	0	0	0.87	0.13	0	0.77	0.23	0	1	0	0	0.69	0.31	0	1	0	0	0
6	0.37	0.63	0	0	0.28	0.72	0	0.61	0.39	0	0.25	0.75	0.69	0.31	0	0.9	0.1	0	0
7	0	0	1	0.87	0.13	0	0.77	0.23	0	0	0	1	0	0	1	0.47	0.53	0	0
8	0	0.21	0.79	0	1	0	0	0.65	0.35	0	0	1	0.4	0.6	0	0	0.5	0.5	1
9	0.61	0.39	0	0	0	1	0	0.91	0.09	0.03	0.97	0	0.69	0.31	0	0.47	0.53	0	0
10	1	0	0	0.87	0.13	0	0.77	0.23	0	0	0.87	0.13	0.69	0.31	0	0	0	1	0

Adım 1. Eş. 4.8 kullanılarak A ve B nitelikleri için oluşturulan bulanık benzerlik matrisleri hesaplanarak Çizelge 4.2 ve 4.3'de verilmiştir.

Çizelge 4.2 A niteliği için benzerlik matrisi

	1	2	3	4	5	6	7	8	9	10
1	1.00	0.92	0.78	0.00	0.00	0.05	0.95	0.84	0.05	0.00
2	0.92	1.00	0.86	0.00	0.00	0.13	0.87	0.92	0.13	0.00
3	0.78	0.86	1.00	0.00	0.00	0.27	0.73	0.94	0.27	0.00
4	0.00	0.00	0.00	1.00	1.00	0.37	0.00	0.00	0.61	1.00
5	0.00	0.00	0.00	1.00	1.00	0.37	0.00	0.00	0.61	1.00
6	0.05	0.13	0.27	0.37	0.37	1.00	0.00	0.21	0.76	0.37
7	0.95	0.87	0.73	0.00	0.00	0.00	1.00	0.79	0.00	0.00
8	0.84	0.92	0.94	0.00	0.00	0.21	0.79	1.00	0.21	0.00
9	0.05	0.13	0.27	0.61	0.61	0.76	0.00	0.21	1.00	0.61
10	0.00	0.00	0.00	1.00	1.00	0.37	0.00	0.00	0.61	1.00

Çizelge 4.3. B niteliği için benzerlik matrisi

	1	2	3	4	5	6	7	8	9	10
1	1.00	0.00	0.52	0.70	1.00	0.13	1.00	0.13	0.00	1.00
2	0.00	1.00	0.00	0.00	0.00	0.72	0.00	0.00	1.00	0.00
3	0.52	0.00	1.00	0.82	0.52	0.28	0.52	0.61	0.00	0.52
4	0.70	0.00	0.82	1.00	0.70	0.28	0.70	0.43	0.00	0.70
5	1.00	0.00	0.52	0.70	1.00	0.13	1.00	0.13	0.00	1.00
6	0.13	0.72	0.28	0.28	0.13	1.00	0.13	0.28	0.72	0.13
7	1.00	0.00	0.52	0.70	1.00	0.13	1.00	0.13	0.00	1.00
8	0.13	0.00	0.61	0.43	0.13	0.28	0.13	1.00	0.00	0.13
9	0.00	1.00	0.00	0.00	0.00	0.72	0.00	0.00	1.00	0.00
10	1.00	0.00	0.52	0.70	1.00	0.13	1.00	0.13	0.00	1.00

Adım 2;

Eş. 4.9 kullanılarak bütünleştirilen bulanık benzerlik matrisi aşağıda sunulmuştur.

$$sim(\Re)_*(x_i, x_j) = \begin{pmatrix} - & 0,29 & 0,41 & 0,17 & 0,39 & 0,14 & 0,59 & 0,22 & 0,14 & 0,18 \\ & - & 0,31 & 0,2 & 0,21 & 0,39 & 0,2 & 0,41 & 0,52 & 0,26 \\ & & - & 0,14 & 0,25 & 0,36 & 0,35 & 0,33 & 0,3 & 0,21 \\ & & & - & 0,41 & 0,35 & 0,28 & 0,39 & 0,21 & 0,63 \\ & & & & - & 0,32 & 0,26 & 0,14 & 0,3 & 0,42 \\ & & & & & - & 0,23 & 0,41 & 0,6 & 0,28 \\ & & & & & & - & 0,34 & 0,12 & 0,21 \\ & & & & & & & - & 0,29 & 0,24 \\ & & & & & & & & - & 0,32 \\ & & & & & & & & & - \end{pmatrix}$$

Adım 3.

Karar sınıfları Çizelge 4.1'den karar niteliğini gösteren Q sütunundan aşağıdaki şekilde bulunur.

$Q_1 = \{x_1, x_4, x_5, x_6, x_7, x_9, x_{10}\}$ (karar değerlerinin "1" olduğu gözlemler)

$Q_0 = \{x_2, x_3, x_8\}$ (karar değerlerinin "0" olduğu gözlemler)

Bölüm 4.1.7'de açıklandığı gibi λ_i değerleri, elde edilen karar sınıflarına göre bütünleştirilmiş benzerlik matrisinden bulunmaktadır. Örneğin λ_1 değeri birinci

satırın maksimum değerinin tersi alınmak suretiyle bulunmaktadır. Aynı işlemler eldeki karar bölüntülerine göre yapıldığında;

$$\boldsymbol{sim}(\Re)_*(\boldsymbol{Q}_1)(\boldsymbol{x}) = \begin{cases} 0{,}41, \boldsymbol{x} = \boldsymbol{x}_1 \\ 0{,}37, \boldsymbol{x} = \boldsymbol{x}_4 \\ 0{,}58, \boldsymbol{x} = \boldsymbol{x}_5 \\ 0{,}40, \boldsymbol{x} = \boldsymbol{x}_6 \\ 0{,}41, \boldsymbol{x} = \boldsymbol{x}_7 \\ 0{,}40, \boldsymbol{x} = \boldsymbol{x}_9 \\ 0{,}37, \boldsymbol{x} = \boldsymbol{x}_{10} \end{cases} \qquad \boldsymbol{sim}(\Re)_*(\boldsymbol{Q}_0)(\boldsymbol{x}) = \begin{cases} 0{,}48\,,\ x\ =\ x_2 \\ 0{,}59\,,\ x\ =\ x_3 \\ 0{,}59\,,\ x\ =\ x_8 \end{cases}$$

λ_i değerlerinin bulunmasından sonra, her bir nitelik için bulunan bulanık benzerlik matrislerinin ilgili satırı ile karşılaştırılmaktadır. Örneğin ayırtedilebilirlik matrisinin (1,6) hücresi bulunurken, $\boldsymbol{sim}(\Re)_*(\boldsymbol{Q}_1)(\boldsymbol{x})$ deki x_6'ya denk gelen 0,40 değeri ile bulanık benzerlik matrislerinin (1,6) hücrelerinin bu değeri geçip geçmediği kontrol edilir. Değerin geçilmiş ise bu iki nitelik arasında belirgin bir farklılığın olduğu anlamına gelmektedir. Örnekte, (1,6) hücreleri için A matrisinde 0,05 değeri B matrisinde 0,13 değeri görülmektedir. Eş. 4-12'den,

A(1,6) için 1-0,05= 0,95;

B(1,6) için 1-0,13= 0,87;

0,95 ve 0,87 değerlerinin λ_i=0,40 değerini geçmesi dolayısı ile bu iki gözlem arasında belirgin bir farkın olduğu anlaşılmaktadır. Bu işlem diğer benzerlik matrisleri için yapıldığında Şekil 4.3'deki ayırtedilebilirlik matrisi elde edilmektedir.

Adım 4. Ayırtedilebilirlik matrisi içerisindeki (1,1), (2,2) gibi içeriği boş küme olanlar silinir.

Adım 5. Ayırtedilebilirlik matrisi içerisindeki girdileri kardinal sayılarına göre sıralanır. Örneğin, matris içerisinde eleman sayısı 1 olan eleman; {D} (aynı zamanda

bu nitelik çekirdektir.), eleman sayısı 2 olan AC (1), AE (1), BD (2), CD (2), DF (6) olarak bulunur. Diğer eleman sayıları gösterilen şekilde sıralanmaktadır.

Adım 6. Öncelikle, P indirgeme kümesine en sık görüldüğünden (önem derecesi fazla olduğundan) D niteliği alınır. İkili seviyelerdeki D niteliği ile beraber görülen nitelik çiftleri silinir. Böylece 2 nci seviye için, AC (1), AE (2) olarak bulunur. Burada A niteliği 2, C niteliği 1 ve E niteliği 1 kez görülmektedir. En büyük önemlilik derecesi A niteliğine ait olduğundan A niteliği indirgeme kümesine atılır. A niteliği ile kesişim yapan girdiler silinir. 2 nci seviyede elde başka nitelik kalmadığından indirgeme {A,D} olarak bulunur.

Böylece, örnek sistem için; Çekirdek {D} niteliği, indirgeme kümesi ise {A,D} olarak bulunur. Bulunan indirgeme kümesi, herhangi bir veri madenciliği algoritmasında kullanılabilir.

4.4. Önerilen Algoritmanın Diğer Algoritmalarla Karşılaştırması

Önerilen algoritma, literatürde bulunan; kaba küme nitelik indirgeme ve Jensen ve Shen [2005] tarafından önerilen FRQR nitelik indirgeme algoritmaları ile karşılaştırılmıştır. Algoritmaların karşılaştırılması için gerekli veriler, UCI makine öğrenimi web sitesinden alınmıştır [UCI Machine Learning Repository, 2007]. Nitelik indirgeme algoritmaları sonucu elde edilen indirgenmiş nitelik kümeleri, tek başlarına herhangi bir anlam ifade etmemelerinden dolayı, J.48, SMO ve Naive Bayes sınıflandırma algoritmalarının sınıflandırma doğrulukları hesaplanmış ve buna göre bir karşılaştırma yapılmıştır.

4.4.1. UCI makine öğrenimi verilerin tanıtımı ve hazırlanması

Deney yapılan veri tabanları, genellikle literatürde bu alanda yapılan çalışmalarda yaygın olarak kullanılması nedeniyle UCI makine öğrenimi deposundan alınmış ve değişik 10 veri tabanı seçilmiştir.

$$M_D(U,\Re)=\begin{pmatrix}
\phi & \{B,C,E,F\} & \{B,C,D\} & \{A,D,E,F\} & \{A,E,F\} & \{A,B,C,D,E\} & \{D\} & \{B,C,D,E,F\} & \{A,B,C,D,E\} & \{A,D,E,F\} \\
\{B,C,E,F\} & \phi & \{B,C,E,F\} & \{A,B,C,D,F\} & \{A,B,C,F\} & \{A,D,F\} & \{B,C,D,E,F\} & \{B,D\} & \{A,D,F\} & \{A,B,C,D,F\} \\
\{C,D\} & \{B,C,E,F\} & \phi & \{A,C,D,E,F\} & \{A,C,D,E\} & \{A,B,C,D,E\} & \{C,D\} & \{C,D,E,F\} & \{A,B,C,E\} & \{A,C,E,F\} \\
\{A,D,E,F\} & \{A,B,C,D,F\} & \{A,C,D,E,F\} & \phi & \{D,F\} & \{A,B,C,F\} & \{A,E,F\} & \{A,B,C,F\} & \{A,B,C,D,F\} & \{D\} \\
\{A,E\} & \{A,B,C,F\} & \{A,C,D,E\} & \{D,F\} & \phi & \{A,B,C,D\} & \{A,D,E\} & \{A,B,C,D,F\} & \{B,C,D\} & \{D,F\} \\
\{A,B,C,D,E\} & \{A,D,F\} & \{A,B,C,D,E\} & \{A,B,C,F\} & \{A,B,C,D\} & \phi & \{A,B,C,E,F\} & \{A,B,F\} & \{D,F\} & \{A,B,C,D,F\} \\
\{D\} & \{B,C,D,E,F\} & \{B,C,D\} & \{A,E,F\} & \{A,D,E,F\} & \{A,B,C,E,F\} & \phi & \{B,C,E,F\} & \{A,B,C,D,E\} & \{A,D,E,F\} \\
\{B,C,D,E\} & \{B,D\} & \{C,D,E,F\} & \{A,C\} & \{A,B,C,D,F\} & \{A,B,F\} & \{B,C,E\} & \phi & \{A,B,D\} & \{A,B,C,D\} \\
\{A,B,C,D,E\} & \{D,F\} & \{A,B,C,E\} & \{B,C,D,F\} & \{B,C,D,F\} & \{D,F\} & \{A,B,C,D,E\} & \{A,B,D,F\} & \phi & \{B,C,F\} \\
\{A,D,E,F\} & \{A,B,C,D,F\} & \{A,B,C,E,F\} & \{D\} & \{D,F\} & \{A,B,C,D,F\} & \{A,D,E,F\} & \{A,B,C,D,F\} & \{AB,C,F\} & \phi
\end{pmatrix}$$

Şekil 4.3. Ayırtedilebilirlik matrisi

Burada, veri kümelerinin seçiminde herhangi bir ölçüt kullanılmamış olup, veri kümelerinin kategorik ve sürekli değerler alabilmesine olanak sağlanmıştır. Kullanılan veri kümelerine ait özet bilgiler Çizelge 4.4'de sunulmuştur. Veri tabanlarının detaylı açıklaması Ek 2-11'da sunulmuştur.

Statlog (Australian Credit Approval) veri kümesi

Veri kümesi 690 örneği içermektedir. Veri kümesinde (6) nümerik ve (8) kategorik olmak üzere toplam (15) nitelik bulunmaktadır. Veri kümesinde (2) değişik karar sınıfı bulunmaktadır. Veri kümesinde 16 örnek eksik değerlidir. Toplam 37 gözlem değerinde eksik değerler bulunmaktadır. Eksik veri olması nedeniyle bu verilerin gözlendiği satır değerleri veri kümesinden çıkartılmış ve indirgeme işlemi bu veri kümesine göre yapılmıştır. Statlog veri kümesi ile ilgili bilgiler EK-2'de sunulmuştur.

Çizelge 4.4. Veri tanımlamaları

No	**Veri**	**Örnek büyüklüğü**	**Nümerik Nitelik sayısı**	**Kategorik Nitelik sayısı**	**Toplam**	**Sınıf Sayısı**
1	Statlog (Australian Credit Approval)	690	6	9	15	2
2	Pima Indians Diabetes	768	8	0	8	2
3	Ecoli	336	5	2	7	7
4	Heart Disease	270	7	6	13	2
5	Ionosphere	351	34	0	34	2
6	Connectionist Bench (Sonar, Mines vs. Rocks)	208	60	0	60	2
7	Small Soybean	47	35	0	35	4
8	Breast Cancer Wisconsin (Diagnostic)	569	30	0	30	2
9	Breast Cancer Wisconsin (Prognostic)	198	33	0	33	2
10	Wine Recognition	178	13	0	13	3

Pima Indians Diabetes veri kümesi

Veri kümesi 768 örneği içermektedir. Veri kümesindeki 8 nitelikte sürekli değerler almaktadır. Veri kümesinde (2) değişik karar sınıfı bulunmaktadır. Veri kümesinde eksik veri bulunmamaktadır. Pima Indians Diabetes veri kümesinin özellikleri EK-3'de verilmiştir.

Ecoli veri kümesi

Veri kümesi 336 örneği içermektedir. Veri kümesinde (5) nümerik ve (2) kategorik olmak üzere toplam (7) nitelik bulunmaktadır. Veri kümesinde (7) değişik karar sınıfı bulunmaktadır. Veri kümesinde eksik veri bulunmamaktadır. Ecoli veri kümesinin özellikleri EK-4'de verilmiştir.

Heart Disease veri kümesi

Veri kümesi 270 örneği içermektedir. Veri kümesinde (7) nümerik ve (6) kategorik olmak üzere toplam (13) nitelik bulunmaktadır. Veri kümesinde (2) değişik karar sınıfı bulunmaktadır. Veri kümesinde eksik veri bulunmamaktadır. Heart Disease veri kümesinin özellikleri EK-5'de verilmiştir.

Ionosphere veri kümesi

Veri kümesi 351 örneği içermektedir. Veri kümesinde (34) nümerik nitelik bulunmaktadır. Veri kümesinde (2) değişik karar sınıfı bulunmaktadır. Veri kümesinde eksik veri bulunmamaktadır. Ionosphere veri kümesinin özellikleri EK-6'da verilmiştir.

Connectionist Bench (Sonar, Mines vs. Rocks) veri kümesi

Veri kümesi 208 örneği içermektedir. Veri kümesinde (60) nümerik nitelik bulunmaktadır. Veri kümesinde (2) değişik karar sınıfı bulunmaktadır. Veri

kümesinde eksik veri bulunmamaktadır. Connectionist Bench (Sonar, Mines vs. Rocks) veri kümesinin özellikleri EK-7'de verilmiştir.

Soybean (Small) veri kümesi

Veri kümesi 47 örneği içermektedir. Veri kümesinde (35) nümerik nitelik bulunmaktadır. Veri kümesinde (4) değişik karar sınıfı bulunmaktadır. Veri kümesinde eksik veri bulunmamaktadır. Soybean (Small) veri kümesinin özellikleri EK-8'de verilmiştir.

Breast Cancer Wisconsin (Diagnostic) veri kümesi

Veri kümesi 569 örneği içermektedir. Veri kümesinde (30) nümerik nitelik bulunmaktadır. Veri kümesinde (2) değişik karar sınıfı bulunmaktadır. Veri kümesinde eksik veri bulunmamaktadır. Breast Cancer Wisconsin (Diagnostic) veri kümesinin özellikleri EK-9'da verilmiştir.

Breast Cancer Wisconsin (Prognostic) veri kümesi

Veri kümesi 198 örneği içermektedir. Veri kümesinde (33) nümerik nitelik bulunmaktadır. Veri kümesinde (2) değişik karar sınıfı bulunmaktadır. Veri kümesinde (4) adet eksik veri bulunmaktadır. Eksik veri olması nedeniyle bu verilerin gözlendiği satır değerleri veri kümesinden çıkartılmış ve indirgeme işlemi bu veri kümesine göre yapılmıştır. Breast Cancer Wisconsin (Prognostic) veri kümesinin özellikleri EK-10'da verilmiştir.

Wine veri kümesi

Veri kümesi 178 örneği içermektedir. Veri kümesinde (33) nümerik nitelik bulunmaktadır. Veri kümesinde (4) adet eksik veri bulunmaktadır. Eksik veri olması nedeniyle bu verilerin gözlendiği satır değerleri veri kümesinden çıkartılmış ve

indirgeme işlemi bu veri kümesine göre yapılmıştır. Breast Cancer Wisconsin (Prognostic) veri kümesinin özellikleri EK-11'de verilmiştir.

4.4.2. Algoritmaların karşılaştırılması

Doğruluk yüzdesi ve nitelik kümesi genişliği açısından karşılaştırma

Önerilen algoritmada, bulanık benzerlik matrislerinin bütünleştirmesi için minimum operatörüne yaklaştığı için p=2 indisi ile bulanıklaştırma işleminde standart sapma çarpanı 0,9 olarak kullanılmıştır.

Kaba küme nitelik indirgemesi için RSAR [Chouchoulas ve Shen, 2001] algoritması kullanılmıştır. Burada kesikli hale getirme işlemi için ROSETTA (EK-12) paket programının içinde bulunan Boole Nedenlemesi algoritması uygulanmıştır.

Çizelge 4.5'de veri kümelerinin indirgeme öncesi ve indirgeme sonrası nitelik kümesi genişlikleri verilmiştir.

Çizelge 4.5. Seçilen nitelik kümesi genişlikleri

Veri kümesi	**Tüm nitelikler**	**RSAR ile indirgeme**	**FRQR ile indirgeme**	**Önerilen Algoritma ile indirgeme**
Statlog (Australian Credit Approval)	15	12	14	8
Pima Indians Diabetes	8	7	8	5
Ecoli	7	6	6	5
Heart Disease	13	10	7	10
Ionosphere	34	3	7	8
Connectionist Bench (Sonar, Mines vs. Rocks)	60	1	5	6
Small Soybean	35	2	2	3
Breast Cancer Wisconsin (Diagnostic)	30	2	6	8
Breast Cancer Wisconsin (Prognostic)	33	1	6	2
Wine Recognition	13	2	5	6
Ortalama	24,8	4,6	6,6	6,1

Seçilen niteliklerin sınıflandırma performansı; J48, Naive Bayes ile SMO algoritmaları ile ölçülmüştür.

J48; karar ağaçları üreten C4.5 [Quinlan, 1993] algoritmasının 8'nci sürümüdür. Bu algoritma çoğu pratik makine öğrenimi uygulamasında kullanılmaktadır. Naive Bayes [Langley ve ark., 1992] sınıflandırma algoritması; hem tahmin edici hem de tanımlayıcı bir sınıflama tekniğidir. Her ilişkide koşullu bir olasılık türetmek için bağımlı ve bağımsız değişkenler arasındaki ilişkiyi analiz eder. SMO [Platt, 1998] ise makine öğreniminde kullanılan destek vektör makinesi (SVM) algoritmasının geliştirilmiş halidir. "Sıralı Minimal Optimizasyon" anlamına gelmektedir.

Sınıflandırma doğruluklarının hesaplanması için WEKA paket programı kullanılmıştır. WEKA paket programı ile ilgili bilgiler EK-13'de sunulmuştur.

İndirgenen niteliklerle yapılan sınıflandırma işleminde 10 katlı çapraz doğrulama metodu kullanılmıştır.10 katlı çapraz doğrulama işlemine göre, her bir veri kümesi 10 parçaya bölünmüş, her bir parça için algoritma 10 kez çalıştırılmıştır. Her bir seferde test kümesi olarak farklı bir parça kullanılmış, kalan 9 parça eğitim için kullanılmıştır. 10 kez çalıştırma sonucu bulunan ortalama değerler belirlenmiştir.

Literatürde nitelik indirgeme algoritmaları tek başlarına kullanılamadığından dolayı, karşılaştırma için genellikle sınıflandırma algoritmaları kullanılmaktadır. Karşılaştırmalar yapılırken de indirgeme kümelerinden bulunan nitelik altkümelerinin sınıflandırma doğruluklarının ortalamaları esas alınmaktadır [Jensen ve Shen, 2007], [Hu ve ark.,2007].

Çizelge 4.6'da SMO Algoritmasının; Çizelge 4.7'de Naive Bayes Algoritmasının ve Çizelge 4.8'de J.48 Algoritmasının sınıflandırma doğruluk sonuçları gösterilmektedir. Köşeli parantez içerisinde algoritmaya göre seçilen niteliklerin küme genişlikleri verilmiştir.

Çizelge 4.6 incelendiğinde, tüm nitelikler için bulunan ortalama değer %86,8; Kaba Küme ile bulunan ortalama değer %79; FRQR ile bulunan ortalama değer %82,9 ve önerilen algoritma ile bulunan değer %85 olduğu görülmektedir. İndirgeme öncesi ortalama değer yüzdesi ile önerilen algoritmanın ortalama değer yüzdesi bir birine oldukça yakındır. Bu durumda, nitelik kümesi genişliklerine bakılması, veri kümesinin boyutunun yüksek olmasının veri madenciliği algoritmalarının performansını düşürdüğü gerçeğinden hareketle daha doğru bir yaklaşım olacaktır.

Çizelge 4.6. Seçilen nitelik kümelerine göre SMO algoritması sınıflandırma sonuçları

Veri kümesi	**SMO sınıflandırma algoritması tahmini doğruluğu**							
	Tüm nitelikler		**RSAR ile indirgeme**		**FRQR ile indirgeme**		**Önerilen Algoritma ile indirgeme**	
Statlog (Australian Credit Approval)	84,6377	[15]	**85,0725**	[12]	84,6377	[14]	84,7826	[8]
Pima Indians Diabetes	**77,4740**	[8]	**77,4740**	[7]	**77,4740**	[8]	77,0833	[5]
Ecoli	**84,2262**	[7]	83,0357	[6]	**84,2262**	[6]	83,631	[5]
Heart Disease	**84,0741**	[13]	83,3333	[10]	81,4815	[7]	82,963	[10]
Ionosphere	**88,6040**	[34]	64,1026	[3]	83,4758	[7]	85,1852	[8]
Connectionist Bench (Sonar, Mines vs. Rocks)	**75,9615**	[60]	68,7500	[1]	72,5962	[5]	73,0769	[6]
Small Soybean	**100**	[35]	**100**	[2]	78,7234	[2]	97,8723	[3]
Breast Cancer Wisconsin (Diagnostic)	**97,7153**	[30]	90,1582	[2]	95,7821	[6]	95,2548	[8]
Breast Cancer Wisconsin (Prognostic)	**77,3196**	[33]	76,2887	[1]	76,2887	[6]	76,2887	[2]
Wine Recognition	**98,3146**	[13]	64,0449	[2]	94,3820	[5]	96,6292	[6]
ORTALAMA	86,8327		79,22599		82,90676		85,2767	

Çizelge 4.6'da indirgeme öncesi 10 veri kümesi için ortalama küme genişliği (24,8) iken önerilen algoritmada (6,1) olduğu görülmektedir. Ayrıca, önerilen algoritmanın doğruluk ortalama değerinin diğer iki algoritmadan daha yüksek olduğu görülmektedir. Buna ilave olarak, kaba küme ile indirgemede ortalama nitelik

kümesi genişliği (4,6) iken önerilen algoritmada (6,1)'dir. Bu durum ise nitelik sayısını azaltmanın her zaman doğruluk yüzdesini arttırmadığını da göstermektedir.

Çizelge 4.6'da 3 algoritmada doğruluk derecesi yüzdeleri 10 veri kümesi için teker teker karşılaştırıldığında, 4 veri kümesinde küçük; 2 veri kümesinde eşit ve diğer 4 veri kümesinde ise daha iyi olduğu görülmektedir.

Çizelge 4.7. Seçilen nitelik kümelerine göre Naive Bayes algoritması sınıflandırma sonuçları

Veri kümesi	**Naive Bayes sınıflandırma algoritması tahmini doğruluğu**							
	Tüm nitelikler		**RSAR**		**FRQR ile indirgeme**		**Önerilen Algoritma ile indirgeme**	
Statlog (Australian Credit Approval)	77,5362	[15]	76,087	[12]	77,5362	[14]	**84,7826**	[8]
Pima Indians Diabetes	76,5625	[8]	76,5625	[7]	76,5625	[8]	**77,2135**	[5]
Ecoli	84,2262	[7]	84,8214	[6]	85,4167	[6]	**86,0119**	[5]
Heart Disease	**83,7037**	[13]	82,9630	[10]	80,0000	[7]	83,3333	[10]
Ionosphere	82,6211	[34]	71,7949	[3]	87,7493	[7]	**90,0285**	[8]
Connectionist Bench (Sonar, Mines vs. Rocks)	**67,7885**	[60]	65,8654	[1]	65,3846	[5]	**67,7885**	[6]
Small Soybean	97,8723	[35]	**100**	[2]	**100**	[2]	97,8723	[3]
Breast Cancer Wisconsin (Diagnostic)	92,9701	[30]	89,2794	[2]	**95,7821**	[6]	93,1459	[8]
Breast Cancer Wisconsin (Prognostic)	63,4021	[33]	74,7423	[1]	76,2887	[6]	**81,4433**	[2]
Wine Recognition	**96,6292**	[13]	70,7865	[2]	96,0674	[5]	95,5056	[6]
ORTALAMA	82,33119		79,29024		84,07875		85,71254	

Çizelge 4.7 incelendiğinde, tüm nitelikler için bulunan ortalama değer %82,33; Kaba Küme ile bulunan ortalama değer %79,29; FRQR ile bulunan ortalama değer %84,07 ve önerilen algoritma ile bulunan değer %85,71 olduğu görülmektedir. Önerilen Algoritmanın doğruluk ortalama değerinin, hem hiçbir indirgeme yapılmamış

durumla elde edilen sınıflandırma doğrulukları hem de iki nitelik indirgeme algoritmasından daha yüksek olduğu görülmektedir.

Çizelge 4.7'de 3 algoritmada doğruluk derecesi yüzdeleri 10 veri kümesi için teker teker karşılaştırıldığında, 4 veri kümesinde küçük; 1 veri kümesinde eşit ve diğer 5 veri kümesinde ise daha iyi olduğu görülmektedir.

Çizelge 4.8. Seçilen nitelik kümelerine göre J48 algoritması sınıflandırma sonuçları

Veri kümesi	**J48 sınıflandırma algoritması tahmini doğruluğu**							
	Tüm nitelikler		**RSAR**		**FRQR ile indirgeme**		**Önerilen Algoritma ile indirgeme**	
Statlog (Australian Credit Approval)	**86,087**	[15]	85,5072	[12]	86,087	[14]	85,3623	[8]
Pima Indians Diabetes	73,9583	[8]	73,9583	[7]	73,9583	[8]	**74,0885**	[5]
Ecoli	**85,4167**	[7]	82,4405	[6]	84,2262	[6]	83,0357	[5]
Heart Disease	76,6667	[13]	**80**	[10]	79,2593	[7]	**80**	[10]
Ionosphere	**91,453**	[34]	83,7607	[3]	90,8832	[7]	90,8832	[8]
Connectionist Bench (Sonar, Mines vs. Rocks)	71,1538	[60]	69,2308	[1]	69,2308	[5]	**74,5192**	[6]
Small Soybean	97,8723	[35]	**100**	[2]	**100**	[2]	97,8723	[3]
Breast Cancer Wisconsin (Diagnostic)	93,1459	[30]	90,6854	[2]	**94,7276**	[6]	94,3761	[8]
Breast Cancer Wisconsin (Prognostic)	73,7113	[33]	**76,2887**	[1]	74,2268	[6]	75,2577	[2]
Wine Recognition	72,4719	[13]	93,8202	[2]	**95,5056**	[5]	92,1348	[6]
ORTALAMA	82,19369		83,56918		84,81048		84,75298	

Çizelge 4.8 incelendiğinde, tüm nitelikler için bulunan ortalama değer %82,2; Kaba Küme ile bulunan ortalama değer %83,5; FRQR ile bulunan ortalama değer %84,8 ve önerilen algoritma ile bulunan değer %84,7 olduğu görülmektedir. FRQR algoritmasından elde edilen indirgenmiş niteliklerin J48 algoritmasına göre sınıflandırma doğruluğu ortalama değer yüzdesi ile önerilen algoritmanın ortalama

değer yüzdesi bir birine oldukça yakındır. Bu durumda, nitelik kümesi genişliklerine bakılması, veri kümesinin boyutunun yüksek olmasının veri madenciliği algoritmalarının performansını düşürdüğü gerçeğinden hareketle daha doğru bir yaklaşım olacaktır. Çizelge 4.8'de de indirgeme öncesi 10 veri kümesi için ortalama küme genişliği (6,6) iken önerilen algoritmada (6,1) olduğu görülmektedir. Bu durum ise, nitelik indirgeme kavramında indirgeme olarak seçilmiş alt küme genişliği ile sınıflandırma doğruluğu arasında bir değiş tokuşun olduğunu göstermesi açısından önemlidir.

Özetle; değişik nitelik indirgemelerine göre seçilen nitelik alt kümeleri hem genişlik hemse sınıflandırma doğruluğu açılarından incelenmiştir. Karşılaştırma işlemi için literatürde bilinen J48, Naive Bayes ve SMO sınıflandırma algoritmaları kullanılmıştır.

Standart sapma çarpan değerlerinin karşılaştırılması

Önerilen algoritmada nitelik değerleri normalleştirilmesinden sonra bulanıklaştırma işlemi yapılırken nitelik kümesi içindeki standart sapma değeri kullanılmış ve bir eşik değer olarak 0,9 çarpanı kullanılmıştır. Bu değer farklı alındığı durumda hesaplanan indirgenmiş küme genişliklerinin değişebileceği düşünülmüştür. Bu kısımda standart sapma çarpan değeri 0,6; 0,8; 0,85; 0,9; 0,95 ve 0,99 alınarak indirgenmiş küme genişliklerine ve sınıflandırma algoritmalarına etkileri incelenmiştir.

Çizelge 4.9'da farklı standart sapma çarpanları kullanılarak indirgenmiş küme genişlikleri ve sınıflandırma algoritmalarına göre hesaplanan doğruluk yüzde ortalamaları verilmektedir.

Çizelge 4.9 Farklı standart sapma çarpan değerleri ile bulanıklaştırmanın sınıflandırma algoritmalarına etkisi

Veri kümesi	**Standart sapma çarpan değeri**	**Seçilen küme genişliği**	**SMO**	**Naive Bayes**	**J48**	**Ortalama**
Statlog (Australian Credit Approval)	0.6	7	85.5072	83.1884	85.5072	84.7343
	0.8	7	85.3623	85.5072	85.2174	**85.3623**
	0.85	10	85.0725	84.3478	84.4928	84.6377
	0.9	8	84.7826	84.7826	85.3623	**84.9758**
	0.95	8	84.7826	84.7826	85.3623	**84.9758**
	0.99	9	84.7826	84.7826	84.7826	84.7826
Pima Indians Diabetes	0.6	6	77.0833	77.2135	74.0885	76.1284
	0.8	6	77.0833	77.2135	74.0885	76.1284
	0.85	6	77.474	77.0833	74.7396	**76.4323**
	0.9	6	77.0833	77.2135	74.0885	76.1284
	0.95	6	77.0833	77.2135	74.0885	76.1284
	0.99	6	74.0885	77.0833	77.2135	76.1284
Ecoli	0.6	4	83.3333	85.4167	81.8452	83.5317
	0.8	5	82.4405	85.119	84.5238	84.0278
	0.85	5	83.631	86.0119	83.0357	**84.2262**
	0.9	5	83.631	86.0119	83.0357	**84.2262**
	0.95	5	83.631	86.0119	83.0357	**84.2262**
	0.99	5	83.0357	83.631	86.0119	**84.2262**
Heart Disease	0.6	9	83.3333	82.5926	80.3704	**82.0988**
	0.8	10	81.4815	80.7407	78.1481	80.1234
	0.85	10	82.963	83.3333	80	**82.0988**
	0.9	10	82.963	83.3333	80	**82.0988**
	0.95	9	76.2963	76.2963	73.7037	75.4321
	0.99	9	78.5185	79.6296	77.4074	78.5185
Ionosphere	0.6	7	83.1909	90.3134	90.3134	87.9392
	0.8	8	86.3248	89.4587	91.1681	**88.9839**
	0.85	7	83.7607	88.8889	88.8889	87.1795
	0.9	8	85.1852	90.0285	90.8832	88.6990
	0.95	8	82.3362	91.4530	88.6040	87.4644
	0.99	8	89.1738	87.1795	90.5983	**88.9839**

Çizelge 4.9 (Devam) Farklı standart sapma çarpan değerleri ile bulanıklaştırmanın sınıflandırma algoritmalarına etkisi

Veri kümesi	**Standart sapma çarpan değeri**	**Seçilen küme genişliği**	**SMO**	**Naive Bayes**	**J48**	**Ortalama**
Connectionist Bench (Sonar, Mines vs. Rocks)	0.6	6	75.4808	64.4231	74.5192	71.4744
	0.8	6	76.4423	66.8269	75.4808	72.9167
	0.85	6	74.0385	68.2692	73.5577	71.9551
	0.9	6	73.0769	67.7885	74.5192	71.7949
	0.95	5	73.5577	74.0385	72.1154	**73.2372**
	0.99	5	70.6731	75.9615	74.0385	**73.5577**
Breast Cancer Wisconsin (Diagnostic)	0.6	8	95.9578	94.2004	95.0791	95.0791
	0.8	8	95.7821	91.9156	92.6186	93.4388
	0.85	7	96.1336	94.3761	97.1880	**95.8992**
	0.9	5	95.2548	93.1459	94.3761	94.2589
	0.95	7	95.9578	94.5518	95.0791	95.1962
	0.99	8	97.1880	95.9578	93.8489	**95.6649**
Breast Cancer Wisconsin Prognostic	0.6	3	76.2887	77.8351	76.2887	**76.8042**
	0.8	3	76.2887	78.3505	75.7732	**76.8041**
	0.85	2	76.2887	76.2887	76.2887	76.2887
	0.9	4	76.2887	81.4433	75.2577	77.6632
	0.95	2	76.2887	77.3196	76.2887	76.6323
	0.99	3	75.7732	76.2887	78.3505	**76.8041**
Wine Recognition	0.6	6	96.6292	97.7528	93.2584	**95.8801**
	0.8	7	95.5056	95.5056	96.6292	95.8801
	0.85	6	96.6292	92.1348	95.5056	94.7565
	0.9	6	96.6292	95.5056	92.1348	94.7565
	0.95	7	94.9438	96.6292	93.82	95.1310
	0.99	8	92.6966	93.2584	95.5056	**93.8202**

Çizelge 4.9 incelendiğinde, her bir veri kümesinin tabloda en son sütunda bulunan ortalama değerlerindeki en yüksek iki değere karşılık gelen standart sapma çarpan değerleri incelendiğinde, en iyi sonuçların, standart sapma çarpan değerlerinin 0,85 ve 0,99 alındığı durumlarda elde edildiği görülmektedir. Bu analiz neticesinde önerilen algoritma için standart sapma çarpan değerinin 0,85 ile 0,99 aralığında daha iyi sonuçlar vereceği değerlendirilmektedir.

5. SONUÇ VE ÖNERİLER

Veritabanları aslında kısmen istenmeyen, ilgisiz ve keşif amacına yönelik olmayan veya belirli bir amaca yönelmemiş verilerden oluşur. Düşük kalitedeki veri, veri madenciliği aşamasında kullanılan tekniğin başarısının da düşmesine sebep olmaktadır.

Bu olumsuzluğun giderilmesi için veri kalitesini arttıracak ve sadece ilgili verilerin bir veri madenciliği algoritmasına girmesini sağlayacak değişik metotlar mevcuttur. Bunlardan biri de nitelik seçimidir. Nitelik seçimi (indirgeme) belirli bir ölçüt ile orijinal veri kümesini daha az nitelikle ifade etmek amacı ile geliştirilmiş metotlar ve yaklaşımlardır.

Klasik bir nitelik seçimi problemi nitelik alt kümesinin oluşturulması, seçilen niteliğinin belirli bir ölçüte göre değerlendirmesinin yapılması, seçilen alt kümenin uyumluluk kalitesine göre seçiminin yapılması ve elde edilen sonucun geçerli olup olmadığına bakılması olmak üzere dört aşamadan oluşmaktadır.

Belirsizliğin modellenmesi, son yıllarda bilginin temsili alanında önemli bir araştırma sahası olmuştur. Zadeh'in bulanık küme teorisi ve Pawlak'ın kaba küme teorisi gibi birçok yaklaşım belirsizlik problemini işaret etmektedir. Bulanık kümeler ve kaba kümeler belirsizliğin modellenmesi için klasik küme teorisinin genelleştirilmesidir. Bu iki teori farklı fakat ilişkili ve birbirini tamamlar niteliktedir.

Bulanık kaba küme yaklaşımını nitelik indirgemede kullanan diğer yaklaşımlar sadece tip-1 bulanık kümelerle ilgilenmektedir. Sonuçta elde edilen indirgenmiş nitelik kümeleri yine bulanık değerler taşımakta ve mutlaka bu verilerle bir sınıflandırma işlemi yapılmaktadır. Bu anlamda tip-2 bulanık kümelerin kullanılması, istenilen veri madenciliği görevi ile ilgili algoritmanın kullanılması, uzman görüşüne dayanarak bazı niteliklere ağırlık verilmesi açılarından özellikle pratik uygulamalar için uygun bir temel sağlamaktadır.

Tez'de bulanık kaba kümelerin nitelik indirgeme için kullanılması anlatılarak bölüm 3'de tip-2 bulanık kümeler tanımlanmış, özellikle pratik uygulamalar için her bir niteliğe ağırlık fonksiyonu da eklenebilecek şekilde bir nitelik indirgeme algoritması sunulmuştur.

Önerilen algoritmanın karşılaştırılmasını yapmak için UCI makine öğrenimi veri seti kullanılmıştır. Bu sette farklı nitelik sayılarına sahip 10 veri kümesi bulunmaktadır. Bu verilere literatürdeki RSAR, FRQR ve önerilen algoritma ile nitelik indirgeme uygulanmış ve elde edilen nitelik alt kümelerinin etkinliğini test etmek için sınıflandırma işlemine tabi tutulmuştur. Sınıflandırmadaki etkinliği de karşılaştırmak için J48, Naive Bayes, SMO algoritmaları kullanılmıştır. İndirgenmiş nitelikler bu algoritmalarda kullanılarak tahmini doğruluk yüzdeleri bulunmuştur. Elde edilen sonuçlar karşılaştırıldığında önerilen algoritmanın diğer algoritmalara göre kullanılan veri tabanları için genelde daha iyi sonuçlar verdiği görülmüştür.

Ayrıca, verilerin bulanıklaştırılması için kullanılan standart sapma çarpan değerlerinin analizi yapılarak, bu değerin 0,85 ila 0,99 arası tanımlanmasının uygun olacağı gösterilmiştir.

Gelecek dönemde, eksik verilere daha kolay adapte edilebilen bulanık kümeler ile tolerans tabanlı kaba kümelerin melezlenmesi konusunda çalışılması planlanmaktadır.

KAYNAKLAR

Akpınar H., "Veri Tabanlarında Bilgi Keşfi", ***İ.Ü. İşletme Fakültesi Dergisi***, 29:1-22 (2000).

Almuallim H., Dietterich T.G., "Learning with many irrelevant features". ***The 9th National Conference on Artificial Intelligence,*** Oregon,USA, 547–552 (1991).

Asuncion, A., Newman D.J. UCI Machine Learning Repository [http://www.ics.uci.edu/~mlearn/MLRepository.html]. Irvine, CA: University of California, School of Information and Computer Science (2007).

Bakar A.A., Sulaiman M.N., Othman M., Selamat M.H., "Finding minimal reduct with binary integer programming in data mining", ***Proceedings of the TENCON***, Kuala Lumpur, Malaysia, 141–149 (2000).

Beynon M., "Reducts within the variable precision rough sets model: a further investigation", ***European Journal of Operational Research*** 134, 592–605, (2001).

Beliakov G., Pradera A., Calvo T., "Aggregation Functions: A Guide for Practitioners" Studies in Fuzziness and Soft Computing, Volume 221, ***Springer-Verlag,*** Berlin Heidelberg, 1-8 (2007).

Bhatt R.,Gopal M., "On Fuzzy rough Set Approach to feature selection.", ***Pattern Recognition Letters*** 26: 965-975 (2005a).

Bhatt R.,Gopal M., "On the compact computational domain of fuzzy rough sets", ***Pattern Recognition Letters*** 26: 1632-1640 (2005b).

Bhatt R.,Gopal M., "Improved feature selection algortihm with fuzzy rough sets on compact computational domain", ***International Journal of General Systems*** 34(4): 485-505 (2005c).

Binay S.H., "Yatırım kararlarında kaba küme yaklaşımı" Basılmamış Doktora Tezi, ***Ankara Üniversitesi Sosyal Bilimler Enstitüsü***, Ankara, 23 (2002).

Bruha I., "From machine learning to knowledge discovery: Survey of preprocessing and postprocessing", ***Intelligent Data Analysis*** 4: 363–374 (2000).

Bodjanova S., "Approximation of fuzzy concepts in decision making" ***Fuzzy Sets and Systems*** 85: 23-29 (1997).

Chen D., Wang X., Zhao S.,"Attribute Reduction Based on Fuzzy Rough Sets", Rough Sets and Intelligent Systems Paradigms, Kryszkiewicz M. ve ark. (Editörler), ***Springer-Verlag,*** Berlin Heidelberg, 381–390, (2007).

Chouchoulas A., Shen Q., "Rough set-aided keyword reduction for text categorisation" ***Applied Artificial Intelligence,*** 15(9): 843–873 (2001).

Dash M., Liu H., "Feature Selection for Classification". ***Intelligent Data Analysis***, 1(3): 131–156 (1997).

Dinçer B, Özaslan M,, Satılmış E,"İllerin sosyo-ekonomik gelişmişlik sıralaması araştırması" Ankara: DPT Bölgesel Gelişme ve Yapısal Uyum Genel Müdürlüğü http://ekutup.dpt.gov.tr/bolgesel/dincerb/il/ (1996).

Dubois D., Prade H., "Putting rough sets and fuzzy sets together", , Intelligent Decision Support:Handbook of Applications and Advances ofthe Rough Sets Theory, R. Slowinski (Editör), ***Kluwer***, Hollanda, 203–222 (1992).

Dubois D., Prade H., "Rough fuzzy sets and fuzzy rough sets", ***International Journal of General Systems***, 17: 191–209 (1990).

Dong J., Zhong N., Ohsuga S. "Using Rough Sets with Heuristics for Feature Selection" ***New Directions in Rough Sets, Data Mining, and Granular- Soft Computing, 7th International Workshop*** (RSFDGrC 99), 178–187, (1999).

Duoqian M., Jue W., "Information based algorithm for reduction of knowledge", ***1997 IEEE international conference on intelligent processing systems***, Beijing, China, 1155-1158 (1997).

Duntsch I., Gediga G., "The rough set engine GROBIAN", ***Proceedings 15th IMACS World Congress*** (4), Berlin, 613–619 (1997).

Fayyad U., Piatetsky S.G., Smyth P., "From data mining to knowledge discovery in databases". ***AI Magazine***, 17(3): 37–54 (1996).

Freitas A., "Generic, Set-Oriented Primitives to Support Data-Parallel Knowledge Discovery in Relational Database Systems", Doktora tezi, ***Department of Computer Science,University of Essex***, Essex,USA, 58 (1997).

Friedman J.H., Tukey J.W., "A projection pursuit algorithm for exploratory data analysis" ***IEEE Transactions on Computers***, 23(9): 881– 890 (1974).

Geng Z, Zhu Q, "A new rough set based heuristic algorithm for attribute reduct", ***Proceedings of the 6th World Congress on Intelligent Control and Automation***, Dalian, China, 3085-3089 (2006).

Golan R.H., Ziarko W., "A methodology for stock market analysis utilizing rough set theory", ***Proceedings of the IEEE/IAFE Computational Intelligence for Financial Engineering***, New York USA, 32–40 (1995).

Han, B., Wu, T.J. "Information Entropy Based Reduct Searching Algorithm", ***American Control Conference, 2002. Proceedings of the 2002*** (6), Hangzhou, China 4577-4582 (2002).

Han, J., Kamber, M., "Data Mining: Concepts and Techniques" ***Academic Press, Morgan Kaufmarm Publishers***, 5-7, 95-101 (2001).

Han J., Hu X., Lin T.Y., "Feature Subset Selection Based on Relative Dependency between Attributes", Rough Sets and Current Trends in Computing, 3066/2004, ***Springer,*** Berlin, 176-185 (2004).

Hedar A, Wang J., Fukushima M., "Tabu Search Attribute Reduction in Rough Set Theory", ***Department of Applied Mathematics & Physics***, ***Kyoto University***, 1-15 (2006).

Holland J., "Adaptation In Natural and Artificial Systems" ***The University of Michigan Press***, Ann Arbour 1-5 (1975).

Hu K., Lu Y., Shi C., "Sampling for Approximate Reduct in Very Large Datasets", http://www.citeseer.ist.psu.edu/587308.html, (2000).

Hu, K., Lu, Y., Shi, C., "Feature ranking in rough sets". ***AI Communication,*** 16(1): 41-50 (2003).

Hu Q.,Xie Z., Yu D., "Hybrid attribute reduction based on a novel fuzzy- rough model and information granulation***", Pattern Recognition,*** 40: 3509-3521, (2007).

Hu X., Lin T.Y., Jianchao J., "A new rough sets model based on database systems", ***Fundamenta Informaticae,*** 59(2): 1–18 (2004).

İnternet: Arditi D, Pulket T., "Predicting the outcome of construction litigation using an intergrated artificial intelligence model"., http://www.iit.edu/ (2007).

İnternet: Fodor K., "A survey of dimension reduction techniques", http://www.llnl.gov/CASC/sapphire/pubs/148494.pdf , (2002).

Jensen R., "Combining rough and fuzzy sets for feature selection" Doktora Tezi, ***University of Edinburgh***, Edinburg,UK 13,48,87 (2005).

Jensen R., Shen Q., "A Rough Set-Aided System for Sorting WWW Bookmarks". Web Intelligence: Research and Development, Zhong N. ve ark. (Editörler),***Springer,*** Berlin 95–105 (2001).

Jensen, R., Shen, Q., "Aiding Fuzzy Rule Induction with Fuzzy Rough Attribute Reduction" ***Proceedings of the 2002 UK Workshop on Computational Intelligence***, Birmingham, UK, 81-88 (2002a).

Jensen, R., Shen Q., "Fuzzy-rough sets for descriptive dimensionality reduction" ***Proceedings of IEEE International Conference on Fuzzy Systems***, Honolulu, USA, 29-34 (2002b).

Jensen, R., Shen, Q. "Selecting informative features with fuzzy-rough sets and its application for complex systems monitoring", ***Pattern recognition***, 37(7): 1351-1363 (2004a).

Jensen, R., Shen, Q., "Semantics-Preserving Dimensionality Reduction: Rough and Fuzzy-Rough-Based Approaches" ***IEEE Transactions on Knowledge and Data Engineering,*** 16(12): 1457-1471 (2004b).

Jensen, R., Shen Q. "Fuzzy-rough attribute reduction with application to web categorization", ***Fuzzy Sets and Systems***, 141(3): 469-485 (2004c).

Jensen, R., Shen Q., "Fuzzy-Rough Feature Significance for Fuzzy Decision Trees" ***Proceedings of the 2005 UK Workshop on Computational Intelligence***, Birkbeck, University of London 89-96 (2005a).

Jensen, R., Shen Q, "Fuzzy-rough data reduction with ant colony optimization". ***Fuzzy Sets and Systems,*** 149(1): 5-20 (2005b).

Jensen, R., Shen Q, "Tolerance-based and Fuzzy-Rough Feature Selection**"** ***Fuzzy Systems Conference,*** London,UK**,** 23(26): 1-6 (2007)

John G.H., Kohavi R., Pfleger K., "Irrelevant Features and the Subset Selection Problem". ***Proceedings of the 11th International Conference on Machine Learning ICML94***, 121–129 (1994). (http://citeseer.ist.psu.edu/john94irrelevant.html)

Kızılkaya E., "Veri Madenciliğinde sınıflandırma problemleri için evrimsel bir algoritma tabanlı yeni bir yaklaşım: Rough-MEP algoritması" Doktora Tezi, ***Gazi Üniversitesi Fen Bilimleri Enstitüsü***, Ankara, 2-3 (2008).

Kira K., Rendell L.A., "The feature selection problem: Traditional methods and a new algorithm" ***Proceedings of Ninth National Conference on Artificial Intelligence***, Anaheim, USA, 129–134 (1992).

Kirkpatrick S., Gelatt C., Vecchi M., "Optimization by simulated annealing". ***Science,*** 220(4598): 671–680 (1983).

Klementa E.P., Mesiarb R., Papc E.,"Triangular norms. Position paper I: basic analytical and algebraic properties" ***Fuzzy Sets and Systems*** 143: 5–26 (2004).

Klir G.J., Yuan B., "Fuzzy Sets and fuzzy logic Theory and Applications", ***Prentice Hall***, New Jersey, 19 (1995).

Kruskal, J.B., "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis" ***Psychometrika***, 29: 1-27 (1964).

Kuncheva L.I., "Fuzzy rough sets: application to feature selection", ***Fuzzy Sets and Systems***, 51(2): 147-153 (1992).

Labib N., Malek M., "Data Mining for Cancer Management In Egypt Case Study: Childhood Acute Lymphoblastic Leukemia" ***Proceedings of world academy of science, Engineering and Technology,*** Budapest, Hungary, 309-314 (2005).

Langley P.,"Areas of Application for Machine Learning" ***Proc. of Fifth international symposium on knowledge engineering***, Sevilla,Spain, 326-331(1992).

Langley P., Iba W., Thompson K., "**An analysis of Bayesian classiffiers**," ***Proc. of the Tenth National Conference on Artifcial Intelligence***, San Jose, California,USA, 223-228 (1992).

Lee H.Y., Ong H.L., Quek L-H., "Exploiting visualization in knowledge discovery". ***Proc. 1st International Conference on Knowledge Discovery and Data Mining***, Québec, Canada, 198-203 (1995).

Lenzerini, M. "Data integration: a theoretical perspective" ***Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems,*** Wisconsin,USA, 233-246. (2002).

Li K., Liu S., "Rough set based attribute reduction approach in data mining", ***Proceedings of the First International Conference on Machine Learning and Cybernetics,*** Beijing, China 60- 63 (2002).

Lin T.Y., Cercone N., "Rough sets and Data Mining: Analysis of Imprecise Data", ***Springer***, ***Kluwer Academic Publishers***, 214 (1997).

Liu H., Setiono R., "A Probabilistic Approach To Feature Selection - A Filter Solution". ***Proceedings of the 9th International Conference on Industrial and Engineering Applications of AI and ES***, Fukuoka, Japan, 284–292 (1996a).

Liu H., Setiono R., "Feature selection and classification - a probabilistic wrapper approach". ***Proceedings of the 9th International Conference on Industrial and Engineering Applications of AI and ES***, Fukuoka, Japan, 419–424. (1996b)

Liu H., Yu L., Motoda H., "Feature extraction selection and construction" The Handbook of Data mining, Ye N. (Editör) ***Lavrence Erbaum Associates***, London, 409-422 (2003).

Manila H., "Data mining: machine learning, statistics, and databases**"** ***Eighth International Conference on Scientific and Statistical Database Management***, Stockholm, Sweden, 1-8 (1996).

Miguel A., "A Review of Dimension Reduction Techniques", ***CS Dept. of Computer Science, University of Sheffield***, 1-69 (1997).

Mladenic D.,"Feature selection for dimensionality reduction" Subspace, Latent Structure and Feature Selection, 3940, Saunders C. ve arkadaşları (Editörler), ***Springer,*** Berlin, 84–102 (2006).

Morsi N., Yakout M., "Axiomatics for fuzzy rough set," ***Fuzzy Sets and Systems***, 100: 327–342 (1998).

Ortega J., "Issues on the application of data mining" (özel basım) ***Artifical Intelligence Review-An International Science and Engineering Journal,*** 14 (2000).

Ovchinnikov S., "Aggregating transitive fuzzy binary relations", ***International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems***, 3(1): 47–55 (1992).

Parthalain N., Jensen R. Shen Q., "Fuzzy entropy-assisted fuzzy-rough feature selection" ***2006 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE'06),*** Vancouver Canada, 423-430 (2006).

Pawlak, Z. "Rough sets", ***International Journal of Computational Informormation Sciences***, 11: 341-356 (1982).

Pawlak, Z., Skowron, A. "Rough membership functions" Advances in the Dempster-Shafer Theory of Evidence, Yager, R., Fedrizzi, M., Kacprzyk, J. (editörler), ***John Wiley & Sons***, New York, 251–271 (1994).

Pawlak Z, "Some Issues on Rough Sets", Transactions on Rough Sets I, Peters J.F. ve ark. (Editörler), ***Springer-Verlag***, Berlin, 1–58 (2004).

Pawlak Z., Skowron A., "Rough Sets: Some Extensions," ***Information Sciences***, 177: 28-40 (2007).

Piatetsky-Shapiro G., "Knowledge Discovery in Real Databases" ***AI Magazine***,11(5): 68-70 (1991).

Platt J., "Fast training of support vector machines using sequential minimal ptimization", Advances in Kernel Methods Support Vector Learning, B. Schlkopf, Burges C., Smola A. (Editörler), ***MIT Press***, Cambridge, UK (1998).

Quinlan J.,"Induction of decision trees", ***Machine Learning*** 1: 81–106 (1986).

Quinlan, J.R., "Constructing Decision Tree. In C4.5": Programs for Machine Learning., ***Morgan Kaufman Publishers*** 5-8 (1993).

Radzikowska A. M., Kerre E. E., "A comparative study of fuzzy rough sets," ***Fuzzy Sets Systems,*** 126: 137–155 (2002).

Raman B., Ioerger T.R., "Instance-based filter for feature selection". ***Journal of Machine Learning Research***, 1: 1–23 (2002).

Roweis S.T., Saul L.K., "Nonlinear Dimensionality Reduction by Locally Linear Embedding" ***Science***, 290(5500): 2323–2326 (2000).

Salido J.M.F., Murakami S., "Rough set analysis ofa general type off uzzy data using transitive aggregations of fuzzy similarity relations", ***Fuzzy Sets and Systems***, 139: 635–660 (2003).

Sammon J. W., "A nonlinear mapping for data structure analysis", ***IEEE Trans. Computers***, C-18(5): 401-409 (1969).

Scheffer J., "Dealing with Missing Data", ***Res. Lett. Inf. Math. Sci***. 3: 153-160 (2002).

Setiono R., Liu H., "Neural network feature selector" IEEE Transactions on Neural Networks, 8(3): 645–662 (1997).

Sever H., Oğuz B., "Veri tabanlarında bilgi keşfine formel bir yaklaşım-1" ***Bilgi Dünyası (Information Word)***, 3(2), 173-204 (2002).

Shen Q., Chouchoulas A., "A modular approach to generating fuzzy rules with reduced attributes for the monitoring of complex systems". ***Engineering Applications of Artificial Intelligence***, 13(3): 263-278 (2000).

Shi Z., Liu S., Zheng Z., "Efficient Attribute Reduction Algorithm" ***First IFIP International Conference on Artificial Intelligence Applications and Innovations***, 859-868 (2004).

Skowron, A., Rauszer, C., "The discernibility matrices and functions in information systems" Intelligent Decision Support. Handbook of Applications and advances of the rough set theory, R.Slowinsky (Editör), ***Kluwer, Dordrecht***, 331–362 (1992).

Slowinski R., Vanderpooten D., "Similarity Relations As a Basis for Rough Approximations", Advances in Machine Intelligence and Soft Computing, Wang P.P., (Editör), ***Bookwrights***, Raleigh, 17-33 (1997).

Slowinski R., Vanderpooten D., "A generalized definition of rough approximations based on similarity" ***IEEE Transactions on Data and Knowledge Engineering***, 12: 331-336 (2000).

Suraj Z., "An Introduction to Rough Set Theory and Its Applications: A tutorial" ***1st International Computer Engineering Conference New Technologies for the Information SocietyICENCO'2004***, Cairo, Egypt, 27-30 (2004).

Tenenbaum J.B., De Silva V., Langford J. C., "A global geometric framework for nonlinear dimensionality reduction". ***Science***, 290(5500): 2319–2323 (2000).

Thiele H., "Fuzzy rough sets versus rough fuzzy sets - an interpretation and a comparative study using concepts of modal logics". Technical report no:CI-30/98, ***University of Dortmund***, 1-12 (1998).

Thiele H., "On axiomatic characterisations of crisp approximation operators", ***Information Sciences,*** 129: 221–226, (2000a).

Thiele H., "On axiomatic characterisation of fuzzy approximation operators I, the fuzzy rough set based case", ***Second International Conference on Rough Sets***, Bariff, Kanada, 239–247 (2000b).

Thiele H., "On axiomatic characterisation of fuzzy approximation operators II, the rough fuzzy set based case", ***Proceedings of the 31st IEEE International Symposium on Multiple-Valued Logic***, Warsaw, Poland, 330–335, (2001).

Torgerson, W. S., "Theory and Methods of Scaling", ***Wiley***, New York, 1-5 (1958).

Traina Jr C., Traina A., L. Wu L., Faloutsos C., "Fast feature selection using the fractal dimension" ***Proceedings of the 15th Brazilian Symposium on Databases (SBBD)***, João Pessoa, Brasil, 158–171 (2000).

Tsang G.C.Y., Degang Chen, Tsang E.C.C., Lee J.W.T., Yeung D.S., "On attributes reduction with fuzzy rough sets" ***IEEE International Conference on Systems, Man and Cybernetics***, Hawaii, USA, 2775 – 2780 (2005).

Valverde L., "On the structure of F-indistinguishability operators", ***Fuzzy Sets and Systems***, 17: 313–328 (1985).

Wang B., Chen S.,"A Complete Algorithm for Attribute Reduction", Lecture Notes In Control And Information Sciences, ***Springer***, New York, 345-352 (2004).

Wang X.Z., Ha Y., Chen D., "On the reduction of fuzzy rough sets," ***Proceedings of the 2005 International Conference on Machine Learning and Cybernetics*** (5), Guangzhou, China, 3174–3178, (2005).

Wang R., Miao D., Hu Y., "Discernibility Matrix Based Algorithm for Reduction of Attributes", ***Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology,*** Hong Kong (2006).

Wu,W.Z., Zhang,W.X., "Constructive and axiomatic approaches of fuzzy approximation operators", ***Information Sciences*** 159: 233–254 (2004).

Wyse, N., Dubes, R. Jain A. "A Critical Evaluation of Instrinsic Dimensionality Algorithms" Pattern recognition In Practice Gelsema E. ve Kanal L. (Editörler), ***North Hlland Publishing***, 415-425 (1980).

Yao J., "Feature selection for fluoresence image classification". ***KDD Lab Proposal, CALD, SCS, CMU.,*** North Carolina USA, (2001).

Yao Y., Zhao Y.,Wang J., "On reduct construction algorithms", Lecture Notes in Computer Science, ***Springer Verlag***, Berlin, 297-304 (2006).

Yenidoğan ,T.,"Pazarlama araştırmalarında çok boyutlu ölçekleme analizi: üniversite öğrencilerinin marka algısı üzerine bir araştırma" ***Akdeniz İ.İ.B.F. Dergisi,*** (15): 138-169 (2008).

Yeung D., Chen D., Tsang E., Lee J., Xizhao W., "On the Generalization of Fuzzy Rough Sets" ***IEEE Transactıons On Fuzzy Systems***,13(3): 343-361 (2005).

Zhang M.,Yao J.T., "A rough sets based approach to feature selection", ***NAFIPS 2004 Annual Meeting of the North American Fuzzy Information Processing Society***, Banff, Alberta, Canada, (2004).

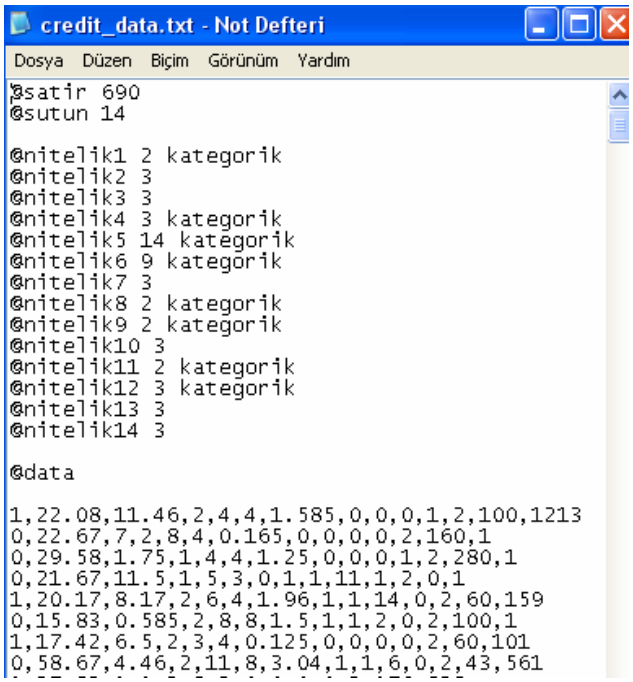
Zadeh L.A., "Fuzzy sets", ***Information control***, 8(3): 338-353 (1965).

Zadeh L.A. "The concept of a linguistic variable and its application to approximate reasoning" , ***Information science***, 8(3): 199-249 (1975a).

Zadeh L.A. "The concept of a linguistic variable and its application to approximate reasoning" , ***Information science***, 8(4): 301-357 (1975b).

EKLER

EK-1 Önerilen Algoritma'nın Programı



Şekil 1.1 Önerilen Algoritma için verilerin düzenlenmesi

Bulanıklaştırma

	0		1			2			3									4														5									6		
S.No	1	2	K	O	B	K	O	B	1	2	3	1	2	3	4	5	6	7	8	9	10	11	12	13	14	1	2	3	4	5	6	7	8	9	K	O	B
1	1	0	0.88941	0.11059	0	0	0	1	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0.21196	0.78804	0
2	0	0	0.83411	0.16589	0	0	0.49975	0.50025	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0.68343	0.31657	0
3	0	0	0.18637	0.81363	0	0.67154	0.32846	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0.32319	0.67681	0
4	0	0	0.92785	0.07215	0	0	0	1	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0.73822	0.26178	0
5	1	0	1	0	0	0	0.23861	0.76139	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0.08746	0.91254	0
6	0	0	1	0	0	0.93156	0.06844	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0.24019	0.75981	0
7	1	0	1	0	0	0	0.61135	0.38865	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0.69671	0.30329	0
8	0	0	0	0	1	0.06667	0.93333	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0.72887	0.27113
9	1	0	0.35041	0.64959	0	0.83894	0.16106	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0.74215	0.25785
10	0	0	0	0	1	0	0.4819	0.5181	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1
11	1	0	0	0.81892	0.18108	0.67154	0.32846	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0.24412	0.75588
12	1	0	0	0.0765	0.9235	0	0.94615	0.05385	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0.07811	0.92189
13	1	0	1	0	0	0.78314	0.21686	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0.28169	0.71831	0
14	1	0	0	0.68581	0.31419	0	0.94615	0.05385	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	1
15	1	0	0	0	1	0.45727	0.54273	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0.93639	0.06361
16	1	0	0	0	1	0	0.71402	0.28598	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0.72494	0.27506	0
17	1	0	0.18637	0.81363	0	0.05775	0.94225	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1
18	0	0	1	0	0	0	0.05336	0.94664	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0.4892	0.5108	0
19	1	0	1	0	0	0.78314	0.21686	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0.69671	0.30329	0
20	0	0	0.85754	0.14246	0	0	0.79772	0.20228	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0.87994	0.12006
21	0	0	0.31854	0.68146	0	0.93156	0.06844	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0.72494	0.27506	0

Şekil 1.2 Bulanıklaştırma işlemi

EK-1(Devam) Önerilen Algoritma'nın Programı

Şekil 1.3 Önerilen Algoritma ile elde edilen indirgenmiş küme

EK-2 Statlog (Australian Credit Approval) veri kümesi

Veri kümesi özellikleri	:	Çok değişkenli
Nitelik özellikleri	:	Kategorik, Tamsayılı, Sürekli
İlgili Görev	:	Sınıflandırma
Örnek Sayısı	:	690
Nitelik Sayısı	:	14
Eksik Değer	:	Var
Konu alanı	:	Mali

Veri kümesi ile ilgili bilgiler

Bu veri tabanı kredi kartı uygulamaları ile ilgilidir. Bütün nitelik isimler ve değerleri verinin elde edildiği şahısların mahremiyetini korumak için anlamsız sembollere dönüştürülmüştür.

Bu veri kümesi, sürekli değerlerin küçük değer alması ile nominal değerlerin büyük değerler almasından dolayı ilginçtir.

Nitelikler ile ilgili bilgiler
Veri kümesinde 6 nümerik ve 8 kategorik değer alan nitelik bulunmaktadır. İstatistiksel algoritmaların tarafından daha iyi anlaşılması bakımından bazı kategoriler numaralandırılmıştır.

A1: 0,1 Kategorik
A2: Sürekli.
A3: Sürekli.
A4: 1,2,3 Kategorik
A5: 1, 2,3,4,5, 6,7,8,9,10,11,12,13,14 Kategorik
A6: 1, 2,3, 4,5,6,7,8,9 kategorik
A7: Sürekli.
A8: 1, 0 Kategorik
A9: 1, 0 Kategorik
A10: Sürekli.
A11: 1, 0 Kategorik
A12: 1, 2, 3 Kategorik
A13: Sürekli.
A14: Sürekli.
A15: 1,2 Sınıf niteliği

EK-3 Pima Indians Diabetes veri kümesi

Veri kümesi özellikleri	:	Çok değişkenli
Nitelik özellikleri	:	Tamsayılı, Sürekli
İlgili Görev	:	Sınıflandırma
Örnek Sayısı	:	768
Nitelik Sayısı	:	8
Eksik Değer	:	Yok
Konu alanı	:	Yaşam

Veri kümesi ile ilgili bilgiler

Daha büyük bir veri tabanından belirli kısıtlar konarak seçilmiş bir veri kümesidir. Burada bütün hastalar, Pima Kızılderili kabilesinden seçilmiş 21 yaşından büyük kadınlara ait bilgilerdir.

Nitelikler ile ilgili bilgiler

Veri kümesinde 8 Nümerik değer alan nitelik bulunmaktadır.

A1: Gebe kalma sayısı

A2: 2 saatlik oral glikoz tolerans testindeki şeker değeri ölçümü

A3: Küçük tansiyon (mm Hg)

A4: Kolun arkasındaki sarkma (mm)

A5: 2 Saatlik İnsülin serumu (Mu U/ml)

A6: Beden kütle indeksi

A7: Şeker şecere fonksiyonu

A8: Yaş

A9: Sınıf değişkeni (0,1)

EK-4 Ecoli veri kümesi

Veri kümesi özellikleri	:	Çok değişkenli
Nitelik özellikleri	:	Sürekli
İlgili Görev	:	Sınıflandırma
Örnek Sayısı	:	336
Nitelik Sayısı	:	8
Eksik Değer	:	Yok
Konu alanı	:	Yaşam

Veri kümesi ile ilgili bilgiler

Bu veri tabanı proteinin toplanması ile ilgili yapılan bazı bilimsel çalışmalardan elde edilmiştir.

A1: Sıra adı

A2: mcg (Sürekli)

A3: gvh (Sürekli)

A4: lip (0,1)

A5: chg (0,1)

A6: aac (Sürekli)

A7: alm1 (Sürekli)

A8: alm2 (Sürekli)

A9: Sınıf değişkeni (1,…,7)

EK-5 Heart Disease veri kümesi

Veri kümesi özellikleri	:	Çok değişkenli
Nitelik özellikleri	:	Kategorik, Tamsayılı, Sürekli
İlgili Görev	:	Sınıflandırma
Örnek Sayısı	:	303
Nitelik Sayısı	:	14
Eksik Değer	:	Var
Konu alanı	:	Yaşam

Veri kümesi ile ilgili bilgiler

Bu veri tabanı gerçekte 76 nitelikten oluşmasına rağmen 14 niteliği kullanılmaktadır. Bu veri tabanında amaç kalp hastası olan şahısların tespitidir.

Nitelikler ile ilgili bilgiler

A1: age; Tamsayılı

A2: sex; (0,1)

A3: cp; (0,1,2,3)

A4: trestbps; Sürekli

A5: chol; Sürekli

A6: fbs; (0,1)

A7: restecg ; (0,1)

A8: thalach ; Sürekli

A9: exang ; (0,1)

A10: oldpeak; Sürekli

A11: slope ;(0,1,2)

A12: ca ; Sürekli

A13: thal; (0,1,2)

A14: num ; (0,1)

EK-6 Ionosphere veri kümesi

Veri kümesi özellikleri	:	Çok değişkenli
Nitelik özellikleri	:	Tamsayılı, Sürekli
İlgili Görev	:	Sınıflandırma
Örnek Sayısı	:	351
Nitelik Sayısı	:	34
Eksik Değer	:	Yok
Konu alanı	:	Fizik

Veri kümesi ile ilgili bilgiler

Bu radar verisi Labrador, Goose körfezindeki bir sistemden toplanmıştır. Bu sistem 6,4 KW'lık güç ile güçlendirilmiş 16 yüksek frekanslı antenden oluşmaktadır.

Hedefler iyonosferdeki serbest elektronlardır. Sınıf değerindeki "iyi" iyonosferdeki yapının tipini ifade eder. "kötü" ise, sinyallerin iyonosferden çıkmasını gösterir.

Nitelikler ile ilgili bilgiler

Veri kümesindeki bütün değerler süreklidir.

Karar niteliği farklı iki değer almaktadır.

EK-7 Connectionist Bench (Sonar, Mines vs. Rocks) veri kümesi

Veri kümesi özellikleri	:	Çok değişkenli
Nitelik özellikleri	:	Sürekli
İlgili Görev	:	Sınıflandırma
Örnek Sayısı	:	208
Nitelik Sayısı	:	60
Eksik Değer	:	Yok
Konu alanı	:	Fizik

Veri kümesi ile ilgili bilgiler

Sonardan elde edilen verilerin madenden yada kayalardan geldiğini ayırtetmek için oluşturulmuş veri kümesidir.

Nitelikler ile ilgili bilgiler

Veri kümesindeki bütün nitelikler sürekli olup karar niteliği 2 farklı değer almaktadır.

EK-8 Soybean (Small) Data Set veri kümesi

Veri kümesi özellikleri	:	Çok değişkenli
Nitelik özellikleri	:	Kategorik
İlgili Görev	:	Sınıflandırma
Örnek Sayısı	:	47
Nitelik Sayısı	:	35
Eksik Değer	:	Yok
Konu alanı	:	Yaşam

Veri kümesi ile ilgili bilgiler

Soybean hastalığı ile ilgili toplanan verilerden oluşturulmuştur.

Nitelikler ile ilgili bilgiler

Veri kümesindeki bütün nitelikler kategoriktir.

Karar niteliği 4 ayrı karardan oluşmaktadır.

EK-9 Breast Cancer Wisconsin (Diagnostic) veri kümesi

Veri kümesi özellikleri	:	Çok değişkenli
Nitelik özellikleri	:	Sürekli
İlgili Görev	:	Sınıflandırma
Örnek Sayısı	:	569
Nitelik Sayısı	:	32
Eksik Değer	:	Yok
Konu alanı	:	Yaşam

Veri kümesi ile ilgili bilgiler

Nitelikler; göğüs kanserindeki kütle ile ilgili çekilen dijital görüntüden elde edilmiştir.

Nitelikler ile ilgili bilgiler

1) Sıra numarası
2) Teşhis (M = malignant, B = benign)
3-32)

10 sürekli değer her bir kanserli hücrenin çekirdeği için hesaplanmıştır;

a) radius

b) texture

c) perimeter

d) area

e) smoothness

f) compactness

g) concavity

h) concave points

i) symmetry

j) fractal dimension

EK-10 Breast Cancer Wisconsin (Prognostic) veri kümesi

Veri kümesi özellikleri	:	Çok değişkenli
Nitelik özellikleri	:	Kategorik, Tamsayılı, Sürekli
İlgili Görev	:	Sınıflandırma, Regresyon
Örnek Sayısı	:	198
Nitelik Sayısı	:	34
Eksik Değer	:	Var
Konu alanı	:	Yaşam

Veri kümesi ile ilgili bilgiler

Her bir kayıt göğüs kanseri ile ilgili takip verilerinden oluşmaktadır. 1984'den itibaren Dr. Wolberg tarafından tutulmuştur.

Nitelikler ile ilgili bilgiler

1) Sıra numarası
2) Teşhis (R = recur, N = nonrecur)
3) Teşhis zamanı
4-33)

a) radius

b) texture

c) perimeter

d) area

e) smoothness

f) compactness

g) concavity

h) concave points

i) symmetry

j) fractal dimension

EK-11 Wine veri kümesi

Veri kümesi özellikleri	:	Çok değişkenli
Nitelik özellikleri	:	Tamsayılı, Sürekli
İlgili Görev	:	Sınıflandırma
Örnek Sayısı	:	178
Nitelik Sayısı	:	13
Eksik Değer	:	Yok
Konu alanı	:	Fizik

Veri kümesi ile ilgili bilgiler

Bu veri tabanı İtalya'da aynı bölgede yapılan 3 değişik şarabın kimyasal analizi sonuçlarını vermektedir. Analiz neticesinde 3 çeşit şarapta 13 değişik nitelik bulmuşlardır.

Nitelikler ile ilgili bilgiler

Veri kümesindeki 13 nitelikte sürekli değerler almaktadır.

Karar niteliği 3 farklı değer alabilmektedir.

EK-12 ROSETTA Programının Tanıtımı

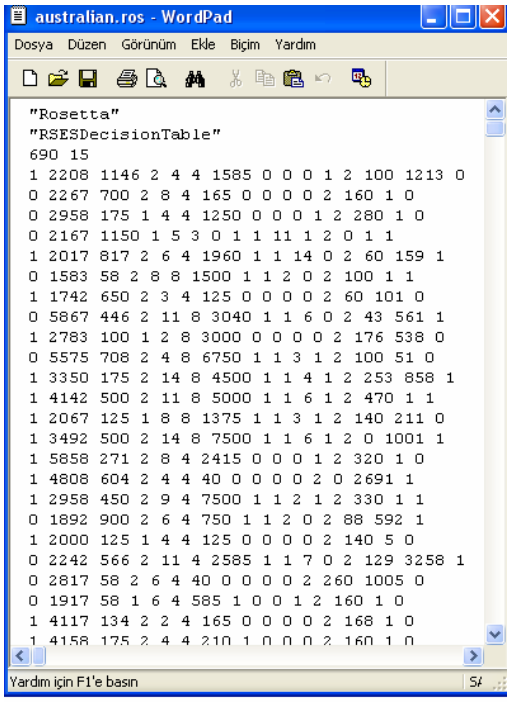
RSES kaba küme ile hesaplama için Polonya, Warşova Üniversitesindeki Mantık Grubu tarafından geliştirilmiş, veri yapıları ve algoritmaların toplandığı bir programdır.

ROSETTA ise RSES altyapısı üzerine kurulmuştur. RSES kütüphanesi üzerine istenilen şekilde kodlar yazılabilmektedir.

ROSETTA veri madenciliği ve bilgi keşfi işlemlerini desteklemek için tasarlanmıştır. Verilerin görüntülenmesinden verilerin ön işlemesine, minimal nitelik altkümesinin bulunmasından karar kurallarının çıkarılmasına kadar kaba küme tabanlı bir çok işlemi destekleyen bir yapıya sahiptir.

Bu kısımda ROSETTA hakkında detaylı bilgiler verilmeyecek olup, sadece kullanımı ile ilgili temel bilgiler verilecektir.

ROSETTA programında verilerin programa uygun olarak hazırlanması programın çalışması açısından önem taşımaktadır. Kullanılacak veri tabanı RSES karar tablosu formatında Şekil 13-1'de gösterilen şekilde herhangi bir editör'e girilmelidir.

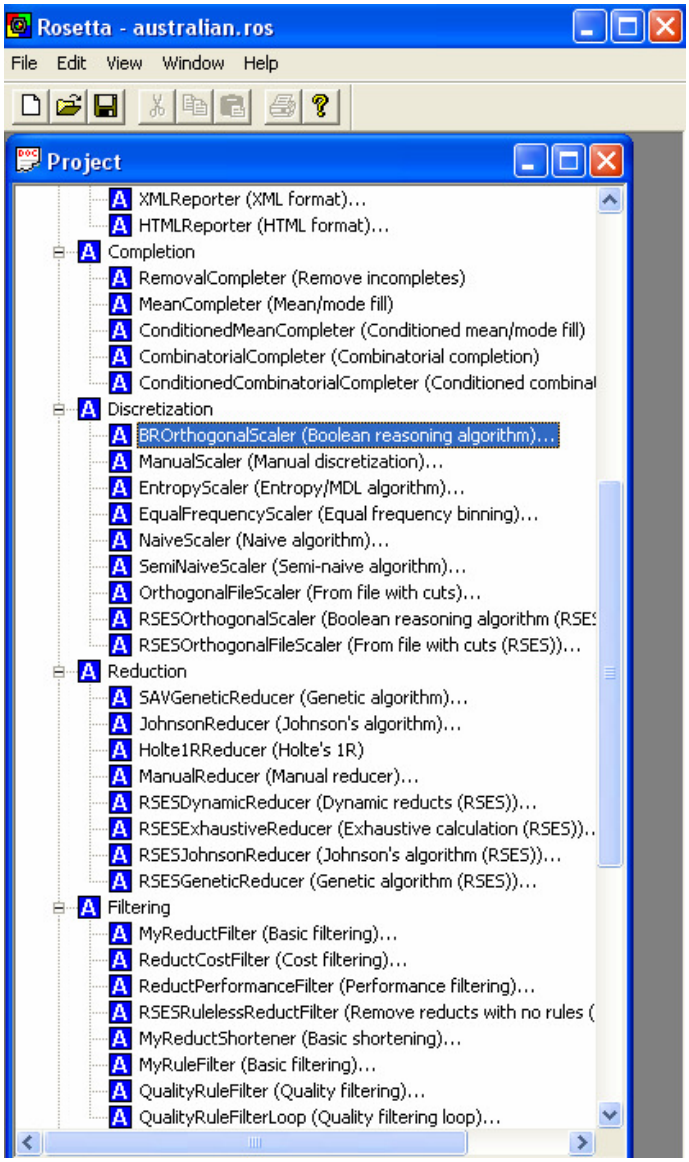


Şekil 12.1 ROSETTA programına veri girişi

EK-12 (Devam) ROSETTA Programının Tanıtımı

ROSETTA programında eksik veriler için çeşitli metotlar bulunmaktadır. Tez'de bu tür gözlemler diğer algoritmalarla karşılaştırmada herhangi bir problemle karşılaşmamak için silinmiştir. Kaba kümelerle nitelik indirgeme yapılmadan önce bir kesikli hale getirme işlemi yapılmalıdır. Program, kesikli hale getirme işlemi için kullanılabilecek 9 ayrı metot sunmaktadır. Tez'de ilgili veri kümesi için nitelik indirgemesi işleminden önce daha iyi sonuçlar elde edilebileceğinden Boolean Reasoning algoritması kullanılmıştır.

Aşağıda ROSETTA programı için kesikli hale getirme ve indirgeme metotları gösterilmektedir.



Şekil 12.2 ROSETTA algoritmaları

EK-13 WEKA programının tanıtımı

WEKA, çeşitli veri madenciliği görevleri için makine öğrenimi algoritmalarını içinde barındıran bir programdır. WEKA, önişleme, sınıflandırma, regresyon, kümeleme, birliktelik kuralları ve görselleştirme için bir ç.ok algoritma içermektedir. Ayrıca kullanıcı kendi ihtiyaçları doğrultusunda yeni algoritmalar oluşturabilmektedir.

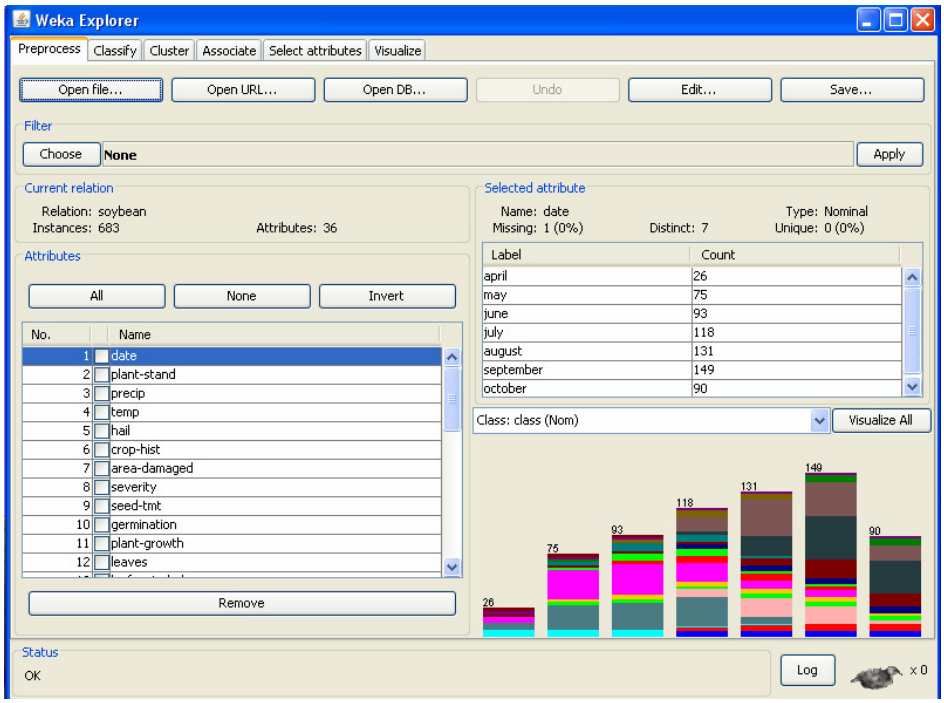
Program ayrıca, sınıflandırma görevi için 10 çarpraz geçerlik testini kendisi yapmaktadır.

WEKA programında çalışabilmek için veriler arff şeklindeki formata getirilmelidir. Bu formatta, ilgi veri tabanının ismi verildikten sonra, nitelikler ve özellikleri tanımlanmaktadır.

@RELATION soybean

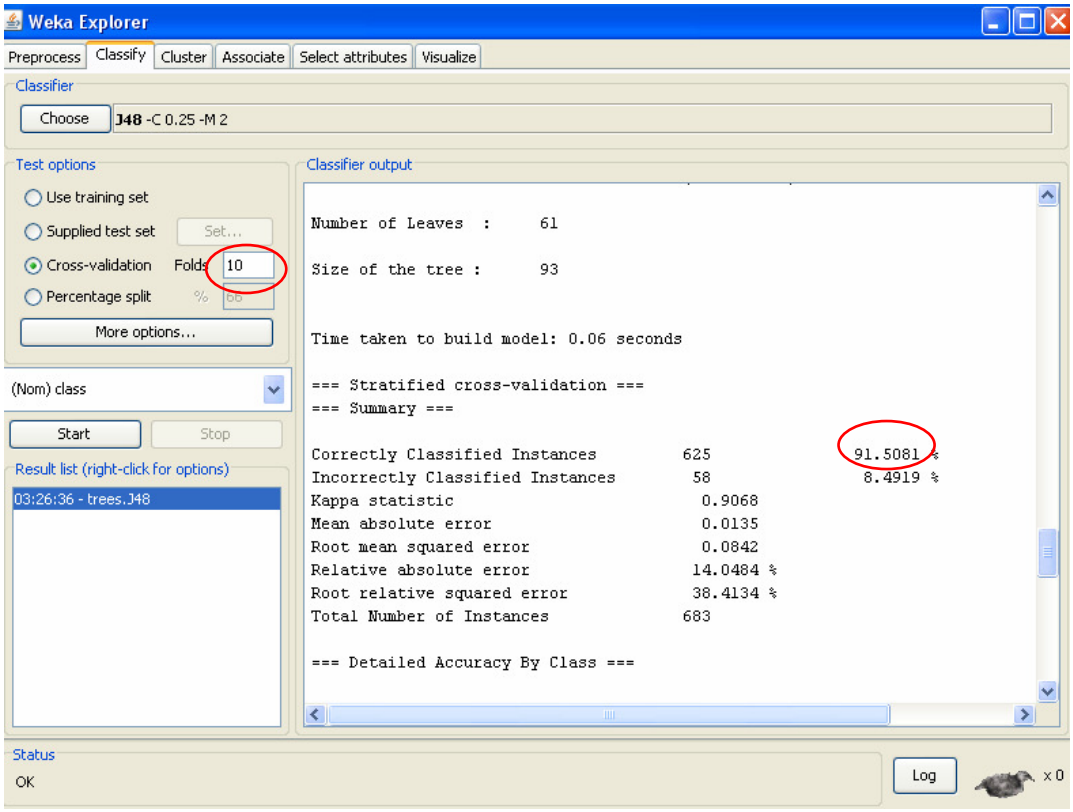
@ATTRIBUTE shriveling	{absent,present}
@ATTRIBUTE roots	{norm,rotted,galls-cysts}
@ATTRIBUTE class	{diaporthe-stem-canker, charcoal-rot}

WEKA programının arayüzü aşağıdaki ekran çıktılarında sunulmuştur.



Şekil 13.1 WEKA programı arayüzü

EK-13 (Devam) WEKA programının tanıtımı



Şekil 13.2 WEKA ile yapılan sınıflandırma

ÖZGEÇMİŞ

Kişisel Bilgiler

Soyadı, adı	: AĞIRGÜN, Bekir
Uyruğu	: T.C.
Doğum tarihi ve yeri	: 07.10.1975 Ankara
Medeni hali	: Evli
Telefon	: 0 (312) 456 43 78
E-Posta	: bekiragi@gmail.com

Eğitim

Derece	Eğitim Birimi	Mezuniyet tarihi
Yüksek lisans	Gazi Üniversitesi /Endüstri Müh	2000
Lisans	Kara Harp Okulu	1996
Lise	Kuleli Askeri Lisesi	1992

Yabancı Dil

İngilizce

Hobiler

Bilgisayar teknolojileri ve Adli bilişim.