BÜYÜK BOYUTLU SINIFLANDIRMA PROBLEMLERİ İÇİN MATEMATİKSEL PROGRAM YAKLAŞIMLARI

Mehmet Tahir Çiftçi

Yüksek Lisans Tezi

Endüstri Mühendisliği Anabilim Dalı

Ekim 2011

ANADOLU ÜNİVERSİTESİ

JÜRİ VE ENSTİTÜ ONAYI

Mehmet Tahir Çiftçi'nin **"Büyük Boyutlu Sınıflandırma Problemlerinin Çözümü için Yeni Bir Matematiksel Programlama Yaklaşımı"** başlıklı bu çalışma **Endüstri Mühendisliği** Anabilim Dalındaki, Yüksek Lisans Tezi 19.08.2011 tarihinde, aşağıdaki jüri tarafından Anadolu Üniversitesi Lisansüstü Eğitim Öğretim ve Sınav Yönetmeliğinin ilgili maddeleri uyarınca değerlendirilerek kabul edilmiştir.

		Adı-Soyadı	İmza
Üye(Tez Danışmanı)	:	Yard. Doç. Dr. GÜRKAN ÖZTÜRK	.....
Üye	:	Yard. Doç. Dr. ŞEREF TÜZEMEN	.....
Üye	:	Yard. Doç. Dr. ÖZGÜR YILMAZEL	.....

Anadolu Üniversitesi Fen Bilimleri Enstitüsü Yönetim Kurulu'nun ..... tarih ve.....sayılı kararıyla onaylanmıştır.

Enstitü Müdürü

ÖZET

Yüksek Lisans Tezi

BÜYÜK BOYUTLU SINIFLANDIRMA PROBLEMLERİN ÇÖZÜMÜ İÇİN YENİ BİR MATEMATİKSEL YAKLAŞIM

Mehmet Tahir ÇİFTÇİ

Anadolu Üniversitesi

Fen Bilimleri Enstitüsü

Endüstri Mühendisliği Anabilim Dalı

Danışman: Yard. Doç. Dr. Gürkan ÖZTÜRK

2011,61 sayfa

Sınıf etiketleri bilinen bir veri kümesi üzerinden oluşturulan modeller yardımıyla, yeni örneklerin hangi sınıfa atanacağının tahmin edilmesi, sınıflandırma problemi olarak adlandırılmaktadır. Birçok alanda karşımıza çıkan bu problemlerin çözümü için farklı disiplinlerden araştırmacılar, yeni yöntemler üzerine çalışmalar yapmaktadır. Böylece her geçen gün yeni yaklaşımlar ve çözüm yöntemleri bu çalışma alanına sunulmaktadır.

Bu yüksek lisans tezinde, büyük boyutlu sınıflandırma problemlerin çözümü için temelinde çokyüzlü konik fonksiyonlar olan yeni bir matematiksel programlama yaklaşımı sunulmuştur. Yeni önerilen yaklaşımda, problemlerin etkin ve hızlı şekilde çözümü için K-Ortalamalar ve gürbüz doğrusal programlama yaklaşımları kullanılmıştır. Literatürde en sık karşılan büyük boyutlu problemler, hem geliştirilen yeni yaklaşım ile hem de alanda en yaygın kullanılan ve başarıları kanıtlanmış yöntemler ile çözdürülmüştür. Elde edilen sonuçlar yeni yaklaşımın belirgin şekilde seçilen diğer yöntemlere göre üstün geldiğini göstermiştir.

Anahtar Kelimeler: Sınıflandırma, Çokyüzlü Konik Fonksiyonlar
K-Ortalamalar, Gürbüz Doğrusal Programlama,
Matematiksel Programlama

ABSTRACT

Master of Science Thesis

A NEW MATHEMATICAL APPROACHES TO LARGE SCALE CLASSIFICATION PROBLEMS

Mehmet Tahir ÇİFTÇİ

Anadolu University

Graduate School of Sciences

Industrial Engineering Program

Supervisor: Assistant Professor Gürkan ÖZTÜRK

2011,61 pages

Classification problem is called that the problem is to estimate class of new instances by using models that is constructed on the data sets which are known class labels. Researchers from different disciplines study on new approaches to solve this problems encountered in many areas. Hence new approaches and solution methods are presented in this field.

In this thesis, a new mathmematical programming approach based on polyhedral conic functions is presented to solve large scale classification problems. In order to solve problems effectively and quickly with this approach; k-means and robust linear programming are used. Frequently used large scale test problems from the literature are solved either proposed approach or commonly used methods with proven success in the field. Obtained results are shown that new approach is clearly superior than the others.

Keywords : Classification , Polyhedral Conic Functions, K-Means, Robust Linear Programming, Mathematical Programming

TEŞEKKÜR

Gerek lisans, gerekse yüksek lisans dönemimde bana her türlü desteği veren , veri madenciliği ve sınıflandırma konularında çalışmamı teşvik eden, değerli hocam sayın Yard. Doç. Dr Gürkan ÖZTÜRK ' e teşekkürü bir borç bilirim.

Yüksek Lisans yapmam konusunda bana gerekli bütün desteği veren müdürüm sayın Hamit GÜNAŞAN'a, bu zorlu dönemde benden yardımlarını hiç bir zaman esirgemeyen aileme teşekkür ederim.

Mehmet Tahir Çiftçi

Ekim 2011

İÇİNDEKİLER

ŞEKİLLER DİZİNİ

ÇİZELGELER DİZİNİ

1. GİRİŞ

Bilgi çağını yaşadığımız bugünlerde teknolojinin gelişmesi ile birlikte veriler dijital ortamda saklanmaya başlanmıştır. Verilerin çeşit ve özelliklerinin artması ile birlikte çok büyük veri tabanları ortaya çıkmıştır. Bununla birlikte her geçen gün kararların hızlı ve doğru bir şekilde verilmesi oldukça önem kazanmıştır. İşte bu aşamada amaç, hızlı bir şekilde verilere ulaşarak değerli bilgiler türetmek ve karar vericiye her yeni durum için belirli bir oranda doğru sonucu sunan modelleri elde etmek olmuştur. Veri yığınları içindeki değerli olan bu bilgilerin çıkarılarak önemli karar problemlerinin çözümünde kullanılması, araştırmaların veri madenciliği alanında yoğunlaşmasına sebep olmuştur.

Veri madenciliğinde genel olarak üç tip problem karşımıza çıkmaktadır: kümeleme, birliktelik analizi ve sınıflandırma problemleri. Kümeleme, sınıf etiketleri olmayan nesnelerin birbirine olan benzerliklerine göre kümelere ayrılması olarak tanımlanırken özellikle istatistiksel tahmin, DNA analizi ve coğrafi bilişim sistemlerinde kullanılmaktadır. Birliktelik analizi, nesnelerin birbiri ile olan ilişkilerini tanımlayan bir modeldir. Genel olarak perakendecilik sektöründe uygulanmaktadır. Sınıflandırma, bir veri kümesinin önceden bilinen sınıflara atanması anlamına gelmektedir. Sınıflandırma, veri madenciliğinin yanı sıra makine öğrenmesi (machine learning) alanında da sıklıkla karşılaşılan temel problemlerden biridir. Hayatımızın bir çok alanında sınıflandırma problemleri ile karşılaşmaktayız. Veri madenciliği bakış açısıyla kredi başvurularının değerlendirilmesi, biyopsi olmadan bir kaç tahlil ile hastalıkların teşhis edilmesi, e-postaların spam olarak tespit edilmesi sınıflandırma problemlerinin uygulama alanlarına örnek olarak gösterilebilir. İnternette beğendiğiniz herhangi bir şeyin fotoğrafını temel alarak buna benzer olanları bulmaya yarayan bir arama motoru olan görüntülü arama tabanlı akıllı alışveriş sitesi like.com'un çalışma mantığı bilgi erişim sistemlerine dayanmaktadır. Bir Türk tarafından geliştirilen bu site 100 milyon dolara Google tarafından satın alınmıştır. Bu durum bu alanda geliştirilen yöntemlerin ne derece önemli olduğunu da gözler önüne sermektedir.

Sınıflandırma problemleri sınıf sayısına göre iki ve çok sınıflı problemler olarak

karşımıza çıkmaktadır. Araştırmacılar çoğunlukla iki sınıflı problemleri çözmeye çalışmaktadır. Bunun sebebi, çok sınıflı problemlerin çözümünde iki sınıflı problemleri çözmek üzere geliştirilen yöntemlerin kullanılmasıdır. İki sınıflı problemler için geliştirilen yöntemler farklı şekillerde bir araya getirilmektedir. En sık karşılaşılan bir araya getirme yöntemleri bire-karşı-bir$(1-e-1)$, bire-karşı-hepsi $(1-e-h)$ ve yönlü çevrimsiz serimdir. Gerek iki gerekse çok sınıflı problemlerin çözümü için geliştirilen matematiksel programlama temelli yaklaşımların bazıları; doğrusal ayırma, h-çok yüzlü ayırma, enb-enk ayırma, bütünsel ağaç eniyileme ve çok yüzlü konik fonksiyonlar temelli yaklaşımlardır.

Son yıllarda, sınıflandırma problemlerinin çözümünde kullanılan matematiksel programlama temelli yaklaşımların arasına, çok yüzlü konik fonksiyonlar(ÇKF) olarak adlandırılan yeni bir fonksiyon sınıfını esas alan yeni yaklaşımlar eklenmiştir. ÇKF, grafiği koni, seviye kümesi dış bükey polihedron olan bir fonksiyondur. Seviye kümesi, n boyutlu uzayı elde edilen dış bükey polihedronun içi ve dışı olmak üzere ikiye ayırır. Bu özellik sınıflandırma problemlerinin çözümü için sıklıkla kullanılan hiper düzlemlere göre farklı üstünlükler sağlamaktadır. ÇKF'lerde elde edilen koninin tepe noktası bir anlamda merkez noktası olarak düşünülmektedir. Bu noktanın belirlenmesi sınıflandırma başarısı ve çözüm süresi açısından oldukça kritiktir. Bugüne kadar önerilen ÇKF temelli yaklaşımlar esas olarak farklı merkez noktası belirleme stratejisine göre oluşturulmuştur.

ÇKF temelli bir yaklaşım olan ÇKF algoritması ile iki küme birden çok fonksiyon kullanılarak yüzde yüz başarı ile ayrılabilmektedir. Ancak elde edilen test başarıların literatür sonuçları ile rekabetçi değerler sunmasına rağmen, test ve eğitim başarılar arasındaki farkın fazla olması istenmeyen bir durumdur. Eğitim başarısının yüksek, test başarısının düşük olduğu durum literatürde aşırı uyum (overfitting) olarak adlandırılmaktadır. Aşırı uyum sınıflandırma problemlerinin çözümünde istenmeyen bir durumdur.

Bu çalışmada, büyük boyutlu sınıflandırma problemlerinin çözümü için ÇKF temelli etkin bir matematiksel programlama yaklaşımının geliştirilmesine odaklanılmıştır. Bu sebeple elde edilecek olan fonksiyonların merkezlerini hızlı ve doğru bir şekilde belirlemek üzere k-ortalamalar yaklaşımı kullanılmıştır. Aşırı uyum sorununu en aza indirmek için ise gürbüz doğrusal programlama (Robust Linear Programming)

yaklaşımından yararlanılmıştır.

Çalışmanın ikinci bölümünde, veri madenciliği hakkında genel bilgiler, sınıflandırma problemlerinin tanımı, sınıflandırma problemlerinin çözümünde literatürde kullanılan yaklaşımlar ve bu yaklaşımların önerdiği sınıflandırıcılar hakkında bilgiler verilmiştir.

Üçüncü bölümde ise büyük boyutlu sınıflandırma problemlerinin çözümü için önerilen yeni yaklaşımın temellindeki teknikler hakkında genel bilgiler verilmiştir. Devam eden kısımda, önerilen yeni yaklaşımın algoritması ve açıklayıcı bir örnek üzerindeki uygulaması anlatılmıştır. Son kısımda ise, literatürdeki büyük boyutlu problemlerin bu yeni yaklaşım ile çözümü ve diğer çözümler ile karşılaştırılmasından bahsedilmiştir.

Gelecekte bu çalışmanın devamında yapılması planlanan çalışmalar hakkında çeşitli bilgiler ve öneriler, dördüncü bölüm olan son bölümde sunulmuştur.

2. SINIFLANDIRMA PROBLEMLERİ ve ÇÖZÜM YAKLAŞIMLARI

Bu bölümde veri madenciliği ve sınıflandırma hakkında genel bilgilendirmeler yapılmıştır. Bölümün devamında sınıflandırma problemlerinin tanımı, bu problemlerin çözümü için literatürde geliştirilmiş olan yöntemler hakkında bilgiler verilmiştir.

2.1 Veri Madenciliği

Veri madenciliği; önceden bilinmeyen, geçerli ve uygulanabilir bilginin veri yığınlarından dinamik bir süreç ile elde edilmesi olarak da tanımlanmaktadır. Bir diğer tanım ise, veri madenciliği, istatistik ve matematik tekniklerle birlikte ilişki tanıma teknolojilerini kullanarak, depolama ortamlarında saklanmış bulunan veri yığınlarının elenmesi ile anlamlı yeni ilişki ve eğilimlerin keşfedilmesi süreci olarak tanımlanmıştır [4]. Genel olarak veri madenciliği, büyük boyutlu verilerden kullanışlı ve gelecek hakkında tahminlerde bulunmamızı sağlayabilecek verilerin ortaya çıkarılması olarak tanımlanabilir. İstenen yararlı verilere ulaşmak için konu ile ilgili olan bütün verilere sahip olunması gerekmektedir. Bütün verilere sahip olunması ise günümüzde ancak bilgisayar teknolojisi ile sağlanabilmektedir.

Bilgisayar endüstrisindeki hızlı gelişim ile birlikte verilerin saklanması ve depolanması için veri tabanları sistemlerinin geliştirilmesi de kaçınılmaz olmuştur. Her geçen gün yeni eklenen veriler ile birlikte veri tabanları çok büyük boyutlara ulaşmıştır. Bu da verilerin analiz edilerek yararlı ve anlaşılabilir bilgilerin türetilmesi için veri madenciliği tekniklerinin gelişmesine ve her geçen gün artan bir araştırma alanı haline gelmesinde etkili olmuştur [5]. Fakat burada karıştırılmaması gereken nokta, veri madenciliğinin kendi başına bir çözüm değil çözüme ulaşmak için verilecek karar sürecini destekleyen, problemi çözmek için gerekli bilgileri sağlamaya yarayan bir araç olduğudur. Veri madenciliğinin görevi; analistin'e, iş yapma aşamasında oluşan veriler arasındaki şablonları ve ilişkileri bulması konusunda yardım etmektedir [4].

Veri madenciliği veritabanı ve veri ambarı teknolojisi, istatistik, makine öğrenmesi, örüntü tanıma, yapay sinir ağları, veri görüntüleme, bilgi erişimi, görüntü ve sinyal işleme gibi birçok disipline ait teknik yaklaşımlar içerir[5].

Veri madenciliğinin kullanıldığı bir diğer alan ise metin madenciğiliğidir. Metin madenciliği, veri madenciliği teknikleri ile yazılı belgeler arasındaki ilişkileri, örtüleri bulmak olarak tanımlanmaktadır. Metin madenciliğinin çözümü için kullanılan tekniklerden birisi de bilgi erişim sistemleridir. Kütüphane veri tabanları (anahtar kelime, başlık, yazar, konu vs ile büyük veri tabanlarında arama), Metin tabanlı arama motorları (Google, Yahoo vs), Multimedya arama (Görsel öğelerle arama), Soru cevap sistemleri (AskJeeves, Answerbus) örnek bilgi erişim sistemleridir. Bilgi erişim sistemlerinin çözümünde veri madenciliği teknikleri kullanılmaktadır.

Veri madenciliği teknikleri özellikle işletmelerde çeşitli alanlarda başarı ile kullanılmaktadır. Başlıca kullanım alanları pazarlama, bankacılık, sigortacılık, perakendecilik, borsa, telekomünikasyon, sağlık ve ilaç, endüstri, bilim ve mühendislik uygulamalarıdır. Veri madenciliği tekniklerinden bazıları aşağıdaki gibidir[5];

- Birliktelik Analizi
- Sınıflandırma
- Kümeleme Analizi
- Tanımlama ve Ayrımlama
- Sıradışılık Analizi
- Evrimsel Analiz

2.2 Veri Madenciliği Teknikleri

Günümüzde verilerin elde edilmesi, saklanması ve ulaşılabilirliği teknolojideki gelişmelere paralel olarak kolaylaşmış ve ucuzlamıştır. Büyük boyutlarda ve hızlı bir şekilde toplanan verilerin çeşitli analizler sonucunda anlamlı bilgilere dönüştürülmesi süreci olarak tanımlanan veri madenciliğinin, günümüzde en çok kullanılan teknikleri kümeleme ve sınıflandırma problemleridir.

Veri madenciliğinde vurgulanan unsurlar istatistiğin tanımı içinde zaten yer almaktadir. İstatistik, verilerin toplanması, sınıflandırılması, özetlenmesi, grafik ve tablolarla sunulması, analiz edilerek ana kütle hakkında anlamlı bilgiler elde edilmesi ve yorumlar yapılmasıdır. Veri madenciliğinde ulaşılmak istenen amaç aslında istatistik biliminin amacı ile aynı doğrultudadır. Verilerden bilgiyi keşfetmek. Zaten veri madenciliğinde kullanılan temel aracın istatistiksel yöntemler oldugu birçok tanımda ve uygulamada vurgulanmaktadır. Her ikisinde de temel olan ögeler veri ve bilgidir. Bu nedenle birbiriyle oldukça örtüşen konulardır. Bu yüzden bir kişi tarafından veri madenciliği olarak adlandırılan bir problem başka biri için istatistik problemi olabilir[4].

Kredi başvurularının değerlendirilmesi, banka kartı harcamalarında sahtekarlık olup olmadığının kararının verilmesi, kara para aklama ve buna benzer finansal suçların belirlenmesi, ses tanıma, gazete haberlerini ayırma veri madenciliği tekniklerinin başarı ile kullanıldığı durumlardır[6].

Sınıflandırma problemi ise, nesnelerin her bir nitelik kümesi ve önceden tanımlanmış olan sınıf etiketlerine atanmasından oluşur. Veri kümesindeki her bir veri için nitelik sınıfı ve sınıf etiketi bilgisi bulunmaktadır. Bu verilere göre elde edilen sınıflandırıcı model, devamında gelen kayıtları sınıflandırmak için kullanılabilecek kısa ve anlamlı veriler türetir. Denetimli sınıflandırma problemlerinde verilerin sınıf etiketleri mevcuttur. Buradaki amaç, sınıf etiketi mevcut olan nesneler üzerinde belirli bir amaca uygun modeller türetmektir[7].

Sınıflandırma modeli, “Tanımlama”, “Tahmin”, “Birliktelik Analizi”, “Kümeleme Analizi” gibi veri madenciliği amaçları için kullanılmaktadır[1].

Tanımlama: Sınıflandırma modeli, farklı sınıfların objelerini ayırt etmek için açıklayıcı bir araç olarak da hizmet edebilir. Örneğin, hem biyologlar hem de diğer kişiler için vücut sıcaklığı, deri, doğurganlık gibi tür özelliklerin bir omurgalıyı memeli, sürüngen, kuş veya balık olarak tanımladığını açıklayan, tanımlayıcı bir modele sahip olmak yararlı olacaktır[1].

Tahmin: Sınıflandırma, evet/hayır, memeli/sürüngen/kuş gibi kesikli çıktılar ile ilgilenir. Tahmin ise sürekli değerler alan çıktılar ile ilgilenir. Tahmin, bazı girdi verileri verildiğinde gelir düzeyi, oy miktarı, gelecek dönem satış tahmini gibi bilinmeyen

sürekli değişkenlere ilişkin değerlerin bulunması için kullanılır[1].

Birliktelik Analizi: Birliktelik analizi, belirli türlerdeki veri ilişkilerinin tanımlayan bir modeldir. Herhangi bir ürün alındığında bu ürünün yanında bir başka ürünün de satın alınması bir birliktelik kuralı verir. Birliktelik analizi çoğunlukla perakendecilik sektöründe faaliyet gösteren işletmelerde uygulanmaktadır. Örneğin, bir süpermarkette yapılan alışverişlerin incelenip hangi ürünün hangi ürünle birlikte satın alındığının belirlenmesi birliktelik kurallarını ilgilendirir[6].

Birliktelik analizi, herhangi bir veritabanında birliktelik kurallarının tanımlanması veritabanı bilgi sürecinin ilk adımıdır. Veritabanındaki herhangi bir X'in aynı zamanda Y'yi içermesi bir birlikteliktir. Bu durum, "Bira içeren %30 alışverişin, %2'si aynı zamanda çocuk bezi de içermektedir." Burada %30 güven seviyesini, %2 ise bu güven seviyesine olan desteği belirtmektedir[6].

Uç Değer Analizi: Bir veritabanı, genel davranışa veya verilerin modelini uymayan nesneler içerebilir. Bu veriler uç verileridir. Çoğu veri madenciliği yöntemi, uç değerleri gürültü veya istisna diye göz ardı eder. Ancak sahtekarlık tespiti gibi bazı uygulamalarda nadiren gerçekleşen olaylar düzenli olarak meydana gelen olaylardan daha ilginç olabilir. Uç değer verilerinin analizleri uç değer veri madenciliği olarak adlandırılır. Örneğin, bir hileli kredi kartı kullanımı, olağan kredi kartı hareketlerinden farklı, aşırı miktarda ürün satın alma gibi uç değer durumlarının tespiti ile ortaya çıkarılabilir[8].

2.3 Kümeleme

Kümeleme analizi özellikle bilim ve iş alanında, birçok durumda uygulanan etkili ve kolay yorumlanabilen bir yöntemdir. Kümeleme analizi veri madenciliğinin en önemli alanlarından birisidir; amacı, nesneleri birbirine olan benzerliklerine göre benzeyenler bir kümeye, benzemeyenler ise bir başka kümeye toplamaktır. Benzersizlikler ise nesneleri tanımlayan özelliklerin değerleri temel alınarak belirlenir. Birbirine benzer nesne gruplarının işaretlenmesi ya da başka gruplarla olan farklılıklarının bulunması ile kümeler oluşturulur. Verilerin kümeleme analizine göre modellenmesinde matematik, istatistik, makine öğrenimi ve yapay zeka gibi bir çok alandan yararlanılır.

Makine öğrenimi açısından, her bir küme gizli bir örüntüyü temsil eder ve uygulanan öğrenme ise bir denetimsiz öğrenmedir. İstatistikte çok değişkenli istatiksel tahmin, ses ve resim tanıması, DNA analizi, coğrafi bilişim sistemleri ve bunlarla ilgili alanlarda kullanılmaktadır [6, 5].

Kümeleme analizi, veri kümesindeki nesnelerin sınıflandırılmasını ayrıntılı bir şekilde açıklamak amacıyla geliştirilmiştir. Bu amaca yönelik olarak, ele alınan örnekte yer alan özellikler, aralarındaki benzerliklere göre gruplara ayrılır, daha sonra bu gruplara dahil edilen bireylerin profili ortaya konur. Bir başka ifade ile kümelemenin amacı, öncelikle ele alınan örnekte gerçekte var olduğu bilinen, varlıklar (birey ya da nesne) arasındaki benzerliklere dayanan az sayıdaki karşılıklı özel grupları oluşturmak, daha sonra bu gruplara giren özellik profilini ortaya koymaktır. Diğer bir hedef ise benzer elemanların gruplanmasıyla veri setini küçültmektir. Satış hareketleri veya çağrı merkezi kayıtları gibi çok fazla parametre içeren çok büyük miktarlardaki verileri analiz etmede en uygun yöntemlerden biri kümelemedir [5].

Kümeleme analizi, sonuçların grafiksel olarak görüntülenebiliyor olması sayesinde benzerliklerin kolay tespit edilmesini sağlar. Yine grafiksel gösterim sayesinde aykırı olan durumların ve sıra dışı verilerin belirlenmesinde etkilidir. Diğer veri madenciliği tekniklerine göre çok büyük veriler üzerinde çalışabildiği için önemli bir avantaj sağlar. Hatta kümeleme analizi, karar ağaçları gibi teknikler için çok büyük boyutlu verilerin bölünmesine en uygun başlangıç noktalarının belirlenmesini sağlar. Kümeleme analizinin bu gibi avantajlarının dışında farklı tiplerde özelliklere sahip (sayısal, sözel gibi) nesnelerin karşılaştırmasına pek olanak sağlayamamaktadır. Kümeleme analizinde benzerlik kriteri olarak genelde uzaklık kavramı kullanılmaktadır. Uzaklık hesaplamak için kullanılan bazı ölçüler ise[5];

- Minkowski uzaklığı
- Manhattan (City-Blok) uzaklığı (n=1)
- Öklid (Euclidean) uzaklığı (n=2)
- "Supremum" (L_{max} norm, $L_{\propto}$ norm) uzaklığı (n=$\propto$)
- Tchebyschev uzaklığı

- Mahalanobis uzaklığı
- Canberra uzaklığı
- Bray Curtis (Sorensen) uzaklığı
- Kosinüs benzerliği
- Genişletilmiş Jaccard benzerliği
- Pearson İlişkisi
- Spearman benzerliği

Literatürde bir çok kümeleme algoritmasının adı geçmektedir. Algoritmalar birbirinden, kümelemenin oluşturuluş şekline göre ayrıldıkları gibi kullanılan veri türüne, yapılacak olan çalışmanın amacına göre de farklılıklar gösterir. Kümeleme algoritmaları, genel olarak hiyerarşik ve bölümlemeli olarak ikiye ayrılırken, bu konuda yapılmış olan yöntemler genel olarak şunlardır[6];

- Bölümlemeli Yöntemler
- Hiyerarşik Yöntemler
- Grid Temelli Yöntemler
- Kategorik Verinin Yinelenmesine Dayanan Yöntemler
- Kısıtlara Dayalı Yöntemler
- Makine Öğrenmesi Alanında Kullanılan Yöntemler

2.4 Sınıflandırma ve Sınıflandırma Problemleri

Sınıflandırma bir veri kümesinin belirli sayıdaki sınıfa atanması anlamına gelmektedir. Sınıflandırma problemleri ise bu atama işleminin yapılması için sınıflandırıcıların geliştirilmesidir. Sınıflandırma problemleri genel olarak iki aşamadan oluşmaktadır. İlk aşama nitelikler ile tanımlanmış olan veri kümesindeki noktaların sınıf etiketlerine atanması için sınıflandırıcıların belirlenmesi, ikinci aşama ise elde edilen sınıflandırı-

cılara göre yeni noktaların sınıflara atanmasıdır. Denetimli veri sınıflandırma problemi, mesaj, başlık ve içeriğine göre spam e-postaların belirlenmesi, hastalıklı hücrelerin belirlenmesi gibi bir çok kullanım alanı mevcuttur. Sınıf etiketleri bilinen veri kümesi kullanılarak, yeni verilerin sınıf etiketlerinin belirlenmesine literatürde denetimli (supervised) sınıflandırma olarak adlandırılmaktadır. Denetimli sınıflandırmada genel olarak verilerin sınıf etiketleri mevcuttur ve bu etiketleri elde etmek kolaydır. Fakat bazı veri kümelerinde sınıf etiketlerini elde etmek hem maliyetli hem de zordur. Bu gibi problemlere litaratürde yarı-denetimli (semi-supervised) sınıflandırma problemleri olarak adlandırılır. Müzik, web sayfası, protein, doküman sınıflandırma problemleri yarı-denetimli sınıflandırma problemlerine örnektir [9].

Sınıflandırma problemlerinde öğrenme denetimlidir ve amaç yeni örnekleri mümkün olan en yüksek doğruluk oranı ile sınıflara atayacak modelleri elde etmektedir.

2.5 Sınıflandırma Problemlerinin Çözümünde Kullanılan Yöntemler

Sınıflandırma problemlerinin çözümü için ayırma analizi, bayes sınıflandırması, sinir ağları, karar ağaçları ve destek vektör makinaları geliştirilmiş olan yöntemlerden bazılarıdır. Herbir yöntem, verilerin özellikleri ve sınıf etiketlerine göre modelin tanımlanması için bir öğrenme algoritması kullanır. Bir öğrenme algoritması tarafından üretilen model, hem giriş verilerine en iyi şekilde temsil etmeli, hem de daha önce hiç gözlemlenmemiş kayıtların sınıf etiketlerini doğru şekilde tahmin edilebilmelidir. Bu yüzden, bir bilgi algoritmasının ana amacı, modelleri iyi bir genelleme yeteneğiyle oluşturmaktır; yani, modeller daha önceden bilinmeyen kayıtların sınıf etiketlerini en doğru şekilde bildirebilmelidir[1].

2.5.1 Ayırma analizi

Ayırma ve sınıflandırma, farklı kümelerdeki nesnelerin ayrılması ve yeni bir nesnenin önceden tanımlı gruplara atanması ile ilgili çok değişkenli tekniklerdir. Ayırma analizi, nesnelerin özelliklerinden dolayı gözlenen farklılıkları araştırmak için kullanılır. Sınıflandırma ise, yeni nesnelerin atanmasında kullanılan iyi tanımlanmış kuralları yönetme anlamında daha az açıklayıcıdır. Ayırma ve sınıflandırmanın amaçları[10]:

- *Amaç 1.* Gözlemlerin ayırıcı özelliklerini grafiksel yada matematiksel olarak tanımlar. Ana kütleleri, sayısal özelliklerine göre mümkün olduğunca ayıran ayırıcı fonksiyonlar bulunmaya çalışılır.

- *Amaç 2.* Nesneleri (gözlemleri), iki veya daha fazla sınıf için sıralar. Burada önemli olan nokta, yeni nesneleri en iyi şekilde etiketli sınıflara atayabilecek bir kuralın çıkarılabilmesidir.

Nesneleri ayıran bir fonksiyon, bir sınıflandırıcı fonksiyon olarak hizmet verebilir ya da tersine, nesneleri sınıflara atayan bir fonksiyon ayırıcı bir yordam olarak önerilebilir. Uygulamada *Amaç* 1 ve 2 çoğunlukla birbirine karışır ve dolayısıyla ayırma ile sınıflandırma arasındaki fark çok net değildir [8].

Doğrusal, lojistik ve karesel ayırma analizlerinin yanında k-en yakın komşu, bayes algoritmaları ve ana bileşenler analizi yine sınıflandırma problemlerinde kullanılan diğer istatistiksel yaklaşımlardandır [11].

2.5.2 Bayes sınıflandırma

Sınıflandırma işleminde istatiksel teknikler de kullanılmaktadır. Bunlardan birisi de Bayes teoremine dayanmaktadır. Bazı uygulamalarda, özellik kümesi ve sınıf değişkeni arasındaki ilişki deterministik olmayan yapıdadır. Bazı dış faktörler yüzünden, özellik kümesi eğitim örneklerine özgü olmasına rağmen, bir test kaydı için sınıf etiketini kesinlikle doğru tahmin edileceği söylenemez. Diyetine ve egzersiz yapma sıklığına göre kalp krizi geçirme riski dış etkenlerden (kalıtım, alkol kullanımı vb) dolayı kesinlikle doğru tahmin edilemez [12].

Bayes sınıflandırıcıları istatiksel sınıflandırma teknikleri arasında yer alır. Bu sınıflandırma işlemine başlarken X kümesi sınıf etiketi bilinmeyen veri kümesi olarak kabul edilsin. H ise bu X veri örneğinin C sınıfına ait olduğu iddia edilen hipotez olsun. O halde, H'nin C sınıfına ait olduğun varsayımıyla $P(H|X)$ olasılığını hesaplamamız söz konusudur. Burada $P(H|X)$, H hipotezinin X üzerinde koşullandırılmasına "sonrasal olasılık" olarak kabul edilir[12].

Örneğin bir torbada bazı cisimlerin bulunduğunu varsayalım. Elimizdeki bilgiler X'i tanımlar. Cisimlerin yuvarlak ve kırmızı olduğunu da bildiğimizi varsayarsak bu

durumda $P(H)$ bir "önsel olasılık" olarak karşımıza çıkacaktır. Yani başlangıçta bu olasılığın ne olduğunu biliyoruz. Ancak $P(X|H)$ olasılığı ise H üzerine kurulduğunda bir "sonrasal olasılık" olarak değerlendirilir. Yani X'in ilgili sınıfı bu durumda "Bayes" bağlantısı şu şekli alır [12]:

$$P(H|X) = \frac{P(X|H) \times P(H)}{P(X)} \tag{2.1}$$

Sınıf koşullu olasılıklarının tahmin edilmesi için "Bayes" sınıflandırma yönteminin iki farklı uygulaması bulunmaktadır[12]: Saf Bayes ve Bayes güven ağı.

Saf Bayes Sınıflandırıcılar: Saf Bayes sınıflandırıcıları, özelliklerin şartlı olarak bağımsız olduğu varsayımı altında sınıf koşullu olasılığını tahmin eder ve şu özelliklere sahiptirler [1]:

- Saf Bayes sınıflandırıcıları izole edilmiş gürültü noktalarına karşı gürbüzdür. Çünkü, bu gibi noktalar verilerden koşullu olasılıklar tahmin edildiğinde ortalama dışında kalır. Ayrıca saf Bayes sınıflandırıcıları, model kurma ve sınıflandırma aşamalarında eksik değerlere sahip örnekleri ihmal ederek bu gibi durumların üstesinden gelirler[8].
- İlgisiz özelliklere karşı gürbüzlerdir. Eğer X_i ilgisiz bir özellik ise $P(X_i|Y)$ hemen hemen normal dağılır. X_i'nin sınıf koşullu olasılığı ardıl olasılıkların toplamı üzerinde hiç bir etkiye sahip olmaz [8].
- Koşullu bağımsızlık varsayımı ilişkili özellikler için sağlanmadığından, ilişkili özellikler saf Bayes sınıflandırıcılarının performansını düşürebilir[8].

Bayes Güven Ağları: Özellikleri bir şekilde ilişkili olan sınıflandırma problemleri için, saf Bayes sınıflandırıcıları tarafından yapılan koşullu bağımsızlık varsayımı çok katı gibi görünebilir. Bayes güven ağları, sınıf koşullu olasılıklarını modellemek için daha esnek bir yaklaşım sunar. Bu yaklaşım, verilen sınıftaki tüm özelliklerin koşullu olarak bağımsız olması yerine, hangi nitelik çiftlerinin koşullu olarak bağımsız olduğunu belirtmemizi mümkün kılar [1].

Bayes güven ağları genel olarak, izleyen özelliklere sahiptir [1]:

- Bayes güven ağları, grafiksel bir model kullanarak, belirli bir tanım kümesinin önsel bilgisini yakalayan bir yaklaşım sağlar. Ağ aynı zamanda, değişkenler arasındaki nedensel bağımsızlıkları kodlamak için de kullanılabilir[8].

- Ağı kurmak hem zaman alıcı hem de fazla çaba gerektiren bir iştir. Ancak, ağın yapısı bir kere belirlendikten sonra yeni bir değişken eklemek oldukça kolaydır[8].

- Bayes ağları, eksik veri ile uğraşmak için oldukça uygundur. Tüm özellik değerlerinin olasılıklarının toplanması veya birleştirilmesi sayesinde eksik özelliklere sahip örnekler ile başa çıkılır[8].

- Veriler önsel bilgi ile olasılıklı olarak birleştirildiği için, bu yöntem modelin aşırı uyumuna karşı oldukça gürbüzdür[8].

2.5.3 En yakın komşu

En yaygın algoritmalardan birisidir. Sınıflandırma yapılırken veritabanındaki her bir kayıdın diğer kayıtlarla olan uzaklığı hesaplanır. Ancak, bir kayıt için diğer kayıtlardan sadece k adedi gözönüne alınır. Algoritmanın isminden de anlaşılabileceği gibi bu k adet kayıt, başka bir deyişle veritabanındaki nokta, mesafesi hesaplanan noktaya diğer kayıtlara nazaran en yakın olan kayıtlardır. Bu yöntem coğrafi bilgi sistemilerinde çok kullanılan yöntemlerderndir. Belirlenen bir noktaya en yakın şehir, istasyon vs belirlenmesi aslında $k - en$ yakın komşu algoritmasının temelini oluşturur[6]

Algoritmada k değeri başlangıçte belirlenir. K değerinin yüksek olması birbirine benzemeyen noktaların bir araya toplanmasına sebep olabilir. Çok küçük seçilmesiyse birbirine benzemesine rağmen bazı noktaların ayrı sınıflara konmasına ya da o tür noktalar için ayrı sınıfların açılmasına neden olabilir. Tipik k değeri 3,5 ve 7'dir[6].

En yakın komşu sınıflandırıcısının özellikleri ise şu şekildedir[6].

- En yakın komşu sınıflandırması, örnek temelli öğrenme olarak bilenen çok genel bir tekniktir. Yani, verilerden elde edilen bir çıkarsama yapmaksızın tahmin

yapmak için özel örnekleri kullanır. Örnek temelli öğrenme algoritmalar genel olarak örnekler arasındaki uzaklığı ya da benzerliği belirlemek için bir yakınlık ölçüsüne ve diğer örneklere yakınlığına dayanan bir sınıflandırma fonksiyonuna gereksinim duyar.

- En yakın komşu sınıflandırıcıları model kurmaya gerek duymazlar. Fakat, eğitim örnekleri arasındaki yakınlık değerlerini ayrı ayrı hesaplamak zorunda olduğundan oldukça maliyetlidir. Buna karşın, model kuran sınıflandırma teknikleri için, model bir kere kurulduktan sonra, veri kümesini sınıflandırmak son derece hızlıdır.

- Karar ağacı ve kural temelli sınıflandırıcılar tüm girdi uzayına uyan bütünsel bir model bulmaya çalışırken, en yakın komşu sınıflandırıcıları tahminlerini yerel bilgilere dayanarak yaparlar. Sınıflandırma kararları yerel olarak yapıldığından, küçük k değerine sahip en yakın komşu sınıflandırıcıları göz önüne alınmayan etkenlere karşı oldukça hassastırlar.

- En yakın komşu sınıflandırıcıları keyfi olarak şekillendirilmiş karar sınırları üretebilir. Bu tarz sınırlar, çoğunlukla doğrusal karar sınırlarına kısıtlanmış olan karar ağacı ve kural temelli sınıflandırıcılarla karşılaştırıldığında, daha esnek bir model gösterimi sağlarlar. En yakın komşu sınıflandırıcılarının karar sınırları eğitim örneklerinin bileşimine dayandığı için, aynı zamanda yüksek değişkenliğe de sahiptir. En yakın komşu sayısının artmasıyla bu değişkenlik azalabilir.

- En yakın komşu sınıflandırıcıları, yaklaşık yakınlık ölçüsü ve veri ön işleme adımları gerçekleşmez ise yanlış tahminler üretebilir[1].

2.5.4 Karar ağaçları

Verilerin içerdiği ortak özellikleri kullanılarak söz konusu verileri sınıflandırmak mümkündür. Sınıflandırma bir öğrenme algoritmasına dayanır. Tüm veriler kullanılarak eğitme işi yapılmaz. Bu veri topluluğuna ait bir örnek veri üzerinde gerçekleştirilir. Öğrenmenin amacı bir sınıflandırma modelinin yaratılmasıdır. Bir başka deyişle sınıflandırma, hangi sınıfa ait olduğu bilinmeyen bir kayıt için bir sınıf belirleme sürecidir. Verileri sınıflandırma yöntemlerinden biri karar ağaçları ile sınıflan-

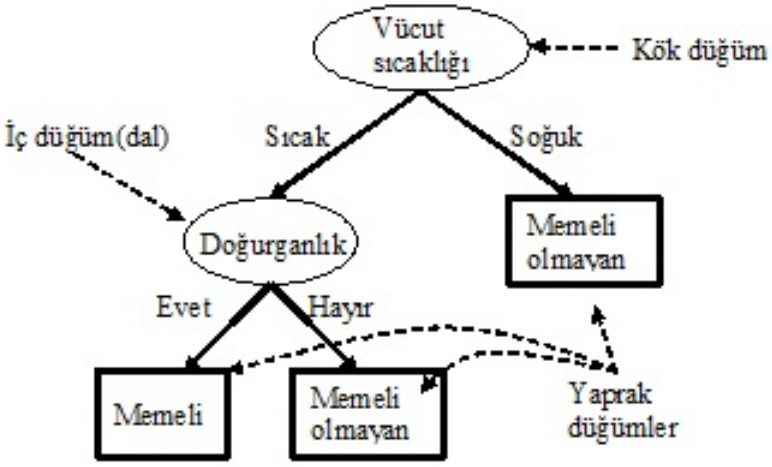
dırma adını taşımaktadır. Denetimli(supervised) öğrenme için karar ağaçları yaygın kullanılan bir yapıdır. Sınıflandırma problemlerinde karar ağacı oluşturma, makine öğrenme ve istatiksel alanlarında kullanımı oldukça fazladır. Diğer yöntemlere göre yapılandırılması ve uygulanması daha kolaydır denilebilir. Bu teknikte sınıflandırma için bir ağaç oluşturulur; daha sonra veritabanındaki her bir kayıt bu ağaca uygulanır ve çıkan sonuca göre de bu kayıt sınıflandırılır. Temel olarak karar ağaçları, ağacın kurulması ve verilerin teker teker ağaca uygulanarak sınıflandırılması olarak iki adımdan oluşmaktadır.

Karar ağaçları akış şemasına benzer bir yapıdırlar. Her bir nitelik bir düğüm tarafından temsil edilir. Dallar ve yapraklar ağaç yapısının elemanlarıdır. En son yapı "yaprak düğüm", en üst yapı "kök düğüm" ve bunların arasında kalan yapıda "dal(iç düğüm) " olarak adlandırılır.

Bir karar ağacı ile sınıflandırmanın nasıl gerçekleştiği omurgalı hayvanların sınıflandırılması problemi ile izleyen şekilde açıklanabilir. Omurgalılar, beş kesin tür grubunda sınıflandırılmak için memeliler ve memeli olmayanlar olmak üzere iki kategoride ele alınsın. Karar ağacı için yapısı için yapılması gereken keşfedilen yeni bir türün memeli olup olmadığını belirlemek için türün özellikleri hakkında sorular sormaktır. Sorulabilecek ilk soru, türün sıcakkanlı mı yoksa soğukkanlı mı olduğudur. Türün soğukkanlı olması durumunda kesinlikle bir memeli olmadığı söylenir. Aksi halde yeni tür, ya bir kuş ya da bir memelidir. Sorulması gereken bir sonraki soru da bu yeni türün dişilerinin doğurganlık özelliğinin olup olmadığıdır. Doğurganlık özelliği varsa, bu tür kesinlikle memelidir. Eğer bu özelliğe sahip değilse, muhtemelen memeli olmayan bir türdür [1].

Verilen bu örnek, sınam kayıtlarının özellikleri hakkında dikkatlice hazırlanmış bir dizi soru sorarak, bir sınıflandırma probleminin nasıl çözüleceğini göstermektedir. Kaydın sınıf etiketi hakkında bir sonuca erişene kadar, her bir cevabın ardından takip eden sorular sorulur. Sorular ve olası cevapları serisi, düğümler ve yönlü ayrıtlardan oluşan hiyerarşik bir yapı olan bir karar ağacı formunda düzenlenebilir. Memeli sınıflandırma probleminin karar ağacı Şekil 2.1 ile verilmiştir [1].

Bir karar ağacında, her yaprak düğüme bir sınıf etiketi atanmıştır. Kök düğümü ya da iç düğüm olan ve yaprak olmayan tüm düğümler, farklı özelliklere sahip kayıt-

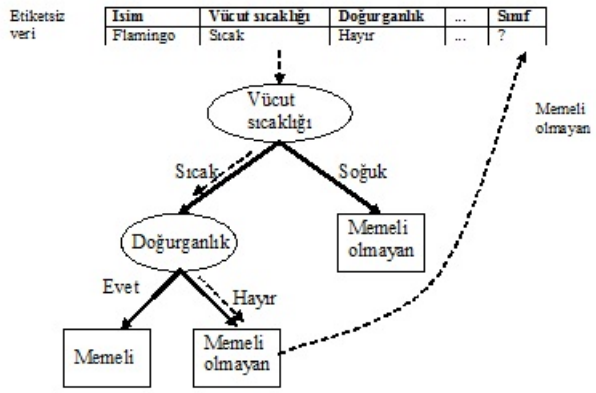


Şekil 2.1: Memelileri sınıflandırma problemi için bir karar ağacı [1]

ları ayırabilmek için öznitelik test koşulları içerirler. Örneğin, Şekil 2.1'de gösterilen kök düğümü, sıcakkanlı ve soğukkanlı omurgalıları ayırabilmek için vücut sıcaklığı özniteliğini kullanır. Bütün soğukkanlı omurgalılar, memeli olmayanlar sınıfından olduğundan dolayı, kök düğümün sağ çocuğu olarak memeli olmayanlar etiketli yaprak düğümü oluşturulmuştur. Eğer omurgalı bir sıcakkanlı ise, bir sonraki öznitelik olarak doğurganlık, memelileri sıklıkla kuşlar sınıfından olabilecek olan diğer sıcakkanlı canlılardan ayırmak için kullanılmıştır.

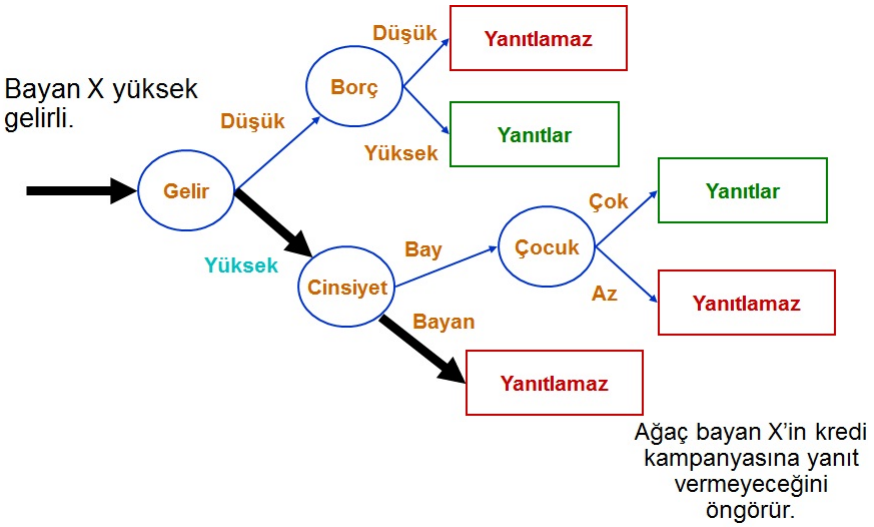
Karar ağacı çizildikten sonra bir test kaydını sınıflandırmak oldukça nettir. Kök düğümden başlayarak kayıda test koşulunu uygularız ve her sonuç için ona ait uygun iç düğümü(dal) takip ederiz. Bu bizi ya yeni test koşulunun uygulanacağı başka bir iç düğüme, ya da bir yaprak düğüme ulaştırır. Şekil 2.2'de flamingonun sınıf etiketini bulmak için kullanılan karar ağacındaki yolu göstermektedir. Bu yol, memeli olmayanlar olarak etiketlenen yaprak düğümünde son bulacaktır.

Karar ağaçları ile ilgili olarak bir diğer örnek ise kredi kampanyasında yeni bir müracaatın sınıflandırılmasını dikkate alabiliriz. Kredi başvurusu sonucu olumlu Yanıtlamaz/Yanıtlar olarak iki farklı kategoride dikkate alınsın. Yeni başvuruda sorulacak olan ilk soru gelir düzeyi düşük mü yüksek mi olduğudur. Gelirin düşük olması durumunda sorulacak olan soru ise borç düzeyi düşük mü yüksek mi olduğudur. Eğer başvuru sahibinin borcu yüksek ise bu başvuru olumlu olarak yanıtlanmaz.



Şekil 2.2: Etiketsiz bir omurgalının sınıflandırılması [1]

Benzer bir şekilde gelir düzeyi yüksek olan bir başvuru sahibi için bir sonraki soru cinsiyet olur. Cinsiyeti Bayan olan bir başvuru sahibi için kampanyaya yanıt vermeyeceğini öngörür. (Şekil2.3)



Şekil 2.3: Kredi kampanyasında başvuru sonucunun sınıflandırılması

Kural olarak, verilen bir öznitelikler kümesinden birçok farklı karar ağacı oluşturmak mümkündür. Bazı karar ağaçları diğerlerinden daha doğru sonuç vermesine rağmen, en iyi (optimal) ağacı bulmaya çalışmak, arama alanının büyüklüğünün üstel olması nedeniyle mantıklı olmayacaktır. Yine de, her ne kadar optimal olmasa da tamlık derecesi makul bir karar ağacını, uygun zaman süresinde sağlayabilecek etkin algoritmalar geliştirilmiştir. Bu algoritmalar genellikle, veriyi bölmek için hangi

özniteliğin kullanılması gerektiği ile ilgili bir dizi yerel optimal karar alarak bir karar ağacı geliştiren, açgözlü (greedy) strateji kullanırlar.

Karar ağaçları oluşturabilmek için belirli bir yol izlenir. Öncelikle veri arasından bir kısmı seçilerek eğitme işi yerine getirilir. Yani karar ağacının, belirli bir örneğe göre, yani eğitim kümesindeki veriye göre oluşturulması söz konusudur. Karar ağacı oluşturuldaktan sonra bu ağaçtan karar kuralları türetilir ve test verisi üzerinde denenir. Olumlu sonuç elde edilirse yeni gözlemleri sınıflandırmak için bazı kurallar kullanılır.

Karar ağaçları oluşturmak için bir çok yöntem geliştirilmiştir. Bunlar temel olarak Entropiye dayalı algoritmalar, sınıflandıma ve regresyon ağaçları, bellek tabanlı modelleri biçimindedir. Entropiye dayalı yöntemler arasında ID3 ve C4.5 algoritmaları sayılabilir [12].

2.5.5 Yapay Sinir Ağları

Beynin üstün özellikleri, bilim adamlarını üzerinde çalışmaya zorlamış ve beynin nörofiziksel yapısından esinlenerek matematiksel modeli çıkarılmaya çalışılmıştır. Beynin bütün davranışlarını tam olarak modelleyebilmek için fiziksel bileşenlerinin dogru olarak modellenmesi gerektiği düşüncesi ile çeşitli yapay hücre ve ağ modelleri geliştirilmiştir. Böylece Yapay Sinir Ağları denen yeni ve günümüz bilgisayarlarının algoritmik hesaplama yönteminden farklı bir bilim alanı ortaya çıkmıştır. Yapay sinir ağları; yapısı, bilgi işleme yöntemindeki farklılık ve uygulama alanları nedeniyle çeşitli bilim dallarının da kapsam alanına girmektedir [13].

Sinir hücreleri bir grup halinde işlev gördüklerinde ağ olarak adlandırılırlar ve böyle bir grupta binlerce nöron bulunur. Nöronların aynı doğrultu üzerinde bir araya gelmeleriyle katmanlar oluşmaktadır. Bu katmanların bir araya gelmeleri yapay sinir ağını ve dolayısıyla yapay sinir ağı modelini oluşturmaktadır[13].

- Girdi Katmanı : Bu katmandaki proses elemanları dış dünyadan bilgileri alarak ara katmanlara transfer ederler. Bazı ağlarda girdi katmanında herhangi bir bilgi işleme olmaz.

- Ara Katman (Gizli Katman) : Girdi katmanından gelen bilgiler işlenerek çıktı katmanına gönderilirler. Bu bilgilerin islenmesi ara katmanlarda gerçeklestirilir. Bir ağ içinde birden fazla ara katman olabilir.

- Çıktı Katmanı : Bu katmandaki proses elemanları ara katmandan gelen bilgileri işleyerek ağın girdi katmanından sunulan girdi seti için üretmesi gereken çıktıyı üretirler. Üretilen çıktı dış dünyaya gönderilir[13].

Yapay sinir ağları, yapılarına göre ileri beslemeli ve geri beslemeli ağlar olmak üzere iki şekilde sınıflandırılır. İleri beslemeli bir yapay sinir ağı, birden fazla katmandan oluşan bir ağdır. Her bir katmanda en az bir nöron vardır. Katmanlardan birisi girdi katman, birisi çıktı katman ve diğer katman veya katmanlar gizli katmanlar olarak adlandırılır. Geri beslemeli yapay sinir ağları da, ileri beslemeli yapay sinir ağlarındakinin tersine durağan hale ulaşıncaya kadar çevrimler devam eder. Ayrıca bütün nöronlar birbiriyle bağlantılıdır. Yapay sinir ağları, öğrenme algoritmalarına göre ise denetimli ve denetimsiz olarak iki farklı şekilde sınıflandırılır[14].

2.5.6 Destek vektör makineleri

Veri madenciliğinde sınıflandırma problemlerinde kullanılan bir diğer yöntem ise destek vektör makineleri adını taşımaktadır. Destek vektör makinesi yöntemi, veriyi birbirinden ayırmak için en uygun fonksiyonun tahmin edilmesi esasına dayanır. Bu yöntem, sınıflandırmayı, doğrusal ve doğrusal olmayan bir fonksiyon yardımıyla gerçekleştirilir. Daha çok makine öğrenmesi yöntemleri arasında yer alan bu yöntem günümüzde veri madenciliğinde sık bir şekilde kullanılmaktadır [12].

Destek vektör makineleri doğrusal olarak ayrılabilir ve ayrılamayan durumlar olarak ikiye ayrılmaktadır.

2.5.6.1 Doğrusal ayrılabilir durum

Verileri birbirinden ayırmak için bir çok hiperdüzlem bulunabilir. Bu hiperdüzlemlerden en iyi ayırıcı hiperdüzlem, genelleştirme başarımı eniyi olan hiperdüzlemdir. Sınıflandırma problemlerinde örnekleri hiperdüzlemin doğru tarafında sınıflandırmanın yanında, daha iyi genelleştirme için örneklerin hiperdüzlemden belli bir mesafe

uzaklıkta olması da istenir. En iyi genelleştirme için en büyüklenmeye çalışılan bu uzaklık, marjin olarak adlandırılmaktadır [8].

2.5.6.2 Doğrusal olarak ayrılamayabilir durum

Veriler doğrusal olarak ayrılamıyor ise daha önce sunulan DVM yaklaşımı geçerli olmaz. Böyle bir durumda eğer iki sınıf doğrusal ayrılabilir değil ise onları tam olarak ayıracak bir hiperdüzlem yoktur. Bu durumda marjinden sapmayı ifade eden bir aylak değişkeni tariflenir. Burada odaklanan iki farklı sapma türü vardır. Bir örnek hiperdüzlemin yanlış tarafında yer alabilir ve yanlış sınıflandırılabilir veya örnek hiperdüzlemin doğru tarafında yer alır ancak aralık içinde kalabilir. Bu durumda hiperdüzlemden yeterince uzak değildir. Bu sebeple enküçük hatayı veren hiper düzlem aranır [8].

2.5.7 Diğer sınıflandırma yöntemleri

Veri madenciliğinde sınıflandırma yöntemleri kadar çok sık kullanılmayan, ancak rağbet gören diğer yöntemler ise; olay temelli çıkarsama, genetik algoritmalar, kaba kümeler ve bulanık küme yaklaşımlarıdır.

2.6 Matematiksel Programlama Yöntemleri

Sınıflandırma problemlerinin çözümünde 1960'lardan bu yana çok çeşitli yaklaşımlar geliştirilmiş ve kullanılmıştır. Matematiksel programlama temelli yöntemler de, sınıflandırma problemlerinin çözümü için sıkça kullanılan yaklaşımlardandır. Bu yaklaşımlar temel olarak R^n'de A ve B gibi belirli sayıda noktaya sahip iki ayrık kümenin ayrılması amacıyla geliştirilmiştir[15]. Mangasarian iki kümeyi ayıracak olan doğrusal ve doğrusal olmayan düzlemler oluşturmak için doğrusal programlama yaklaşımı kullanılmıştır. Daha sonraki yıllarda Bennet ve Mangasarian doğrusal ayırma için gürbüz(robust) bir yaklaşım geliştirmişlerdir[16]. Astorino ve Gaudioso [17] ise doğrusal programlama ile belirlenen sayıda hiper düzlem oluşturarak A ve B kümelerini ayırmaya çalışmışlardır. Bir diğer yaklaşım ise Bagirov [18] tarafından sunulan enb-enk ayırma yaklaşımıdır. Erenguc ve Koehler [19] ise matematik-

sel program kullanımı ile yayınladıkları makalede literatürde amaç fonksiyonları ve kullanılan tekniklere göre farklılaşan 22 matematiksel modeli incelemişlerdir. Amaç fonksiyonlarının farklılaşmasının, yanlış sınıflandırılan nokta sayısı, küme dışı sapma, küme içi sapma, küme içi ve dışı toplam sapmanın enküçüklenmesi biçimde olduğunu göstermişlerdir.

Sınıflandırma problemlerinin çözümü için kullanılan matematiksel program yaklaşımlarından bir diğeri de karma tamsayılı programlama mantığıdır. Glen [20, 21] çok çeşitli amaç fonksiyonları içeren karma tamsayılı yaklaşımları önermiş ve doğrusal ayırma analizinde önemsiz çözümden kaçınacak, sınıflandırma doğruluk oranını artıracak amaçlar üzerine durmuştur. Glen geliştirdiği bu yöntemleri finansal oranlara göre şirketlerin durumlarının tespiti ve kredi başvuru sonuçlarının tespitinde kullanmıştır. Benzer bir şekilde Üney ve Türkay [22]'da çok sınıflı problemlerin çözümü için tüm sınıflara ait örnekleri ayıracak çok boyutlu kutuların kullanılmasına dayalı karma tamsayılı matematiksel bir model geliştirilmişlerdir.

Matematiksel programların sınıflandırma problemlerinde kullanımına ait bir çok uygulama vardır. Bunlar içinde Bennet ve Mangasarian [16] tarafından sunulan gürbüz yaklaşımın önemli bir yeri vardır. Bu makalede herhangi iki kümeyi ayırmak için hata fonksiyonunu enküçükleyerek bir hiper düzlemin bulunabileceğini gösterilmiştir. Bu çalışma daha sonra yapılan bir çok çalışmaya temel olmuştur.

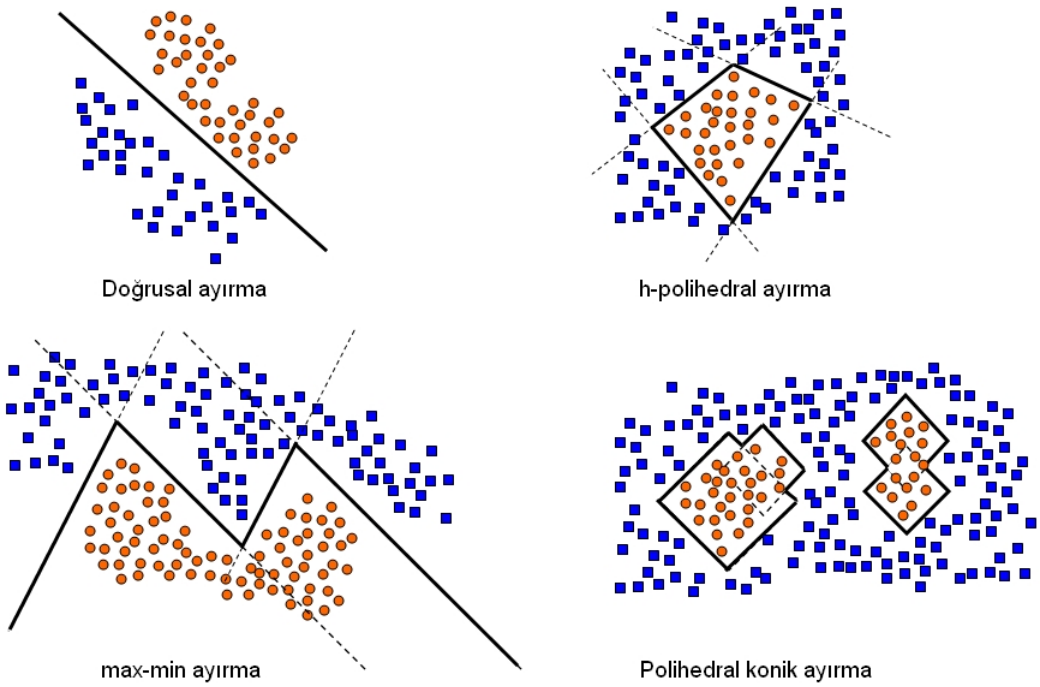
Sonuç olarak sınıflandırma probleminin çözümü için en çok kullanılan matematiksel programlama yaklaşımları doğrusal ayırma, h-çok yüzlü ayırma, enb-enk ayırma ve bütünsel ağaç enyilemedir.

2.7 Çokyüzlü Konik Fonksiyonlar ile Sınıflandırma

Sınıflandırma problemi, sonlu sayıda noktadan oluşan ayrık iki kümenin ayrılması problemi olarak tanımlanmaktadır. Veri kümelerinin tamamı dış bükey bir yapıya sahip ise doğrusal, herhangi bir tanesi dış bükey bir yapıya sahip ise h-çokyüzlü ayırma ile tam olarak ayrılabilmektedir. Ancak bu iki yaklaşım da dışbükey olmayan ayırıcı yüzeyler oluşturamamaktadır. $Enb-enk$ ayırma, belirli sayıda hiperdüzlemin alt kümelerini kullanarak dışbükey olmayan ayırıcı yüzeyler oluşturmaktadır. Gasi-

mov ve Öztürk [23] tarafından geliştirilen Çokyüzlü konik ayırma ile, dışbükey olmayan ayırıcı yüzeyler oluşturulabilmekte ve bunun yanında $enb - enk$ yaklaşımından farklı olarak, birden çok dışbükey olmayan kümenin diğer kümeden tam olarak ayrılması da sağlanabilmektedir.

Gasimov ve Öztürk, bu yaklaşımların geometrik gösterimlerini şekil 2.4'de gösterilmiştir.



Şekil 2.4: Bazı ayırma yaklaşımlarının grafiksel görünümü[2]

Çokyüzlü konik fonksiyonlar(ÇKF) temeline dayalı sınıflandırma, sonlu sayıda ardışık adım ile, her adımda A kümesinin bir kısmını B kümesinden ayıran bir ÇKF oluşturarak gerçekleştirilmektedir. ÇKF'ler ile iki kümenin ayrılması, A kümesine ait mümkün olduğunca çok noktanın ÇKF ile oluşturulan dışbükey polihedronun içinde, tüm B kümesine ait noktaların ise bu polihedronun dışında kalmasını sağlayarak gerçekleştirmektedir. ÇKF algoritması her ardıştırmada en çok noktayı ayırmayı hedeflediğinden ve o an verilebilecek en iyi kararı aradığından aç gözlü bir bir yaklaşım olarak nitelendirilmiştir[8].

2.8 Sonuç Karşılaştırma Yöntemleri

Genel olarak literatürde geliştirilen teknik ve yaklaşımlardan elde edilen modellerin performanslarının karşılaştırılması için çeşitli yaklaşımlar geliştirilmiştir. Bu bölümde, bir sınıflandırıcının performansını değerlendirmede yaygın olarak kullanılan farklı yöntemler hakkında kısa bilgiler verilmiştir.

Denetimli sınıflandırma problemlerinin çözümü için geliştirilen yaklaşımların başarımının ölçümü için farklı yöntemler bulunmaktadır. Bu yöntemler de, veriler sınam ve eğitim kümeleri olarak ikiye ayrılır. Eğitim kümesindeki veriler ile geliştirilen yöntem uygulanır. Daha sonra, elde edilen kurallar veya model sınam kümesindeki veri grubu üzerinde denenir. Sonuçta, bir ikili sınıflandırma problemi için Çizelge 2.1 ile gösterildiği gibi, kaç adet hatalı ve doğru sınıflandırılmış nokta olduğunu gösteren bir çizelge elde edilir. Bu çizelgedeki sonuçlar yardımıyla da tekniğin başarımı yüzde olarak belirlenmiş olur [8].

Çizelge 2.1: Hatalı sınıflandırma matrisi

		Tahmin edilin sınıf	
		A	B
Gerçek	A	Doğru (D_1)	Yanlış(Y_1)
Sınıf	B	Yanlış (Y_2)	Doğru (D_2)

Bir eğitim bir sınam kümesi: Bir eğitim bir sınam kümesi yönteminde etiketlenmiş örnekleri olan orijinal veriler, eğitim ve sınam kümeleri olmak üzere iki ayrık kümeye bölünür. Ardından, eğitim kümesinden bir sınıflandırma modeli elde edilir ve modelin performansı sınam kümesi üzerinde değerlendirilir. Eğitim ve sınam için ayrılmış olan bir kısım veri, genellikle analizcinin inisiyatifinde belirlenir[8].

Rassal alt örnekleme: Bir eğitim bir sınam kümesi yönteminin bir sınıflandırıcının performansının tahminini güçlendirmek için birkaç defa tekrarlandığı yaklaşım, rassal alt örnekleme olarak adlandırılır[8].

***K*-kez çarpraz doğrulama:** K-kez çarpraz doğrulama yöntemi, veri grubunun eşit sayıda örnek içeren k adet parçaya bölünerek genelleştirilir. Daha sonrasında her sınamda $k-1$ parça eğitim kümesi, k. parça ise sınam kümesi olarak kabul

edilir. Eğitim kümesine uygulanan yaklaşım ile elde edilen model sınam kümesine uygulanır. Bu işlem k defa gerçekleştirilir. Toplam hata ise k aşamadaki hataların toplamıdır. Başarı oranı ise doğru olarak sınıflandırılan nokta sayısının veri kümesindeki örnek sayısına bölünerek bulunur. Genel olarak k değeri 10 kabul edilir[8].

Biri dışarıda kalsın: Biri dışarıda kalsın yöntemi k-kez çapraz doğrulamanın özel bir halidir. k'nın kümelerdeki toplam örnek sayısına eşit olduğu bir durumdur. Bütün verileri kullanabildiği için modellerin elde edilmesinde oldukça avantaj sağlar. Buna karşın veri kümesinin büyüklüğü kadar bir tekrar sözkonusu olduğu için oldukça maliyetlidir[8].

Sınıflandırma probleminin çözümü için oluşturulan modelin başarısını değerlendirirken kullanılan temel kavramlar ise doğruluk/hata oranı, duyarlılık, kesinlik ve F-ölçütüdür. Modelin başarısı belirlenirken kullanılan Çizelge 2.1'de gösterilen parametreler ile hesaplanır. Buna göre;

Doğruluk Oranı: Modelin başarısını ölçmek için kullanılan en popüler, basit ve belirleyici ölçüt doğruluk oranıdır.

$$Dogruluk = \frac{D_1 + D_2}{D_1 + D_2 + Y_1 + Y_2} \tag{2.2}$$

Kesinlik: Kesinlik, doğru olarak tahminlenmiş olan doğru sayısının, doğru olarak tahminlenen tüm örnek sayısına oranıdır.

$$Kesinlik = \frac{D_1}{D_1 + Y_2} \tag{2.3}$$

Duyarlılık: Doğru sınıflandırılmış pozitif örnek sayısının toplam pozitif örnek sayısına oranıdır. Doğruluk ile kesinlik birbiri ile ters orantılıdır.

$$Duyarlilik = \frac{D_1}{D_1 + Y_1} \tag{2.4}$$

F-ölçütü: Kesinlik ve duyarlılık ölçütü tek başına anlamlı bir karşılaştırma sonucu çıkarmamıza yeterli değildir. Her iki ölçütü beraber değerlendirmek daha doğru

sonuçlar verir. Bunun için F-ölçütü tanımlanmıştır. F-ölçütü, kesinlik ve duyarlılığın harmonik ortalamasıdır.

$$F - Olcut = \frac{2 * Duyarlilik * Kesinlik}{Duyarlilik + Kesinlik} \tag{2.5}$$

Bu çalışmada modelin başarısını değerlendirmek için doğruluk oranı kullanılmıştır.

3. BÜYÜK BOYUTLU SINIFLANDIRMA PROBLEMLERİNİN ÇÖZÜMÜ İÇİN YENİ BİR YAKLAŞIM

Bu bölümde büyük boyutlu ve çoklu sınflandırma problemlerinin çözümü için geliştirilen matematiksel program temelli yaklaşım açıklanmıştır. Geliştirilen bu yeni yaklaşım için temel olarak ÇKF'ler kullanılmaktadır. Bu yeni yaklaşımın ilk kısmında ÇKF ve ÇKF algoritması anlatılmıştır. Devam eden kısımda ÇKF'ların merkez noktalarının belirlenmesi için k-ortalamalar, ÇKF parametreleri içinse gürbüz doğrusal programlama yaklaşımı hakkında bilgiler verilmiştir. Son kısımda ise yeni geliştirilen yaklaşım, bu yaklaşımın açıklayıcı bir örnek üzerinde uygulaması ve literatürdeki veri kümeleri üzerindeki uygulamaları yapılmıştır.

3.1 Çokyüzlü Konik Fonksiyonlar

Özturk'ün [8] önerdiği iki ve çok sınıflı yaklaşımların temelini oluşturan çokyüzlü ayırma fonksiyonu $g_{(w,\xi,\gamma,a)} : R^n \to R$ Denklem 3.1 ile tanımlanmaktadır.

$$g_{(w,\xi,\gamma,a)}(x) = w(x-a) + \xi \left\|x-a\right\|_1 - \gamma, \tag{3.1}$$

Burada $w \in R^n, \quad \xi \in R_+ = [0,+\infty), \quad \gamma \geq 1, \quad wx = w_1x_1 + \cdots + w_nx_n$ ifadesi w ile x vektörlerinin skaler çarpımı ve $\|x\|_1 = |x_1| + \cdots + |x_n|$ ise x vektörünün 1 normudur. [8]

Denklem 3.1 ile tanımlanan $g_{(w,\xi,\gamma,a)}$ fonksiyonunun temel özelliği tepe noktası olan çok yüzlü bir konidir. $g_{(w,\xi,\gamma,a)}$ fonksiyonunun tepe noktası $(a,-\gamma)$'dır. Bu fonksiyonların seviye kümesi bir dış bükey polihedrondur.

Sınıflandırma probleminin çözümü için Gasimov ve Öztürk tarafından önerilen ÇKF algoritmasının [23] her adımında Denklem 3.1'de verildiği şekilde bir fonksiyon, belirli bir doğrusal programlama probleminin çözümü olarak w, ξ ve γ parametrelerinin bulunmasıyla elde edilmektedir. İlk aşamada fonksiyonların oluşması için gerekli olan tepe noktası $\mathcal{A}$ kümesinden rassal olarak seçilir. Bu fonksiyonun bir

alt seviye kümesi olarak tanımlanan dışbükey polihedron tüm uzayı, $\mathcal{B}$ kümesinin tüm elemanları bu polihedronun "dışında" mümkün olduğu kadar çok $\mathcal{A}$ kümesi noktasının da "içinde" olacak şekilde, ikiye böler. Algoritma, polihedronun içinde kalan bu noktaları $\mathcal{A}$ kümesinden çıkarak, sonraki ardıştırmaya geçer ve boş küme elde edilinceye kadar her ardıştırmada $\mathcal{A}$ kümesinin kalan kısmını kullanarak yeni bir polihedron üretir. Nihai ayırma fonksiyonu türetilen tüm bu ayırma fonksiyonlarının noktasal enküçüğü olarak hesaplanır[8].

3.1.1 ÇKF Algoritması

ÇKF algoritması Öztürk'ün [8] çalışmasında şu şekilde anlatılmaktadır.

$\mathcal{A}$ ve $\mathcal{B}$ kümeleri R^n'de verilmiş iki küme olsun:

$$A = \{a^i \in R^n : i \in I\}, \qquad I = \{1, \ldots, m\},$$
$$B = \{b^j \in R^n : j \in J\}, \qquad J = \{1, \ldots, p\},$$

Bu durumda ÇKF algoritması izleyen şekilde ifade edilir.

ÇKF Algoritması

Adım 0: $l = 1, I_l = I, A_l = A$ atamalarını yap, Adım 1'e git.

Adım 1: a^l noktası, A_l kümesinin herhangi bir noktası olsun. P_l problemini çöz.

$$w\left(a^i - a^l\right) + \xi \left\|a^i - a^l\right\|_1 - \gamma + 1 \leq y_i, \qquad \forall i \in I_l, \tag{3.2}$$

$$-w\left(b^j - a^l\right) - \xi \left\|b^j - a^l\right\|_1 + \gamma + 1 \leq 0, \qquad \forall j \in J, \tag{3.3}$$

$$y = (y_1, \ldots, y_m) \in R^m_+, w \in R^n, \xi \in R, \gamma \geq 1 \tag{3.4}$$

kısıtları altında

$$(P_l) \qquad \text{enk } \left(\frac{ye_m}{m}\right) \tag{3.5}$$

(P_l) probleminin bir çözümünü $w^l, \xi^l, \gamma^l, y^l$ bul. Bu çözüme karşı gelen ÇKF'yi Denklem 3.6 ile gösterildiği şekilde oluştur ve Adım 2'ye geç.

$$g_l(x) = g_{(w^l,\xi^l,\gamma^l,a^l)}(x) \tag{3.6}$$

Adım 2: $I_{l+1} = \{i \in I_l : g_l(a^i) + 1 > 0\}, A_{l+1} = \{a^i \in A_l : i \in I_{l+1}\}$, $l = l + 1$ güncellemelerini yap. Eğer $A_l \neq \emptyset$ ise Adım 1'e git.

Adım 3: $\mathcal{A}$ ve $\mathcal{B}$ kümelerini ayıran $g(x)$ fonksiyonunu Denklem 3.7 ile tanımla ve dur.

$$g(x) = \operatorname*{enk}_l g_l(x) \tag{3.7}$$

Algoritma, her $l.$ adımda, bir a_l noktası seçer ve (P_l) doğrusal alt problemini çözerek (w^l, ξ^l, γ^l) parametrelerini hesaplar. Bu parametrelerin tümü ile Denklem 3.6 ile verilen g_l fonksiyonu tanımlanır. g_l fonksiyonun grafiğinin $z = g_l(x)$ olmak üzere (x, z) noktalarından oluştuğu ve tepe noktası ise $(a^l, -\gamma^l)$ noktasıdır. Denklem 3.4 ile verilen kısıt kümesinde bulunan $\gamma \geq 1$ kısıtı bu koninin tepe noktasının $z = 0$ hiperdüzleminin "alt" bölgesinde yerleşmesini sağlar. Denklem 3.2 ile verilen kısıt, a^l noktasının yakınında bulunan mümkün olduğunca çok sayıda A_l kümesine ait noktanın, g_l fonksiyonun seviye kümesi ile elde edilen polihedronun içine alarak $\{x : g_l(x) \leq 0\}$, $\mathcal{B}$ kümesinden ayrılmasını sağlar[8].

ÇKF Algoritması için durma kriteri $\mathcal{A}$ kümesinde ayrılmayan noktaların oluşturduğu kümenin boş küme olmasıdır. Her bir ayırma fonksiyonu için $\mathcal{A}$ kümesinin ayrılmayan bir noktasını tepe noktası olarak seçileceği için bunun bir durdurma kritesi olabileceği açıktır. Fakat burada ÇKF'ler için tepe noktaları ne kadar iyi seçilirse algoritmanın etkinliği o kadar fazla olacaktır.

3.2 K-Ortalama

İlk olarak 1967 yılında MacQuen tarafından ortaya atılan k-ortalama algoritması, sürekli olarak kümelerin yenilendiği ve en uygun çözüme ulaşana kadar devam eden döngüsel bir algoritmadır. Bölümlemeli algoritmaların tipik özelliklerini taşır. Bu

alandaki benzer algoritmaların çoğu ya k-ortalama algoritmasından esinlenerek ya da algoritmanın geliştirilmesiyle ortaya çıkmıştır. 1967 yılından bu yana bir çok k-ortalama algoritması temelli yaklaşım geliştirilmiştir [6].

3.2.1 Temel K-Ortalama Algoritması

K-Ortalama, sınıf bilgisi olmayan verilerin özelliklerine göre k sayıda sınıfa kümeleme işlemidir. Kümeleme, ilgili kümenin merkez değeri ile veri setindeki her nesnenin arasındaki farkın kareleri toplamının minimumu alınarak gerçekleştirilir. Nesnelerin sınıflandırılması işlemi gerçekleştikten sonra her bir sınıfa veya kümeye ilgili etiketin verilmesi uzman bir kişi tarafından yapılır. K-ortalamalar kümelemesinde amaç, gerçekleştirilen bölümleme işlemi sonunda elde edilen kümelerin, küme içi benzerliklerinin maksimum ve kümeler arası benzerliklerinin minimum olmasını sağlamaktır. Küme benzerliği, kümenin ağırlık merkezi olarak kabul edilen bir nesne ile kümedeki diğer nesneler arasındaki uzaklıkların ortalama değeri ile ölçülmektedir.

A noktalar kümesi $\mathbb{R}^n$ n-boyutlu uzayda tanımlı olsun.

$A = \{a^1, \ldots, a^m\}$ olmak üzere $a^i \in \mathbb{R}^n$, $i = \{1, \ldots, m\}$
A veri kümesindeki elemanların, verilen k sayısı kadar altkümeye atanması problemi hard unconstrained sınıflandırma problemi olarak düşünülmektedir.
A^j , j=1, . . . , k

(1)$A^j \neq \emptyset$ $j = 1, \ldots, k$;
(2)$A^j \cap A^l = \emptyset$ $j, l = 1, ..., k$,j$\neq$ l;
(3)$A = \bigcup_{j=1}^{j} A^j$,
(4)$A^j, j = 1, ..., k$. kısıtlara maruz kalmayan kümeler olmak üzere,

A^j, j= 1, . . . , k kümeler olarak adlandırılır. Her bir A^j kümesinin merkezinin ise $x^j \in$ $j = 1, ..., k$. olarak gösterilmektedir[24].

K-Ortalamalar Algoritması:

Adım 1: Toplam küme sayısını oluşturacak olan k değerini seç,

Adım 2: Veri kümesinin içinden rasgele olarak k adet başlangıç noktasını seç,

Adım 3: Veri kümesindeki her bir noktayı bir benzerlik yöntemine göre k. kümeye ata,

Adım 4: Herbir kümenin merkezi yeniden hesapla,

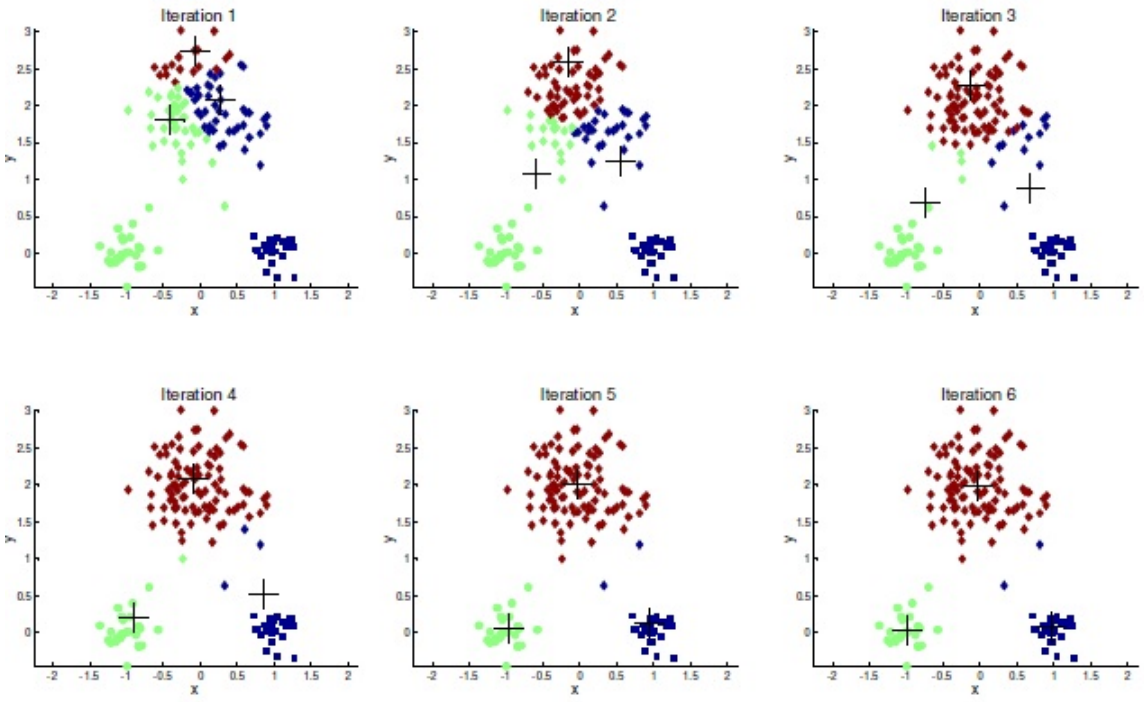
Adım 5: Eğer küme merkezi ve benzerlik değeri değişmiyor ise algoritma ***DUR***.

Diğer durumlarda yeni küme merkezlerine göre ***Adım 3***'e git.

Algoritmanın ilk basamağı data içerisinde sınıflandırılmasını düşündüğümüz küme sayısına karar verilmesiyle başlar, sonra algoritma rassal bir şekilde k adet başlangıç noktasını veri noktalarından seçer. Her örnek en çok benzer olan küme içerisine yerleştirilir. Bütün örnekler bu benzerlik yöntemine göre onlara uygun kümeye yerleştirildikten sonra küme merkezleri her bir yeni kümenin ortalaması olarak hesaplanır ve yenilenir. Örneklerin sınflandırılması süreci ile küme merkezinin hesaplanması, döngüden çıkış kriterine gerçekleşene kadar devam eder[25]. (Şekil 3.1)

K-ortalamalar sınıflandırma metodunda, kümelerin merkezleri genel olarak kümedeki noktaların ortalaması olarak alınır. Verilerin atanacakları kümelerin tespiti için kullanılan benzerlik ise Euclidean uzaklığı, cosine benzerliği, correlation vs ile hesaplanabilir. Özellikle uzaklık hesaplamada sonucundaki kümelerin belirlenmesi için genel hesap sayılarını azaltmak için kullanılan iki ana yaklaşım vardır.

- Özellikle bir kaç döngüden sonra kümelere atanan elemanların değişmediği durumda, bir daha uzaklık hesaplama ihtiyacına gerek olmaz ve tekrardan çıkılarak kümeler elde edilir.
- Her bir döngüde hata fonksiyonları hesaplanır ve döngüden çıkma kriteri olarak hata fonksiyonu değişkeni olarak kullanılır. K-ortalama algoritmasında döngüden çıkma kriteri için sabit bir değer yoktur. Bu değer kullanıcı tarafından tespit edilir [26].



Şekil 3.1: K-Ortalama ile Sınıflandırma

Burada en genel hata fonksiyonu değeri hataların kareleri toplamıdır.(Sum of Squared Error(SSE))Her bir nokta için, hata en yakın kümeye olan uzaklıktır. SSE hesabı için bu hataların karesini hesaplar sonra toplanır.

3.2.1.1 K-Ortalama ile ilgili genel düşünceler

K-Ortalamalar anlaması ve uygulaması kolay bir algoritmadır. Buna rağmen bazı özellikleri üzerine düşünülmesi gerekiyor.

- Algoritma sadece reel değerlere göre çalışır. Eğer veri kümesi kategorilerden oluşuyorsa bunları sayısal değerler çevrilmelidir[25].
- k değeri başlangıçta seçilmesi gerekiyor. Eğer başlangıçta doğru seçilmez ise model beklenen etkili çözümü elde edemez. Farklı başlangıç değerleri için farklı sonuçlar elde edileceğinden dolayı başlangıç noktalarının seçimi oldukça önemlidir. Bunun için farklı sezgisel algoritmalar geliştirilmiştir[27].
- K-Ortalamalar algoritması verilerin içindeki küme büyüklüklerinin yaklaşık

olarak eşit olduğu durumda en iyi çalışır. En iyi çözümün eşit olmayan boyuttaki kümeleri işaret ediyorsa en iyi çözümü bulamayabilir[27].

- Eğitimli kümeler için bu algoritmayı uygulamak modelin çözümünün etkinliği için eğitimsiz olan kümelere uygulamaktan daha fazla sonuç elde etmemizi sağlayacaktır[25].

Sonuç olarak k-ortalamalar algoritması kullanımı, anlaşılırlık ve maliyet açısından oldukça avantajlı bir sınıflandırma yaklaşımıdır. Buna rağmen k değerinin değişemiyor olması ve başlangıç noktalarının rassal olarak seçilmesi k-ortalamalar algoritmasının dezavantajlarıdır. Bu yüzden litaratürde sezgisel yaklaşımlar ile geliştirilmiş bir çok k-ortalamalar algoritması vardır[24, 27, 26, 28].

3.3 Gürbüz Doğrusal Programlama

Gürbüz doğrusal programlama (RLP) Bennet ve Mangasarian [16] tarafından geliştirilmiştir. RLP, n boyutlu uzayda iki ayrık nokta kümesinin yanlış sınıflandırılan noktaların toplamını minimum yapacak bir düzlem, tek bir doğrusal model ile formula edilebileceğini iddia etmektedir. RLP minimum hatayı eniyileyen bir model olarak kurulmuştur.

A ve B n boyutlu $\mathbb{R}^n$ uzayında iki ayrı küme olsun. A mxn matrisi ile B kümesi ise kxn matrisi ile temsil edilsin. RLP, aşağıdaki özellikleri sağlayan tek bir doğrusal programın formula edilebileceğini kabul eder;
i) Eğer A ve B ayrık dışbükey kümeler ise kesin bir şekilde ayırabilen bir düzlem elde edilebilir.
ii) Eğer A ve B kesişen dışbükey kümeler ise bütün olası durumlar için yanlış sınıflandırılan noktaların ölçümlerini minimize edecek bir düzlem elde edilebilir.
iii) İlgisiz kısıtlamalar olmaksızın düşünülenin dışındaki herhangi özel durumu çıkaracak doğrusal programı bulmak zordur[16].

Çoğu doğrusal program *i* özelliğine sahiptir. Buna rağmen *ii* ve *iii* özelliğine sahip olan doğrusal programlama ise oldukça zordur. RLP ise bu üç özellliğin tamamını sağlayan bir formulasyonu önermektedir. RLP genel olarak sıkıntılı olan durumlarda hataları minimize eden hiper düzlem her zaman türetebilebilir. Her iki kümesinin

hatalarının ortalamasının anlamlı olduğu durumda boş kümenin olduğu muhtemel çözüler türetilebilir. Buna rağmen bu boş çözüm bizim doğrusal programlama için eşsiz değildir ve yararlı bir çözüm her zaman bulunabilir[16].

$\mathbb{R}^n$ 'de n boyutlu reel uzayında x bir vektör olmak üzere x_+

$(x_+)_i := \max\{x_+, 0\}$, $i = 0, 1..n$ gösteren bir vektör olsun,

$A \in R^m xn$ formülü mxn gerçek matrisi, A' ise i. satırını gösterecek A_i iken transpose olarak,

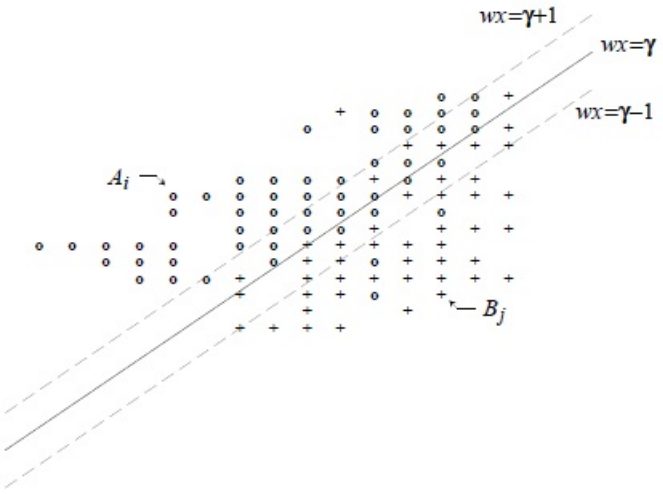
x'in 1-normunu $\sum_{i=1}^{n} |x_i|$, y'in 1-normunu $\sum_{i=1}^{n} |y_i|$,

x'in ∞-normunu ise $max_{1 \leq i \leq n} |x_i|$ $\|\mathrm{x}\|_\infty$ olarak gösterilirsin[16].

RLP'nin hataları minimize eden eniyileme modeli,

$$\min_{w,\gamma} \frac{1}{m} ||(Aw - e\gamma + e)_+||_1 + \frac{1}{m} ||(Bw + e\gamma + e)_+||_1 \tag{3.8}$$

$A \in R^{mxn}$, A kümesindeki m noktaları olmak üzere $B \in R^{kxn}$, B kümesindeki k noktaları olmak üzere w en iyi ayırma uazayında normalini sunan n boyutlu bir ağırlık vektörü ve γ ayırma uzayının lokasyonunu veren eşik değeridir. $w.x = \gamma$ (Şekil 3.2)[16].



Şekil 3.2: Doğrusal ayrılamayan $A(o)$ ve $B(o)$ için en iyi ayırma $w.x = \gamma$

Bennet ve Mangasarian [16] bu modeli hataların enküçükleneceği A ve B kümelerini ayırabilecek olan modeli 3.9'deki gibi olduğunu göstermişlerdir[16].

$$\min_{w,\gamma,y,z} \left\{ \frac{ey}{m} + \frac{ez}{k} \| Aw - e\gamma + y \geq e, -Bw + e\gamma + z \geq e, y \geq 0, z \geq 0 \right\} \tag{3.9}$$

RLP modelini incelediğimizde hatalı sınıflandırılan noktaların toplam sayısını en küçükleyen bir matematiksel modeldir. Özellikle tamamen dış bükey olmayan kümelerin birbirinden ayrılması için RLP'nin diğer doğrusal programlamalara göre en önemli avantaj sağlamaktadır. Bu nedenlerden dolayı RLP özellikle h-polihedral ayırma gibi bir çok yaklaşımın temelini oluşturmaktadır.

3.4 K-Ortalama-ÇKF-RLP Yaklaşımı

Gasimov ve Öztürk'ün önerdiği ÇKF algoritmasının ilk adımında tepe noktalarının (merkez noktaları) belirlendiği, bir sonraki adımda ise fonksyon parametrelerinin (w, ξ ve γ) bulunduğundan bahsedilmişti. ÇKF temelli sınıflandırma yaklaşımında, $\mathcal{A}$ kümesinden $\mathcal{B}$ kümesini ayıracak olan ÇKF'lerin merkez noktalarının belirlenmesinin sınıflandırma başarısı üzerinde önemli bir etkiye sahiptir. ÇKF merkezleri karar değişkeni olarak ele alınıp modellendiğinde problem dış bükey ve doğrusal olmadığından daha karmaşık hale gelmektedir. Gasimov ve Öztürk geliştirdikleri algoritma birden çok ÇKF kullanarak iki kümenin eğitim aşamasında yüzde yüz başarı ile ÇKF algoritması yardımıyla ayrılabileceği göstermişlerdir. Ancak ÇKF yaklaşımı eğitim aşamasında rekabetçi başarılar sunarken, göz önüne alınmayan diğer özelliklerinde etkisiyle test aşamasında aynı eğitim başarısı elde edilememektedir. Bu da eğitim ile test başarıları arasındaki belirgin bir farkın olmasına sebep olmaktadır. Literatürde bu tür durumlara aşırı uyum (overfitting) denilmektedir. Sınıflandırma problemlerinde aşırı uyum tercih edilen bir durum değildir. Bu sebeple, ÇKF için merkez noktaları ve doğrusal programlama ile elde edilen parametrelerin belirlenmesi oldukça kritiktir. Bu çalışmada, merkez noktaları ve fonksiyon parametreleri w, ξ ve γ üzerine odaklanılarak daha etkin ve aşırı uyum sorunu aşılmış bir yaklaşım elde edilmiştir.

Çalışmanın ilk aşamasında, merkez noktalarının daha etkin bir şekilde belirlenmesi için $\mathcal{A}$ kümesinin elemanlarının yoğun olduğu bölgeleri temsil edecek noktaların tespit edilmesi amaçlanmıştır. Bunun için k-ortalamalar algoritmasının değiştirilmiş

bir hali $\mathcal{A}$ kümesine uygulanmaktadır. Burada amaç, k-ortalamalar algoritması ile $\mathcal{A}$ kümesini temsil eden minimum sayıda noktanın elde edilerek, minimum sayıda ÇKF ile daha etkin çözümler bulunmasıdır. Geliştirilen k-ortalamalar algoritmasında, k ve artış olmak üzere iki tane parametre tanımlanmıştır. k değeri temel k-ortalamalar algoritmasından farklı olarak küme merkezleri için bir üst sınır olarak alınmıştır. Algoritmanın her bir adımında kullanıcı tarafından belirlenen bir artış miktarı kadar merkez sayında artırım yapılmaktadır. Her bir aşamada ise modelin eğitim başarısı hesaplanmaktadır. Burada algoritmanın durma kriteri ise eğitim başarısının bir önceki k değerine göre azalma olması durumudur. Geliştirilen k-ortalamalar algoritması ile A kümesine uygulanması sonucunda j adet merkez noktası ve bu merkez noktalarına göre atanan j adet veri kümesi elde edilmektedir.

Sınıflandırıcı fonksiyonunu oluşturan ÇKF'leri elde etmek için, gerekli olan merkez noktaları k-ortalamalar algoritması ile elde edilirken, w, ξ ve γ parametrelerini elde etmek için gürbüz doğrusal program(RLP) yaklaşımı kullanılmıştır. RLP modellenirken $\mathcal{A}$ kümesinin $\mathcal{B}$ kümesinden tamamıyla ayırmak yerine polihedronların dışında kalan $\mathcal{A}$ kümesi elemanları toplamı ile polihedronların iç kısımda kalan $\mathcal{B}$ kümesi elemanlarının toplamının en küçüklenmesi amaçlanmıştır. Bu şekilde hem daha hızlı ve etkin ÇKF'ler elde edilmiş hem de aşırı uyum sorununun önüne geçilmiştir. K-ortalama-RLP temelli ÇKF yaklaşımı algoritması izleyen adımları şekildedir:

D noktalar kümesi R^n'de n boyutlu uzayda tanımlı olsun.
$D = \{d_l : d_1, d_2, ...d_m\}$ $d_l \in R^n, l = \{1, 2, ..m\}$
$D = \bigcup_{j=1}^{c} D^j$,
D^j Veri kümesindeki sınıf bilgisi olmak üzere c ise sınıf sayısı olsun.
Buna göre k-ort-RLP-ÇKF algoritması;
Adım 0: Veri kümesindeki sınıf sayısı(C) olmak üzere j= 0 yap

Adım 1: j= j + 1 yap.

Adım 2.0: j kümesini A kumesi olarak belirle, k=0 yap, E_k=0 yap

Adım 2.1: k=k+$artis$ yap

Adım 2.2: A kümesine k-ortalamaları uygula ve A_i kümelerini elde et.

$A = \bigcup_{i=1}^{k} A_i$.

Adım 2.3: i=1 yap.

Adım 2.3.1: $i \leq k$ iken devam et değilse *Adim 2.4*'e git.

Adım 2.3.2: i. merkez noktası ve A_i veri kümesine gore RLP uygulayarak

$g_i(x)$ ÇKF fonksiyonlarını elde et.

Adım 2.3.3: i=i+1 yap ve *Adim 2.3.1*'e git.

Adım 2.4: Elde edilen i adet ÇKF ile sınıflandırıcı fonksiyonu oluştur.

$g^j(x) = \min_{1 \leq i \leq k} g_i^j(x)$

Adım 2.5: E_k eğitim başarısını hesapla

Adım 2.6: $E_k < E_{k-1}$ ise *Adım 2.1* 'e degilse *Adım 2.7* 'ye git.

Adım 2.7: $g^j(x)$ fonksiyonunu kaydet.

Adım 3: $j = C$ ise *Adım 4*'e git değilse *Adım 1*'e git.

Adım 4: Dur.

3.5 Açıklayıcı Örnek

Bu bölümde k-ort-RLP-ÇKF yaklaşımının daha iyi bir şekilde anlaşılması için iki boyutlu bir örnek incelenecektir. D kümesi A, B ve C üç sınıftan oluşan ve R^2 de tanımlı bir veri kümesi olsun. Buna göre,

D noktalar kümesi R^2'de 2 boyutlu uzayda tanımlı olsun.

$D = \{d_l : d_1, d_2, ...d_{185}\}$ $d_l \in R^2, l = \{1, 2, ..185\}$

$D = \bigcup_{j=1}^{3} D^j$,. Veri kümesinde c=3 sınıf sayısıdır.

Veri kümesindeki 185 noktanın 115 adeti eğitim, 70 adeti ise test verilerini oluşturmaktadır. Veri kümesinin elemanları ve sınıf bilgileri şekil 3.3 ve 3.4 'de verilmiştir. Bu örnek GAMS programında yazılmış ve çözülmüştür.

Örnek	x	y	Sınıf bilgisi
1	4,59	32,70	A
2	19,26	23,95	A
3	2,85	15,25	B
4	15,90	46,30	B
5	35,63	41,43	B
6	2,60	9,36	B
7	12,70	16,22	C
8	11,66	6,83	C
9	13,54	7,21	C
10	14,63	20,93	A
11	16,29	15,57	A
12	9,60	15,31	A
13	17,67	32,46	A
14	22,29	11,77	A
15	7,68	27,58	A
16	23,33	28,68	A
17	33,83	49,55	B
18	44,20	33,68	B
19	11,32	47,74	B
20	10,50	14,33	C
21	7,36	11,16	A
22	15,92	5,06	C
23	-13,30	-3,05	C
24	9,35	-14,61	C
25	11,58	13,77	A
26	8,53	21,44	A
27	14,98	19,45	A
28	49,64	10,44	B
29	26,57	38,45	B
30	36,31	5,77	B
31	-2,53	9,37	C
32	4,11	4,73	C
33	23,51	18,66	A
34	15,62	25,44	A
35	10,20	25,04	A
36	21,64	27,71	A
37	11,61	25,36	A
38	13,37	13,34	A
39	7,45	33,15	A
40	19,80	45,38	B

Örnek	x	y	Sınıf bilgisi
41	28,50	50,68	B
42	43,53	17,66	B
43	-9,90	7,40	C
44	-10,76	-0,32	C
45	1,51	7,63	C
46	-7,88	7,29	C
47	11,00	14,25	A
48	7,21	7,87	A
49	6,47	18,32	A
50	6,72	28,81	A
51	20,17	23,61	A
52	8,32	4,76	A
53	11,48	33,41	A
54	8,74	34,20	A
55	6,56	6,92	A
56	51,34	48,94	B
57	44,24	21,55	B
58	14,08	11,32	C
59	13,67	-7,87	C
60	-16,52	2,56	C
61	-15,56	-8,00	C
62	10,11	25,19	A
63	13,30	25,39	A
64	17,40	20,39	A
65	4,39	12,71	A
66	9,07	31,91	A
67	8,54	20,29	A
68	27,77	11,42	B
69	11,42	9,59	B
70	-6,31	6,82	C
71	4,20	10,99	C
72	12,22	-10,04	C
73	11,83	11,79	C
74	10,26	24,52	A
75	8,38	23,33	A
76	13,61	19,16	A
77	15,58	7,16	A
78	19,21	13,90	A
79	5,86	13,09	A
80	9,27	20,70	A

Örnek	x	y	Sınıf bilgisi
81	19,13	22,89	B
82	11,74	1,59	C
83	1,13	-11,17	C
84	5,54	-8,71	C
85	10,73	21,14	A
86	13,71	15,96	A
87	13,11	16,39	A
88	8,32	27,08	A
89	19,30	27,40	A
90	26,60	23,59	B
91	42,85	20,63	B
92	4,87	31,93	B
93	34,09	40,26	B
94	9,59	51,89	B
95	-14,69	-4,03	C
96	3,02	14,25	C
97	18,61	8,27	A
98	7,10	35,43	A
99	20,03	16,42	A
100	22,83	20,42	A
101	7,09	21,96	A
102	11,67	5,78	A
103	12,42	4,52	A
104	11,90	31,82	A
105	12,41	10,26	A
106	17,32	26,05	A
107	4,26	7,71	A
108	20,71	28,41	A
109	11,82	23,55	A
110	6,35	18,59	B
111	13,47	28,33	B
112	-10,22	5,69	C
113	-12,09	-16,64	C
114	-2,78	-12,69	C
115	4,16	11,78	C

Şekil 3.3: Eğitim kümesinin verileri

Veri kümesinin grafiksel olarak görünümü şekil 3.5 de verilmiştir. Problemin çözümünde ilk aşamasında merkez sayısına (k)üst sınır olarak 10 değeri verilmiştir. Merkez sayısı için artış ise 1 olarak belirlenmiştir. Çözüm yaklaşımında $1 - e - h$ yöntemi kullanıldığından dolayı başlangıçta A kümesinin B ve C kümelerinden ayıracak olan ÇKF'larının bulunması için A kümesine k-ortalamalar algoritması uygulanmıştır. K-ortalamalar sadece A kümesine uygulandığı için merkezler, kümenin

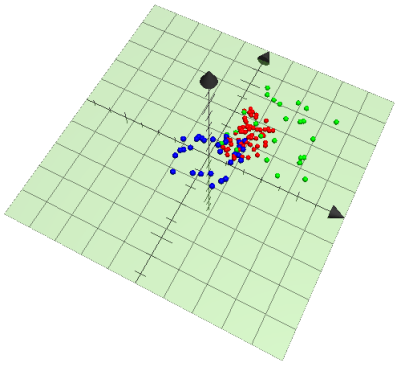
Örnek	x	y	Sınıf bilgisi
1	21,05	25,98	A
2	21,29	24,04	A
3	31,21	24,83	B
4	9,40	14,83	B
5	51,88	34,73	B
6	35,67	33,77	B
7	23,88	11,88	B
8	16,58	11,32	C
9	5,40	-11,92	C
10	8,48	-15,35	C
11	29,03	37,51	B
12	43,98	27,80	B
13	29,07	41,62	B
14	49,40	46,30	B
15	8,66	14,11	C
16	3,26	7,65	C
17	-10,31	-6,31	C
18	-13,93	15,90	C
19	12,86	10,12	A
20	15,73	23,30	A
21	17,01	31,01	A
22	16,76	26,00	A
23	7,01	31,61	A
24	16,42	29,18	A
25	19,12	30,05	B

Örnek	x	y	Sınıf bilgisi
26	6,75	35,85	B
27	43,24	32,90	B
28	-9,32	0,65	C
29	15,16	4,73	C
30	-0,71	13,39	C
31	11,66	14,46	A
32	5,33	10,94	A
33	15,23	31,57	A
34	5,20	14,78	C
35	2,12	-1,97	C
36	10,01	13,49	A
37	22,26	14,72	A
38	29,34	12,46	B
39	6,13	16,99	B
40	-8,75	15,80	C
41	13,77	23,10	A
42	20,70	16,86	A
43	16,12	15,46	A
44	5,32	36,28	A
45	-1,60	11,94	C
46	9,14	1,60	C
47	7,35	20,70	A
48	10,40	33,82	A
49	19,57	5,30	A
50	21,45	14,29	A

Örnek	x	y	Sınıf bilgisi
51	11,16	25,15	A
52	20,81	3,47	A
53	13,85	10,91	A
54	14,09	13,57	A
55	18,69	18,21	A
56	25,39	39,05	B
57	47,50	22,41	B
58	6,51	11,16	C
59	16,44	33,96	A
60	10,68	35,61	A
61	10,60	13,55	B
62	-16,02	-14,28	C
63	2,91	15,53	C
64	24,55	44,07	B
65	6,34	33,43	B
66	24,28	34,24	B
67	-11,04	0,34	C
68	-10,03	10,57	C
69	16,53	8,64	A
70	23,32	35,18	A

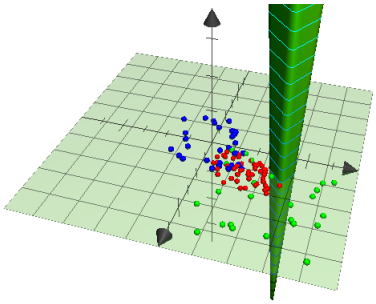
Şekil 3.4: Test kümesinin verileri

elemanlarının yoğun olduğu noktalara doğru yoğunlaşmıştır. Bu merkez noktalarına göre RLP uygulanarak PCF'in parametreleri elde edilmiştir. A kümesinin B ve C kümelerinden ayrılması için en iyi sonucu $k = 5$ değeri için bulunmuştur. A kümesinin B ve C kümesinden ayrılması için elde edilen fonksiyonlar ve grafikleri şekil 3.6, 3.7, 3.8, 3.9, 3.10 'de verilmiştir.



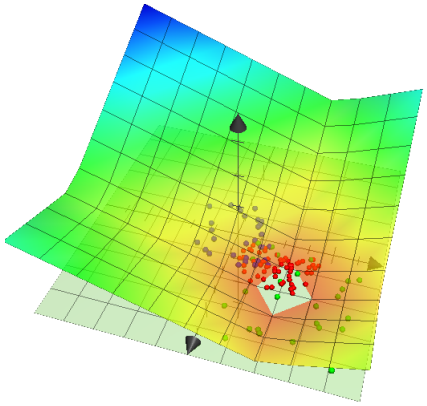
Şekil 3.5: Veri Kümesi

$$g_1(x, y) = -4.50(x - 8.38) + 3.12(y - 32.67) + 12.66(|x - 8.38| + |y - 32.67|) - 66.43$$



Şekil 3.6: $g_1(x, y)$ Fonksiyonu

$$g_2(x,y) = -0.04(x - 20.12) + 0.03(y - 25.70) + 0.23(|x - 20.12| + |y - 25.70|) - 3.31$$

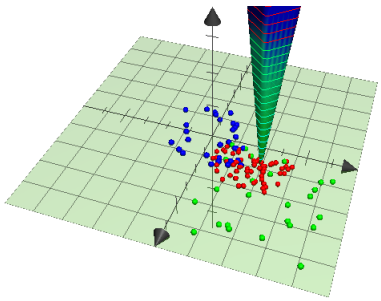


Şekil 3.7: $g_2(x, y)$ Fonksiyonu

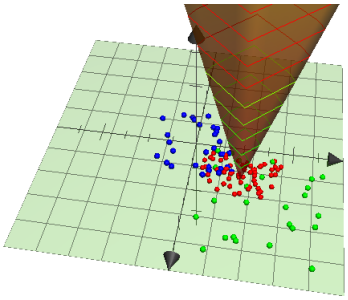
$$g_3(x,y) = 0.19(x - 9.48) + 0.01(y - 23.93) + 0.62(|x - 9.48| + |y - 23.93|) - 5.40$$

$$g_4(x,y) = -0.05(x - 15.15) - 0.04(y - 19.98) + 0.50(|x - 15.15| + |y - 19.98|) - 2.31$$

$$g_5(x,y) = -0.04(x - 12.38) + 0.01(y - 11.22) + 0.17(|x - 12.38| + |y - 11.22|) - 2.50$$



Şekil 3.8: $g_3(x, y)$ Fonksiyonu



Şekil 3.9: $g_4(x, y)$ Fonksiyonu

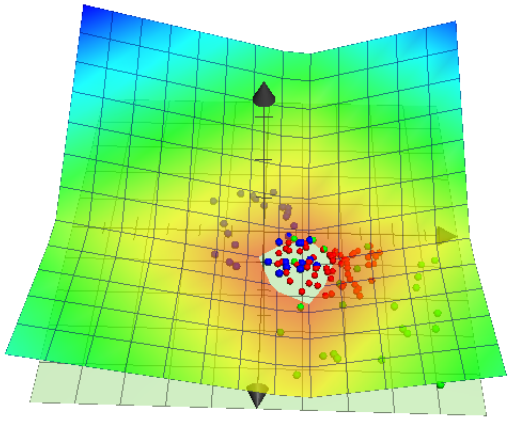
$$g(x) = \underset{l}{\text{enk}}\ g_l(x) \tag{3.10}$$

Problem çözüldükten sonra elde edilen $g_l(x)$ grafiği ise şekil 3.11 'de verilmiştir.

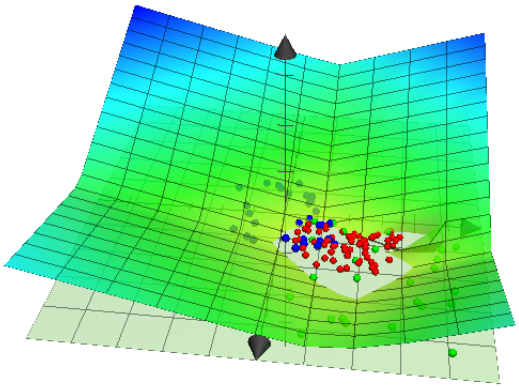
B kümesinin A ve C kümesinden ayrılması için en iyi çözüm k=2 değeri için bulunmuştur. Elde edilen fonksiyonlar;

$$\begin{aligned} g_1(x,y) &= -0.06(x-17.44) - 0.06(y-16.35) - 0.06(|x-17.44| + |y-16.35|) + 1.00 \\ g_2(x,y) &= -1.13(x-31.19) - 0.51(y-38.51) - 0.50(|x-31.19| + |y-38.51|) - 4.06 \end{aligned}$$

C kümesinin A ve B kümelerinden ayrılması için en iyi çözüm k=3 için bulun-



Şekil 3.10: $g_5(x, y)$ Fonksiyonu



Şekil 3.11: A kümesinin B ve C kümesinden ayıran $g(x, y)$ Fonksiyonu

muştur. Elde edilen fonksiyonlar;

$$
\begin{aligned}
g_1(x,y) &= 0.07(x+8.35)+0.10(y+7.55)-1.59 \\
g_2(x,y) &= 0.30(x-2.84)+0.19(y-5.13)-0.06(|x-2.84|+|y-5.13|)-2.16 \\
g_3(x,y) &= -0.15(x-13.61)-0.01(y-10.32)+0.54(|x-13.61|+|y-10.32|)-4.81
\end{aligned}
$$

3.6 Hesapsal Sonuçlar

Literatürdeki veri kümeleri, veri büyüklükleri, sınıf sayısı ve nitelik sayısına göre küçük boyutlu, orta büyüklü ve büyük boyutlu olarak ayrılmaktadır. Machine Learn-

Çizelge 3.1: Veri kümelerinin parametreleri[3]

	Shuttle	**Letters**	**Page-blocks**	**Abalone**
Örnek Sayısı	58000	20000	5473	4177
Egitim Sayısı	43500	15000	4000	3133
Test Sayısı	14500	5000	1473	1044
Sınıf Sayısı	7	26	5	3
Nitelik sayısı	10	17	11	9

ing Repository sitesinde veri kümeleri bu niceliklerine göre sınıflandırılmaktadır. Yeni geliştirilen K-Ort-RLP-ÇKF yaklaşımı literatürdeki dört büyük boyutlu çok sınıflı veri kümesine uygulanmıştır. Bu veri kümeleri;

- Deniz Kabuğu (Abalone)
- Uçak iniş kontrolu (Shuttle)
- Sayfa Blokları (Page-Blocks)
- Harf Tanıma(Letter recognition)

Bu kısımda ilk olarak veri kümeleri hakkında genel bir bilgilendirme yapılmıştır. Bir sonraki adımda ise veri kümelerinin literatürdeki yaklaşımlar çözülmesi ile elde edilen sonuçların K-Ort-RLP-ÇKF yaklaşımının sonuçları ile karşılaştırılmıştır. Önerilen bu yaklaşım ile literatürdeki yaklaşımlardan daha iyi sonuçlar elde edilmiştir. Ayrıca aşırı uyum probleminde önüne geçilmiştir. Bu yaklaşım ile bu probleminde önüne geçilmiştir.

Litaratürde en çok kullanılan çok boyutlu kümelerin bazılarının örnek büyüklükleri çizelge3.1'de bulabilirsiniz.

Deniz Kabuğu (Abalone) Veri Kümesi: Deniz kabuğu veri kümesi litaratürdeki büyük boyutlu sınıflandırma problemlerinin içinde en fazla rağbet gören veri kümelerindendir. 1994 yılında J. Nash ve arkadaşları tarafından Tazmanya'daki deniz kabuklarının biyolojik yapıları isimli çalışmadan alınmıştır. J.Nash ve arkadaşları, deniz kabuklarının fiziksel niteliklerine göre yaşlarını tahmin etmek istemişlerdir.

Çizelge 3.2: Deniz kabuğu veri kümesinin nitelikleri[3]

Nitelik	**Veri tipi**	**Ölçü**	**Tanım**
Cinsiyet	Sözel		M,F ve I
Uzunluk	Sayısal	mm	En uzun kabul uzunluğu
Çap	Sayısal	mm	Kabuğun diklemesine uzunluğu
Yükseklik	Sayısal	mm	Kabuğun et ile yüksekliği
Bütün Ağırlık	Sayısal	gr	Deniz kabuğunun tamamının ağırlığı
Kabuksuz Ağırlığı	Sayısal	gr	İç kısmın ağırlığı
İç Organ Ağırlığı	Sayısal	gr	Bağırsak ağırlığı
Kabuk Ağırlığı	Sayısal	gr	Kurutulduktan sonraki ağırlık
Halka	Tam sayı		+1.5 yıl içindeki yaşını verir

Normal şartlar altında deniz kabuklarının yaşları koni ile kesme, boyama yöntemleriyle mikroskop aracılığıyla kabuklardaki halka sayısına göre belirlenmektedir. Hava, lokasyon gibi ilave bilgilerde bu bilgileri çözmek için yararlı olmaktadır. Bu işlem hem zaman alıcı hemde sıkıcı bir işlemdir. Bu yüzden sınıflandırma problemlerinde deniz kabuklarının yaşlarının tespiti çok kullanılan veri kümelerindendir [3].

Deniz kabukları veri kümesi, sekiz niteliğe sahip ve 4177 veriden oluşmaktadır. Nitelikler kategorik, tamsayı ve reel sayılardan oluşmaktadır. Verilerin 3133 adeti eğitim geri kalan 1044 adeti ise test kümelerini oluşturmaktadır. Nitelik bilgilerini çizelge 3.2'de görebilirsiniz [3].

Uçak İniş(Shuttle) Veri Kümesi: Litaratürde en büyük boyutlu problemlerin başında gelmektedir. Avustralya Sidney üniversitesinden Jason Catlett tarafından geliştirilmiştir. Uçakların inişleri etkileyen dokuz niteliğin belirlendiği ve buna göre kararın verildiği bir veri kümesidir. Veri kümesinin yaklaşık olarak %80'i aynı kümeye aittir. Bu nedenle varsayılan doğruluk oranı %80'dir. Amaç ise %99 - %99, 90 oranında bir doğruluk elde etmektir. Orjinal veri kümesinde sipariş zamanı gibi bir bilgide bulunmaktadır ve bu bilgininde sınıflandırma ile ilgili olduğu düşünülmektedir. Ancak bu Shuttle'ın amaçları ile ilgili görünmemektedir. Orjinal veri kümesindeki örneklerin zamanlarıda sıralarıda rassal olarak seçilmiştir. Shuttle veri kümesindeki tüm nitelikler sayısaldır. Toplam 58000 veriden oluşmaktadır. bu verilerden 43500 adeti eğitim, 14500 adeti ise testtir. Diğer veri nitelikleri ise çizelge 3.3'de gösterilmiştir[3].

Çizelge 3.3: Shuttle veri kümesinin nitelikleri[3]

Nitelik	**Veri tipi**
Rad Flow	Sayısal
Fpv Close	Sayısal
Fpv Open	Sayısal
High	Sayısal
Bypass	Sayısal
Bpv Close	Sayısal
Bpv Open	Sayısal

Sayfa Blokları (Page-Blocks) Veri Kümesi: Sayfa blokları (Page-blocks) veri kümesi 1995 yılında Bari üniversitesinden Danota Malerba tarafından yayınlanmıştır. Bir dökümanın sayfa düzenindeki bütün blokların sınıflandırılması ve segmentayonunun belirlenmesi problemi olarak tanımlanmaktadır. 5473 adet örnek 54 farklı dökümandan meydana gelmiştir. Her gözlem bir blok ile ilgilidir ve niteliklerin tamamı sayısaldır. Bu çalışma grafik alanlarından metinlerin ayırmak için belgelerin analiz edilmesinde önemli bir adımdır. Gerçekteki beş sınıf şu şekildedir [3].

- Metin
- Yatay Çizgi
- Resim
- Dikey Çizgi
- Grafik

Sayfa blokları veri kümesindeki bütün örnekler mevcuttaki düşük görüntülemeye göre kontrol edilmiştir. Bazı verilere Uzunluk pürüzsüzleştirici algoritması (RSLA) uygulanarak farklı nitelikler elde edilmektedir. Bu veri kümesindeki nitelikler ise çizelge 3.4 sıralanmaktadır[3].

Harf Tanıma (Letter Recognition) Veri Kümesi: 1991 yılında David J. Slate tarafından yayınlanmıştır. David J. Slate'in bu çalışmasında ızgara tarama görüntülerinden harflerin doğru bir şekilde tahmin edilmesi için çeşitli varyasyonlar ile hollanda tarzı adaptif sınıflandırma yapabilen 16 nitelik vektörünün yetenekleri araştırılmıştır. Burada amaç ingiliz alfabesindeki 26 temel harfin görüntülerdeki

Çizelge 3.4: Sayfa blokları veri kümesinin nitelikleri[3]

Nitelik	**Veri tipi**	**Tanım**
Yükseklik	Tamsayı	Blokların yüksekliği
Genişlik	Tamsayı	Blokların genişliği
Alan	Tamsayı	Bloğun alanı (yükseklik*genişlik)
Açıklık	Sürekli	Blokların açıklığı (genişlik/yükseklik)
P-black	Sürekli	Blok içindeki siyah piksellerin oranı
P-and	Sürekli	RSLA uygulandıktan sonra siyah piksel oranı
Ortalama	Sürekli	Siyah beyaz geçişlerin ortalama sayısı
Siyah piksel	Tamsayı	Bloktaki orjinal haritanın siyah piksel sayısı
Siyah piksel2	Tamsayı	RLSA uygulandıktan sonra siyah piksellerin toplam sayısı
wb-Geçiş	Tamsayı	Blokların orjinal haritasındaki siyah beyaz geçişlerinin sayısı

siyah-beyaz dikdörtgen piksellerin sayılarına göre her birinin tanımlanmasıdır. Resimlerin analizinde, 20 farklı yazı fontu ve her bir harfin 20 yazı tipi ile rassal bir şekilde çarpılarak elde edilen 20000 benzersiz uyaran dikkate alınmaktadır. Her uyarana 0-15 arasındaki tamsayı dizisine sığacak ölçekli 16 temel sayısal özelliğe çevrilmesidir. Genel olarak ilk 15000 veri eğitim için geri kalan 5000 adet ise test için kullanılmaktadır. Bu veri kümesinde elde edilen en yüksek doğruluk oranı %80'den biraz fazladır. Diğer yaklaşımların bu oranı ne kadar geliştirebileceği merak edilen bir konudur. Harf tanıma veri kümesinin nitelikleri çizelge 3.5'da verilmiştir[3].

3.6.1 Veri Kümelerinin Sayısal Sonuçları

Bir önceki bölümde ayrıntılı olarak bahsedilen veri kümeleri sürekli yada tamsayıları nitelikleri içeren ve hiç bir eksik değeri olmayan kümelerdir. Bu veri kümeleri Waikoto, Yeni Zelanda Üniversitesi tarafından Java'da yazılmış olan veri madenciliğinin makine öğrenme paketi olan WEKA'da çözdürülmüştür. Veri kümelerinin WEKA'da çözümü için kullanılan sınıflandırıcılardan Sade Bayes algoritması(SB), lojistik ayırma (Lojistik), çok katmanlı algılayıcı (Multi-Layer Perceptron) (ÇKA), destek vektör makineleri (DVM), doğrusal destek vektörleri makineleri (DDVM), normalleştirilmiş polinomlar ile destek vektör makineleri (NDVM) ve DVM(PUK) kullanılmıştır. Bu sınıflandırıcıların WEKA'da çözümü sırasında varsayılan parametreler kullanılmıştır. Sınıflandırıcıların seçimi içinse Bagirov ve arkadaşlarının çalışmasından [29] faydalanılmıştır.

Yukarıda belirttiğimiz dört veri kümesinin k-ort-RLP-ÇKF algoritması ile çözümü

Çizelge 3.5: Harf Tanıma veri kümesinin nitelikleri[3]

Nitelik	**Veri tipi**	**Tanım**
Harf	Sözel	
x-box	Sayısal	Kutunun yatay pozisyonu
y-box	Sayısal	Kutunun dikey pozisyonu
Genişlik	Sayısal	Kutunun genişliği
Yükseklik	Sayısal	Kutunun yüksekliği
onpix	Sayısal	Toplam piksel sayısı
x-bar	Sayısal	Kutudaki x'lerin ortalaması
y-bar	Sayısal	Kutudaki y'lerin ortalaması
x2bar	Sayısal	x'lerin varyantı
y2bar	Sayısal	y'lerin varyantı
y2bar	Sayısal	x y'lerin korelasyonu
x2ybr	Sayısal	x*x*y nin ortalaması
xy2br	Sayısal	x*y*y nin ortalaması
x-ege	Sayısal	Sağdan sola köşe sayısının ortalaması
xegvy	Sayısal	x-ege ile y'nin korelasyonu
y-ege	Sayısal	Alttan üste doğru köşe sayısının ortalaması
xegvy	Sayısal	y-ege ile x'nin korelasyonu

Çizelge 3.6: Veri kümelerinin K-Ort-RLP-ÇKF ile çözülmesi ile edilen sonuçlar

	Başarı oranları	Abalone	Shuttle	Letter	Page-Block
K-ort-RLP-ÇKF	Eğitim	64.70	99.64	89.70	97.95
	Test	64.27	99.69	84.28	92.87

Çizelge 3.7: K-Ort-RLP-PCF ile diğer yaklaşımların çözümlerinin karşılaştırılması

	Abalone	Shuttle	Letter	Page-Block
SB	57.85	98.32	74.12	88.39
Lojistik	**64.27**	96.83	77.40	91.72
ÇKA	63.51	**99.75**	83.20	92.80
DVM	60.73		82.40	87.03
NDVM	60.25	96.81	82.34	89.48
DVM(PUK)	64.18	99.50		88.53
K-ort-RLP-ÇKF	**64.27**	99.69	**84.28**	**92.87**

ile elde edilen çözümler çizelge 3.6'da verilmiştir.

Veri kümelerinin Weka'da ilgili sınıflandırıcılar ile çözümü ile elde edilen sonuçlar ise çizelge 3.7'de verilmiştir.

K-ort-RLP-ÇKF algoritması GAMS programında kodlanmış ve dört veri kümesi GAMS'te çözülmüştür. K-ort-RLP-ÇKF yaklaşımı ile elde edilen sonuçları incelediğimizde eğitim sonuçları ile test sonuçlarının birbirine yakın olduğu görülmüştür. Bu da özellikle sınıflandırma problemlerinde karşılaşılan aşırı uyum sorunun çözüldüğünü göstermektedir. Elde edilen sonuçlar ile literatürdeki sınıflandırma problemi çözüm yaklaşımlarından karşılaştırılan yöntemler içerisinde Abalone, Letter ve Page-Blocks veri kümeleri için en iyi çözümler elde edilmiştir. Shuttle veri kümesi içinse ikinci en iyi çözüm elde edilmiştir. Özellikle veri kümelerinin doğruluk oranı düşük olan veri kümelerinde daha iyi sonuçlar elde edilmiştir. Ayrıca çözüm süreleri karşılaştırılmamasına rağmen süreler büyük boyutlu problemlere göre oldukça düşüktür.

4. SONUÇ ve ÖNERİLER

Günümüzde neredeyse bütün bilim dallarında büyük ve önemli çalışmalar yapılmaktadır. Her bir çalışma ile yeni veriler ve bilgiler elde edilmektedir. Bilgisayar sistemlerinin ve altyapılarının gelişmesi ile birlikte bu verilerin saklanmasıda kolaylaşmıştır. Bu aşamadan sonra geriye sadece bu verilerin ihtiyaçlara göre analiz edilmesi ve daha doğru ve hızlı karar almak için gerekli olan bilgilerin çıkarılması kalmaktadır. Günümüzde internet arama motorlarından, e-maillerin ayrılması, finans ve sağlık sektöründe bu verilerin analiz edilmektedir ve sınıflandırma yaklaşımları kullanılmaktadır. İşte bu yüzden sınıflandırma problemlerinin çözümü günümüzde olduğu gibi gelecekte de oldukça önemli bir yere sahip olacaktır.

Bu çalışmada da yukarıdaki bilgiler ışığında özellikle büyük boyutlu ve çok sınıflı sınıflandırma problemlerinin etkin ve hızlı bir şekilde çözümü için yeni bir yaklaşım üzerine çalışılmıştır. Literatürdeki sınıflandırma problemleri ve yaklaşımları incelenmiş bu yaklaşımların daha etkin hale nasıl getirilebilir sorusu üzerine araştırmalar yapılmıştır. Özellikle sınıflandırma problemlerinin çözümünde çok yüzlü konik fonksiyonların (ÇKF) etkin yaklaşımlardan birisi olduğu görülmüştür. Fakat bu yaklaşımın temel olarak başlangıç noktasının seçimi ve aşırı uyum gibi iki noktada dezavantajları olduğu görülmüş ve bunları geliştirecek yaklaşımlar tasarlanması amaçlanmıştır. Bu çerçevede kümeleme için kullanılan k-ortalamalar algoritması tek bir sınıfa uygulanarak daha etkin merkez noktaları elde edilmeye çalışılmıştır. Aşırı uyum için ise bir sınıfa ait bütün noktaları %100 ayırmak yerine sınıflandırma hatasını enküçükleyecek bir amaç üzerine çalışılmıştır. Sonuç olarak k-ort-RLP-ÇKF isimli yeni bir yaklaşım elde edilmiştir. Çalışmanın devamında literatürdeki veri setleri incelenmiş ve bu veri setlerinin literatürdeki çözümleri ile elde edilen sonuçlar karşılaştırılmıştır. Sonuçlara baktığımızda literatürde en çok karşılaşılan dört büyük boyutlu sınıflandırma probleminin üçü için en iyi sonuçlar elde edilmiştir.

Bu yeni yaklaşımda k-ortalama sadece bir sınıfa uygulandığı için merkez noktaları, küme elemanlarının yoğun olduğu bölgelerden oluşmaktadır. Fakat k-ortalamalar

için başlangıç noktaları hala rasgele seçilmektedir ki bu da merkez noktalarının seçimi için farklı stratejiler araştırılabilir. Günümüzde literatürde global k-means gibi farklı stratejiler geliştirilmiştir. Bu yaklaşımlar incelenebilir ve bu yeni yaklaşıma entegre edilebilir.

Bu çalışmada sonuç karşılaştırma yöntemlerinden $1-e-h$ yöntemi kullanılmıştır. Çok sınıflı sınıflandırma problemleri için $1-e-1$ gibi, iki sınıflı sınıflandırma problemleri içinde k-kez çapraz doğrulama gibi diğer yaklaşımlarda uygulanabilir. Daha etkin yaklaşımların bulunabilmesi için bu tür çalışamalara devam edilmelidir. Bu yaklaşımların yanı sıra litaratürde kullanılan veri kümelerinin yanısıra gerçek hayat problemleri üzerine de çalışmalar yapılmalıdır.

KAYNAKÇA

[1] P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*. Boston: Pearson Education, 2006.

[2] R. N. Gasimov and G. Ozturk, "Separation via polihedral conic functions," *Optimization Methods and Software*, no. 21, 527–540, 2006.

[3] Anonim, *Machine Learning Repository*. http://archive.ics.uci.edu/ml/datasets.html: UC Irvine University, 2010.

[4] S. TİRYAKİ, *Lojistik Alanında Bir Veri Madenciliği Uygulaması*. PhD thesis, İstanbul Teknik Üniversitesi, İstanbul, 2006.

[5] M. IŞIK, *Bölünmeli Kümeleme Yöntemleri İle Veri Madenciliği Uygulamaları*. PhD thesis, Marmara Üniversitesi, İstanbul, 2006.

[6] G. Silahtaroglu, *Kavram ve Algoritmalarıyla Temel Veri Madenciliği*. İstanbul: Papatya Yayincilik Egitim, 2008.

[7] R. Rastogi and K. Shim, "PUBLIC: A decision tree classifier that integrates," *Data mining and Knowledge Discovery*, no. 4, 315–344, 2000.

[8] G. Ozturk, *Sınıflandırma Problemleri İçin Yeni bir Matematiksel Progmram Yaklaşımı*. PhD thesis, Eskişehir Osmangazi Üniversitesi, Eskişehir, 2007.

[9] M. S. KAYGULU, *Supervised and Unsupervised Learning Techniques in Data Mining*. PhD thesis, Dokuz Eylül Üniversitesi, İzmir, 1999.

[10] R. A. Johnson and D. W. Wichern, *Applied Multivariate Statistical Analysis*. NewJersey: Prentice-Hall, 2002.

[11] K. Özdamar, *Paket Programlar ile İstatistiksel Veri Analizi-1*. Eskişehir: Kaan Kitabevi, 5 ed., 2004.

[12] Y. Özkan, *Veri Madenciliği Yöntemleri*. İstanbul: Papatya Yayincilik Egitim, 2008.

[13] G. Aynekin, *İnternet İçerik Madenciliğinde Yapay Sinir Ağları Ve Bir Uygulama*. PhD thesis, Uludağ Üniversitesi, Bursa, 2006.

[14] T. Cura, *Modern Sezgisel Teknikler Ve Uygulamaları*. İstanbul: Papatya Yayincilik Egitim, 2008.

[15] O. Mangasarian, "Linear and nonlinear separation of patterns by linear programming," *Operations Research*, no. 13, 444–452, 1965.

[16] K. P. Bennett and O. L. Mangasarian, "Robust linear programming discrimination of two linearly inseparable sets," *Optimization Methods and Software*, no. 1, 23–34, 1992.

[17] A. Astorino and M. Gaudioso, "Polyhedral separability through successive lp," *Journal of Optimization Theory and Applications*, no. 112-2, 265–293, 2002.

[18] A. M. Bagirov, "Max-min separability," *Optimization Methods and Software*, no. 20, 2–3, 271–290, 2005.

[19] S. S. Erenguc and G. J. Koehler, "Survey of mathematical programming models and experimental results for linear discriminant analysis," *Managerial and Decision Economics*, no. 11-4, 215–225, 1990.

[20] J. J. Glen, "Integer programming methods for normalisation and variable selection in mathematical programming discriminant analysis models," *The Journal of the Operational Research Society*, no. 50-10, 1043–1053, 1999.

[21] J. J. Glen, "Classification accuracy in discriminant analysis: A mixed integer programming approach," *The Journal of the Operational Research Society*, no. 52-3, 328–339, 2001.

[22] F. Uney and M. Turkay, "A mixed-integer programming approach to multiclass data classification problem," *European Journal Of Operational Research*, no. 173-3, 910–920, 2006.

[23] R.Gasimov and G.Öztürk, "Separation via polyhedral conic functions," *Optimization Methods and Software*, no. 21, 527-540, 2006.

[24] A.M.Bagirov, "Modified global k-means algorithm for minimum sum-of-squares clustering problems," *Elsevier Science Inc.*, no. 41-1, 3192–3199,2008.

[25] R.J.Roiger and M.W.Geatz, *Data Mining A Tutorial-Based Primer*. United States: Pearson Education Inc, 2003.

[26] K.Alsabti, S.Ranka, and V.Singh, "An efficient k-means clustering algorithm," in *In Proceedings of IPPS/SPDP Workshop on High Performance Data Mining*, 1998.

[27] A.Likas, N.Vlassis, and J.J.Verbeek, "The global k-means clustering algorithm," *Pattern Recognation*, no. 36-2, 451-461, 2003.

[28] G.F.Tzortzis and A.C.Likas, "The global kernel k-means algorithm for clustering in feature space," *IEEE Transactions on Neural Networks*, no. 20-7, 1181-1194, 2009.

[29] A.Bagirova, J.Ugona, D.Webba, G.Ozturk, and R.Kasimbeyli, "A novel piecewise linear classifier based on polyhedral conic and max-min separabilities," *TOP*, no. Submitted Paper, 2011.