Örüntü Tanıma Uygulamalarında Altuzay Analiziyle Öznitelik Seçimi ve Sınıflandırma

Serkan Günal

DOKTORA TEZİ

Elektrik ve Elektronik Mühendisliği Anabilim Dalı

Haziran 2008

Feature Selection and Classification by Subspace Analysis in Pattern Recognition Applications

Serkan Günal

DOCTORAL DISSERTATION

Department of Electrical and Electronics Engineering

June 2008

Örüntü Tanıma Uygulamalarında Altuzay Analiziyle Öznitelik Seçimi ve Sınıflandırma

Serkan Günal

Eskişehir Osmangazi Üniversitesi
Fen Bilimleri Enstitüsü
Lisansüstü Yönetmeliği Uyarınca
Elektrik ve Elektronik Mühendisliği Anabilim Dalı
Telekomünikasyon Bilim Dalında
DOKTORA TEZİ
Olarak Hazırlanmıştır

Danışman: Yrd. Doç. Dr. Rifat EDİZKAN

Haziran 2008

ONAY

Elektrik ve Elektronik Mühendisliği Anabilim Dalı Doktora öğrencisi Serkan Günal'ın DOKTORA tezi olarak hazırladığı "Örüntü Tanıma Uygulamalarında Altuzay Analiziyle Öznitelik Seçimi ve Sınıflandırma" başlıklı bu çalışma, jürimizce lisansüstü yönetmeliğin ilgili maddeleri uyarınca değerlendirilerek kabul edilmiştir.

Danışman **:** Yrd. Doç. Dr. Rifat EDİZKAN

İkinci Danışman : ——————

Doktora Tez Savunma Jürisi:

Üye : Yrd. Doç. Dr. Rifat EDİZKAN

Üye : Yrd. Doç. Dr. Erol SEKE

Üye : Doç. Dr. Ömer Nezih GEREK

Üye : Prof. Dr. M. Bilginer GÜLMEZOĞLU

Üye : Prof. Dr. Atalay BARKANA

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

Prof. Dr. Nimetullah BURNAK

Enstitü Müdürü

Örüntü Tanıma Uygulamalarında Altuzay Analiziyle Öznitelik Seçimi ve Sınıflandırma

Serkan GÜNAL

ÖZET

Bu tez çalışmasında, örüntü tanımanın üç temel öğesi olan öznitelik çıkarımı, öznitelik seçimi ve sınıflandırma konuları üzerinde durulmuştur.

Öznitelik çıkarımı konusunda, ses ve konuşma sinyalleri için Fourier dönüşümüne alternatif olabilecek dalgacık dönüşümü temelli bir öznitelik çıkarım yöntemi önerilmektedir. Farklı karakteristikteki ses sinyallerini içeren veritabanları üzerinde yapılan deneylerde, dalgacık özniteliklerinin Fourier özniteliklerine göre, özellikle durağan olmayan ve ani frekans değişimleri içeren sinyalleri daha iyi temsil ettiği gözlenmiştir.

Öznitelik seçimi konusunda, özniteliklerin bireysel ayırdedicilik derecelerini belirleyen altuzay temelli iki yeni ayrılabilirlik ölçüsü geliştirilmiştir. Bu ölçüler daha sonra çok sınıflı örüntü tanıma problemlerinde öznitelik seçimi amacıyla kullanılmıştır. Farklı sayı ve yapıda özniteliği barındıran veritabanları üzerinde yapılan deneyler, altuzay temelli ölçülerle yapılan öznitelik seçiminin, uzaksaklık ve Bhattacharyya gibi klasik ayrılabilirlik ölçüleriyle yapılan seçime göre gerek sınıflandırma hassasiyeti gerekse boyut indirgeme açısından daha başarılı olduğunu ortaya koymuştur.

Sınıflandırma konusunda genetik algoritma temelli yeni bir altuzay sınıflandırıcı geliştirilmiştir. Yeni sınıflandırıcı hem sınıf-içi hem de sınıflar-arası ilişkileri değerlendirmekte; ayrıca, klasik altuzay sınıflandırıcıların aksine altuzay izdüşümü için gerek büyük gerekse küçük özdeğerlere karşılık gelen özyönleri birlikte kullanabilmektedir. Çeşitli veritabanları üzerinde yapılan deneylerde, genetik altuzay sınıflandırıcı klasik altuzay yöntemlerine göre bu özelliği ile öne çıkarak karşılaştırılabilir ya da daha yüksek bir sınıflandırma başarımı sağlamıştır.

Anahtar Kelimeler: Örüntü tanıma, altuzay, öznitelik çıkarma, öznitelik seçimi ve sınıflandırma

Feature Selection and Classification by Subspace Analysis in Pattern Recognition Applications

Serkan GÜNAL

SUMMARY

In this thesis study, feature extraction, feature selection and classification subjects, which are the three fundamental topics of pattern recognition, are studied.

In feature extraction, a wavelet transform based feature extraction method for sound and speech signals is proposed as alternative to classic Fourier transform. Experiments on several datasets containing signals with different characteristics indicate that the wavelet features represent signals better than the Fourier features in case of non-stationary structure and instantaneous frequency changes.

In feature selection, two novel separability measures, which detect the individual discriminatory powers of the features, are developed. These measures are then employed for feature selection in multi-class pattern recognition problems. Experiments on several datasets with different characteristics and different number of features reveal that the subspace based feature selection is better than classic separability measures such as Divergence and Bhattacharyya in terms of both classification accuracy and dimension reduction.

In classification, a genetic algorithm based subspace classifier is developed. New classifier evaluates both within-class and between-class relationships; moreover, it is capable of using the eigendirections corresponding to both large and small eigenvalues together for subspace projection unlike the classic subspace classifiers. Due to these properties, the genetic subspace classifier provides comparable or even better classification performance with respect to the classic subspace methods in the experiments that are carried out for different datasets.

Keywords: Pattern recognition, subspace, feature extraction, feature selection and classification

TEŞEKKÜR

Bu tez çalışmasının ortaya çıkmasındaki katkılarından dolayı başta danışmanım Yrd. Doç. Dr. Rifat EDİZKAN olmak üzere tez izleme komitesinde yer alan Doç. Dr. Ömer Nezih GEREK ve Yrd. Doç. Dr. Erol SEKE'ye teşekkür ederim.

Ayrıca, doktora eğitimim süresince benden desteklerini esirgemeyen sevgili ailemin tüm fertlerine ve bu süreçte aramızdan ayrılan değerli babama minnetlerimi sunarım.

İÇİNDEKİLER

Sayfa

İÇİNDEKİLER (devam)

Sayfa

İÇİNDEKİLER (devam)

Sayfa

ŞEKİLLER DİZİNİ

Sayfa

ŞEKİLLER DİZİNİ (devam)

Sayfa

ÇİZELGELER DİZİNİ

Sayfa

ÇİZELGELER DİZİNİ (devam)

Sayfa

SİMGELER VE KISALTMALAR DİZİNİ

B	Bhattacharyya
CLAFIC	Sınıf Özellikli Bilgi Sıkıştırma (Class Featuring Information Compression)
CS	Ortak Altuzay (Common Subspace)
CVA	Ortak Vektör Yaklaşımı (Common Vector Approach)
D	Uzaksaklık (Divergence)
DCT	Kesikli Kosinüs Dönüşümü (Discrete Cosine Transform)
ES	Tam Kapsamlı Arama (Exhaustive Search)
FLDA	Fisher'in Doğrusal Ayırtaç Analizi (Fisher's Linear Discriminant Analysis)
FS	Fisher Altuzayı (Fisher Subspace)
FT	Fourier Dönüşümü (Fourier Transform)
GA	Genetik Algoritma (Genetic Algorithm)
GSBS	Genelleştirilmiş Ardışık Geri Yönde Seçim (Generalized Sequential Backward Selection)
GSFS	Genelleştirilmiş Ardışık İleri Yönde Seçim (Generalized Sequential Forward Selection)
MFCC	Mel Süzgeçlenmiş Kepstrum Katsayıları (Mel Filtered Cepstrum Coefficient)
NDA	Parametrik Olmayan Ayırtaç Analizi (Nonparametric Discriminant Analysis)
PCA	Temel Bileşen Analizi (Principal Component Analysis)
PTA	Artı *l* - Çıkar *r* Seçim (Plus *l* - Takeaway *r* Selection)
SBFS	Ardışık Geri Yönde Kayan Seçim (Sequential Backward Floating Selection)
SBS	Ardışık Geri Yönde Seçim (Sequential Backward Selection)
SFFS	Ardışık İleri Yönde Kayan Seçim (Sequential Forward Floating Selection)
SFS	Ardışık İleri Yönde Seçim (Sequential Forward Selection)
STFT	Kısa Süreli Fourier Dönüşümü (Short Time Fourier Transform)
WT	Dalgacık Dönüşümü (Wavelet Transform)

BÖLÜM 1

GİRİŞ

1.1 Örüntü

Örüntü, olay veya nesnelerin düzenli bir biçimde birbirini takip ederek gelişmesi şeklinde tanımlanır (Türk Dil Kurumu, 2008). Başka bir ifadeyle örüntü, kendini tekrarlayan olay veya nesnelere karşılık gelir. Ses sinyali, uzaktan algılama verisi, insan yüzü, retina, parmak izi, bir metin içerisindeki karakterler ve biyomedikal cihazlardan elde edilen görüntüler, örüntüye dair bazı örneklerdir.

1.2 Örüntü Tanıma

Örüntü tanıma, ortak özelliğe sahip veya aralarında bir ilişki kurulabilen karmaşık olayları ve nesneleri çeşitli yöntemler vasıtasıyla tanımlayıp bir kategoriye ya da sınıfa koyma işlemidir. İnsanoğlu, sahip olduğu görme, işitme, dokunma, tat ve koku alma duyuları sayesinde çevresindeki örüntüleri belli oranda tanıma yeteneğine sahiptir. Gerek bu işlemin otomatik hale getirilmesi gerekse insan duyularının yeterli olmadığı durumlarda, tanıma ve sınıflandırmanın sağlanabilmesi için "makina ortamında örüntü tanıma" kavramı ortaya çıkmıştır. Konuşma, konuşmacı, insan yüzü, parmak izi, el yazısı, imza ve biyomedikal görüntü tanıma, örüntü tanıma uygulamalarının yaygın örnekleridir. Yarım yüzyılı aşkın süredir üzerinde çalışılmakta olan bu bilim kolunun önemi, her geçen gün ortaya çıkan yeni uygulama alanlarıyla birlikte giderek artmaktadır.

1.3 Örüntü Tanıma Süreci

Örüntü tanıma sürecinin temel öğeleri, öznitelik çıkarımı, öznitelik seçimi ve sınıflandırmadır (Bkz. Şekil 1.1). Bu temel öğelerin yanında, bazı durumlarda birtakım önişlem (gürültü azaltma, önvurgulama, v.b.) ve sonişlemlerde süreç içerisinde yer alabilir.

Şekil 1.1. Örüntü tanıma süreci

1.4 Öznitelik Çıkarımı

Öznitelik, temel olarak örüntüye dair ölçülebilir ya da gözlenebilir bilgidir. Öznitelik çıkarımı, ilgisiz ve fazla bilgiyi eleyerek örüntüye ait karakteristik özelliklerin elde edilmesini sağlar. Gereksiz bilgilerin elenmesi, tanıma işleminin süresini kısaltmak açısından da büyük önem taşımaktadır. Eğer bir örüntü birden fazla öznitelik ile temsil ediliyorsa, tek bir öznitelik yerine bir "öznitelik kümesi" söz konusudur. d adet özniteliğe sahip öznitelik kümesi ise d-boyutlu "öznitelik vektörü" ile temsil edilir. Özniteliklerin içinde bulunduğu d-boyutlu $\mathbb{R}^d$ uzayı ise "öznitelik uzayı" olarak isimlendirilir (Kuncheva, 2004). Örüntülere ait öznitelikler, nicel (sayısal) ya da nitel (kategorik) olabilir. Örneğin, bir otomobilin maksimum hız bilgisi nicel, model bilgisi ise nitel bir özniteliktir. İstatistiksel örüntü tanıma daha çok nicel özniteliklerden faydalanırken, sözdizimsel örüntü tanıma ağırlıklı olarak nitel öznitelikleri kullanır.

1.5 Öznitelik Seçimi

Öznitelik çıkarma işleminde, örüntü hakkındaki ilgisiz bilgilerin elenip karakteristik özelliğin elde edilmesiyle belli oranda boyut indirgeme sağlanır. Öznitelik seçiminde ise çıkarılmış olan özniteliklerin ayırdedicilikleri çeşitli yöntemlerle incelenerek mevcut öznitelik kümesinden daha ayırdedici bir altküme bulunması amaçlanır. Bu işlem, boyut indirgeme oranını artırmanın yanısıra "boyutun laneti" (Bellman, 1961; Theodoridis and Koutroumbas, 2003; Jain and Zongker, 1997) etkisini de azaltmaya yarar.

1.6 Sınıflandırma

Bilinmeyen bir örüntüyü tanıyabilmek için öznitelik çıkarımı ve öznitelik seçimini takiben sınıflandırma işlemi yürütülür. Sınıflandırma için, hangi sınıfa ait olduğu önceden bilinen belirli sayıdaki öznitelik vektörünün oluşturduğu "veri kümeleri" bir eğitim sürecinden geçirilir. Bu eğitim sonucunda, bilinmeyen örüntüyü uygun sınıfa atamakta kullanılan bir karar kuralı ya da mekanizması oluşturulur. Bir eğitim kümesi ve birtakım önsel olasılıklar yardımıyla gerçekleştirilen bu sınıflandırma yaklaşımı "güdümlü sınıflandırma" olarak isimlendirilir. Herhangi bir önsel bilgi ve eğitim kümesi kullanılmadan gerçekleştirilen sınıflandırma ise "güdümsüz sınıflandırma" dır (Duda et al., 2001; Theodoridis and Koutroumbas, 2003).

1.7 Tez Çalışması

Öncelikli olarak, bu tez çalışması, istatistiksel örüntü tanıma çerçevesinde nicel öznitelikler ve güdümlü sınıflandırma yöntemleri kullanılarak gerçekleştirilmiştir. Çalışmada, temel olarak örüntü tanıma sürecinin yukarıda bahsi geçen üç ana başlığı

(öznitelik çıkarımı, öznitelik seçimi ve sınıflandırma) üzerinde durulmuştur. Tez çalışması kapsamında,

- Ses ve konuşma tanıma uygulamalarında kullanılabilecek dalgacık dönüşümü temelli bir öznitelik çıkarma yöntemi
- Altuzay analizine dayalı sınıf ayrılabilirlik ölçüleri ve çok sınıflı örüntü tanıma uygulamaları için bu ölçüleri kullanan öznitelik seçim yöntemi
- Genetik algoritma temelli bir altuzay sınıflandırıcı

geliştirilmiştir. Çalışmanın düzeni şu şekildedir:

Bölüm 2'de, başlangıç olarak öznitelik çıkarımı hakkında temel bilgiler verilmiş ve bazı örüntü tanıma uygulamaları için kullanılan öznitelik çıkarma yöntemleri hakkında örnekler sunulmuştur. Daha sonra, özellikle ses ve konuşma tanıma uygulamalarında kullanılabilecek dalgacık dönüşümü temelli bir öznitelik çıkarma yöntemi önerilmiş ve bu yöntemin Fourier dönüşümü temelli klasik yöntem ile karşılaştırması yapılmıştır. Konuyla ilgili gerçekleştirilen deneysel çalışmalar (Bölüm 5), dalgacık dönüşümü temelli özniteliklerin, ani frekans değişimleri içeren ve durağan olmayan sinyallerde doğru tanıma oranı açısından daha başarılı sonuçlar sağladığını ortaya koymuştur.

Bölüm 3'te, öncelikle sınıf ayrılabilirliği ve öznitelik seçimi konusunda temel bilgiler verilmiş, literatürde yaygın olarak kullanılan öznitelik seçim yöntemleri açıklanmıştır. Daha sonra, klasik ayrılabilirlik ölçülerine alternatif olarak, bu tez kapsamında geliştirilmiş olan altuzay temelli iki yeni ayrılabilirlik ölçüsü ve bu ölçülerin çok sınıflı örüntü tanıma problemlerinde öznitelik seçimi amacıyla kullanılabilmesi için önerilen yöntem tanıtılmıştır. Konuyla ilgili yapılan deneysel çalışmalarda (Bölüm 5), altuzay temelli ayrılabilirlik ölçülerinin hem doğru sınıflandırma hem de boyut indirgeme oranı açısından etkili olduğu gözlenmiştir.

Bölüm 4, örüntü tanımanın son aşaması olan sınıflandırma üzerinedir. Bu bölümde, altuzay temelli doğrusal sınıflandırıcılar esas alınmıştır. İlk olarak, altuzay

sınıflandırma mantığı açıklanmış ve klasik altuzay sınıflandırıcılar hakkında bilgiler verilmiştir. Bunu takiben, bu tez çalışması kapsamında, klasik altuzay sınıflandırıcılara alternatif olarak geliştirilen genetik algoritma temelli altuzay sınıflandırıcı tanıtılmaktadır. Geliştirilen bu yeni sınıflandırıcının başarımı, deneysel çalışmalarda (Bölüm 5) diğer sınıflandırıcılar ile karşılaştırılmış ve başarılı sonuçlar elde edilmiştir.

Bölüm 5, tez kapsamında yapılan tüm deneysel çalışmaları içermektedir. Bu çalışmalarda altı farklı veritabanı kullanılmıştır. Bu doğrultuda, öncelikle veritabanları tanıtılmış, her bir veritabanında kullanılan öznitelikler açıklanmış, sınıf, öznitelik ve örnek sayıları belirtilmiştir. Daha sonra, öznitelik çıkarımı, öznitelik seçimi ve sınıflandırma konularının her biri içi yapılan deneyler ayrı ayrı açıklanmış, her deneye ait sonuç ve yorumlara yer verilmiştir.

Bölüm 6, tez çalışması kapsamında yapılan araştırmalara ve geliştirilen yöntemlere dair genel sonuçları içermektedir.

BÖLÜM 2

ÖZNİTELİK ÇIKARIMI

Öznitelik çıkarımı, en basit ifadeyle bir boyut indirgeme işlemidir. Sınıflandırılacak bir örüntü genellikle çok fazla miktarda ve gereksiz bilgi içerir. Bu durum, sınıflandırma hassasiyetini düşürürken işlem süresini de yükseltir. Bu olumsuzluğu gidermek için, örüntü bilgisi daha düşük miktardaki başka bir veriye dönüştürülür. Örüntüye ait fazla ve gereksiz verinin elenip, sadece örüntüyü temsil eden ve toplam veriden çok daha az sayıdaki karakteristik bilginin elde edildiği bu dönüşüme öznitelik çıkarımı adı verilir. Karakteristik özniteliklerin çıkarımı, örüntü tanıma sistemlerinin kritik tasarım aşamalarından biridir. Çıkartılan özniteliklerin ayırdedici ve mümkün olduğunca az sayıda olması, tanıma işleminin daha basit sınıflandırıcılarla, daha yüksek hassasiyetle ve daha kısa sürede gerçekleştirilmesini sağlar.

Öznitelik çıkarımı için kullanılacak yöntem, problem alanına bağlıdır. Örneğin, ses ve konuşma tanıma problemlerinde, sinyalin spektrumu incelenerek sinyalin tümü yerine çeşitli frekans bantlarındaki bilgiler öznitelik olarak kullanılabilir (Rabiner and Juang, 1993). Görüntü tanıma problemlerinde, bir görüntünün tamamı yerine, görüntünün spektrum analizi bilgileri ve görüntüye ait renk, parlaklık, köşe ve kenar gibi bilgiler öznitelik kümesini oluşturabilir (Gonzalez and Woods, 2007). Metin sınıflandırma uygulamalarında, metin içerisindeki kelime ve sembollerin tamamından ziyade sınıflara özgü birtakım anahtar kelime ve sembollerin frekansları (metin içerisinde görülme sıklıkları) öznitelik olabilir (Günal et al., 2006). Problem alanı için mantıksal ya da algoritmik bir öznitelik çıkarım yöntemi bulunamadığı durumlarda ise Temel Bileşen Analizi (PCA) gibi boyut indirgeme yöntemlerinden faydalanılabilir (Duda et al., 2001).

Tez çalışmasının bu bölümünde, ses ve konuşma tanıma uygulamalarında, Fourier dönüşümüne alternatif olarak kullanılabilecek dalgacık dönüşümü temelli bir öznitelik çıkarım yöntemi önerilmektedir. Literatürde, ses sinyallerinden öznitelik çıkarmada en

çok tercih edilen yöntemlerden biri Fourier dönüşümü temelli Mel Süzgeçlenmiş Kepstrum Katsayıları (MFCC)'dır (Rabiner and Juang, 1993). Ancak, Fourier dönüşümü, durağan olmayan ve ani frekans değişimleri içeren sinyalleri temsil etmekte yetersiz kalabilmektedir (Mallat, 1999; Proakis and Manolakis, 2006). Buna ilaveten, MFCC parametreleri sinyallerin frekans bantlarına dair direkt olarak bir bilgi sunmazlar. MFCC özniteliklerin çıkarılmasına dair detaylı bilgi, (Günal, 2003) çalışmasında bulunabilir. Bu tez çalışmasında önerilen yöntem ile elde edilen öznitelikler, hem uzun süreli ve durağan olmayan hem de ani frekans değişimleri içeren sinyalleri temsil etmede Fourier temelli özniteliklere göre daha başarılıdır. Bu öznitelikler, aynı zamanda sinyallerin frekans bantlarına dair bilgileri taşımaktadır. Bu yapı sayesinde dalgacık temelli öznitelikler, tez çalışmasının öznitelik seçme bölümüne ait bazı deneylerde ayırdedici altbantların bulunması için de kullanılmıştır.

Bölüm 5'teki konuyla ilgili deneysel çalışmalarda, önerilen dalgacık temelli özniteliklerin Fourier dönüşümü temelli öznitelikler ile sınıflandırma hassasiyeti açısından karşılaştırması yapılmıştır. Bu karşılaştırmanın eşit koşullarda gerçekleşmesini sağlamak için, Fourier temelli MFCC parametreleri yerine bunların frekans bölgesindeki karşılıkları (kesikli kosinüs dönüşümü (DCT) uygulanmamış halleri) kullanılmıştır.

2.1 Fourier ve Dalgacık Dönüşümleri

Fourier analizi, basit bir ifadeyle, sinyalleri değişik frekanslardaki sinüs dalgalarına ayırır. Başka bir deyişle, Fourier dönüşümü sayesinde zaman bölgesindeki sinyalin frekans bölgesindeki karşılığı elde edilir (Bkz. Şekil 2.1).

Şekil 2.1. Fourier dönüşümü

Ancak, bu dönüşüm sırasında zaman bilgisi kaybolur. Yani, sinyal içerisindeki farklı frekansların hangi zaman diliminde varolduğu kestirilemez. Bu durum, durağan sinyaller için önemli olmamakla birlikte, konuşma gibi durağan olmayan - zaman içerisinde değişim gösterebilen – sinyaller açısından problem teşkil etmektedir. Bu olumsuzluğu gidermek için pencereleme işlemiyle sinyal küçük çerçevelere ayrılır ve çerçeve içinde kalan kısa süreli sinyallerin durağan olduğu kabul edilerek, her çerçeve için Fourier dönüşümü hesaplanır (Şekil 2.2). Bu işleme, Kısa Süreli Fourier Dönüşümü (STFT) adı verilir (Misiti et al., 2005; Proakis and Manolakis, 2006).

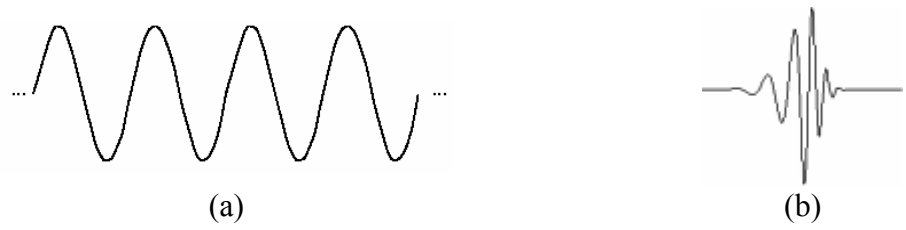
Şekil 2.2. Kısa süreli Fourier dönüşümü

STFT, sinyalin zaman ve frekans eksenindeki durumunun bileşkesini verir. Sinyalin, hangi zaman diliminde hangi frekansa sahip olduğuna dair bilgiler sunar. Ancak, bu bilgiler pencereleme boyutuyla sınırlı bir hassasiyete sahiptir. Sinyalin spektrumunda yer alan düşük ya da yüksek tüm frekans bileşenleri aynı zaman penceresinde incelenebilir. Bununla birlikte, konuşma sinyali ve benzeri sinyal

çeşitlerinde zaman ya da frekans bilgisinin daha hassas belirlenmesi gerektiği için STFT yetersiz kalmakta; dolayısıyla, daha esnek bir yapıya ihtiyaç duyulmaktadır.

Şekil 2.3. Dalgacık dönüşümü

Dalgacık dönüşümü (Şekil 2.3), gerekli olan bu esnekliği sağlamak için kullanılan bir dönüşüm tipidir. Dalgacık dönüşümü de Fourier gibi sinyali küçük parçalara ayırır. Ancak, Fourier dönüşümü bu işlem için sonsuz uzunlukta olduğu varsayılan ve değişik frekanslardaki düzenli sinüs dalgalarını (Şekil 2.4-a) kullanırken, dalgacık dönüşümü "ana dalgacık" (Şekil 2.4-b) adı verilen sınırlı süreli, düzensiz ve asitmetrik sinyal parçalarının, ölçeklenmiş ve kaydırılmış hallerini kullanır (Mallat, 1999). Şekil 2.4'den anlaşılacağı gibi, sinyallerdeki kısa süreli ve keskin değişiklikler, pürüzsüz ve düzgün bir sinüs dalgasından ziyade düzensiz bir dalgacık ile daha iyi analiz edilebilir. Başka bir ifadeyle, dalgacık dönüşümü daha iyi zaman - frekans lokalizasyonu sağlar. Dalgacık dönüşümünde yaygın olarak kullanılan ana dalgacık türleri Haar, Daubechies, Coiflet, Symlet, Morlet ve Meyer'dir (Mallat, 1999).



Şekil 2.4. a) Sinüs dalgası b) Dalgacık (Daubechies-8)

Matematiksel olarak,

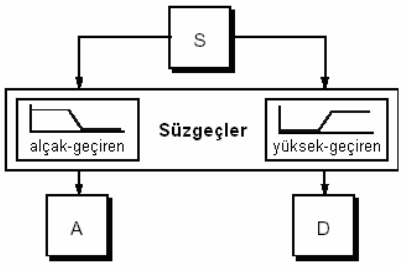
$$F(\omega) = \int_{-\infty}^{\infty} f(t) e^{-j\omega t} dt \tag{2.1}$$

şeklinde ifade edilen Fourier dönüşümü (Proakis and Manolakis, 2006), $f(t)$ sinyalinin bütün zaman zarfında karmaşık üstel bir çarpanla çarpımlarının toplamına karşılık gelmektedir. Karmaşık üst, gerçek ve sanal sinüs biçimli bileşenlere ayrılabilir. Dalgacık dönüşümü ise,

$$C(a,b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} x(t) \psi(\frac{t-b}{a}) dt \tag{2.2}$$

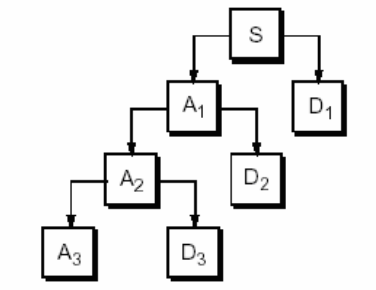
şeklindedir (Mallat, 1999). Burada, $\psi(t)$ ana dalgacığı, a ölçek faktörünü, b ise kayma faktörünü göstermektedir.

Standart dalgacık dönüşümü ile bir sinyal (S), düşük frekans bileşenlerini içeren "yaklaşım" (A) ve yüksek frekans bileşenlerini içeren "detay" (D) kısımlarına, yani altbantlarına, ayrılır (Şekil 2.5).



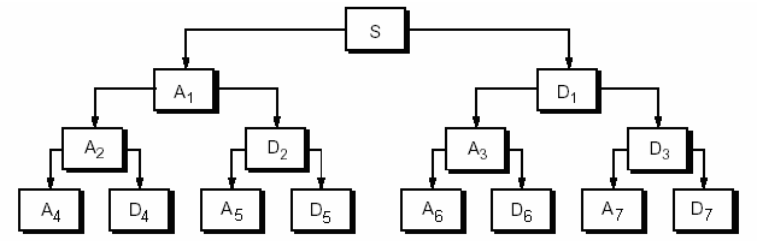
Şekil 2.5. Dalgacık dönüşümünde süzgeçleme işlemi

Daha sonra, sinyalin yaklaşım bileşeni aynı işlem ile tekrar altbantlarına ayırılır ve bu işleme, istenen çözünürlüğe ulaşıncaya kadar devam edilir (Bkz. Şekil 2.6).



Şekil 2.6. Standart dalgacık analizi

Tam altbant ayrışımı elde etmek için ise dalgacık paket analizi ile sinyalin yaklaşım kısmının yanında detay kısmına da dönüşüm uygulanır (Şekil 2.7). Bu işlem, düşük frekans bileşenlerinin yanısıra yüksek frekans bileşenlerinde de önemli bilgiler taşıyan sinyal türlerinin analizinde oldukça faydalıdır.

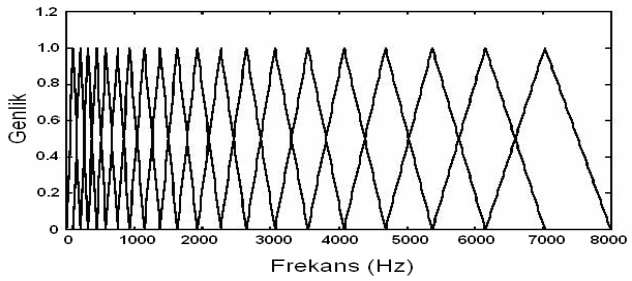


Şekil 2.7. Dalgacık paket analizi

İki dönüşüm tipi işlem karmaşıklığı açısından incelendiğinde ise $O(n)$ karmaşıklığa sahip dalgacık dönüşümü, $O(n \log n)$ karmaşıklıklı Fourier dönüşümüne göre öne çıkmaktadır.

2.2 Dalgacık Dönüşümü Temelli Öznitelik Çıkarımı

Öncelikle, ses ya da konuşma sinyallerinin hem düşük hem de yüksek frekans bileşenlerinde önemli bilgiler taşıyabileceği düşünülerek standart dalgacık dönüşümünden ziyade paket analizi daha uygun olacaktır. Bu noktada, Mel ölçekli süzgeç bankası (Şekil 2.8) insan işitme sistemini en iyi temsil eden yapı olarak bilindiği için (Davis and Mermelstein, 1980) bu ölçekteki altbant ayrıştırması, doğrusal paket analizinden çok daha faydalı olabilir. Zira, Fourier temelli MFCC öznitelikleri de bu ölçeği kullanmaktadır. Bu şekildeki altbant ayrıştırmasının sınıflandırma başarımı üzerindeki faydaları geçmişte yapılan çalışmalarda da gözlenmiştir (Farooq and Datta, 2001; Ricotti, 2005; Günal and Edizkan, 2006; Günal and Edizkan, 2007). Dolayısıyla, bu tez çalışmasında önerilen dalgacık temelli öznitelik çıkarım yönteminde 24 bantlı Mel ölçeği tercih edilmiştir.



Şekil 2.8. Mel ölçekli süzgeç bankası örneği

Dalgacık dönüşümü temelli öznitelik çıkarım yönteminde şu adımlar izlenir:

i) İlk olarak sinyal, daha kısa süreli, dolayısıyla daha durağan bir yapı elde etmek için belirli sayıda çerçeveye bölünür ve öznitelik çıkarım işlemi her çerçeve için ayrı ayrı yürütülür. Bu noktada, aynı sınıfa ait farklı sinyaller için farklı uzunluklar söz konusu olabileceğinden çerçeve sayısı sabit, çerçeve uzunluğu ise değişken tutulmuştur. Benzer yaklaşımlar (Edizkan et al., 2005; Günal and Edizkan, 2007) çalışmalarında da mevcuttur. Böylece, her sinyal için eşit sayıda

öznitelik çıkarılabilmektedir. Değişken çerçeve uzunluğundan kaynaklanabilecek olası problemler de dalgacık dönüşümünün etkin zaman – frekans lokalizasyon özelliği sayesinde giderilebilir. Bu aşamadaki diğer bir nokta ise çerçeve sayısının belirlenmesidir. Bu sayıya, sinyallerin ortalama uzunluğu gözönüne alınarak karar verilmelidir. Daha yüksek sınıflandırma başarımı elde edebilmek için çözünürlüğü artırmak gerektiğinden, daha fazla sayıda çerçeve kullanılmalıdır. En uygun çerçeve sayısı, sinyal çeşidine göre deneysel olarak belirlenebilir (Günal and Edizkan, 2007).

ii) Her bir çerçeve için, 24 bantlı Mel ölçeğine uygun şekilde dalgacık dönüşümü uygulanır. Bu ayrıştırma sonucunda 24 altbant ve bu altbantlara ait dalgacık katsayıları elde edilir. Çizelge 2.1'de 8 kHz örnekleme frekansı için, ayrıştırılmış 24 altbanta ait frekans bilgisi görülmektedir.

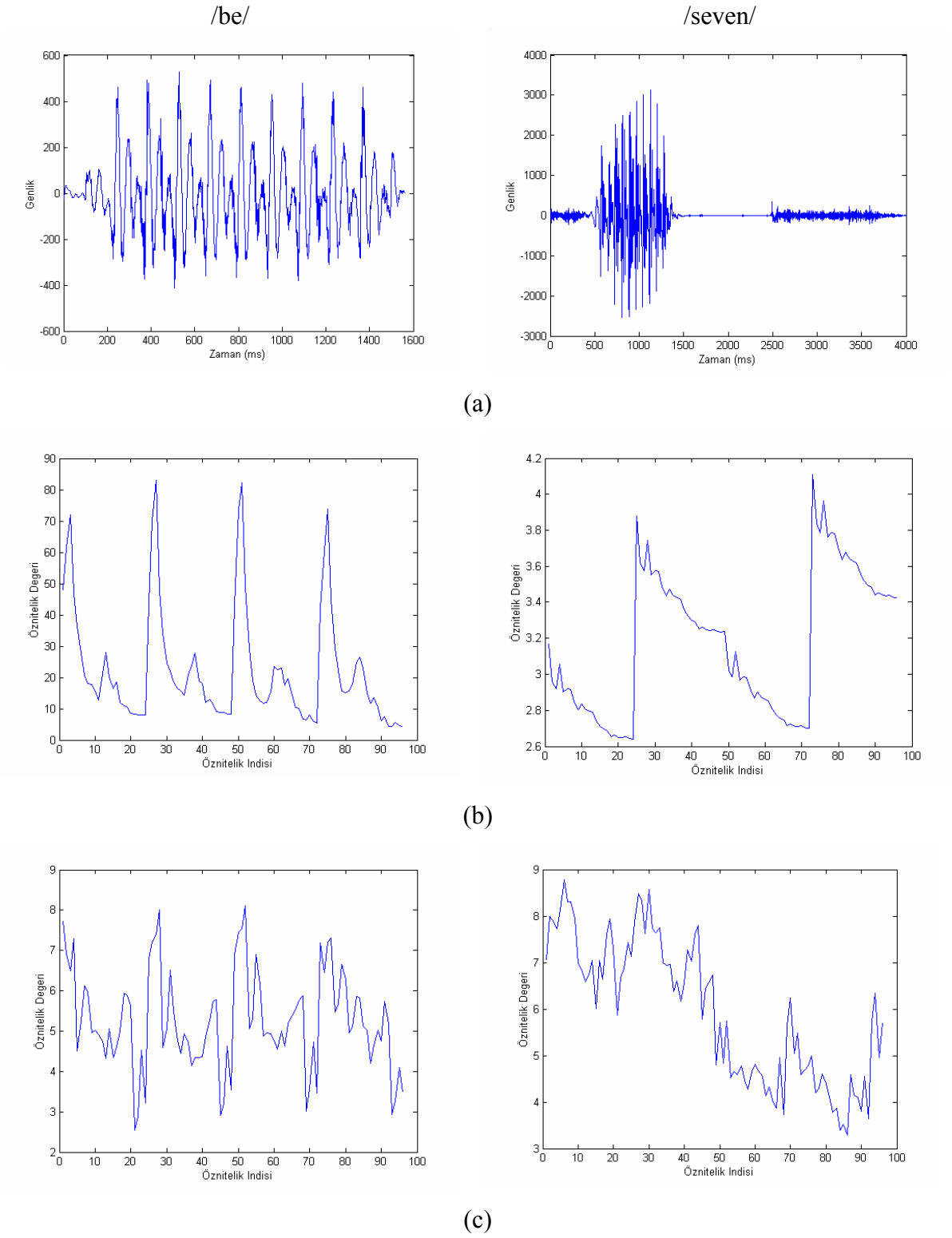
iii) 24 altbantın her birinde dalgacık katsayılarının log enerjileri hesaplanarak her bir çerçeve için 24 skaler değer bulunur. Böylelikle, her çerçeve kendi altbantlarındaki enerjiler ile temsil edilir.

iv) Son olarak, her bir çerçeve için bulunan 24 değer biraraya getirilir ve ilgili sinyalin tamamını temsil eden, (çerçeve sayısı × 24) uzunluğunda satır ya da sütun öznitelik vektörü elde edilir.

Şekil 2.9'da, farklı yapıdaki iki konuşma sinyaline ait zaman bölgesi görüntüsü, 4 çerçeve için elde edilmiş Fourier ve dalgacık temelli öznitelikler görülmektedir.

Çizelge 2.1. Altbant frekans bilgisi

Altbant	Frekans Aralığı (Hz)			Bant Genişliği (Hz)
1	0	-	62,5	62,5
2	62,5	-	125	62,5
3	125	-	187,5	62,5
4	187,5	-	250	62,5
5	250	-	312,5	62,5
6	312,5	-	375	62,5
7	375	-	437,5	62,5
8	437,5	-	500	62,5
9	500	-	562,5	62,5
10	562,5	-	625	62,5
11	625	-	687,5	62,5
12	687,5	-	750	62,5
13	750	-	875	125
14	875	-	1000	125
15	1000	-	1125	125
16	1125	-	1250	125
17	1250	-	1375	125
18	1375	-	1500	125
19	1500	-	1750	250
20	1750	-	2000	250
21	2000	-	2500	500
22	2500	-	3000	500
23	3000	-	3500	500
24	3500	-	4000	500



Şekil 2.9. Farklı yapıdaki iki konuşma sinyali: /be/ ve /seven/
(a) Zaman bölgesi görüntüsü (b) Fourier öznitelikleri (c) Dalgacık öznitelikleri

BÖLÜM 3

ÖZNİTELİK SEÇİMİ

Giriş bölümünde de ifade edildiği üzere öznitelik seçim işleminde, önceden tanımlanmış bir öznitelik kümesi içerisinden ayırtedici öznitelikler seçilip ilgisiz olanlar elenerek başlangıçtakine oranla daha yüksek ayırdediciliğe sahip ve daha düşük boyutlu bir altküme bulunması hedeflenir. Bu durum, özellikle görüntü ve ses tanıma gibi yüksek sayıda öznitelik kullanan uygulamalar açısından büyük önem taşımaktadır.

Öznitelik seçimiyle sağlanan boyut indirgeme sayesinde hem işlem süresinden hem de örüntü tanıma uygulamasının çalıştırılacağı sistem belleğinden tasarruf edilir. Bu avantajların yanısıra "boyutun laneti" (Bellman, 1961; Jain and Zongker, 1997) etkisi de azaltılmış olur. Bu etki, öznitelik kümesi boyutunun artırılmasının başlangıçta sınıflandırma hassasiyetini de artırdığını, ancak belli bir noktadan sonra eğitim kümesindeki veri seyrekliğinin çoğalmasıyla dağılımların iyi temsil edilemediğini; dolayısıyla, sınıflandırma başarımının düşeceğini ifade eder. Bu sebeple, öznitelik kümesi boyutuyla eğitim kümesindeki örnek sayısı arasında uygun bir denge sağlanmalıdır. (Jain and Chandrasekaran, 1982) çalışmasında, eğitim kümesindeki örnek sayısının, öznitelik kümesi boyutunun en az 5 ila 10 katı kadar olması gerektiği vurgulanmıştır. Bu yüzden, öznitelik seçim işlemi, özellikle limitli sayıda örneğe sahip veritabanlarında bu dengeyi sağlayabilmek açısından oldukça önemlidir.

Öznitelik seçimi için literatürde süzgeçlemeden, sarıcı yaklaşımlara (Kohavi and John, 1997) kadar pek çok yöntem bulunmaktadır. Bu yöntemlerin arasından yalnızca tam kapsamlı arama (ES) ve dal-ve-sınır yöntemi (tekdüze ölçüt fonksiyonu kullanılması durumunda) (Narendra and Fukunaga, 1977) eniyi öznitelik altkümesine ulaşmayı garantiler. Ancak, iki yöntem de küçük ve orta büyüklükteki öznitelik kümeleri için bile oldukça yüksek işlem süresine ihtiyaç duyar. Bu durum, nispeten daha kısa sürelerde sonuç veren alt eniyi öznitelik seçim yöntemlerini öne çıkarır. Yaygın olarak kullanılan alt eniyi seçim yöntemlerinden bazıları şunlardır: Bireysel En

İyi Öznitelik Seçimi, Ardışık İleri Yönde Seçim (SFS) (Whitney, 1971), Ardışık Geri Yönde Seçim (SBS) (Marill and Green, 1963), Artı l – Çıkar r Seçim (PTA) (Stearns, 1976), Ardışık İleri Yönde Kayan Seçim (SFFS), Ardışık Geri Yönde Kayan Seçim (SBFS) (Pudil et al., 1994), ve Genetik Seçim (Siedlecki and Sklansky, 1989; Yang and Honavar, 1998). Literatürde bu yöntemleri kullanan çok sayıda öznitelik seçme çalışması bulunmaktadır (Kavzoglu and Mather, 2000; Lai et al., 2006; Reyes-Aldasoro and Bhalerao, 2006; Rokach, 2008; Uncu and Turksen, 2007). Ayrıca, farklı seçim yöntemlerinin başarımlarını karşılaştıran çalışmalar da bulunmaktadır (Jain and Zongker, 1997; Kudo and Sklansky, 2000; Guyon and Elisseeff, 2003).

Yukarıda bahsi geçen öznitelik seçim yötemlerinin tamamı, özniteliklerin ilgililiğine ya da ilgisizliğine karar verebilmek için bir ölçüt fonksiyonu (ayrılabilirlik ölçüsü) kullanır. Yaygın olarak tercih edilen ölçütler, hatalı sınıflandırma olasılığı, ve Uzaksaklık, Dönüşmüş Uzaksaklık, Bhattacharyya, Jeffries – Matusita gibi olasılıksal uzaklık ölçüleridir (Webb, 2002; Theodoridis and Koutroumbas, 2003). Hatalı sınıflandırma olasılığı, sınıflandırıcı türüne bağlı bir ölçüttür. Bu ölçüt ile özniteliklerin seçiminde, hatalı sınıflandırma olasılığını enküçükleyecek öznitelik altkümesine ulaşılması hedeflenir. Olasılıksal uzaklık ölçüleri ise tanımada kullanılacak sınıflandırıcı hakkında bir ön bilgiye ihtiyaç duymazlar. Bu ölçüler ile yapılan öznitelik seçiminde, sınıflar arasındaki olasılıksal uzaklığı enbüyükleyecek öznitelik altkümesine ulaşılmaya çalışılır.

Özniteliklerin değerlendirilmesi, her bir özniteliği bireysel olarak inceleyen "tek-değişkenli" yaklaşımlarla veya öznitelikler arasındaki olası ilintileri göz önüne alarak özniteliklerin tümünü bir grup şeklinde inceleyen "çok-değişkenli" yaklaşımlar ile yapılabilir. Tek-değişkenli yaklaşımlar, öznitelik ilintilerini değerlendirmemekle birlikte özniteliklerin bireysel ayırdedicilik güçlerini gösterebilir ve çok-değişkenli yaklaşımlara kıyasla çok daha hızlı çalışırlar. Öznitelikler arasındaki ilintilerin düşük olduğu ya da toplam sınıf ayrılabilirliğine hangi özniteliklerin daha çok hangilerinin daha az katkı yaptığının bilinmesi gerektiği örüntü tanıma uygulamalarında, tek değişkenli yaklaşımlar tercih edilmelidir. (Su and Lee, 1994; Khan et al., 2001; Tibshirani et. al, 2002; Günal et al., 2006), bu yönde yapılmış çalışmalardan bazılarıdır.

Daha önce belirtilmiş olan alt eniyi seçim yöntemlerinin büyük çoğunluğu çok-değişkenli yaklaşımlardır. Bununla birlikte, bu yöntemlerin ölçüt fonksiyonu olarak kullanılan olasılıksal uzaklık ölçüleri, tek başlarına kullanıldığında, özniteliklerin bireysel ayırdediciliklerini gösterebilmekte; dolayısıyla, direkt olarak tek-değişkenli yaklaşımlar şeklinde kullanılabilmektedir. (Su and Lee, 1994; Theodoridis and Koutroumbas, 2003; Günal vd., 2008) çalışmaları, bu tarz uygulamaların bazı örneklerini içermektedir.

Uzaksaklık ve Bhattacharyya gibi klasik ayrılabilirlik ölçülerinin temelinde Bayes hatasını enküçükleme kuralı yatmaktadır (Theodoridis and Koutroumbas, 2003). Ancak, altuzay analizi de uygun altuzaylarda varolan bilgiyi kullanarak özniteliklerin ayrılabilirlik derecelerini gösterebilir. Bu doğrultuda, tez çalışmasının bu bölümünde özniteliklerin bireysel ayırdediciliklerini gösteren, alt uzay temelli iki yeni ayrılabilirlik ölçüsü geliştirilmiştir. Bu ölçüler, Ortak Altuzay (CS) ve Fisher Altuzay (FS) ölçüsü olarak adlandırılmaktadır. Geliştirilen bu ölçülerin çok-sınıflı örüntü uygulamalarında öznitelik seçimi için kullanılabilmesi amacıyla da ayrıca bir yöntem önerilmektedir. Bölüm 5'te yer alan deneysel çalışmalarda, altuzay temelli yeni ölçüler ve klasik ölçülerin tek-değişkenli formları ile öznitelik seçme işlemleri farklı veritabanları üzerinde, sınıflandırma hassasiyeti ve boyut indirgeme açısından karşılaştırılmıştır. Ayrıca, tek-değişkenli ve çok-değişkenli öznitelik seçme yöntemleri arasında çeşitli açılardan kıyaslamalar yapılmıştır.

Altbölümlerde, öncelikle sınıf ayrılabilirliğini belirlemede yaygın olarak kullanılan olasılıksal uzaklık ölçüleri, eniyi ve alt eniyi öznitelik seçim yöntemleri anlatılmıştır. Daha sonra, tez kapsamında geliştirilmiş olan altuzay temelli yeni ayrılabilirlik ölçüleri tanıtılmış ve tek-değişkenli yaklaşımların çok-sınıflı uygulamalarda öznitelik seçiminde kullanılabilmesi için önerilen yöntem verilmiştir.

3.1 Olasılıksal Uzaklık Ölçüleri

Bu bölümde, öznitelik seçme yöntemlerinde ölçüt olarak kullanılan Uzaksaklık, Bhattacharyya, Dönüşmüş Uzaksaklık ve Jeffries-Matusita olasılıksal uzaklık ölçüleri tanıtılmaktadır.

3.1.1 Uzaksaklık

Bayes karar kuralı doğrultusunda, (c_i, c_j) sınıf çifti ve öznitelik vektörü x için

$$P(c_i \mid x) > P(c_j \mid x) \tag{3.1}$$

koşulu sağlanıyorsa, x'in c_i sınıfına ait olduğu kabul edilir (Duda et al., 2001; Theodoridis and Koutroumbas, 2003). Dolayısıyla, hatalı sınıflandırma olasılığı, bu iki koşullu olasılığın farkına bağlıdır. Bu sebeple,

$$\frac{P(c_i \mid x)}{P(c_j \mid x)} \tag{3.2}$$

oranı, öznitelik vektörü x'in iki sınıf arasındaki ayırdedicilik derecesine dair önemli bilgiler taşıyabilir. Benzer şekilde,

$$d_{ij}(x) \equiv \ln \frac{P(x \mid c_i)}{P(x \mid c_j)} \tag{3.3}$$

oranı içinde bulunan aynı bilgi, c_i sınıfının c_j sınıfına göre ayırdedici bilgisinin bir ölçüsü olarak kullanılabilir (Kailath, 1967; Webb, 2002; Theodoridis and Koutroumbas, 2003). Sınıfların tamamen örtüşmesi durumunda,

$$d_{ij}(x) = 0 \tag{3.4}$$

eşitliği geçerlidir. Öznitelik vektörü x farklı değerler alabileceği için, c_i ve c_j sınıfları üzerindeki ortalama değeri dikkate alınmalıdır:

$$\begin{aligned} d_{ij}(x) &= \int_{-\infty}^{+\infty} p(x \mid c_i) \ln \frac{P(x \mid c_i)}{P(x \mid c_j)} dx \\ d_{ji}(x) &= \int_{-\infty}^{+\infty} p(x \mid c_j) \ln \frac{P(x \mid c_j)}{P(x \mid c_i)} dx \end{aligned} \tag{3.5}$$

Bu durumda,

$$D_{ij} = d_{ij} + d_{ji} \tag{3.6}$$

toplamı uzaksaklık olarak isimlendirilir ve sınıfların öznitelik vektörü x'e göre bir ayrılabilirlik ölçüsü olarak kullanılabilir (Theodoridis and Koutroumbas, 2003). Çok sınıflı problemlerde, uzaksaklık her bir sınıf çifti için hesaplanmalıdır.

Yoğunluk fonksiyonlarının Gauss tipinde olduğu kabul edilirse, uzaksaklık,

$$D_{ij} = \frac{1}{2} trace(\Phi_i^{-1}\Phi_j + \Phi_j^{-1}\Phi_i - 2I) + \frac{1}{2}(\mu_i - \mu_j)^T(\Phi_i^{-1} + \Phi_j^{-1})(\mu_i - \mu_j) \tag{3.7}$$

eşitliğiyle daha basit bir şekilde hesaplanabilir. Burada, Φ_i ve Φ_j, sınıfların ortak değişinti matrislerini; μ_i ve μ_j, sınıf ortalama vektörlerini; trace işlemi ise ilgili matrisin ana köşegen öğelerinin toplamını ifade etmektedir. Bir boyutlu durumda bu formül,

$$D_{ij} = \frac{1}{2}\left(\frac{\sigma_j^2}{\sigma_i^2} + \frac{\sigma_i^2}{\sigma_j^2} - 2\right) + \frac{1}{2}(m_i - m_j)^2\left(\frac{1}{\sigma_i^2} + \frac{1}{\sigma_j^2}\right) \tag{3.8}$$

haline indirgenir. Burada, σ^2 ve m, ilgili sınıflardaki verinin sırasıyla değişinti ve ortalama değerini temsil etmektedir. Bu formül yardımıyla, çok boyutlu bir öznitelik vektörünün her bir boyutuna ait uzaksaklık ölçüsü, ilgili sınıf çiftleri için bireysel olarak hesaplanabilir.

3.1.2 Bhattacharyya

Bhattacharyya, sınıfların ayrılabilirliğini hesaplamak için kullanılan diğer bir olasılıksal uzaklık ölçüsüdür. Bhattacharyya, uzaksaklık tipinde bir ölçüdür. Aynı X tanım kümesindeki olasılık dağılımları $p(x)$ ve $q(x)$ için Bhattacharyya katsayısı,

$$b(p,q) = \int \sqrt{p(x)q(x)dx} \tag{3.9}$$

şeklinde tanımlanır. Başka bir ifadeyle Bhattacharyya, $x \in X$ noktaların olasılıklarının kare kökü şeklinde bileşenlere sahip iki vektörün skalar çarpımı şeklinde ifade edilir. Bu sebeple, geometrik yorumlamaya uygundur: Bhattacharyya, bu iki vektör arasındaki açının kosinüsüne eşittir. Bhattacharyya uzaklığı (B) ise $(-\ln b)$ şeklinde hesaplanmakta ve $0 \le B \le \infty$ aralığında değer alabilmektedir (Kailath, 1967; Theodoridis and Koutroumbas, 2003).

Gauss dağılımı için c_i ve c_j sınıflarına ait Bhattacharyya uzaklık ölçüsü

$$B_{ij} = \frac{1}{8}(\mu_i - \mu_j)^T \left(\frac{\Phi_i + \Phi_j}{2} \right)^{-1} (\mu_i - \mu_j) + \frac{1}{2} \ln \frac{\left| (\Phi_i + \Phi_j)/2 \right|}{\sqrt{\left|\Phi_i\right| \left|\Phi_j\right|}} \tag{3.10}$$

şeklinde tanımlanır. Burada, Φ_i ve Φ_j, sınıfların ortak değişinti matrislerini; μ_i ve μ_j ise sınıf ortalama vektörlerini ifade etmektedir. Tek boyutlu durum için, bu eşitlik

$$B_{ij} = \frac{1}{8}(m_i - m_j)^2 \left(\frac{2}{\sigma_i^2 + \sigma_j^2} \right) + \frac{1}{2} \ln \frac{\left| (\sigma_i^2 + \sigma_j^2)/2 \right|}{\sqrt{\sigma_i^2 \sigma_j^2}} \tag{3.11}$$

haline indirgenir. Burada, σ^2 ve m, ilgili sınıflardaki verinin sırasıyla değişinti ve ortalama değerini temsil etmektedir.

3.1.3 Dönüşmüş uzaksaklık

Dönüşmüş uzaksaklık ölçüsü, sınıfların birbirlerinden belirgin şekilde ayrık olması durumunda, ortalama uzaksaklığın artması ve uzaksaklık ölçüsünün yanıltıcı olma etkisini azaltmak için kullanılır (Kavzoglu and Mather, 2000). Başka bir ifadeyle, dönüşmüş uzaksaklık, klasik uzaksaklığı belli bir değer aralığına ölçekler. Bu ölçü,

$$TD_{ij} = c \left[1 - e^{\frac{-D_{ij}}{8}} \right] \tag{3.12}$$

şeklinde tanımlanır. Bu formülde, c değeri, dönüşmüş uzaksaklığın değer aralığını belirleyen sabiti temsil etmektedir.

3.1.4 Jeffries-Matusita

Uzaksaklık ile dönüşmüş uzaksaklık arasındaki ilişkiye benzer şekilde, Jeffries-Matusita, Bhattacharyya uzaklık değerini belli bir aralığa ölçekler (Kavzoglu and Mather, 2000). Jeffries-Matusita ölçüsü, yüksek ayrılabilirlik değerlerini bastırırken düşük olanları vurgulama eğilimine sahiptir. Bu ölçü, şu şekilde tanımlanmaktadır:

$$JM_{ij} = \sqrt{2(1 - e^{-B_{ij}})} \tag{3.13}$$

3.2 En İyi Öznitelik Seçim Yöntemleri

Eniyi öznitelik altkümesine ulaşmayı garanti eden iki yöntemden biri tam kapsamlı arama diğeri ise tekdüze ölçüt fonksiyonu kullanılması koşuluyla dal-ve-sınır yöntemidir. Bu bölümde, iki yöntem de kısaca anlatılmaktadır.

3.2.1 Tam kapsamlı arama

Bu yöntemde, N elemanlı bir öznitelik kümesinden en iyi sonucu veren d elemanlı bir altküme elde etmek için $\binom{N}{d}$ adet olası altkümelerin tamamı incelenir. Tam kapsamlı arama en iyi sonucu vermesine rağmen, işlem süresi orta büyüklükteki öznitelik kümeleri için bile oldukça uzundur. Bu kısıtlayıcı faktörden dolayı tam kapsamlı arama yöntemi, öznitelik seçme işlemlerinde çok fazla tercih edilmemektedir.

3.2.2 Dal ve sınır

Narendra ve Fukunaga tarafından geliştirilen bu yöntem (Narendra and Fukunaga, 1977), eniyi sonuca tam kapsamlı aramadan daha kısa sürede ulaşmaktadır. Ancak, yöntemin en iyi sonuca ulaşabilmesi için ölçüt fonksiyonu tekdüze olmalıdır. Başka bir ifadeyle, öznitelik altkümeleri X ve Y için,

$$X \subset Y \Rightarrow J(X) < J(Y) \tag{3.14}$$

durumu sağlanmalıdır. Buna göre, bir altkümeye yeni bir özniteliğin eklenmesiyle elde edilen daha yüksek boyutlu yeni altkümeye ait ölçüt fonksiyonu değeri mutlaka daha büyük olmalıdır. Aksi takdirde, bu yöntem ile eniyi çözüme ulaşılması mümkün

olamaz. Geniş öznitelik kümeleri söz konusu olduğunda, işlem süresi açısından bu seçim yöntemi de çok elverişli değildir.

3.3 Alt En İyi Öznitelik Seçim Yöntemleri

Çok uzun işlem sürelerine ihtiyaç duyan eniyi öznitelik seçim yöntemlerine alternatif olarak kullanılan alt eniyi yöntemler bu bölümde anlatılmaktadır.

3.3.1 Bireysel en iyi öznitelik seçimi

Bu seçim yöntemi, tek değişkenli bir yaklaşımdır. Öznitelikler, belirlenen bir ölçüt fonksiyonuna göre bireysel olarak değerlendirilir ve sıralanır. İstenen sayıda özniteliğin seçiminde ise sıralı listedeki en önemli öznitelikten başlanıp sıradaki diğer özniteliklerle devam edilir. Bu yöntem oldukça hızlı olmasına rağmen öznitelikler arasındaki olası ilintileri değerlendirmediği için her zaman çok etkili olamayabilir. Öznitelik kümesindeki elemanların düşük ilintili ya da ilintisiz olması durumunda ise oldukça iyi sonuçlar alınabilmektedir.

3.3.2 Ardışık ileri yönde seçim

Bu seçim yöntemi, "aşağıdan yukarıya" doğru çalışır. Yöntem, ilk olarak (Whitney, 1971) çalışmasında sunulmuştur. Boş bir öznitelik kümesiyle işleme başlayarak, her bir adımda o anki öznitelik altkümesinin ölçüt fonksiyonu değerini eniyileyen öznitelik, altkümeye eklenir. Bu işlem, istenen öznitelik boyutuna ulaşıncaya kadar tekrarlanır. Ardışık ileri yönde seçim yönteminde, her adımda tek bir öznitelik altkümeye eklenir. Her adımda altkümeye n adet özniteliğin eklendiği yöntem ise Genelleştirilmiş İleri Yönde Seçim'dir (Kittler, 1978). Her iki yöntem de içiçelik

etkisine maruz kalmaktadır. Yani, öznitelik(ler) bir kez seçildikten sonra bir daha kümeden çıkarılamaz. Bu sebepten ötürü iki yöntem de çoğunlukla alt eniyi sonuca ulaşabilmektedir.

3.3.3 Ardışık geri yönde seçim

Bu seçim yöntemi, "yukarıdan aşağıya" doğru çalışır. Yöntem, ilk olarak (Marill and Green, 1963) çalışmasında önerilmiştir. Burada, ardışık ileri yönde seçimin tersi bir durum söz konusudur. Başlangıçta öznitelik kümesinin tamamı göz önüne alınarak, her bir adımda o anki öznitelik altkümesinin ölçüt fonksiyonu değerini eniyileyecek şekilde bir öznitelik kümeden çıkarılır. Çıkarma işlemi, istenen öznitelik boyutuna ulaşıncaya kadar tekrarlanır. Her adımda bir yerine n adet özniteliğin elendiği yöntem ise Genelleştirilmiş Geri Yönde Seçim olarak isimlendirilir (Kittler, 1978). İleri yönde seçim algoritmalarında olduğu gibi bu yöntemlerde de içiçelik etkisi söz konusudur. Seçim sürecinde, öznitelik(ler) kümeden bir kez çıkarıldıktan sonra bir daha giremez. Bu durum, yöntemlerin alt eniyi sonuç vermesine sebep olur.

3.3.4 Artı *l* – çıkar *r* seçim

Ardışık ileri ve geri yönde seçim yöntemlerinin maruz kaldığı içiçelik etkisi, seçim esnasında belli oranda ters yöne hareket ederek kısmen giderilebilir. Bunun için, seçim işleminin bir adımında l adet öznitelik ileri yönde seçim yöntemiyle kümeye eklendikten sonra, r adet öznitelik geri yönde seçim ile kümeden çıkarılır. Bu yöntem, "Artı l – Çıkar r" olarak isimlendirilmiştir (Stearns, 1976). Burada, içiçelik etkisi belli oranda azalmasına rağmen, yöntem halen ileri ve geri yönde seçimi temel aldığı için alt eniyi sonuç verir.

3.3.5 Ardışık ileri yönde kayan seçim

Artı l – Çıkar r seçim yöntemi, içiçelik etkisini azaltmak için öznitelik kümesi üzerinde l ve r parametreleriyle belirlenen miktarda ardışık ekleme ve çıkarma işlemleri uygular. Ancak, en iyi öznitelik kümesini elde edebilmek için bu parametrelerin hangi değeri alması gerektiğini belirleyen teorik bir yöntem mevcut değildir. Bu yüzden, (Pudil et al., 1994) çalışmasında önerilen ardışık ileri yönde kayan seçim yöntemi l ve r parametrelerini sabitlemek yerine kaymalarını sağlar. Böylece seçimin herhangi bir adımında, ölçüt fonksiyonu mevcut değerinden daha iyi bir değere ulaşıncaya kadar aynı yönde hareket edilir. Bu esnek yapı, öznitelik kümesi boyutunun her adımda tekdüze olmayan bir şekilde değişmesini sağlar.

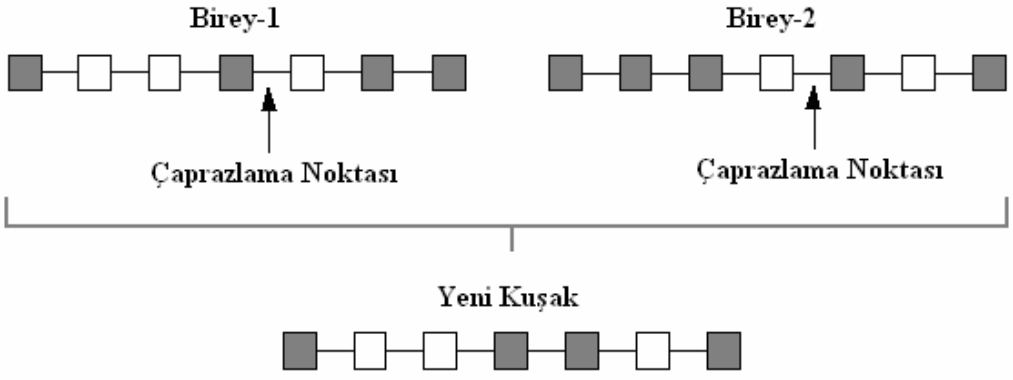
3.3.6 Ardışık geri yönde kayan seçim

Yine (Pudil et al., 1994) çalışmasında önerilmiş olan bu seçim yöntemi, ardışık ileri yönde kayan seçimle aynı prensibe dayanır. Ancak, seçim işlemi bu yöntemde ters yöne doğru çalışmaktadır.

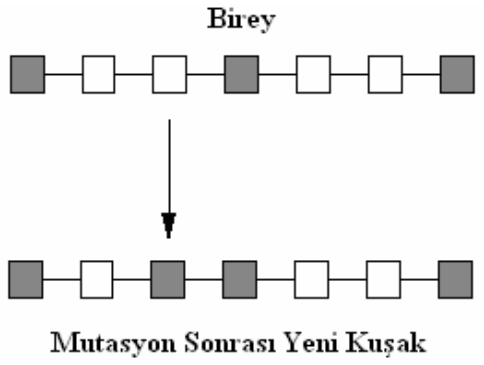
3.3.7 Genetik seçim

Genetik Algoritma (GA), biyolojik evrimleşme sürecini taklit eden stokastik bir arama yöntemidir (Goldberg, 1989). GA, potansiyel çözümler nüfusu arasından en uygun olanların yaşaması ve bunlardan türeyen yeni kuşakların daha iyi çözümler üretmesi prensibine göre çalışır. Çözümler, uygun bir alfabe ile kodlanmış kromozomlara karşılık gelmektedir. Çözümlerin uygunluk derecesi ise tanımlanan bir uyum fonksiyonu ile belirlenir. Yeni kuşaklar, mevcut nüfus üzerine belirli olasılıklarla doğal genetiğe ait çaprazlama ve mutasyon işlemleri uygulanarak türetilir. Çaprazlama işleminde, farklı bireylere ait kromozomlardaki veriler, bir çaprazlama noktası referans

alınarak birleştirilir ve yeni kuşak oluşturulur (Şekil 3.1). Mutasyon işleminde ise bir bireye ait kromozomun içeriği değiştirilerek yeni kuşak oluşturulur (Şekil 3.2).



Şekil 3.1. Çaprazlama işlemi örneği



Şekil 3.2. Mutasyon işlemi örneği

GA, ilk kez (Siedlecki and Sklansky, 1989) çalışmasında öznitelik seçimi için kullanılmıştır. Bu seçim yönteminde kromozomların uzunluğu öznitelik setinin boyutuna eşittir. Kromozomlar {0,1} alfabesi ile kodlanır. "1" ile ifade edilen indisler seçilen öznitelikleri; "0" ile ifade edilenler ise seçilmeyen öznitelikleri temsil eder. Örnek olarak,

$$\{1\,0\,1\,0\,1\,1\,0\,0\,0\,1\,0\,0\,0\,0\,0\,0\,0\,1\,1\,1\} \tag{3.15}$$

biçiminde tanımlanmış 20 boyutlu öznitelik setine ait kromozoma göre 1, 3, 5, 6, 10, 18, 19, ve 20 numaralı öznitelikler seçilmiş, diğerleri ise elenmiştir. Herhangi bir kromozoma ait uyum fonksiyonu değeri ise kromozomda seçilen özniteliklere ait ölçüt fonksiyonu değerine karşılık gelmektedir.

GA temelli seçim yönteminde, nüfus boyutu, kuşak sayısı ve çaprazlama - mutasyon olasılık değerleri deneysel olarak belirlenir. Ölçüt fonksiyonu olarak ise genellikle doğru sınıflandırma olasılığı kullanılır. (Yang and Honavar, 1998; Ishibuchi and Nakashima, 2000; Rokach, 2008), genetik seçime dair bazı örnek çalışmalardır.

3.4 Altuzay Temelli Öznitelik Seçim Yöntemleri

Bu bölümde, özniteliklerin bireysel ayırdediciliklerini belirlemek için tez çalışması kapsamında geliştirilmiş olan iki yeni ayrılabilirlik ölçüsü açıklanmaktadır. Bu ölçüler, Ortak Altuzay (CS) ve Fisher Altuzayı (FS) ölçüsü olarak isimlendirilmiştir. Bu iki ölçünün altında yatan fikir, altuzaylardaki sınıf-içi ve sınıflar-arası ayırdedici bilgilerden faydalanmaktır (Günal and Edizkan, 2008). Altbölümlerde, iki ölçü de detaylı şekilde açıklanmakta ve ilgili formülasyonlar verilmektedir.

3.4.1 Ortak altuzay ölçüsü

Bu ölçü aslen, altuzay temelli ve sınıf-içi ortak değişinti bilgisini kullanan bir örüntü sınıflandırıcı olan Ortak Vektör Yaklaşımı (CVA) yönteminden (Gülmezoğlu et al., 1999; Gülmezoğlu et al., 2001) faydalanmaktır. CVA, bugüne kadar pek çok farklı örüntü tanıma uygulamasında başarıyla kullanılmıştır (Çevikalp et al., 2005; Günal et al., 2006; Gülmezoğlu et al., 2007). Bu yöntemde, sınıf-içi ortak değişintilerin sıfır ya da küçük özdeğerlerine karşılık gelen özvektörler ile bir farksızlık altuzayı oluşturulur.

Bir sınıfın ortalama vektörünün kendi farksızlık altuzayına izdüşümü ise “ortak vektör” adı verilen ve o sınıfı temsil eden eşsiz bir vektör oluşturur. CS ölçüsü, sınıfların bu eşsiz özelliğini kullanır.

Bir öznitelik kümesindeki her bir özniteliğin iki sınıf arasındaki CS ölçüsünü hesaplamak için öncelikle bir sınıfın ortalama vektörünün diğer sınıfın farksızlık altuzayına izdüşümü bulunur. Daha sonra her bir öznitelik için izdüşürülmüş ortalama vektör ile diğer sınıfın ortak vektörünün ilgili indisleri arasındaki uzaklık hesaplanır. Bu noktada, bir özniteliğin iki sınıf için ayırdediciliği fazla ise hesaplanan uzaklık büyük olacaktır. Böylelikle, CS ölçüsünün değeri de büyük çıkacaktır. Tersi durumda, düşük ayırdediciliğe sahip bir öznitelik, izdüşürülmüş ortalama vektörünün diğer sınıfın ortak vektörüne yakın olmasına yol açacak, böylece CS ölçüsü küçük değerli olacaktır (Günal and Edizkan, 2008).

CS ölçüsünün hesaplanması adım adım şu şekilde gerçekleşmektedir:

a_i^c vektörünün, c sınıfına ait d boyutlu i’nci öznitelik vektörünü temsil ettiği düşünülürse, ilgili sınıfa ait ortalama vektörü a_{ave}^c,

$$a_{ave}^c = \frac{1}{m}\sum_{i=1}^{m} a_i^c \tag{3.16}$$

formülüyle bulunur. Burada m değeri, bir sınıftaki öznitelik vektörü sayısını göstermektedir. Daha sonra, sınıf-içi ortak değişinti matrisleri Φ_c,

$$\Phi_c = \sum_{i=1}^{m} (a_i^c - a_{ave}^c)(a_i^c - a_{ave}^c)^T \tag{3.17}$$

formülüyle hesaplanır. Bu matrislerin elde edilmesini takiben, özdeğer ayrıştırması

$$\left(\Phi_c - \lambda_j^c I\right) u_j^c = 0 \qquad j = 1, 2, ..., d \tag{3.18}$$

ile her sınıfın ortak değişinti matrisine ait {özdeğer, özvektör}çiftleri $\left\{\lambda_j^c, u_j^c\right\}$ bulunur ve özdeğerler küçükten büyüğe doğru sıralanır. Sıralanmış özdeğerlerden en küçük z adedine karşılık gelen özvektörler kullanılarak,

$$P_{Common}^c = \sum_{j=1}^{z} u_j^c u_j^{cT} \tag{3.19}$$

formülü yardımıyla her sınıfın farksızlık altuzayına izdüşüm matrisi, P_{Common}^c, hesaplanır. Burada z değeri, küçük özdeğerlerin toplamının tüm özdeğerlerin toplamına oranı belli bir L eşik değerinden küçük olacak şekilde belirlenir.

$$\frac{\sum_{j=1}^{z} \lambda_j^c}{\sum_{j=1}^{d} \lambda_j^c} < L \tag{3.20}$$

Bu eşik değeri çeşitli örüntü tanıma uygulamalarında %5 ile %10 arasında değişmektedir (Oja, 1983; Günal et al., 2006; Gülmezoğlu et al., 2007). En uygun değer, deneysel olarak belirlenmelidir.

Farksızlık altuzayına izdüşüm matrislerinin elde edilmesinden sonra,

$$a_{ave}^{c,k} = P_{Common}^k a_{ave}^c \tag{3.21}$$

formülü ile sınıf ortalamalarının ayrı ayrı her bir sınıfın farksızlık altuzayına izdüşümü hesaplanır. Burada $a_{ave}^{c,k}$ vektörü, a_{ave}^c vektörünün k sınıfına izdüşümünü ifade etmektedir. Doğal olarak $c = k$ durumunda bu vektör, ilgili sınıfın ortak vektörüne karşılık gelir.

Bu işlemlerin sonucunda, öznitelik kümesindeki her bir öznitelik için (c,k) sınıf çiftine dair CS ölçüsü,

$$CS_{ck}(j) = \left\| a_{ave}^{c,k}(j) - a_{ave}^{k,k}(j) \right\| \qquad j = 1,...,d \tag{3.22}$$

formülü yardımıyla, izdüşürülmüş ortalama vektörlerin ilgili indisleri (j) arasındaki öklid uzaklığı hesaplanarak elde edilir.

3.4.2 Fisher altuzayı ölçüsü

FS ölçüsü, adından da anlaşılacağı üzere, gerek sınıf-içi gerekse sınıflar-arası ortak değişinti bilgilerinden faydalanan Fisher'in Doğrusal Ayırtaç Analizi (FLDA) ölçütünü temel almaktadır. FLDA, temel olarak öznitelik boyutundan daha düşük boyuttaki bir altuzayda, aynı sınıfa ait örneklerin birbirine yakın, farklı sınıflardaki örneklerinse birbirinden uzak olmalarını sağlayacak şekilde bir dönüşüm elde etmeyi hedefler (Duda et al., 2001; Webb, 2002; Theodoridis and Koutroumbas, 2003). FS ölçüsü, işte bu altuzayı kullanarak özniteliklerin ayırtedicilik derecelerini tespit eder.

Bu bağlamda, bir öznitelik kümesindeki özniteliklerin iki sınıfa dair FS ayrılabilirlik ölçülerini hesaplayabilmek için öncelikle sınıfların ortalama vektörlerinin Fisher altuzayına izdüşümü alınır. Daha sonra, her bir öznitelik için iki sınıfın izdüşürülmüş ortalama vektörlerinin ilgili indisleri arasındaki uzaklık ölçülür. Bu uzaklık FS ölçüsünü temsil eder. Aynen CS ölçüsünde olduğu gibi, yüksek ayırdediciliğe sahip öznitelikler için ölçülen uzaklık fazla olacak; böylece, FS ölçüsünün değeri de yüksek çıkacaktır. Tersi durumda ise düşük ayırdediciliğe sahip bir öznitelik, izdüşürülmüş ortalama vektörünün ilgili indisindeki değerin diğer sınıfa ait ilgili değere yakın olmasına yol açacak; dolayısıyla, FS ölçüsünün değeri küçük olacaktır (Günal and Edizkan, 2008).

FS ölçüsünün hesaplanması adım adım şu şekilde gerçekleşir:

Öncelikle, (3.17) denklemi yardımıyla sınıf-içi ortak değişinti matrisleri (Φ_c) bulunur. Daha sonra, bütün sınıflara ait ortak değişinti matrisleri toplanarak toplam sınıf-içi ortak değişinti matrisi (Φ_w) elde edilir:

$$\Phi_w = \sum_{c=1}^{n} \Phi_c \tag{3.23}$$

Burada, n değeri toplam sınıf sayısını ifade etmektedir. Bu işlemi takiben, sınıfların ortalama vektörlerini (a_{ave}^{c}) ve genel sınıf ortalama vektörünü (a_{ave_all}) kullanarak,

$$\Phi_b = \sum_{i=1}^{n} (a_{ave}^{c} - a_{ave\_all})(a_{ave}^{c} - a_{ave\_all})^T \tag{3.24}$$

formülü yardımıyla sınıflar-arası ortak değişinti matrisi (Φ_b) bulunur. Sonraki adımda, FLDA oranı doğrultusunda,

$$\Phi_f = \Phi_w^{-1}\Phi_b \tag{3.25}$$

denklemi ile Φ_f matrisi hesaplanır ve özdeğer ayrıştırması ile bu matrise ait özdeğer – özvektör çiftleri (λ_j, u_j) elde edilir. Ayrıştırma sonrasında özdeğerler sıralanarak en büyük l adet özdeğere karşılık gelen özvektörler ile Fisher altuzayına izdüşüm matrisi elde edilir:

$$P_{Fisher} = \sum_{j=1}^{l} u_j u_j^{\ T} \tag{3.26}$$

Bu işlemi takiben, sınıfların ortalama vektörlerinin Fisher altuzayına izdüşümü ($a_{ave}^{c,Fisher}$) hesaplanır:

$$a_{ave}^{c,Fisher} = P_{Fisher} a_{ave}^{c} \tag{3.27}$$

Son olarak, öznitelik kümesindeki her bir öznitelik için (c,k) sınıf çiftine dair FS ölçüsü,

$$FS_{ck}(j) = \left\| a_{ave}^{c,Fisher}(j) - a_{ave}^{k,Fisher}(j) \right\| \qquad j = 1,...,d \tag{3.28}$$

formülü yardımıyla, izdüşürülmüş ortalama vektörlerin ilgili indisleri (j) arasındaki öklid uzaklığı hesaplanarak elde edilir.

3.5 Çok Sınıflı Öznitelik Seçimi

Yukarıda detayları açıklanmış olan CS ve FS ölçüleri, klasik ayrılabilirlik ölçülerinde olduğu gibi özniteliklerin sadece sınıf çiftleri arasındaki ayırdedicilik bilgisini sağlamaktadır. İki sınıflı öznitelik problemlerinde tüm öznitelikler, kullanılan ayrılabilirlik ölçüsü ile belirlenmiş bireysel ayırdedicilik derecelerine göre sıralanır. Daha sonra, bu sıralama üzerinden istenen sayıda öznitelik seçilir. Ancak, bu ölçüler ve seçme yöntemi, çok sınıflı uygulamalarda direkt olarak kullanılamaz.

Ayrılabilirlik ölçülerinin çok sınıflı duruma uyarlanmasında izlenebilecek bir yol, tüm sınıf çiftleri için ayrılabilirliklerin hesaplanması ve bunların ortalamasının alınmasıdır. Fakat bu yaklaşım, sınıflar arasındaki değişintilerin farklı olduğu durumlarda çok faydalı olmayabilir. Bunun yerine, en kötü sınıf dağılım senaryosu düşünülerek her bir özniteliğin ayırtediciliği, olası bütün sınıf çiftleri için elde edilmiş olan ayrılabilirlik ölçülerinin en küçüğü ile temsil edilir:

$$M_{\min}(j) = \min\left\{ M_{ck}(j) \right\} \tag{3.29}$$

Bu denklemde $M_{ck}(j)$, (c,k) sınıf çifti için öznitelik kümesindeki j indisli özniteliğe ait ayrılabilirlik ölçüsünü, $M_{\min}(j)$ ise ilgili özniteliğin çok sınıflı ayrılabilirlik ölçüsünü ifade etmektedir. Başka bir deyişle $M_{\min}(j)$ ölçüsü, ilgili

özniteliğin tüm sınıflar için genel ayırdedicilik gücünü göstermektedir (Günal and Edizkan, 2008). Bu çözüme benzer yaklaşımlar, (Su and Lee, 1994; Theodoridis and Koutroumbas, 2003; Günal vd., 2008) çalışmalarında da mevcuttur.

Her bir özniteliğe ait çok sınıflı ayrılabilirlik ölçüsü hesaplandıktan sonra tüm öznitelikler sahip oldukları ölçülere göre büyükten küçüğe doğru sıralanır. Böylelikle, istenen sayıda öznitelik seçilirken en büyük ölçüye sahip öznitelikten başlanıp sıralamadaki diğer öznitelikler ile devam edilir. Yüksek ayrılabilirlik ölçüsüne sahip özniteliklerin seçilmesiyle ilgisiz öznitelikler elenmiş ve sınıflar arasındaki ayrıştırma artırılmış olur.

BÖLÜM 4

SINIFLANDIRMA

Sınıflandırma temel olarak, bilinmeyen bir örüntünün, o örüntüye ait öznitelikler kullanılarak bir karar mekanizması (sınıflandırıcı) yardımıyla hangi sınıfa ait olduğunun belirlenmesi şeklinde tanımlanabilir.

Örüntü tanıma sistemlerinin başarımında, sınıflandırma yönteminin rolü de en az örüntü çıkarma ve seçme kadar büyüktür. Literatürde, güdümlü ve güdümsüz pek çok türde sınıflandırma yöntemi bulunmaktadır. Yaygın olarak kullanılan yöntemlerden bazıları Bayes, doğrusal sınıflandırıcılar, doğrusal olmayan sınıflandırıcılar, yapay sinir ağları, saklı Markov modeller, çekirdek yöntemleri, karar ağaçları ve topak analizidir (Duda et al., 2001; Schölkopf and Smola, 2001; Webb, 2002; Theodoridis and Koutroumbas, 2003). Ancak, örüntü tanıma problemlerinde genel olarak hangi sınıflandırma yönteminin en iyi ya da en uygun olduğu kesin olarak söylenemez. Örüntünün çeşidi, örüntüden elde edilen özniteliklerin yapısı ve sayısı, işlem süresi ve karmaşıklığı gibi faktörlere göre tercih edilmesi gereken sınıflandırma yöntemi farklılık gösterebilir.

Bir sınıflandırıcının başarımını ölçmek için sınıflandırılacak örüntülere ait belirli sayıda örnek içeren "veritabanı" öncelikle "eğitim" ve "test" kümesi şeklinde ikiye ayrılmalıdır. Daha sonra, eğitim kümesindeki öznitelikler kullanılarak elde edilen karar ya da sınıflandırma kuralı, test kümesine uygulanıp sınıflandırma hata olasılığı hesaplanır. Buradaki amaç, hangi sınıfa ait olduğu bilinen özniteliklerle yapılan modellemenin, bilinmeyen öznitelikler üzerindeki başarımını ölçmektir. Yapılan ölçümün daha güvenilir olmasını sağlamak için k-katlı çapraz doğrulama (Stone, 1974; Kuncheva, 2004) yöntemi de kullanılabilir. Çapraz doğrulamada veritabanı öncelikle k adet parçaya bölünür. Daha sonra, bu parçaların bir tanesi test için ayrılarak, kalan $k-1$ adet parça eğitim için kullanılır. Eğitim sonrasında, ayrılan parça test edilir. Bu işlem, k parçanın herbiri için tekrarlanır ve elde edilen tanıma sonuçlarının ortalaması alınır.

Böylelikle veritabanındaki tüm veri, sırasıyla eğitim ve test için kullanılarak daha güvenilir bir başarım ölçümü yapılmış olur. Az sayıda örnek içeren veritabanlarında bu yöntem oldukça faydalıdır.

Tez çalışmasının bu bölümünde, daha önce ifade edildiği gibi sınıflandırma yöntemleri arasından doğrusal altuzay sınıflandırıcılar üzerinde durulmuştur. Bu bağlamda, ilerleyen altbölümlerde, altuzay ve altuzay sınıflandırma mantığı açıklanmış, klasik altuzay yöntemleri tanıtılmıştır. Bunları takiben, tez çalışması kapsamında geliştirilen genetik algoritma temelli altuzay sınıflandırma yöntemi açıklanmaktadır.

4.1 Altuzay

n-boyutlu $\mathbb{R}^n$ uzayı içindeki d-boyutlu S altuzayı, doğrusal bağımsız d adet birimdik taban vektörü içeren $\{u_1, u_2, ..., u_d\}$ kümesi ile tanımlanır (Hogben, 2006). Başka bir ifadeyle, S altuzayı, $\{u_1, u_2, ..., u_d\}$ taban vektörlerinin doğrusal kombinasyonlarının kümesidir:

$$S = \left\{ x \mid x = \sum_{i=1}^{d} a_i u_i, \quad a_i \in \mathbb{R} \right\} \tag{4.1}$$

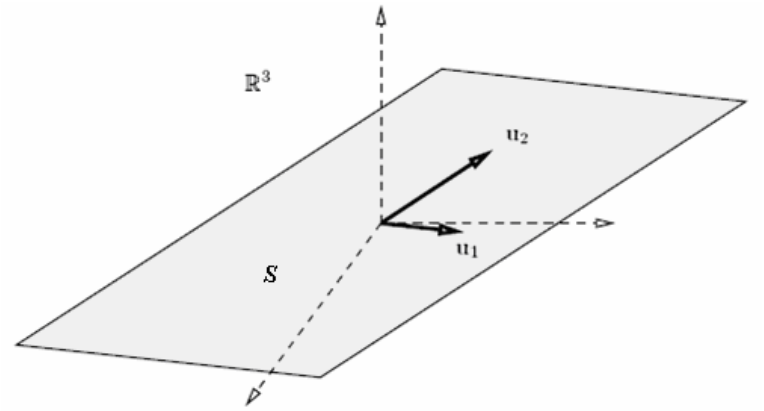
S altuzayına ait $\{u_1, u_2, ..., u_d\}$ taban vektörleri**,**

$$U = \begin{bmatrix} \uparrow & \uparrow & & \uparrow \\ u_1 & u_2 & \dots & u_d \\ \downarrow & \downarrow & & \downarrow \end{bmatrix} \tag{4.2}$$

şeklinde biraraya getirilerek, $(n \times d)$ boyutundaki d kertesli U taban vektör kümesi matrisini meydana getirir. Bu durumda, S altuzayı

$$S = \left\{ x \mid x = Uz, \quad z \in \mathbb{R}^d \right\} \tag{4.3}$$

şeklinde de ifade edilebilir. Burada, z terimi, a_i katsayılarının çarpan vektörü şeklinde biraraya getirilmiş halidir. Şekil 4.1, $\mathbb{R}^3$ uzayı içindeki 2-boyutlu S altuzayını göstermektedir. S altuzayı, $\{u_1, u_2\}$ taban vektörleri ile tanımlanmıştır.



Şekil 4.1. 3-boyutlu uzaydaki 2-boyutlu S altuzayı

$\mathbb{R}^n$ uzayındaki x vektörünün S altuzayındaki karşılığı olan $\hat{x}$, izdüşüm işlemiyle bulunur. Bu işlem, P izdüşüm matrisi kullanılarak

$$\hat{x} = Px \tag{4.4}$$

şeklinde gerçekleştirilir. P matrisi ise U taban vektör kümesi matrisi kullanılarak,

$$P = UU^T \tag{4.5}$$

eşitliğiyle bulunur. İzdüşüm matrisinin iki temel özelliği bulunmaktadır:

i) P matrisi, $\mathbb{R}^n$ uzayındaki her bir x vektörünü S altuzayındaki bir $\hat{x}$ izdüşüm vektörü ile ilişkilendirir.

$$Px = \hat{x} \in S \subset \mathbb{R}^n \tag{4.6}$$

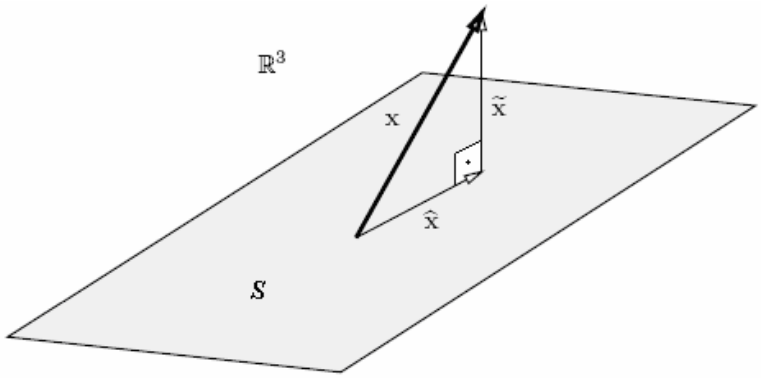
ii) S altuzayındaki her bir vektörün izdüşümü kendisine eşittir.

$$P\hat{x} = \hat{x} \tag{4.7}$$

Şekil 4.2'de görüldüğü üzere x vektörü, $\hat{x}$ ve $\tilde{x}$ şeklindeki iki vektörün toplamına karşılık gelmektedir. Burada, $\hat{x}$ vektörü, x'in S altuzayındaki izdüşümü iken, $\tilde{x}$ vektörü ise

$$\tilde{x} = x - \hat{x} = (I - P)x \tag{4.8}$$

eşitliğiyle elde edilen "kalan" vektörüdür.



Şekil 4.2. 3-boyutlu x vektörünün 2-boyutlu S altuzayına izdüşümü

4.2 Altuzay Sınıflandırma

Altuzay yöntemlerinin tarihçesi 1933'te (Hotelling, 1933) çalışması ile başlamıştır. (Watanabe et al., 1967) çalışması ise altuzay yaklaşımıyla örüntü

sınıflandırmaya dair ilk uygulamadır. İlerleyen yıllarda da altuzay sınıflandırma konusuna olan ilgi giderek artmıştır.

Altuzay sınıflandırma mantığının çıkış noktası, çok boyutlu verinin doğrusal temel bileşenleri kullanılarak küçük miktarda bir kayıp ile sıkıştırılabilmesi ve geri çatılabilmesidir (Oja, 1983). Doğrusal altuzayların sınıf modeli olarak kullanılmasının temelinde, sınıflara ait vektör dağılımlarının (yaklaşık olarak) öznitelik uzayından daha düşük boyuttaki altuzaylarda uzanması varsayımı yatar. Altuzay sınıflandırmada, bilinmeyen bir öznitelik vektörü, en yakın olduğu altuzayın temsil ettiği sınıfa atanır. Sınıflandırılacak özniteliklerin dağılım şekli düşünüldüğünde doğrusallık varsayımı her zaman doğru olmasa da pek çok durumda oldukça yüksek sınıflandırma başarımı elde edilebilmektedir.

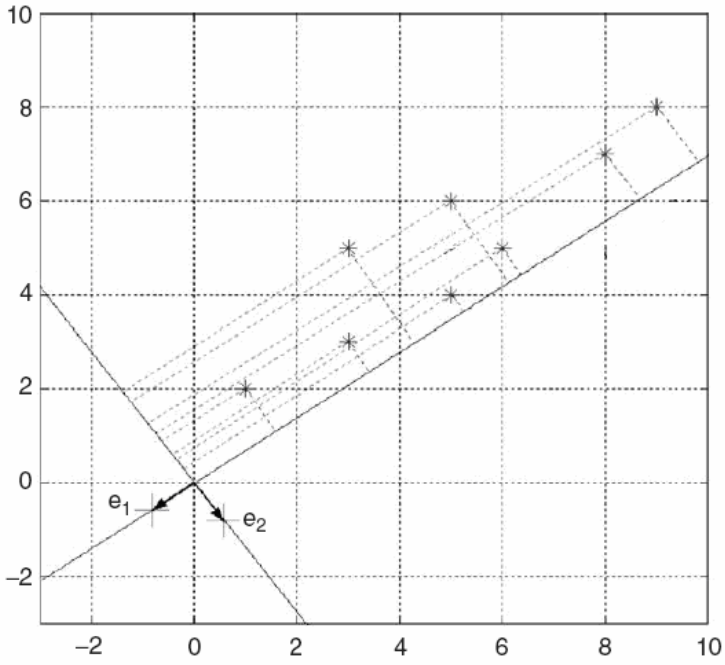
4.3 Klasik Altuzay Sınıflandırıcılar

Bu bölümde, literatürde yaygın olarak kullanılan bazı altuzay sınıflandırıcılar tanıtılmıştır. Bunlar, Temel Bileşen Analizi (PCA), Fisher'in Doğrusal Ayırtaç Analizi (FLDA), Sınıf Özellikli Bilgi Sıkıştırma (CLAFIC) ve Ortak Vektör Yaklaşımı (CVA)'dır.

4.3.1 PCA

Bölüm 2'de belirtildiği üzere, problem alanı için mantıksal ya da algoritmik bir öznitelik çıkarım yöntemi bulunamadığı durumlarda veya çok yüksek boyutlu örüntülerde sıkıştırma ve boyut indirgeme için PCA yönteminden faydalanılabilir. Bu klasik yöntem (Pearson, 1901) çalışmasında geliştirilmiş, genelleştirmesi ise (Loéve, 1963) çalışmasında yapılmıştır.

Bu yöntem, öznitelik çıkarmanın yanında bir altuzay sınıflandırıcı olarak da kullanılabilmektedir. PCA, sınıf-içi ortak değişinti matrisinin en büyük özdeğerlerine karşılık gelen (yani sınıf dağılımını en iyi temsil eden) özvektörlerin (temel bileşenler) kapsadığı altuzaylar oluşturur. Şekil 4.3'te, örnek bir dağılıma ait temel bileşen yönleri $\{e_1, e_2\}$ görülmektedir. Ancak, örüntü tanıma uygulamalarında dağılımları en iyi temsil eden özyönler her zaman en ayırdedici yönler olmamaktadır. Bu sebeple, özellikle karmaşık dağılımlar söz konusu olduğunda, PCA yönteminin sınıflandırma başarımı düşük kalmaktadır. Bu sınıflandırıcıda sınıflar-arası ilişkiler de göz önüne alınmaz.



Şekil 4.3. Örnek bir dağılıma ait temel bileşen yönleri (Akay, 2006)

4.3.2 CLAFIC

Klasik altuzay sınıflandırıcılardan biri olan CLAFIC, (Watanabe et al., 1967) çalışmasında geliştirilmiştir. CLAFIC, temel olarak sınıf ilintilerinin özvektörlerini

kullanarak sınıflandırıcı altuzayları için taban matrislerini oluşturur. Bu sınıflandırıcıda sınıflar-arası ilinti değerlendirilmez.

Sınıflandırma işlemine başlamadan önce, isteğe bağlı olarak, eğitim setindeki tüm sınıfların genel ortalaması öznitelik vektörlerinden çıkarılarak sınıf dağılımları orijin bölgesine kaydırılabilir. Sınıf-içi ilintiler, ortak değişintilerden farklı olduğu için her bir sınıftaki ilk özyön sadece sınıf ortalaması yönünün genel ortalamadan orijin noktasına dönüştürülmüş şeklini yansıtır.

CLAFIC sınıflandırma yordamı şu şekildedir:

Her bir c sınıfı için,

$$R_c = \frac{1}{N_c} \sum_{i=1}^{N_c} a_i^c a_i^{cT} \tag{4.9}$$

formülü yardımıyla, R_c ilinti matrisleri elde edilir. Burada, N_c ilgili sınıftaki öznitelik vektörü sayısını, a_i^c ise i nci öznitelik vektörünü göstermektedir. Daha sonra ilinti matrislerinin özdeğer (λ_i^c) ayrıştırması yapılarak l adet en büyük özdeğere karşılık gelen özvektörler (u_i^c) ile taban matrisleri oluşturulur:

$$U_c = \left(u_i^c \mid (R_c - \lambda_i^c I) u_i^c = 0, \quad \lambda_i^c \geq \lambda_{(i+1)}^c, \quad i = 1, ..., l \right) \tag{4.10}$$

Sonraki aşamada, hangi sınıfa ait olduğu bilinmeyen bir öznitelik vektörünün (a_x) sınıflandırılması için ilgili vektörün her bir sınıf altuzayındaki izdüşüm uzunluğu hesaplanır ve en büyük uzunluğu veren sınıfa atanır:

$$c^* = \arg\max_{1 \leq c \leq T} \left\| U_c a_x \right\|^2 \tag{4.11}$$

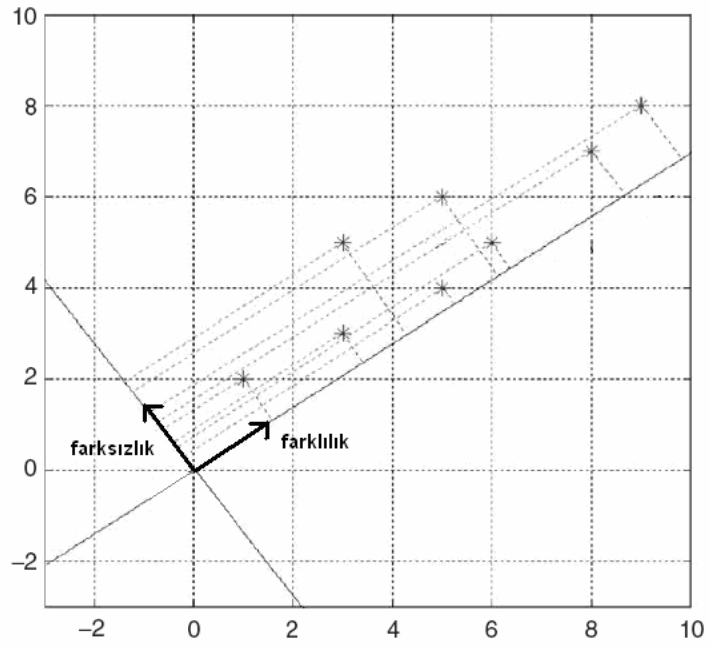
Bu denklemde, T toplam sınıf sayısını belirtmektedir.

4.3.3 CVA

CVA, sınıf-içi ortak değişinti bilgisini kullanan bir altuzay sınıflandırıcıdır. Bu anlamda PCA yöntemine benzemekle, oluşturduğu altuzaylar tamamen farklıdır. PCA, sınıf dağılımlarını en iyi temsil eden özvektörlerle altuzay oluştururken, CVA ise en ayırdedici özvektörleri kullanmaktadır.

CVA ile bir sınıfın ortak ve değişmeyen özelliklerini temsil eden "ortak vektör" elde edilir. Bu yöntemde, öznitelik uzayı, "farklılık altuzayı" ve "farksızlık altuzayı" (Bkz. Şekil 4.4) olmak üzere iki dikgen altuzaya bölünür. Bu altuzayları kapsayan özvektörler, sınıfların ortak değişinti matrislerinden elde edilir. Bir sınıfa ait ortalama vektörünün o sınıfın farksızlık altuzayına olan izdüşümü, o sınıfın ortak vektörünü verir.

CVA, yeterli ($m \geq n$) ve yetersiz ($m < n$) veri durumları için uygulanabilir. Burada, m ve n, sırasıyla eğitim kümesindeki öznitelik vektörü sayısını ve öznitelik vektörü boyutunu temsil etmektedir. Bu tez çalışmasında kullanılan veritabanlarında yeterli veri durumu söz konusudur.



Şekil 4.4. Örnek bir dağılıma ait farklılık ve farksızlık altuzayı yönleri

Sütun vektörleri $\left\{a_1^c, a_2^c, ..., a_m^c\right\} \in \mathbb{R}^n$, C sınıfına ait öznitelik vektörleri; a_{ave}^c ilgili sınıfın ortalama vektörü; $P_c^{\perp}$ ise yine ilgili sınıfın farksızlık altuzayına izdüşüm matrisi olsun. Bu durumda, her bir sınıfa ait ortak vektör ve altuzay bölümlemesi,

$$F^c = \sum_{i=1}^{m} \left\| P_c^{\perp}(a_i^c - a_{ave}^c) \right\|^2 \tag{4.12}$$

ölçütünün enküçüklenmesiyle elde edilir. Bu kriter, bir sınıfın farksızlık altuzayında, o sınıfın özniteliklerinin sınıf ortalamasına yakın olmasını ifade etmektedir. Bu eniyileme problemi,

$$(\Phi^c - \lambda_j^c I)u_j^c = \vec{0} \qquad j = 1, 2, .., m \tag{4.13}$$

şeklindeki genelleştirilmiş özdeğer probleminin sonucunu bularak çözülebilir. Burada, Φ^c, sınıf-içi ortak değişinti matrisini, $\left\{\lambda_j^c, u_j^c\right\}$ çifti ise bu matrisin özdeğer ve özvektör çiftlerini göstermektedir.

Yeterli veri durumunda ($m \geq n$), F^c ölçütünü enküçüklemek için farksızlık altuzayı, en küçük özdeğerlere karşılık gelen özvektörler ile oluşturulur (Gülmezoğlu et al., 2007). Sınıf-içi ortak değişinti matrisi Φ^c'nin özayrışımı ile elde edilen özdeğerler, küçükten büyüğe doğru sıralanırsa, bunların arasından en küçük k tanesine karşılık gelen özvektörler, o sınıfın farksızlık altuzayını kapsarken, geriye kalan ($n-k$) adet özdeğere karşılık gelen özvektörler ise farklılık altuzayını kapsar.

$$\underbrace{\lambda_1 < \lambda_2 < ..... < \lambda_k}_{Farksızlık} < \underbrace{\lambda_{k+1} < \lambda_{k+2} < ..... < \lambda_n}_{Farklılık} \tag{4.14}$$

Bu şekildeki ayrıştırmanın amacı, sınıf dağılımındaki büyük değişimleri eleyerek, küçük olanları ortaya çıkarmaktır. Bunun altında yatan sebep ise özyönleri fazla değişim göstermeyen altuzayın, o sınıfa ait ortak karakteristiği taşımasıdır. Bu sayede, sınıflandırma açısından daha elverişli bir durum ortaya çıkmaktadır.

Burada, k parametresi, enerji oranı göz önünde bulundurularak, en küçük k adet özdeğerin toplamının tüm özdeğerlerin toplamına oranı belli bir eşik değeri (L) ile sınırlandırılacak şekilde belirlenir:

$$\frac{\sum_{j=1}^{k} \lambda_j^c}{\sum_{j=1}^{n} \lambda_j^c} < \mathrm{L} \tag{4.15}$$

Eşik değeri, örüntü türüne göre %5 ile %10 arasında değişebilmektedir (Oja, 1983; Günal et al., 2006; Gülmezoğlu et al., 2007). k parametresinin belirlenmesini takiben, farksızlık altuzayını kapsayan özvektörler kullanılarak,

$$P_c^{\perp} = \sum_{j=1}^{k} u_j^c u_j^{cT} \tag{4.16}$$

formülü ile farksızlık altuzayına izdüşüm matrisi elde edilir. Böylece, ilgili sınıfa ait ortak vektör (a_{com}^c), öznitelik uzayındaki sınıf ortalamasının farksızlık altuzayına izdüşümü alınarak,

$$a_{com}^c = P_c^{\perp} a_{ave}^c \tag{4.17}$$

şeklinde bulunur. Bu yöntem ile sınıflandırmada ise bilinmeyen bir a_x öznitelik vektörü, enküçük öklid uzaklığı prensibi kullanılarak,

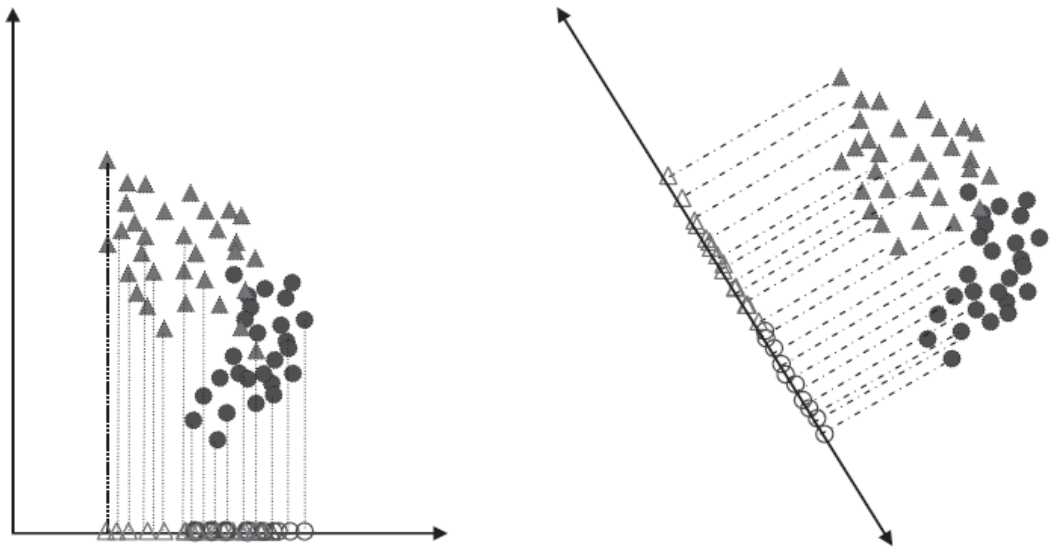
$$c^* = \arg\min_{1 \le c \le T} \left\| P_c^{\perp} a_x - a_{com}^c \right\| \tag{4.18}$$

formülü yardımıyla en yakın olduğu ortak vektörün ait olduğu sınıfına atanır. Bu eşitlikte T parametresi, toplam sınıf sayısını ifade etmektedir.

4.3.4 FLDA

(Fisher, 1936) çalışmasında geliştirilmiş olan FLDA, sınıf-içi ortak değişintinin yanısıra sınıflar-arası ortak değişinti bilgisini de kullanır. Bu yöntemde izdüşüm yönleri, aynı sınıfa ait örneklerin birbirlerine olan uzaklıklarını enküçükleyecek ve aynı zamanda farklı sınıflara ait örneklerin birbirlerine olan uzaklıklarını ise enbüyükleyecek şekilde belirlenir (Duda et al., 2001; Theodoridis and Koutroumbas, 2003). FLDA ile elde edilen altuzay, gerek sınıflandırma gerekse boyut indirgeme amaçlı olarak kullanılabilmektedir. Şekil 4.5-a'da görülen izdüşüm yönü, iki sınıfa ait örneklerin karışmasına yol açmakta, Şekil 4.5-b'de ise FLDA ile bulunan izdüşüm yönü sınıfları oldukça başarılı şekilde ayırmaktadır.

Bu yöntem ile en çok $(Sınıf\ Sayısı - 1)$ boyutlu altuzaya dönüşüm elde edilir. Dönüşmüş koordinat sisteminin eksenleri, ayrımsama önemine göre sıralanabilir. Sınıflandırma kuralı olarak "en yakın sınıf ortalaması" kullanıldığında, doğrusal karar sınırları elde edilir.



(a) (b)

Şekil 4.5. (a) Uygunsuz izdüşüm (b) FLDA ile uygun izdüşüm (Akay, 2006)

Sınıf-içi saçılım matrisi S_W ,

$$S_W = \sum_{i=1}^{C} p_i \Phi_i \tag{4.19}$$

şeklinde tanımlanır. Bu eşitlikte, Φ_i sınıf ortak değişinti matrisini, p_i ilgili sınıfın öncül olasılığını, C ise toplam sınıf sayısını ifade etmektedir. Sınıflar-arası saçılım matrisi S_B ise

$$S_B = \sum_{i=1}^{C} p_i (\mu_i - \mu)(\mu_i - \mu)^T \tag{4.20}$$

formülü yardımıyla hesaplanır. Burada, μ_i ve μ vektörleri sırasıyla ilgili sınıfın ortalaması ve tüm sınıfların ortalamasını temsil etmektedir. S_W ve S_B saçılımlarını kullanan Fisher ölçütü,

$$J(W) = \frac{\left| W^T S_B W \right|}{\left| W^T S_W W \right|} \tag{4.21}$$

şeklindedir. Bu ölçütü enbüyükleyen W dönüşüm matrisi,

$$S_W^{-1} S_B W = \lambda W \tag{4.22}$$

eşitliğiyle gösterilen, genelleştirilmiş özdeğer problemini çözerek elde edilir. Bunun sonucunda, W matrisi $S_W^{-1} S_B$ matrisinin en büyük $C-1$ adet özdeğerine karşılık gelen özvektörler ile oluşturulur.

Yetersiz veri durumu söz konusu olduğunda ise S_W tekil olacağı için tersi alınamaz ve özdeğer problemi çözülemez. Bu durumda, S_W matrisinin tersi yerine sözde tersini hesaplamak, uygulanabilecek yöntemlerden biridir (Tian et al., 1988).

FLDA ile sınıflandırmada ise W dönüşüm matrisi ile elde edilen ve tüm sınıflar için ortak olan altuzayda, en yakın sınıf ortalaması kuralı kullanılabilir. Bu doğrultuda tanımlanmış

$$c^* = \arg\min_{1 \le c \le T} \left\| W(a_x - a_{ave}^c) \right\| \tag{4.23}$$

kuralı ile bilinmeyen a_x vektörü uygun sınıfa atanır.

4.4 Genetik Altuzay Sınıflandırıcı

Önceki bölümlerde bahsedilen klasik altuzay sınıflandırıcılar, çeşitlerine göre yalnızca sınıf-içi ilinti/değişinti bilgisini (PCA, CLAFIC ve CVA) ya da sınıf-içi ve sınıflar-arası değişinti bilgisini birlikte (FLDA) kullanmaktadır. Sadece sınıf-içi ilişkileri değerlendirmek çoğu durumda yeterince yüksek bir başarım sağlamamaktadır. Bunun yanısıra, altuzay sınıflandırıcıların altuzay tabanı oluştururken kullandıkları özyönlerde farklılık göstermektedir. CVA sınıflandırıcı, değişintinin az olduğu; yani, küçük özdeğerlere karşılık gelen özyönlerden altuzay tabanı oluştururken diğerleri büyük özdeğerlere karşılık gelen özyönleri temel almaktadır. Halbuki, ayırdedicilik açısından gerek küçük gerekse büyük özdeğerlere karşılık gelen özyönler birlikte kullanıldığında daha etkili olabilir. Ayrıca, bu sınıflandırıcıların bazılarında uygun özyön sayısının belirlenmesi de bir problem teşkil etmektedir.

Yukarıda bahsedilen problemlere çözüm üretmek amacıyla, tez çalışmasının bu bölümünde hem sınıf-içi hem de sınıflar-arası ilişkileri değerlendiren, aynı zamanda, özyön seçimindeki limitleri kaldıran, genetik algoritma temelli bir altuzay sınıflandırıcı geliştirilmiştir.

Daha önce bahsedildiği gibi hem sınıf-içi hem de sınıflar-arası ilişkileri değerlendiren altuzay sınıflandırıcı FLDA'dir. Ancak, bu sınıflandırıcıda kullanılan saçılım matrislerinin kertes özellikleri incelendiğinde, öznitelik vektörlerinin boyutu ne

kadar yüksek olursa olsun, altuzay izdüşümü en fazla (sınıf sayısı - 1) adet boyut üzerine gerçekleştirilebilir. Bu durum, özniteliklerin çoğunun ayırdedici bilgi içerdiği yüksek boyutlu problemlerde sınıflandırma başarımının düşmesine yol açar. Ayrıca bu yaklaşım, sınıf dağılımlarının tek doruklu olduğu varsayımını kabullenir. Ancak, çok doruklu ya da fazlasıyla üst üste binen dağılımlar söz konusu ise, FLDA etkisiz kalabilir (Akay, 2006). FLDA sınıflandırıcının farklı bir versiyonu, Parametrik Olmayan Doğrusal Ayırtaç Analizi (NDA), sınıflar-arası saçılım matrisini hesaplarken farklı bir yol izleyerek bu matrisi tam kertesli hale getirir. Böylece, hem tek doruk kabullenmesi hem de izdüşümdeki (sınıf sayısı – 1) adet boyut sınırı ortadan kalkar (Fukunaga and Mantock, 1983).

FLDA yönteminde, sınıflar-arası saçılım matrisinin hesaplanmasında sadece sınıf ortalamaları gözönünde bulundurulurken (4.20), NDA yönteminde ise sınıflardaki örneklerin tamamı kullanılır. Bunu yaparken aynı zamanda örneklerin sınıf-içi ve sınıflar-arası komşulukları da değerlendirilir (Bkz. Şekil 4.6). Bu doğrultuda tanımlanan S_B sınıflar-arası saçılım matrisi,

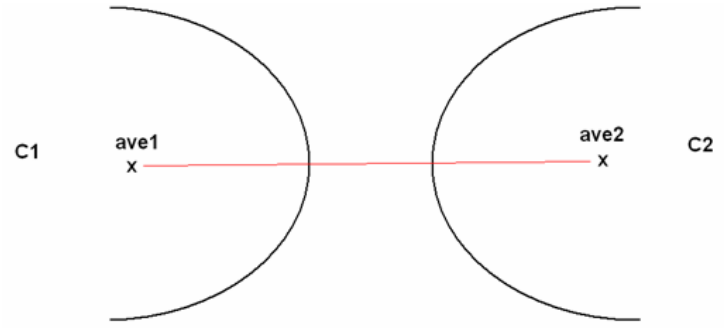
$$S_B = \frac{1}{N}\sum_{i=1}^{C}\sum_{j=1}^{C}\sum_{a\in C_i} w_{ija}(a-m_{ij})(a-m_{ij})^T \tag{4.24}$$

formülü yardımıyla hesaplanır. Bu formülde, N, toplam örnek sayısını; m_{ij}, i sınıfındaki a öznitelik vektörünün j sınıfındaki k-enyakın komşularıyla uzaklığının ortalamasını; w_{ija} ise

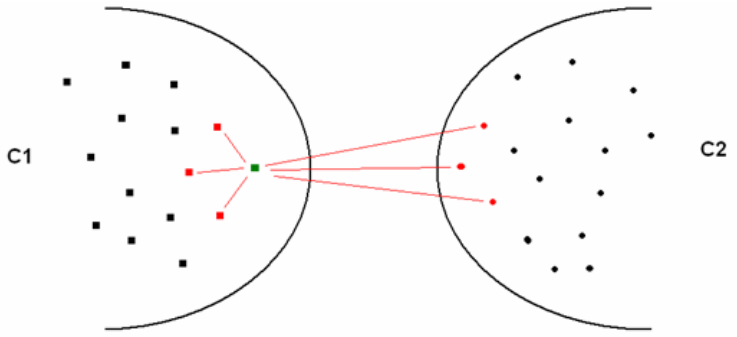
$$w_{ija} = \frac{\min(d(a^i_{KNN}), d(a^j_{KNN}))}{d(a^i_{KNN}) + d(a^j_{KNN})} \tag{4.25}$$

şeklinde tanımlanan, i sınıfındaki a öznitelik vektörünün j sınıfına göre ağırlık miktarını göstermektedir. Burada, $d(a^i_{KNN})$, a öznitelik vektörünün i sınıfında bulunan k-enyakın komşularıyla arasındaki öklid uzaklığıdır. En uygun k değeri, deneysel olarak

belirlenir. Bu formüle göre, bir sınıfa ait örnek, diğer sınıfa ait örneklerden uzak ise w_{ija} ağırlığı küçük olacaktır.



(a)



(b)

Şekil 4.6. Sınıflar-arası ilişkinin değerlendirilmesi (a) FLDA (b) NDA

NDA sınıflandırıcının, sınıflar-arası saçılımın hesaplanması haricindeki aşamaları FLDA ile aynıdır. Dolayısıyla, yeni elde edilen S_B matrisi,

$$S_W^{-1} S_B W = \lambda W \tag{4.26}$$

eşitliğinde kullanılarak, genelleştirilmiş özdeğer problemi çözülür ve $S_W^{-1} S_B$ matrisinin en büyük özdeğerlerine karşılık gelen özvektörler ile izdüşüm matrisi oluşturulur.

NDA sınıflandırıcının kullanılmasıyla sınıflar-arası ilişkiler FLDA yöntemine kıyasla daha doğru bir şekilde analiz edilebilir. Ancak, özvektörlerin seçimi büyüklük ve sayı açısından halen bir problem teşkil etmektedir. Bu problemi gidermek için ikinci aşamada genetik algoritmadan yararlanılmıştır. Bu amaçla, Bölüm 3'te anlatılan genetik algoritma ile öznitelik seçim işlemine benzer şekilde özyön seçimi yapılır. Bunun için, {0,1} alfabesi ile kodlanmış ve öznitelik uzayı boyutundaki kromozom yapısı kullanılır. Bu kromozom yapısında, 1 değerli indisler, seçilmiş özyönleri; 0 değerli olanlar ise seçilmeyen özyönleri temsil etmektedir. Bu problemde, uyum fonksiyonu değeri ise seçilen özyönler üzerine yapılan izdüşüm ile ulaşılan sınıflandırma hassasiyeti olarak belirlenmiştir.

Böylelikle, GA-NDA sınıflandırıcı ile sınıf-içi ve sınıflar-arası ilişkiler uygun bir şekilde değerlendirilmekte; bunun yanısıra, büyük ya da küçük özdeğerlere karşılık gelen, farklı sayıdaki özvektörün seçimi de mümkün kılınmaktadır.

BÖLÜM 5

DENEYSEL ÇALIŞMALAR

Bu bölüm, tez kapsamında gerçekleştirilen tüm deneysel çalışmaları ve bunların sonuçlarını içermektedir. Alt bölümlerde ilk olarak deneylerde kullanılan veritabanları tanıtılmıştır. Daha sonra ise sırasıyla Öznitelik Çıkarma (Bölüm 2), Öznitelik Seçme (Bölüm 3) ve Sınıflandırma (Bölüm 4) konularına ait deneyler yer almaktadır.

5.1 Veritabanları

Tez kapsamında yapılan deneysel çalışmalarda, çeşitli sayıda sınıf, öznitelik ve örneğe sahip toplam 6 farklı veritabanı kullanılmıştır. Bu veritabanlarına ait bilgiler Çizelge 5.1'de özetlenmiştir. Tüm veritabanlarında yeterli veri durumu söz konusudur.

Çizelge 5.1. Veritabanı listesi

No	Veritabanı	Sınıf Sayısı	Öznitelik Sayısı	Örnek Sayısı / Sınıf
1	TI-DIGIT	10	96	450
2	E-SET	8	96	150
3	VOWEL	5	96	200
4	POWER	4	19	30
5	VEHICLE	4	18	200
6	PROTEIN	4	8	160

8 kHz'de örneklenmiş olan TI-DIGIT veritabanı, {/ow/, /zero/, /one/, /two/, /three/, /four/, /five/, /six/, /seven/, /eight/, /nine/} İngilizce rakamlardan oluşan 11 sınıfa sahiptir (Leonard, 1984). Ancak bu çalışma da /ow/ sınıfı haricindeki diğer 10 sınıf kullanılmıştır. Öznitelik çıkarımı için Bölüm 2'de konuşma tanıma uygulamaları için

önerilen dalgacık tabanlı yöntem kullanılmış ve 96 adet öznitelik elde edilmiştir (Çizelge 5.2). Her bir sınıfa ait örnek sayısı 450 olarak belirlenmiştir.

16 kHz'de örneklenmiş olan E-SET veritabanı, {/b-e/, /s-e/, /d-e/, /jh-e/, /p-e/, /t-e/, /v-e/, /z-e/} fonem çiftlerinden oluşan 8 sınıfa sahiptir. E-SET, konuşma tanıma uygulamalarında yaygın olarak tercih edilen TIMIT veritabanındaki (Zue et al., 1990; Farooq and Datta, 2003; Reynolds and Antoniou, 2003; Günal and Edizkan, 2006) lehçe bölgelerinin tamamı kullanılarak elde edilmiştir. Öznitelik çıkarımı için, TI-DIGIT veritabanına benzer şekilde Bölüm 2'deki yöntem kullanılmış ve 96 adet öznitelik elde edilmiştir (Çizelge 5.2). Her bir sınıfa ait örnek sayısı ise 160 olarak tanımlanmıştır.

VOWEL veritabanı, {/aa/, /eh/, /iy/, /ow/, /uw/} sesli fonemlerinden oluşan 5 sınıfa sahiptir. Bu veritabanı da E-SET'te olduğu gibi, TIMIT veritabanından elde edilmiştir. Öznitelik çıkarımı için, TI-DIGIT ve E-SET veritabanlarındakine benzer şekilde, Bölüm 2'de önerilen yöntem kullanılmış ve 96 adet öznitelik elde edilmiştir (Çizelge 5.2). Her bir sınıfa ait örnek sayısı 200 olarak alınmıştır.

Çizelge 5.2. TI-DIGIT, E-SET ve VOWEL veritabanlarının öznitelik listesi

No	Öznitelik
1	(Çerçeve-1, Altbant-1) Enerji değeri
.	...
.	...
24	(Çerçeve-1, Altbant-24) Enerji değeri
25	(Çerçeve-2, Altbant-1) Enerji değeri
.	...
.	...
48	(Çerçeve-2, Altbant-24) Enerji değeri
49	(Çerçeve-3, Altbant-1) Enerji değeri
.	...
.	...
72	(Çerçeve-3, Altbant-24) Enerji değeri
73	(Çerçeve-4, Altbant-1) Enerji değeri
.	...
.	...
96	(Çerçeve-4, Altbant-24) Enerji değeri

POWER veritabanı, güç kalitesi olaylarının sınıflandırılması için (Gerek et al., 2006) çalışması kapsamında oluşturulmuş 4 sınıflı bir veritabanıdır. Çizelge 5.3'te sınıfların açıklamaları verilmiştir. Veritabanındaki örneklerden çıkarılan 19 öznitelik, Çizelge 5.4'te listelenmektedir. Her bir sınıfa ait örnek sayısı ise 30'dur.

Çizelge 5.3. POWER veritabanı sınıfları

Sınıf	Açıklama
1	Direnil, indüklemeli ve hız ayarlı sürücü yükü ile ark hatası
2	Direnil ve indüklemeli yük ile ark hatası
3	Direnil ve indüklemeli yük ile motor başlatma
4	Direnil, indüklemeli ve hız ayarlı sürücü yükü ile motor başlatma

Çizelge 5.4. POWER veritabanının öznitelik listesi

No	Öznitelik
1	Dalgacık Detay Seviyesi-1 enküçük katsayı
2	Dalgacık Detay Seviyesi-1 enbüyük katsayı
3	Dalgacık Detay Seviyesi-2 enküçük katsayı
4	Dalgacık Detay Seviyesi-2 enbüyük katsayı
5	Dalgacık Detay Seviyesi-3 enküçük katsayı
6	Dalgacık Detay Seviyesi-3 enbüyük katsayı
7	Dalgacık Detay Seviyesi-4 enküçük katsayı
8	Dalgacık Detay Seviyesi-5 enbüyük katsayı
9	50 Hz'de orantısal sinyal enerjisi
10	2. derece logaritmik momentin enbüyük değeri
11	2. derece logaritmik momentin enküçük değeri
12	Enbüyük yamukluk değeri
13	Enküçük yamukluk değeri
14	Enbüyük savrukluk değeri
15	Enküçük savrukluk değeri
16	3. derece logaritmik momentin enbüyük değeri
17	3. derece logaritmik momentin enküçük değeri
18	4. derece logaritmik momentin enbüyük değeri
19	4. derece logaritmik momentin enküçük değeri

VEHICLE veritabanı, {otobüs, minibüs, Opel, Saab} şeklindeki dört farklı aracın silüetlerinden oluşan 4 sınıfa sahiptir (Asuncion and Newman, 2007). Bu veritabanından elde edilmiş olan 18 adet öznitelik Çizelge 5.5'te listelenmiştir. Her bir sınıfa ait örnek sayısı 199 olarak tanımlanmıştır.

Çizelge 5.5. VEHICLE veritabanının öznitelik listesi

No	Öznitelik
1	Tıkızlık
2	Yuvarlaklık
3	Uzaklık yuvarlaklığı
4	Yarıçap oranı
5	Ana eksen en-boy oranı
6	Enbüyük uzunluk en-boy oranı
7	Saçılma oranı
8	Uzatılmışlık
9	Ana eksen dikdörtgenselliği
10	Enbüyük uzunluk dikdörtgenselliği
11	Büyük eksende ölçeklenmiş değişinti
12	Küçük eksende ölçeklenmiş değişinti
13	Fırdolanımın ölçeklenmiş yarıçapı
14	Büyük eksen üzerindeki yamukluk
15	Küçük eksen üzerindeki yamukluk
16	Büyük eksen üzerindeki basıklık
17	Küçük eksen üzerindeki basıklık
18	İçi boşluk (oyukluk) oranı

PROTEIN (Yeast) veritabanı, proteinlerin hücresel lokalizasyon bölgelerini tahmin etmek için oluşturulmuştur (Asuncion and Newman, 2007). Bu veritabanı, aslen 10 sınıf içermekle birlikte, bazı sınıflarındaki örnek sayısı çok az olduğu için 10 sınıfın sadece 4'ü kullanılmıştır: {CYT (cytosolic or cytoskeletal), NUC (nuclear), MIT (mitochondrial) and ME3 (membrane protein, no N-terminal signal)}. Veritabanından elde edilmiş olan 8 adet öznitelik, Çizelge 5.6'da listelenmektedir. Her bir sınıfa ait örnek sayısı ise 160 olarak belirlenmiştir.

Çizelge 5.6. PROTEIN veritabanının öznitelik listesi

No	Öznitelik
1	Sinyal dizisi tanıma için McGeoch'nin yöntemi
2	Sinyal dizisi tanıma için Von Heijne'nin yöntemi
3	ALOM zar kapsama bölgesi tahmin programı skoru
4	Mitokondriyel ve mitokondriyel olmayan proteinlerin N-terminal bölgesindeki amino asit içeriğinin ayırtaç analizi skoru
5	"HDEL" altdizgisinin mevcudiyeti
6	C-terminus'taki peroksisomal hedefleme sinyali
7	Ekstrahücre proteinlerinde ve boşluklardaki amino asit içeriğinin ayırtaç analizi skoru
8	Çekirdeksel ve çekirdeksel olmayan proteinlerin çekirdeksel lokalizasyon sinyallerinin ayırtaç analizi skoru

5.2 Öznitelik Çıkarma Deneyleri

Deneysel çalışmaların bu kısmında, ses ve konuşma tanıma uygulamaları için Bölüm 2'de önerilmiş olan dalgacık dönüşümü temelli öznitelik çıkarma yönteminin Fourier dönüşümü temelli yöntem ile sınıflandırma başarımı açısından karşılaştırması yapılmıştır. Burada bahsi geçen Fourier dönüşümü temelli yöntem, Bölüm 2'de de ifade edildiği üzere MFCC özniteliklerinin DCT uygulanmamış şeklidir. DCT uygulanmamasındaki amaç, sinyale ait altbant frekans bilgilerinin korunması ve dalgacık temelli yöntem ile birebir karşılaştırma yapılabilmesidir.

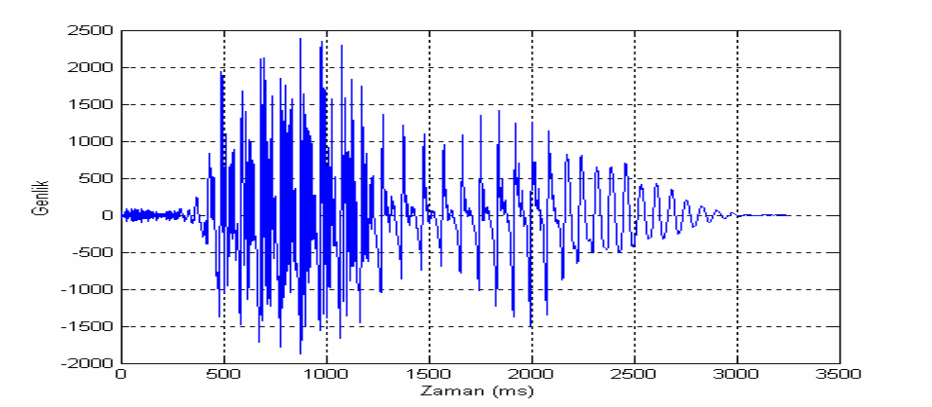
Önerilen öznitelik çıkarma yönteminin değişik koşullarda sınanabilmesi için deneysel çalışmalarda farklı karakteristikteki ses sinyallerini içeren TI-DIGIT, E-SET ve VOWEL veritabanları kullanılmıştır. Bu yönteminin farklı çerçeve sayıları için sağladığı başarımlar incelendiğinde en uygun çerçeve sayısının 4 olduğu gözlenmiştir (Günal and Edizkan, 2007). Çerçeve sayısının 4'ten daha yüksek olması durumunda, artan işlem süresine karşılık sınıflandırma başarımındaki artış çok fazla olmamaktadır. Deneysel çalışmalarda yalnızca 4 çerçeve için sonuçlar verilmiştir. Ana dalgacık tipi olarak ise Daubechies-32 tercih edilmiştir.

Özniteliklerin sınıflandırma başarımlarının elde edilmesi için CLAFIC, CVA ve FLDA altuzay sınıflandırıcılar kullanılmış, sınıflandırma sonuçlarının güvenilirliğini sağlamak içinse çapraz doğrulamadan faydalanılmıştır.

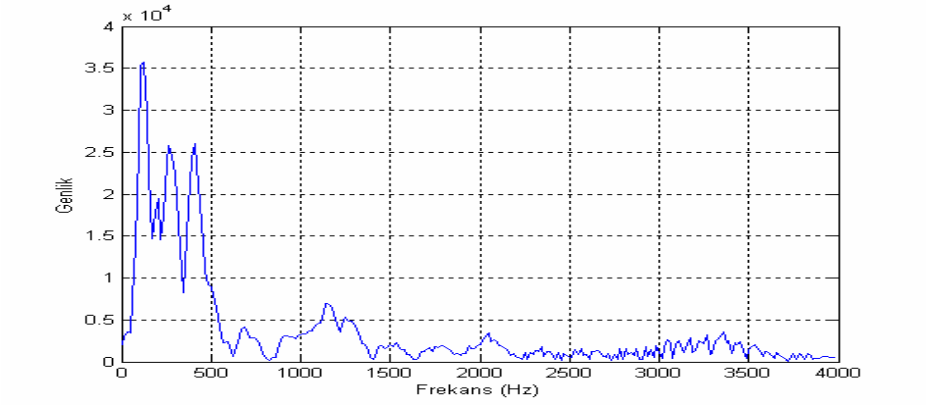
Öznitelik çıkarma deneylerinde kullanılan ilk veritabanı TI-DIGIT'dir. Şekil 5.1'de, bu veritabanından seçilmiş örnek bir sinyalin zaman bölgesi, Fourier dönüşümü ve dalgacık katsayıları görülmektedir.

Çizelge 5.7'de TI-DIGIT veritabanı için CLAFIC, CVA ve FLDA altuzay sınıflandırıcılar üzerinde Dalgacık ve Fourier temelli öznitelikler ile sınıf bazında elde edilen tanıma oranları görülmektedir. Şekil 5.2'de ise üç sınıflandırıcıya ait ortalama tanıma oranları karşılaştırmalı olarak verilmiştir. Bu sonuçlardan da açıkça görüldüğü üzere Dalgacık öznitelikleri Fourier özniteliklerine karşı tüm sınıflandırıcılarda belirgin bir üstünlük sağlamaktadır.

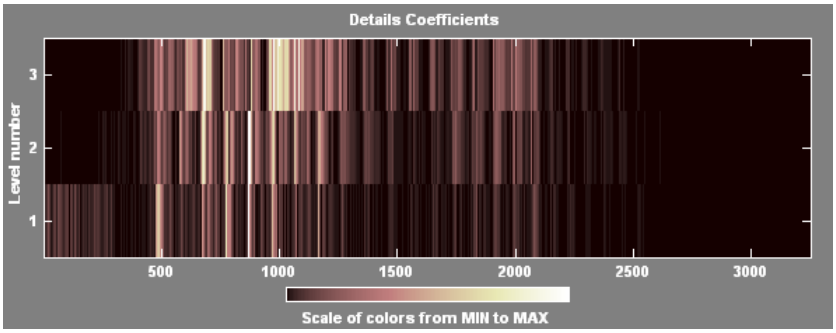
Bu durum şu şekilde yorumlanabilir: TI-DIGIT veritabanı, rakam telaffuzlarını içerdiği için ses sinyalleri, E-SET ve VOWEL fonem veritabanlarına kıyasla daha uzundur. Bu durum, çerçeve sürelerinin de uzun olmasına yol açmaktadır. Aynı zamanda, rakamlar sesli ve sessiz çeşitli fonemlerin birleşiminden oluştuğu için farklı frekans bileşenleri ve sinyal içerisinde ani frekans değişimleri söz konusudur. Fourier dönüşümü, Bölüm 2'de belirtildiği üzere, sinyalleri uzun bir zaman sürecinde iyi temsil edememekte ve frekanslardaki ani ve kısa süreli değişimleri çok iyi tespit edememektedir. Dalgacık dönüşümü ise daha iyi zaman – frekans lokalizasyonu sağladığı için her iki durumda da sinyalleri başarıyla temsil etmektedir. Bu sebeple, TI-DIGIT veritabanındaki sinyallerden, dalgacık dönüşümü ile elde edilen öznitelikler, Fourier temelli özniteliklere göre daha iyi sınıflandırma başarımı sağlamıştır.



(a)



(b)



(c)

Şekil 5.1. TI-DIGIT veritabanından örnek bir sinyal: /seven/ (a) Zaman bölgesi (b) Fourier dönüşümü (c) Dalgacık katsayıları

Çizelge 5.7. TI-DIGIT veritabanı: Dalgacık ve Fourier dönüşümü temelli özniteliklerin (a) CLAFIC (b) CVA (c) FLDA tanıma oranları (%)

Sınıf	Dalgacık	Fourier
'zero'	100.00	96.00
'one'	92.00	96.00
'two'	100.00	84.00
'three'	98.00	86.00
'four'	100.00	98.00
'five'	96.00	98.00
'six'	100.00	72.00
'seven'	98.00	96.00
'eight'	92.00	88.00
'nine'	92.00	94.00
Ortalama	96.80	90.80

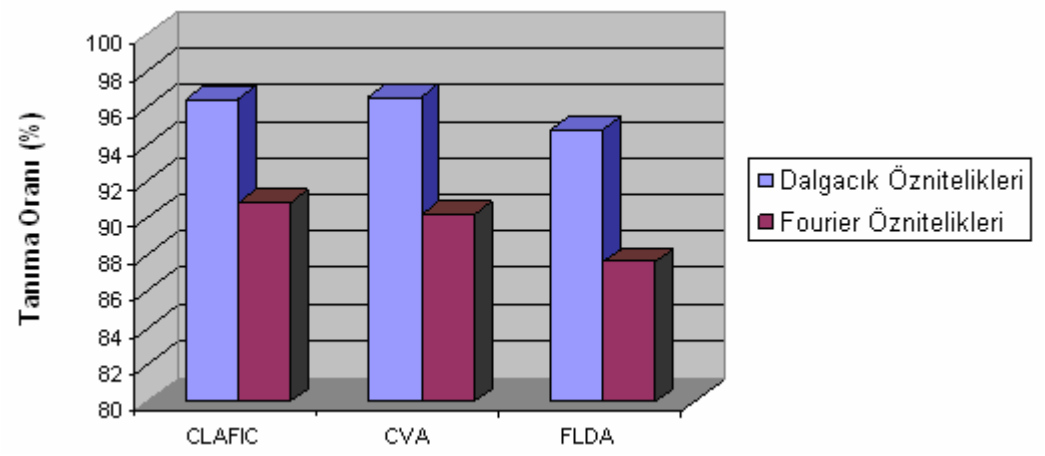
(a)

Sınıf	Dalgacık	Fourier
'zero'	98.00	88.00
'one'	94.00	90.00
'two'	98.00	84.00
'three'	98.00	92.00
'four'	100.00	100.00
'five'	96.00	96.00
'six'	100.00	82.00
'seven'	98.00	90.00
'eight'	100.00	86.00
'nine'	84.00	94.00
Ortalama	96.60	90.20

(b)

Sınıf	Dalgacık	Fourier
'zero'	100.00	90.00
'one'	88.00	94.00
'two'	96.00	86.00
'three'	100.00	96.00
'four'	100.00	98.00
'five'	96.00	92.00
'six'	100.00	82.00
'seven'	92.00	70.00
'eight'	90.00	76.00
'nine'	86.00	92.00
Ortalama	94.80	87.60

(c)



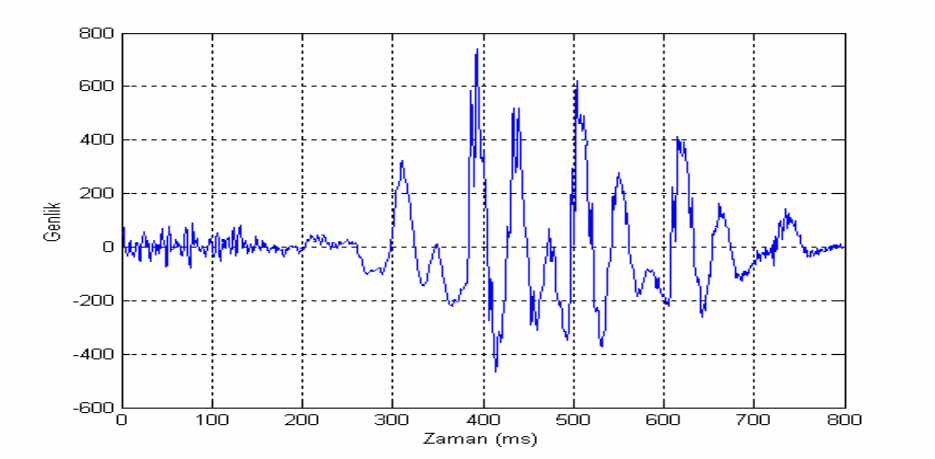
Şekil 5.2. TI-DIGIT veritabanı: Dalgacık ve Fourier temelli özniteliklerin farklı sınıflandırıcılar ile sağladığı ortalama tanıma oranları

E-SET, bu bölümde kullanılan ikinci veritabanıdır. E-SET içerisinden seçilmiş bir örneğin zaman bölgesi, Fourier dönüşümü ve dalgacık katsayıları Şekil 5.3'te verilmektedir.

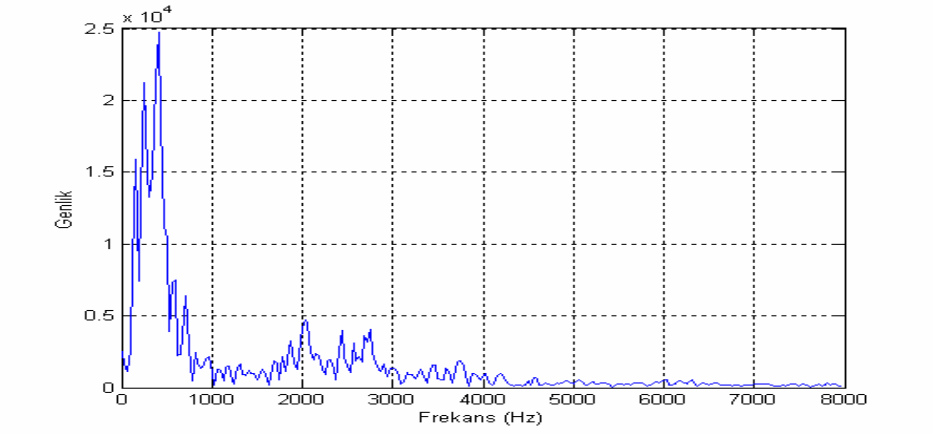
E-SET veritabanı üzerinde yapılan sınıflandırma deneylerinin sonuçları Çizelge 5.8 ve Şekil 5.4'te verilmektedir. Çizelge 5.8, CLAFIC, CVA ve FLDA altuzay sınıflandırıcılar üzerinde Dalgacık ve Fourier temelli öznitelikler ile sınıf bazında elde edilen tanıma oranlarını, Şekil 5.4'te ise üç sınıflandırıcıya ait ortalama tanıma oranlarını karşılaştırmalı olarak göstermektedir. Bu verilerden anlaşıldığı üzere, CLAFIC ve CVA sınıflandırıcılarda dalgacık öznitelikleriyle elde edilen tanıma oranları Fourier özniteliklerine oranla biraz daha yüksektir. FLDA sınıflandırıcıda ise iki tip öznitelik ile de hemen hemen aynı başarıma ulaşılmıştır.

Bu veritabanında sinyaller, TI-DIGIT veritabanına göre nispeten daha kısadır. Ancak, buradaki sinyaller gürültü biçimindeki sessiz fonemleri takip eden periyodik sesli fonemlerden oluşmakta ve özellikle iki fonemin geçiş noktalarında ani ve kısa süreli frekans değişimleri bulunmaktadır (Bkz. Şekil 5.3-a). Bu kez, sinyal uzunlukları fazla olmamasına rağmen ani frekans değişimleri söz konusu olduğu için dalgacık öznitelikleri Fourier özniteliklere göre biraz daha iyi başarım sağlamıştır. Dolayısıyla,

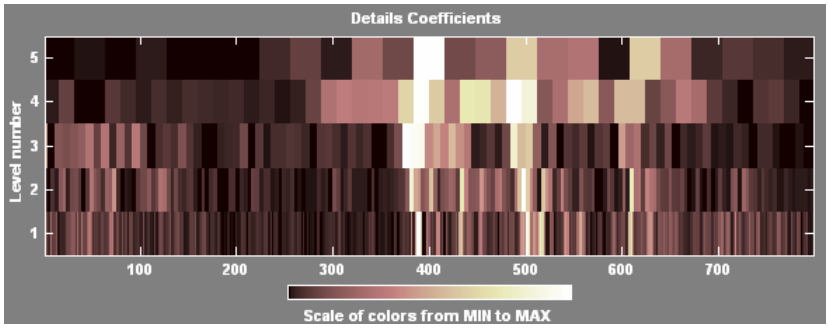
elde edilen başarımlar ile E-SET veritabanındaki sinyallerin karakteristik yapıları örtüşmektedir.



(a)



(b)



(c)

Şekil 5.3. E-SET veritabanından örnek bir sinyal: /b-iy/ (a) Zaman bölgesi (b) Fourier dönüşümü (c) Dalgacık katsayıları

Çizelge 5.8. E-SET veritabanı: Dalgacık ve Fourier dönüşümü temelli özniteliklerin (a) CLAFIC (b) CVA (c) FLDA tanıma oranları (%)

Sınıf	Dalgacık	Fourier
/b-e/	56.60	63.30
/s-e/	70.00	93.30
/d-e/	53.30	40.00
/jh-e/	90.00	93.30
/p-e/	50.00	30.00
/t-e/	73.30	70.00
/v-e/	83.30	80.00
/z-e/	86.60	76.60
Ortalama	70.40	68.30

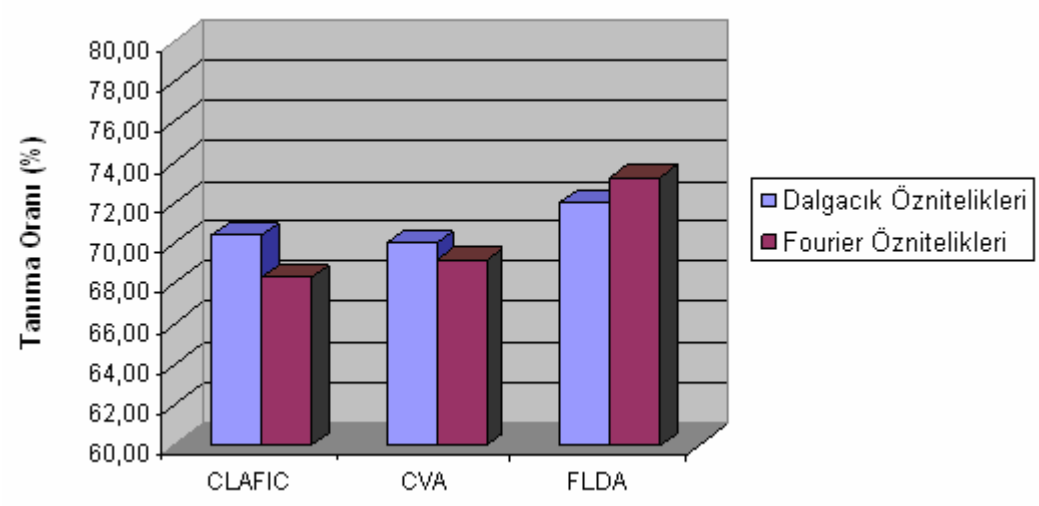
(a)

Sınıf	Dalgacık	Fourier
/b-e/	46.60	66.60
/s-e/	73.30	90.00
/d-e/	53.30	46.60
/jh-e/	96.60	93.30
/p-e/	53.30	33.30
/t-e/	66.60	63.30
/v-e/	86.60	86.60
/z-e/	83.30	73.30
Ortalama	70.00	69.10

(b)

Sınıf	Dalgacık	Fourier
/b-e/	63.30	63.30
/s-e/	66.60	80.00
/d-e/	60.00	56.60
/jh-e/	96.60	86.60
/p-e/	60.00	56.60
/t-e/	73.30	80.00
/v-e/	86.60	86.60
/z-e/	70.00	76.60
Ortalama	72.00	73.30

(c)

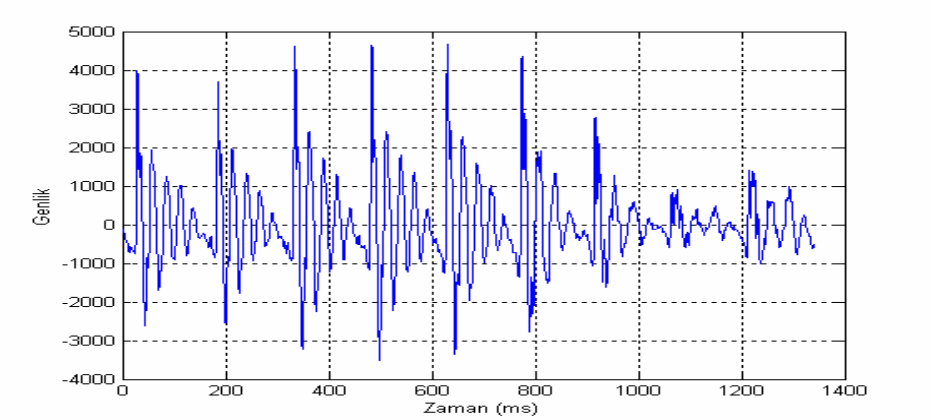


Şekil 5.4. E-SET veritabanı: Dalgacık ve Fourier temelli özniteliklerin farklı sınıflandırıcılar ile sağladığı ortalama tanıma oranları

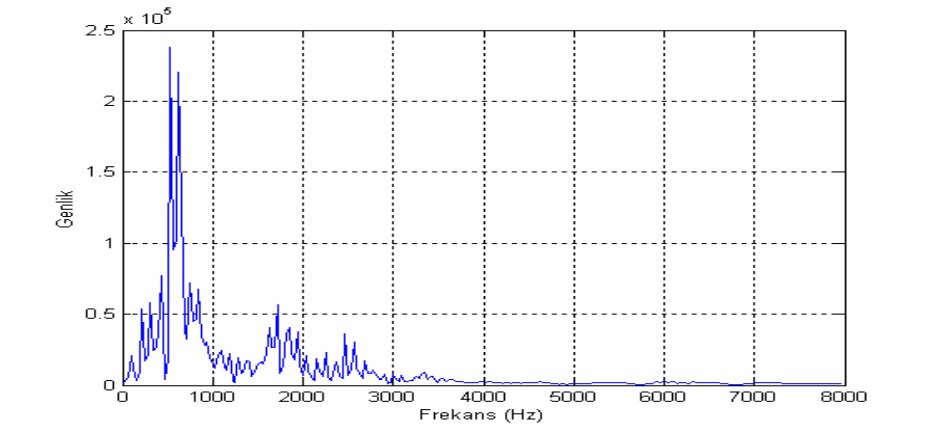
Öznitelik çıkarma deneylerinde kullanılan üçüncü ve son veritabanı VOWEL'dır. Şekil 5.5'te, bu veritabanına ait örnek bir sinyalin zaman bölgesi, Fourier dönüşümü ve dalgacık katsayıları görülmektedir.

VOWEL veritabanına ait sınıflandırma deneylerinin sonuçları Çizelge 5.9 ve Şekil 5.6'da görülmektedir. Çizelge 5.9'da, CLAFIC, CVA ve FLDA altuzay sınıflandırıcılar üzerinde Dalgacık ve Fourier temelli öznitelikler ile sınıf bazında elde edilen tanıma oranları, Şekil 5.6'da ise üç sınıflandırıcıya ait ortalama tanıma oranları karşılaştırmalı olarak verilmiştir. Sonuçlardan anlaşıldığı üzere, bu kez Fourier dönüşümü daha başarılı bir tanıma başarımı sağlamıştır. Üç sınıflandırıcıdan ikisinde, Fourier öznitelikleriyle çok daha yüksek tanıma oranlarına ulaşılmıştır.

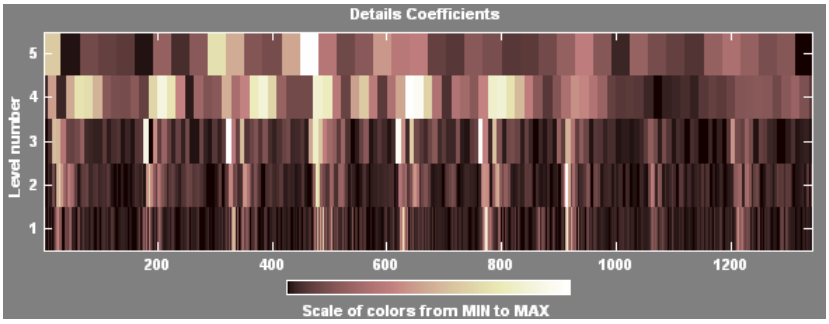
Diğer iki veritabanında elde edilen sonuçlara ters olan bu durum şu şekilde açıklanabilir: VOWEL veritabanı içerisindeki ses sinyalleri hem kısa sürelidir hem de bu sinyaller sesli fonemlere ait olduğu için sinüs benzeri periyodik yapıdadırlar (Bkz. Şekil 5.5-a). Bu sinyal karakteristiğinden ötürü, düzgün sinüs sinyaline kıyasla düzensiz yapıdaki ana dalgacık, sesli fonemleri temsil etmekte biraz daha zayıf kalmıştır. Böylece, Fourier dönüşümü dalgacık dönüşümünden daha başarılı olmuştur.



(a)



(b)



(c)

Şekil 5.5. VOWEL veritabanından örnek bir sinyal: /eh/ (a) Zaman bölgesi (b) Fourier dönüşümü (c) Dalgacık katsayıları

Çizelge 5.9. VOWEL veritabanı: Dalgacık ve Fourier dönüşümü temelli özniteliklerin (a) CLAFIC (b) CVA (c) FLDA tanıma oranları (%)

Sınıf	Dalgacık	Fourier
/aa/	80.00	76.00
/eh/	48.00	72.00
/iy/	84.00	80.00
/ow/	62.00	74.00
/uw/	70.00	82.00
Ortalama	68.80	76.80

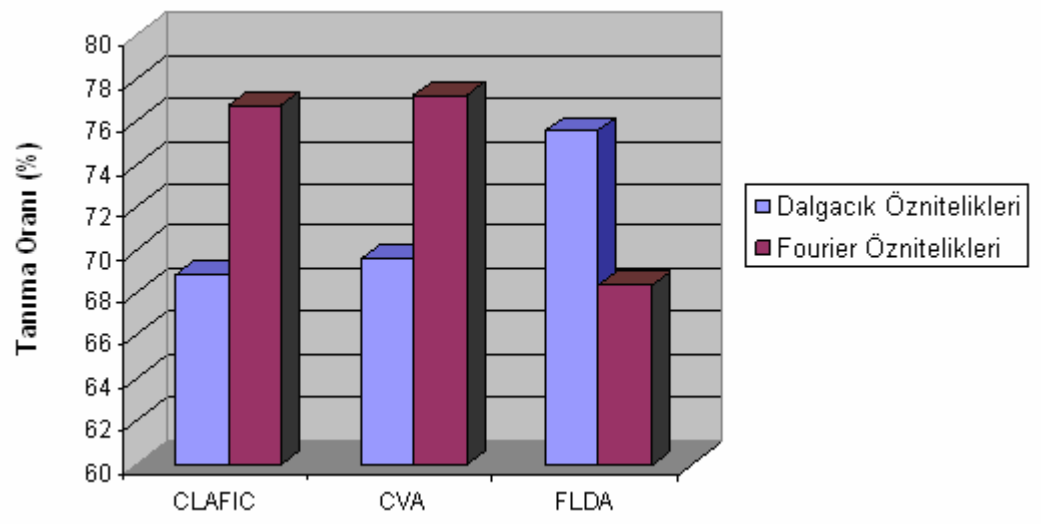
(a)

Sınıf	Dalgacık	Fourier
/aa/	82.00	84.00
/eh/	60.00	68.00
/iy/	80.00	78.00
/ow/	52.00	72.00
/uw/	74.00	84.00
Ortalama	69.60	77.20

(b)

Sınıf	Dalgacık	Fourier
/aa/	82.00	76.00
/eh/	64.00	50.00
/iy/	88.00	74.00
/ow/	58.00	62.00
/uw/	86.00	80.00
Ortalama	75.60	68.40

(c)



Şekil 5.6. VOWEL veritabanı: Dalgacık ve Fourier temelli özniteliklerin farklı sınıflandırıcılar ile sağladığı ortalama tanıma oranları (%)

5.3 Öznitelik Seçme Deneyleri

Deneysel çalışmaların bu bölümü, öznitelik seçme konusuyla ilgili yapılan deneyleri ve bu deneylere dair sonuçları içermektedir. Bu doğrultuda, Bölüm 3'te tanıtılmış olan altuzay temelli ayrılabilirlik ölçüleri ile klasik ayrılabilirlik ölçülerinin tek-değişkenli formlarının öznitelik seçme başarımları, sınıflandırma hassasiyeti ve boyut indirgeme oranı açısından karşılaştırılmıştır. Bu karşılaştırma için farklı sayı ve yapıda özniteliğe sahip E-SET, VEHICLE, PROTEIN ve POWER veritabanları kullanılmıştır. Bunun yanısıra, POWER veritabanı referans alınarak çok-değişkenli öznitelik seçme yöntemlerinin başarımı incelenmiş ve elde edilen bu sonuçlar ile tez çalışmasında önerilen tek-değişkenli yaklaşım sonuçları çeşitli açılardan karşılaştırılmıştır.

Özniteliklerin Ayırdedicilik Güçleri

Öncelikle, uzaksaklık, Bhattacharyya, CS ve FS ayrılabilirlik ölçüleri yardımıyla özniteliklerin bireysel ayırdedicilik dereceleri belirlenmiştir. Deneysel çalışmalarda,

dönüşmüş uzaksaklık ve Jeffries-Matusita ölçülerinin sırasıyla uzaksaklık ve Bhattacharyya ile aynı sonuçlara sahip olduğu belirlendiği için deney sonuölarında bu ölçülere yer verilmemiştir. Çizelge 5.10 – 5.13 , sırasıyla E-SET, VEHICLE, PROTEIN ve POWER veritabanındaki özniteliklerin, her bir ayrılabilirlik ölçüsü için ayırdedicilik (önem) sırasını göstermektedir. Görüldüğü üzere, uzaksaklık ve Bhattacharyya genelde benzer sıralamayı vermiş, ancak CS ve FS ölçüleri ile farklı sıralamalar elde edilmiştir. Başka bir ifadeyle, altuzay temelli ölçüler ile bulunan yüksek ya da düşük ayırdedici öznitelikler, klasik ölçülerle bulunanlardan farklıdır.

Tanıma ve Boyut İndirgeme Analizi

Öznitelik önem sıralamasını takiben, bu sıralamalardan hangisinin daha uygun olabileceğini doğrulamak için her bir sıralamaya göre çeşitli sayıda özniteliğin seçimiyle sağlanan sınıflandırma hassasiyetleri bulunmuştur. Bu noktada bir hatırlatma olarak, Bölüm 3'te bahsedildiği üzere uzaksaklık ve Bhattacharyya ölçüleri Bayes hatasını enküçükleme prensibine göre çalışır. Bu açıdan, altuzay temelli seçim yöntemleriyle elde edilen Bayes sınıflandırma başarımlarının klasik yöntemlerle karşılaştırmasını yapmak akıllıca olacaktır. Buna ilaveten, önerilen yöntemlerin farklı sınıflandırıcılar için de geçerli olduğunu doğrulamak için, Bayes yanında FLDA sınıflandırıcı ile de tanıma analizi yapılmıştır. Bunun sonucunda, sırasıyla E-SET, VEHICLE, PROTEIN ve POWER veritabanları için çeşitli sayılarda öznitelik seçilerek Çizelge 5.14 – 5.17'de verilmekte olan ortalama tanıma oranlarına ulaşılmıştır. Veritabanlarının çoğunda, düşük boyutlardaki öznitelik altkümeleri, tüm öznitelik kümesine göre daha yüksek sınıflandırma hassasiyeti sağlamıştır.

Farklı öznitelik sayılarıyla ulaşılan çok sayıda tanıma oranını ayrı ayrı değerlendirerek hangi seçim yönteminin daha etkili olduğunu söylemek kolay olmayabilir. Bu sebeple, sınıflandırma hassasiyeti ve boyut bilgisini birleştiren bir puanlama sistemi geliştirilmiştir. Bu sistemde, düşük boyutlarda yüksek tanıma oranı elde etmenin genel puana katkısı yüksek boyutlara göre daha yüksektir. Bu kural doğrultusunda,

$$Puan = \frac{1}{k}\sum_{i=1}^{k}\left(\frac{\dim_{toplam}}{\dim_i}\right)R_i \tag{5.1}$$

şeklinde bir formül tanımlanmıştır. Bu formülde, k , toplam kaç adet öznitelik altkümesi boyutunda tanıma işlemi yapıldığını; $\dim_i$, i'nci tanıma işlemindeki öznitelik altkümesi boyutunu; $\dim_{toplam}$, tüm öznitelik kümesinin boyutunu; R_i ise i'nci tanıma işleminde elde edilen tanıma oranını ifade etmektedir. Böylece, her boyutta elde edilen tanıma oranları, o boyuta göre ağırlıklandırılır ve tüm ağırlıklı tanıma oranlarının ortalaması bulunur. Bu ağırlıklı ortalama puanı, ilgili öznitelik seçme yönteminin genel başarımını temsil eder.

Tasarlanan bu değerlendirme sistemi ile her bir seçim yöntemine ait puanlar, tüm veritabanları için hem Bayes hem de FLDA sınıflandırıcıyla elde edilmiştir. Çizelge 5.14 – 5.17'de sırasıyla E-SET, VEHICLE, PROTEIN ve POWER veritabanları için tanıma oranlarının yanısıra altküme boyutların ağırlıkları ve ağırlıklı tanıma oranlarıyla hesaplanan puanlar da görülmektedir. Her durumda ulaşılan en yüksek puan, ilgili tabloda koyu renk ile gösterilmiştir. Şekil 5.7'de ise, bu puanlar her bir veritabanı için karşılaştırmalı olarak yer almaktadır. Bu şekilden anlaşılacağı gibi veritabanlarının tamamında, altuzay temelli öznitelik seçim yöntemleri diğer yöntemlerden daha yüksek başarıma sahiptir. Bu durum, kullanılan sınıflandırıcıların her ikisi için de geçerlidir.

Çizelge 5.10. E-SET veritabanı: (a) Uzaksaklık (b) Bhattacharyya (c) CS (d) FS ayrılabilirlik ölçüleri için azalan yönde öznitelik önem sıralaması

23, 24, 22, 28, 21, 17, 16, 18, 15, 72, 1, 26, 2, 37, 27, 19, 39, 34, 44, 30, 79, 4, 14, 8, 10, 9, 12, 76, 66, 78, 69, 36, 47, 40, 29, 20, 86, 48, 13, 57, 50, 94, 75, 25, 31, 87, 3, 33, 45, 58, 43, 93, 95, 64, 42, 52, 41, 59, 38, 80, 77, 90, 84, 46, 82, 54, 53, 92, 6, 70, 7, 35, 5, 85, 63, 32, 83, 68, 51, 89, 96, 11, 55, 81, 73, 67, 74, 65, 56, 62, 91, 49, 61, 88, 71, 60

(a)

23, 24, 22, 28, 21, 16, 17, 18, 1, 15, 72, 26, 2, 27, 19, 37, 39, 8, 79, 4, 34, 14, 10, 30, 44, 9, 36, 78, 29, 12, 76, 66, 40, 69, 47, 20, 25, 94, 31, 75, 13, 48, 50, 86, 87, 95, 33, 3, 57, 93, 58, 45, 41, 52, 77, 43, 64, 42, 80, 84, 59, 46, 38, 90, 53, 82, 54, 6, 92, 35, 32, 83, 70, 85, 63, 7, 5, 68, 89, 51, 11, 96, 81, 55, 73, 74, 67, 65, 56, 91, 62, 61, 49, 71, 88, 60

(b)

21, 3, 2, 5, 20, 42, 16, 59, 45, 4, 71, 11, 10, 96, 88, 18, 39, 80, 46, 19, 29, 62, 35, 22, 13, 38, 48, 33, 79, 6, 92, 27, 32, 94, 82, 58, 68, 65, 83, 9, 44, 69, 57, 15, 95, 61, 24, 72, 36, 75, 86, 37, 8, 50, 66, 93, 28, 51, 1, 31, 91, 87, 47, 74, 84, 7, 81, 17, 90, 85, 23, 14, 53, 34, 70, 41, 76, 25, 49, 64, 73, 67, 60, 63, 26, 56, 77, 54, 89, 52, 12, 78, 55, 40, 30, 43

(c)

38, 32, 22, 61, 95, 29, 62, 69, 15, 57, 21, 53, 71, 7, 26, 65, 94, 19, 79, 5, 52, 41, 82, 45, 23, 35, 87, 47, 33, 88, 89, 44, 11, 14, 83, 12, 28, 90, 6, 17, 81, 54, 40, 37, 59, 93, 13, 68, 66, 75, 55, 67, 31, 74, 48, 85, 80, 91, 1, 39, 9, 50, 60, 18, 42, 2, 77, 46, 43, 58, 70, 56, 72, 8, 3, 34, 27, 92, 64, 96, 86, 30, 25, 63, 36, 16, 49, 78, 76, 24, 84, 10, 51, 20, 4, 73

(d)

Çizelge 5.11. VEHICLE veritabanı: (a) Uzaksaklık (b) Bhattacharyya (c) CS (d) FS ayrılabilirlik ölçüleri için azalan yönde öznitelik önem sıralaması

1, 5, 2, 18, 15, 6, 3, 17, 12, 13, 8, 7, 16, 11, 10, 9, 4, 14
(a)
1, 5, 18, 15, 2, 6, 3, 17, 13, 10, 16, 12, 8, 7, 11, 9, 14, 4
(b)
8, 17, 14, 9, 7, 6, 18, 11, 12, 5, 2, 4, 3, 16, 1, 10, 13, 15
(c)
2, 1, 7, 5, 9, 8, 6, 17, 18, 4, 12, 13, 15, 10, 3, 16, 14, 11
(d)

Çizelge 5.12. PROTEIN veritabanı: (a) Uzaksaklık (b) Bhattacharyya (c) CS (d) FS ayrılabilirlik ölçüleri için azalan yönde öznitelik önem sıralaması

6, 2, 8, 7, 1, 4, 3, 5
(a)
6, 2, 8, 1, 7, 4, 3, 5
(b)
1, 2, 4, 8, 3, 7, 5, 6
(c)
8, 3, 7, 2, 1, 4, 5, 6
(d)

Çizelge 5.13. POWER veritabanı: (a) Uzaksaklık (b) Bhattacharyya (c) CS (d) FS ayrılabilirlik ölçüleri için azalan yönde öznitelik önem sıralaması

11, 1, 9, 15, 13, 18, 4, 16, 3, 14, 19, 2, 12, 17, 6, 10, 7, 8, 5
(a)
11, 1, 9, 18, 15, 13, 16, 14, 3, 4, 19, 2, 17, 6, 12, 10, 7, 8, 5
(b)
12, 15, 2, 1, 17, 14, 7, 10, 13, 8, 4, 3, 5, 9, 6, 18, 11, 19, 16
(c)
11, 12, 13, 15, 10, 9, 8, 3, 1, 2, 6, 5, 7, 4, 14, 17, 16, 18, 19
(d)

Çizelge 5.14. E-SET veritabanı: Ayrılabilirlik ölçüleri için farklı öznitelik sayılarıyla elde edilen sınıflandırma başarımı (a) Bayes (b) FLDA sınıflandırıcı

Ağırlık	**Boyut**	**D**	**B**	**CS**	**FS**
9,60	10	57,08	58,75	70,83	51,67
4,80	20	69,17	68,75	69,58	65,00
3,20	30	71,67	70,42	69,17	69,17
2,40	40	70,83	74,58	69,58	63,33
1,92	50	68,33	69,17	70,00	65,83
1,60	60	65,83	63,75	67,08	38,75
1,37	70	34,58	32,92	25,00	59,17
1,20	80	31,67	32,50	22,08	28,75
1,07	90	28,75	32,08	27,50	37,50
1,00	96	37,92	37,92	37,92	37,92
	Puan	167	169	**177**	156

(a)

Ağırlık	**Boyut**	**D**	**B**	**CS**	**FS**
9,60	10	53,75	52,08	63,75	50,42
4,80	20	64,58	65,83	71,25	61,67
3,20	30	67,08	66,25	73,75	69,17
2,40	40	67,92	69,17	71,67	68,75
1,92	50	70,83	67,50	72,92	68,75
1,60	60	71,67	71,67	72,08	68,33
1,37	70	70,83	70,42	72,50	71,67
1,20	80	72,92	72,92	71,25	71,25
1,07	90	70,42	72,50	72,08	70,83
1,00	96	72,08	72,08	72,08	72,08
	Puan	179	177	**195**	174

(b)

Çizelge 5.15. VEHICLE veritabanı: Ayrılabilirlik ölçüleri için farklı öznitelik sayılarıyla elde edilen sınıflandırma başarımı (a) Bayes (b) FLDA sınıflandırıcı

Ağırlık	**Boyut**	**D**	**B**	**CS**	**FS**
9,00	2	43,00	43,00	43,00	48,50
6,00	3	46,50	48,00	57,00	60,50
4,50	4	53,00	52,50	62,50	60,50
3,60	5	60,00	60,00	64,50	59,50
3,00	6	59,00	59,00	67,50	61,50
2,57	7	63,00	63,00	69,50	63,00
2,25	8	67,00	67,00	69,50	67,00
2,00	9	74,00	68,50	72,00	74,50
1,80	10	75,00	74,00	76,50	81,00
1,64	11	79,00	79,00	76,00	81,00
1,50	12	82,00	85,00	77,50	82,00
1,38	13	83,50	84,00	78,00	83,00
1,29	14	82,50	86,00	79,50	86,50
1,20	15	85,00	85,00	76,50	86,00
1,13	16	86,00	86,00	83,00	85,50
1,06	17	88,50	87,00	89,00	87,50
1,00	18	89,00	89,00	89,00	89,00
	Puan	162	162	169	**173**

(a)

Ağırlık	**Boyut**	**D**	**B**	**CS**	**FS**
9,00	2	34,50	34,50	32,50	35,50
6,00	3	48,50	44,00	45,50	55,50
4,50	4	47,00	41,50	52,50	55,50
3,60	5	53,00	53,00	56,00	55,50
3,00	6	50,00	50,00	57,50	56,50
2,57	7	57,50	57,50	57,00	58,50
2,25	8	59,50	59,50	57,00	59,50
2,00	9	67,00	62,00	63,00	70,00
1,80	10	71,50	67,00	70,50	72,00
1,64	11	70,00	68,50	68,00	72,00
1,50	12	71,00	72,50	71,50	73,50
1,38	13	71,00	76,00	72,00	70,50
1,29	14	71,00	75,00	74,50	77,50
1,20	15	74,50	74,50	74,00	77,50
1,13	16	74,50	74,50	76,00	77,00
1,06	17	77,50	75,50	78,50	78,50
1,00	18	78,50	78,50	78,50	78,50
	Puan	144	140	144	**152**

(b)

Çizelge 5.16. PROTEIN veritabanı: Ayrılabilirlik ölçüleri için farklı öznitelik sayılarıyla elde edilen sınıflandırma başarımı (a) Bayes (b) FLDA sınıflandırıcı

Ağırlık	**Boyut**	**D**	**B**	**CS**	**FS**
4,00	2	25,00	25,00	41,88	56,88
2,67	3	25,00	25,00	53,75	56,88
2,00	4	25,00	25,00	50,63	57,50
1,60	5	25,00	25,00	72,50	62,50
1,33	6	25,00	25,00	71,88	71,88
1,14	7	25,00	25,00	54,38	54,38
1,00	8	25,00	25,00	25,00	25,00
	Puan	49	49	102	**111**

(a)

Ağırlık	**Boyut**	**D**	**B**	**CS**	**FS**
4,00	2	35,63	35,63	35,63	48,13
2,67	3	41,25	41,25	48,13	53,13
2,00	4	41,88	45,00	54,38	59,38
1,60	5	48,75	48,75	73,13	63,75
1,33	6	51,88	51,88	71,88	71,88
1,14	7	71,88	71,88	72,50	72,50
1,00	8	71,88	71,88	71,88	71,88
	Puan	91	92	107	**115**

(b)

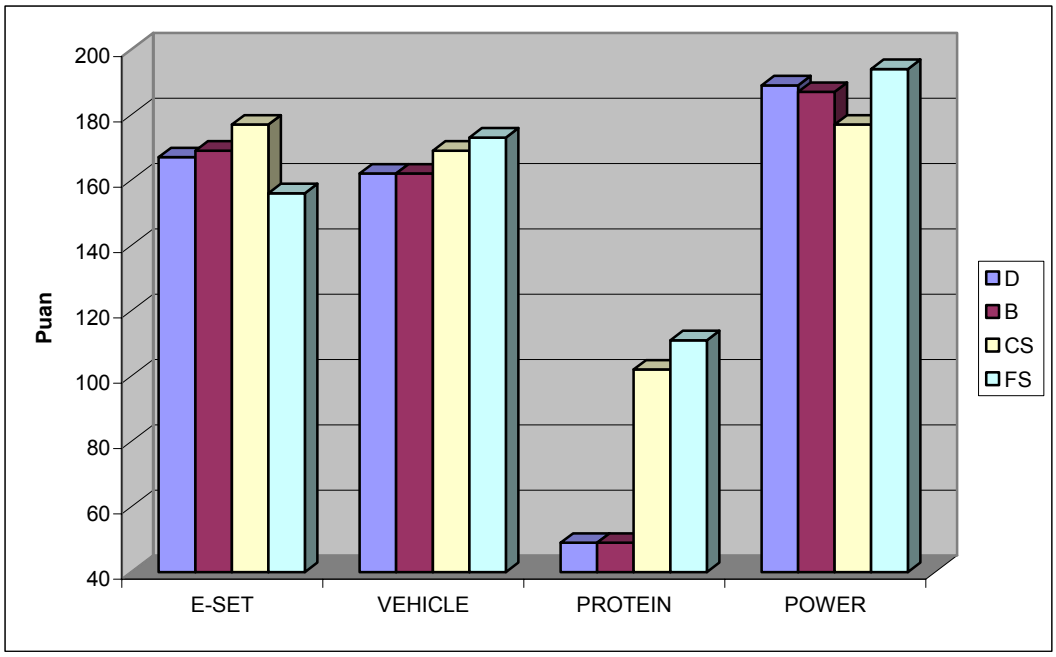
Çizelge 5.17. POWER veritabanı: Ayrılabilirlik ölçüleri için farklı öznitelik sayılarıyla elde edilen sınıflandırma başarımı (a) Bayes (b) FLDA sınıflandırıcı

Ağırlık	Boyut	D	B	CS	FS
9,50	2	64,17	64,17	55,00	60,00
6,33	3	72,50	72,50	64,17	72,50
4,75	4	68,33	64,17	66,67	78,33
3,80	5	75,00	67,50	65,00	75,00
3,17	6	75,00	75,00	70,83	74,17
2,71	7	72,50	71,67	70,00	77,50
2,38	8	72,50	74,17	70,00	75,83
2,11	9	73,33	75,83	70,83	78,33
1,90	10	76,67	76,67	70,83	74,17
1,73	11	77,50	77,50	70,83	77,50
1,58	12	74,17	74,17	72,50	77,50
1,46	13	74,17	70,00	68,33	75,00
1,36	14	68,33	69,17	65,83	72,50
1,27	15	67,50	67,50	70,00	76,67
1,19	16	65,00	65,00	70,83	75,00
1,12	17	63,33	63,33	75,00	70,83
1,06	18	64,17	64,17	71,67	70,00
1,00	19	67,50	67,50	67,50	67,50
	Puan	189	187	177	**194**

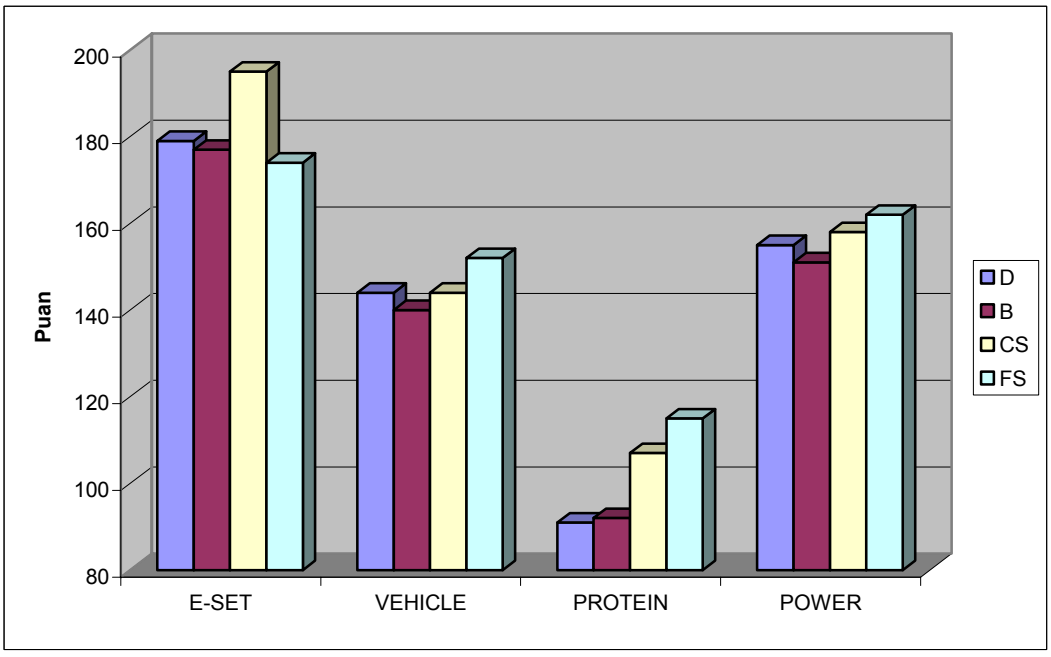
(a)

Ağırlık	Boyut	D	B	CS	FS
9,50	2	50,00	50,00	50,00	50,83
6,33	3	48,33	48,33	61,67	61,67
4,75	4	50,83	41,67	58,33	62,50
3,80	5	59,17	55,00	62,50	59,17
3,17	6	61,67	61,67	65,83	59,17
2,71	7	62,50	67,50	64,17	60,83
2,38	8	65,83	61,67	66,67	63,33
2,11	9	67,50	62,50	59,17	65,83
1,90	10	67,50	67,50	56,67	68,33
1,73	11	71,67	71,67	55,00	66,67
1,58	12	70,83	70,83	62,50	67,50
1,46	13	63,33	63,33	60,00	66,67
1,36	14	61,67	60,83	60,83	65,00
1,27	15	61,67	61,67	59,17	60,00
1,19	16	59,17	59,17	60,00	61,67
1,12	17	56,67	56,67	60,00	60,83
1,06	18	58,33	58,33	56,67	57,50
1,00	19	58,33	58,33	58,33	58,33
	Puan	155	151	158	**162**

(b)



(a)



(b)

Şekil 5.7. Tüm veritabanları için karşılaştırmalı puanlar (a) Bayes (b) FLDA sınıflandırıcı

Öznitelik seçimi konusunda tek-değişkenli yöntemlerin birbirleriyle karşılaştırmasını takiben çok-değişkenli yöntemlerin başarımları da analiz edilmiştir. Bölüm 2'de de ifade edildiği üzere çok-değişkenli yaklaşımlar, özniteliklerin bireysel ayırdediciliklerini göstermez; bunun yerine, öznitelikler arasındaki olası ilintileri de göz önüne alıp öznitelikleri bir grup olarak inceler.

Bu çalışmaya özel olarak, sadece POWER veritabanına ait sonuçlara yer verilmektedir. Bu yöntemlerin ölçüt fonksiyonu olarak Bayes sınıflandırma hassasiyeti kullanılmıştır. Dolayısıyla, farklı bir sınıflandırıcı kullanarak burada verilenlerden farklı sonuçlara ulaşılması söz konusu olabilir.

Öncelikle, eniyi sonucu elde etmek ve bu sonucu diğer yöntemlerin referansı olarak kullanabilmek için ES yöntemi uygulanmış; daha sonra, alt eniyi çok-değişkenli seçim yöntemlerinin her birisi için sonuçlar alınmıştır. Elde edilen bulgular hem çok-değişkenli yöntemlerin kendi aralarında hem de tek-değişkenli yöntemler ile çeşitli açılardan karşılaştırılmıştır. Sırasıyla ES, SFS, SBS, SFFS, PTA(1,2), GSFS(2), GSBS(2) ve GA öznitelik seçim yöntemleriyle değişik boyutlarda ulaşılan en yüksek ortalama tanıma oranları ve bu oranları sağlayan öznitelik altkümeleri Çizelge 5.18'de görülmektedir. GA temelli öznitelik seçiminde, veritabanı 19 boyutlu öznitelik kümesiyle temsil edildiği için kromozom boyutu 19'dur. Bu çalışma için nüfus boyutu ve kuşak sayısı 50, çaprazlama olasılığı % 90, mutasyon olasılığı ise % 8 olarak tanımlanmıştır. Her bir seçim yöntemi için elde edilen en yüksek tanıma oranı, çizelgede koyu renkle belirtilmiştir.

ES yöntemine ait sonuçlardan görüldüğü üzere eniyi öznitelik altkümesi 8 boyutlu {1, 7, 8, 11, 12, 13, 18, 19} indisli özniteliklerden oluşmakta ve %85.83 tanıma oranı sağlamaktadır. Bu oran, 19 özniteliğin tamamının kullanılmasıyla ulaşılan orandan (%67.50) oldukça yüksektir. Dolayısıyla, etkili öznitelik seçiminin önemi net olarak görülmektedir. Alt eniyi seçim yöntemlerinde ise 8 ya da daha düşük boyutlarda bu tanıma oranına ulaşmak mümkün olmamıştır. Ancak, bu yöntemlerin bazıları eniyi sonuca yakın değerlere ulaşabilmiştir.

Eniyi ES yöntemiyle hem çok-değişkenli hem de tek-değişkenli seçim yöntemlerini genel olarak karşılaştırabilmek için değişik analizlerden faydalanılmaktadır.

Benzerlik Oranı

Bu doğrultuda, ilk olarak, seçim yöntemlerinin 8 boyutta seçtiği öznitelikler değerlendirilerek, ES yöntemiyle seçilen özniteliklere benzerlik oranları bulunmuştur. Benzerlik oranı, bir öznitelik yöntemiyle seçilen özniteliklerin ne kadarının, ES yöntemiyle seçilen eniyi özniteliklerle uyuştuğunu gösterir:

$$Benzerlik = \frac{1}{\dim_{toplam}} \sum_{i=1}^{d} m_i \qquad m_i = \begin{cases} 0 & f^i_{yöntem} \notin f_{ES} \\ 1 & f^i_{yöntem} \in f_{ES} \end{cases} \tag{5.2}$$

Bu formülde, d, benzerlik oranının hesaplandığı boyutu (8); $f^i_{yöntem}$, belirli bir seçim yöntemiyle elde edilen d-boyutlu öznitelik altkümesindeki $i$nci özniteliği; f_{ES}, ES yöntemiyle seçilen öznitelik altkümesini; $\dim_{toplam}$ ise toplam öznitelik boyutunu ifade etmektedir. Bu yolla elde edilen benzerlik oranları, Çizelge 5.19'da verilmektedir. Bu çizelgeye göre, ES yöntemine en yüksek benzerlikteki (%62.50) öznitelikleri öneren yöntemler çok-değişkenli PTA ve GSFS olmakla birlikte, tek-değişkenli yaklaşımların benzerlikleri de (%50) çok-değişkenli yöntemlerin bazılarından daha yüksektir.

Sınıflandırma Başarımı

ES yöntemiyle karşılaştırmanın ikinci adımında, alt eniyi yöntemlerle elde edilen en yüksek tanıma oranları ve bu oranları sağlayan öznitelik sayıları değerlendirilmiştir. Çizelge 5.20'de görülen veriler, bu değerlendirmeleri özetlemektedir. Görüldüğü üzere, çok-değişkenli yaklaşımlar ile ulaşılan tanıma oranları, tek-değişkenli yaklaşımlara göre daha yüksektir. Bu durumun ortaya çıkmasının pek çok nedeni vardır. Öncelikle, çok-değişkenli yaklaşımlar öznitelikler arasındaki ilintileri de değerlendirmektedir. Ayrıca, bu yöntemlerin ölçüt fonksiyonu direkt olarak Bayes sınıflandırma hassasiyeti olarak

belirlenmiştir. Oysaki, tek-değişkenli yaklaşımlar sınıflandırıcıdan bağımsız çalışmaktadır. Buna rağmen, FS temelli tek-değişkenli öznitelik seçim yöntemi yardımıyla sadece 4 boyutta %78,33 tanıma oranına ulaşılmıştır. Bu değer ise eniyi ES yöntemiyle seçilen özniteliklerin sağladığı %81.67 oranına çok yakındır. Daha farklı bir sınıflandırıcı kullanarak çok daha iyi sonuçlar elde etmekte mümkün olabilir.

Son olarak, her bir seçim yönteminin işlem süresi değerlendirilmiştir. İşlem süresi ölçümleri, Intel Pentium Dual Core 2 GHz işlemciye sahip bir bilgisayar üzerinde yapılmıştır. Çizelge 5.20'de verilen değerlere göre, altuzay temelli yöntemler 1 saniyenin altındaki işlem süreleriyle diğer yöntemlerin açık ara önüne geçmektedir. ES yöntemi eniyi çözümü sunmasına rağmen, 18 günü geçen işlem süresi yüzünden tercih edilebilir bir yöntem olmaktan uzaktır. Bu sebeple, pratik uygulamalarda çoğu zaman alt eniyi öznitelik seçim yöntemleri ön plana çıkmaktadır. Bunların başarımını etkileyen faktörler ise görüldüğü üzere öznitelikler arasındaki ilintilerin derecesi ve kriter fonksiyonu seçimidir. Sonuç olarak, gerek tek-değişkenli gerekse çok değişkenli öznitelik seçim yöntemleri yardımıyla hem öznitelik kümesinin boyutu indirgenip hem de sınıflandırma başarımının önemli oranlarda artırılabildiği gösterilmiştir.

Çizelge 5.18. POWER veritabanı: Çok değişkenli öznitelik seçim yöntemleriyle elde edilen öznitelik altkümeleri ve bunların sınıflandırma başarımları: (a) ES (b) SFS (c) SBS (d) SFFS (e) PTA (f) GSFS (g) GSBS (h) GA

Boyut	**Tanıma (%)**	**Seçilen Öznitelikler**
1	50,00	18
2	73,33	1, 13
3	78,33	1, 11, 13
4	81,67	1, 8, 11, 13
5	82,50	1, 8, 11, 12, 13
6	83,33	1, 3, 11, 13, 14, 18
7	83,33	1, 7, 11, 12, 13, 18, 19
8	**85,83**	1, 7, 8, 11, 12, 13, 18, 19
9	85,83	1, 3, 7, 8, 11, 13, 14, 18, 19
10	85,00	1, 3, 6, 7, 8, 11, 12, 14, 18, 19
11	85,00	1, 3, 6, 7, 8, 11, 12, 14, 15, 18, 19
12	85,00	1, 2, 3, 6, 7, 8, 10, 11, 12, 14, 18, 19
13	85,83	1, 2, 3, 6, 7, 8, 10, 11, 12, 13, 14, 15, 18
14	83,33	1, 2, 3, 5, 6, 7, 8, 11, 12, 14, 15, 16, 18, 19
15	81,67	1, 2, 3, 4, 5, 6, 7, 8, 10, 11, 12, 13, 14, 15, 18
16	80,00	1, 2, 3, 4, 5, 6, 7, 8, 10, 11, 12, 13, 14, 15, 17, 18
17	78,33	1, 2, 3, 4, 5, 6, 7, 8, 10, 11, 12, 13, 14, 15, 17, 18, 19
18	72,50	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 19
19	67,50	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19

(a)

Boyut	**Tanıma (%)**	**Seçilen Öznitelikler**
1	50,00	18
2	65,83	17, 18
3	68,33	8, 17, 18
4	70,83	4, 8, 17, 18
5	73,33	4, 8, 14, 17, 18
6	71,67	4, 7, 8, 14, 17, 18
7	73,33	4, 7, 8, 14, 15, 17, 18
8	74,17	3, 4, 7, 8, 14, 15, 17, 18
9	75,83	1, 3, 4, 7, 8, 14, 15, 17, 18
10	78,33	1, 3, 4, 7, 8, 14, 15, 16, 17, 18
11	78,33	1, 3, 4, 6, 7, 8, 14, 15, 16, 17, 18
12	**79,17**	1, 3, 4, 6, 7, 8, 11, 14, 15, 16, 17, 18
13	79,17	1, 3, 4, 6, 7, 8, 11, 14, 15, 16, 17, 18, 19
14	77,50	1, 3, 4, 6, 7, 8, 11, 12, 14, 15, 16, 17, 18, 19
15	77,50	1, 3, 4, 5, 6, 7, 8, 11, 12, 14, 15, 16, 17, 18, 19
16	75,83	1, 2, 3, 4, 5, 6, 7, 8, 11, 12, 14, 15, 16, 17, 18, 19
17	73,33	1, 2, 3, 4, 5, 6, 7, 8, 10, 11, 12, 14, 15, 16, 17, 18, 19
18	70,00	1, 2, 3, 4, 5, 6, 7, 8, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19
19	67,50	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19

(b)

Boyut	Tanıma (%)	Seçilen Öznitelikler
1	30,00	14
2	66,67	13, 14
3	75,00	11, 13, 14
4	75,83	10, 11, 13, 14
5	78,33	3, 10, 11, 13, 14
6	79,17	3, 10, 11, 13, 14, 15
7	82,50	2, 3, 10, 11, 13, 14, 15
8	80,83	2, 3, 6, 10, 11, 13, 14, 15
9	81,67	2, 3, 5, 6, 10, 11, 13, 14, 15
10	82,50	2, 3, 4, 5, 6, 10, 11, 13, 14, 15
11	82,50	2, 3, 4, 5, 6, 10, 11, 13, 14, 15, 19
12	**83,33**	2, 3, 4, 5, 6, 8, 10, 11, 13, 14, 15, 19
13	81,67	2, 3, 4, 5, 6, 7, 8, 10, 11, 13, 14, 15, 19
14	80,00	2, 3, 4, 5, 6, 7, 8, 10, 11, 13, 14, 15, 16, 19
15	77,50	2, 3, 4, 5, 6, 7, 8, 10, 11, 13, 14, 15, 16, 17, 19
16	75,83	2, 3, 4, 5, 6, 7, 8, 10, 11, 12, 13, 14, 15, 16, 17, 19
17	74,17	2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 19
18	72,50	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 19
19	67,50	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19

(c)

Boyut	Tanıma (%)	Seçilen Öznitelikler
1	50,00	18
2	70,00	4, 15
3	75,00	4, 11, 12
4	76,67	4, 5, 11, 12
5	78,33	3, 4, 13, 14, 15
6	80,00	3, 4, 11, 13, 14, 15
7	80,83	2, 4, 5, 6, 8, 11, 12
8	80,83	2, 3, 4, 5, 6, 8, 11, 12
9	81,67	2, 3, 4, 5, 6, 8, 11, 12, 13
10	82,50	2, 3, 4, 5, 6, 8, 11, 13, 14, 15
11	83,33	2, 3, 4, 5, 6, 8, 11, 12, 13, 14, 15
12	**84,17**	2, 3, 4, 5, 6, 10, 11, 13, 14, 15, 18, 19
13	83,33	2, 3, 4, 5, 6, 7, 10, 11, 13, 14, 15, 18, 19
14	83,33	1, 2, 3, 4, 5, 6, 7, 8, 11, 13, 14, 15, 18, 19
15	81,67	1, 2, 3, 4, 5, 6, 7, 8, 11, 12, 13, 14, 15, 18, 19
16	80,00	1, 2, 3, 4, 5, 6, 7, 8, 11, 12, 13, 14, 15, 17, 18, 19
17	78,33	1, 2, 3, 4, 5, 6, 7, 8, 10, 11, 12, 13, 14, 15, 17, 18, 19
18	72,50	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 19
19	67,50	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19

(d)

Boyut	Tanıma (%)	Seçilen Öznitelikler
1	48,33	16
2	73,33	1, 13
3	78,33	1, 11, 13
4	78,33	1, 11, 13, 14
5	79,17	2, 11, 13, 14, 15
6	83,33	1, 3, 11, 13, 14, 18
7	80,83	1, 3, 11, 13, 14, 15, 18
8	81,67	1, 3, 11, 12, 13, 14, 15, 18
9	80,83	1, 2, 3, 11, 13, 14, 15, 18, 19
10	**85,00**	2, 3, 5, 6, 11, 13, 14, 15, 18, 19
11	83,33	2, 3, 4, 5, 6, 11, 13, 14, 15, 18, 19
12	84,17	2, 3, 4, 5, 6, 8, 11, 13, 14, 15, 18, 19
13	83,33	2, 3, 4, 5, 6, 8, 10, 11, 13, 14, 15, 18, 19
14	83,33	1, 2, 3, 4, 5, 6, 7, 8, 11, 13, 14, 15, 18, 19
15	81,67	1, 2, 3, 4, 5, 6, 7, 8, 10, 11, 13, 14, 15, 18, 19
16	80,00	1, 2, 3, 4, 5, 6, 7, 8, 11, 12, 13, 14, 15, 17, 18, 19
17	78,33	1, 2, 3, 4, 5, 6, 7, 8, 10, 11, 12, 13, 14, 15, 17, 18, 19
18	71,67	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 17, 18, 19
19	67,50	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19

(e)

Boyut	Tanıma (%)	Seçilen Öznitelikler
2	73,33	1, 13
4	81,67	1, 8, 11, 13
6	**82,50**	1, 5, 8, 11, 12, 13
8	80,00	1, 3, 5, 8, 9, 11, 12, 13
10	80,00	1, 2, 3, 5, 7, 8, 9, 11, 12, 13
12	76,67	1, 2, 3, 5, 7, 8, 9, 10, 11, 12, 13, 18
14	75,00	1, 2, 3, 5, 7, 8, 9, 10, 11, 12, 13, 14, 15, 18
16	76,67	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 18
18	71,67	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 17, 18, 19
19	67,50	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19

(f)

Boyut	Tanıma (%)	Seçilen Öznitelikler
1	25,00	12
3	75,83	7, 11, 12
5	77,50	1, 7, 8, 11, 12
7	81,67	1, 7, 8, 11, 12, 18, 19
9	**85,00**	1, 3, 7, 8, 11, 12, 14, 18, 19
11	85,00	1, 3, 6, 7, 8, 11, 12, 14, 15, 18, 19
13	84,17	1, 2, 3, 6, 7, 8, 11, 12, 13, 14, 15, 18, 19
15	81,67	1, 2, 3, 4, 5, 6, 7, 8, 11, 12, 13, 14, 15, 18, 19
17	78,33	1, 2, 3, 4, 5, 6, 7, 8, 10, 11, 12, 13, 14, 15, 17, 18, 19
19	67,50	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19

(g)

Boyut	Tanıma (%)	Seçilen Öznitelikler
13	**85,83**	1, 2, 3, 6, 7, 8, 10, 11, 12, 13, 14, 15, 18

(h)

Çizelge 5.19. Alt eniyi yöntemlerle seçilen özniteliklerin ES yöntemiyle seçilen eniyi öznitelik altkümesine benzerlik oranları (%)

Yöntem	ES ile Benzerlik Oranı (%)
CS	37,50
FS	50,00
ES	100,00
SFS	37,50
SBS	25,00
SFFS	37,50
PTA	62,50
GSFS	62,50

Çizelge 5.20. Tek-değişkenli ve çok-değişkenli öznitelik seçim yöntemleriyle elde edilen en iyi tanıma sonuçları

Yöntem	Tanıma (%)	Boyut	Seçilen Öznitelikler	Süre (sn)
CS	75,00	17	1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,17,18	< 1
FS	78,33	4	11,12,13,15	< 1
ES	85,83	8	1,7,8,11,12,13,18,19	1.600.000
SFS	79,17	12	1,3,4,6,7,8,11,14,15,16,17,18	478
SBS	83,33	12	2,3,4,5,6,8,10,11,13,14,15,19	502
SFFS	84,17	12	2,3,4,5,6,10,11,13,14,15,18,19	2.122
PTA	85,00	10	2,3,5,6,11,13,14,15,18,19	3.484
GSFS	82,50	6	1,5,8,11,12,13	1.543
GSBS	85,00	9	1,3,7,8,11,12,14,18,19	1.652
GA	85,83	13	1,2,3,6,7,8,10,11,12,13,14,15,18	7.503

5.4 Sınıflandırma Deneyleri

Deneysel çalışmaların bu bölümünde, altuzay sınıflandırma konusuyla ilgili yapılan deneyler ve bu deneylere ait sonuçlar yer almaktadır. Bu doğrultuda, tez kapsamında geliştirilmiş olan genetik algoritma temelli altuzay sınıflandırıcı (GA-NDA), farklı sayı ve yapıda özniteliğe sahip VEHICLE, PROTEIN ve POWER veritabanları üzerinde sınıflandırma hassasiyeti açısından klasik altuzay sınıflandırıcılar ile karşılaştırılmıştır.

GA-NDA yönteminde tanımlanmış olan genetik algoritma parametreleri Çizelge 5.21'de görülmektedir. Bölüm 4'te belirtildiği üzere GA uyum fonksiyonu değeri, seçilen özyönler üzerine yapılan izdüşüm ile elde edilen sınıflandırma hassasiyeti olarak belirlenmiştir.

Çizelge 5.21. GA-NDA sınıflandırıcının GA parametreleri

Kromozom Boyutu	Öznitelik uzayı boyutu
Nüfus Sayısı	20
Kuşak Sayısı	20
Çaprazlama Olasılığı	%90
Mutasyon Olasılığı	%5

Her bir veritabanı için GA-NDA yöntemiyle elde edilen eniyi özyönlere ait indisler (artan yöndeki özdeğer sıralamasına göre) Çizelge 5.22 – 5.24'te verilmiştir. Görüldüğü üzere seçilen özyönler hem küçük hem de büyük özdeğerlere karşılık gelebilmekte ve toplam sayı, (sınıf sayısı - 1) limitinin üzerinde olabilmektedir. Bu durum, dağılımlardaki ayırdediciliğin, değişimin hem az hem de çok olduğu yönlerde olabileceğinin açık bir göstergesidir.

Çizelge 5.22. VEHICLE veritabanı: Seçilen özyönlerin indisleri

10, 14, 15, 17, 18

Çizelge 5.23. PROTEIN veritabanı: Seçilen özyönlerin indisleri

1, 2, 3, 4, 5, 6, 7, 8

Çizelge 5.24. POWER veritabanı: Seçilen özyönlerin indisleri

5, 6, 9, 12, 14, 16, 17, 18, 19

GA-NDA ve klasik altuzay sınıflandırıcılar kullanılarak her bir veritabanı için elde edilmiş olan ortalama tanıma oranları Çizelge 5.25'te görülmektedir. Sonuçlardan anlaşılacağı üzere veritabanlarının tamamında, GA-NDA sınıflandırıcı diğerlerine göre karşılaştırılabilir ya da daha yüksek bir başarım sunmaktadır.

Çizelge 5.25. Altuzay sınıflandırıcıların ortalama tanıma oranları (%)

	PCA	**CLAFIC**	**CVA**	**FLDA**	**NDA**	**GA-NDA**
VEHICLE	42,13	74,63	75,63	78,50	79,25	80,00
PROTEIN	70,79	69,22	70,79	71,88	71,41	71,41
POWER	41,67	69,17	70,83	58,33	72,50	75,83

BÖLÜM 6

SONUÇLAR

Bu tez çalışmasının temel amacı örüntü tanımanın temel öğeleri olan öznitelik çıkarma, öznitelik seçme ve sınıflandırma konularında literatürde mevcut bulunan yaklaşımlarla karşılaştırılabilir ya da daha yüksek başarım sağlayan yöntemler geliştirmektir.

Öznitelik çıkarımı konusunda ses ve konuşma sinyalleri referans alınarak dalgacık dönüşümü temelli bir yöntem önerilmiştir. Farklı karakteristiğe sahip sinyaller üzerinde yapılan deneysel çalışmalarda, bu yöntemin özellikle uzun süreli, durağan olmayan ve ani frekans değişimleri içeren sinyalleri temsil etmede Fourier dönüşümü temelli klasik öznitelik çıkarma yöntemine göre daha başarılı olduğu gözlenmiştir. Bunun yanısıra dalgacık temelli öznitelikler, sinyallerin frekans bileşenlerine dair bilgileri de taşımaktadır. Bu yapıdaki özniteliklerin sonraki aşamada öznitelik seçme işlemine tabi tutulmasıyla sinyallerin ayırdedici altbantları da bulunabilmektedir. MFCC gibi Fourier temelli özniteliklerde ise bu bilginin direkt olarak elde edilmesi mümkün değildir.

Çalışmanın ikinci aşaması olan öznitelik seçimi, mevcut öznitelik kümesi içindeki elemanların çeşitli ölçütler yardımıyla ayırdedicilik güçlerinin belirlenerek elemeye tabi tutulması ve başlangıç kümesinden daha düşük boyutlu ve daha ayırdedici bir alt küme elde edilmesi şeklinde açıklanabilir. Tez çalışmasının bu bölümünde, özniteliklerin bireysel ayırdedicilik derecelerini belirlemek için altuzay temelli iki yeni ayrılabilirlik ölçüsü (CS ve FS) geliştirilmiş ve bu ölçülerin, çok sınıflı örüntü uygulamalarında öznitelik seçimi amacıyla kullanılabilmesi için bir yöntem önerilmiştir. Yeni ölçülerle yapılan öznitelik seçim başarımı, farklı sayı ve yapıda özniteliği barındıran çeşitli veritabanlarında, sınıflandırma hassasiyeti ve boyut indirgeme oranı açısından uzaksaklık ve Bhattacharyya gibi klasik ayrılabilirlik ölçülerinin tek-değişkenli versiyonları ile karşılaştırılmıştır. Altuzay temelli öznitelik seçimi iki durumda da başarılı sonuçlar vermiştir. Bu karşılaştırmayı takiben, kullanılan veritabanları üzerinde

çok-değişkenli öznitelik seçme yöntemlerinin başarımı da analiz edilmiştir. Daha sonra, gerek tek-değişkenli gerekse çok-değişkenli yaklaşımların sağladığı sonuçlar, eniyi ES yöntemiyle elde edilen eniyi sonuç ile benzerlik oranı ve işlem süresi gibi açılardan karşılaştırılmıştır. Bunun sonucunda, öznitelikler arası ilintilerin düşük olduğu durumlarda altuzay temelli tek-değişkenli öznitelik seçme yöntemlerinin, çok-değişkenli yöntemlerle karşılaştırılabilir başarım sağladığı, işlem süresi açısından çok daha avantajlı olduğu ve ayrıca özniteliklerin ayırdedicilik derecelerinin belirlenmesinde sınıflandırıcıdan bağımsız sonuçlar verdiği görülmüştür.

Tez çalışmasının son bölümü olan sınıflandırma konusunda, doğrusal altuzay sınıflandırma yöntemleri esas alınmıştır. Literatürde, yalnızca sınıf-içi (PCA, CLAFIC, CVA) ya da hem sınıf-içi hem de sınıflar-arası ilişkileri değerlendiren (FLDA) altuzay sınıflandırıcılar bulunmaktadır. Bunlardan bazıları altuzay tabanı olarak ilinti ya da ortak değişintilerin küçük özdeğerlerine karşılık gelen (değişimin az olduğu) özvektörleri yada özyönleri kullanırken, bazıları ise büyük özdeğerlere karşılık gelen yönleri (değişimin fazla olduğu) kullanmaktadır. Buna ilaveten, altuzayı oluşturan özyönlerin sayıları bazı sınıflandırıcılarda limitli iken (FLDA) bazılarında ise en uygun sayıyı belirlemek bir sorun teşkil etmektedir. Altuzay sınıflandırıcılardaki bu farklılıkları ve sorunları değerlendirerek, tez çalışmasının sınıflandırma bölümünde, genetik algoritma temelli yeni bir altuzay sınıflandırıcı (GA-NDA) geliştirilmiştir. Yeni sınıflandırıcı gerek sınıf-içi gerekse sınıflar-arası ilişkileri daha uygun bir şekilde değerlendirmekte; ayrıca, klasik altuzay sınıflandırıcıların aksine altuzay izdüşümü için hem büyük hem de küçük özdeğerlere karşılık gelen özyönlerin birlikte kullanılmasını mümkün kılmaktadır. Çeşitli veritabanları üzerinde yapılan deneysel çalışmalarda, genetik altuzay sınıflandırıcı klasik altuzay yöntemlerine göre bu özellikleriyle öne çıkarak genelde daha yüksek bir başarım sağlamıştır. Bu sonuçlar, sınıflandırma başarımında sınıf-içi ve sınıflar-arası ilişkilerin uygun şekilde analiz edilmesinin gerekliliğini ve de sınıf dağılımlarında, değişimin hem az hem de çok olduğu özyönlerin ayırdedicilik açısından önemli olabileceğini ortaya koymuştur.

KAYNAKLAR DİZİNİ

Akay, M., 2006, Wiley Encyclopedia of Biomedical Engineering, Wiley-Interscience, 4152 p.

Asuncion, A. and Newman, D.J., 2007, UCI Machine Learning Repository. Irvine, CA: University of California, School of Information and Computer Science.

Bellman, R., 1961, Adaptive control processes: a guided tour, Princeton University Press, 255 p.

Çevikalp, H., Neamtu, M., Wilkes, M. and Barkana, A., 2005, Discriminative common vectors for face recognition, , IEEE Trans. on Pattern Analysis and Machine Intelligence, 27(1), 4–13.

Davis, S. B. and Mermelstein, P., 1980, Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences, IEEE Trans. on Acoustic, Speech and Signal Processing, 28(4), 357–366.

Duda, R. O., Hart, P. E. and Stork, D. G., 2001, Pattern classification, John Wiley & Sons Inc., USA, 654 p.

Edizkan, R., Gülmezoğlu, M. B., Ergin, S and Barkana, A., 2005, Improvements on common vector approach for multiclass problems, European Signal Processing Conference, Antalya, Türkiye.

Farooq, O. and Datta, S., 2001, Mel filter-like admissible wavelet packet structure for speech recognition, IEEE Signal Processing Letters, 8(7), 196–198.

Farooq, O. and Datta, S., 2003, Phoneme recognition using wavelet based features, Information Sciences, 150, 5–15.

KAYNAKLAR DİZİNİ (devam)

Fisher, R. A, 1936, The use of multiple measurements in taxonomic problems, Annals of Eugenics, 7, 179–188.

Fukunaga, K. and Mantock, J., 1983, Nonparametric discriminant analysis, IEEE Trans. on Pattern Analysis and Machine Intelligence, 5, 671–678.

Fukunaga, K., 1990, Introduction to statistical pattern recognition, Academic Press, 592 p.

Gerek, Ö. N., Ece, D. G. and Barkana, A., 2006, Covariance analysis of voltage waveform signature for power-quality event classification, IEEE Trans. on Power Delivery, 21(4), 2022–2031.

Goldberg, D. E., 1989, Genetic algorithms in search, optimization, and machine learning, Addison-Wesley, Reading, MA, 432 p.

Gonzalez, R. C. and Woods, R. E., 2007, Digital image processing, Prentice Hall, 976 p.

Guyon, I. and Elisseeff, A., 2003, An introduction to variable and feature selection, Journal of Machine Learning Research, 3, 1157–1182.

Gülmezoğlu, M. B., Keskin, M., Dzhafarov, V. and Barkana, A., 1999, A novel approach to isolated word recognition, IEEE Trans. on Speech and Audio Processing, 7(6), 620–628.

Gülmezoğlu, M. B., Dzhafarov, V. and Barkana, A., 2001, The common vector approach and its relation to principal component analysis, IEEE Trans. on Speech and Audio Processing, 9(6), 655–662.

KAYNAKLAR DİZİNİ (devam)

Gülmezoğlu, M. B., Dzhafarov, V., Edizkan, R. and Barkana, A., 2007, The common vector approach and its comparison with other subspace methods in case of sufficient data, Computer Speech and Language, 21(2), 266–281.

Günal, S., 2003, Ortak vektör yaklaşımı yöntemiyle TI TMS320C6711 DSK platformunda konuşmacıdan bağımsız gerçek zamanlı rakam tanıma, Yüksek lisans tezi, Osmangazi Üniversitesi, 102 s.

Günal, S., Ergin, S., Gülmezoğlu, M. B. and Gerek, Ö. N., 2006, On feature extraction for spam e-mail detection, Lecture Notes in Computer Science, 4105, 635–642.

Günal, S. and Edizkan, R., 2006, Wavelet based discriminative feature extraction for speech recognition, International Conference on Modeling and Simulation, Konya, Türkiye, 621–624.

Günal, S. and Edizkan, R., 2007, Use of novel feature extraction technique with subspace classifiers for speech recognition, IEEE International Conference on Pervasive Services, İstanbul, Türkiye, 80–83.

Günal, S., Edizkan, R., Gerek, Ö.N. ve Ece, D.G., 2008, Güç kalitesi olaylarının sınıflandırılması için öznitelik seçimi, IEEE 16. Sinyal İşleme ve İletişim Uygulamaları Kurultayı, Didim, Türkiye.

Günal, S. and Edizkan, R., 2008, Subspace based feature selection for pattern recognition, Information Sciences (in press).

Hogben, L., 2006, Handbook of Linear Algebra, Chapman & Hall/CRC, 1400 p.

Hotelling, H., 1933, Analysis of a complex of statistical variables into principal components, Journal of Educational Psychology, 24, 498–520.

KAYNAKLAR DİZİNİ (devam)

Ishibuchi, H. and Nakashima, T., 2000, Multi-objective pattern and feature selection by a genetic algorithm, Genetic and Evolutionary Computation Conference, Las Vegas, USA, 1069–1076.

Jain, A.K. and Chandrasekaran, B., 1982, Dimensionality and sample size considerations in pattern recognition practice, Handbook of Statistics, vol.2, Amsterdam: North Holland.

Jain, A.K. and Zongker, D., 1997, Feature selection: evaluation, application, and small sample performance, IEEE Trans. on Pattern Analysis and Machine Intelligence, 19(2), 153–158.

Kailath, T., 1967, The divergence and Bhattacharyya distance measures in signal selection, IEEE Trans. on Communication Technology, 15(1), 52–60.

Kavzoglu, T. and Mather, P.M., 2000, The use of feature selection techniques in the context of artificial neural networks, 26th Annual Conference of Remote Sensing Society, Leicester, UK.

Khan, J., Wei, J. S., Ringner, M., Saal, L. H., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C. R., Peterson, C. and Meltzer, P. S., 2001, Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks, Nature Medicine, 7(6), 673–679.

Kittler, J., 1978, Feature set search algorithms. In: C.H. Chert, Ed., Pattern Recognition and Signal Processing. Sijthoff and Noordhoff, Mphen aan den Rijn, Netherlands, 41-60.

Kohavi, G. and John, R., 1997, Wrappers for feature subset selection, Artificial Intelligence, 97, 273–324.

KAYNAKLAR DİZİNİ (devam)

Kudo, M. and Sklansky, J., Comparison of algorithms that select features for pattern classifiers, 2000, Pattern Recognition, 33, 25–41.

Kuncheva, L. I., 2004, Combining pattern classifiers, John Wiley & Sons Inc., New Jersey, 350 p.

Lai, C., Reinders, M. J. T. and Wessels, L., 2006, Random subspace method for multivariate feature selection, Pattern Recognition Letters, 27, 1067–1076.

Leonard, R. G., 1984, A database for speaker independent digit recognition, IEEE International Conference on Acoustics, Speech, and Signal Processing, 9, 328–331.

Loéve, M., 1963, Probability Theory, Van Nostrand, New York.

Mallat, S., 1999, A wavelet tour of signal processing, Academic Press, 637 p.

Marill, T. and Green, D. M., 1963, On the effectiveness of receptors in recognition system, IEEE Trans. Inform. Theory, 9, 11–17.

Misiti, M., Misiti, Y., Oppenheim, G. and Poggi, J., 2005, Wavelet toolbox user's guide, The MathWorks, Inc.

Narendra, P. M. and Fukunaga, K., 1977, A branch and bound algorithm for feature subset selection, IEEE Trans. on Computers, 26(9), 917–922.

Oja, E., 1983, Subspace methods of pattern recognition, John Wiley & Sons Inc., USA, 200 p.

Pearson, K., 1901, On lines and planes of closest fit to systems of points in space, Philosophical Magazine, 2(6), 559–572.

KAYNAKLAR DİZİNİ (devam)

Proakis, J. G. and Manolakis, D. K., 2006, Digital signal processing, Prentice Hall, 1004 p.

Pudil, P., Novovicova, J. and Kittler, J., 1994, Floating search methods in feature selection, Pattern Recognition Letters, 15, 1119–1125.

Rabiner, L. and Juang, B. H., 1993, Fundamentals of speech recognition, Prentice Hall, 496 p.

Reyes-Aldasoro, C. C. and Bhalerao, A., 2006, The Bhattacharyya space for feature selection and its application to texture segmentation, Pattern Recognition, 39(5), 812–826.

Reynolds, T. J. and Antoniou, C. A., 2003, Experiments in speech recognition using a modular MLP architecture for acoustic modelling, Information Sciences, 156, 39–54.

Ricotti, L. P., 2005, Multitapering and a wavelet Variant of MFCC in speech recognition, IEE Proceedings - Vision, Image, and Signal Processing, 152(1), 29–35.

Rokach, L., 2008, Genetic algorithm-based feature set partitioning for classification problems, Pattern Recognition, 41, 1693–1717.

Schölkopf, B. and Smola, A. J., 2001, Learning with kernels: support vector machines, regularization, optimization, and beyond, MIT Press, 644 p.

Siedlecki, W. and Sklansky, J., 1989, A note on genetic algorithms for large-scale feature selection, Pattern Recognition Letters, 10, 335–347.

KAYNAKLAR DİZİNİ (devam)

Stearns, S. D., 1976, On selecting features for pattern classifiers, 3rd lnternational Conference on Pattern Recognition, Coronado, CA, 71-75.

Stone, M., 1974, Cross-validatory choice and assessment of statistical predictions, Journal of the Royal Statistical Society. Series B (Methodological), 36(2), 111–147.

Su, K. and Lee, C., 1994, Speech recognition using weighted HMM and subspace projection approaches, IEEE Trans. on Speech and Audio Processing, 2(1), 69–79.

Theodoridis, S. and Koutroumbas, K., 2003, Pattern recognition, Academic Press, USA, 689 p.

Tian, Q., Fainman, Y. and Lee, S. H., 1988, Comparison of statistical pattern-recognition algorithms for hybrid processing. II. Eigenvector-based algorithm, Journal of the Optical Society of America A, 5(10), 1670–1682.

Tibshirani, R., Hastie, T., Narasimhan, B. and Chu, G., 2002, Diagnosis of multiple cancer types by shrunken centroids of gene expression, Proc. National Academy of Sciences, 99(10), 6567–6572.

Türk Dil Kurumu, 2008, Web Sitesi: http://www.tdk.org.tr.

Uncu, O. and Turksen, I. B., 2007, A novel feature selection approach: combining feature wrappers and filters, Information Sciences, 177, 449–466.

Watanabe, S., Lambert, P. F., Kulikowski, C. A., Buxton, J. L. and Walker, R., 1967, Evaluation and selection of variables in pattern recognition, In J. Tou (Ed.), Computer and Information Sciences II. New York: Academic Press.

KAYNAKLAR DİZİNİ (devam)

Webb, A., 2002, Statistical pattern recognition, John Wiley & Sons Ltd., England, 496 p.

Whitney, A.W., 1971, A direct method of nonparametric measurement selection, IEEE Trans. Comput, 20, 1100–1103.

Yang, J. and Honavar, V., 1998, Feature subset selection using a genetic algorithm, IEEE Intelligent Systems, 13(2), 44–49.

Zue, V., Seneff, S. and Glass, J., 1990, Speech database development at MIT: TIMIT and beyond, Speech Communication, 9(4), 351–356.

ÖZGEÇMİŞ

Serkan Günal, 1978 yılında Eskişehir'de doğdu. İlk, orta ve lise öğrenimini aynı ilde tamamladı. Daha sonra, sırasıyla 1999 ve 2003 yıllarında Eskişehir Osmangazi Üniversitesi, Elektrik - Elektronik Mühendisliği Anabilim dalında lisans ve yüksek lisans derecelerini aldı. Aynı üniversitede, 1999 - 2001 yılları arasında Araştırma Görevlisi olarak çalışan Günal, sonraki birkaç yılda yerli ve yabancı çeşitli teknoloji şirketlerinde AR-GE Mühendisi pozisyonunda bulundu. Halen, Anadolu Üniversitesi Bilgisayar Mühendisliği Bölümü'nde görev yapmaktadır. Başlıca akademik çalışma alanları, örüntü tanıma ve sayısal sinyal işlemedir.