

İSTANBUL TEKNİK ÜNİVERSİTESİ ★ FEN BİLİMLERİ ENSTİTÜSÜ

KARARLI VE BAŞARIMI YÜKSEK ÖZNİTELİK SEÇİMİ

**YÜKSEKLİSANS TEZİ
Gökhan GÜLGEZEN**

Anabilim Dalı : Bilgisayar Mühendisliği

Programı : Bilgisayar Mühendisliği

HAZİRAN 2009

İSTANBUL TEKNİK ÜNİVERSİTESİ ★ FEN BİLİMLERİ ENSTİTÜSÜ

KARARLI VE BAŞARIMI YÜKSEK ÖZNİTELİK SEÇİMİ

**YÜKSEK LİSANS TEZİ
Gökhan GÜLGEZEN
(504071516)**

Tezin Enstitüye Verildiği Tarih : 04 Mayıs 2009

Tezin Savunulduğu Tarih : 03 Haziran 2009

**Tez Danışmanı : Doç. Dr. Zehra ÇATALTEPE(İTÜ)
Diğer Jüri Üyeleri : Yrd. Doç. Dr. Şule GÜNDÜZ
ÖĞÜDÜCÜ(İTÜ)
Doç. Dr. Tunga GÜNGÖR(BÜ)**

HAZİRAN 2009

ÖNSÖZ

Danışmanım Zehra Çataltepe'ye tez çalışmam boyunca gösterdiği anlayış ve samimiyetten ötürü teşekkür ederim. Bana her konuda desteğini hissettirdi ve engin bilgisiyle araştırmalarımızı yönlendirerek bu çalışmayı mümkün kıldı.

Sağladığı burs imkanları ile bana ve benim gibi bir çok öğrenciye destek veren, akademik yayınlarıyla ve müsbet bilimlerde araştırma geliştirme faaliyetleriyle öncü bir kurum olan TÜBİTAK'a ayrıca teşekkür etmek istiyorum.

Tez çalışmam boyunca her zaman yanımda olan ve bir çok konuda yardımlarına başvurduğum arkadaşlarım Barış Şenliol, Yusuf Yaslan, Eser Aygün ve Sanjarbek Hudaiberdiev'e çok müteşekkirim. Ayrıca çalışmalarını bizimle paylaşan, bizden yardımlarını esirgemeyen Prof. Dr. Lei Yu'ya da teşekkürü bir borç biliyorum.

Son olarak sevgileri ve özverileriyle bana her daim yol gösteren, kendimi güvende hissetmemi sağlayan aileme, Dilber Gülgezen, Fedai Gülgezen ve Neslihan Gülgezen'e sonsuz kez teşekkür ederim.

Mayıs 2009

Gökhan Gülgezen

İÇİNDEKİLER

Sayfa

ÖNSÖZ.....	iii
İÇİNDEKİLER	v
ÇİZELGE LİSTESİ.....	vii
ŞEKİL LİSTESİ.....	ix
ÖZET	xi
SUMMARY	xiii
1. GİRİŞ	1
2. VERİ KÜMELERİ.....	5
2.1 Parkinson Veri Kümesi	6
2.2 Sonar Veri Kümesi	6
2.3 İyonosfer Veri Kümesi	7
2.4 Musk (Versiyon 1) Veri Kümesi	8
2.5 Şarap Veri Kümesi	8
2.6 Ses Türü Veri Kümesi	9
2.7 Çoklu Öznitelik Veri Kümesi	10
2.8 El Yazısı Rakamlar Veri Kümesi	10
3. ÖZNİTELİK SEÇME ALGORİTMALARI.....	13
3.1 MRMR Algoritması	13
3.2 FCBF Algoritması	15
3.3 MRMR ve FCBF Algoritmalarının Karşılaştırılması.....	18
4. MID VE MIQ YÖNTEMLERİNİN KARARLILIK YÖNÜNDEN KARŞILAŞTIRILMASI.....	21
4.1 Teorik Analiz.....	21
4.2 Deneysel Analiz	23
5. MRMR ALGORİTMASI İÇİN YENİ BİR SEÇİM KRİTERİ, MID_{α}	25
6. KARARLILIK ÖLÇÜTÜ	27
7. KARARLILIK VE BAŞARIM KARŞILAŞTIRILMASI İÇİN DENEYSEL METODOLOJİ.....	31
8. DENEY SONUÇLARI.....	33
8.1 MID ve MIQ Seçim Kriterlerinin Karşılaştırılması	33
8.2 MID_{α} Tekniğinin Kararlılık ve Başarım Sonuçları	34
9. TARTIŞMA VE SONUÇ	49
KAYNAKLAR.....	51
EKLER.....	53

ÇİZELGE LİSTESİ

Sayfa

Çizelge 2.1 : Deneylerde kullanılan veri kümeleri hakkında özet bilgiler.	5
Çizelge 3.1 : MRMR algoritmasının öznitelik seçim kriterleri.	15
Çizelge 3.2 : FCBF ve MRMR algoritmalarının farklı veri kümelerinde başarımlarından karşılaştırılması.	20
Çizelge 8.1 : MRMR, MID_{α} seçim kriterinin farklı α ve r değerleri için Parkinson veri kümesi üzerindeki kararlılık sonuçlarının ortalama sıralanma değerleri ve standart sapmaları.	36
Çizelge 8.2 : MRMR, MID_{α} seçim kriterinin farklı α ve r değerleri için Sonar veri kümesi üzerindeki kararlılık sonuçlarının ortalama sıralanma değerleri ve standart sapmaları.	37
Çizelge 8.3 : MRMR, MID_{α} seçim kriterinin farklı α ve r değerleri için Sonar veri kümesi üzerindeki başarımlar (k-en-yakın-komşu) sonuçlarının ortalama sıralanma değerleri ve standart sapmaları.	38
Çizelge 8.4 : MRMR, MID_{α} seçim kriterinin farklı α ve r değerleri için İyonosfer veri kümesi üzerindeki kararlılık sonuçlarının ortalama sıralanma değerleri ve standart sapmaları.	39
Çizelge 8.5 : MRMR, MID_{α} seçim kriterinin farklı α ve r değerleri için Musk veri kümesi üzerindeki kararlılık sonuçlarının ortalama sıralanma değerleri ve standart sapmaları.	40
Çizelge 8.6 : MRMR, MID_{α} seçim kriterinin farklı α ve r değerleri için Şarap veri kümesi üzerindeki kararlılık sonuçlarının ortalama sıralanma değerleri ve standart sapmaları.	42
Çizelge 8.7 : MRMR, MID_{α} seçim kriterinin farklı α ve r değerleri için Şarap veri kümesi üzerindeki başarımlar (svm) sonuçlarının ortalama sıralanma değerleri ve standart sapmaları.	42
Çizelge 8.8 : MRMR, MID_{α} seçim kriterinin farklı α ve r değerleri için Ses Türü veri kümesi üzerindeki kararlılık sonuçlarının ortalama sıralanma değerleri ve standart sapmaları.	43
Çizelge 8.9 : MRMR, MID_{α} seçim kriterinin farklı α ve r değerleri için Çoklu Öznitelik veri kümesi üzerindeki kararlılık sonuçlarının ortalama sıralanma değerleri ve standart sapmaları.	44
Çizelge 8.10 : MRMR, MID_{α} seçim kriterinin farklı α ve r değerleri için Çoklu Öznitelik veri kümesi üzerindeki başarımlar (k-en-yakın-komşu) sonuçlarının ortalama sıralanma değerleri ve standart sapmaları.	45
Çizelge 8.11 : MRMR, MID_{α} seçim kriterinin farklı α ve r değerleri için Çoklu Öznitelik veri kümesi üzerindeki kararlılık sonuçlarının ortalama sıralanma değerleri ve standart sapmaları.	46

Çizelge 8.12 : MRMR, MID_{α} seçim kriterinin farklı α ve r değerleri için El Yazısı Rakamlar veri kümesi üzerindeki başarımlar (k-en-yakın-komşu) sonuçlarının ortalama sıralanma değerleri ve standart sapmaları.....	47
Çizelge A.1 : FCBF, MRMR-MID ve MRMR-MIQ öznitelik seçme yöntemlerinin Parkinson veri kümesi üzerindeki sonuçları.....	54
Çizelge A.2 : FCBF, MRMR-MID ve MRMR-MIQ öznitelik seçme yöntemlerinin Sonar veri kümesi üzerindeki sonuçları.....	54
Çizelge A.3 : FCBF, MRMR-MID ve MRMR-MIQ öznitelik seçme yöntemlerinin İyonosfer veri kümesi üzerindeki sonuçları.....	54
Çizelge A.4 : FCBF, MRMR-MID ve MRMR-MIQ öznitelik seçme yöntemlerinin Musk veri kümesi üzerindeki sonuçları.....	55
Çizelge A.5 : FCBF, MRMR-MID ve MRMR-MIQ öznitelik seçme yöntemlerinin Şarap veri kümesi üzerindeki sonuçları.....	55
Çizelge A.6 : FCBF, MRMR-MID ve MRMR-MIQ öznitelik seçme yöntemlerinin Ses Türü veri kümesi üzerindeki sonuçları.....	55
Çizelge A.7 : FCBF, MRMR-MID ve MRMR-MIQ öznitelik seçme yöntemlerinin Çoklu Öznitelik veri kümesi üzerindeki sonuçları.....	56
Çizelge A.8 : FCBF, MRMR-MID ve MRMR-MIQ öznitelik seçme yöntemlerinin El Yazısı Rakamlar veri kümesi üzerindeki sonuçları.....	58
Çizelge B.1 : MRMR, MID_{α} seçim kriterinin farklı α ve r değerleri için Parkinson veri kümesi üzerindeki başarımlar (k-en-yakın-komşu) sonuçlarının ortalama sıralanma değerleri ve standart sapmaları.....	59
Çizelge B.2 : MRMR, MID_{α} seçim kriterinin farklı α ve r değerleri için Parkinson veri kümesi üzerindeki başarımlar (svm) sonuçlarının ortalama sıralanma değerleri ve standart sapmaları.....	60
Çizelge B.3 : MRMR, MID_{α} seçim kriterinin farklı α ve r değerleri için Sonar veri kümesi üzerindeki başarımlar (svm) sonuçlarının ortalama sıralanma değerleri ve standart sapmaları.....	61
Çizelge B.4 : MRMR, MID_{α} seçim kriterinin farklı α ve r değerleri için İyonosfer veri kümesi üzerindeki başarımlar (k-en-yakın-komşu) sonuçlarının ortalama sıralanma değerleri ve standart sapmaları.....	62
Çizelge B.5 : MRMR, MID_{α} seçim kriterinin farklı α ve r değerleri için İyonosfer veri kümesi üzerindeki başarımlar (svm) sonuçlarının ortalama sıralanma değerleri ve standart sapmaları.....	63
Çizelge B.6 : MRMR, MID_{α} seçim kriterinin farklı α ve r değerleri için Musk veri kümesi üzerindeki başarımlar (k-en-yakın-komşu) sonuçlarının ortalama sıralanma değerleri ve standart sapmaları.....	64
Çizelge B.7 : MRMR, MID_{α} seçim kriterinin farklı α ve r değerleri için Şarap veri kümesi üzerindeki başarımlar (k-en-yakın-komşu) sonuçlarının ortalama sıralanma değerleri ve standart sapmaları.....	65
Çizelge B.8 : MRMR, MID_{α} seçim kriterinin farklı α ve r değerleri için Ses Türü veri kümesi üzerindeki başarımlar (k-en-yakın-komşu) sonuçlarının ortalama sıralanma değerleri ve standart sapmaları.....	66
Çizelge B.9 : MRMR, MID_{α} seçim kriterinin farklı α ve r değerleri için El Yazısı Rakamlar veri kümesi üzerindeki başarımlar (svm) sonuçlarının ortalama sıralanma değerleri ve standart sapmaları.....	67

ŞEKİL LİSTESİ

Sayfa

Şekil 2.1 : Parkinson veri kümesindeki özniteliklerin sınıf etiketleri vektörü ile arasındaki SU (solda) ve kendi dışındaki diğer öznitelikler ile olan ortalama SU (sağda) değerleri.....	6
Şekil 2.2 : Sonar veri kümesindeki özniteliklerin sınıf etiketleri vektörü ile arasındaki SU (solda) ve kendi dışındaki diğer öznitelikler ile olan ortalama SU (sağda) değerleri.....	7
Şekil 2.3 : İyonosfer veri kümesindeki özniteliklerin sınıf etiketleri vektörü ile arasındaki SU (solda) ve kendi dışındaki diğer öznitelikler ile olan ortalama SU (sağda) değerleri.....	7
Şekil 2.4 : Musk veri kümesindeki özniteliklerin sınıf etiketleri vektörü ile arasındaki SU (solda) ve kendi dışındaki diğer öznitelikler ile olan ortalama SU (sağda) değerleri.....	8
Şekil 2.5 : Şarap veri kümesindeki özniteliklerin sınıf etiketleri vektörü ile arasındaki SU (solda) ve kendi dışındaki diğer öznitelikler ile olan ortalama SU (sağda) değerleri.....	9
Şekil 2.6 : Ses Türü veri kümesindeki özniteliklerin sınıf etiketleri vektörü ile arasındaki SU (solda) ve kendi dışındaki diğer öznitelikler ile olan ortalama SU (sağda) değerleri.....	9
Şekil 2.7 : Çoklu Öznitelik veri kümesindeki özniteliklerin sınıf etiketleri vektörü ile arasındaki SU (solda) ve kendi dışındaki diğer öznitelikler ile olan ortalama SU (sağda) değerleri.....	10
Şekil 2.8 : El Yazısı Rakam veri kümesindeki özniteliklerin sınıf etiketleri vektörü ile arasındaki SU (solda) ve kendi dışındaki diğer öznitelikler ile olan ortalama SU (sağda) değerleri.....	11
Şekil 4.1 : Cauchy dağılımının olasılık yoğunluk fonksiyonu (solda) ve kümülatif dağılım fonksiyonu (sağda).....	23
Şekil 4.2 : Musk veri kümesi üzerinde her bir öznitelik seçimi aşamasında V , W ve $V - W$ (MID) değerlerinin ortalama ve standart sapma değerleri.....	24
Şekil 4.3 : Musk veri kümesi üzerinde her bir öznitelik seçimi aşamasında V , W ve V/W (MIQ) değerlerinin ortalama ve standart sapma değerleri.....	24
Şekil 6.1 : İki öznitelik kümesi R_1 ve R_2 'nin tam iki parçalı çizge olarak modeli....	28
Şekil 7.1 : Öznitelik seçme algoritmalarının kararlılık ve başarımlarından karşılaştırılmasında kullanılacak deneysel metodoloji modeli.....	32
Şekil 8.1 : MRMR, MID ve MIQ öznitelik seçim kriterlerinin Musk veri kümesinde kararlılık yönünden karşılaştırılması.....	33
Şekil 8.2 : MRMR, MID ve MIQ öznitelik seçim kriterlerinin Musk veri kümesi üzerinde başarımlarından karşılaştırılması.....	34
Şekil 8.3 : MRMR, MID_α seçim kriterinin farklı α ve r değerleri için Parkinson veri kümesi üzerinde kararlılık sonuçları.....	36

Şekil 8.4 : MRMR, MID_{α} seçim kriterinin farklı α ve r değerleri için Sonar veri kümesi üzerinde kararlılık sonuçları.	37
Şekil 8.5 : MRMR, MID_{α} seçim kriterinin farklı α ve r değerleri için Sonar veri kümesi üzerinde başarımlık (k-en-yakın-komşu) sonuçları.	38
Şekil 8.6 : MRMR, MID_{α} seçim kriterinin farklı α ve r değerleri için İyonosfer veri kümesi üzerinde kararlılık sonuçları.	39
Şekil 8.7 : MRMR, MID_{α} seçim kriterinin farklı α ve r değerleri için Musk veri kümesi üzerinde kararlılık sonuçları.	40
Şekil 8.8 : MRMR, MID_{α} seçim kriterinin farklı α ve r değerleri için Şarap veri kümesi üzerinde kararlılık sonuçları.	41
Şekil 8.9 : MRMR, MID_{α} seçim kriterinin farklı α ve r değerleri için Şarap veri kümesi üzerinde başarımlık (svm) sonuçları.	42
Şekil 8.10 : MRMR, MID_{α} seçim kriterinin farklı α ve r değerleri için Ses Türü veri kümesi üzerinde kararlılık sonuçları.	43
Şekil 8.11 : MRMR, MID_{α} seçim kriterinin farklı α ve r değerleri için Çoklu Öznitelik veri kümesi üzerinde kararlılık sonuçları.	44
Şekil 8.12 : MRMR, MID_{α} seçim kriterinin farklı α ve r değerleri için Çoklu Öznitelik veri kümesi üzerinde başarımlık (k-en-yakın-komşu) sonuçları.	45
Şekil 8.13 : MRMR, MID_{α} seçim kriterinin farklı α ve r değerleri için Çoklu Öznitelik veri kümesi üzerinde kararlılık sonuçları.	46
Şekil 8.14 : MRMR, MID_{α} seçim kriterinin farklı α ve r değerleri için El Yazısı Rakamlar veri kümesi üzerinde başarımlık (k-en-yakın-komşu) sonuçları.	47
Şekil B.1 : MRMR, MID_{α} seçim kriterinin farklı α ve r değerleri için Parkinson veri kümesi üzerinde başarımlık (k-en-yakın-komşu) sonuçları.	59
Şekil B.2 : MRMR, MID_{α} seçim kriterinin farklı α ve r değerleri için Parkinson veri kümesi üzerinde başarımlık (svm) sonuçları.	60
Şekil B.3 : MRMR, MID_{α} seçim kriterinin farklı α ve r değerleri için Sonar veri kümesi üzerinde başarımlık (svm) sonuçları.	61
Şekil B.4 : MRMR, MID_{α} seçim kriterinin farklı α ve r değerleri için İyonosfer veri kümesi üzerinde başarımlık (k-en-yakın-komşu) sonuçları.	62
Şekil B.5 : MRMR, MID_{α} seçim kriterinin farklı α ve r değerleri için İyonosfer veri kümesi üzerinde başarımlık (svm) sonuçları.	63
Şekil B.6 : MRMR, MID_{α} seçim kriterinin farklı α ve r değerleri için Musk veri kümesi üzerinde başarımlık (k-en-yakın-komşu) sonuçları.	64
Şekil B.7 : MRMR, MID_{α} seçim kriterinin farklı α ve r değerleri için Şarap veri kümesi üzerinde başarımlık (k-en-yakın-komşu) sonuçları.	65
Şekil B.8 : MRMR, MID_{α} seçim kriterinin farklı α ve r değerleri için Ses Türü veri kümesi üzerinde başarımlık (k-en-yakın-komşu) sonuçları.	66
Şekil B.9 : MRMR, MID_{α} seçim kriterinin farklı α ve r değerleri için El Yazısı Rakamlar veri kümesi üzerinde başarımlık (svm) sonuçları.	67

KARARLI VE BAŞARIMI YÜKSEK ÖZNETELİK SEÇİMİ

ÖZET

Öznitelik seçme yöntemleri girdi boyutunu azaltmaları ve ilgisiz, artık veriyi yok etmeleri gibi özellikleriyle makine öğrenimi algoritmalarının daha hızlı çalışmalarına ve daha başarılı sonuçlar elde etmelerine olanak sağlarlar. Özellikle son on yılda öznitelik seçme yöntemleri örüntü tanıma, sinyal işleme, metin sınıflandırma ve biyoenformatik gibi alanlarda tanınır olmuşlardır. Bu ilginin ana sebebi, bu alanlarda üzerinde çalışılan veri kümelerinin örnek ve öznitelik sayısı bakımından çok yüksek boyutlara ulaşmalarıdır.

MRMR (*Minimum Redundancy Maximum Relevance*) ve FCBF (*Fast Correlation-Based Filter*) gibi entropi tabanlı öznitelik seçme yöntemleri yüksek başarımlar ve hızlı çalışmaları nedeniyle tercih edilen öznitelik seçme yöntemleri arasındadırlar. Bu algoritmaların başarımları çeşitli veri kümeleri üzerinde sınanmış ve gözlenmiştir ki FCBF algoritması az sayıda özneteliğin seçildiği durumlarda gereğinden fazla öznitelik eleyerek sınıflandırma başarısının düşmesine yol açmaktadır.

Öznitelik seçme yöntemlerinin başarısını ölçen bir diğer ölçüt ise kararlılık (*stability*), yani veri kümesi üzerindeki küçük değişikliklere rağmen aynı ya da yakın özniteliklerin seçilmesidir. Özellikle seçilen özniteliklerin kimliklerinin çok önemli olduğu biyolojik dizi analizi alanında, bir veri kümesi için verdiği sonuçlar ile aynı dağılımdan gelen başka bir veri kümesi için verdiği sonuçlar arasında büyük farklar bulunan öznitelik seçme algoritmaları tercih edilmemektedir. Kararlılık, veri kümesinin örnek sayısının az ve boyut sayısının fazla olduğu durumlarda da büyük önem teşkil etmektedir. Eğitim kümesi üzerindeki ufak değişikliklerin seçilen öznitelikler açısından büyük farklar yaratmadığı bir öznitelik seçme algoritması başarılı bir algoritma olarak değerlendirilmektedir.

Bu çalışmada iki öznitelik kümesi arasındaki kararlılık değerini ölçen yeni bir kararlılık ölçütü ortaya konulmuştur. MRMR algoritmasının kullandığı iki seçim kriteri olan MID (*Mutual Information Difference*) ve MIQ (*Mutual Information Quotient*) yöntemleri hem başarımlar hem de kararlılık yönünden karşılaştırılmıştır. Bu iki metodun başarımlar yönünden yakın sonuçlar verdiği görülmüştür. Diğer taraftan özellikle küçük veri kümeleri için MID yöntemi MIQ yöntemine göre daha kararlı öznitelik kümeleri ile sonuçlanmıştır. MID yönteminin neden daha kararlı bir öznitelik seçim kriteri olduğu teorik olarak ortaya konulmaya çalışılmış ve bu sav deneysel olarak desteklenmiştir.

Ayrıca eğitim veri kümelerinin örnek sayısını azaltarak öznitelik seçme algoritmalarının kararlılığını ve başarımlarını ölçen bir deneysel metodoloji modeli sunulmuştur. MRMR algoritması için, özniteliklerin sınıf etiketleri ile ilişkisinin (*feature relevance*) ve öznitelik artıklığının (*feature redundancy*) öznitelik seçimindeki etkisini bir α parametresi yardımıyla kontrol eden yeni bir seçim kriteri, MID_{α} , ortaya konularak değişik α değerleri için farklı veri kümelerinde yeni seçim

krterinin kararlılık ve başarıım değeri gözlenmiştir. Yeni seçim kriteri MID_{α} ile öznitelik seçiminde kararlılığın kontrol edilebileceği deneysel olarak gösterilmiştir.

STABLE AND ACCURATE FEATURE SELECTION

SUMMARY

Feature selection methods can help machine learning algorithms produce faster and more accurate solutions because they reduce input dimensionality and they can eliminate irrelevant or redundant features. Especially for the last decade, feature selection methods became well known in the fields like pattern recognition, signal processing, text categorization and bioinformatics. The main reason of this interest is the dramatical increase both in instance and feature sizes of the data sets used in these research areas.

Entropy based feature selection algorithms, such as MRMR (*Minimum Redundancy Maximum Relevance*) and FCBF (*Fast Correlation-Based Filter*) are preferred feature selection methods because they are very fast and produce sets of features that result in quite accurate classifiers. Accuracy measurements of MRMR and FCBF feature selection algorithms are performed on various data sets and it is observed that FCBF may throw away good features when there is a smaller number of features left after elimination process and this situation leads to decrease in classification accuracy for some data sets.

In addition to accuracy, stability, having not too significant changes in the selected features when the identity of samples change, is also a measure of success for a feature selection algorithm. Especially in domains like microarray analysis, where the identity of selected features are important, a feature selection algorithm which returns a set of features for a certain data set and a completely different set of features for another data set from the same distribution would not be preferable. Stability could especially be a concern when the number of samples in a data set is small and the dimensionality is high. A feature selection algorithm is said to be stable if small changes in the identity of training samples does not result in a big change in the set of selected features.

In this study, a new stability measure that can measure the stability between two sets of features is introduced. The two feature evaluation methods within MRMR, MID (*Mutual Information Difference*) and MIQ (*Mutual Information Quotient*) are compared both in stability and accuracy. These two methods result in features with similar accuracy. On the other hand, especially for small data sets, MID results in more stable feature sets than MIQ. Theoretical explanation for why MID is a better criteria for small sample size is given and it is supported with experimental test results.

A new experimental model is also introduced in order to measure the accuracy and the stability of a feature selection algorithm for the case of a reduction of training samples. A new feature selection criterion, MID_α , where redundancy and relevance of selected features are controlled by parameter α is presented and using the predefined model, its accuracy and stability values are obtained for different α values for various datasets. It is empirically shown that stability of feature selection can be controlled using MID_α criterion.

1. GİRİŞ

Son yıllarda sınıflandırma algoritmalarının üzerinde çalıştıkları veri kümeleri gerek örnek sayısı gerekse öznitelik sayısı bakımından oldukça büyük rakamlara ulaşmışlardır. Bu nedenle araştırmacılar öznitelik seçme yöntemlerine her zamankinden fazla ihtiyaç duymaya başlamışlar ve dosya sınıflandırmadan sinyal işlemeye, biyoenformatikten örüntü tanıma işlemlerine kadar pek çok alanda kullanım yöntemlerini aramaya koyulmuşlardır. Öznitelik seçme algoritmalarının kullanım alanları her geçen gün artmaktadır.

Özellikle birbirleri ile yüksek korelasyona sahip bir çok özneliğin bulunduğu ve örnek sayısının az olduğu veri kümelerinde öznitelik seçme algoritmalarının önemi bir kat daha artmaktadır. Bu veri kümelerinde öznitelik seçme algoritmaları hem toplanan veri içindeki gürültülü, sapkın ve gereksiz öznitelikleri eleyerek veri kümesinin daha sağlıklı bir şekilde ifade edilmesini sağlarlar hem de düşük örnek sayılarında dahi sınıflandırma algoritmalarının yüksek başarımları göstermelerine katkıda bulunabilirler [1].

Öznitelik seçme yöntemleri temel olarak iki ana kategoride incelenmektedir: Sarmal yöntemler (*wrapper methods*) ve filtreleme yöntemleri (*filter methods*). Geriye doğru öznitelik seçimi (*backward feature selection*) ve ileri doğru öznitelik seçimi (*forward feature selection*) sarmal yöntemlere verilecek en iyi örneklerdir [2]. Bu metodlar, bir özneliğin seçimi ya da elenmesi kararını sınıflandırma algoritmalarının başarımlarını göz önünde bulundurarak vermeleri nedeniyle seçilen öznitelik kümeleri ile yüksek başarımlar elde edilebilir. Ancak her bir öznitelik seçimi ya da elenmesi aşamasında veri kümesi eğitimi ve testi gerçekleştirdikleri için zamanın kısıtlı olduğu problemlerde tercih edilmemektedirler. Filtreleme yöntemleri ise öznitelikler arası korelasyon/artıklık (*feature redundancy*) ve özniteliklerin sınıf etiketleri vektörü ile ilişkilerini (*feature relevance*) temel alan bir seçim kriteri yardımıyla öznitelikleri sıralayan ya da ağırlıklandıran algoritmalarlardır. ICA (*Independent Component Analysis*, [3]) ve PCA (*Principal Component Analysis*) öznitelikler arası korelasyon değerlerini kullanarak veri kümesinin daha düşük sayıda boyuta iz düşümünü alan

projeksiyon algoritmaları olarak bu gruba örnek gösterilebilir. Girdi verisinin normal dağılım gösterdiği durumlarda PCA, normal olmayan dağılım gösterdiği durumlarda ise ICA iyi sonuçlar vermektedir. Diğer taraftan bu iki yöntem de özniteliklerin sınıf etiketleri ile olan ilişkisini ya da sınıflandırıcı başarımını göz önünde bulundurmamak için elde edilen sonucun yüksek sınıflandırıcı başarısı vereceğini garanti etmek doğru olmaz. MRMR (*Minimum Redundancy Maximum Relevance*, [4]) ve FCBF (*Fast Correlation-Based Filter*, [5]) gibi filtreleme yöntemleri ise hem özniteliklerin sınıf etiketleri ile ilişkisini hem de özniteliklerin birbirleri arasındaki ilişki ve artıklık değerlerini hesaplayarak öznitelik seçimi yapmaları nedeniyle PCA ve ICA yöntemlerinden bir adım öndedirler. Bu algoritmalar bahsi geçen korelasyon değerlerini elde etmek için bilgi teorisi tabanlı entropi (*entropy*) hesabını kullanırlar. Gürültülü ve ilgisiz özniteliklerin yoğun olduğu ya da sınıflandırma algoritmalarının başedebileceğinden çok daha fazla özneliğin bulunduğu veri kümelerinde FCBF ve MRMR yöntemleri projeksiyon algoritmalarına göre çok daha hızlı çalışmaktadırlar. ICA ve PCA gibi FCBF ve MRMR öznitelik seçme algoritmaları da sınıflandırıcı başarımını öznitelik seçiminde hesaba katmamaları nedeniyle bazı durumlarda tercih edilmemektedirler. Filtreleme yöntemleri ile sarmal yöntemlerin birbirlerine göre avantajlarını kullanan hibrit öznitelik seçme yöntemleri son zamanlarda popülerleşmeye başlamışlardır [6,7].

Geleneksel inancıya göre, bir özneliğin sınıf etiketleri ile ilişkisi, korelasyonu öznitelik seçiminde en önemli kriterdir çünkü sınıf etiketleri ile yüksek ilişkili özniteliklerin kullanılması sınıflandırıcı başarımını arttırmaktadır [8]. Bu nedenle öznitelik seçme algoritmalarının büyük çoğunluğu sınıf etiketleri ile olan ilişki üzerine yoğunlaşmıştır [9]. Diğer taraftan veri kümesini en iyi şekilde genellemek amacıyla, minimum sayıda öznitelik ile veri kümesinin öznitelik uzayının büyük bir kısmının kapsanması istenmektedir. Bu arayış aralarındaki artıklığın minimum olduğu diğer bir deyişle birbirine benzemeyen, örtüşmeyen özniteliklerin seçilmesi anlamına gelmektedir. Özniteliklerin artıklığı ve özniteliklerin sınıf etiketleriyle olan ilişkileri arasında kurulan denge üzerine de bir çok çalışma bulunmaktadır ([4,10]).

Gözard edilen önemli bir konu ise veri kümesi üzerindeki ufak değişimler karşısında öznitelik seçme algoritmalarının gösterdiği kararlılıktır (*stability*). Öznitelik seçme algoritmalarının temelde aynı dağılımdan gelen farklı sayıda örnek içeren veri kümelerinde dahi tutarlı seçimler yapması beklenmektedir. Bu durum yüksek boyutlu

veri içeren her uygulama için geçerlidir. Özellikle öznitelik seçme algoritmalarının, gözlenen fenomenin karakteristik belirleyicilerini ortaya çıkarmak için kullanıldığı alanlarda büyük önem teşkil etmektedir [11]. Örnek olarak biyolojik dizi analizi (*microarray data analysis*) alanında, eğitim kümesi üzerindeki ufak değişiklikler çok farklı özniteliklerin, bu çalışmadaki diğer adıyla genlerin, seçilimine yol açabilmektedir [12,13]. Bu tutarsızlık bu alanda çalışan uzmanların biyolojik belirleyici olarak gördükleri özniteliklere, genlere duydukları güvenin azalmasına yol açmaktadır.

Literatürde öznitelik seçme algoritmalarının kararlılığı üzerine kısıtlı sayıda çalışma bulunmaktadır. Bu yöndeki bilinen ilk çalışmalardan biri Kalousis tarafından gerçekleştirilmiştir ve bu çalışma öznitelik seçme algoritmalarının kararlılığı üzerine deneysel verilerin bulunduğu en iyi çalışma olarak nitelendirilebilir [12]. İlgili çalışma birçok kararlılık ölçütü kullanarak öznitelik sıralama ve ağırlıklandırma algoritmalarının kararlılığını değişken veri kümeleri üzerinde test etmiş ve algoritmaların benzer sınıflandırma başarımları göstermelerine rağmen kararlılık yönünden çok farklı sonuçlar verebileceğini göstermiştir. Birlik öznitelik seçimi (*ensemble feature selection*, [14]) ve grup tabanlı öznitelik seçimi (*group-based feature selection*, [13]) olarak adlandırılan güncel iki çalışma sınıflandırma başarımından feragat etmeden kararlı öznitelik seçimi üzerine yoğunlaşmaktadır. Bahsi geçen çalışmalar kararlı öznitelik seçimi üzerinde durmalarına rağmen öznitelik seçiminde sınıf etiketleri ile ilişki ve öznitelikler arası artıklığa farklı oranda değer verilmesinin kararlılık ve başarımlar üzerindeki etkisi henüz tespit edilmemiştir. Öznitelik seçiminde kararlı sonuçlar elde etmek ve sınıf etiketleri ile ilgililik ve öznitelik artıklığı kavramlarının kararlı öznitelik seçiminde sahip olduğu etkiyi araştırmak bu çalışmanın motivasyonudur.

Bu bölümün ardından deneylerde kullanılan veri kümeleri hakkında kısa bilgiler verilecektir. Bölüm 3'te üstünde çalışılan algoritmaları tanıtmak amacıyla MRMR ve FCBF algoritmalarının çalışma prensipleri anlatılacak ve başarımlar yönünden karşılaştırılması yapılacaktır. Bölüm 4'te basit ama etkili bir öznitelik seçme yöntemi olan MRMR algoritmasının kararlılığı üzerine teorik ve deneysel çıkarımlar sunulacak ve MRMR algoritmasının ayrık veri kümeleri üzerinde uygulanması için geliştirilen iki yöntemden MID (*Mutual Information Difference*) ve MIQ (*Mutual Information Quotient*) arasındaki fark vurgulanarak, yapılan deneyler sonucunda

MID tekniğinin daha kararlı sonuçlar verdiği gösterilecektir. Bölüm 5'te MRMR algoritması için yeni bir öznitelik seçim kriteri olan MID_{α} tanıtıldıktan sonra Bölüm 6'da öznitelik seçiminde kararlılığın önemi vurgulanacak ve iki öznitelik kümesi arasındaki kararlılık seviyesini ölçen yeni bir ölçüt ortaya konulacaktır. Bölüm 7'de bir öznitelik seçme algoritmasının başarımlı ve kararlılık açısından başarısını ölçen bir deneysel metodoloji sunulacak ve Bölüm 8'de MID, MIQ ve MID_{α} yöntemlerinin geliştirilen deneysel metodoloji yardımıyla çeşitli veri kümeleri üzerindeki sonuçları gözlemlenecektir. Çalışma Tartışma ve Sonuç bölümüyle sonlanacaktır.

2. VERİ KÜMELERİ

Bu bölümde deneylerde kullanılan veri kümeleri hakkında kısa bilgiler verilecektir. Tüm veri kümelerine ait özet bilgiler Çizelge 2.1’de görülebilir.

Çizelge 2.1 : Deneylerde kullanılan veri kümeleri hakkında özet bilgiler.

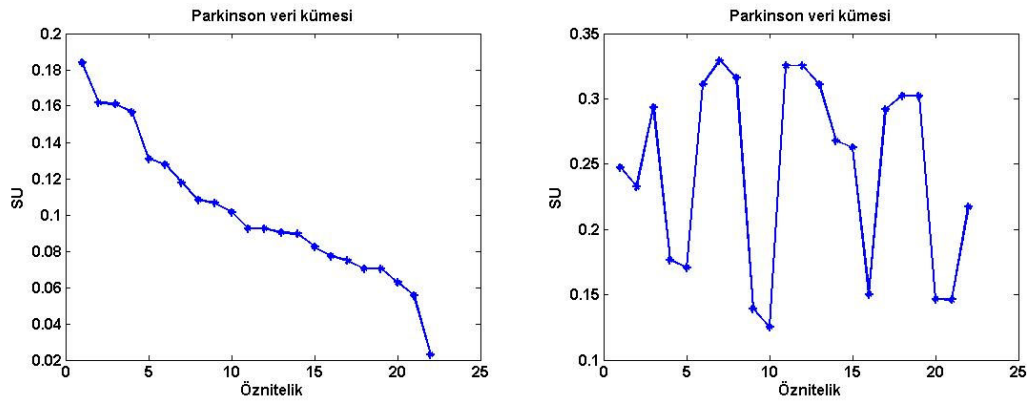
Veri Kümesi	Öznitelik Sayısı	Örnek Sayısı	Sınıf Sayısı
Parkinson	23	195	2
Sonar	60	208	2
İyonosfer	34	351	2
Musk (Versiyon 1)	166	476	2
Şarap	13	178	3
Ses Türü	50	500	5
Çoklu Öznitelik	649	2000	10
El Yazısı Rakamlar	64	3823	10

Bu çalışmada kullanılan öznitelik seçme yöntemleri olan MRMR ve FCBF algoritmaları, öznitelikleri ve sınıf etiketlerini ayrık rastlantı değişkenleri olarak görürler. Bu nedenle sürekli veri içeren veri kümelerinde bu algoritmalar uygulanmadan önce her bir öznitelik kendi içinde ayrıklaştırma (*discretization*) işlemine tabi tutulmalıdır. Bu işlem en basit haliyle her özniteliği maksimum ve minimum değerleri arasında 10 eşit bölmeye ayırarak ve her bölme için sınır değerleri elde edildikten sonra öznitelik değerlerini düştüğü bölmeyi ifade edecek şekilde numaralandırarak yapılmıştır.

Veri kümeleri ve barındırdıkları öznitelikler hakkında bilgi vermesi amacıyla her bir özniteliğin sınıf etiketleriyle olan simetrik belirsizlik (*Symetrical Uncertainty*), *SU*, değerleri hesaplanmış ve bu değerler büyükten küçüğe doğru sıralanmıştır. Veri kümelerinin açıklamalarının ardından gelen şekillerin solundaki grafikler bu değerleri göstermektedir. Bu sıralama esas alınarak her bir özniteliğin kendisi dışında kalan tüm öznitelikler ile arasındaki ortalama simetrik belirsizlik değerleri ise aynı şeklin sağ tarafındaki grafikte gösterilmiştir. Simetrik belirsizlik hesabı FCBF algoritmasının tanıtıldığı bölüm 3.2’de detaylı olarak açıklanmaktadır.

2.1 Parkinson Veri Kümesi

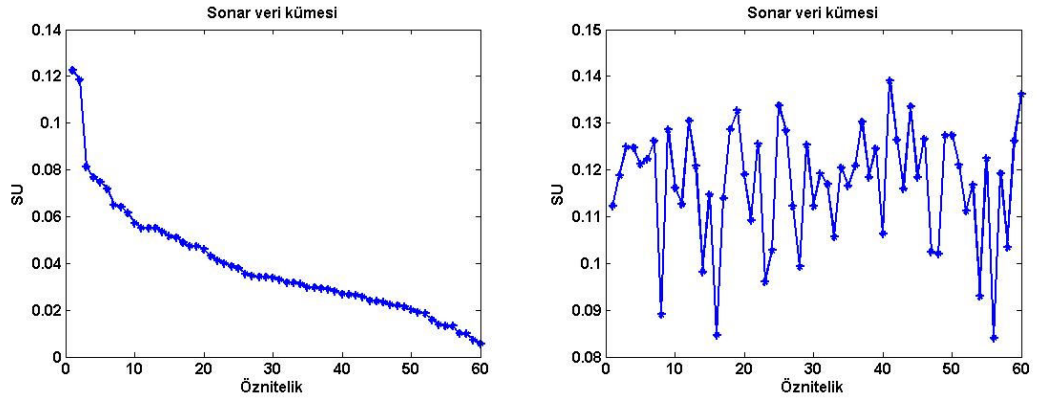
23'ü Parkinson hastası 31 kişiden alınmış biyomedikal ses ölçümlerini barındıran bir veri kümesidir. Veri kümesindeki 22 sütundan her biri belirli bir ses ölçütüne, 195 satırdan her biri ise bireylerden alınan ses kayıtlarına denk düşmektedir. Ana amaç hasta insanları sağlıklılardan ayırmaktır. 2 sınıflı bu veri kümesi UCI internet sitesinden [15] alınmıştır. Parkinson veri kümesindeki özniteliklerin sınıf etiketleri vektörü ile arasındaki simetrik belirsizlik değerleri ve kendi dışındaki diğer öznitelikler ile olan ortalama simetrik belirsizlik değerleri Şekil 2.1'de verilmiştir.



Şekil 2.1 : Parkinson veri kümesindeki özniteliklerin sınıf etiketleri vektörü ile arasındaki SU (solda) ve kendi dışındaki diğer öznitelikler ile olan ortalama SU (sağda) değerleri.

2.2 Sonar Veri Kümesi

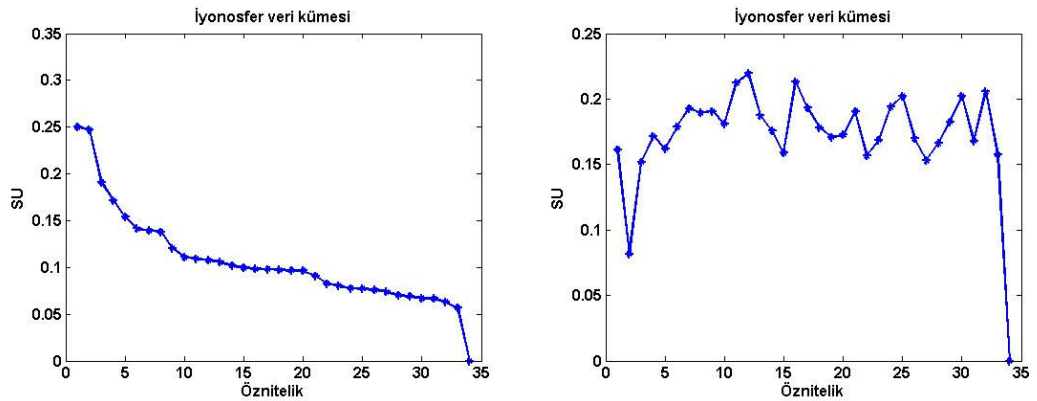
Veri kümesi iki farklı cisme değişik açılarla ve değişik koşullarda sonar sinyaller yansıtılarak oluşturulmuştur. 111 örnek metal silindirlerden, 97 örnek ise kayalardan elde edilmiştir. Her örneğe 0.0 ile 1.0 arasında değişen 60 öznitelik karşılık gelmektedir. Her bir değer belirli bir frekans aralığındaki enerji seviyesini ifade etmektedir. Toplamda 208 örnek içeren 60 özniteliğe sahip 2 sınıflı bir veri kümesidir ve UCI internet sitesinden [15] alınmıştır. Sonar veri kümesindeki özniteliklerin sınıf etiketleri vektörü ile arasındaki simetrik belirsizlik değerleri ve kendi dışındaki diğer öznitelikler ile olan ortalama simetrik belirsizlik değerleri Şekil 2.2'de verilmiştir.



Şekil 2.2 : Sonar veri kümesindeki özniteliklerin sınıf etiketleri vektörü ile arasındaki SU (solda) ve kendi dışındaki diğer öznitelikler ile olan ortalama SU (sağda) değerleri.

2.3 İyonosfer Veri Kümesi

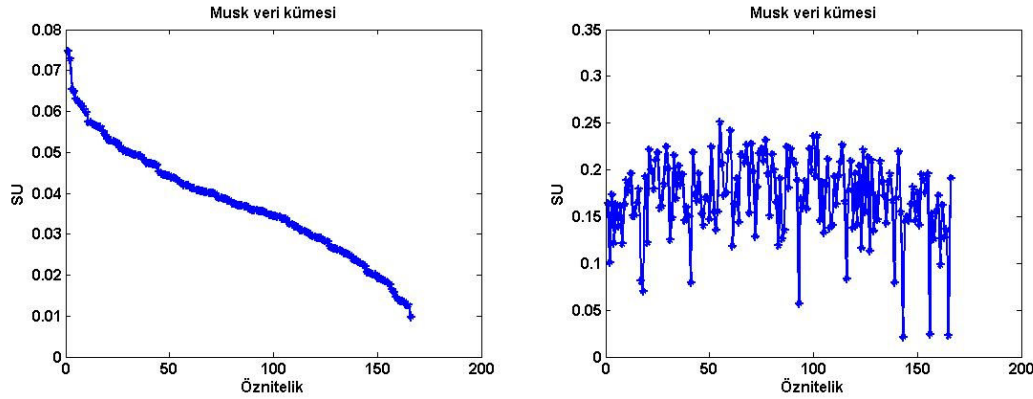
İyonosfer veri kümesi 2 sınıfa ait 351 gözlemin 34 özniteliğinden oluşmaktadır. Veri Labrador, Goose Bay'daki sistem tarafından toplanmıştır. Hedefler iyonosfer tabakasındaki serbest elektronlardır. Herhangi bir çeşit yapının iyonosferde olduğunun kanıtı olan durumlar “iyi”, olmadığını kanıtlayan durumlar ise “kötü” olarak işaretlenerek sınıf etiketleri oluşturulmuştur. Bu veri kümesi UCI internet sitesinden [15] alınmıştır. İyonosfer veri kümesindeki özniteliklerin sınıf etiketleri vektörü ile arasındaki simetrik belirsizlik değerleri ve kendi dışındaki diğer öznitelikler ile olan ortalama simetrik belirsizlik değerleri Şekil 2.3'te verilmiştir.



Şekil 2.3 : İyonosfer veri kümesindeki özniteliklerin sınıf etiketleri vektörü ile arasındaki SU (solda) ve kendi dışındaki diğer öznitelikler ile olan ortalama SU (sağda) değerleri.

2.4 Musk (Versiyon 1) Veri Kümesi

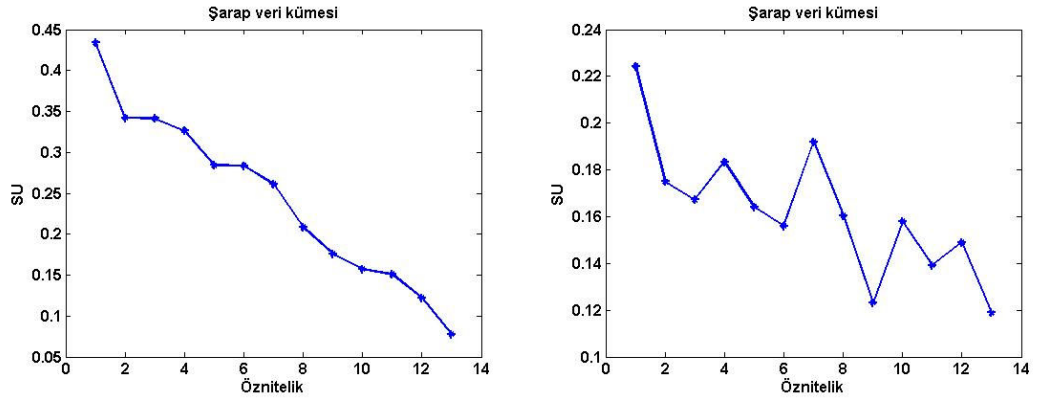
Musk (Versiyon 1) veri kümesi uzmanlar tarafından 47'si musk, 45'i musk olmayan olarak belirtilen 92 molekülü tanımlar. Burada amaç yeni moleküllerin musk olup olmadığını anlamaya çalışmaktır. Molekülleri tanımlayan 166 öznelik molekülün şekline ya da yapısına bağlıdır. Veri kümesi oluşturulurken moleküllerin düşük enerjili yapıları üretilmiş ve benzer yapılar filtrelenmiştir. Bu işlem sonucunda geriye 476 yapı kalmıştır. Daha sonra her bir yapı için öznelik vektörleri çıkartılmıştır. Bu veri kümesi UCI internet sitesinden [15] alınmıştır. Musk veri kümesindeki özneliklerin sınıf etiketleri vektörü ile arasındaki simetrik belirsizlik değerleri ve kendi dışındaki diğer öznelikler ile olan ortalama simetrik belirsizlik değerleri Şekil 2.4'te verilmiştir.



Şekil 2.4 : Musk veri kümesindeki özneliklerin sınıf etiketleri vektörü ile arasındaki *SU* (solda) ve kendi dışındaki diğer öznelikler ile olan ortalama *SU* (sağda) değerleri.

2.5 Şarap Veri Kümesi

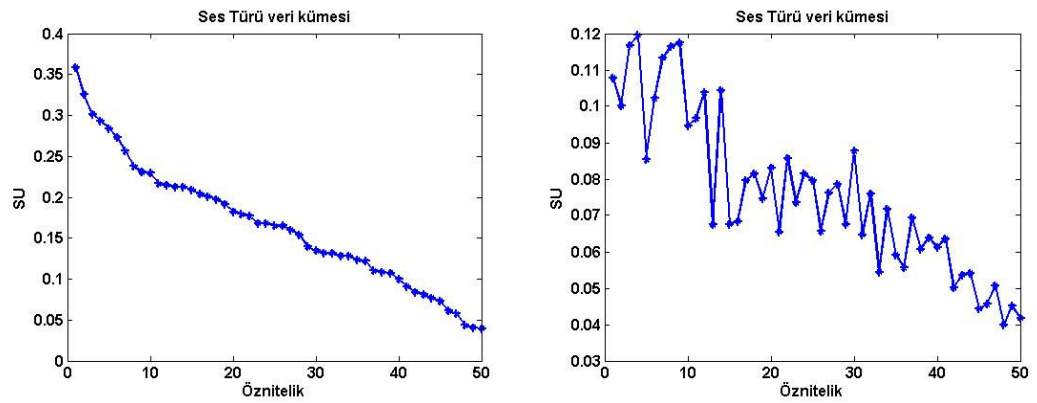
Bu veri kümesi İtalya'da aynı bölgede üç farklı zirai yöntem kullanılarak yetiştirilen şarapların kimyasal analizi sonucu elde edilmiştir. Bu analiz üç şarap tipinde de bulunan 13 bileşen değerinin elde edilmesini sağlamıştır. 13 sürekli özneliğe ve 178 örnek sayısına sahip bu veri kümesi UCI internet sitesinden [15] alınmıştır. Şarap veri kümesindeki özneliklerin sınıf etiketleri vektörü ile arasındaki simetrik belirsizlik değerleri ve kendi dışındaki diğer öznelikler ile olan ortalama simetrik belirsizlik değerleri Şekil 2.5'te verilmiştir.



Şekil 2.5 : Şarap veri kümesindeki özniteliklerin sınıf etiketleri vektörü ile arasındaki SU (solda) ve kendi dışındaki diğer öznitelikler ile olan ortalama SU (sağda) değerleri.

2.6 Ses Türü Veri Kümesi

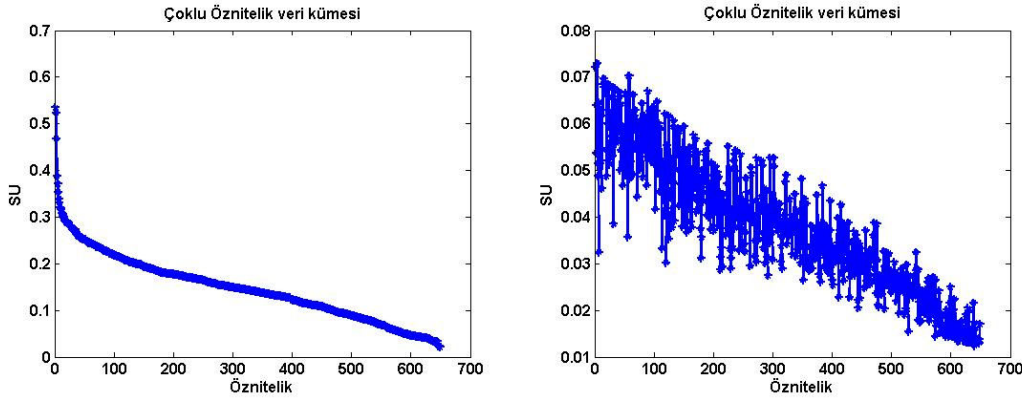
Ses Türü veri kümesi Tzanetakis [16] veri kümesinin en belirgin 5 sınıfı olan klasik, caz, pop, hiphop ve rege müzik türlerine ait 500 örneği içeren bir alt kümesidir. İki farklı öznitelik grubunun bir araya gelmesinden oluşmuştur. 30 öznitelik MARYAS yazılımı kullanılarak çıkartılmıştır [16]. 20 öznitelik ise [17]'te verilen yapı ile elde edilmiştir. Ses Türü veri kümesindeki özniteliklerin sınıf etiketleri vektörü ile arasındaki simetrik belirsizlik değerleri ve kendi dışındaki diğer öznitelikler ile olan ortalama simetrik belirsizlik değerleri Şekil 2.6'da verilmiştir.



Şekil 2.6 : Ses Türü veri kümesindeki özniteliklerin sınıf etiketleri vektörü ile arasındaki SU (solda) ve kendi dışındaki diğer öznitelikler ile olan ortalama SU (sağda) değerleri.

2.7 Çoklu Öznitelik Veri Kümesi

Çoklu öznitelik veri kümesi el yazısı rakamların sayısallaştırılmış ikili imgelerinden oluşmaktadır. Çeşitli yöntemlerle elde edilmiş 649 öznitelik 2000 örnek için çıkartılmıştır. Her rakam için 200 adet örnek içeren 10 sınıflı bu veri kümesi UCI internet sitesinden [15] alınmıştır. Çoklu Öznitelik veri kümesindeki özniteliklerin sınıf etiketleri vektörü ile arasındaki simetrik belirsizlik değerleri ve kendi dışındaki diğer öznitelikler ile olan ortalama simetrik belirsizlik değerleri Şekil 2.7’de verilmiştir.

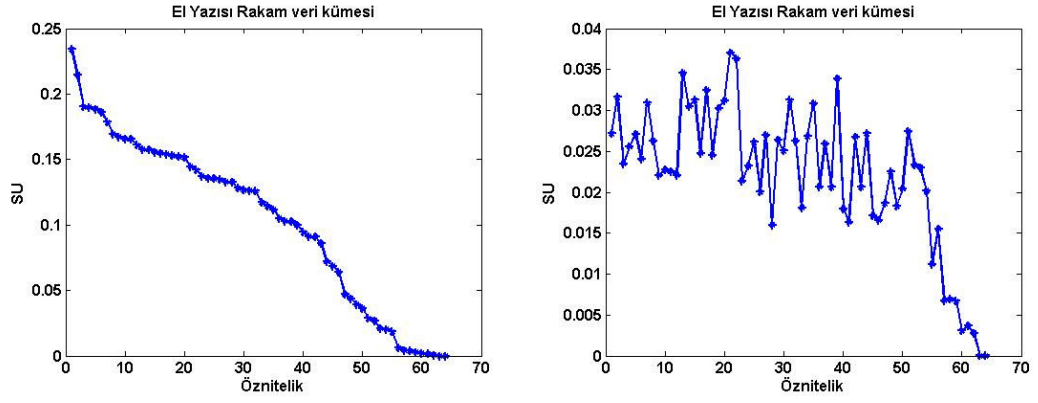


Şekil 2.7 : Çoklu Öznitelik veri kümesindeki özniteliklerin sınıf etiketleri vektörü ile arasındaki SU (solda) ve kendi dışındaki diğer öznitelikler ile olan ortalama SU (sağda) değerleri.

2.8 El Yazısı Rakamlar Veri Kümesi

Bu veri kümesi baskı formatındaki el yazısı rakamların NIST tarafından sağlanan önışlem programları sayesinde oluşturulmuş, normalize edilmiş bit haritalarından elde edilmiştir. 43 insanın 30’undan elde edilen örnekler ile eğitim kümesi kalan 13’ünden elde edilen örnekler ile test kümesi oluşturulmuştur. Her bir rakam için 32×32 ’lik bit haritaları üstüste gelmeyecek şekilde 4×4 ’lük bloklara ayrılmış ve her bloktaki piksel sayısı sayılmıştır. Bu işlem her bir örnek için 1 ile 16 arasında bir tam sayı değeri içeren 8×8 ’lik bir girdi verisi üretmiştir. Bu çalışmada deneyler eğitim kümesi üzerinde gerçekleştirilmiştir. Eğitim kümesindeki örnek sayısı 3823’tür. Öznitelik sayısı yukarıdan da anlaşılacağı gibi 64, sınıf sayısı ise 10’dur (0,1,2, ..., 9). Bu veri kümesi UCI internet sitesinden [15] alınmıştır. El Yazısı Rakamlar veri kümesindeki özniteliklerin sınıf etiketleri vektörü ile arasındaki simetrik belirsizlik

değerleri ve kendi dışındaki diğer öznitelikler ile olan ortalama simetrik belirsizlik değerleri Şekil 2.8’de verilmiştir.



Şekil 2.8 : El Yazısı Rakamlar veri kümesindeki özniteliklerin sınıf etiketleri vektörü ile arasındaki SU (solda) ve kendi dışındaki diğer öznitelikler ile olan ortalama SU (sağda) değerleri.

3. ÖZNİTELİK SEÇME ALGORİTMALARI

Bu bölümde entropi tabanlı iki öznitelik seçme algoritması, MRMR ve FCBF, tanıtılarak çalışma prensipleri anlatılacak ve birbirlerine göre karşılaştırılmaları sunulacaktır.

3.1 MRMR Algoritması

MRMR (*Minimum Redundancy Maximum Relevance*) algoritması, sınıf etiketleriyle en ilişkili (*relevance*) öznitelikleri seçmeye çalışırken eş zamanlı olarak seçilen öznitelikler arasındaki artıklığı (*redundancy*), fazlalığı minimize etmeye çalışan bir filtreleme yöntemidir [4]. Bu bölümde MRMR algoritmasının detayları sunulacaktır. Algoritma her bir özniteliği ve sınıf etiketleri vektörünü ayrı bir rastlantı değişkeni olarak görür ve iki öznitelik arasındaki ya da bir öznitelik ile sınıf etiketleri vektörü arasındaki benzerlik seviyesini ölçmek için aralarındaki ortak bilgiyi (*mutual information*), $I(X, Y)$, kullanır.

$$I(X, Y) = \sum_{y \in Y} \sum_{x \in X} p(x, y) \log \left(\frac{p(x, y)}{p_1(x)p_2(y)} \right) \quad (3.1)$$

Yukarıdaki denklemde $p(x, y)$, X ve Y rastlantı değişkenlerinin birleşik olasılık dağılım fonksiyonunu, $p_1(x)$ ve $p_2(y)$ ise sırasıyla X ve Y rastlantı değişkenlerinin marjinal olasılık dağılım fonksiyonlarını göstermektedir. Ortak bilgi iki rastlantı değişkeninin tamamıyla bağımsız olduğu durumlarda 0 değerini almaktadır. Bu değer negatif olamaz ve simetriktir ($I(X, Y) \geq 0$, $I(X, Y) = I(Y, X)$).

Gösterimde kolaylık sağlamak amacıyla, her bir öznitelik f_i , n tane gözlemin bir araya gelerek oluşturduğu bir vektör olarak tanımlanacaktır ($f_i = [f_i^1, f_i^2, f_i^3, \dots, f_i^n]$). Bir öznitelik vektörünün reel değerlerini barındıran f_i vektörü ise F_i ayrı rastlantı değişkeninin bir örneği olarak ele alınacak ve öznitelik i ve j arasındaki ortak bilgi, $I(F_i, F_j)$, bu ayrı rastlantı değişkenleri üzerinden hesaplanacaktır. Burada $i, j = 1, 2, \dots, d$, d ise veri kümesinin boyut sayısını diğer bir

değişle veri kümesindeki öznitelik sayısını göstermektedir. Ortak bilgi kavramı sadece iki öznitelik arasındaki benzerlik seviyesini değil herhangi bir öznitelik i ile sınıf etiketleri vektörü, h ($h = [h^1, h^2, h^3, \dots, h^n]$) ve ona denk düşen ayırık rastlantı değişkeni, H , arasındaki benzerliğin, $I(H, F_i)$, ölçülmesinde de kullanılmaktadır.

S seçilmek istenen öznitelik kümesi olsun ve $|S|$ bu kümenin eleman sayısını göstere. Seçilmek istenen kümenin seçilebilecek en iyi küme olmasını garanti etmek için yukarıda bahsedilen iki koşulun karşılanması gerekir. Bunlardan ilki maksimum ilişki (*maximum relevance*):

$$\max W, W = \frac{1}{|S|} \sum_{F_i \in S} I(F_i, H) \quad (3.2)$$

İkincisi ise minimum artıklık (*minimum redundancy*):

$$\min V, V = \frac{1}{|S|^2} \sum_{F_i, F_j \in S} I(F_i, F_j) \quad (3.3)$$

koşuludur. MRMR algoritması bu iki koşulu iki basit kombinasyon ile birleştirmektedir:

$$\max (V - W) \quad (3.4)$$

$$\max (V/W) \quad (3.5)$$

İdeal olarak yukarıda tanımlanan iki koşulun karşılanması sonucunda en iyi öznitelik kümesinin elde edilmesi $O(N^{|S|})$ karmaşıklığa sahip bir algoritma gerektirmektedir. Arama uzayının çok büyük olması nedeniyle, MRMR algoritması bu optimizasyon sorununu şu şekilde çözer: Öncelikle ilk öznitelik Eşitlik 3.2'ye göre seçilir. Bu demektir ki seçilen ilk öznitelik sınıf etiketleri ile maksimum seviyede ortak bilgi içeren öznitelik olacaktır. Bundan sonraki her basamakta Eşitlik 3.6 veya 3.7'yi sağlayan öznitelik F_i seçilir ve seçilen öznitelikler S kümesinde tutulur. Burada S algoritmanın her basamağında seçilen özniteliklerin tutulduğu kümeyi, Ω tüm öznitelik kümesini, Ω_S ise seçilmiş öznitelikler dışında kalan tüm öznitelikleri ifade etmektedir ($\Omega_S = \Omega - S$).

$$\max_{F_i \in \Omega_S} \left[I(F_i, H) - \frac{1}{|S|} \sum_{F_j \in S} I(F_i, F_j) \right] \quad (3.6)$$

$$\max_{F_i \in \Omega_S} \left\{ I(F_i, H) / \left[\frac{1}{|S|} \sum_{F_j \in S} I(F_i, F_j) \right] \right\} \quad (3.7)$$

İlk öznitelik seçilmesinin ardından Eşitlik 3.6 ya da 3.7’yi sağlayan öznitelik S kümesine dahil edilir ve istenilen öznitelik sayısına ulaşıldığında algoritma sonlanır. Maksimum ilişki ve minimum artıklık koşullarının bu iki basit kombinasyonu MRMR algoritmasının ayırık veriler üstünde kullanılan iki seçim kriterini ortaya çıkarır. Bu iki seçim kriteri Çizelge 3.1’de özetlenmiştir. Yukarıda geçen algoritmanın $O(Big O)$ notasyonuna göre karmaşıklığı $O(|S|N)$ ’dir.

Çizelge 3.1 : MRMR algoritmasının öznitelik seçim kriterleri.

Seçim Kriteri	Formül
MID, Ortak Bilgi Farkı (<i>Mutual Information Difference</i>)	$\max_{F_i \in \Omega_S} \left[I(F_i, H) - \frac{1}{ S } \sum_{F_j \in S} I(F_i, F_j) \right]$
MIQ, Ortak Bilgi Oranı (<i>Mutual Information Quotient</i>)	$\max_{F_i \in \Omega_S} \left\{ I(F_i, H) / \left[\frac{1}{ S } \sum_{F_j \in S} I(F_i, F_j) \right] \right\}$

3.2 FCBF Algoritması

FCBF (*Fast Correlation-Based Filter*) L. Yu and H. Liu tarafından 2003 yılında sunulan bir öznitelik seçme algoritmasıdır [5]. Algoritma öznitelikler arası ilişki (*relevance*) ve artıklık (*redundancy*) değerlerini temel alarak geliştirilmiştir. İlk olarak öznitelikler ile sınıf etiketleri, H , arasındaki ilişki, diğer bir deyişle, her bir öznitelik ile sınıf etiketleri ile olan korelasyonu hesaplanır. Korelasyon değeri belli bir eşik değerine eşit ya da bu değerin altında olan öznitelikler elenmektedir. Hesaplanan doğrusal olmayan korelasyon değerinin her zaman pozitif olması nedeniyle, 0 eşik değeri durumunda, tüm öznitelikler, sınıf etiketleri ile en ilişkiliden en ilişkisize doğru sıralanmış olur. 0’dan büyük bir eşik değerinin kullanılması bazı özniteliklerin ilk etapta elenmesine yol açacağından algoritmanın hızlanmasında etkili olmaktadır.

Sınıf etiketleri ile düşük korelasyona sahip özniteliklerin elenmesinin ardından, kalan öznitelikler arasından sınıf etiketleri ile en ilişkili olanlar kullanılarak artık öznitelikler saptanmaya çalışılır. Artıklık iki özniteliğin birbiriyle gereğinden fazla ilişkili olması ya da iki öznitelik arasındaki korelasyon değerinin yükek olması olarak değerlendirilebilir. Algoritmanın artık öznitelikleri eleme aşaması olan ikinci etap şu şekilde çalışmaktadır: Diyelim ki öznitelik i 'nin sınıf etiketleri ile korelasyonu öznitelik j 'den yüksektir. Eğer öznitelik i ile öznitelik j arasındaki korelasyon değeri, öznitelik j ile sınıf etiketleri vektörü arasındaki korelasyon değerinden yüksek ya da bu değere eşit ise öznitelik j artık bir öznitelik olarak tanımlanır ve öznitelik i 'nin varlığında öznitelik j 'ye gerek duyulmadığı için öznitelik kümesinden atılır.

$$SU_{i,j} \geq SU_{j,h} \rightarrow Ele F_j \quad (3.8)$$

Burada SU Eşitlik 3.9'da tanımlanan iki ayrık rastlantı değişkeni arasındaki simetrik belirsizlik (*symmetrical uncertainty*) değerini ifade etmektedir. Bu işlem sınıf etiketleriyle aralarındaki korelasyonlarına göre sıralanmış öznitelikler için ilk öznitelikten başlayarak kalan tüm öznitelikler için tekrarlanır. İlk döngü bittiğinde bir sonraki öznitelik ile kalan öznitelikler arasında yine aynı işlem uygulanır ve hiçbir özniteliğin artık olarak değerlendirilemediği döngü ile algoritma sonlanır. İkinci etap sonucunda elimizde kalan öznitelikler, sınıf etiketleri ile en ilişkili, kendi aralarında ise en az artıklığa sahip öznitelik kümesi olarak değerlendirilmektedir.

İki öznitelik arasındaki ilişki, artıklık ve bir öznitelik ile sınıf etiketleri arasında ilişki değerleri entropi tabanlı doğrusal olmayan bir korelasyon ölçütü olan simetrik belirsizlik, SU :

$$SU(X,Y) = 2 \left[\frac{IG(X;Y)}{H(X) + H(Y)} \right] \quad (3.9)$$

kullanılarak hesaplanmaktadır [18]. Simetrik belirsizlik birbirlerine göre bağımsız iki rastlantı değişkeni için minimum değeri olan 0 değerini, bir rastlantı değişkeninin diğerini tamamiyle öngördüğü, ifade ettiği durumda ise maksimum değeri olan 1 değerini alan simetrik bir ölçüttür ($SU(X,Y) = [0,1]$, $SU(X,Y) = SU(Y,X)$). Eşitlik 3.9'da, X ve Y herhangi iki özniteliği ya da bir öznitelik ve sınıf etiketleri vektörü çiftini ifade etmektedir. Simetrik belirsizlik ölçütünün temeli bir bilgi teorisi

(*information theory*) terimi olan entropi (*entropy*) değerine, bir rastlantı değişkeninin belirsizlik ölçümüne dayanmaktadır. X rastlantı değişkeninin entropisi $H(X)$:

$$H(X) = - \sum_i P(x_i) \log_2(P(x_i)) \quad (3.10)$$

Y rastlantı değişkeni gözlemlendikten sonra X rastlantı değişkeninin entropisi $H(X|Y)$ ise şu şekilde:

$$H(X|Y) = - \sum_j P(y_j) \sum_i P(x_i|y_j) \log_2(P(x_i|y_j)) \quad (3.11)$$

hesaplanmaktadır. X rastlantı değişkeni beraberinde Y rastlantı değişkeninin bilinmesinin X 'in entropi değerinde meydana getirdiği düşüş, diğer bir deyişle Y biliniyorken X 'in bilgi kazancı (*information gain*), $IG(X; Y)$ olarak tanımlanır [19]:

$$IG(X; Y) = H(X) - H(X|Y) \quad (3.12)$$

X ve Y rastlantı değişkenleri için doğrusal korelasyon katsayısı r :

$$r = \frac{\sum_i (x_i - \bar{x}_i)(y_i - \bar{y}_i)}{\sqrt{\sum_i (x_i - \bar{x}_i)^2} \sqrt{\sum_i (y_i - \bar{y}_i)^2}} \quad (3.13)$$

yerine simetrik belirsizlik ölçütünün kullanılmasının nedeni öznitelikler arası ya da bir öznitelik ile sınıf etiketleri arasındaki korelasyonun doğrusal olmadığı durumlarda doğrusal korelasyon ölçümünün yanlış sonuçlar verebileceği düşünülmektedir.

FCBF algoritmasının diğer öznitelik seçme algoritmalarına göre en avantajlı yanlarından biri hızlı çalışmasıdır. Bir filtreleme yöntemi olan FCBF algoritması, öznitelikler ile sınıf etiketleri arasındaki ya da iki öznitelik arasındaki korelasyon değerlerini hesaplayarak bir özneliğin sınıflandırmadaki değerine sınıflandırıcı algoritmalarının sonucuna gerek duymadan saptayabilmektedir. Bu yönüyle sarmal öznitelik seçme algoritmalarına göre hız açısından büyük avantaj sağlamaktadır. Diğer taraftan, algoritma öznitelikler ile sınıf etiketleri arasındaki ilişkiyi hesaplarken, her özneliği tek başına değerlendirmeye alması sebebiyle tek başına kötü ancak grup olarak çok daha iyi sonuç verebilecek öznitelikleri göz ardı etmektedir. Algoritmanın O (*Big O*) notasyonuna göre karmaşıklığı n örnek içeren ve

d öznitelik sayısına sahip bir veri kümesi için ikinci etapta her eleme döngüsünde kalan özniteliklerin ortalama olarak yarısının elendiği varsayımı yapılarak $O(nd \log d)$ olarak ifade edilebilir [5].

3.3 MRMR ve FCBF Algoritmalarının Karşılaştırılması

Görüldüğü gibi FCBF algoritması ile MRMR algoritması arasında iki önemli fark bulunmaktadır. MRMR algoritması özünde tüm öznitelikleri en değerliden değersize doğru sıralamaya çalışır ve sınıflandırma problemi için kaç adet öznitelik kullanılması gerektiği kararını kullanıcıya bırakır. Bu nedenle MRMR algoritması bir öznitelik seçme algoritmasından ziyade öznitelik sıralama algoritması olarak değerlendirilmelidir. FCBF algoritması ise bir öznitelik kümesi seçme algoritmasıdır. Algoritma, sınıflandırma probleminde kullanılmak üzere belli sayıda özniteliği gruplar ve bu öznitelik kümesinin bir değer sırasına sahip olduğunu söylemek doğru olmaz.

FCBF algoritmasının sınıflandırma problemlerinde kullanılacak öznitelik sayısını otomatik olarak saptaması, verilerin dağılımı hakkında fazla bilginin bulunmadığı ya da sınıflandırma algoritması için öznitelik sayısının sorun teşkil etmediği durumlarda yararlıdır. Ancak FCBF algoritmasının verdiği öznitelik grubunun, veri kümesini en iyi şekilde ifade eden ya da en iyi sınıflandırma başarısı göstermesi beklenen öznitelik kümesi olduğu iddia edilemez. Çizelge 3.2’de görülebileceği gibi FCBF algoritması bazı durumlarda gereğinden fazla özniteliği devre dışı bırakarak başarımların değerlerinin ciddi şekilde düşmesine yol açmıştır. Ayrıca deneysel açıdan değerlendirildiğinde, algoritmanın veri kümesinden bootstrap gibi örnekleme yöntemleriyle elde edilen alt kümeler üzerinde uygulandığında farklı sayıda öznitelik içeren ve değer sıralamasına sahip olmayan kümeler vermesi, kararlılığının ve başarımlarının kendisiyle ya da farklı yöntemlerle karşılaştırılmasını zorlaştırmaktadır.

Diğer taraftan sınıflandırma algoritmasının zaman ya da bellek kullanımı gibi sorunlarla başa çıkması gerektiği durumlarda MRMR öznitelik seçme algoritmasının istenilen sayıda öznitelik ile sonuçlanma özelliği büyük avantaj sağlamaktadır. MRMR algoritmasında sınıflandırma problemi için en başarılı ve kararlı sonucu verecek öznitelik sayısını saptamak bir ek iş olarak karşımıza çıkmasına rağmen sonuç olarak sabit bir kümeden ziyade bir sıralama vermesi kullanıcıya seçim şansı

vermekte ve deneysel çalışmaların sonunda sınıflandırıcı için en uygun öznitelik kümesinin saptanmasına olanak sağlamaktadır.

Çizelge 3.2’de deneylerde kullanılan veri kümeleri üzerinde bootstrap yöntemi kullanılarak eğitim kümeleri oluşturulmuş ve seçilmeyen örnekler test verisi olarak kullanılmıştır. Eğitim kümeleri üzerinde sırasıyla FCBF, MID seçim kriteri ile MRMR ve MIQ seçim kriteri ile MRMR algoritmaları uygulanmıştır ve elde edilen öznitelik kümeleri test veri kümeleri üzerinde k-en-yakın-komşu ($k = 3$) algoritması kullanılarak başarımlar yönünden test edilmiştir. Sayı değerleri içeren ilk sütun FCBF algoritmasının eğitim kümesi üzerinde hangi sayıda öznitelik seçerek sonlandığını ve bu öznitelikler kullanılarak elde edilen başarımlar değerlerini göstermektedir. İkinci sütun aynı sayıda öznitelik kullanılarak elde edilen MRMR-MID sonucunun başarımlar değerini, üçüncü sütun ise yine aynı sayıda öznitelik kullanılarak elde edilen MRMR-MIQ sonucunun başarımlar değerini göstermektedir. Son iki sütun ise MRMR algoritmasının iki seçim kriteri kullanılarak elde edilen sonuçlar arasında maksimum başarımlar veren öznitelik sayılarını ve başarımlar değerlerini göstermektedir. Görüldüğü gibi FCBF ve MRMR aynı sayıda öznitelik ile sonuçlandığında başarımlar değerleri birbirine çok yakın olmasına rağmen çoğu veri kümesinde FCBF gereğinden fazla özniteliği eleyerek başarımlar değerlerinin daha yüksek olabilecek iken düşmesine yol açmıştır. Diğer taraftan bu deneyde yapıldığı gibi MRMR algoritması sonucu oluşabilecek tüm öznitelik kümeleri için sınıflandırıcı başarımlar hesaplayarak en başarılı öznitelik sayısını ve karşılık gelen başarımlar değerini bulmak hayli zaman ve bellek harcayan bir yöntemdir. Algoritmaların veri kümeleri üzerinde elde ettikleri sonuçlar özniteliklerin indeks numaraları kullanılarak Ekler bölümünde Çizelge A.1-A.8 çizelgelerinde verilmiştir.

Çizelge 3.2 : FCBF ve MRMR algoritmalarının farklı veri kümelerinde başarımlarının yönünden karşılaştırılması.

Veri Kümesi	FCBF, Başarım (%)	MRMR, MID Başarım (%)	MRMR, MIQ Başarım (%)	MRMR, MID En Yüksek Başarım (%)	MRMR, MIQ En Yüksek Başarım (%)
Parkinson	2/0.6957	2/0.6957	2/0.6957	7/0.8116	5/0.8261
Sonar	2/0.5333	2/0.6533	2/0.6533	31/0.8667	58/0.8667
İyonosfer	2/0.8583	2/0.8583	2/0.8583	5/0.8898	6/0.9055
Musk	1/0.6108	1/0.6108	1/0.6108	49/0.8919	81/0.8973
Şarap	6/0.6721	6/0.6721	6/0.6721	3/0.9508	3/0.9508
Ses Türü	14/0.5172	14/0.5440	14/0.6322	3/0.6034	12/0.6322
Çoklu Öznitelik	151/0.9289	151/0.9412	151/0.9275	131/0.9439	253/0.9453
El Yazısı Rakamlar	24/0.9696	24/0.9646	24/0.9703	43/0.9795	39/0.9781

4. MID VE MIQ YÖNTEMLERİNİN KARARLILIK YÖNÜNDEN KARŞILAŞTIRILMASI

Bu bölümde MRMR öznitelik seçme algoritmasının iki seçim kriteri olan MID ve MIQ tekniklerinin kararlılık yönünden deneysel ve teorik olarak karşılaştırılması yapılmaya çalışılacaktır. Bu bölümün motivasyonu MRMR algoritmasında hesaplanan entropi değerlerinin sınırlı sayıda örnek içeren veri kümeleri kullanılması nedeniyle hata paylarına sahip olmaları ve bu hata dağılımlarının varyansını ya da hesaplanan entropi değerlerinin varyansını minimum seviyede tutan hesaplama yönteminin aranmasıdır.

Öznitelik seçme ve sınıflandırma algoritmaları doğaları gereği kısıtlı sayıda örnek içeren veri kümeleri üzerinden hesaplamalar yaparak verinin gerçek dağılımını dolayısıyla da aynı dağılımdan gelen yeni veriler için sınıf etiketlerini kestirmeye çalışırlar. Bu çalışmada üstünde durulan öznitelik seçme algoritması olan MRMR algoritması, öznitelikleri ayrık rastlantı değişkenleri olarak görmekte ve bu varsayım üzerinden gerekli korelasyon değerlerini hesaplamaktadır. Bir rastlantı değişkeninin gerçek dağılımının saptanabilmesi için o dağılımdan gelen tüm verilerin elimizde olması gerekmektedir ancak bu çoğu durumda olası değildir. Ayrıca MRMR algoritması daha hızlı çalışmasını sağlamak amacıyla, sürekli değerler içeren öznitelikler için dahi, entropi hesabından önce ayrıklaştırma (*discretization*) işlemi uygulanmasını gerektirmektedir. Bir öznitelik için tüm örnek uzayına ve olasılık değerlerine sahip olsak dahi bu ayrıklaştırma işlemi sebebiyle hesaplanan entropi değerleri için hata payı kaçınılmaz olmaktadır. Bu hata payının minimum seviyede tutulması algoritmanın kararlılığını ve başarımını etkileyen bir özellik olarak karşımıza çıkmaktadır.

4.1 Teorik Analiz

Bölüm 3.1’de bahsedildiği üzere, MRMR algoritması bir sonraki özniteliği tespit etmeye çalışırken birbirine çok benzer iki yöntem kullanmaktadır: MID (*mutual information difference*) ve MIQ (*mutual information quotient*). MID, öznitelik F_i ile

sınıf etiketleri vektörü H arasındaki ortak bilginin (V) öznitelik F_i ile öznitelik F_j arasındaki ortalama ortak bilgi (W) ile farkı ($V - W$), MIQ ise öznitelik F_i ile sınıf etiketleri vektörü H arasındaki ortak bilginin öznitelik F_i ile öznitelik F_j arasındaki ortalama ortak bilgiye oranı (V/W) olarak tanımlanmıştı ($F_i \in \Omega_S$, $F_j \in S$ ve $j = 1, 2, \dots, |S|$).

Algoritmanın sınırlı sayıda örnek içeren veri kümeleri üstünde uygulanması nedeniyle özniteliklerin ya da sınıf etiketleri vektörünün gerçek olasılık dağılım fonksiyonu elde edilememektedir. Öznitelikler ve sınıf etiketleri vektörü ayrık rastlantı değişkenleri olarak tanımlandıkları için iki öznitelik arasındaki ortak bilgi hesabı ile bir öznitelik ve sınıf etiketleri vektörü arasındaki ortak bilgi hesabının benzer tipte hata payı içerdiği kabul edilebilir.

Diyelimki V ve W , yukarıda özetlenmeye çalışılan hesaplamaların gerçek değerlerini, $V + \varepsilon$ ve $W + \delta$ ise bu hesaplamaların n sayıda örnek içeren veri kümesi üstünde hesaplanmasından kaynaklanan gerçek değerler ve hata paylarını ifade etsin. Bu aşamada ε hata payının dağılımını 0 ortalamalı ve σ_n^2 varyanslı normal dağılım olarak tanımlamak yanlış olmaz, $N(0, \sigma_n^2)$. W değerinin $|S|$ sayıda ortak bilgi değerinin ortalaması olması, δ dağılımında $|S|$ sayıda $N(0, \sigma_n^2)$ dağılımın ortalaması olmasını gerektirir ve bu ortalama dağılımda yine $N(0, \sigma_n^2)$ olarak kabul edilebilir.

$(V + \varepsilon) - (W + \delta)$ ve $(V + \varepsilon)/(W + \delta)$ değerlerinin varyansları sırasıyla MID ve MIQ tekniklerinin kararlılığına ilişkin göstergelerin başındadır çünkü yüksek varyans seviyesi öznitelik seçim aşamasında yüksek hata payına sahip hesaplamalar sonucunda farklı öznitelikler seçilmesine yol açabilir. MID tekniğinin varyansı Eşitlik 4.1' de görüldüğü gibi kolayca hesaplanabilmektedir:

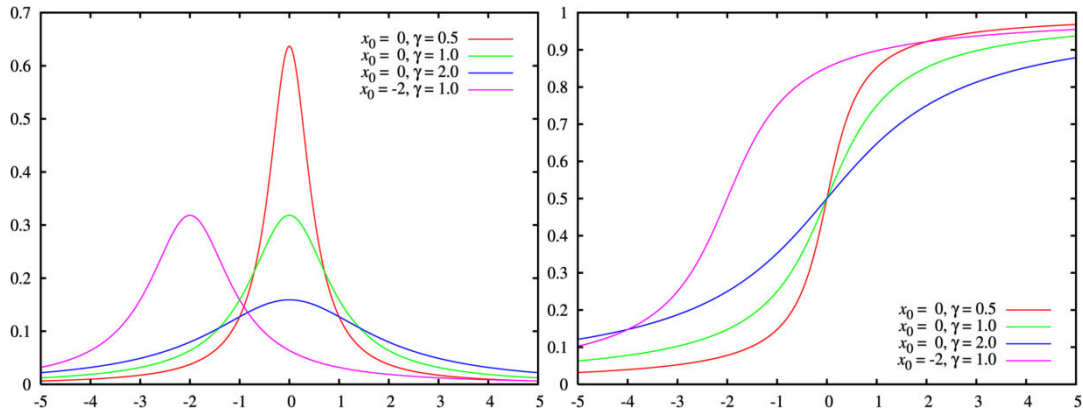
$$var(MID) = var((V + \varepsilon) - (W + \delta)) = 2\sigma_n^2 \quad (4.1)$$

Diğer taraftan MIQ tekniğinin yani $(V + \varepsilon)/(W + \delta)$ hesabının varyansını hesaplamak çok daha karmaşıktır. İlk olarak $W + \delta$ değeri 0 etrafında ihmal edilemeyen bir dağılıma sahipse, $(V + \varepsilon)/(W + \delta)$ oranı ortalaması ve varyansı tanımsız ve ikinci momenti sonsuz olan Cauchy dağılımı göstermektedir [20]. Eğer $(V + \varepsilon)/(W + \delta)$ oranında pay ve payda 0'dan uzak değerler ise bu oran V/W ortalamalı ve birim varyanslı normal dağılım göstermektedir.

Cauchy dağılımının olasılık dağılım fonksiyonu Eşitlik 4.2’de verilmiştir.

$$f(x; x_0, \gamma) = \frac{1}{\pi} \left[\frac{\gamma}{(x - x_0)^2 + \gamma^2} \right] \quad (4.2)$$

Burada x_0 dağılımın tepe noktasının konumunu belirleyen konum parametresini (*location parameter*), γ ise ölçekleme parametresini (*scale parameter*) göstermektedir. Cauchy dağılımının olasılık yoğunluk fonksiyonu ve kümülatif dağılım fonksiyonu Şekil 4.1’de görülebilir [21].

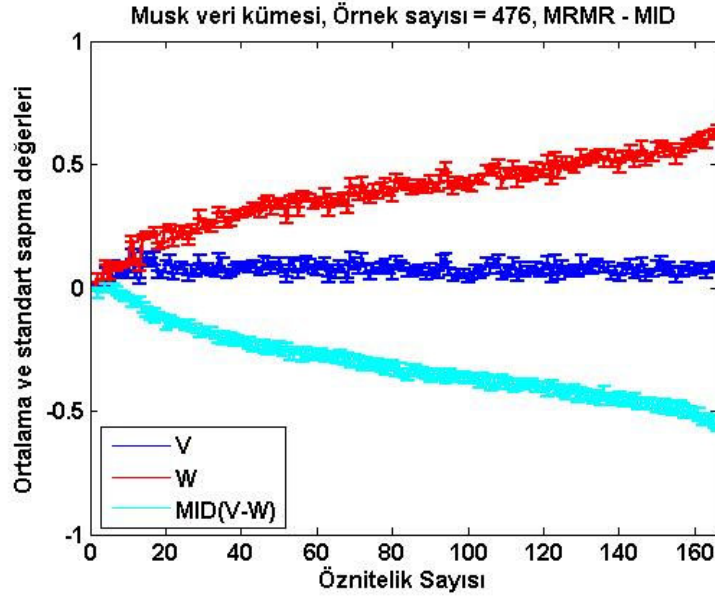


Şekil 4.1 : Cauchy dağılımının olasılık yoğunluk fonksiyonu (solda) ve kümülatif dağılım fonksiyonu (sağda).

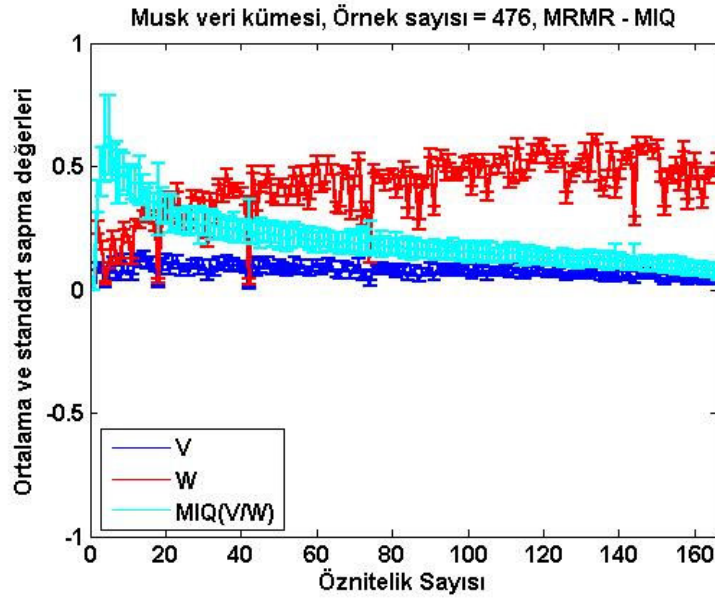
Bölüm 4.2’de görüleceği gibi düşük sayıda öznitelik seçildiği durumlarda W değeri 0’a yaklaşmakta ve $(V + \varepsilon)/(W + \delta)$ (MIQ) değerleri yüksek varyans göstermektedir.

4.2 Deneysel Analiz

V , W , $V - W$ (MID) ve V/W (MIQ) değerlerinin ortalamalarını ve varyanslarını deneysel olarak göz önüne serebilmek amacıyla bootstrap yönteminden yararlanılmıştır. İlk olarak iki farklı seçim kriteri kullanılarak tüm veri kümesi üzerinde MRMR algoritması uygulanmış ve her bir kriter için bir öznitelik sırası elde edilmiştir. Daha sonra bootstrap yöntemi kullanılarak tüm veri kümesinden n sayıda örnek içeren 50 adet alt veri kümesi elde edilmiş ve elde edilen her veri kümesi için daha önce elde edilmiş öznitelik sıralamaları kullanılarak V , W , $V - W$ (MID) ve V/W (MIQ) değerleri hesaplanmıştır. Musk (Versiyon 1) veri kümesi üzerinde yapılan bu deneyin sonuçları Şekil 4.2 ve Şekil 4.3’te verilmiştir.



Şekil 4.2 : Musk veri kümesi üzerinde her bir öz nitelik seçimi aşamasında V , W ve $V - W$ (MID) değerlerinin ortalama ve standart sapma değerleri.



Şekil 4.3 : Musk veri kümesi üzerinde her bir öz nitelik seçimi aşamasında V , W ve V/W (MIQ) değerlerinin ortalama ve standart sapma değerleri.

Şekil 4.2 ve Şekil 4.3'te görüleceği gibi MID tekniği ile hesaplanan entropi değerlerinin standart sapması MIQ tekniği ile hesaplanan entropi değerlerinin standart sapmasından daha küçüktür. Özellikle ilk 20 öz niteliğin seçiminde MIQ tekniğinin standart sapmasının büyük değerler alması, bu tekniğin kararlılık yönünden MID tekniğine göre daha az güvenilir olduğunu göstermektedir.

5. MRMR ALGORİTMASI İÇİN YENİ BİR SEÇİM KRİTERİ, MID_{α}

MID ve MIQ tekniklerinin teorik ve deneysel olarak karşılaştırıldığı Bölüm 4'ten anlaşılacağı gibi MID tekniği MIQ tekniğine göre daha kararlı (*stable*) sonuçlar vermektedir. MRMR algortmasının sınıf etiketleri vektörü ile maksimum ilişkili, daha önceden seçilmiş öznitelikler ile minimum artıklığa sahip özniteliği seçme aşamasında uyguladığı bu yöntem üzerinde uygulanabilecek küçük bir modifikasyon algoritmanın başarımlı ve kararlılık yönünden farklı davranışlar göstermesine yol açabilir. Şöyle ki, her öznitelik seçimi aşamasında seçilecek özniteliğin sınıf etiketleri vektörü ile ilişkisini ölçen V değerine bir α parametresini çarpan olarak kullanarak daha fazla ağırlık verilmesi, sınıf etiketleri vektörü ile yüksek ilişkili özniteliklerin seçim şansını artırarak algoritmanın başarımlı yönünden daha iyi sonuçlar vermesine yardımcı olabilir. Diğer yandan bu α parametresi W değerine bir çarpan olarak kullanılarak minimum artıklık koşuluna ağırlık verilebilir ve sınıf etiketleri vektörü ile düşük ilişkili özniteliklerin seçilme şansını arttırmamıza rağmen kendi aralarında minimum artıklığa sahip en iyi öznitelik kümesi elde edilebilir. Bu modifikasyon algoritmanın kararlılığını da etkileyerek bu çalışmanın motivasyonu olan daha kararlı öznitelik kümeleri seçilmesine yardımcı olabilir.

Algoritmanın başarımlı ve kararlılığını değiştirebilecek bir parametre olarak α değeri, MID tekniğinde yer alan $V - W$ hesabı yerine Eşitlik 5.1'de yer alan kriterin kullanılması ile algoritmaya dahil edilmiştir.

$$MID_{\alpha} = \alpha \cdot V - (1 - \alpha) \cdot W \quad (5.1)$$

Sunulan yeni seçim kriteri farklı α parametleri ($\alpha = [0, 0.25, 0.5, 0.75, 1]$) kullanılarak birçok veri kümesinde başarımlı ve kararlılık yönünden sınanmış ve sonuçlar Bölüm 8'de sunulmuştur.

6. KARARLILIK ÖLÇÜTÜ

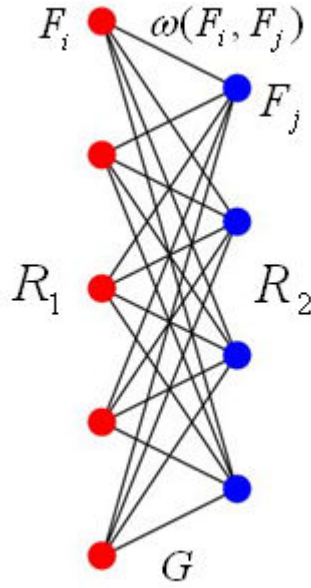
Geçmişte sunulan bir çok öznitelik seçme algoritması için başarı kriteri olarak sınıflandırıcı başarıyı baz alınmasına rağmen, kararlı sonuçlar üretme özelliği de öznitelik seçme algoritmaları için büyük önem teşkil eden bir unsurdur. Öznitelik seçme algoritmalarının kararlılığı aynı dağılımdan gelen farklı veri kümeleri üzerinde dahi aynı ya da yakın özniteliklerin seçilmesi olarak tanımlanabilir. Bir öznitelik seçme algoritması için bir öznitelik önemli ise aynı dağılımdan gelen farklı veri kümeleri içinde aynı durum geçerli olmalıdır. Özellikle küçük veri kümelerinde öznitelik seçme yöntemlerinin kararlılığı bir adım öne çıkmaktadır.

Öznitelik seçme algoritmasının kararlılığını ölçmek amacıyla algoritmanın aynı dağılımdan gelen farklı iki veri kümesi üzerinde uygulanmasından elde edilen iki öznitelik kümesi arasındaki tutarlılığın seviyesini belirleyebilecek bir ölçüte ihtiyaç vardır. Bu bölümde [13] tarafından sunulan yöntemle benzer bir kararlılık ölçütünün tanımı yapılacaktır.

Varsayalım $R_1 = \{F_i\}_{i=1}^{|R_1|}$ ve $R_2 = \{F_j\}_{j=1}^{|R_2|}$ öznitelik seçme algoritmalarının sonucu olan iki öznitelik kümesi olsun ve her bir F_i ve F_j tek bir özniteliği temsil etsin. R_1 ve R_2 arasındaki kararlılığı ölçmek amacıyla [13] R_1 ve R_2 'nin ağırlıklı tam iki parçalı çizge, $G = (V, E)$, olarak modellenmesini önermekte, $V = R_1 \cup R_2$, $E = \{(F_i, F_j) | F_i \in R_1, F_j \in R_2\}$, ve öznitelikler arasındaki her bir bağlantının, (F_i, F_j) , bir ağırlık ölçütü, $\omega(F_i, F_j)$ ile ilişkilendirilmesi gerektiğini belirtmektedir. Bu çalışmada, çizgedeki iki nokta ya da diğer bir deyişle iki öznitelik, F_i ve F_j , arasındaki ağırlık, $\omega(F_i, F_j)$, bu öznitelik çifti arasındaki simetrik belirsizlik (*symmetrical uncertainty*) değerli olarak atanmıştır. Entropi tabanlı doğrusal olmayan bu korelasyon FCBF algoritmasının açıklandığı Bölüm 3.2, Eşitlik 3.9'da açıkça tanımlanmaktadır. Hatırlatma amacıyla Eşitlik 6.1'de tekrar edilmiştir.

$$SU_{i,j} = 2 \left[\frac{IG(F_i; F_j)}{H(F_i) + H(F_j)} \right] \quad (6.1)$$

Daha önce tanımlandığı gibi F_i ve F_j iki ayrık rastlantı değişkeni olarak ele alınmakta ve sırasıyla öznitelik i ve öznitelik j 'yi temsil etmektedirler.



Şekil 6.1 : İki öznitelik kümesi R_1 ve R_2 'nin tam iki parçalı çizge olarak modeli.

Şekil 6.1'de gösterilen modelde $\omega(F_i, F_j)$, $SU_{i,j}$ 'ye eşittir ve R_1 ve R_2 arasındaki genel benzerlik (*kararlılık seviyesi*) Eşitlik 6.2'de verilen formül ile hesaplanmaktadır.

$$Sim^M(R_1, R_2) = \frac{\sum_{(F_i, F_j) \in M} \omega(F_i, F_j)}{|M|} \quad (6.2)$$

Burada M , G çizgesindeki iki parçaya denk gelen iki öznitelik seti, R_1 ve R_2 , arasındaki maksimum ağırlıklı eşleşmeyi, $|M|$ ise bu eşleşmedeki öznitelik çifti sayısını temsil etmektedir. Maksimum ağırlıklı iki parçalı çizge eşleştirme problemi birden farklı sayıda algoritma ile çözülebilir. Bu çalışmada optimum sonuca ulaşabilmek için Hungarian algoritması [22] kullanılmıştır. Algoritma Tennessee Üniversitesi'nden Alexander Melin tarafından koda dökülmüş olup Matlab Central [23] internet sitesinden alınmıştır.

İki öznitelik kümesi arasındaki benzerliği ölçmek amacıyla böyle bir yöntemin kullanılmasının nedeni özniteliklerin indeks numaralarından ziyade iki öznitelik arasındaki simetrik belirsizlik değeri kullanılarak özniteliklerin içerdiği değerlerinde karşılaştırma işlemine dahil edilmesinin daha doğru olduğu düşünülmektedir. Bu

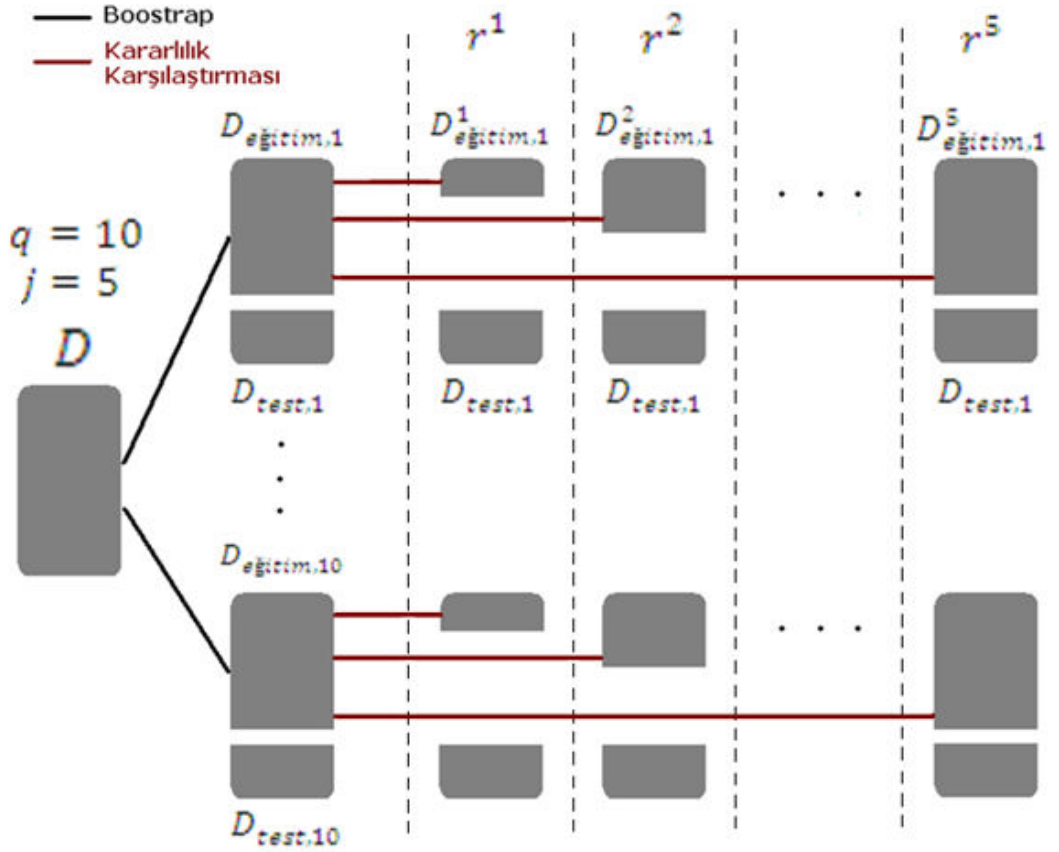
yöntem iki öznitelik kümesinin çok sayıda farklı indekse sahip öznitelik içermesine rağmen bu özniteliklerin aralarında yüksek derecede bağlılık, korelasyon olduğu durumlarda daha mantıklı sonuçlar vermektedir.

Bu çalışmada tanımlanan kararlılık ölçütü [13]'te açıklanan yöntemden iki açıdan farklıdır. İlk olarak ilgili makalede tek tek özniteliklerden oluşan iki öznitelik kümesi arasındaki kararlılıktan farklı olarak öznitelik gruplarından oluşan öznitelik kümeleri arasındaki kararlılık ölçülmektedir. Dahası, modellenen çizgede öznitelik grupları arasındaki ağırlıklar simetrik belirsizlik değerinden farklı olarak her bir grubun merkezinde yer alan ya da grubu en iyi şekilde tanımlayan öznitelikler arasındaki korelasyon katsayısına bakılarak hesaplanmıştır. Bu çalışmadaki yöntem [13]'te açıklanan metodun özel bir hali olarak algılanabilir.

7. KARARLILIK VE BAŞARIM KARŞILAŞTIRILMASI İÇİN DENEYSEL METODOLOJİ

Bir öznitelik seçme algoritmasının kararlılığı, aynı dağılımdan gelen farklı veri kümeleri üzerinde elde edilen sonuçların birbirlerine göre tutarlı olması olarak ifade edilmiştir. Bu çalışmada kullanılan veri kümeleri için aynı dağılımdan gelen yeni veriler toplama şansına sahip olunmaması nedeniyle ortaya konulan metodoloji, aynı dağılıma ait farklı veri kümeleri elde etmek amacıyla bootstrap yöntemini kullanmaktadır.

n sayıda örnek içeren D veri kümesi için belli bir öznitelik seçme algoritmasının başarımını ve kararlılığını ölçmek amacıyla D veri kümesinden bootstrap yöntemi kullanılarak n sayıda örnek içeren q eğitim alt kümesi, $D_{\text{eğitim},i}$, oluşturulmuştur ($i = 1, 2, \dots, q$). Her bootstrap yöntemi sonrasında seçilmeyen örnekler ise test veri kümesi, $D_{\text{test},i}$, olarak kullanılacaktır. Ana öznitelik sıralamasının elde edilmesi amacıyla ilk olarak eğitim kümesi $D_{\text{eğitim},i}$ üzerinde öznitelik seçme algoritması uygulanmıştır ve R_i öznitelik kümesi elde edilmiştir. Algoritmanın kararlılığı, düşük sayıda örnek içeren veri kümelerinde dahi tutarlı sonuçlar vermesi olarak tanımlandığından ikinci etapta $D_{\text{eğitim},i}$ eğitim kümesinden $r \cdot |D_{\text{eğitim},i}|$ ($r \cdot n$) sayıda örnek içeren yeni eğitim veri kümesi, $D_{\text{eğitim},i}^j$, oluşturulmuştur ($r = [r^1, r^2, \dots, r^j]$). Daha sonra değişik oranlarda örnek sayısına sahip bu küçük veri kümeleri üzerinde öznitelik seçme algoritması uygulanmış ve her r^j oranı için öznitelik kümesi R_i^j elde edilmiştir. Örnek olması ve bu çalışmada gerçekleştirilen deneylerde kullanılan değerlere tekabül etmesi nedeniyle Şekil 7.1’de q için 10 ve j için 5 değerleri için sunulan modelin grafiksel anlatımı gösterilmiştir.



Şekil 7.1 : Öznitelik seçme algoritmalarının kararlılık ve başarımlarından karşılaştırılmasında kullanılacak deneysel metodoloji modeli.

Sonuç olarak r^j oranı için q alt küme kullanılarak gerçekleştirilen deneyde algoritmanın kararlılığı Eşitlik 7.1’de verilmiştir.

$$\frac{1}{q} \sum_{i=1}^q Sim^M(R_i, R_i^j) \quad (7.1)$$

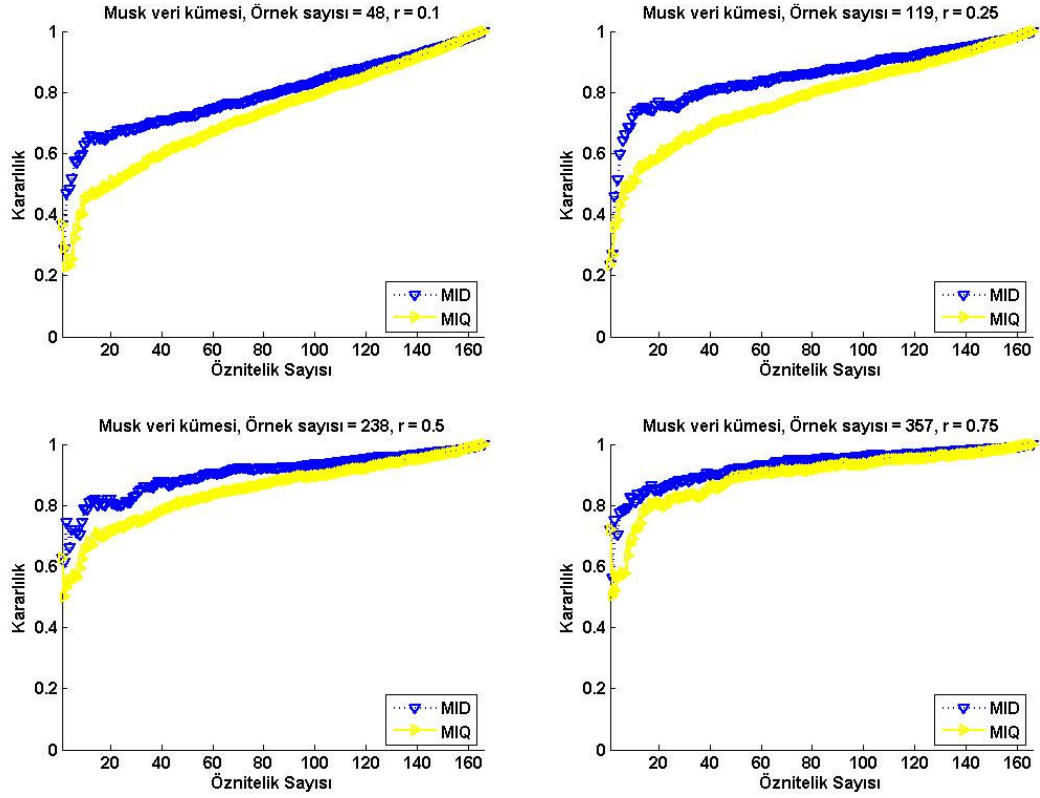
Algoritmanın başarımlarları ise eğitim kümelerinden elde edilen öznitelik kümelerinin bu kümelere karşılık gelen test kümeleri üzerinde k-en-yakın-komşu (*k-nearest neighbor*) veya svm (*support vector machine*) sınıflandırıcıları kullanılarak elde edilmiştir. Kararlılık ölçme yönteminin örnek alındığı makalede [13] sunulan yöntemde bir defa bootstrap yöntemi kullanılmış ve tüm veri kümesi ile bootstrap yöntemi kullanılarak elde edilen veri kümesi arasındaki kararlılık hesaplanmıştır. Bu yöntem bu çalışmada sunulan yöntemde r için 0.632 kullanılmasına denktir [24]. Ancak tek bootstrap kullanılması ve daha küçük oranda örnek içeren veri kümeleri üzerinde deney gerçekleştirilmemesi nedeniyle yetersiz görülmüştür.

8. DENEY SONUÇLARI

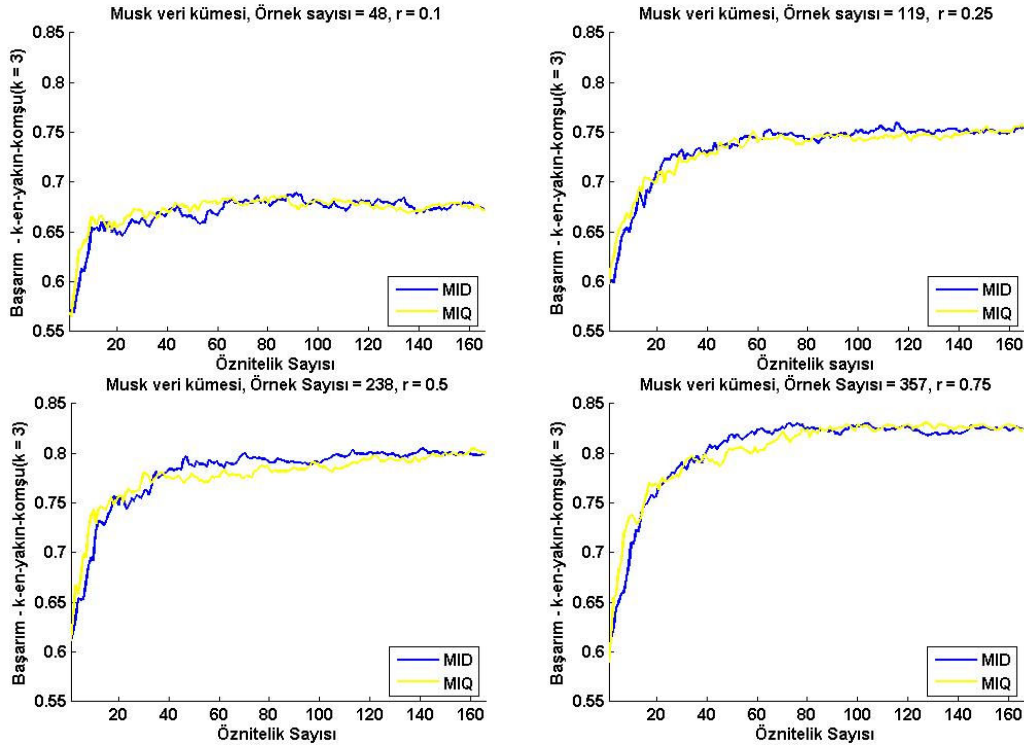
Bu bölümde Bölüm 2’de tanıtılan 8 veri kümesi üzerinde gerçekleştirilen kararlılık ve başarımlar deneylerinin sonuçları sunulacaktır.

8.1 MID ve MIQ Seçim Kriterlerinin Karşılaştırılması

Bu bölümde MRMR algoritmasının iki seçim kriteri olan MID ve MIQ tekniklerinin kararlılık ve başarımlar yönünden karşılaştırılmasının sonuçları verilecektir.



Şekil 8.1 : MRMR, MID ve MIQ öznitelik seçim kriterlerinin Musk veri kümesinde kararlılık yönünden karşılaştırılması.



Şekil 8.2 : MRMR, MID ve MIQ öznitelik seçim kriterlerinin Musk veri kümesi üzerinde başarımlarından karşılaştırılması.

Şekil 8.1 ve Şekil 8.2’de MRMR algoritmasının iki farklı seçim kriteri olan MID ve MIQ tekniklerinin kararlılık ve başarımlarından karşılaştırılması verilmiştir. Bölüm 7’de sunulan deneysel metodoloji kullanılarak, Musk veri kümesi üzerinde farklı r ($r = [0.1, 0.25, 0.5, 0.75]$) değerleri için gerçekleştirilen deney sonuçlarında bu iki tekniğin sınıflandırıcı başarımlarından yakını sonuçlar vermesine rağmen tüm r oranları için MID tekniğinin MIQ’ya göre daha kararlı sonuçlar verdiği gözlenmiştir. MID ve MIQ yöntemlerinde hesaplanan entropi dağılımlarının elde edildiği ve entropi hesaplamalarındaki ortalama ve standart sapma değerlerinin sunulduğu bölüm 4’te alınan sonuçları destekleyen sonuçlar elde edilmiş ve MID tekniğinin MIQ tekniğine göre daha kararlı öznitelik kümeleri ile sonuçlandığı tezi doğrulanmıştır.

8.2 MID_{α} Seçim Kriterinin Kararlılık ve Başarım Sonuçları

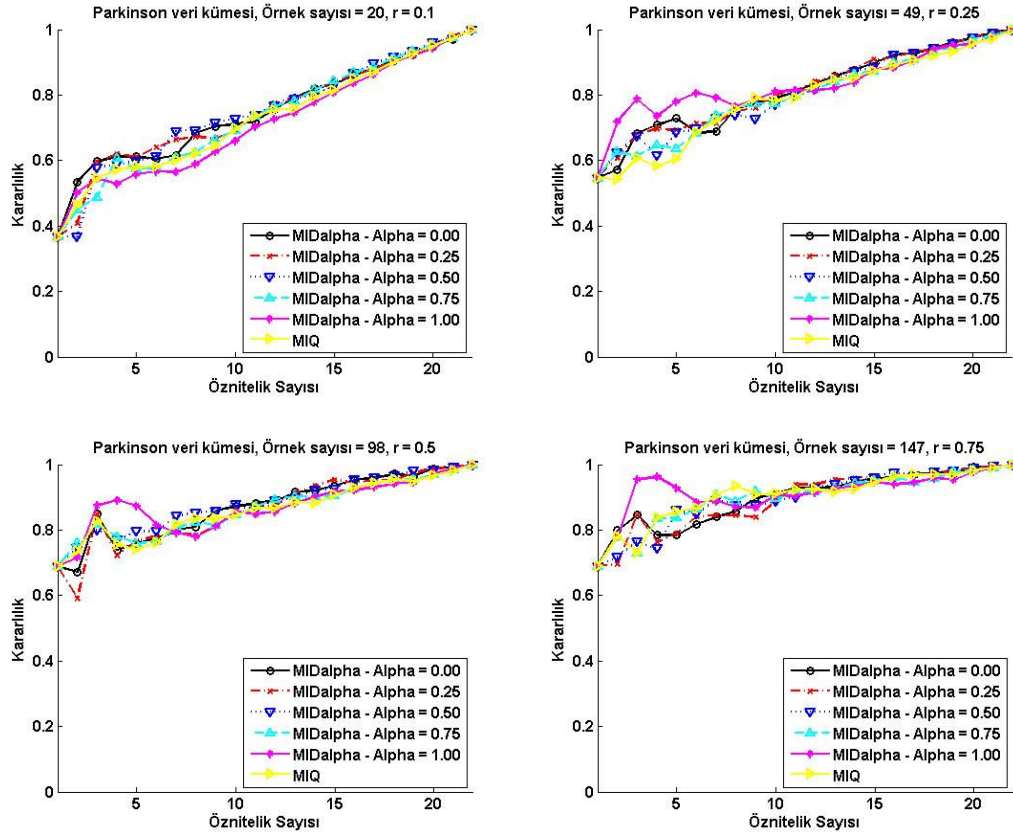
Bölüm 7’de başarımlar ve kararlılık karşılaştırılması için geliştirilen deneysel metodoloji modeli sunulmuştu. Bu model kısaca bootstrap edilerek oluşturulan eğitim veri kümelerinden farklı r oranlarında örnek içeren yeni eğitim veri kümeleri oluşturularak bu küme ile oluşturulan yeni veri kümeleri arasındaki kararlılık ve

bootstrap işleminden sonra elde edilen test veri kümelerinde başarımlar ölçülmesi olarak tanımlanmıştır. Bu bölümde, önerilen yeni seçim kriteri MID_{α} 'nın farklı α değerleri için başarımlar ve kararlılık sonuçları sergilenecektir ($\alpha = [0, 0.25, 0.5, 0.75, 1]$). q değeri için 10, j değeri için 5 alınmış ve r^j değerleri 0.1, 0.25, 0.5, 0.75 ve 1 olarak seçilmiştir. Bölüm 2'de kısa açıklamaları verilmiş 8 veri kümesi üzerinde gerçekleştirilen deneylerin sonuçları şekillerle açıklanacak ve yorumlanacaktır. Veri kümelerinin ve r değerlerinin fazlalığı nedeniyle deney sonuçlarının bir kısmı Ekler bölümünde verilmiştir. $r^j = 1$, oranı deneylerin doğruluğunu test etmek amacıyla kullanıldığı için grafiklerde bu oran değeri için veriler gösterilmeyecektir.

Deneylerde öznitelik seçme yöntemlerinin sınıflandırıcı başarımlarını ölçmek amacıyla PRtools [25] kütüphanesinden k -en-yakın-komşu (KNN , *k-nearest neighbor*) algoritması ($k=3$) ve LibSVM [26] kütüphanesinden svm (*support vector machine*) sınıflandırıcısı kullanılmıştır. Svm sınıflandırıcıları için kullanılan parametreler varsayılan (*default*) parametrelerdir.

Grafiklerde betimlenen sonuçları daha anlaşılır bir şekilde ifade etmesi amacıyla test edilen farklı seçim kriterleri her bir öznitelik sayısı için kararlılık ve başarımlar yönünden sıralanarak, ilgili çizelgelerde bu sıra değerlerinin ortalamaları ve standart sapmaları verilmiştir. Hangi r oranı için hangi α değerinin daha başarılı sonuçlar verdiği bu çizelgeler yardımıyla daha kolay saptanabilmektedir.

Şekil 8.3 ve Çizelge 8.1'de Parkinson veri kümesi için elde edilen kararlılık sonuçları bulunmaktadır. Şekil 8.3'te farklı α değerleri için kararlılık yönünden kayda değer bir fark elde edilememiştir ancak az sayıda öz niteliğin seçildiği durumlarda (0-10 arası) $\alpha = 1$ değeri $r = 0.25, 0.5, 0.75$ değerleri için diğerlerine nazaran yüksek değerler göstermektedir. Diğer taraftan Çizelge 8.1 algoritmanın genel olarak $\alpha \leq 0.5$ değerleri için diğer α değerlerine nazaran daha kararlı sonuçlar verdiğini ortaya koymaktadır.



Şekil 8.3 : MRMR, MID_{α} seçim kriterinin farklı α ve r değerleri için Parkinson veri kümesi üzerinde kararlılık sonuçları.

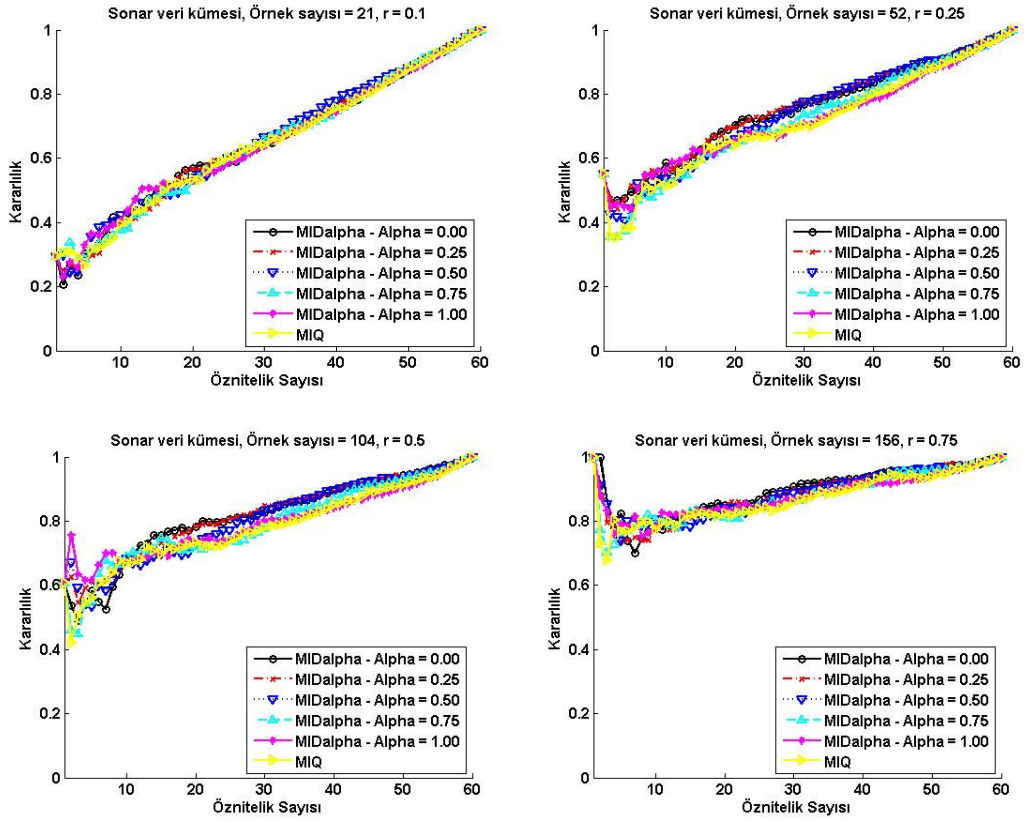
Çizelge 8.1 : MRMR, MID_{α} seçim kriterinin farklı α ve r değerleri için Parkinson veri kümesi üzerindeki kararlılık sonuçlarının ortalama sıralanma değerleri ve standart sapmaları.

r oranı	$\alpha = 0$	$\alpha = 0.25$	$\alpha = 0.5$	$\alpha = 0.75$	$\alpha = 1$
$r = 0.1$	2.5 ± 1.2	2.3 ± 1.2	2.1 ± 1.3	2.9 ± 1.1	4.3 ± 1.4
$r = 0.25$	2.5 ± 1.1	2.2 ± 1.1	3.0 ± 1.4	3.6 ± 1.3	2.7 ± 1.9
$r = 0.5$	2.3 ± 1.2	2.9 ± 1.3	2.0 ± 1.1	3.2 ± 1.4	3.5 ± 1.6
$r = 0.75$	2.7 ± 1.3	2.6 ± 1.6	2.5 ± 1.2	2.9 ± 1.4	3.3 ± 1.6

Parkinson veri kümesinde gerçekleştirilen sınıflandırma başarımları testlerinde herhangi bir α değerinin diğerlerine nazaran üstünlük gösterdiği söylenemez. Bu deneylerle ilgili sonuçlar Ek B.1 bölümünde Şekil B.1, Çizelge B.1, Şekil B.2 ve Çizelge B.2’de görülebilir.

Şekil 8.4 ve Çizelge 8.2’de Sonar veri kümesi üzerinde gerçekleştirilen kararlılık deneylerinin sonuçları sergilenmiştir. Bu veri kümesinde kararlılık yönünden herhangi bir α değerinin üstünlük gösterdiği söylenemez. Sonuçlar birbirine çok yakın olmasına rağmen $\alpha \leq 0.5$ değerleri için ortalama sıralanma değerleri daha

küçüktür diğer bir deyişle bu α değerleri diğerlerine nazaran daha kararlı sonuçlar göstermektedir.

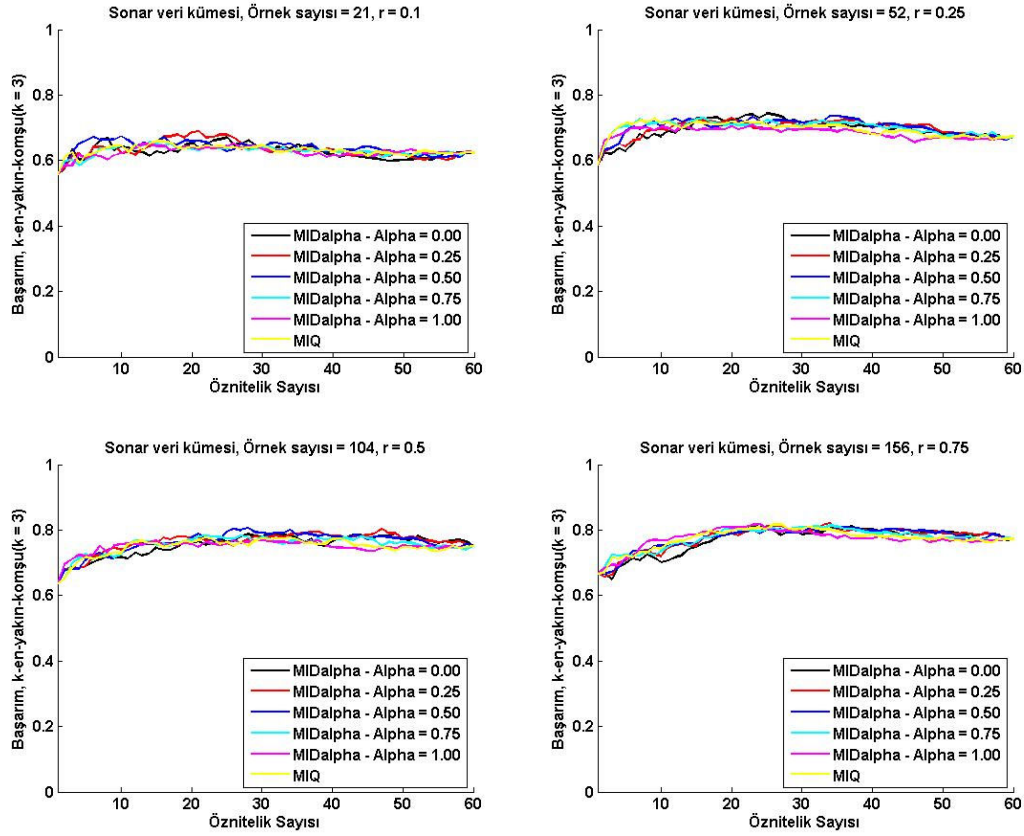


Şekil 8.4 : MRMR, MID_{α} seçim kriterinin farklı α ve r değerleri için Sonar veri kümesi üzerinde kararlılık sonuçları.

Çizelge 8.2 : MRMR, MID_{α} seçim kriterinin farklı α ve r değerleri için Sonar veri kümesi üzerindeki kararlılık sonuçlarının ortalama sıralanma değerleri ve standart sapmaları.

r oranı	$\alpha = 0$	$\alpha = 0.25$	$\alpha = 0.5$	$\alpha = 0.75$	$\alpha = 1$
$r = 0.1$	3.1 ± 1.3	2.6 ± 1.3	2.2 ± 1.3	3.2 ± 1.3	3.7 ± 1.4
$r = 0.25$	2.2 ± 0.8	1.7 ± 0.8	2.6 ± 1.2	4.2 ± 0.8	4.0 ± 1.4
$r = 0.5$	2.2 ± 1.2	2.0 ± 0.9	2.9 ± 1.2	3.7 ± 1.2	3.9 ± 1.4
$r = 0.75$	1.8 ± 1.2	2.6 ± 1.2	2.8 ± 1.1	3.8 ± 1.3	3.7 ± 1.4

Sonar veri kümesi için gerçekleştirilen başarımlar deneylerinde görülmüştür ki $\alpha = 0.25$ değeri diğer α değerlerine göre daha iyi sonuçlar vermektedir ancak fark çok küçüktür. Bahsedilmesi gereken diğer bir nokta ise bu veri kümesinde k-en-yakın-komşu algoritmasının başarımlar oranlarının svm algoritmasına oranla çok daha yüksek değerlere ulaştığıdır. Şekil 8.5 ve Çizelge 8.3'te k-en-yakın-komşu algoritması için elde edilen sonuçlar görülebilir. Sınıflandırıcı olarak svm algoritmasının kullanıldığı sonuçlar Ek B.3 bölümünde Şekil B.3 ve Çizelge B.3'te sergilenmiştir.

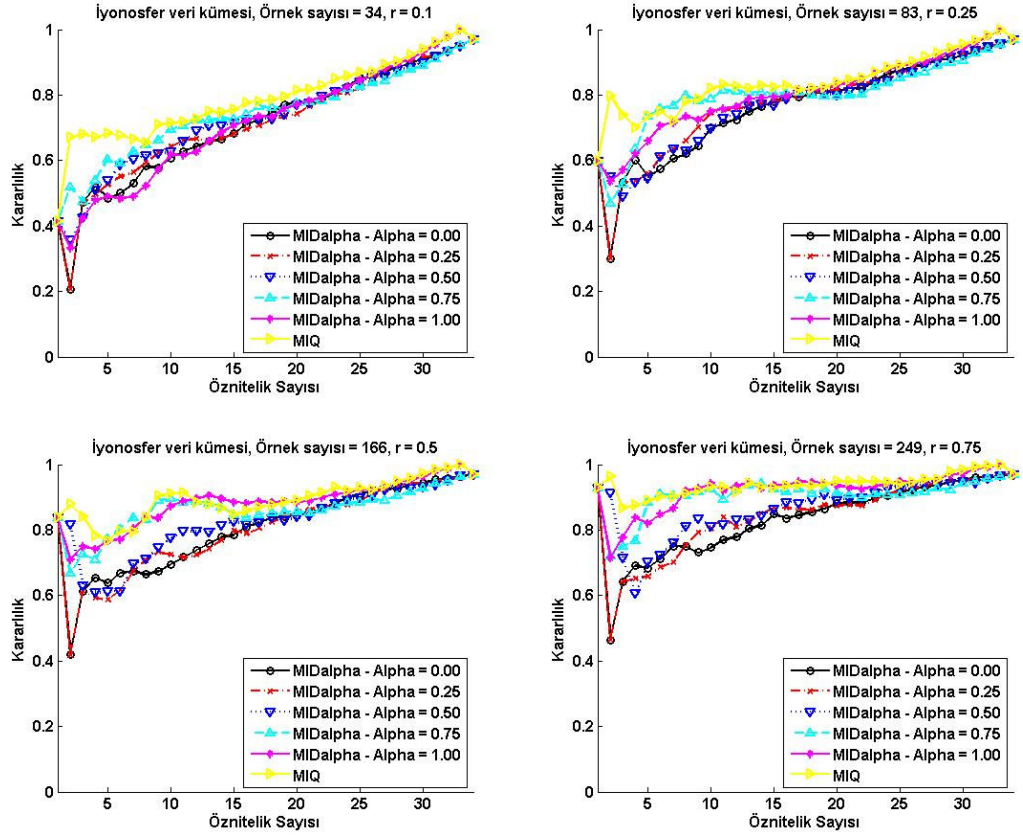


Şekil 8.5 : MRMR, MID_{α} seçim kriterinin farklı α ve r değerleri için Sonar veri kümesi üzerinde başarımlar (k-en-yakın-komşu) sonuçları.

Çizelge 8.3 : MRMR, MID_{α} seçim kriterinin farklı α ve r değerleri için Sonar veri kümesi üzerindeki başarımlar (k-en-yakın-komşu) sonuçlarının ortalama sıralanma değerleri ve standart sapmaları.

r oranı	$\alpha = 0$	$\alpha = 0.25$	$\alpha = 0.5$	$\alpha = 0.75$	$\alpha = 1$
$r = 0.1$	3.0 ± 2.0	2.7 ± 1.2	2.3 ± 1.3	3.1 ± 1.1	3.6 ± 1.0
$r = 0.25$	3.1 ± 1.9	1.8 ± 0.7	2.5 ± 1.2	3.7 ± 1.0	3.7 ± 1.0
$r = 0.5$	3.4 ± 1.7	1.8 ± 1.1	2.7 ± 1.4	3.3 ± 1.1	3.6 ± 0.9
$r = 0.75$	4.9 ± 0.5	1.9 ± 1.1	2.0 ± 0.9	3.0 ± 1.0	2.9 ± 1.1

İyonosfer veri kümesi için elde edilen kararlılık sonuçları Şekil 8.6 ve Çizelge 8.4'te gösterilmiştir. Kararlılık grafiklerinde ilk göze çarpan nokta diğer veri kümelerinin aksine bu veri kümesinde MIQ yönteminin MID_{α} tekniğine nazaran daha yüksek kararlılık değerleri gösterdiğiidir. Bunun yanında $\alpha = 0.75, 1$ değerleride diğer α değerlerine göre yüksek kararlılık göstermişlerdir.



Şekil 8.6 : MRMR, MID_{α} seçim kriterinin farklı α ve r değerleri için İyonosfer veri kümesi üzerinde kararlılık sonuçları.

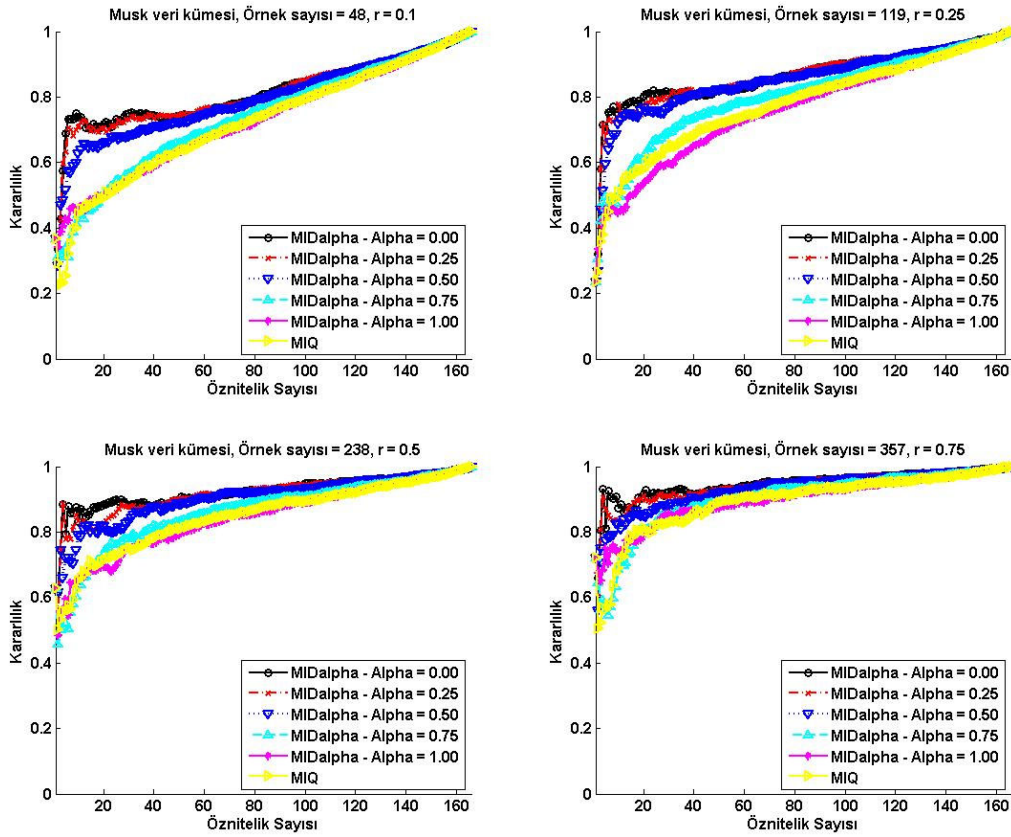
Çizelge 8.4 : MRMR, MID_{α} seçim kriterinin farklı α ve r değerleri için İyonosfer veri kümesi üzerindeki kararlılık sonuçlarının ortalama sıralanma değerleri ve standart sapmaları.

r oranı	$\alpha = 0$	$\alpha = 0.25$	$\alpha = 0.5$	$\alpha = 0.75$	$\alpha = 1$
$r = 0.1$	3.1 ± 1.2	3.4 ± 1.4	2.8 ± 1.1	2.3 ± 1.7	2.7 ± 1.5
$r = 0.25$	3.5 ± 1.2	3.0 ± 0.9	3.4 ± 1.1	3.0 ± 1.9	1.4 ± 0.6
$r = 0.5$	3.3 ± 1.1	3.7 ± 1.3	3.3 ± 1.0	2.9 ± 1.5	1.2 ± 0.4
$r = 0.75$	4.0 ± 1.1	3.5 ± 1.1	3.0 ± 1.0	2.6 ± 1.4	1.2 ± 0.5

İyonosfer veri kümesi üzerinde elde edilen başarımlar tüm yöntemler için birbirine çok yakın sonuçlar vermiştir. Sıralanma değerlerine ve standart sapmalarına bakılarak denilebilir ki $\alpha = 0.75$ ve 1 değerleri diğerlerine nazaran daha üst sıralardadır ancak aradaki fark çok küçüktür. Bu veri kümesi için elde edilen başarımlar ve sıralanma değerleri Ek B.4 ve Ek B.5 bölümlerinde Şekil B.4, Çizelge B.4, Şekil B.5 ve Çizelge B.5'te görülebilir.

Musk veri kümesi için elde edilen sonuçlar Şekil 8.7 ve Çizelge 8.5'te gösterilmektedir. Diğer veri kümelerinin aksine bu veri kümesinde $\alpha = 0, 0.25$ değerleri düşük sayıdaki öznitelik seçimi durumlarında diğerlerine göre yüksek

değerler almışlardır. $\alpha = 0.75, 1$ değerleri ise kötü kararlılık sonuçları vermişlerdir. Başarım yönünden ise diğer veri kümelerinde olduğu gibi gözle görülür bir fark bulunmamaktadır. Bu veri kümesi için çok uzun zaman alması nedeniyle svm sonuçları elde edilememiştir. K-en-yakın-komşu algoritması için başarım sonuçları Ek B.6 bölümünde Şekil B.6 ve Çizelge B.6'da görülebilir ve ilgili çizelge göstermektedir ki $\alpha = 0.25$ değeri diğer α değerlerine nazaran daha yüksek başarım elde etmiştir. Ancak grafiklerden anlaşılacağı gibi başarım değerleri arasında büyük farklar bulunmamaktadır.

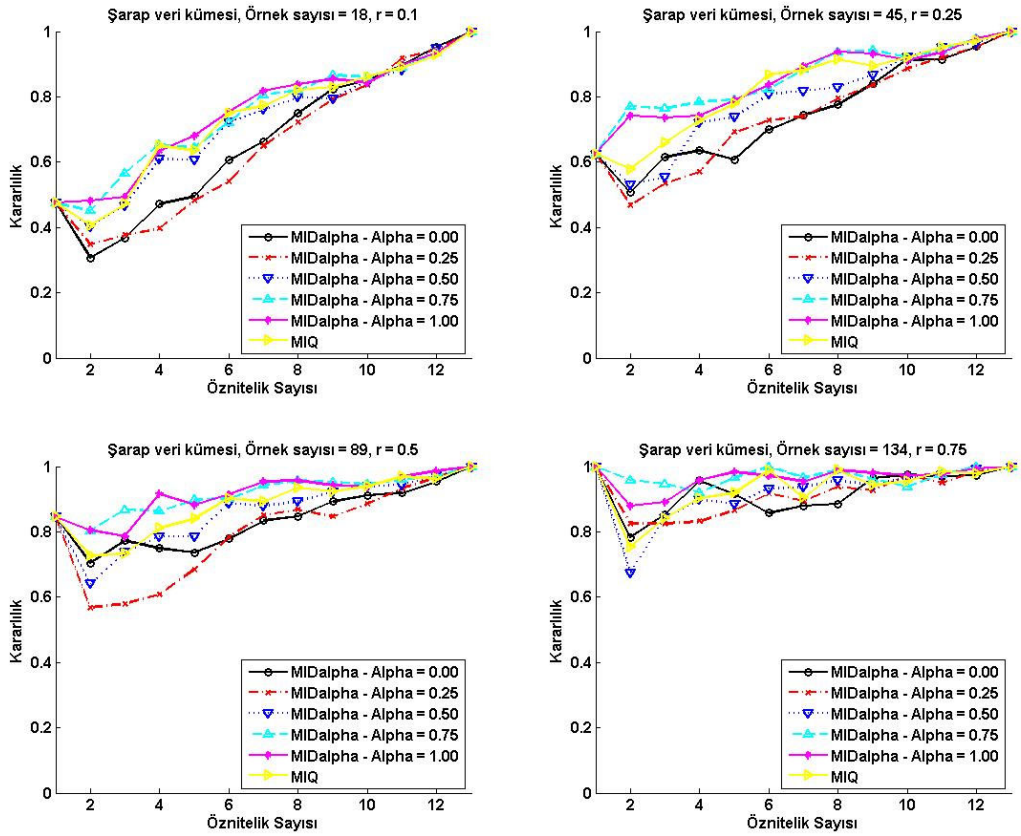


Şekil 8.7 : MRMR, MID_α seçim kriterinin farklı α ve r değerleri için Musk veri kümesi üzerinde kararlılık sonuçları.

Çizelge 8.5 : MRMR, MID_α seçim kriterinin farklı α ve r değerleri için Musk veri kümesi üzerindeki kararlılık sonuçlarının ortalama sıralanma değerleri ve standart sapmaları.

r oranı	$\alpha = 0$	$\alpha = 0.25$	$\alpha = 0.5$	$\alpha = 0.75$	$\alpha = 1$
$r = 0.1$	1.7 ± 0.9	1.8 ± 0.8	2.7 ± 0.8	3.9 ± 0.7	4.7 ± 0.8
$r = 0.25$	1.9 ± 0.8	1.4 ± 0.6	2.7 ± 0.6	4.0 ± 0.4	4.8 ± 0.8
$r = 0.5$	1.7 ± 0.7	1.5 ± 0.6	2.8 ± 0.6	4.1 ± 0.5	4.8 ± 0.7
$r = 0.75$	1.4 ± 0.7	1.8 ± 0.6	2.9 ± 0.6	4.1 ± 0.6	4.7 ± 0.7

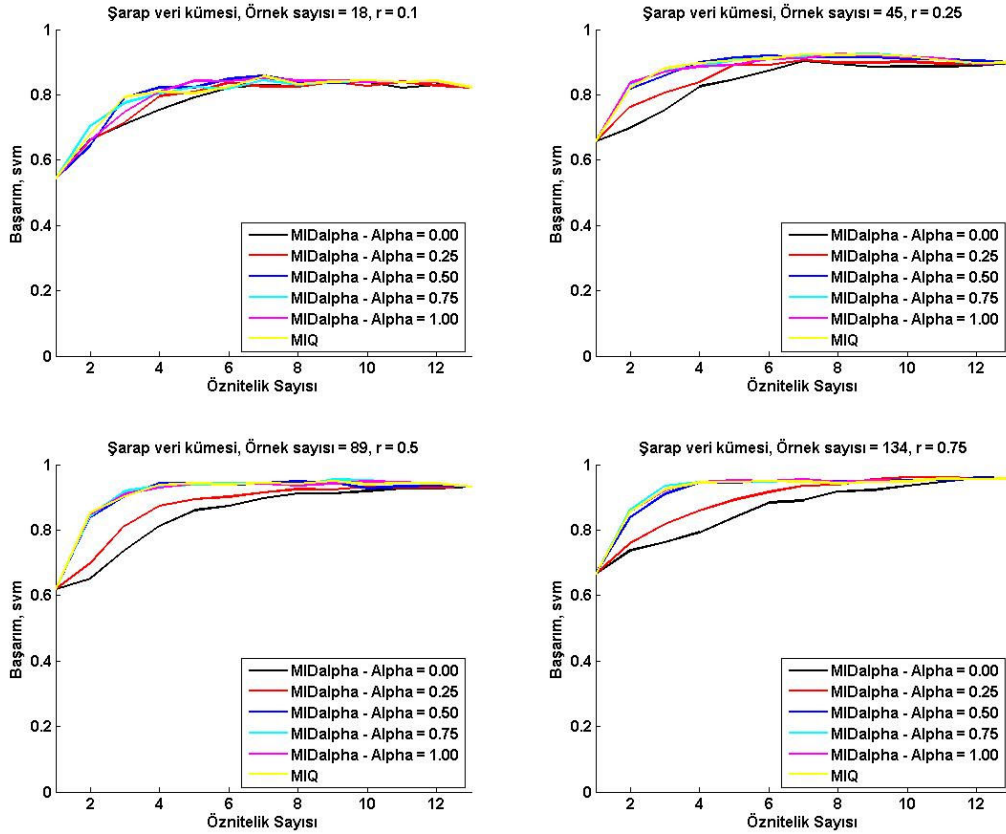
Şekil 8.8 ve Çizelge 8.6 Şarap veri kümesi için gerçekleştirilen deneyler sonucunda elde edilen kararlılık değerlerini göstermektedir. Bu veri kümesi için $\alpha = 0.75, 1$ değerleri kararlılık yönünden diğerlerine nazaran çok daha iyi sonuçlar vermişlerdir. Tüm r değerleri için bu α değerlerinin üstünlüğünden söz edilebilir. Başarım sonuçları ise Şekil 8.9 ve Çizelge 8.7’de sergilenmiştir. Svm algoritmasının sonuçları k-en-yakın-komşu algoritmasının sonuçlarına göre çok daha başarılıdır ve bu nedenle burada sunulmuştur. K-en-yakın-komşu algoritması için sonuçlar Ek B.7 bölümünde Şekil B.7 ve Çizelge B.7’de bulunabilir. Sınıflandırma başarımı için de yüksek kararlılık değerleri veren α değerlerinin başarısından söz edilebilir. $\alpha \geq 0.5$ değerleri diğer α değerlerine göre daha yüksek başarımlar göstermişlerdir.



Şekil 8.8 : MRMR, MID_{α} seçim kriterinin farklı α ve r değerleri için Şarap veri kümesi üzerinde kararlılık sonuçları.

Çizelge 8.6 : MRMR, MID_α seçim kriterinin farklı α ve r değerleri için Şarap veri kümesi üzerindeki kararlılık sonuçlarının ortalama sıralanma değerleri ve standart sapmaları.

r oranı	$\alpha = 0$	$\alpha = 0.25$	$\alpha = 0.5$	$\alpha = 0.75$	$\alpha = 1$
$r = 0.1$	3.2 ± 1.4	3.8 ± 1.6	2.8 ± 1.0	2.0 ± 1.2	1.8 ± 1.0
$r = 0.25$	3.7 ± 1.3	4.1 ± 1.4	2.5 ± 1.0	1.4 ± 0.5	1.8 ± 0.9
$r = 0.5$	3.7 ± 1.3	3.9 ± 1.4	2.6 ± 0.9	1.5 ± 0.6	1.4 ± 0.6
$r = 0.75$	3.1 ± 1.5	3.6 ± 1.4	3.0 ± 1.3	1.8 ± 1.2	1.7 ± 0.7



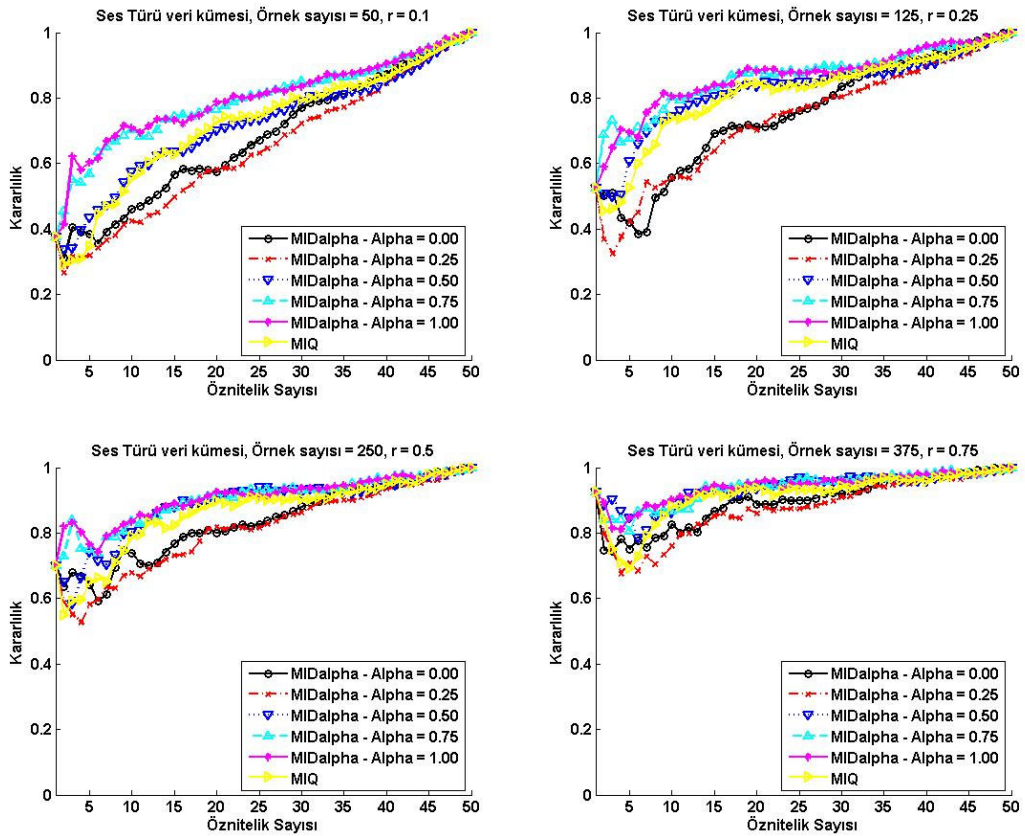
Şekil 8.9 : MRMR, MID_α seçim kriterinin farklı α ve r değerleri için Şarap veri kümesi üzerinde başarımlı (svm) sonuçları.

Çizelge 8.7 : MRMR, MID_α seçim kriterinin farklı α ve r değerleri için Şarap veri kümesi üzerindeki başarımlı (svm) sonuçlarının ortalama sıralanma değerleri ve standart sapmaları.

r oranı	$\alpha = 0$	$\alpha = 0.25$	$\alpha = 0.5$	$\alpha = 0.75$	$\alpha = 1$
$r = 0.1$	2.2 ± 1.5	3.9 ± 1.6	3.2 ± 1.2	2.5 ± 1.0	2.2 ± 0.7
$r = 0.25$	2.5 ± 1.4	4.3 ± 1.3	3.1 ± 1.3	1.8 ± 0.7	2.2 ± 0.7
$r = 0.5$	3.3 ± 1.0	4.2 ± 1.4	2.9 ± 1.1	1.5 ± 0.6	1.4 ± 0.6
$r = 0.75$	4.4 ± 1.1	3.2 ± 1.2	2.6 ± 1.1	1.5 ± 0.8	1.6 ± 0.6

Ses Türü veri kümesi de Şarap veri kümesine benzer sonuçlar vermiştir. Şekil 8.10 ve Çizelge 8.8 bu veri kümesinde gerçekleştirilen kararlılık deney sonuçlarını

göstermektedir. $\alpha = 0.75, 1$ değerlerinin başarısı ilk etapta göze çarpmaktadır. Bu değerler tüm r değerleri için diğerlerine nazaran çok daha yüksek kararlılık seviyelerine ulaşmışlardır. Başarım yönünden de durum farklı değildir. Başarım değerlerinin genel olarak çok düşük olmasına rağmen 0.75 ve 1 α değerleri diğerlerine nazaran daha iyi sonuçlar vermektedir. Ses Türü veri kümesi için çok uzun zama alması nedeniyle svm başarım sonuçları elde edilememiştir. K-en-yakın-komşu algoritması kullanılarak elde edile başarım sonuçları Ek B.8 bölümünde Şekil B.8 ve Çizelge B.8’de verilmiştir.

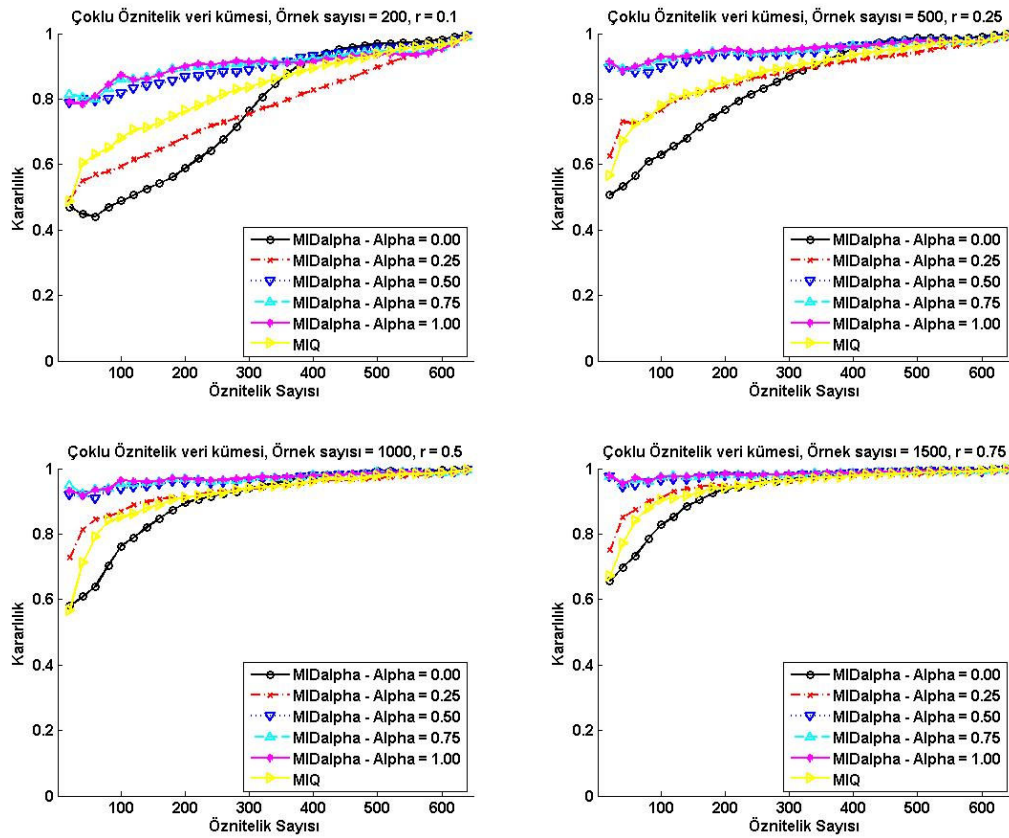


Şekil 8.10 : MRMR, MID_α seçim kriterinin farklı α ve r değerleri için Ses Türü veri kümesi üzerinde kararlılık sonuçları.

Çizelge 8.8 : MRMR, MID_α seçim kriterinin farklı α ve r değerleri için Ses Türü veri kümesi üzerindeki kararlılık sonuçlarının ortalama sıralanma değerleri ve standart sapmaları.

r oranı	$\alpha = 0$	$\alpha = 0.25$	$\alpha = 0.5$	$\alpha = 0.75$	$\alpha = 1$
$r = 0.1$	3.5 ± 0.8	4.7 ± 0.9	3.3 ± 0.8	1.8 ± 0.6	1.3 ± 0.5
$r = 0.25$	3.7 ± 1.1	4.5 ± 0.9	3.2 ± 0.7	1.8 ± 0.8	1.4 ± 0.5
$r = 0.5$	3.5 ± 1.1	4.7 ± 0.8	2.7 ± 1.0	1.9 ± 0.8	1.8 ± 0.8
$r = 0.75$	3.5 ± 1.2	4.7 ± 0.9	2.6 ± 1.1	1.9 ± 0.9	2.0 ± 0.9

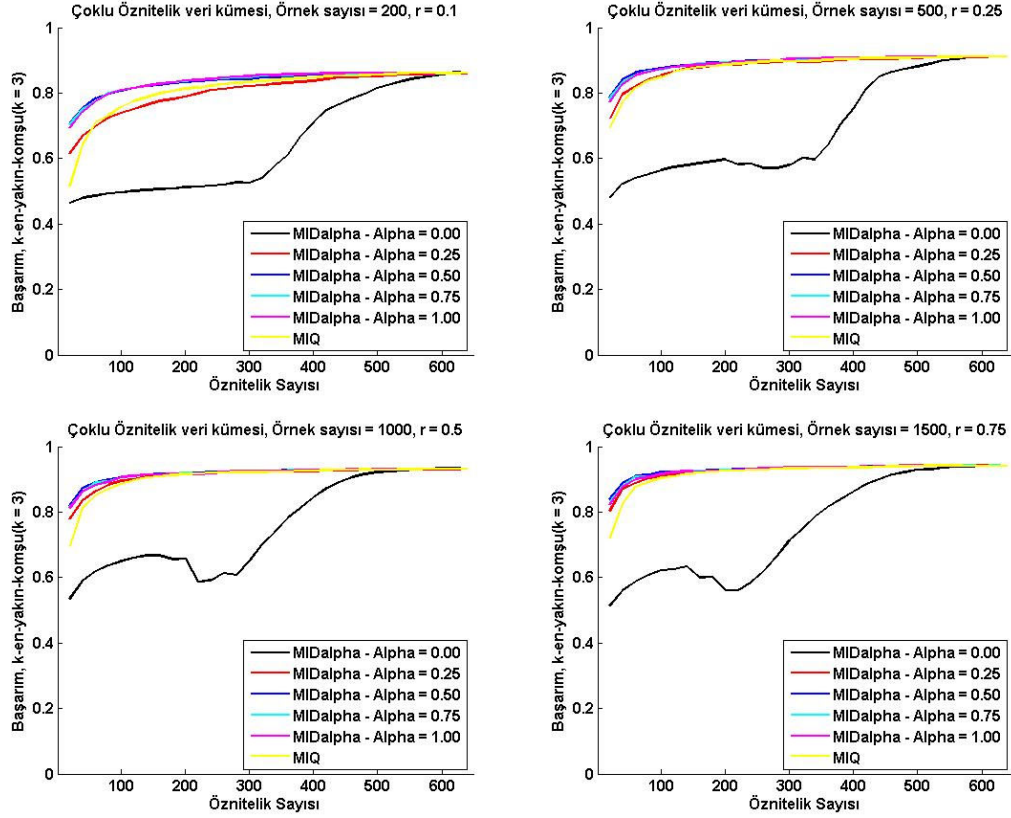
Çoklu Öznitelik veri kümesi de Şarap ve Ses Türü veri kümelerine benzer sonuçlar göstermektedir. Şekil 8.11 ve Çizelge 8.9’da bu veri kümesi üzerinde gerçekleştirilen kararlılık deneylerinin sonuçları sergilenmiştir. $\alpha \geq 0.5$ değerleri için kararlılık sonuçları oldukça iyidir. Başarım yönünden ise tüm r değerleri için $\alpha = 0$ değerinin çok kötü sonuçlar vermesinin dışında gözle görülür bir fark bulunmamaktadır. Çoklu Öznitelik veri kümesi için k-en-yakın-komşu algoritmasının başarım sonuçları Şekil 8.12 ve Çizelge 8.10’da gösterilmiştir. Sınıflandırıcı olarak svm algoritmasının kullanıldığı başarım sonuçları deneylerin çok uzun zaman alması nedeniyle elde edilememiştir.



Şekil 8.11 : MRMR, MID_α seçim kriterinin farklı α ve r değerleri için Çoklu Öznitelik veri kümesi üzerinde kararlılık sonuçları.

Çizelge 8.9 : MRMR, MID_α seçim kriterinin farklı α ve r değerleri için Çoklu Öznitelik veri kümesi üzerindeki kararlılık sonuçlarının ortalama sıralanma değerleri ve standart sapmaları.

r oranı	$\alpha = 0$	$\alpha = 0.25$	$\alpha = 0.5$	$\alpha = 0.75$	$\alpha = 1$
$r = 0.1$	3.3 ± 1.8	4.4 ± 0.6	2.4 ± 0.6	2.2 ± 0.8	2.7 ± 1.5
$r = 0.25$	3.4 ± 1.9	4.5 ± 0.5	2.8 ± 0.6	2.2 ± 0.7	2.1 ± 1.3
$r = 0.5$	3.3 ± 1.8	4.5 ± 0.5	3.2 ± 0.6	2.1 ± 0.8	1.9 ± 0.9
$r = 0.75$	3.2 ± 1.8	4.5 ± 0.5	3.3 ± 0.6	2.1 ± 0.8	1.9 ± 1.0



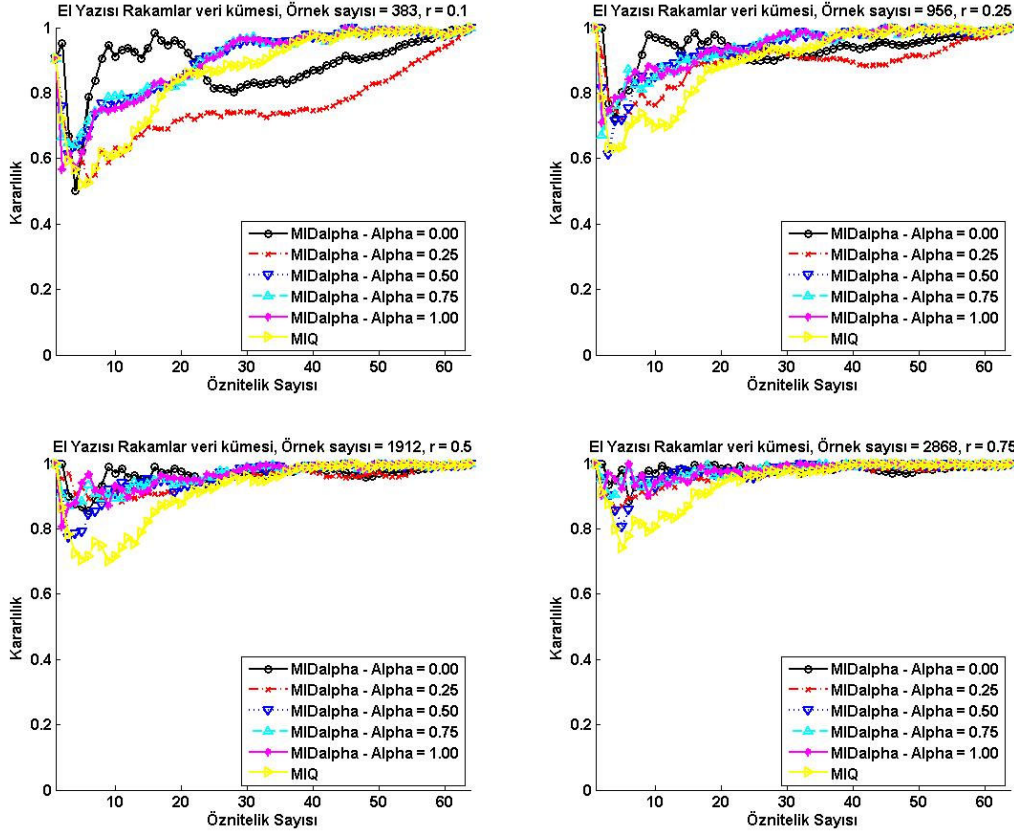
Şekil 8.12 : MRMR, MID_{α} seçim kriterinin farklı α ve r değerleri için Çoklu Öznitelik veri kümesi üzerinde başarımlar (k-en-yakın-komşu) sonuçları.

Çizelge 8.10 : MRMR, MID_{α} seçim kriterinin farklı α ve r değerleri için Çoklu Öznitelik veri kümesi üzerindeki başarımlar (k-en-yakın-komşu) sonuçlarının ortalama sıralanma değerleri ve standart sapmaları.

r oranı	$\alpha = 0$	$\alpha = 0.25$	$\alpha = 0.5$	$\alpha = 0.75$	$\alpha = 1$
$r = 0.1$	4.8 ± 0.7	3.9 ± 0.4	2.4 ± 0.9	1.8 ± 0.6	1.7 ± 0.8
$r = 0.25$	4.7 ± 0.7	3.8 ± 0.7	1.9 ± 0.9	2.1 ± 0.6	2.1 ± 0.9
$r = 0.5$	4.6 ± 1.0	3.6 ± 0.7	1.9 ± 0.9	2.1 ± 0.8	2.1 ± 1.0
$r = 0.75$	4.7 ± 0.7	3.4 ± 1.0	1.8 ± 0.9	2.0 ± 0.7	2.2 ± 1.0

Son olarak El Yazısı Rakamlar veri kümesi için deney sonuçları sunulacaktır. Şekil 8.13 bu veri kümesine ait kararlılık deney sonuçlarını göstermektedir. Düşük öznitelik sayıları için $\alpha = 0$ değeri yüksek kararlılık sonuçları elde etsede, genel olarak $\alpha \geq 0.5$ değerleri için yüksek kararlılık değerleri elde edildiği Çizelge 8.11'den anlaşılmaktadır. Bu veri kümesi başarımlar sonuçları yönünden Çoklu Öznitelik veri kümesine çok benzemektedir. Yine $\alpha = 0$ değeri için çok kötü başarımlar değerleri elde edilmiş, diğer α değerleri arasında ise başarımlar yönünden kayda değer bir fark olmadığı gözlenmiştir. K-en-yakın-komşu algoritması kullanılarak elde edilen başarımlar ve sıralanma değerleri Şekil 8.14 ve Çizelge 8.12'de

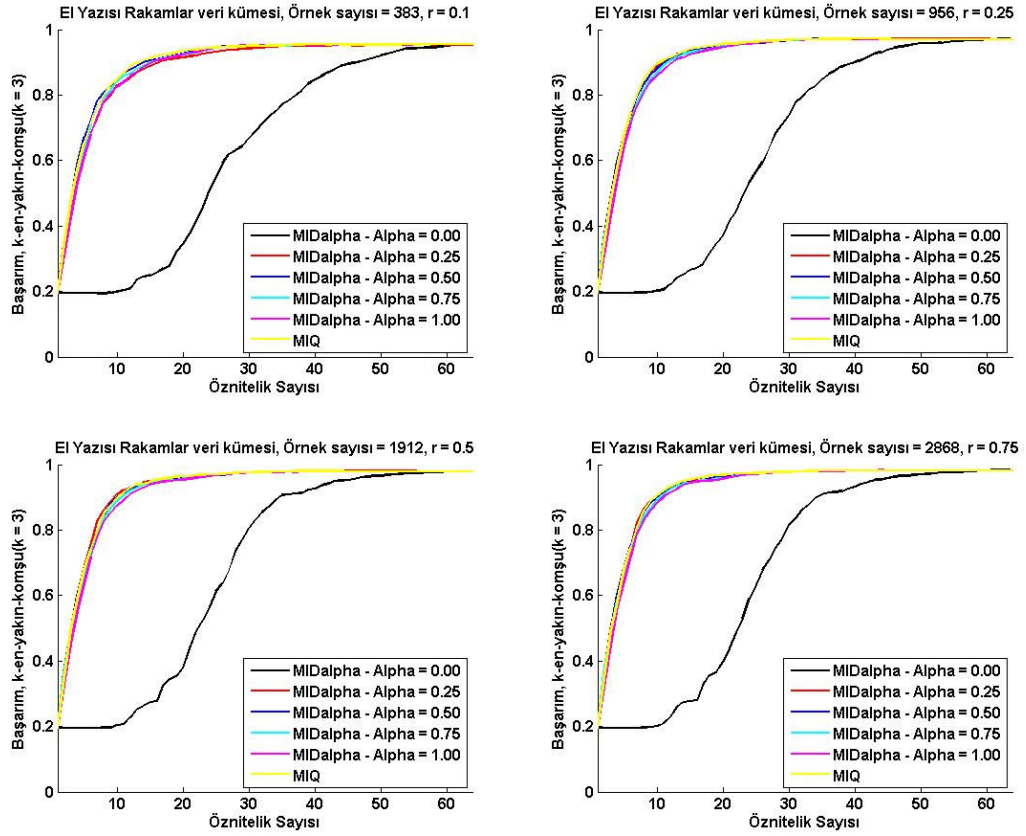
görülebilir. Svm başarımları sonuçları ise Ek B.9 bölümünde Şekil B.9 ve Çizelge B.9’da verilmiştir.



Şekil 8.13 : MRMR, MID_{α} seçim kriterinin farklı α ve r değerleri için El yazısı Rakamlar veri kümesi üzerinde kararlılık sonuçları.

Çizelge 8.11 : MRMR, MID_{α} seçim kriterinin farklı α ve r değerleri için El yazısı Rakamlar veri kümesi üzerindeki kararlılık sonuçlarının ortalama sıralanma değerleri ve standart sapmaları.

r oranı	$\alpha = 0$	$\alpha = 0.25$	$\alpha = 0.5$	$\alpha = 0.75$	$\alpha = 1$
$r = 0.1$	2.8 ± 1.5	4.6 ± 1.0	2.3 ± 0.9	2.2 ± 0.9	2.3 ± 1.1
$r = 0.25$	2.9 ± 1.5	4.4 ± 1.0	2.4 ± 1.0	2.3 ± 1.1	2.0 ± 1.0
$r = 0.5$	3.0 ± 1.6	3.6 ± 1.3	2.7 ± 1.4	2.4 ± 0.9	2.1 ± 1.2
$r = 0.75$	2.5 ± 1.4	3.8 ± 1.3	2.5 ± 1.4	2.4 ± 1.1	2.3 ± 1.2



Şekil 8.14 : MRMR, MID_{α} seçim kriterinin farklı α ve r değerleri için El Yazısı Rakamlar veri kümesi üzerinde başarımlı (k-en-yakın-komşu) sonuçları.

Çizelge 8.12 : MRMR, MID_{α} seçim kriterinin farklı α ve r değerleri için El Yazısı Rakamlar veri kümesi üzerindeki başarımlı (k-en-yakın-komşu) sonuçlarının ortalama sıralanma değerleri ve standart sapmaları.

r oranı	$\alpha = 0$	$\alpha = 0.25$	$\alpha = 0.5$	$\alpha = 0.75$	$\alpha = 1$
$r = 0.1$	4.3 ± 0.8	4.1 ± 1.1	2.0 ± 0.9	1.9 ± 0.8	1.9 ± 0.9
$r = 0.25$	4.4 ± 0.9	3.8 ± 1.0	2.0 ± 0.9	2.0 ± 0.9	1.7 ± 0.8
$r = 0.5$	4.5 ± 0.8	3.1 ± 1.2	2.3 ± 1.1	2.0 ± 0.8	1.8 ± 1.0
$r = 0.75$	4.5 ± 0.8	3.1 ± 1.0	2.0 ± 1.1	1.9 ± 0.9	1.9 ± 0.9

9. TARTIŞMA VE SONUÇ

Bu çalışmada öznitelik seçme algoirtmalarının sonucu olan öznitelik kümeleri arasındaki kararlılık seviyesini belirleyen bir ölçüt geliştirilerek sunulan deneysel metodoloji yardımıyla bir çok veri kümesi üzerinde kararlı öznitelik seçimi gerçekleştirilmeye çalışılmıştır. MRMR algoritmasının MIQ kriterine göre daha kararlı olduğu ispatlanan MID seçim kriterinin üzerinde yapılan modifikasyon sonucu yeni bir seçim kriteri, MID_{α} , ortaya konulmuş ve bir çok veri kümesinde kararlılık ve başarımlar yönünden test edilmiştir. Üstünde deneyler yapılan veri kümeleri için kararlılığı arttıran α değerleri sunulurken bu yöntemi ve önerilen değerleri kullanmanın daha kararlı sonuçlar elde edilmesinin yanında başarımlar yönünden herhangi bir kötü sonuca ya da düşüşe yol açmadığı gösterilmiştir.

Veri kümelerinin tanıtıldığı bölümde veri kümeleri hakkında verilen kısa bilgilerden sonra her öznitelik için sınıf etiketleri ile olan simetrik belirsizlik değerleri hesaplanmış ve solda bulunan grafiklerde büyükten küçüğe doğru sıralanarak gösterilmiştir. Elde edilen bu grafiğin azalan olması doğası gereğidir. Bununla beraber şekillerin sağ tarafında verilen grafiklerde solda elde edilen sıralama kullanılarak her bir öz niteliğin kendi dışındaki diğer özniteliklerle olan ortalama simetrik belirsizlik değerleri hesaplanmış ve grafiğe dökülmüştür. Parkinson, Sonar, İyonosfer ve Musk veri kümeleri için sağdaki grafikler herhangi bir eğilim göstermemiştir. Bu veri kümelerinde herhangi bir α değeri belirgin bir biçimde öne çıkmamasına rağmen görülmüştür ki bu veri kümeleri üzerinde gerçekleştirilen kararlılık deneylerinde İyonosfer veri kümesi dışında $\alpha \leq 0.5$ değerleri diğer α değerlerine göre azda olsa bir adım öndedirler. Diğer taraftan Şarap, Ses Türü, Çoklu Öznitelik ve El Yazısı Rakam veri kümeleri için elde edilen aynı simetrik belirsizlik grafikleri soldaki grafiklere benzer bir şekilde azalan eğilim göstermektedir. Yine aynı veri kümeleri için elde edilen deney sonuçlarında görülmüştür ki yüksek α değerleri kararlılık ve başarımlar yönünden diğerlerine nazaran daha iyi sonuçlar vermektedir. Bu ilginç çıkarım şu şekilde açıklanabilir: Sınıf etiketleri ile yüksek ilişkili öznitelikler veri kümesindeki diğer öznitelikler ile de yüksek ilişkili değerlerine

sahipseler ya da sınıf etiketleri ile ilişkisiz öznitelikler veri kümesindeki diğer özniteliklerle de ilişkisiz ya da az ilişkili durumdalarsa, bu veri kümesi için kararlılık ve başarımlardan daha iyi sonuçlar elde edilmesi için öznitelik seçiminde sınıf etiketleri vektörü ile olan ilişkiye daha fazla, öznitelik artıklığına ise daha az ağırlık verilmesi gerekmektedir. MID_{α} tekniği kullanılarak yüksek α değerleri ile öznitelik seçimi yapıldığında bu tarz veri kümelerinde daha kararlı ve başarımları yüksek sonuçlar elde edildiği Deney Sonuçları bölümünde gösterilmiştir. Ters mantık uygulayarak Parkinson, Sonar ve Musk veri kümelerine benzer simetrik belirsizlik değerleri gösteren veri kümeleri için düşük α değerleri azda olsa daha iyi sonuç vereceği söylenebilir.

Kararlılık yönünden farklı α değerleri için diğerlerine nazaran daha başarılı sonuçlar alınmasına rağmen çoğu veri kümesinde başarımlardan çok az ilerleme kaydedildiği gözlenmiştir. Ancak bu bir eksi olarak görülmemelidir çünkü kararlılığın artması esnasında başarımlarının düşmemesi de bir artı olarak değerlendirilmelidir.

Yapılacak gelecek çalışmalardan ilki kararlılık değerini arttıran α değerinin bu çalışmada uygulanan deneysel metodoloji yöntemi kullanılarak değil de veri kümesinin daha farklı özelliklerini inceleyerek bu özelliklerin kararlılık değerlerini ne şekilde etkilediğini saptayıp bir kural listesi çıkarmak ve hangi α değerinin kullanılması gerektiğine bu kural listesi kullanılarak karar verilmesi olabilir. Diğer bir deyişle veri kümesi için en uygun α değerinin önceden kestirimi yapılabilir.

KAYNAKLAR

- [1] **Guyon, I., Elisseeff, A.**, 2003: An Introduction to Variable and Feature Selection. In *Journal of Machine Learning Research*. 3, p. 1157-1182.
- [2] **R. Duda, P. Hart and D. Stork**, 2001: *Pattern classification*, 2nd ed. New York: John Wiley and Sons.
- [3] **A. Hyvarinen and E. Oja**, 1999: "Independent component analysis: Algorithms and applications," *Neural Networks*, 13, p. 411-430.
- [4] **Ding, C., Peng, H.**, 2003: Minimum Redundancy Feature Selection from Microarray Gene Expression Data. In *Proceedings of the Computational Systems Bioinformatics conference (CSB'03)*, p. 523-529.
- [5] **L. Yu and H. Liu**, 2003: "Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution". In *Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003)*, Washington DC.
- [6] **Das, S.**, 2001: Filters, wrappers and a boosting-based hybrid for feature selection. *Proceedings of the Eighteenth International Conference on Machine Learning*, p. 74-81.
- [7] **Xing, E., Jordan, M., & Karp, R.**, 2001: Feature selection for high-dimensional genomic microarray data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, p. 601-608.
- [8] **John, G. H., Kohavi, R., Pfleger, K.**, 1994: Irrelevant Feature and The Subset Selection Problem. In *Proceedings of the Eleventh International Conference on Machine Learning*, p. 121-129.
- [9] **Liu, H., Yu, L.**, 2005: Toward Integrating Feature Selection Algorithms for Classification and Clustering. In *IEEE Transactions on Knowledge and Data Engineering* 17(4), p. 491-502.
- [10] **Yu, L., Liu, H.**, 2004: Efficient Feature Selection via Analysis of Relevance and Redundancy. In *Journal of Machine Learning Research* 5, p. 1205-1224.
- [11] **Pepe, M. S., Etzioni, R., Feng, Z. and et al**, 2001: Phases of Biomarker Development for Early Detection of Cancer. In *J Natl Cancer Inst* 93, p. 1054-1060.
- [12] **Kalousis, A., Prados, J., Hilario, M.**, 2007: Stability of Feature Selection Algorithms: A Study on High-Dimensional Spaces. In *Knowledge and Information Systems* 12, p. 95-116.

- [13] **Yu, L., Ding, C., Loscalzo, S.,** 2008: Stable Feature Selection via Dense Feature Groups, In *Proceedings of the 14th ACM International Conference on Knowledge Discovery and Data Mining (KDD-08)*, p. 803-811.
- [14] **Saeys, Y., Abeel, T., Peer, Y. V.,** 2008: Robust Feature Selection Using Ensemble Feature Selection Techniques. In *Proceedings of the ECML Conference*, p. 313-325.
- [15] **Asuncion, A. & Newman, D.J.,** 2007: UCI Machine Learning Repository [<http://www.ics.uci.edu/~mllearn/MLRepository.html>]. Irvine, CA: University of California, School of Information and Computer Science.
- [16] **Tzanetakis G., Cook P.,** 2002: Musical genre classification of audio signals. In *IEEE Transactions on Speech and Audio Processing*, 10(5), p. 293-302.
- [17] **Moerchen F., Ultsch A., Thies M., Loehken I.,** 2006: Modelling timbre distance with temporal statistics from polyphonic music. In *IEEE Transactions on Speech and Audio Processing* 14, p. 81-90.
- [18] **Press, W. H., Flannery, B. P., Teukolsky, S. A., & Vetterling, W. T.,** 1988. *Numerical recipes in C*. Cambridge University Press, Cambridge.
- [19] **Quinlan, J.,** 1993: *C4.5: Programs for machine learning*. Morgan Kaufmann.
- [20] <<http://mathpages.com/home/kmath042/kmath042.htm>> alındığı tarih : 05.04.2009.
- [21] <http://en.wikipedia.org/wiki/Cauchy_distribution> alındığı tarih : 05.06.2009.
- [22] <<http://www.mathworks.com/matlabcentral/fileexchange/11609>> alındığı tarih : 12.10.2008.
- [23] <<http://www.mathworks.com/matlabcentral/>> alındığı tarih : 12.10.2008.
- [24] **Alpaydin, E.,** 2004: *Introduction to Machine Learning*, The MIT Press, Cambridge, Massachussetts.
- [25] **Duin, R. P. W.,** 2000: PR Tools Version 3.0, A Matlab Toolbox for Pattern Recognition. <http://www.ph.tn.tudelft.nl/bob/PRTTOOLS.html>.
- [26] **Chih-Chung Chang and Chih-Jen Lin,** 2001: LIBSVM : a library for support vector machines. Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.

EKLER

- EKA.1 :** Öznitelik Seçme Yöntemlerinin Parkinson Veri Kümesi Üzerindeki Sonuçları
- EKA.2 :** Öznitelik Seçme Yöntemlerinin Sonar Veri Kümesi Üzerindeki Sonuçları
- EKA.3 :** Öznitelik Seçme Yöntemlerinin İyonosfer Veri Kümesi Üzerindeki Sonuçları
- EKA.4 :** Öznitelik Seçme Yöntemlerinin Musk Veri Kümesi Üzerindeki Sonuçları
- EKA.5 :** Öznitelik Seçme Yöntemlerinin Şarap Veri Kümesi Üzerindeki Sonuçları
- EKA.6 :** Öznitelik Seçme Yöntemlerinin Ses Türü Veri Kümesi Üzerindeki Sonuçları
- EKA.7 :** Öznitelik Seçme Yöntemlerinin Çoklu Öznitelik Veri Kümesi Üzerindeki Sonuçları
- EKA.8 :** Öznitelik Seçme Yöntemlerinin El Yazısı Rakamlar Veri Kümesi Üzerindeki Sonuçları
- EKB.1 :** Parkinson Veri Kümesinde K-En-Yakın-Komşu Başarım Sonuçları
- EKB.2 :** Parkinson Veri Kümesinde SVM Başarım Sonuçları
- EKB.3 :** Sonar Veri Kümesinde SVM Başarım Sonuçları
- EKB.4 :** İyonosfer Veri Kümesinde K-En-Yakın-Komşu Başarım Sonuçları
- EKB.5 :** İyonosfer Veri Kümesinde SVM Başarım Sonuçları
- EKB.6 :** Musk Veri Kümesinde K-En-Yakın-Komşu Başarım Sonuçları
- EKB.7 :** Şarap Veri Kümesinde K-En-Yakın-Komşu Başarım Sonuçları
- EKB.8 :** Ses Türü Veri Kümesinde K-En-Yakın-Komşu Başarım Sonuçları
- EKB.9 :** El Yazısı Rakamlar Veri Kümesinde SVM Başarım Sonuçları

EKA.1

Çizelge A.1 : FCBF, MRMR-MID ve MRMR-MIQ öznitelik seçme yöntemlerinin Parkinson veri kümesi üzerindeki sonuçları

FCBF	22, 2
MRMR-MID	22, 2, 13, 15, 20, 6, 1, 12, 3, 19, 5, 17, 10, 18, 8, 21, 4, 9, 7, 16, 11, 14
MRMR-MIQ	22, 2, 13, 1, 19, 20, 6, 3, 12, 18, 10, 17, 5, 11, 16, 9, 4, 14, 8, 7, 21, 15

EKA.2

Çizelge A.2 : FCBF, MRMR-MID ve MRMR-MIQ öznitelik seçme yöntemlerinin Sonar veri kümesi üzerindeki sonuçları

FCBF	11, 51
MRMR-MID	12, 51, 4, 50, 52, 9, 49, 5, 60, 11, 2, 6, 58, 47, 40, 59, 21, 3, 13, 44, 1, 36, 48, 14, 10, 43, 31, 46, 54, 39, 28, 7, 56, 15, 37, 55, 17, 57, 22, 38, 53, 45, 24, 41, 20, 8, 33, 27, 42, 19, 26, 32, 16, 35, 29, 30, 23, 25, 34, 18
MRMR-MIQ	12, 51, 49, 52, 11, 59, 50, 9, 5, 13, 1, 36, 47, 6, 4, 21, 10, 48, 2, 46, 44, 54, 58, 37, 17, 15, 39, 20, 60, 45, 38, 16, 26, 35, 31, 22, 3, 43, 8, 55, 27, 24, 14, 42, 19, 40, 41, 53, 7, 33, 28, 32, 30, 56, 57, 23, 25, 34, 18, 29

EKA.3

Çizelge A.3 : FCBF, MRMR-MID ve MRMR-MIQ öznitelik seçme yöntemlerinin İyonosfer veri kümesi üzerindeki sonuçları

FCBF	1, 5
MRMR-MID	5, 1, 2, 4, 3, 33, 24, 16, 10, 7, 18, 25, 6, 34, 31, 28, 29, 9, 14, 12, 17, 32, 8, 23, 26, 20, 13, 22, 21, 11, 30, 27, 15, 19
MRMR-MIQ	5, 1, 18, 3, 4, 33, 24, 7, 14, 31, 6, 29, 28, 10, 9, 16, 25, 32, 15, 21, 8, 34, 17, 12, 22, 11, 20, 13, 23, 26, 30, 27, 19, 2

EKA.4

Çizelge A.4 : FCBF, MRMR-MID ve MRMR-MIQ öznitelik seçme yöntemlerinin Musk veri kümesi üzerindeki sonuçları

FCBF	36
MRMR-MID	36, 67, 145, 13, 5, 157, 136, 31, 37, 147, 141, 146, 58, 76, 162, 16, 79, 59, 91, 28, 94, 9, 164, 30, 107, 133, 165, 56, 50, 102, 111, 29, 40, 140, 43, 88, 99, 63, 51, 54, 149, 84, 105, 138, 132, 77, 134, 163, 116, 93, 66, 44, 86, 92, 20, 159, 47, 129, 42, 21, 95, 68, 121, 15, 2, 18, 4, 1, 106, 166, 19, 120, 152, 83, 87, 75, 97, 69, 150, 46, 82, 110, 6, 148, 24, 158, 154, 12, 10, 52, 33, 70, 126, 74, 25, 96, 98, 8, 39, 125, 45, 118, 117, 48, 160, 78, 62, 34, 130, 161, 137, 81, 123, 155, 73, 49, 17, 127, 55, 80, 153, 103, 114, 32, 64, 142, 124, 108, 156, 27, 14, 65, 128, 3, 26, 89, 41, 53, 115, 151, 135, 101, 61, 11, 38, 7, 35, 112, 72, 131, 143, 119, 122, 139, 60, 104, 71, 22, 109, 113, 100, 85, 23, 144, 57, 90
MRMR-MIQ	36, 157, 145, 79, 5, 59, 58, 136, 162, 67, 43, 40, 37, 92, 134, 30, 76, 148, 29, 9, 91, 42, 54, 28, 116, 163, 12, 151, 149, 88, 4, 102, 13, 70, 133, 31, 115, 68, 16, 84, 141, 166, 129, 138, 135, 99, 150, 50, 165, 19, 17, 38, 69, 95, 140, 87, 132, 10, 120, 147, 51, 127, 118, 64, 20, 114, 110, 96, 21, 57, 97, 94, 63, 3, 130, 60, 98, 6, 77, 34, 39, 142, 155, 89, 83, 48, 2, 93, 111, 11, 112, 108, 121, 75, 8, 27, 124, 117, 119, 126, 113, 71, 128, 32, 46, 137, 26, 65, 74, 41, 52, 107, 144, 125, 49, 86, 47, 159, 45, 33, 100, 139, 72, 101, 131, 143, 55, 7, 160, 73, 154, 23, 44, 164, 25, 82, 81, 103, 152, 56, 161, 80, 1, 61, 158, 106, 14, 24, 22, 18, 53, 156, 146, 90, 109, 85, 78, 35, 105, 62, 66, 153, 104, 15, 123, 122

EKA.5

Çizelge A.5 : FCBF, MRMR-MID ve MRMR-MIQ öznitelik seçme yöntemlerinin Şarap veri kümesi üzerindeki sonuçları

FCBF	7, 13, 2, 1, 10, 11
MRMR-MID	7, 1, 11, 13, 2, 10, 12, 6, 5, 9, 3, 4, 8
MRMR-MIQ	7, 1, 11, 13, 2, 10, 12, 6, 5, 9, 4, 3, 8

EKA.6

Çizelge A.6 : FCBF, MRMR-MID ve MRMR-MIQ öznitelik seçme yöntemlerinin Ses Türü veri kümesi üzerindeki sonuçları

FCBF	24, 31, 12, 46, 13, 8, 40, 47, 42, 33, 32, 16, 2, 48
MRMR-MID	24, 46, 40, 12, 31, 9, 47, 13, 42, 25, 36, 8, 33, 11, 32, 16, 39, 43, 37, 2, 7, 41, 50, 23, 1, 14, 18, 22, 34, 6, 17, 48, 45, 49, 28, 38, 26, 15, 35, 21, 10, 4, 19, 30, 27, 20, 5, 44, 29, 3
MRMR-MIQ	24, 40, 31, 9, 47, 12, 46, 36, 6, 42, 25, 8, 33, 32, 13, 39, 11, 16, 43, 23, 50, 41, 37, 2, 14, 34, 18, 21, 7, 22, 48, 1, 45, 49, 28, 17, 26, 38, 15, 35, 30, 10, 4, 19, 27, 20, 5, 44, 29, 3

EKA.7**Çizelge A.7 : FCBF, MRMR-MID ve MRMR-MIQ öznitelik seçme yöntemlerinin Çoklu Öznitelik veri kümesi üzerindeki sonuçları**

FCBF	358, 362, 181, 97, 1, 289, 186, 29, 133, 515, 223, 37, 55, 293, 120, 290, 157, 195, 184, 48, 194, 419, 154, 292, 2, 33, 295, 485, 608, 198, 222, 647, 219, 190, 217, 104, 577, 624, 221, 208, 39, 648, 458, 463, 205, 625, 285, 169, 224, 303, 145, 225, 213, 410, 294, 297, 430, 158, 179, 62, 456, 623, 313, 287, 85, 230, 509, 626, 505, 299, 302, 621, 476, 209, 298, 307, 227, 595, 379, 229, 392, 402, 128, 518, 460, 537, 320, 164, 304, 228, 601, 317, 323, 383, 122, 318, 234, 610, 286, 335, 300, 226, 291, 232, 281, 319, 315, 619, 609, 334, 242, 306, 220, 279, 329, 497, 231, 239, 240, 331, 326, 310, 237, 273, 324, 604, 325, 348, 271, 346, 269, 252, 340, 343, 339, 309, 350, 314, 238, 257, 244, 260, 259, 322, 354, 345, 262, 251, 258, 347, 355
MRMR-MID	362, 181, 357, 358, 223, 29, 97, 19, 361, 290, 293, 1, 109, 360, 37, 289, 133, 55, 359, 218, 53, 132, 193, 186, 94, 292, 7, 185, 108, 295, 67, 157, 217, 146, 184, 72, 65, 106, 144, 195, 33, 222, 135, 30, 194, 99, 154, 608, 12, 113, 6, 64, 120, 624, 190, 219, 48, 647, 207, 2, 515, 105, 41, 221, 10, 192, 200, 43, 198, 125, 84, 36, 57, 91, 8, 111, 419, 103, 633, 13, 165, 202, 104, 199, 114, 180, 86, 613, 500, 434, 212, 93, 31, 66, 22, 26, 642, 119, 112, 152, 115, 420, 123, 516, 140, 124, 224, 81, 15, 303, 54, 24, 50, 166, 645, 201, 640, 530, 225, 204, 296, 514, 285, 82, 177, 102, 188, 183, 208, 49, 205, 648, 501, 577, 179, 117, 147, 127, 297, 435, 531, 34, 151, 141, 78, 216, 118, 44, 45, 485, 458, 174, 173, 75, 90, 625, 187, 129, 134, 576, 88, 644, 3, 213, 47, 130, 153, 294, 529, 230, 110, 287, 149, 578, 70, 98, 35, 499, 38, 169, 85, 404, 486, 100, 139, 421, 138, 158, 471, 626, 9, 167, 87, 142, 444, 145, 561, 136, 649, 449, 159, 191, 69, 42, 546, 623, 545, 68, 46, 148, 150, 172, 39, 52, 418, 592, 313, 96, 429, 59, 405, 62, 445, 18, 227, 459, 51, 170, 436, 594, 632, 302, 462, 638, 472, 430, 229, 403, 56, 593, 639, 4, 79, 298, 307, 612, 288, 579, 540, 448, 463, 470, 32, 635, 555, 634, 433, 80, 457, 517, 14, 562, 74, 58, 491, 410, 539, 560, 189, 128, 630, 63, 73, 450, 476, 209, 61, 621, 395, 505, 447, 196, 443, 299, 556, 21, 156, 484, 101, 554, 210, 575, 456, 161, 431, 541, 171, 126, 92, 175, 76, 425, 544, 525, 502, 162, 160, 380, 446, 320, 20, 524, 206, 143, 379, 228, 27, 40, 137, 155, 464, 510, 490, 317, 176, 415, 469, 406, 301, 16, 591, 365, 71, 304, 107, 396, 461, 11, 402, 504, 532, 460, 411, 121, 595, 455, 569, 234, 83, 526, 286, 487, 519, 364, 424, 552, 416, 164, 537, 477, 323, 509, 381, 391, 631, 318, 553, 417, 570, 182, 23, 375, 211, 641, 451, 366, 600, 409, 518, 528, 203, 473, 163, 495, 390, 432, 392, 538, 465, 226, 496, 585, 426, 513, 440, 536, 439, 376, 122, 506, 77, 414, 511, 547, 441, 568, 454, 300, 374, 480, 646, 291, 89, 475, 394, 368, 382, 601, 559, 335, 599, 197, 520, 580, 367, 389, 319, 232, 283, 281, 610, 428, 584, 523, 586, 535, 178, 401, 503, 388, 378, 533, 609, 617, 481, 312, 614, 168, 422, 571, 277, 242, 603, 315, 551, 387, 28, 369, 619, 5, 479, 615, 442, 627, 474, 370, 383, 231, 521, 332, 557, 306, 574, 437, 563, 116, 407, 466, 494, 334, 279, 534, 397, 498, 492, 220, 377, 60, 629, 489, 522, 363, 542,

Çizelge A.7 (devam) : FCBF, MRMR-MID ve MRMR-MIQ öznitelik seçme yöntemlerinin Çoklu Öznitelik veri kümesi üzerindeki sonuçları

MRMR-MID	543, 596, 423, 239, 305, 611, 330, 386, 590, 17, 620, 214, 25, 622, 240, 284, 373, 372, 329, 581, 371, 605, 550, 311, 583, 587, 331, 478, 527, 636, 310, 598, 328, 95, 337, 316, 326, 408, 427, 280, 237, 400, 273, 602, 507, 637, 324, 582, 512, 497, 488, 508, 384, 468, 572, 628, 452, 482, 247, 261, 567, 604, 249, 438, 265, 282, 566, 483, 412, 271, 269, 413, 252, 597, 346, 393, 616, 348, 643, 325, 565, 343, 558, 267, 385, 233, 131, 548, 453, 238, 340, 235, 350, 270, 266, 467, 339, 257, 336, 260, 309, 618, 549, 308, 255, 268, 244, 564, 314, 259, 256, 607, 253, 341, 264, 493, 278, 275, 215, 245, 274, 276, 262, 263, 321, 236, 246, 250, 399, 251, 345, 352, 354, 342, 589, 254, 322, 243, 573, 356, 248, 344, 258, 272, 338, 355, 347, 327, 606, 333, 241, 398, 353, 349, 351, 588
MRMR-MIQ	362, 459, 599, 559, 418, 496, 357, 410, 516, 595, 359, 406, 490, 592, 217, 364, 112, 49, 290, 500, 416, 624, 518, 223, 181, 10, 361, 552, 462, 358, 578, 469, 420, 99, 222, 53, 297, 429, 602, 292, 108, 225, 219, 576, 491, 458, 515, 360, 434, 193, 377, 218, 1, 37, 445, 555, 289, 86, 133, 425, 594, 531, 505, 224, 419, 600, 157, 476, 94, 577, 19, 480, 295, 228, 132, 647, 8, 436, 514, 29, 230, 138, 457, 285, 379, 146, 109, 579, 444, 213, 85, 626, 553, 122, 530, 411, 303, 106, 601, 633, 485, 184, 72, 591, 286, 68, 15, 227, 205, 421, 293, 454, 158, 561, 185, 35, 42, 501, 431, 504, 424, 221, 435, 593, 510, 12, 41, 190, 97, 392, 443, 186, 88, 613, 43, 317, 417, 645, 287, 64, 195, 67, 596, 477, 537, 229, 81, 415, 144, 307, 298, 365, 134, 642, 472, 562, 404, 460, 84, 291, 128, 455, 154, 495, 7, 517, 648, 231, 430, 13, 640, 519, 30, 463, 74, 499, 34, 318, 625, 405, 36, 363, 475, 113, 234, 540, 575, 649, 296, 75, 140, 529, 288, 433, 179, 118, 471, 414, 198, 192, 65, 44, 556, 121, 242, 546, 446, 639, 449, 632, 538, 55, 395, 54, 403, 608, 110, 539, 209, 172, 473, 6, 135, 142, 450, 120, 509, 409, 170, 169, 92, 119, 383, 461, 560, 481, 226, 402, 439, 541, 313, 437, 635, 380, 102, 124, 470, 375, 204, 302, 283, 506, 208, 87, 545, 125, 523, 580, 486, 294, 623, 200, 46, 370, 194, 598, 320, 448, 304, 378, 207, 525, 281, 284, 216, 202, 103, 587, 492, 456, 152, 45, 114, 145, 368, 428, 180, 66, 240, 62, 447, 396, 634, 143, 48, 98, 70, 159, 422, 91, 532, 644, 93, 590, 24, 32, 130, 524, 585, 101, 484, 2, 90, 167, 206, 277, 111, 306, 612, 502, 220, 554, 464, 366, 212, 376, 73, 33, 426, 312, 323, 26, 22, 319, 547, 149, 474, 610, 20, 123, 381, 621, 544, 47, 100, 280, 526, 423, 586, 451, 441, 61, 536, 638, 104, 382, 153, 174, 115, 391, 611, 166, 503, 80, 581, 232, 300, 440, 105, 369, 494, 239, 78, 513, 151, 96, 603, 182, 31, 609, 465, 394, 401, 57, 188, 82, 533, 487, 330, 528, 141, 199, 150, 173, 83, 432, 148, 468, 442, 551, 147, 191, 568, 332, 201, 279, 58, 52, 156, 387, 315, 18, 50, 557, 299, 176, 197, 14, 367, 129, 617, 301, 183, 482, 597, 59, 489, 374, 127, 622, 408, 520, 641, 386, 139, 165, 522, 155, 479, 569, 630, 136, 397, 164, 574, 187, 335, 542, 247, 71, 117, 3, 615, 407, 511, 126, 614, 161, 107, 27, 9, 498, 584, 390, 384, 237, 334, 11, 210, 178, 211, 570, 563, 646, 497, 388, 116, 305, 38, 177, 543, 39, 337, 265, 331, 508, 56, 76,

Çizelge A.7 (devam) : FCBF, MRMR-MID ve MRMR-MIQ öznitelik seçme yöntemlerinin Çoklu Öznitelik veri kümesi üzerindeki sonuçları

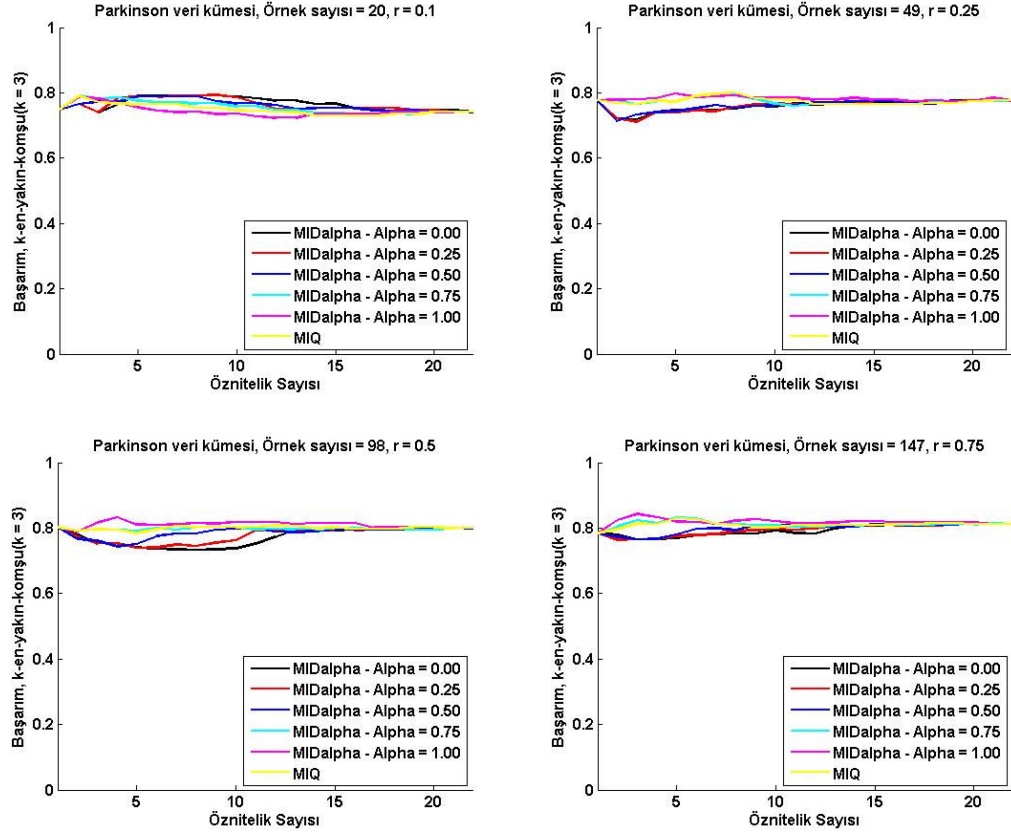
MRMR-MIQ	203, 389, 79, 371, 478, 627, 51, 4, 438, 631, 619, 168, 311, 571, 324, 534, 21, 252, 527, 582, 249, 175, 535, 466, 427, 261, 5, 23, 60, 329, 160, 171, 282, 269, 69, 346, 310, 16, 400, 521, 25, 162, 63, 512, 273, 343, 372, 196, 572, 260, 137, 507, 326, 233, 316, 604, 453, 328, 629, 393, 266, 452, 413, 620, 40, 605, 566, 189, 271, 385, 255, 483, 163, 270, 28, 77, 238, 348, 262, 583, 267, 567, 17, 214, 636, 373, 488, 550, 412, 235, 253, 89, 616, 325, 558, 250, 278, 350, 467, 257, 589, 244, 264, 251, 95, 259, 256, 637, 274, 268, 339, 245, 493, 263, 548, 276, 628, 340, 565, 236, 275, 309, 246, 243, 643, 314, 341, 254, 607, 321, 131, 399, 345, 354, 549, 336, 248, 352, 573, 564, 342, 308, 272, 258, 618, 215, 356, 347, 355, 322, 338, 344, 327, 241, 333, 398, 353, 349, 606, 351, 588
----------	--

EKA.8

Çizelge A.8 : FCBF, MRMR-MID ve MRMR-MIQ öznitelik seçme yöntemlerinin El Yazısı Rakamlar veri kümesi üzerindeki sonuçları

FCBF	43, 31, 29, 35, 44, 22, 27, 21, 37, 62, 47, 19, 11, 59, 20, 38, 28, 6, 36, 45, 51, 52, 53, 64
MRMR-MID	43, 22, 62, 44, 29, 27, 31, 11, 35, 37, 21, 14, 19, 39, 3, 47, 30, 20, 28, 55, 34, 51, 38, 6, 59, 52, 54, 36, 45, 63, 61, 42, 46, 23, 13, 53, 60, 10, 26, 7, 18, 4, 5, 15, 12, 2, 50, 64, 58, 56, 16, 8, 48, 24, 41, 49, 9, 32, 17, 25, 33, 57, 1, 40
MRMR-MIQ	43, 39, 11, 62, 22, 28, 27, 44, 6, 21, 37, 53, 31, 19, 35, 52, 59, 29, 47, 20, 14, 51, 63, 38, 34, 54, 30, 36, 13, 10, 45, 42, 61, 26, 55, 23, 46, 3, 7, 60, 18, 5, 12, 2, 64, 4, 15, 48, 49, 50, 56, 58, 16, 41, 8, 24, 25, 32, 9, 17, 33, 57, 1, 40

EK B.1

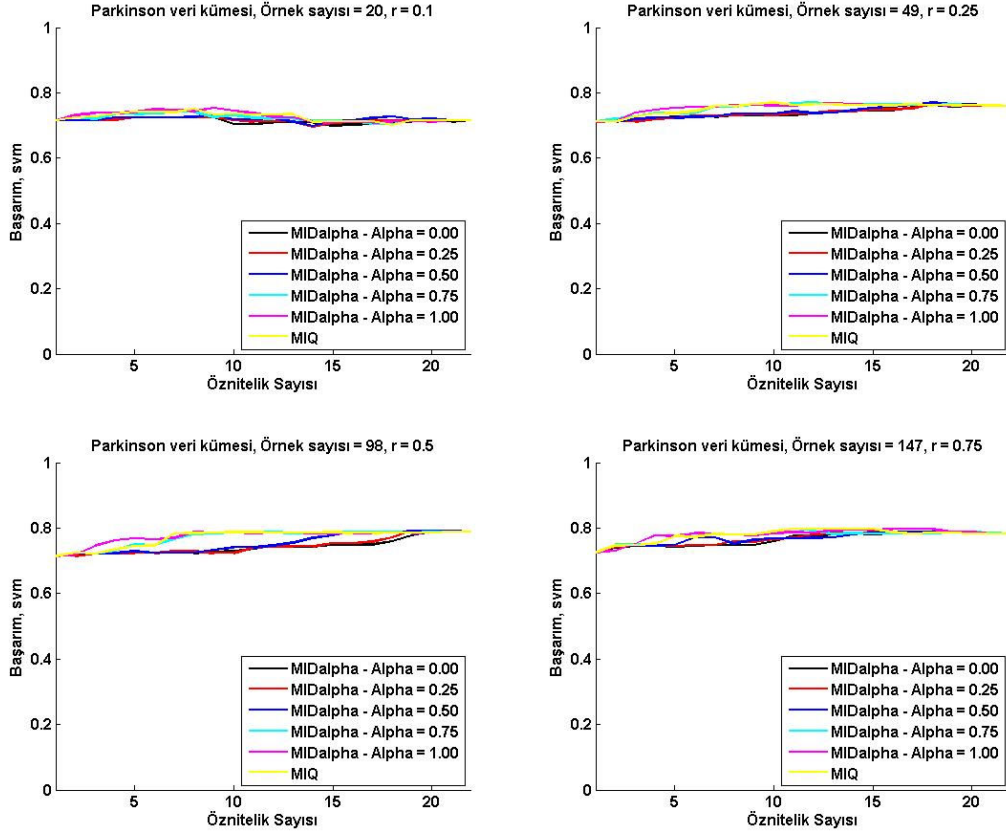


Şekil B.1 : MRMR, MID_{α} seçim kriterinin farklı α ve r değerleri için Parkinson veri kümesi üzerinde başarımlar (k-en-yakın-komşu) sonuçları.

Çizelge B.1 : MRMR, MID_{α} seçim kriterinin farklı α ve r değerleri için Parkinson veri kümesi üzerindeki başarımlar (k-en-yakın-komşu) sonuçlarının ortalama sıralama değerleri ve standart sapmaları.

r oranı	$\alpha = 0$	$\alpha = 0.25$	$\alpha = 0.5$	$\alpha = 0.75$	$\alpha = 1$
$r = 0.1$	2.6 ± 1.9	2.4 ± 0.9	2.4 ± 1.1	2.9 ± 1.1	4.0 ± 1.2
$r = 0.25$	3.6 ± 1.6	2.2 ± 1.2	2.9 ± 1.4	3.2 ± 1.0	2.4 ± 1.3
$r = 0.5$	4.5 ± 1.3	2.5 ± 1.2	1.8 ± 0.9	2.7 ± 1.1	2.9 ± 1.2
$r = 0.75$	4.4 ± 1.2	2.4 ± 1.4	2.3 ± 1.1	2.5 ± 1.1	2.7 ± 1.2

EK B.2

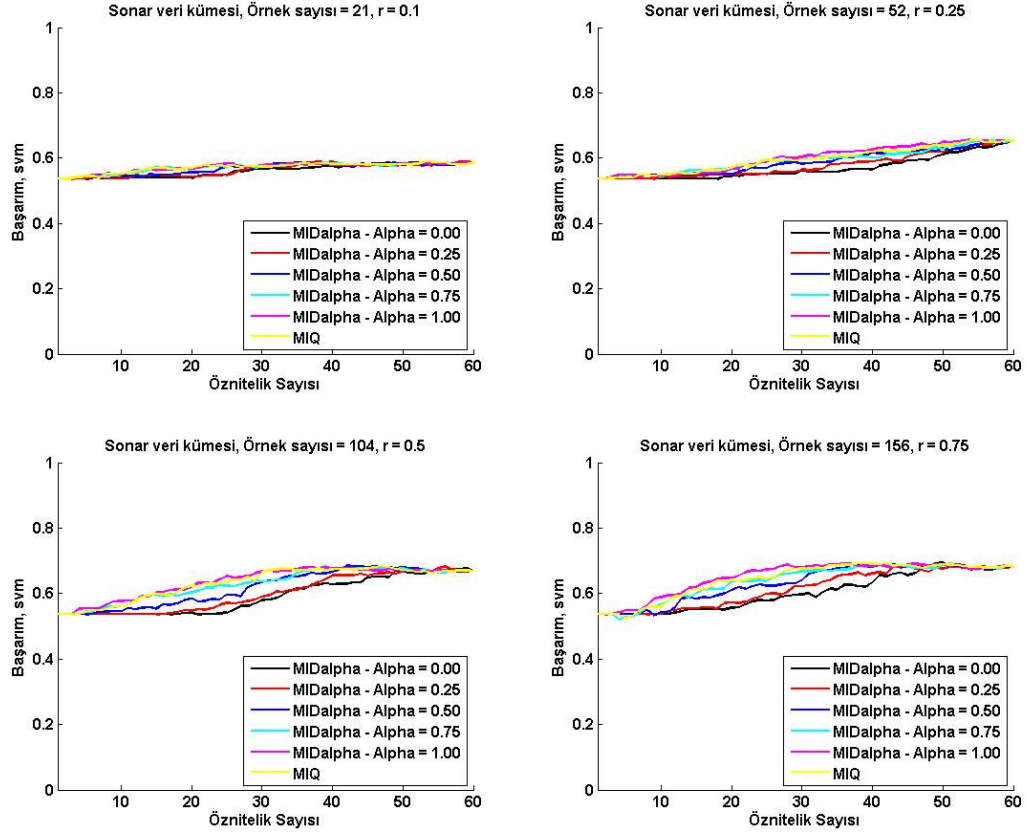


Şekil B.2 : MRMR, MID_α seçim kriterinin farklı α ve r değerleri için Parkinson veri kümesi üzerinde başarıml (svm) sonuçları.

Çizelge B.2 : MRMR, MID_α seçim kriterinin farklı α ve r değerleri için Parkinson veri kümesi üzerindeki başarıml (svm) sonuçlarının ortalama sıralanma değerleri ve standart sapmaları.

r oranı	$\alpha = 0$	$\alpha = 0.25$	$\alpha = 0.5$	$\alpha = 0.75$	$\alpha = 1$
$r = 0.1$	3.0 ± 1.9	2.3 ± 0.9	2.2 ± 1.1	2.8 ± 1.1	3.9 ± 1.2
$r = 0.25$	3.9 ± 1.4	2.2 ± 1.2	2.8 ± 1.3	3.1 ± 1.1	2.3 ± 1.4
$r = 0.5$	4.6 ± 1.0	2.4 ± 1.1	1.8 ± 0.8	2.6 ± 1.1	2.9 ± 1.2
$r = 0.75$	4.5 ± 1.1	2.4 ± 1.4	2.3 ± 1.0	2.5 ± 1.1	2.7 ± 1.2

EK B.3

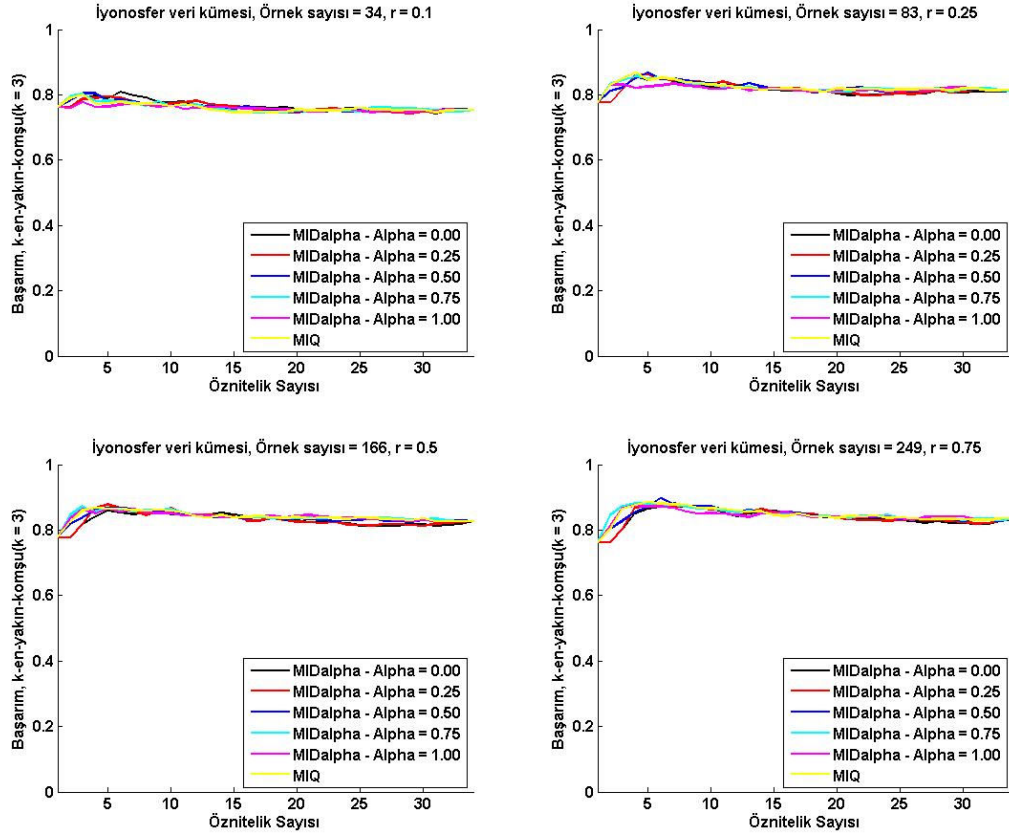


Şekil B.3 : MRMR, MID_α seçim kriterinin farklı α ve r değerleri için Sonar veri kümesi üzerinde başarımlar (svm) sonuçları.

Çizelge B.3 : MRMR, MID_α seçim kriterinin farklı α ve r değerleri için Sonar veri kümesi üzerindeki başarımlar (svm) sonuçlarının ortalama sıralanma değerleri ve standart sapmaları.

r oranı	$\alpha = 0$	$\alpha = 0.25$	$\alpha = 0.5$	$\alpha = 0.75$	$\alpha = 1$
$r = 0.1$	3.6 ± 1.8	2.5 ± 1.3	2.2 ± 1.2	3.0 ± 1.1	3.5 ± 0.8
$r = 0.25$	4.4 ± 1.3	1.5 ± 0.6	2.2 ± 1.0	3.4 ± 0.9	3.3 ± 1.1
$r = 0.5$	4.9 ± 0.6	1.6 ± 0.8	2.3 ± 1.0	2.9 ± 1.0	3.1 ± 1.1
$r = 0.75$	4.9 ± 0.5	1.9 ± 1.1	2.0 ± 0.9	3.0 ± 1.0	2.9 ± 1.1

EKB.4

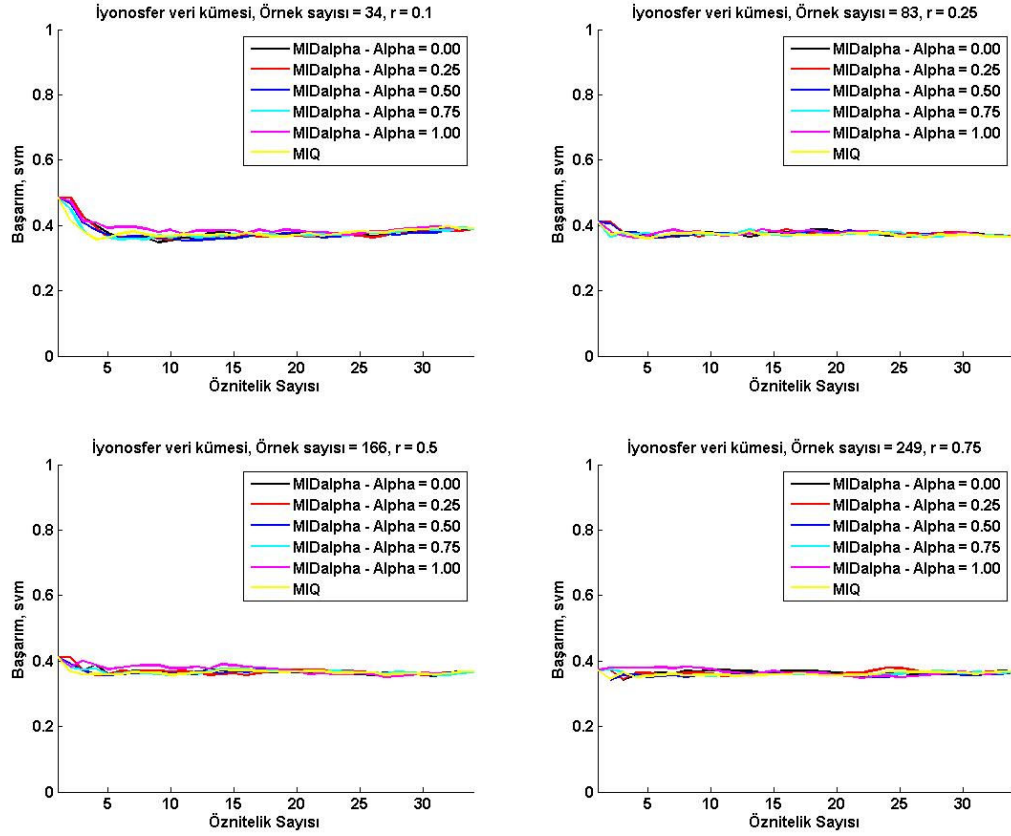


Şekil B.4 : MRMR, MID_α seçim kriterinin farklı α ve r değerleri için İyonosfer veri kümesi üzerinde başarımlı (k-en-yakın-komşu) sonuçları.

Çizelge B.4 : MRMR, MID_α seçim kriterinin farklı α ve r değerleri için İyonosfer veri kümesi üzerindeki başarımlı (k-en-yakın-komşu) sonuçlarının ortalama sıralanma değerleri ve standart sapmaları.

r oranı	$\alpha = 0$	$\alpha = 0.25$	$\alpha = 0.5$	$\alpha = 0.75$	$\alpha = 1$
$r = 0.1$	2.7 ± 1.9	3.5 ± 1.2	3.0 ± 0.9	2.4 ± 0.9	2.9 ± 1.7
$r = 0.25$	2.7 ± 1.9	3.2 ± 1.2	3.5 ± 1.2	3.1 ± 1.1	1.9 ± 1.0
$r = 0.5$	3.4 ± 1.6	3.6 ± 1.5	3.3 ± 1.0	2.8 ± 1.0	1.4 ± 0.7
$r = 0.75$	3.8 ± 1.4	3.6 ± 1.2	3.2 ± 1.1	2.5 ± 1.0	1.4 ± 0.8

EK B.5

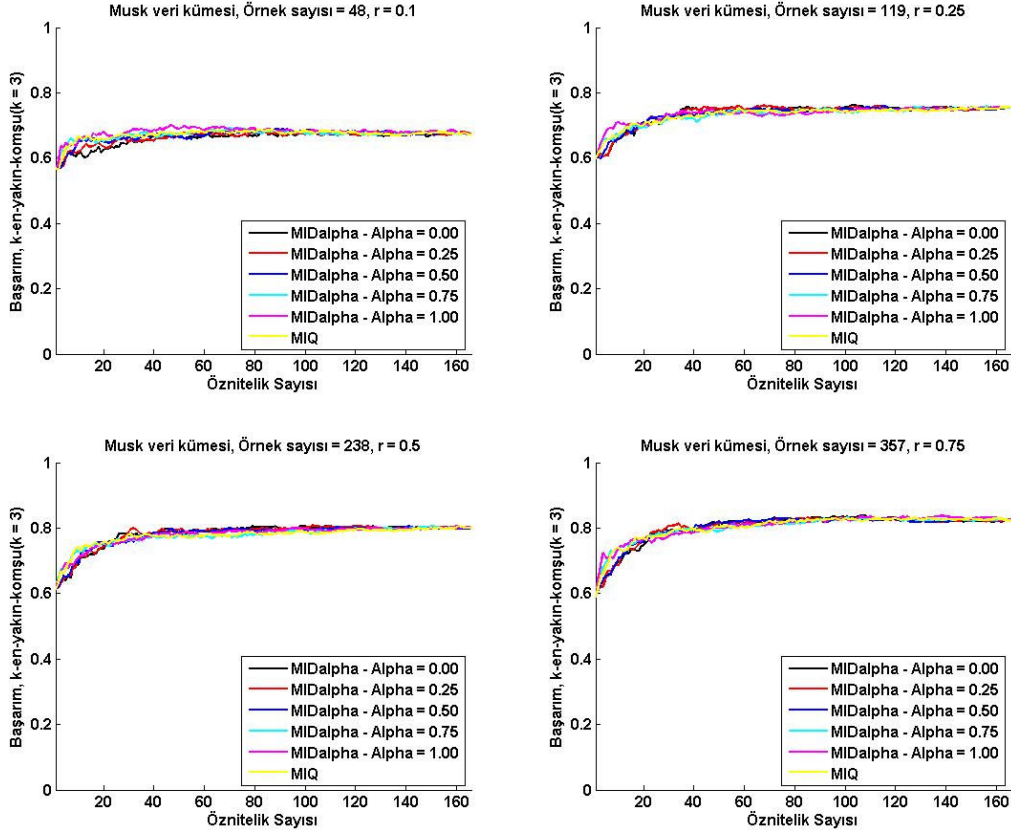


Şekil B.5 : MRMR, MID_α seçim kriterinin farklı α ve r değerleri için İyonosfer veri kümesi üzerinde başarımlar (svm) sonuçları.

Çizelge B.5 : MRMR, MID_α seçim kriterinin farklı α ve r değerleri için İyonosfer veri kümesi üzerindeki başarımların (svm) sonuçlarının ortalama sıralanma değerleri ve standart sapmaları.

r oranı	$\alpha = 0$	$\alpha = 0.25$	$\alpha = 0.5$	$\alpha = 0.75$	$\alpha = 1$
$r = 0.1$	4.6 ± 1.0	3.0 ± 1.0	2.5 ± 0.8	1.9 ± 1.2	2.5 ± 1.3
$r = 0.25$	4.8 ± 0.7	2.7 ± 0.9	3.0 ± 0.9	2.6 ± 1.4	1.4 ± 0.5
$r = 0.5$	4.8 ± 0.7	3.1 ± 1.0	2.9 ± 0.8	2.5 ± 1.1	1.2 ± 0.4
$r = 0.75$	4.8 ± 0.7	3.2 ± 0.9	2.9 ± 0.8	2.4 ± 1.0	1.2 ± 0.5

EKB.6

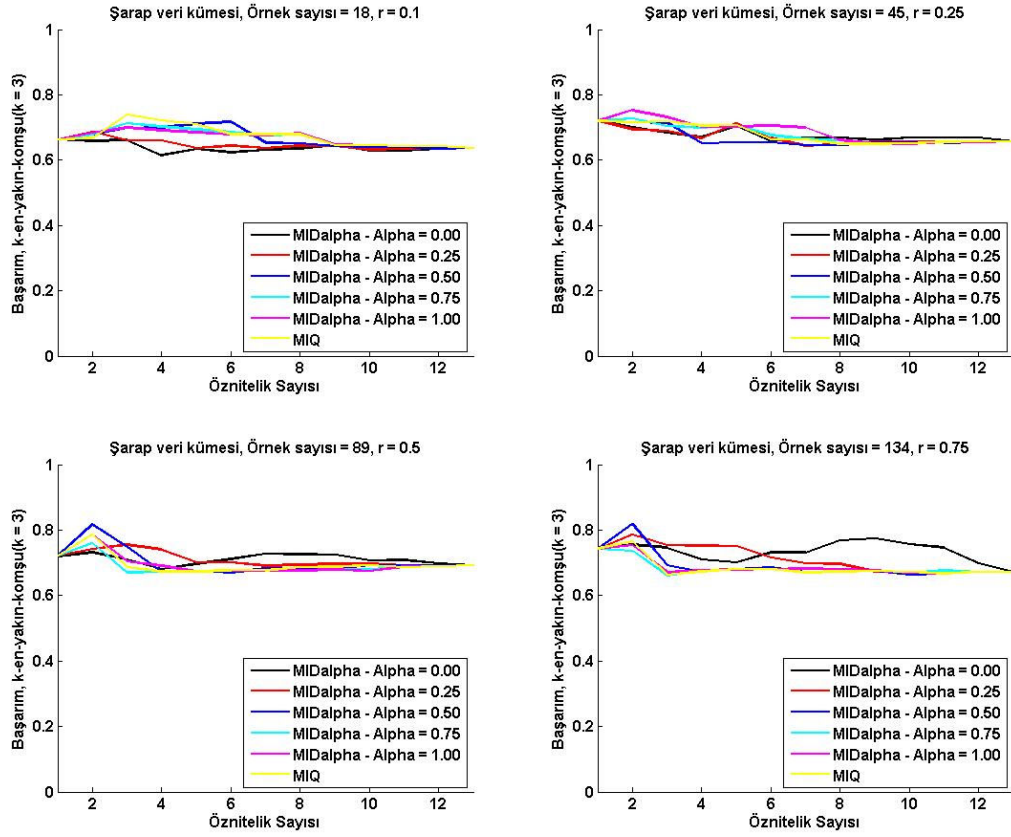


Şekil B.6 : MRMR, MID_α seçim kriterinin farklı α ve r değerleri için Musk veri kümesi üzerinde başarımlar (k-en-yakın-komşu) sonuçları.

Çizelge B.6 : MRMR, MID_α seçim kriterinin farklı α ve r değerleri için Musk veri kümesi üzerindeki başarımlar (k-en-yakın-komşu) sonuçlarının ortalama sıralanma değerleri ve standart sapmaları.

r oranı	$\alpha = 0$	$\alpha = 0.25$	$\alpha = 0.5$	$\alpha = 0.75$	$\alpha = 1$
$r = 0.1$	4.2 ± 1.1	1.3 ± 0.6	1.9 ± 0.6	3.3 ± 0.8	4.1 ± 0.7
$r = 0.25$	4.3 ± 1.0	1.1 ± 0.3	2.0 ± 0.4	3.3 ± 0.5	4.2 ± 0.8
$r = 0.5$	4.5 ± 0.8	1.1 ± 0.4	1.9 ± 0.3	3.3 ± 0.7	4.1 ± 0.7
$r = 0.75$	4.9 ± 0.4	1.1 ± 0.4	2.0 ± 0.5	3.2 ± 0.7	3.7 ± 0.6

EK B.7

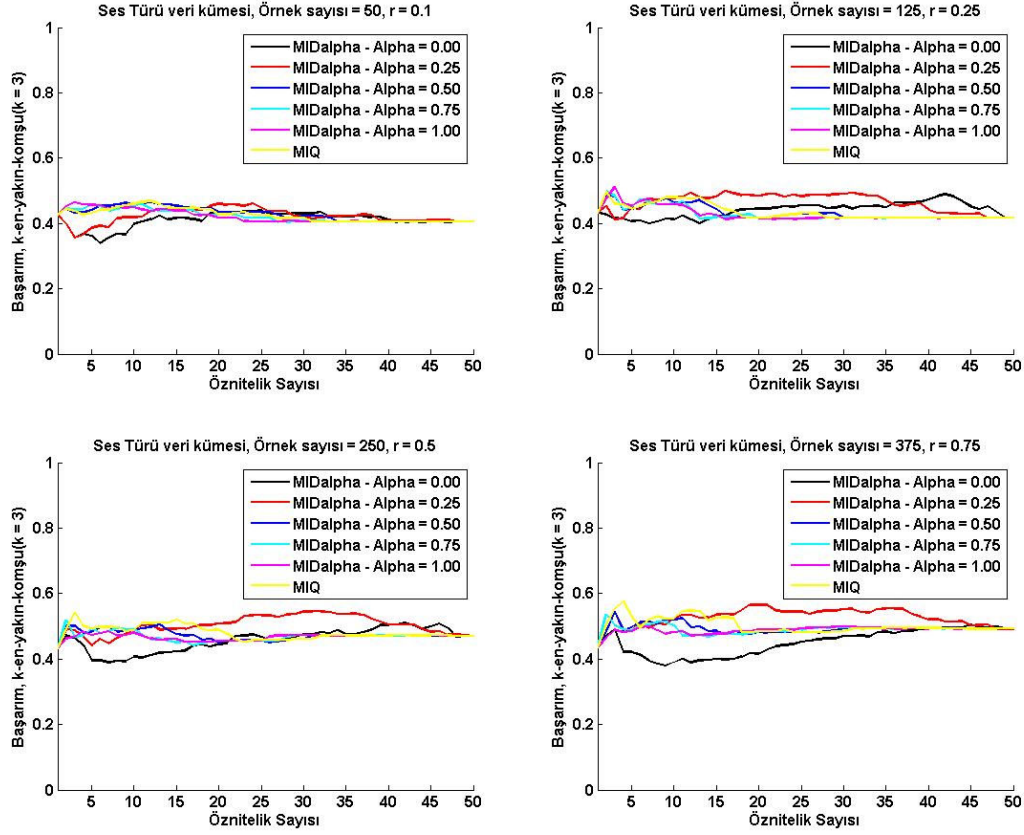


Şekil B.7 : MRMR, MID_α seçim kriterinin farklı α ve r değerleri için Şarap veri kümesi üzerinde başarımlar (k-en-yakın-komşu) sonuçları.

Çizelge B.7 : MRMR, MID_α seçim kriterinin farklı α ve r değerleri için Şarap veri kümesi üzerindeki başarımlar (k-en-yakın-komşu) sonuçlarının ortalama sıralanma değerleri ve standart sapmaları.

r oranı	$\alpha = 0$	$\alpha = 0.25$	$\alpha = 0.5$	$\alpha = 0.75$	$\alpha = 1$
$r = 0.1$	3.5 ± 1.6	3.6 ± 1.5	2.8 ± 1.1	2.1 ± 0.9	1.8 ± 0.8
$r = 0.25$	4.0 ± 1.3	3.9 ± 1.1	2.6 ± 1.0	1.5 ± 0.5	1.8 ± 0.7
$r = 0.5$	4.1 ± 1.1	3.8 ± 1.3	2.5 ± 0.8	1.5 ± 0.6	1.4 ± 0.6
$r = 0.75$	4.4 ± 1.1	3.2 ± 1.2	2.6 ± 1.1	1.5 ± 0.8	1.6 ± 0.6

EKB.8

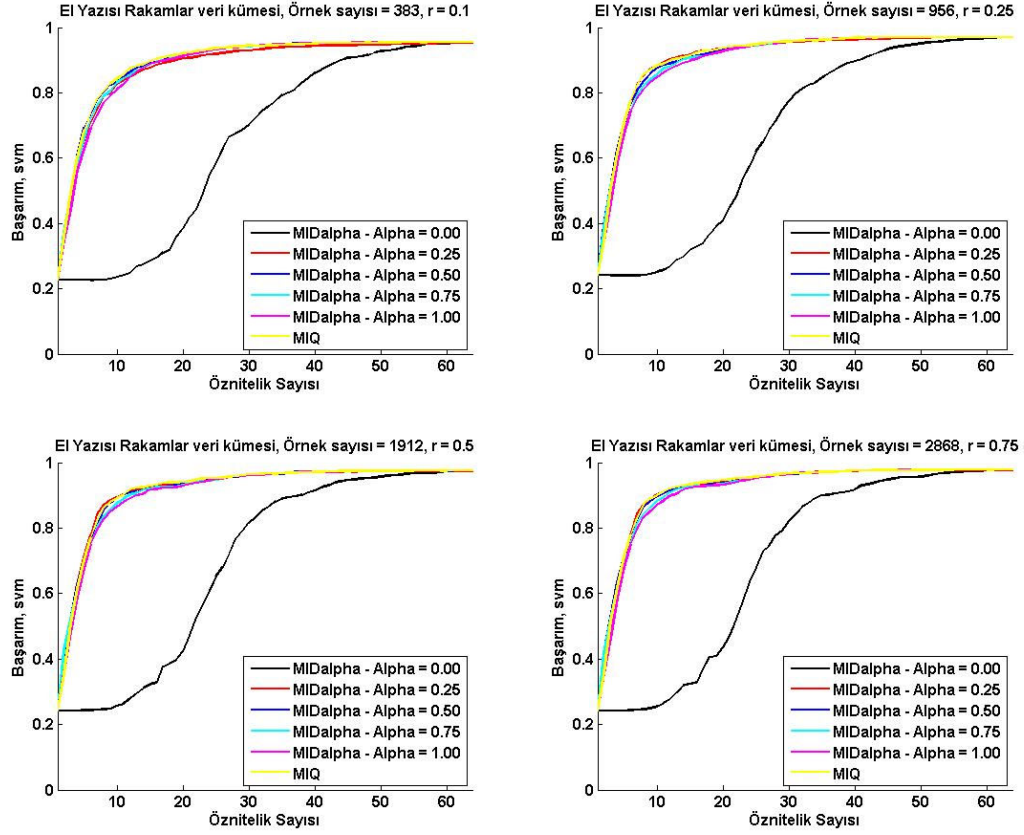


Şekil B.8 : MRMR, MID_α seçim kriterinin farklı α ve r değerleri için Ses Türlü veri kümesi üzerinde başarımlar (k-en-yakın-komşu) sonuçları.

Çizelge B.8 : MRMR, MID_α seçim kriterinin farklı α ve r değerleri için Ses Türlü veri kümesi üzerindeki başarımlar (k-en-yakın-komşu) sonuçlarının ortalama sıralanma değerleri ve standart sapmaları.

r oranı	$\alpha = 0$	$\alpha = 0.25$	$\alpha = 0.5$	$\alpha = 0.75$	$\alpha = 1$
$r = 0.1$	4.7 ± 0.8	3.8 ± 0.7	3.1 ± 0.5	1.7 ± 0.5	1.3 ± 0.5
$r = 0.25$	4.8 ± 0.6	3.9 ± 0.7	2.9 ± 0.4	1.7 ± 0.6	1.4 ± 0.5
$r = 0.5$	4.9 ± 0.6	3.9 ± 0.6	2.5 ± 0.8	1.8 ± 0.8	1.7 ± 0.7
$r = 0.75$	4.9 ± 0.6	3.8 ± 0.7	2.3 ± 0.9	1.8 ± 0.8	1.9 ± 0.8

EK B.9

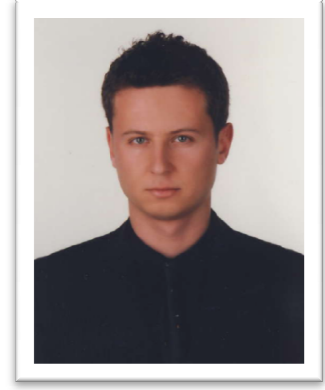


Şekil B.9 : MRMR, MID_{α} seçim kriterinin farklı α ve r değerleri için El Yazısı Rakamlar veri kümesi üzerinde başarımlar (svm) sonuçları.

Çizelge B.9 : MRMR, MID_{α} seçim kriterinin farklı α ve r değerleri için El Yazısı Rakamlar veri kümesi üzerindeki başarımlar (svm) sonuçlarının ortalama sıralanma değerleri ve standart sapmaları.

r oranı	$\alpha = 0$	$\alpha = 0.25$	$\alpha = 0.5$	$\alpha = 0.75$	$\alpha = 1$
$r = 0.1$	4.3 ± 0.8	4.1 ± 1.1	2.0 ± 0.9	1.9 ± 0.8	1.9 ± 0.9
$r = 0.25$	4.4 ± 0.9	3.8 ± 1.0	2.0 ± 0.9	2.0 ± 0.9	1.7 ± 0.8
$r = 0.5$	4.6 ± 0.8	3.1 ± 1.1	2.3 ± 1.1	2.0 ± 0.8	1.8 ± 1.0
$r = 0.75$	4.5 ± 0.8	3.1 ± 1.0	2.0 ± 1.1	1.9 ± 0.9	1.9 ± 0.9

ÖZGEÇMİŞ



Ad Soyad: Gökhan Gülgezen

Doğum Yeri ve Tarihi: Tekirdağ, 22.09.1984

Lisans Üniversitesi: İzmir Yüksek Teknoloji Enstitüsü, Bilgisayar Mühendisliği

Yayın Listesi:

- **G. Gulgezen**, Z. Cataltepe and L. Yu, 2009: "Stable and Accurate Feature Selection", *The European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD)*, 7-11 September 2009, Bled, Slovenia.
- **G. Gülgezen**, Z. Çataltepe and L. Yu, 2009: "Stable Feature Selection Using MRMR Algorithm / MRMR Algoritması Kullanılarak Kararlı Öznitelik Seçimi", *17th IEEE Conference on Signal Processing and Communications Applications (SIU 2009)*, 9-11 April 2009, Antalya, Turkey.
- B. Senliol, **G. Gulgezen**, L. Yu, Z. Cataltepe, 2008: "Fast Correlation Based Filter (FCBF) with a Different Search Strategy", poster presentation at *ISCIS 2008 Conference*, 27-29 October 2008, Istanbul, Turkey.