

# MRMR Algoritması Kullanılarak Kararlı Öznitelik Seçimi

## Stable Feature Selection Using MRMR Algorithm

Gökhan Gülgezen<sup>1</sup>, Zehra Çataltepe<sup>1</sup>, Lei Yu<sup>2</sup>

<sup>1</sup> Bilgisayar Mühendisliği Bölümü  
İstanbul Teknik Üniversitesi  
{gulgezen,cataltepe}@itu.edu.tr

<sup>2</sup> Binghamton Üniversitesi  
Bilgisayar Bilimleri Bölümü  
lyu@cs.binghamton.edu

### Özetçe

Öznitelik seçme yöntemleri girdi boyutunu azaltmaları ve ilgisiz, artık veriyi yok etmeleri gibi özellikleriyle makine öğrenmesi algoritmalarının daha hızlı çalışmalarına ve daha başarılı sonuçlar elde etmelerine olanak sağlar. MRMR (Minimum Redundancy Maximum Relevance, [1]) ve FCBF (Fast Correlation-Based Filter, [2]) gibi entropi tabanlı öznitelik seçme yöntemleri yüksek başarımları ve hızlı çalışmaları nedeniyle tercih edilen öznitelik seçme yöntemleri arasındadırlar. Öznitelik seçme yöntemlerinin iyiliğini ölçen başka bir başarımları ölçütü de kararlılık (stability), yani veri kümesi değişince de aynı ya da yakın özniteliklerin seçilmesidir. Biz bu çalışmada yeni bir kararlılık ölçütünde ortaya koyarak değişik veri kümeleri üzerinde MRMR öznitelik seçme yönteminin başarımları ve kararlılığı ile ilgili ölçümler yaptık. MRMR'ı çalıştırırken kullanılan iki yöntem olan MID ve MIQ yöntemlerinden, MID'nin küçük veri kümelerinde daha kararlı bir öznitelik seçimi yöntemi olduğunu gözlemledik.

### Abstract

Feature selection methods help machine learning algorithms produce faster and more accurate solutions, because they reduce the input dimensionality and they can eliminate irrelevant or redundant features. Entropy based feature selection algorithms, such as MRMR (Minimum Redundancy Maximum Relevance, [1]) and FCBF (Fast Correlation-Based Filter, [2]) are preferred feature selection methods because they are very fast and produce sets of features that result in quite accurate classifiers. Besides accuracy, stability is another measure of goodness for a feature selection algorithm. A feature selection algorithm is said to be stable if changes in the identity of data points available for feature selection still result in the same or similar sets of features. In this study, we first developed a new stability measurement and performed accuracy and stability measurements of MRMR when it is used on different data sets. We found out that, the two feature selection methods within MRMR, MID and MIQ result in features with similar accuracy. On the other hand, MID results in more stable feature sets than MIQ and therefore should be preferred over MIQ, especially for small number of available samples.

### 1. Giriş

Geçtiğimiz son on yılda sınıflandırma algoritmalarının üstünde uygulandığı veri kümelerindeki öznitelik sayıları binleri onbinleri bulmaktadır. Bu nedenle araştırmacılar öznitelik

seçme yöntemlerine her zamankinden daha fazla ihtiyaç duymakta ve dosya sınıflandırmadan sinyal işleme, biyoinformatikten örüntü tanıma işlemlerine kadar pek çok alanda kullanım yöntemlerini aramaya koyulmuşlardır.

Özellikle korelasyonları yüksek bir çok öznitelik bulunduğu ve örnek sayısının az olduğu veri kümelerinde öznitelik algoritmalarının önemi bir kat daha artmaktadır. Bu durumlarda öznitelik seçme algoritmaları hem veri kümesi içindeki gürültülü, sapkın ve gereksiz öznitelikleri eleyerek verilerin daha sağlıklı ifade edilmesini sağlarlar hem de düşük örnek sayılarında dahi sınıflandırıcı algoritmalarının yüksek başarı oranları göstermelerine katkıda bulunabilirler.

Öznitelik seçme algoritmalarının bu işlevleri yerine getirebilmesi için temelde aynı dağılımdan gelen farklı sayıda örnek içeren veri kümelerinde dahi tutarlı, kararlı seçimler yapması beklenmektedir.

Biz bu çalışmamızda basit ama etkili bir öznitelik seçme yöntemi olan MRMR([1]) algoritmasının kararlılığı (stability) üzerinde teorik ve deneysel sonuçlar sunacağız. MRMR algoritmasının ayrı veri kümeleri üzerinde uygulanmak üzere geliştirdiği iki yöntemden MID (Mutual Information Difference) ve MIQ (Mutual Information Quotient) arasındaki farkı vurgulayarak yapılan deneyler sonucunda MID tekniğinin daha kararlı sonuçlar verdiğini göstereceğiz.

Bildirinin 2. bölümünde MRMR algoritması ve MIQ/MID teknikleri anlatılacaktır. 3. bölüm kararlılık ölçütüne ayrılmış olup 4. bölüm deneylerde kullanılan veri kümeleri hakkında bilgi vermektedir. 5. bölüm deney düzenini açıklamaya çalışmaktadır. 6. ve 7. bölümlerde bu deneylerin sonuçları ve çıkarımlar okuyucuya sunulmuştur.

### 2. MRMR Algoritması

MRMR algoritması sınıf etiketleriyle en ilişkili (relevant) öznitelikleri seçmeye çalışırken eş zamanlı olarak seçilen öznitelikler arasındaki artıklığı, fazlalığı minimize etmeye çalışan bir filtreleme yöntemidir [1]. Bu bölümde MRMR algoritmasının detayları sunulacaktır. Algoritma her bir öznitelikli ayrı bir rastlantı değişkeni olarak görür ve iki öznitelik,  $X$  ve  $Y$ , arasındaki benzerlik seviyesini ölçmek için aralarındaki ortak bilgiyi (mutual information),  $I(X, Y)$ , kullanır:

$$I(X, Y) = \sum_{y \in Y} \sum_{x \in X} p(x, y) \log \left( \frac{p(x, y)}{p_1(x)p_2(y)} \right) \quad (1)$$

Yukarıdaki denklemde  $p(x, y)$   $X$  ve  $Y$  nin birleşik olasılık dağılım fonksiyonunu,  $p_1(x)$  ve  $p_2(y)$  ise sırasıyla  $X$  ve  $Y$  rastlantı değişkenlerinin diğer bir deyişle özniteliklerinin marjinal olasılık dağılım fonksiyonlarını göstermektedir.

Gösterimde kolaylık sağlamak amacıyla, her bir öznelik  $f_i$ ,  $N$  tane gözlemin bir araya gelerek oluşturduğu bir vektör olarak tanımlanacaktır ( $f_i = [f_i^1, f_i^2, f_i^3, \dots, f_i^N]$ ).  $f_i$  ise  $F_i$  ayrık rastlantı değişkeninin bir örneği olarak ele alınacak ve öznelik  $i$  ve  $j$  arasındaki ortak bilgi  $I(F_i, F_j)$  olarak gösterilecektir. Burada  $i, j = 1, 2, \dots, d$ ,  $d$  ise veri kümesinin boyut sayısını diğer bir deyişle veri kümesindeki öznelik sayısını göstermektedir. Ortak bilgi kavramı sadece iki öznelik arasındaki benzerlik düzeyini değil herhangi bir öznelik  $i$  ile sınıf etiketleri vektörü,  $h$  ( $h = [h^1, h^2, \dots, h^N]$ ), arasındaki benzerliğin ölçülmesinde de kullanılmaktadır,  $I(H, F_i)$ .

$S$  seçilmek istenen öznelik kümesi olsun ve  $|S|$  ise bu kümenin eleman sayısını göstereyin. Seçilmek istenen kümenin seçilebilecek en iyi küme olmasını garanti etmek için iki koşulun karşılanması gerekir. Bunlardan ilki *minimum artıklık* (*minimum redundancy*) koşuludur:

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

Diğeri ise *maksimum ilişki* (*maximum relevance*) koşuludur:

$$\max V, \quad V = \frac{1}{|S|} \sum_{F_i \in S} I(F_i, H), \quad (3)$$

Bu iki koşulu birleştiren iki basit kombinasyon ise şu şekilde ifade edilebilir:

$$\max(V - W), \quad (4)$$

$$\max(V/W), \quad (5)$$

Yukarıdaki ifadelere göre en iyi kümenin elde edilmesi  $O(N^{|S|})$  arama gerektirdiği için, MRMR algoritması bu optimizasyon sorununu şu şekilde çözer. Öncelikle ilk öznelik Eşitlik (3)'e göre seçilir. Bundan sonraki her basamakta Eşitlik (6) ve (7)'yi sağlayan öznelik  $i$  seçilir ve seçilen öznelik  $S$  kümesinde tutulur.  $\Omega_S = \Omega - S$  seçilmiş öznelikler dışında kalan tüm öznelikleri ifade etmektedir.

$$\max_{F_i \in \Omega_S} I(H, F_i), \quad (6)$$

$$\min_{F_i \in \Omega_S} \frac{1}{|S|} \sum_{F_j \in S} I(F_i, F_j), \quad (7)$$

Eşitlik (6) ve (7)'nin eşitlik (4) ve (5)'e göre kombine edilmesi algoritma için iki seçim kriterini ortaya çıkarmaya yeter:

Tablo 1: MRMR algoritmasının öznelik seçme kriterleri

İsim	Formül
Ortak Bilgi Farkı (MID)	$\max_{F_i \in \Omega_S} \left[ I(F_i, H) - \frac{1}{ S } \sum_{F_j \in S} I(F_i, F_j) \right] \quad (8)$
Ortak Bilgi Oranı (MIQ)	$\max_{F_i \in \Omega_S} \left[ I(F_i, H) / \frac{1}{ S } \sum_{F_j \in S} I(F_i, F_j) \right] \quad (9)$

Görüldüğü gibi Eşitlik (6) Eşitlik (3)'te belirtilen koşulun aynısıdır ve Eşitlik (7) Eşitlik (2)'nin bir benzeridir. Yukarıda geçen algoritmaların karmaşıklığı  $O(|S| \cdot N)$ .

### 3. Kararlılık Hesabı

Öznelik seçme algoritmalarının kararlılığını (stability) [3] ölçmek için iki öznelik alt kümesi arasındaki kararlılık seviyesini belirleyebilecek bir ölçüte ihtiyaç vardır. Bu bölümde [3] tarafından sunulan yöntemle benzer bir yöntem sunulacaktır. Varsayalım  $R_1 = \{F_i\}_{i=1}^{R_1}$  ve  $R_2 = \{F_j\}_{j=1}^{R_2}$  öznelik seçme algoritmalarının sonucu olan iki öznelik kümesi olsun ve her bir  $F_i$  ve  $F_j$  tek bir özneliği temsil etsin.  $R_1$  ve  $R_2$  arasındaki kararlılığı ölçmek amacıyla [3]  $R_1$  ve  $R_2$  'nin ağırlıklı tam iki parçalı çizge,  $G = (V, E)$ , olarak modellenmesini önermekte,  $V = R_1 \cup R_2$ ,  $E = \{(F_i, F_j) | F_i \in R_1, F_j \in R_2\}$ , ayrıca öznelikler arasındaki her bir bağlantının  $(F_i; F_j)$  bir ağırlık ölçütü  $\omega(F_i, F_j)$  ile ilişkilendirilmesi gerektiğini belirtmektedir. Bizim yöntemimizde iki nokta ya da diğer bir deyişle iki öznelik,  $F_i$  and  $F_j$ , arasındaki ağırlık,  $\omega(F_i, F_j)$ , bu öznelik çifti arasındaki simetrik belirsizlik ile ifade edilmektedir. Bu entropi tabanlı doğrusal olmayan korelasyona  $SU$  denir ve şu şekilde hesaplanır:

$$SU_{i,j} = 2 \left[ \frac{IG(F_i | F_j)}{H(F_i) + H(F_j)} \right], \quad (10)$$

Daha önce tanımlandığı gibi  $F_i$  and  $F_j$  iki ayrık rastlantı değişkenidir ve sırasıyla öznelik  $i$  ve öznelik  $j$  'yi temsil ederler.

$$IG(X | Y) = H(X) - H(X | Y), \quad (11)$$

$$H(X) = \sum_{x \in X} p(x) \log_2(p(x)), \quad (12)$$

$$H(X | Y) = \sum_{y \in Y} p(y) \sum_{x \in X} p(x | y) \log_2(p(x | y)), \quad (13)$$

Bizim metodumuzda  $\omega(F_i, F_j)$ ,  $SU_{i,j}$  'ye eşittir ve  $R_1$  ve  $R_2$  arasındaki genel benzerlik (kararlılık seviyesi) aşağıda gösterilen formül ile hesaplanmaktadır:

$$Sim^M(R_1, R_2) = \frac{\sum_{(F_i, F_j) \in M} \omega(F_i, F_j)}{|M|}, \quad (14)$$

Burada  $M$ ,  $G$  çizgesindeki maksimum eşleşmedir. Diğer bir deyişle  $M$  iki öznelik seti,  $R_1$  ve  $R_2$ , arasındaki maksimum ağırlıklı eşleşmeyi temsil eder. Maksimum ağırlıklı iki parçalı çizge eşleştirme problemi birçok şekilde çözülebilmektedir. Biz optimum sonuca ulaşmak için Hungarin algoritmasını kullandık. Algoritma Tennessee Üniversitesi'nden Alexander

Melin tarafından koda dökülmüş ve [Matlab Central](#) internet sitesinden elde edilmiştir.

İki öznitelik seti arasındaki benzerliği (kararlılığı) ölçmek için böyle bir yöntemin kullanılmasının nedeni özniteliklerin indekslerinden ziyade içerdikleri değerlere bakarak benzerlik ölçmenin daha doğru olduğunu saptamamızdır. 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.

Bizim kararlılık ölçütümüz [3]'te açıklanan yöntemden iki açıdan farklıdır. İlk olarak ilgili makalede tek tek özniteliklerden oluşan iki öznitelik kümesi arasındaki kararlılıktan ziyade öznitelik gruplarından oluşan öznitelik kümeleri arasındaki kararlılık ölçülmeye çalışılmaktadır. İkinci olarak ise çizgede öznitelik grupları arasındaki ağırlıklar bizim kullandığımız simetrik belirsizlik yönteminden farklı olarak her bir grubun merkezinde yer alan ya da grubu en iyi şekilde tanımlayan öznitelikler arasındaki korelasyon katsayısına bakılarak hesaplanmaktadır. Bizim metodumuz [3]'te açıklanan yöntemin özel bir halidir ve öznitelik grupları yerine tek tek öznitelikler arasındaki benzerlik yukarıda belirtilen simetrik belirsizlik yöntemiyle saptanmaktadır.

#### 4. Veri Kümeleri

Bu bölümde deneylerin gerçekleştirildiği veri kümeleri tanıtılacaktır. Veri kümelerinin üstünde MRMR algoritması MID ve MIQ teknikleriyle uygulanmadan önce tüm öznitelikler maximum ve minimum değerleri arasında 10 bölmeye ayrılarak ayrık rastlantı değişkenleri haline getirilmiştir.

##### 4.1. Musk(Versiyon 1) Veri Kümesi

Bu 47'si uzmanlar tarafından musk olarak nitelenen ve geri kalan 45 tanesinin musk olmayan diye tanımlandığı 92 molekülü barındıran bir veri kümesidir. Amaç yeni moleküllerin musk olup olmadıklarını saptamaktır. Bu veri kümesini üretmek amacıyla ilk olarak moleküllerin düşük enerjili biçimlenmeleri üretilmiştir, daha sonra bunlardan çok benzer olanlar atılmıştır. Bu işlemin sonunda geriye 476 biçimlenme kalmıştır. Son olarak bu 476 biçimlenme için 166 adet öznitelik çıkarılmıştır. Bu veri kümesi [UCI Machine Learning Repository](#) internet sitesinden alınmıştır [5].

##### 4.2. El Yazısı Rakamlar Veri Kümesi

El yazısı rakamların normalize edilmiş bit haritaları NIST tarafından sağlanan önışleme programlarını kullanarak çıkartılmıştır. Biz bu çalışmada veri kümesinin 43 insandan 30'unun katkıda bulunduğu eğitim setini kullandık. El yazısı rakamların 32x32 bit haritaları 4x4 lük üst üste gelmeyen bloklara bölünmüş ve her bir bloktaki dolu piksel adedi sayılmıştır. Bu her bir rakam için 1...16 aralığında tam sayı değerlerinden oluşan 8x8'lik bir veri matrisi elde edilmesine yol açmıştır. Veri kümesindeki örnek sayısı 3823'tür ve yukarıdan da anlaşılacağı gibi her bir örnek için 64 öznitelik tanımlıdır. Sınıf etiketleri rakamların gerçek tam sayı değerlerinden oluşmaktadır. Bu veri kümesi [UCI Machine Learning Repository](#) internet sitesinden alınmıştır [5].

Tablo 2:Veri kümeleri ve özellikleri

İsim	Örnek Sayısı	Öznitelik Sayısı	Sınıf Sayısı
Musk(Versiyon 1) Veri Kümesi	476	166	2{0, 1}
El Yazısı Rakam Veri Kümesi	3823	64	10{0,...,9}

### 5. Deneysel Metodoloji

Bir öznitelik seçme algoritmasının kararlılığı değişken eğitim kümelerinden elde edilmiş öznitelik kümeleri sonuçlarının ortalama benzerliği olarak tanımlanabilir. Her bir veri alt kümesi rastgele seçim yada bootstrap algoritmasının sonucu olarak oluşturulabilir.

$N$  tane örnek içeren bir  $D$  veri kümesi kullanılarak öznitelik seçme algoritmasının kararlılığı ölçülecek olsun. Bunu gerçekleştirmek için ilk olarak deney kümelerini oluşturmak gerekir. Bu çalışmada tüm veri kümesinden 10 kez  $N$  örnek içerecek şekilde veri bootstrap algoritması kullanılarak çekildi. Bu işlemin sonucunda 10 eğitim kümesi,  $D_{train,i}$ ,  $i = 1, 2, \dots, q$ ,  $q = 10$  oluşturuldu. Her bootstrap işleminden sonra  $D_{train,i}$  veri kümesine ait olmayan tüm örnekler  $D_{test,i}$  veri kümesinde tutuldu ve bu veri kümeleri başarımlarının hesaplanması için kullanıldı. Algoritmanın kararlılığını göstermek için ilk olarak öznitelik seçme algoritması tüm eğitim veri kümelerinin üzerinde uygulanarak, öznitelik kümeleri,  $R_i$ 'ler elde edildi. Öznitelik seçme algoritmasının kararlı davranmasının, eğitim kümesinin örnek sayısındaki azalmaya rağmen tüm veri kümesi kullanılarak elde edilen öznitelik kümesi olan  $R_i$  ile benzer, tutarlı öznitelik kümeleri üretmesi olarak tanımlanmıştır. Bu amaçla her bir eğitim veri kümesinden  $(D_{train,i}, r * |D_{train,i}|)$  ( $r = [r^1, r^2, \dots, r^j]$ ), sayıda örnek alınarak bu veri kümelerinde MRMR algoritması uygulandı ve  $R_i^j$  öznitelik kümeleri elde edildi. Sonuç olarak her bir  $r$  oranı için  $q$  adet  $R_i$  ve  $R_i^j$  öznitelik kümesi oluşturuldu.

Deneylerimizde, 5 tane  $r$  değeri kullanıldı,  $r = [0.1, 0.25, 0.5, 0.75, 1]$ . Sonuç olarak her bir  $r^j$  oranı için,  $q$  veri alt kümesinde algoritmanın kararlılığı şu şekilde hesaplandı:

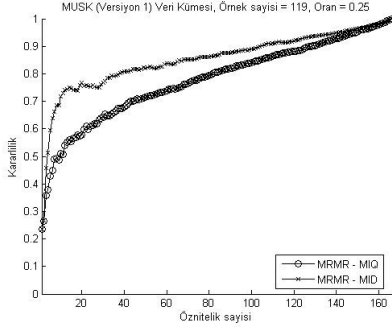
$$\frac{1}{q} \sum_{i=1}^q Sim^M(R_i, R_i^j), \quad (15)$$

Deneylerde kararlılık ve başarımlar hesaplamaları mevcut eğitim veri kümelerindeki örnek sayısını azaltarak ve bootstrap sayısını fazla tutmaya çalışılarak gerçekleştirildi. [3]'te ise kararlılık hesaplamaları sadece tek bir bootstrap yapılmış veri kümesi üzerinden gösterilmeye çalışılmış, bu da  $r$  değerinin yaklaşık olarak 0.632'ye eşit olduğu duruma karşılık gelmektedir ([4] p. 333). MID ve MIQ tekniklerinin kararlılık karşılaştırmaları ve elde edilen öznitelik kümelerinin  $D_{test,i}$  veri kümelerindeki başarımlarını Deney Sonuçları kısmında görülebilir.

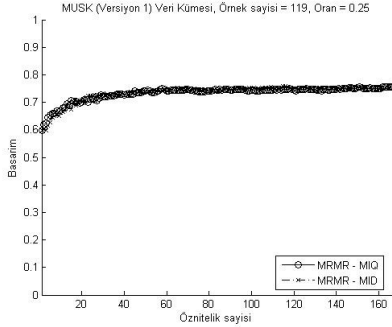
## 6. Deney Sonuçları

Deneylerde sınıflandırıcı algoritması olarak k-en-yakın-komşu (knn: k-nearest neighbor) sınıflandırıcısı kullanılmış ve k değeri 3 olarak alınmıştır. Hesaplamalar her iki veri kümesinde olası tüm öznelik sayıları için yapılmıştır.

Şekil 1 ve 2'de Musk (Versiyon 1) veri kümesi üzerinde, tüm verinin %25'i ( $r=0.25$ ) kullanılarak yapılan deney sonuçları gösterilmiştir. Şekil 1'de görüldüğü gibi, MID yöntemi ile elde edilen kararlılık değerleri bütün seçilen öznelik sayıları için MIQ'dan daha fazladır. Şekil 2'de görüldüğü gibi MID ve MIQ'nun başarımları arasında belirgin bir fark bulunmamaktadır. Yer darlığı dolayısıyla diğer  $r$  değerleri ve el yazısı rakam veri kümesindeki sonuçların grafikleri verilememiştir. Fakat onlarda da aynı durum gözlenmiştir.

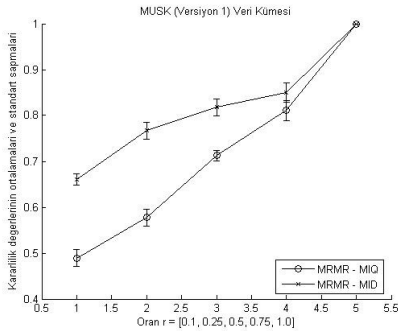


Şekil 1. Musk (Versiyon 1) MID ve MIQ kararlılık değerleri

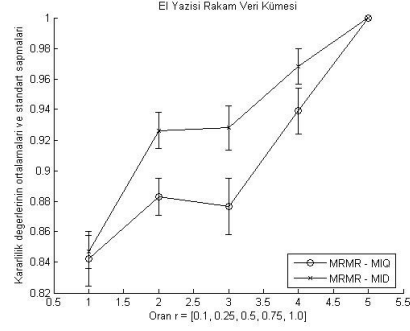


Şekil 2. Musk (Versiyon 1) MID ve MIQ başarımları

Şekil 3 ve 4'de, 20 öznelik seçilmesi durumunda MID ve MIQ yöntemlerinin 10 deneyde buldukları kararlılık değerleri bütün  $r$  değerleri için gösterilmiştir. Bu şekillerde de görüldüğü gibi, MID, MIQ'dan daha kararlı sonuçlar vermektedir.



Şekil 3. Musk (Versiyon 1) veri kümesinde MID ve MIQ'nun ortalama kararlılık değerleri (Öznelik sayısı = 20)



Şekil 4. El Yazısı Rakam veri kümesinde MID ve MIQ'nun ortalama kararlılık değerleri (Öznelik sayısı = 20)

## 7. Sonuçlar

Bu bildiriye, MRMR öznelik seçme yönteminin iki değişik uygulaması, MID ve MIQ, başarımları yönünden olduğu gibi, kararlılık yönünden de incelenmiştir. Kararlılığın değişen veri kümesi boyutlarında ölçülmesi için tüm veri yerine verinin değişik sayıda alt kümeleri alınarak öznelik seçimi yapılmış ve bu durum için bir kararlılık ölçme yöntemi geliştirilmiştir. MID yönteminin, özellikle veri kümelerinin küçük olduğu durumlarda, MIQ'dan daha kararlı olduğu görülmüştür.

## Teşekkür

Yazarlar SUNY Binghamton Üniversitesi Bilgisayar Bilimleri Bölümü'nden Prof. Lei Yu'ya kararlı öznelik seçimi konusundaki yardımlarından, Türkiye Bilimsel ve Teknolojik Araştırma Kurumu'na(TÜBİTAK) sağlanan burstan dolayı teşekkür ederler.

## 8. Kaynakça

- [1] C. Ding and H.C. Peng, "Minimum Redundancy Feature Selection from Microarray Gene Expression Data," Proc. Second IEEE Computational Systems Bioinformatics Conf., pp. 523-528, Aug. 2003.
- [2] Lee, K.-F., *Automatic Speech Recognition: The Development of the SPHINX SYSTEM*, Kluwer Academic Publishers, Boston, 1989.
- [3] L. Yu and H. Liu. "Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution". In Proceedings of The Twentieth International Conference on Machine Learning (ICML-03), pp 856-863, Washington, D.C., August 21-24, 2003.
- [4] Lei Yu, Chris Ding, and Steven Loscalzo. "Stable Feature Selection via Dense Feature Groups". to appear in Proceedings of The 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-08), Las Vegas, NV, August, 2008.
- [5] E. Alpaydin, Introduction to Machine Learning, The MIT Press, Cambridge, Massachusetts, 2004.
- [6] 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.