

Mikro-dizilim Veri Sınıflandırmasında Öznitelik Seçme Algoritmalarının Karşılaştırılması

Abdülkadir Gümüüşçü¹, İbrahim Berkan Aydılek², Ramazan Taşaltın³

^{1,3}Harran Üniversitesi, Mühendislik Fakültesi, Elektrik-Elektronik Mühendisliği Bölümü, Şanlıurfa.

²Harran Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, Şanlıurfa

e-posta: agumuscu@harran.edu.tr, berkanaydilek@harran.edu.tr, rtasaltin@harran.edu.tr

Geliş Tarihi: 07.06.2016

Kabul Tarihi: 26.07.2016

Özet

Mikro-dizilim verileri son zamanlarda kanser hastalığı teşhisinde sıklıkla kullanılmaktadır. Mikro-dizilim verilerinin makina öğrenme algoritmalarında eğitilmesi ve modellenmesi nitelik sayısının fazlalığından ve örnek sayısının azlığından dolayı oldukça zordur. Bu yüzden mikro-dizilim verilerinin değerlendirilmesinde öznitelik seçme algoritmaları büyük önem kazanmaktadır. Öznitelik seçme algoritmaları genel olarak filtre, sarmal ve gömülü olmak üzere üç ana başlıkta incelenmektedir. Bu çalışmada mikro-dizilim verilerini sınıflandırmada filtre modelli öznitelik seçme algoritmaları, sınıflandırma başarı oranı ve uygulanma süreleri bakımından karşılaştırılmıştır. Bu bağlamda 5 farklı filtre modelli öznitelik seçme algoritması, 11 farklı mikro-dizilim veri setine uygulanmıştır. Makina öğrenme algoritması olarak k-NN sınıflandırma algoritması uygulanmış ve 10-katlamalı çapraz yöntemi ile doğrulanmıştır.

Anahtar kelimeler: Mikro-dizilim veri seti; Sınıflandırma; Öznitelik Seçme.

Comparison of Feature Selection Algorithms on Microarray Data Classification

Abstract

Recently, microarray data sets are often used in cancer diagnosis. Training and modeling of microarray data sets is quite hard process because of high number of feature and low number of sample. For this reason feature selection is a pre-processing technique with great importance in microarray classification. In the literature, feature selection techniques in terms of classification can be examined under three titles as Filter, Wrapper, Embed. In this study, filter model feature selection algorithms are compared in terms of classification success rate and processing time on microarray data classification. In this context, 5 different filter model feature selection algorithms are applied to 11 different microarray datasets. The results are tested with k-fold cross validation by using k-nearest neighbor (k-NN) method.

Keywords: Microarray Data Set, Classification, Feature Selection

1. Giriş

Mikro-dizilim verileri son zamanlarda hastalık teşhisinde önemli ölçüde kullanılmaya başlanmıştır. Elde edilen veriler, makina öğrenme algoritmaları ile modellenmektedir. Fakat düşük örnek sayısı ve yüksek nitelik sayısı makina öğrenme algoritmalarının işini oldukça zorlaştırmaktadır. Bu dezavantajı gidermek amacı ile öznitelik algoritmaları kullanılmaktadır. Bu açıdan bakıldığında seçilen öznitelik seçme algoritması, mikro-dizilim veri sınıflandırılmasında çok önemli bir işlem adımı olarak karşımıza çıkmaktadır. Öznitelik seçme işlemi yüz tanıma [1], sosyal ağlarda sınıflandırma [2], duygu analizi [3], hastalık

teşhisi [4] ve bunun gibi birçok veri madenciliği alanında oldukça yaygın şekilde kullanılmaktadır.

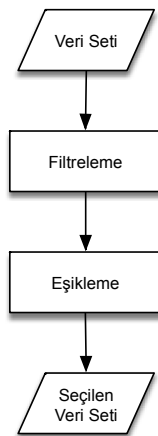
Öznitelik seçme algoritmaları literatürde genel olarak filtre modelli, sarmal modelli ve gömülü olmak üzere üç ana başlıkta incelenmektedir [5]. Filtre modelli öznitelik seçme algoritmalarının en büyük avantajı hızlı ve sınıflandırma algoritmasından bağımsız olmasıdır. Bunun yanında sınıflandırma başarı oranının düşük olması dezavantajdır. Sarmal modelli öznitelik seçme algoritmaları ise belirli bir düzende seçilen öznitelik kümesini değiştirerek sınıflandırma başarı oranı hesaplanır. Hesaplanan sınıflandırma başarı oranına göre eleme veya seçme işlemleri yapılarak

sınıflandırma başarı oranı yükseltilmeye çalışılır. Bu işlem iteratif olarak en ideal öznitelik kümesi buluncaya kadar devam ettirilir. Sarmal modellenli öznitelik seçme algoritmalarına genetik algoritmalar [6] örnek verilebilir. Sarmal modellenli öznitelik seçme algoritmalarının yüksek sınıflandırma başarı oranı sahip olması en büyük avantajıdır. Gömülü öznitelik seçme algoritmaları sınıflandırma algoritmasının içinde gömülü olarak bulunan öznitelik seçme algoritmalarıdır. Bu modele en uygun örnek karar ağaçlarıdır [7]. Literatürde ele alınan öznitelik seçme algoritmaları Tablo 1’de özetlenmiştir.

Öznitelik Seçme Modeli	Avantaj	Dezavantaj
Filtre Modelli	Hızlı, Sınıflandırma Algoritmasından Bağımsız	Başarı Oranı Diğer Modellere göre Düşük
Sarmal Modelli	Başarı Oranı Yüksek, Uygulanması Kolay	Yavaş, İşlem Yükü Fazla
Gömülü Modelli	İşlem Yükü Az, Başarı Oranı Yüksek	Karar Ağaçlarından Bağımsız olarak kullanılmaması

Tablo 1. Öznitelik Seçme Algoritmaları.

Filtre modellenli öznitelik seçme algoritmaları istatistiksel varsayımlar kullanılarak niteliklerin sonuca ulaşmaya katkısını hesaplayarak nitelikleri puanlar ve eşikleme yaparak öznitelikleri seçer. Şekil 1’de filtre modellenli öznitelik seçme algoritmalarının işlem adımları gösterilmektedir.



Şekil 1. Filtre Modelli Öznitelik Seçme Algoritması İşlem Adımları.

Tablo 1’de görüldüğü üzere filtre modellenli öznitelik seçme algoritmalarının avantajları, sınıflandırma algoritmasından bağımsız olması ve hızlı olması şeklinde listelenebilir. Literatürde en yaygın

kullanılan filtre modellenli öznitelik seçme algoritmaları, Korelasyon-bazlı öznitelik seçme (CFS) [8], bilgi kazancı (IG) [9], ortak bilgi [10], Ki-kare [11], ReliefF [12], ve F-skor [13] şeklinde sıralanmaktadır.

Bu çalışmada mikro-dizilim verilerini sınıflandırmada filtre modellenli öznitelik seçme algoritmaları, sınıflandırma başarı oranı ve uygulanma süreleri bakımından karşılaştırılmıştır. Bu bağlamda 5 farklı filtre modellenli öznitelik seçme algoritması, 11 farklı mikro-dizilim veri setine uygulanmıştır. Makina öğrenme algoritması olarak k-NN sınıflandırma algoritması uygulanmış ve 10-katlamalı çapraz yöntemi ile doğrulanmıştır.

Bu çalışmada kullanılan mikro-dizilim veri setleri <http://www.gems-system.org> [14] sitesinden indirilerek kullanılmıştır.

2. Materyal ve Metot

Bu çalışmada 5 farklı filtre modellenli öznitelik seçme algoritması işlem süresi ve sınıflandırma başarı oranı açısından karşılaştırılmıştır. Bunlar Ki-Kare, ReliefF, F-Skor, Korelasyon-bazlı öznitelik seçme (CFS) ve Bilgi Kazancı (IG) olarak sıralanmaktadır. Bu kısımda karşılaştırılacak filtre modellenli öznitelik seçme algoritmaları ve test aşamasında kullanılan k-NN sınıflandırma algoritması özetlenecektir.

2.1. Ki-Kare Öznitelik Seçme

Ki-kare istatistik temelli olup öznitelik seçme işlemlerinde yaygın şekilde kullanılmaktadır. Ki-kare istatistik değeri her nitelik değeri için formül (1) ile hesaplanır.

$$\chi^2 = \sum_{i=1}^m \sum_{j=1}^k \frac{(A_{ij} - \frac{R_i C_j}{N})^2}{\frac{R_i C_j}{N}} \quad (1)$$

Yukarıdaki formülde m verilen aralık sayısı, k sınıf sayısını, R_i i ’nci aralıktaki örnek sayısını, C_j j ’nci sınıftaki örnek sayısını, A_{ij} j ’nci sınıfta ve i ’nci aralıktaki örnek sayısını, N ise veri setinde bulunan toplam örnek sayısını temsil etmektedir.

Bu metot buldukları sınıfa göre tüm niteliklerin Ki-kare’sini hesaplayarak tek tek değerlendirir [11].

2.2. ReliefF öznitelik seçme

ReliefF, Relief istatistiksel modelinin geliştirilmiş versiyonudur. ReliefF metodu, veri setinden bir örnek ele alarak ilgili örneğin, kendi sınıflarındaki diğer örneklerle yakınlığını ve farklı sınıflarla olan uzaklığına bağlı bir model oluşturarak öznitelik seçme işlemini gerçekleştirmektedir [15]. Formül (2)'de ReliefF hesaplama denklemi verilmiştir.

$$S_i = \frac{\sum_{j=1}^m -fark(x_{ij}, \text{enyakin_ayni}_{ij}) + fark(x_{ij}, \text{enyakin_farkli}_{ij})}{m} \quad (2)$$

Formül (2)'de m veri setinde bulunan örnek sayısını, $fark(x_{ij}, \text{enyakin_ayni}_{ij})$ j 'nci örnekte bulunan i 'nci niteliğin aynı sınıfa sahip en yakın örneğe olan uzaklığını, $fark(x_{ij}, \text{enyakin_farkli}_{ij})$ ise j 'nci örnekte bulunan i 'nci niteliğin farklı sınıfa sahip en yakın örneğe olan uzaklığını, S_i ise i 'nci niteliğin ReliefF değerini belirtmektedir.

2.3. F-Skor öznitelik seçme

Fisher-skor filtresi her niteliğin diğer niteliklerden farkını hesaplayan bir filtre modelidir. Dolayısıyla diğer niteliklerden farkı fazla olan nitelik sınıflandırma algoritmaları için daha önemlidir. Bundan dolayı Fisher-skor filtresinin sonucunda yüksek değere sahip nitelikler daha ilişkili niteliklerdir [12]. İki sınıf olma durumunda her bir i 'nci nitelik için formül (3)'deki gibi F-skor hesaplanır.

$$F_i = \frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2} \quad (3)$$

Yukarıdaki formülde F_i i 'nci niteliğin f-skor değerini, μ_k k 'nci sınıfa ait i 'nci niteliklerin ortalama değerlerini, σ_k^2 ise k 'nci sınıfa ait i 'nci niteliklerin standart sapma değerlerini temsil eder.

2.4. Korelasyon-bazlı öznitelik seçme (CFS)

CFS, 1999 yılında Hall tarafından geliştirilmiştir [8]. CFS, alt öznitelik kümelerini korelasyon-bazlı değerlendirerek en iyi alt öznitelik kümeyi bulmayı hedefleyen filtre modeli öznitelik seçme algoritmalarından biridir. Temel prensip olarak kendi aralarında korelasyonu az, sınıf etiketleri ile korelasyonu fazla öznitelik alt kümesini seçmeye

çalışan bir algoritmadır. Her bir alt küme için hesaplanan CFS katsayısı denklem (4)'de verilmiştir.

$$Skor_s = \frac{k\bar{\gamma}_{cf}}{\sqrt{k+k(k-1)\bar{\gamma}_{ff}}} \quad (4)$$

Denklem (4)'de s , k adet öznitelige sahip öznitelik alt kümesini, $\bar{\gamma}_{cf}$ öznitelik alt kümesinin sınıf ile korelasyonunu ve $\bar{\gamma}_{ff}$ ise öznitelik alt kümesindeki öznitelikler arasındaki korelasyonu temsil etmektedir.

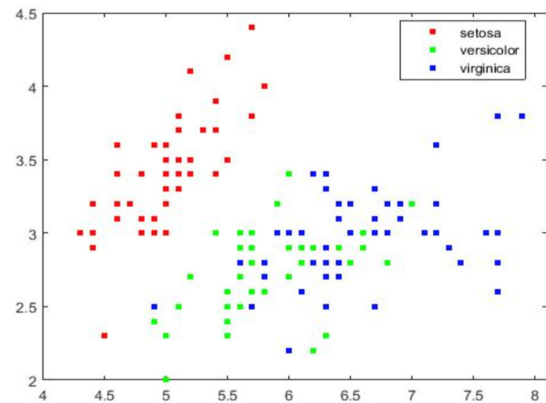
2.5. Bilgi Kazancı öznitelik seçme

Bilgi kazancı, karar ağaçlarında ilk dallanmanın hangi nitelikten başlanmasını belirleyen ölçütlerden biridir. Bunun yanında filtre modeli öznitelik seçme işlemlerinde de yaygın olarak kullanılmaktadır. Öznitelik seçme prensibi, sınıflara ait en çok bilgiye sahip öznitelik kümesini bulmaya dayanır. Entropi tabanlı bir öznitelik seçme algoritması olup her nitelik için IG katsayısı hesaplanıp, en yüksek IG katsayısına sahip öznitelik kümeleri seçilmektedir [9].

2.6. k-NN Sınıflandırma

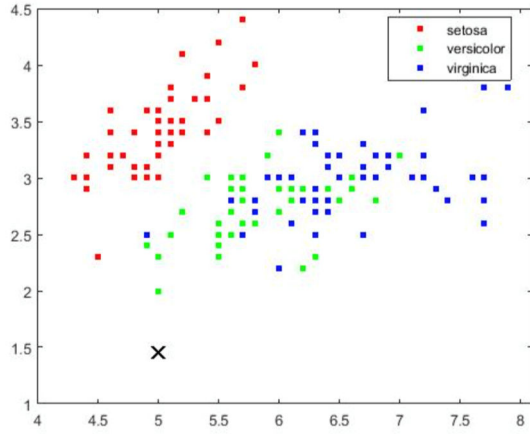
k-NN sınıflandırma algoritması birçok alanda sıklıkla kullanılan sınıflandırma metotlarının başında gelmektedir [16]. k-NN sınıflandırmasının önemli avantajlarından olan basitliği ve kullanım kolaylığı yanında büyük veri setlerinde sağladığı kararlılık da kullanımını yaygınlaştırmıştır.

Şekil 2'de gösterildiği gibi k-NN sınıflandırma algoritması uygulanırken eğitim verisindeki örnekler düzleme yerleştirilerek sınıfları ile ilişkilendirilir.



Şekil 2. Eğitim Setinin Düzlemde Görünümü

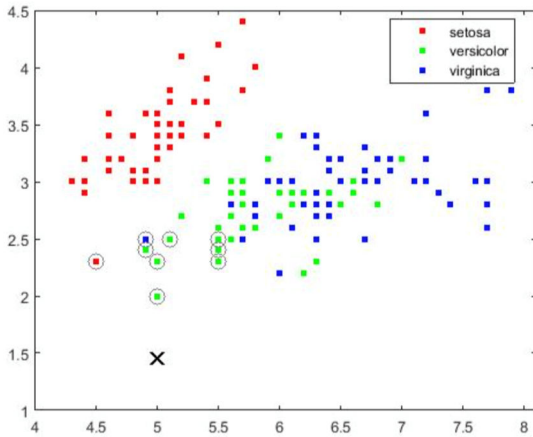
Şekil 3’de görüldüğü gibi kurulan düzleme test için kullanılacak örnekler düzleme yerleştirilir.



Şekil 3. Test Veri Girişi Görünümü

Şekil 3’de görüldüğü üzere eğitim verisi ile oluşturulan düzleme X işareti gösterilen test verisi eklenmiştir.

Daha sonra Şekil 4’de de gösterildiği gibi k değeri kadar test verisinin en yakın komşularına bakılarak test verisinin sınıfı belirlenir. Test verisinin sınıfı, komşu örnekler en çok hangi sınıfa ait ise ilgili sınıfa ait olduğu tahmin edilir.



Şekil 4. Test Veri Girişi Görünümü

Şekil 4’deki örnekte k değeri 9 olduğundan test verisine en yakın 9 adet örnek çember içine alınmış olup, çember içine alınan örneklerden en fazla sınıfa sahip versicolor sınıfı olduğundan test verisinin de sınıfı versicolor olarak tahmin edilir.

Bu çalışmada k değeri 1 olarak seçilip, komşuluk uzaklık hesabı öklid uzaklığa göre hesaplanmıştır.

2.5. Mikro-dizilim Veri Seti

Mikro-dizilim teknolojileri son zamanlarda özellikle hastalık teşhisinde oldukça yaygın kullanılmaktadır. DNA mikro-dizilim teknolojisi, dokulardaki gen ifade profillerindeki global değişikliklerin incelenmesinde kullanılan yeni ve güçlü bir teknolojidir [17]. DNA mikroarray teknolojisinin, birçok genin aktivitesinin aynı zamanda izlenebilmesi; hızlı bir yöntem olması; hasta ve sağlıklı hücrelerdeki genlerin aktivitelerinin karşılaştırılmasını sağlaması gibi avantajları bulunmaktadır [18].

Mikro-dizilim veri setlerini makina öğrenme algoritmaları ile anlamlandırmak örnek sayısının azlığı ve nitelik sayısının çokluğundan dolayı oldukça zordur.

Bu çalışmada 9_Tumors, 11_Tumors, 14_Tumors, Brain_Tumor1, Brain_Tumor2, Leukemia1, Leukemia2, Lung_Cancer, SRBCT, Prostate_Tumors, DLBCL adında toplam 11 adet veri seti üzerinde öznitelik seçme algoritmalarının sınıflandırma başarı oranı ve işlem süreleri bakımından karşılaştırılması yapılmıştır. Çalışmada kullanılacak veri setleri kanser tümör çeşitlerini sınıflandırmaktadır. Tablo 2’de bu çalışmada kullanılan mikro-dizilim veri setlerinin hangi tümör tiplerini belirlediği listelenmiştir.

Veri Seti	Açıklaması
9_Tumors	9 Çeşit İnsan Tümör Tipi
11_Tumors	11 Çeşit İnsan Tümör Tipi
14_Tumors	14 Çeşit İnsan Tümör Tipi ve 12 Normal Doku
Brain_Tumor1	5 Çeşit Beyin Tümör Tipi
Brain_Tumor2	4 Çeşit Beyin Tümör Tipi
Leukemia1	AML, ALL B-cell ve ALL T-cell
Leukemia2	AML, ALL, MLL
Lung_Cancer	4 Çeşit Akciğer Tümör Tipi ve Normal Doku
SRBCT	4 Çeşit Hücre Tümörü
Prostate_Tumor	Prostat Tümörü ve Normal Doku
DLBCL	2 Çeşit Hücre Tümörü

Tablo 2. Mikro-dizilim veri setleri açıklamaları.

Bazı veri setleri bölgesel tümörlere yoğunlaşmış olmakla beraber 9_Tumors, 11_Tumors ve 14_Tumors veri setleri vücut genelinde tümör sınıflandırması yapmaktadır. Veri setlerine ait örnek

sayısı, sınıf sayısı, gen sayısı Tablo 3'de özetlenmiştir.

Veri Seti	Sınıf Sayısı	Örnek Sayısı	Gen Sayısı
9_Tumors	9	60	5726
11_Tumors	11	174	12533
14_Tumors	26	308	15009
Brain_Tumor1	5	90	5920
Brain_Tumor2	4	50	10367
Leukemia1	3	72	5327
Leukemia2	3	72	11225
Lung_Cancer	5	203	12600
SRBCT	4	83	2308
Prostate_Tumor	2	102	10509
DLBCL	2	77	5469

Tablo 3. Mikro-dizilim veri setleri özellikleri.

Tablo 3'de görüldüğü üzere mikro-dizilim veri setleri farklı sınıf, örnek ve gen sayısına sahiptirler. En çok sınıf sayısına sahip veri seti 14_Tumors veri setidir.

3. Filtre Modelli Öznitelik Seçme Algoritmalarının Karşılaştırılması

Bu çalışmada karşılaştırılacak veri setleri Şekil 5'de gösterilen işlem adımlarına tabii tutularak sınıflandırma başarı oranları hesaplanmıştır.



Şekil 5. Veri Setlerine Uygulanan İşlem Adımları.

Veri Seti	Ki-Kare	Relieff	F-Skor	CFS	IG
9_Tumors	1.21	1.97	2.68	9.34	2.32
11_Tumors	3.54	12.60	6.45	28.59	5.46
14_Tumors	7.34	57.12	17.61	138.56	8.65
Brain_Tumor1	1.32	2.72	1.56	13.58	2.34
Brain_Tumor2	1.74	2.98	2.17	8.54	2.98
Leukemia1	1.13	1.96	0.97	5.42	2.12
Leukemia2	1.91	3.99	1.86	7.54	3.45
Lung_Cancer	3.34	12.11	3.32	14.60	5.39
SRBCT	0.72	1.19	0.70	4.15	1.28
Prostate_Tumor	1.77	5.38	1.40	6.76	4.98
DLBCL	1.28	1.91	0.81	3.42	1.97

Tablo 4. Filtre Modelli Öznitelik Seçme İşlemi Yapılmış Mikro-dizilim Veri Setlerinin İşlem Sürelerinin Karşılaştırılması

Tablo 4'de görüldüğü üzere Ki-Kare ve F-Skor filtre modelli öznitelik seçme algoritmaları işlem süresi bakımından diğer filtre modelli öznitelik seçme algoritmasına göre daha iyi performans sağlamaktadır.

Filtre modelli öznitelik seçme algoritmaları karşılaştırılırken ilk işlem adımı olarak veri setinin tüm niteliklerine filtre modelli öznitelik seçme algoritması uygulanarak her niteliğin istatistiksel olarak bulunduğu sınıf etiketi ile ne kadar fazla ilişkili ve diğer nitelikler ile ne kadar az ilişkili olduğunu belirten bir katsayı hesaplanır. Hesaplanan bu katsayılar eşikleme işlemine tabii tutularak seçme işlemi tamamlanmış olur. Seçilen öznitelikler ile k-NN sınıflandırma modeli oluşturularak, 10-katlamalı çapraz doğrulama ile sınıflandırma başarı oranı hesaplanmıştır. Bu işlem 11 farklı mikro-dizilim veri setine uygulanmıştır.

4. Tartışma ve Sonuç

Bu çalışmada mikro-dizilim verilerine uygulanan filtre modelli öznitelik seçme algoritmaları işlem süresi ve sınıflandırma başarı oranı açısından karşılaştırılmıştır. Karşılaştırma sonucunda mikro-dizilim veri setleri için işlem süresi ve sınıflandırma bakımından en uygun filtre modelli öznitelik seçme algoritması belirlenecektir.

Tablo 4'de Ki-kare, Relieff, F-Skor filtre modelli öznitelik seçme algoritmaları işlem süreleri açısından karşılaştırılmıştır.

9_Tumors, 11_Tumors, 14_Tumors, Brain_Tumor1 ve Brain_Tumor2 veri setlerinde en hızlı sonucu Ki-Kare filtre modelli öznitelik seçme algoritması verirken, Leukemia1, Leukemia2, Lung_Cancer, SRBCT, Prostate_Tumor ve DLBCL veri setlerinde ise

en hızlı sonucu F-Skor filtre modelli öznitelik seçme algoritması vermektedir.

Tablo 5’de ise Ki-kare, ReliefF, F-Skor filtre modelli öznitelik seçme algoritmaları sınıflandırma başarı oranları açısından karşılaştırılmıştır.

Veri Seti	Ki-Kare	ReliefF	F-Skor	CFS	IG
9_Tumors	60.00 %	61.67 %	65.00 %	70.00 %	66.67 %
11_Tumors	89.08 %	86.88 %	83.91 %	90.23 %	83.33 %
14_Tumors	56.82 %	56.17 %	56.17 %	64.61 %	56.82 %
Brain_Tumor1	81.11 %	85.56 %	87.78 %	92.22 %	88.89 %
Brain_Tumor2	72.00 %	74.00 %	72.00 %	90.00 %	78.00 %
Leukemia1	94.44 %	94.44 %	94.44 %	98.61 %	93.06 %
Leukemia2	95.83 %	95.83 %	93.06 %	100.0 %	91.67 %
Lung_Cancer	91.13 %	89.66 %	94.58 %	95.07 %	90.15 %
SRBCT	100.0 %	100.0 %	100.0 %	100.0 %	98.80 %
Prostate_Tumor	81.37 %	81.37 %	79.41 %	94.12 %	89.22 %
DLBCL	93.51 %	97.40 %	90.91 %	96.10 %	93.51 %

Tablo 5. Filtre Modelli Öznitelik Seçme İşlemi Yapılmış Mikro-dizilim Veri Setlerinin Sınıflandırma Başarı Oranı Karşılaştırılması

Tablo 5’de verilen sınıflandırma başarı oranları göz önünde bulundurulduğunda mikro-dizilim veri seti için CFS filtre modelli öznitelik seçme algoritmasının başarılı sonuçlar verdiği gözlemlenmiştir.

CFS filtre modelli öznitelik seçme algoritması, DLBCL mikro-dizilim veri seti hariç diğer tüm veri setlerinde en yüksek sınıflandırma başarı oranını sağlamıştır.

ReliefF filtre modelli öznitelik seçme algoritması ise DLBCL mikro-dizilim veri setlerinde en yüksek sınıflandırma başarı oranını sağlamıştır.

Tablo 6’da ise beş farklı filtre modelli öznitelik seçme algoritmasının da hem işlem süreleri hem de sınıflandırma başarı oranları ortalamaları alınarak bir karşılaştırma yapılmıştır.

Filtre Modelli Öznitelik Seçme Algoritması	Sınıflandırma Başarı Oranı Ortalamaları	İşlem Süresi Ortalamaları
Ki-Kare	83.20 %	2.3
F-Skor	83.38 %	3.59
ReliefF	83.90 %	9.44
CFS	90.09 %	21.86
IG	84.56 %	3.72

Tablo 6. Öznitelik Seçme Algoritmalarının Karşılaştırılması

Tablo 6’da bu çalışmada karşılaştırılan Ki-Kare, F-skor, ReliefF, CFS ve IG filtre modelli öznitelik seçme algoritmalarının ortalama sınıflandırma başarı oranları ve ortalama işlem süreleri karşılaştırılmıştır. Bu bağlamda CFS filtre modelli öznitelik seçme

işlem süreleri hesaplanırken, 2.5 GHz dört çekirdekli Intel Core i7 işlemcili, 16 GB 1600 MHz belleğe sahip bilgisayar kullanılmıştır.

algoritması genel olarak mikro-dizilim veri setleri için en iyi sınıflandırma başarı oranına sahiptir. Fakat işlem süresi bakımından Ki-kare filtre modelli öznitelik seçme algoritması diğer iki algoritmaya göre daha başarılı sonuçlar elde etmiştir.

Teşekkür

Bu çalışma HÜBAK tarafından 15101 numaralı proje kapsamında desteklenmiştir.

Kaynaklar

- [1] Yang AY, Wright J, Ma Y, Sastry SS, “Feature Selection in Face Recognition: A Sparse Representation Perspective” Technical Report No. UCB/EECS-2007-99, 2007.
- [2] Ostrowski DA, “Feature Selection for Twitter Classification” In: Proceedings Of The IEEE International Conference On Semantic Computing (ICSC) 267-272, 2014.
- [3] Akba F, “Duygu Analizinde Öznitelik Seçme Metriklerinin Değerlendirilmesi: Türkçe Film Eleştirileri” Yüksek Lisans Tezi, Hacettepe Üniversitesi Bilgisayar Mühendisliği Bölümü, Ankara, 2014.
- [4] Eskidere Ö, “Ses Ölçümlerinden Parkinson Hastalığının Teşhisi İçin Öznitelik Seçme Yöntemlerinin Karşılaştırılması” Sigma Mühendislik ve Fen Bilimleri Dergisi 30:402–414, 2012.
- [5] Saey Y, Inza I, Larranaga P, “A Review Of Feature Selection Techniques In Bioinformatics” Bioinformatics 19: 2507–2517, 2007.
- [6] Holland J, “Adaptation in Natural and Artificial Systems”, University of Michigan Press, Ann Arbor, 1975.
- [7] Quinlan JR, “C4.5: Programs for Machine Learning” Machine Learning 16: 235-240, 1993.

- [8] Hall M, "Correlation-Based Feature Selection For Machine Learning" Phd Thesis, Department Of Computer Science, Waikato University, New Zealand, 1999.
- [9] Quinlan JR, "Induction Of Decision Trees", Machine Learning 1:81–106, 1986.
- [10] Battiti R, "Using Mutual Information For Selecting Features In Supervised Neural Net Learning" IEEE Transactions On Neural Networks 5:537–550, 1994.
- [11] Liu H, Setiono R, "Chi2: Feature Selection And Discretization Of Numeric Attributes" In: Proceedings Of The IEEE 7th International Conference On Tools With Artificial Intelligence 338-391, 1995.
- [12] Robnik-Sikonja M, Kononenko I, "Theoretical And Empirical Analysis Of Relieff And Relieff" Mach. Learn. 53:23–69, 2003.
- [13] Duda RO, Hart PE, ve Stork DG, "Pattern Classification", John Wiley & Sons, New York, 2001.
- [14] Statnikov A, Tsamardinos I, Dosbayev Y, Aliferis CF, "Gems: A System For Automated Cancer Diagnosis And Biomarker Discovery From Microarray Gene Expression Data", International Journal Of Medical Informatics, 74:491-503, 2005.
- [15] Bolón-Canedo V, Sánchez-Marono N, Alonso-Betanzos A, Benitez JM, Herrera F, "A Review Of Microarray Datasets And Applied Feature Selection Methods" Information Sciences 282:111–135, 2014.
- [16] Cover T, Hart P, "Nearest Neighbor Pattern Classification" IEEE Transactions On Information Theory 13:21–27, 1967.
- [17] Bednar M, "DNA Microarray Technology and Application", Med Sci Monid, 6: 796–800, 2000.
- [18] Sassanfar, S, Walker, G, "DNA Microarray Technology. What Is It and How Is It Useful", MIT, Biology Science Outreach, 2003.