Afyon Kocatepe Üniversitesi Fen ve Mühendislik Bilimleri Dergisi

Afyon Kocatepe University Journal of Science and Engineering

AKÜ FEMÜBİD **16** (2016) Özel Sayı (31-35) AKU J. Sci. Eng.**16** (2016) Özel Sayı (31-35)

3 Farklı Filtre Modelli Öznitelik Seçme Algoritmalarının Kombine Edilerek İyileştirilmesi

Abdülkadir Gümüşçü[1], İbrahim Berkan Aydilek[2] - Ramazan Taşaltın[3]

[1,3]*Harran Üniversitesi, Mühendislik Fakültesi, Elektrik-Elektronik Mühendisliği Bölümü, Şanlıurfa.*

[2]*Harran Üniversitesi, Mühendislik Fakültesi, Bİlgisayar Mühendisliği Bölümü, Şanlıurfa*

e-posta: agumuscu@harran.edu.tr

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

Özet

Günümüzde mikro-dizilim verileri hastalık teşhisinde kullanılmaya başlanmıştır. Mikro-dizilim verilerinin sahip olduğu yüksek p ve düşük n parametreleri, makina öğrenme algoritmalarının sınıflandırma başarı oranını olumsuz yönde etkilemektedir. Bu sebeple mikro-dizilim veri setlerinin modellenmesinde öznitelik seçme önemli bir işlem adımı olarak karşımıza çıkmaktadır. Bu çalışmada 3 farklı filtre modelli öznitelik seçme algoritması 7 farklı veri setine uygulanacak ve sonuçlar birleştirilerek öznitelik seçme işlemi gerçekleştirilecektir. Elde edilen sonuçlar k-NN makina öğrenme algoritmasında, 10-katlamalı çapraz doğrulama metodu ile doğrulanarak test edilmiştir. Önerilen metot ile elde edilen sınıflandırma başarı oranları, 3 farklı filtre modelli öznitelik seçme algoritmalarının teker teker uygulanması ile elde edilen başarı oranları ile kıyaslanmıştır. Bu kıyaslamaya ek olarak filtrelerin uygulanma süreleri de kıyaslanmıştır.

Improvement of Filter Model Feature Selection By Combining 3 Different Filter Model Feature Selection

Keywords
Micro-arrray data set; Classificaiton; Feature Selection.

Abstract

Nowadays, microarray data is an important contribution to the diagnosis of illness. High p and low n parameters that have micro-array data, affect the machine learning algorithms classification success rate negatively. For this reason feature selection is a pre-processing technique with great importance in microarray classification. In this study 3 different feature selection algorithms applied to 7 different datasets and results will be achieved by combining feature selection process. The resultsaretested with k-fold cross validation by using k-nearest neighbor (k-NN) method. Classification success rate that obtained by the proposed method is compared withobtained by the filter modelled feature seleciton method individually. In addition, proccesing times are compared.

1. Giriş

Günümüzde mikro-dizilim verileri hastalık teşhisine önemli katkılar sağlamaktadır. Mikro-dizilim verilerini, makina öğrenme algoritmaları ile anlamlandırmak örnek sayısının azlığı ve gen sayısının fazlalığından ötürü çok zordur. Bu açıdan bakıldığında gen analizinde öznitelik seçme algoritmaları çok önemli bir işlem adımıdır.

Literatürde öznitelik seçme algoritmaları filtre modelli, sarmal modelli ve gömülü olmak üzere üç ana başlıkta ele alınmaktadır (Saeys *ve ark*. 2007). Filtre modelli öznitelik seçme algoritmaları istatistiksel varsayımlar kullanılarak niteliklerin sonuca ulaşmaya katkısını hesaplayarak nitelikleri puanlar ve eşikleme yaparak öznitelikleri seçer. Filtre modelli öznitelik seçme algoritmaları sınıflandırma algoritmalarından bağımsız çalışırlar. Filtre modelli öznitelik seçme algoritmalarının avantajları hızlı ve sınıflandırma algoritmalarından bağımsızlık şeklinde özetlenebilir (Saeys *veark*. 2007). En çok kullanılan filtre modelli öznitelik seçme algoritmaları Korelasyon-bazlı öznitelik seçme(CFS) (Hall,1999), bilgi kazancı (Quinlan,1986), ortak bilgi (Battiti,1994), Ki-kare (Liu ve Setiono 1994), ReliefF (Robnik-Sikonjave

Kononenko 2003), ve F-skor (Duda *ve ark*. 2001) şeklinde sıralanmaktadır. Sarmal modelli öznitelik seçme algoritmaları ise sınıflandırma başarı oranını yükseltme amacıyla oluşturulan nitelik gruplarını sınıflandırma algoritmasında deneyerek en ideal öznitelikleri bulma metodlarıdır. Sarmal modelli öznitelik seçme algoritmalarının başarı oranının yüksek olmasının yanında özellikle nitelik sayısı büyük olan veri setlerinde işlem süresi çok uzamaktadır (Kohavi ve John 1997). Sıralı İleri Seçme (SFS) (Kittler,1978), Sıralı Geri Eleme (SBE) (Kittler,1978) ve Genetik algoritma (Holland,1975) kullanılan başlıca sarmal modelli öznitelik seçme algoritmalarıdır. Gömülü modelli öznitelik algoritmaları ise sınıflandırma algoritmasına dahil olan öznitelik seçme metodlarındandır. Bu algoritmalar zaten sınıflandırma yaparken öznitelik seçmek zorundadır. Bu modele ID3 (Quinlan,1986), C4.5 (Quinlan,1993) karar ağaçları örnek verilebilir.

Bu çalışmada mikro-dizilim veri setlerine tek bir filtre modelli öznitelik seçme algoritması uygulamak yerine başarı oranını iyileştirecek 3 farklı filtre modelli öznitelik seçme algoritmasının birleştirilerek uygulanması önerilecektir. Önerilen yöntem ile mikro-dizilim veri setinde sarmal modelli öznitelik seçme algoritmasına göre daha hızlı, tek bir filtre modelli öznitelik seçme algoritmasına göre daha başarılı bir öznitelik seçme işlemi yapılabilecektir.

Önerilen yöntemin başarı oranı ve işlem süresi bilgi kazancı filtre modelli öznitelik seçme, genetik algoritma sarmal modelli öznitelik seçme algoritmaları ile kıyaslanmıştır.

Bu çalışmada http://www.gems-system.org (Statnikov*ve ark*. 2005) sitesinden indirilen mikro-dizilim veri setleri kullanılmıştır.

2. Materyal ve Metot

Önerilen yöntemde 3 farklı filtre modelli öznitelik seçme algoritması kullanılmıştır. Bunlar Ki-Kare, ReliefF, F-Skor olarak sıralanmaktadır.

2.1.Ki-Kare

Ki-kare istatistik temelli olup öznitelik seçme işlemlerinde yaygın şekilde kullanılmaktadır. Bu metot bulundukları sınıfa göre tüm niteliklerin Ki-kare'sini hesaplayarak tek tek değerlendirir (Liu ve Setiono 1994).

2.2.ReliefF

ReliefF,Relief istatistiksel modelinin geliştirilmiş versiyonudur. Relief 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 (Bolón-Canedo *ve ark*. 2014).

2.3.F-Skor

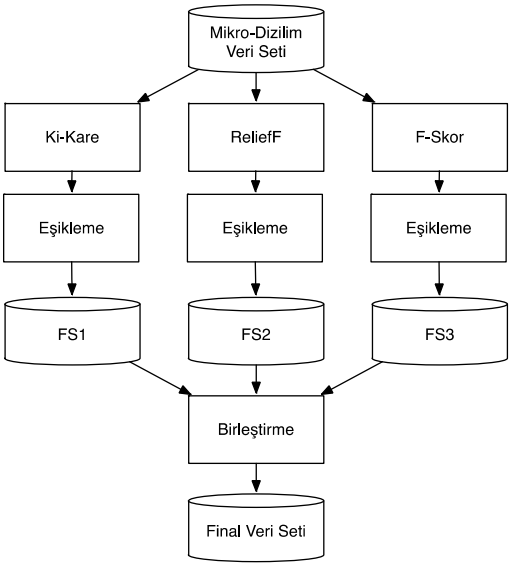
Fisher-skor filtresi her niteliğin diğer niteliklerden farkını hesaplayan bir filtre modelidir. Dolaysıyla diğer niteliklerden farkı fazla olan nitelik sınıflandırma algoritmaları için daha önemlidir. Bundan dolayı Fisher-skor filtresinin sonucunda yüksek değere sahip nitelikler daha ilişkili niteliklerdir (Duda *ve ark.*. 2001).

2.4.k-NN Sınıflandırma

k-NN sınıflandırma algoritması birçok alanda sıklıkla kullanılan metotların başında gelmektedir (Cover ve Hart 1967). k-NN sınıflandırmasının önemli avantajları basitliği ve kullanım kolaylığı yanında büyük veri setlerinde sağladığı kararlılık da kullanımını yaygınlaştırmıştır.

k-NN sınıflandırma algoritması uygulanırken eğitim verisindeki örnekler düzleme yerleştirilerek sınıfları ile ilişkilendirilir. Kurulan düzleme test için kullanılacak örnekler düzleme yerleştirilir. 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. Bu çalışmasında k değeri 1 olarak seçilip, komşuluk uzaklık hesabı öklid uzaklığa göre hesaplanmıştır**.**

3. Önerilen Yöntem



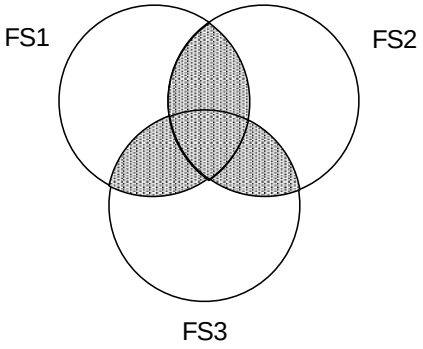
Şekil 1.Önerilen Yöntemin İşlem Adımları

Şekil 1'de gösterildiği gibi mikro-dizilim veri setleri öncelikle 3 farklı filtre modelli öznitelik seçme algoritmalarına tabii tutulmaktadır. Eşikleme yapılarak seçilen öznitelik kümeleri FS1, FS2 ve FS3 şeklinde adlandırılmıştır.

FS1, FS2 ve FS3 öznitelik veri setleri birleştirilerek FS4 oluşturulmaktadır. FS4 oluşturulurken formül (1)'deki kriter gözönünde bulundurulmuştur.

$$FS4 = (FS1 \cap FS2) \cup (FS1 \cap FS3) \cup (FS2 \cap FS3) \quad (1)$$

Bir niteliğin FS4 öznitelik veri setine dahil olabilmesi için FS1, FS2 ve FS3 öznitelik veri setlerinden en az ikisinde bulunması gerekmektedir. Şekil 2'de bu kriter özetlenmiştir.



Şekil 2.FS4 öznitelik veri setine seçilme kriteri

Birleştirme adımından sonra seçilen FS4 öznitelik veri seti k-NN sınıflandırma algoritması ile test edilmektedir. Sınıflandırma başarı oranları hesaplanırken 10-katlamalı doğrulama yapılmaktadır.

4. Tartışma ve Sonuç

Bu çalışmada mikro-dizilim verilerine uygulanan filtre modelli öznitelik seçme algoritması tek bir istatistiksel model ile değil 3 farklı model ile kombine edilerek sarmal modelli öznitelik seçme algoritmalarına göre daha hızlı, tek bir filtre modelli öznitelik seçme algoritmasına göre daha başarılı bir yöntem önerilmiştir. Tablo 1'de önerilen yöntem ile Ki-kare, ReliefF, F-Skor filtre modelli öznitelik seçme algoritmaları ve Genetik algoritma sarmal modelli öznitelik seçme yöntemi işlem süreleri açısından karşılaştırılmıştır.

Tablo 1. Önerilen yöntemin işlem sürelerinin karşılaştırılması

Veri Seti	Ki-Kare	ReliefF	F-Skor	Bilgi Kazancı	Genetik Algoritma	Önerilen Yöntem
11_Tumors	**3.54**	6.45	12.60	11.23	2795.30	25.12
Brain_Tumor1	**1.32**	1.56	2.72	2.56	8428.9	5.51
Brain_Tumor2	**1.74**	2.17	2.98	2.67	17778.0	6.36
Leukemia1	1.13	**0.97**	1.96	1.74	11775.0	3.87
Leukemia2	1.91	**1.86**	3.99	3.06	2946.40	7.05
SRBCT	0.72	**0.70**	1.19	0.93	1921.70	2.09
Prostate_Tumor	1.77	**1.40**	5.38	4.21	122.94	7.89

Tablo 1'de de görüldüğü üzere önerilen yöntem mikro-dizilim veri setlerinde genetik algoritma

sarmal modelli öznitelik seçme yönteminden daha hızlı çalışmaktadır. Filtre modelli öznitelik seçme algoritmalarında ise 11_Tumors, Brain_Tumor1 ve Brain_Tumor2 veri setlerinde Ki-Kare, Leukemia1, Leukemia2, SRBCT ve Prostate_Tumor veri setlerinde ise ReliefF filtre modelli öznitelik seçme algoritmaları daha hızlı sonuç vermiştir.

Önerilen yöntem işlem süresi açısından filtre modelli öznitelik seçme algoritmalarına göre yavaş, fakat sarmal modelli öznitelik seçme algoritmalarına göre hızlı olarak değerlendirilebilir.

Tablo 2'de ise önerilen yöntem sınıflandırma başarı oranına göre karşılaştırılmıştır. Önerilen yöntem, Ki-Kare, ReliefF, F-Skor ve Bilgi Kazancı olmak üzere 4 farklı filtre modelli öznitelik seçme algoritması ve Genetik algoritma sarmal modelli öznitelik seçme yöntemi ile karşılaştırılmıştır.

Tablo 2. Önerilen yöntemin sınıflandırma başarı oranı karşılaştırılması

Veri Seti	Ki-Kare	ReliefF	F-Skor	Bilgi Kazancı	Genetik Algoritma	Önerilen Yöntem
11_Tumors	89.08%	86.88%	83.91%	83.33%	83.33%	**89.66%**
Brain_Tumor1	81.11%	85.56%	87.78%	88.89%	**92.22%**	87.78%
Brain_Tumor2	72.00%	74.00%	72.00%	78.00%	**80.00%**	78.00%
Leukemia1	94.44%	94.44%	94.44%	93.06%	**97.22%**	94.44%
Leukemia2	**95.83%**	**95.83%**	93.06%	91.67%	**95.83%**	**95.83%**
SRBCT	**100.00%**	**100.00%**	**100.00%**	98.80%	96.39%	**100.00%**
Prostate_Tumor	81.37%	81.37%	79.41%	89.22%	**91.18%**	82.35%

Tablo 2'deki sonuçlara göre önerilen yöntem 11_Tumors, Leukemia2 ve SRBCT veri setlerinde Genetik algoritma sarmal modelinden daha yüksek sınıflandırma başarı oranına sahiptir. Belirtilen veri setlerinde önerilen yöntem hem işlem süresi hem de sınıflandırma başarı oranı kategorilerinde iyileştirme sağlamıştır.

Teşekkür

Bu çalışma Harran Üniversitesi Bilimsel Araştırma Komisyonu (15101) tarafından maddi olarak desteklenmektedir. Desteklerinden ötürü Harran Üniversitesi Bilimsel Araştırma Komisyonu'na teşekkür ederiz.

Kaynaklar

Battiti R., 1994. Using Mutual Information For Selecting Features In Supervised Neural Net Learning, IEEE Transactions On Neural Networks 5:537–550.

Bolón-Canedo V., Sánchez-Marono N., Alonso-Betanzos A., Benıtez J.M., Herrera F. 2014. A Review Of Microarray Datasets And Applied Feature Selection Methods Information Sciences 282:111–135

Cover T. , Hart P., 1967. Nearest Neighbor Pattern Classification, IEEE Transactions On Information Theory 13:21–27.

Duda R.O., Hart P.E., And Stork D.G., 2001. Pattern Classification, John Wiley & Sons, New York.

Hall,M., 1999. Correlation-Based Feature Selection For Machine Learning. Phd Thesis., Department Of Computer Science, Waikato University, New Zealand.

Holland J., 1975. Adaptation In Natural And Artificial Systems. University Of Michigan Press, Ann Arbor.

Kittler,J. 1978. Pattern Recognition And Signal Processing, Chapter Feature Set Search Algorithms Sijthoff And Noordhoff, Alphen Aan Den Rijn, Netherlands, Pp. 41–60.

Kohavi R., John G.H., 1997. Wrappers For Feature Subset Selection, Artificial Intelligence 97:273–324.

Lıu H., Setiono R., 1995. Chi2: Feature Selection And Discretization Of Numeric Attributes. In: Proceedings Of The IEEE 7th International Conference On Tools With Artificial Intelligence 338-391.

Robnik-Sikonja M., Kononenko I., 2003. Theoretical And Empirical Analysis Of Relieff And Relieff Mach. Learn. 53:23–69.

Quinlan J.R., 1986. Induction Of Decision Trees, Machine Learning 1:81–106.

Quinlan, J. R., 1993. C4. 5: Programs For Machine Learning, Machine Learning 16: 235-240.

Saeys Y., Inza I., Larranaga P., 2007. A Review Of Feature Selection Techniques In Bioinformatics. Bioinformatics 19: 2507–2517.

Statnikov A., Tsamardınos I., Dosbayev Y., Alıferıs C.F., 2005. Gems: A System For Automated Cancer Diagnosis And Biomarker Discovery From Microarray Gene Expression Data, International Journal Of Medical Informatics, 74:491-503.