C. Ünsalan and A. Erçil, "Comparison of feature selection algorithms a new performance criteria for feature selection", Proceedings of IEEE SIU'98, pp. 60-65, May 1998, Kizilcahamam, Turkey (in Turkish)

Öznitelik Seçme Yöntemlerinin Karşılaştırılması ve Başarı Kriteri

* Cem ÜNSALAN ** Aytül ERÇİL
*Boğaziçi Üniversitesi, Elektrik-Elektronik Mühendisliği Bölümü
unsalan@boun.edu.tr
**Boğaziçi Üniversitesi, Endüstri Mühendisliği Bölümü
ercil@boun.edu.tr

Özetçe

Sınıf bilgisinin elde edileceği özniteliklerin bazıları, bu bilgiyi içermiyor olabilir. Bu özniteliklerin sınıflandırıcıya verilmeleri bir anlam ifade etmeyeceği için, sınıflandırma işleminden önce ayıklanmaları gerekir.Sınıf bilgisini içinde barındıran ve sınıflandırıcının işini kolaylaştıran özniteliklerin bulunması için öznitelik seçme yöntemleri geliştirilmiştir. Bu çalışmada, kullanılan öznitelik seçme yöntemlerine ek olarak yeni bir öznitelik seçme yöntemi önerilmiştir. Tüm öznitelik seçme yöntemlerinin başarılarını hesaplamak için yeni bir yöntem geliştirilmiştir.

1. Giriş

Sınıflandırıcıya verilecek öznitelik sayısı, tüm bileşimleri denemeye elverişi değilse, sınıflandırıcının işini kolaylaştıracak bir ön işleme gerek duyulur. Bu sayede sınıflandırma gücü olmayan öznitelikler ayıklanabilir. Öznitelik uzayının boyutu arttığı zaman, sınıflandırıcı için boyut laneti de sınıflandırıcının başarısını azaltıcı etki yapar. Bu nedenlerden ötürü öznitelik seçme yöntemleri kullanılır.

Her özniteliğin elde edilmesi toplam sistem için fazladan bir işlem yükü demektir ki bu da işlem zamanını arttırır. Gerçek zamanlı uygulamalarda bu sorun da kullanılan öznitelik sayısının sınırlandırılmasını gerektirir. Bu nedenle de öznitelik seçme yöntemlerine ihtiyaç duyulur.

En önemli öznitelik seçme yöntemlerinden biri dallanma ve sınırlama algoritmasıdır. Bu algoritma Narendra ve Fukunaga [1] tarafından önerilmiştir. Kittler [2] değişik öznitelik seçme yöntemlerini incelemiştir. Pudil et al. [3] sıralı ileri kayan ve sıralı geri kayan seçme yöntemlerini kullanmıştır.

2. Öznitelik Seçme Yöntemleri

Bu bölümde kullanılan öznitelik seçme yöntemleri kısaca ele alınacaktır. Literatürde kullanılan öznitelik seçme yöntemleri bunlarla sınırlı değildir.

2.1 Entropi Ölçütü

Entropi ölçütü özniteliğin içindeki bilgi yardımı ile seçim yapar. Bu nedenle öznitelik bir dağılım gibi ele alınıp, entropisi bulunur. Entropi arttıkça sınıfların daha iyi ayrılabiliyor olduğu kabul edilir.

Entropi ölçütünü hesaplamak için öznitelik normalleştirilir, öyle ki tüm değerler sıfırın üstünde olsun ve toplam bir etsin.

M veriye sahip ft özniteliğinin entropisi aşağıdaki şekilde bulunabilir:

$$E = -\sum_{i=1}^{M} ft^*(i)\log(ft^*(i))$$

$$ft^*(i) = \frac{ft(i) + |\min(ft)|}{\sum_{i=1}^{M}(ft(i) + |\min(ft)|)}$$

2.2 Şekil Benzerliği ile Seçme

Bir özniteliğin şekli, verilerin dağılımı olarak alınabilir, öyle ki veriler eldeki tüm öznitelikler için aynı sıraya sahip olsun. Şekil benzerliği ile seçmenin amacı, dağılımları birbirine yakın olan öznitelikleri ayıklamaktır.

Eğer iki öznitelik birbirine yakın şekillere sahip iseler, bu özniteliklerdeki veri dağılımı da birbirine yakındır demektir. Bu özniteliklerden birini atmak seçme bilgisini azaltmaz.

Eldeki iki öznitelik için benzerlik ölçüsü, bu iki özniteliğin yumuşatılmış ve normalleştirilmiş şekilleri arasındaki Euclid uzaklığı olarak alınır. Yumuşatma işlemi şekil üzerindeki gürültüyü azaltmak ve şekle bir genellik vermek amacıyla yapılır.

Benzerlik ölçüsü aşağıdaki gibi hesaplanır:

Aynı i indisine sahip, ft_1 ve ft_2 öznitelikleri için:

$$şekil\ ft_1 \rightarrow ft_1^*(i) = \frac{ft_1(i) + |\min(ft_1)|}{\sum_{i=1}^{M}(ft_1(i) + |\min(ft_1)|)} \quad i = 1...M$$

$$şekil\ ft_2 \rightarrow ft_2^*(i) = \frac{ft_2(i) + |\min(ft_2)|}{\sum_{i=1}^{M}(ft_2(i) + |\min(ft_2)|)} \quad i = 1...M$$

M toplam veri sayısıdır.

Bu iki öznitelik arasındaki uzaklık aşağıdaki gibi bulunur:

$$uzaklik = \sum_{i=1}^{M} |ft_1^*(i) - ft_2^*(i)|$$

Birbirine benzeyen öznitelikleri bulmak için, uzaklık ölçüsü üzeride hiyerarşik kümeleme yapılır [4]. Seçilecek öznitelik sayısı sonuç küme sayısını belirler, ki bu da öznitelikler arasındaki benzerliği doğrudan belirler.

2.3 Fisher'in Ölçütü

Lineer Ayırma Analizi (LDA) projeksiyon algoritmalarından biridir. Projeksiyon sınıf ayrımını arttırmak için yapılır, böylece yeni uzayda sınıflandırma işlemi daha kolay yapılabilir [4]. Bu uzayı bulabilmek için iteratif olarak Fisher'in geliştirdiği ölçü arttırılmaya çalışılır. Bu ölçü:

$$d = \frac{siniflar\ arasi\ uzaklik}{\sum sinif\ \ içi\ dagilim}$$

Fisher'in ölçüsü iki koşulu sağlamaya çalışır:

i- Sınıf merkezleri arası uzaklıkençoklanmalıdır.

ii- Her sınıfın kendi içindeki dağılımı en az düzeye getirilmelidir.

Bu çalışmada bu ölçü doğrudan özniteliğin kalitesini belirtiyor kabul edilmiştir.

2.4 Principal Component Analysis (PCA)

PCA eldeki veri kümesini, yeni veri uzayına geçirir, öyle ki bu yeni uzayda veri dağılımındaki sapma en iyi şekilde korunsun. Eğer geçirme işlemi daha düşük boyutlu bir uzaya yapılırsa, öznitelik uzayının boyutu da azaltılmış olur.

PCA ve yapay sinir ağları arasında bağ bulmak için araştırmalar yapılmıştır. Oja [5] tek bir nöron kullanarak PCA'yı uygulamıştır. Chen [6] uyarlanır PCA bulmak için yapay sinir ağlarını kullanmıştır. Diamantaras ve Kung [7], kitaplarında PCA ve yapay sinir ağlarını birleştirmişlerdir.

3. Öznitelik Seçme Yöntemleri İçin Başarı Kriteri

Kullanılan öznitelik seçme yöntemleri de başarıları ölçüsünde derecelendirilmelidir. Bu bölümde başarı derecelendirmesi için bir yöntem geliştirilmeye çalışılmıştır. Öznitelik seçme yöntemleri için başarı derecesini aşağıdaki gibi belirtebiliriz:

Her sınıflandırıcı, yapısına göre öncel sınıf yapısı için bir tahminde bulunur, ki sınıflandırma başarısı bu tahminle doğrudan ilgilidir. Eğer eldeki öznitelik kümesi yeterince fazla ve farklı yapılara sahip sınıflandırıcı ile denenirse, genel sınıflandırma sonucu bu öznitelik kümesi hakkında bilgi verir. Kullanılan sınıflandırıcı sayısını ne kadar arttırırsak, sınıflandırıcıların öncel tahminlerinden de o kadar bağımsız bir sonuç elde ederiz.

Bu çalışmada, kullanıcı sayısını yeterince arttırmak yerine, öncel tahminleri farklı iki sınıflandırma yöntemi kullanılmıştır. Bu yöntemler Bayes sınıflandırıcısı ve KNN sınıflandırıcısıdır [4]. Bayes sınıflandırıcısı için öncel sınıf dağılımı Normal dağılım olarak alınır. Diğer taraftan KNN sınıflandırıcısı için bir öncel sınıf dağılımı kabul edilmez ve eldeki veri kümesinden bu dağılım bulunmaya çalışılır. Bu yüzden bu iki yöntem istatistiksel veri analizinde iki farklı yaklaşımı gösterir.

Eldeki veri kümesi için, eğer iki sınıflandırıcı da iyi sonuç verirse, buradan bu öznitelik kümesinin kullanılan sınıflandırıcıdan bağımsız olarak yeterince iyi olduğu düşünülebilir. Ya da bir sınıflandırıcı iyi sonuç verirse, eldeki öznitelik kümesinin sınıflandırıcıya bağımlı olarak iyi olduğu düşünülebilir. Öznitelik seçme yöntemleri ile seçilen öznitelikler bu yöntemle sınıflandırılınca, ortalama başarı da öznitelik seçme yönteminin başarısı hakkında bilgi verir.

Başarı kriteri şu şekilde hesaplanır:

i- Bayes sınıflandırıcısı ve KNN sınıflandırıcısı öznitelik seçme yöntemi tarafından şeçilen öznitelikleri sınıflandırır ve sonuç yüzde olarak belirtilir.

Burada iki farklı başarı kriteri bulunabilir.

ii- Eğer sınıflandırıcıdan bağımsız başarı isteniyorsa, her iki sınıflandırma sonucunun ortalaması ve standart sapması hesaplanır. Ortalama, ortak sınıflandırma yüzdesini, standart sapma da bu iki sınıflandırıcının sınıflandırmada ne kadar anlaştığını belirtir. Ortak sınıflandırma yüzdesi ve anlaşma ağırlıklı olarak toplanır ve öznitelik kümesi için başarı derecesi bulunur.

iii- Eğer sınıflandırıcıya bağımlı başarı isteniyorsa, her iki sınıflandırma sonucundan en yüksek olanı alınır ve bu öznitelik için başarı derecesi kabul edilir.

iv- Öznitelik seçme yönteminin başarısını bulabilmek için, yeterince çok öznitelik uzayından seçme yapılır ve her uzaydaki başarı toplanıp, ortalaması alınır bu da öznitelik seçme yönteminin başarısını verir.

Başarı kriterini formül halinde yazmak istersek:
Diyelim ki:

$C_m\ (f_i)$ Bayes sınıflandırıcının öznitelik seçme yöntemi ile seçilen öznitelik f_i için doğru sınıflandırma yüzdesini versin.

$C_k\ (f_i)$ KNN sınıflandırıcının öznitelik seçme yöntemi ile seçilen öznitelik f_i için doğru sınıflandırma yüzdesini versin.

Ortalama doğru sınıflandırma:

$$\mu_i = \frac{C_m(f_i) + C_k(f_i)}{2}$$

Anlaşma veya doğru snıflandırma standart sapması:

$$\sigma_i = (C_m(f_i) - \mu_i)^2 + (C_k(f_i) - \mu_i)^2)^{1/2}$$

Eğer öznitelik seçme yöntemi N öznitelik kümesi için denenmişse:

$$Basari = \frac{1}{N}\sum_{i=1}^{N}(w_1 * \mu_i + w_2 * \sigma_i)$$

4. Deney

Başarı kriteri yöntemini test etmek için dört deney seti hazırlanmıştır. Bu setlerin özellikleri sırasıyla:

Set1, İris veri kümesidir. Fisher tarafından kullanılmıştır ve dört boyutlu uzayda üç sınıf barındırır, öyle ki bir sınıf diğer iki sınıftan doğrusal olarak ayrılabilir değildir.

Set2, en kötü durum olarak alınmıştır. Birbiriyle çakışan Normal dağılıma sahip iki sınıf, bu ikisini kapsayan normal dağılıma sahip bir üçüncü sınıfın içinde alınmıştır. Öznitelik uzayı üç boyutludur.

Set3, iki boyutlu iki öznitelik kümesi birleştirilerek elde edilmiştir. Birinci kümede bir sınıf normal olarak dağılmıştır ve ikinci sınıf bu sınıfı çevreleyecek şekilde ve yarıçapı normal dağılıma sahip dairesel bir yapıya sahiptir. Ikinci küme de normal dağılıma sahip ve çakışan iki kümenin birleşiminden oluşur.

Set4, PCA algoritmasının en zayıf olduğu öznitelik uzayıdır. İki sınıf, en çok sapmaya sahip iki boyutta tamamen çakışır ve en az sapmaya sahip üçüncü botutta doğrusal olarak ayrılabilir.

Bu deney setleri ile ilgili daha geniş bilgi referanslarda bulunabilir [8].

Aşağıdaki tabloda herbir öznitelik seçme yöntemi tarafından seçilen öznitelik kümesinin sınıflandırma sonuçları yüzde olarak verilmiştir. Her öznitelik seçme yönteminden iki tane öznitelik seçmesi istenmiştir.

Tablo 4.1 Sınıflandırma Sonuçları

		Entropi	Fisher	Şekil B.	PCA
Set1	Bayes	80.00	97.33	94.67	89.33
	KNN	77.33	94.67	97.33	94.67
Set2	Bayes	44.00	45.33	45.33	39.56
	KNN	64.44	59.56	59.56	57.11
Set3	Bayes	79.00	79.00	73.00	53.00
	KNN	69.00	69.00	94.50	100.00
Set4	Bayes	50.50	100.00	100.00	58.50
	KNN	47.50	100.00	100.00	54.00

Aşağıdaki tabloda, öznitelik seçme yöntemleri için sınıflandırıcıdan bağımsız başarı dereceleri bulunmuştur.

Tablo 4.2 Sınıflandırıcıdan Bağımsız Başarı

	Entropi	Fisher	Şekil B.	PCA
Set1	78.56	95.89	95.89	91.79
Set2	53.40	51.88	51.88	47.63
Set3	73.60	73.60	82.89	74.62
Set4	48.88	100.00	100.00	56.07
Başarı	**63.61**	**80.34**	**82.66**	**67.53**

Aşağıdaki tabloda, öznitelik seçme yöntemleri için sınıflandırıcıya bağımlı başarı dereceleri bulunmuştur.

Tablo 4.3 Sınıflandırıcıya Bağımlı Başarı

	Entropi	Fisher	Şekil B.	PCA
Set1	80.00	97.33	97.33	94.67
Set2	64.44	59.56	59.56	57.11
Set3	79.00	79.00	94.50	100.00
Set4	50.50	100.00	100.00	58.50
Başarı	**68.49**	**83.97**	**87.85**	**77.57**

Bu sonuçlara göre Fisher'in öznitelik seçme yöntemi ve Şekil Benzerliği ile seçme yöntemi birbirlerine yakın ve yüksek başarı dereceleri almışlardır. Bu sonucu test etmek için gerçek veri kümesi bir sonraki adımda kullanılacak ve bu kümeden seçilen öznitelikler LVQ sınıflandırıcısı ile sınıflandırılıp başarı kriteri hakkında daha sağlıklı bilgi elde edilecektir.

Gerçek veri kümesi çelik yüzey üzerinden, farklı imge işleme yöntemleri ile elde edilen öznitelikler kullanılarak yapılacaktır. Kullanılan imge işleme yöntemleri verilen

referansta daha iyi görülebilir [9].Her doku analizi yönteminden elde edilen öznitelik kümesi, öznitelik seçme yöntemleri kullanılarak, hiyerarşik bir yapıda üç özniteliğe indirilmiştir [8]. Bu öznitelikler LVQ ile sınıflandırılıp iki sınıf bilgisi, paslı yüzey ve kumlanmış yüzey, elde edilmiştir. Sınıflandırma sonucu başarı aşağıdaki tabloda verilmiştir. Her doku analizi yönteminden elde edilen öznitelik sayısı da bu tabloda belirtilmiştir.

Tablo 4.4 Öznitelik Seçme Yöntemlerini Gerçek Veri ile Karşılaştırma

Doku Analizi Yöntemi	Top. Özn.	Entropi	Fisher	Şekil B.	PCA
Gri Düzey Eş Oluşum Matrisi	30	%97.20	%96.11	%92.58	%77.62
Markov Rassal Alanları	125	%88.20	%90.51	%86.50	%85.52
İmgenin Histogramı	20	%97.81	%97.45	%98.05	%97.20
İki Boyutlu Fourier Dönüşümü	20	%92.34	%98.16	%86.74	%90.15
Wavelet Dönüşümü	1792	%88.44	%97.93	%58.03	%51.22
Gabor Dönüşümü	30	%51.34	%98.05	%51.34	%51.34
Radon Dönüşümü	120	%88.56	%99.15	%92.58	%80.54
Yüzey Yoğunluğu Yaklaşımı	240	%96.84	%82.85	%97.20	%93.19

Tablo 4.4 den görüleceği üzere Fisher'in öznitelik seçme ölçüsü sekiz set üzerinden altısında en yüksek sınıflandırma başarısını elde etmiştir. Bunu Şekil Benzerliği ile seçilen öznitelikler izlemektedir. Burada belirtilmesi gereken husus, öznitelik seçiminin hiyerarşik olarak yapılmış olmasıdır [8].

Sonuç

Bu bildiride, kullanılan öznitelik seçme yöntemlerine ek olarak yeni bir öznitelik seçme yöntemi geliştirilmiştir. İncelenen tüm öznitelik seçme yöntemlerinin başarı derecelerini bulmak için yeni bir yöntem geliştirilmiştir. Bu yöntem gerçek veri kümesi ile test edilmiştir. Elde edilen sonuç, başarı kriterinin yeterince iyi derecelendirme yapabildiğidir.

Kaynakça

[1] Narendra P. M., Fukunaga K. "A Branch and Bound Algorithm for Feature Subset Selection" *IEEE Trans on Computers,* vol. c-26 ,no 9 pp 917-922, 1977

[2] Kittler J., "Feature Selection and Extraction" *Handbook of Pattern Recognition and Image Processing* T. Y. Young and K. S. Fu (editors) Academic Press Inc., London, 1986

[3] Pudil P.,Novovicova J., Kittler J. "Floating Search Methods in Feature Selection" *Pattern Recognition Letters* ,vol.15, pp 1119-1125, 1994

[4] Duda R. O. and Hart P. E., *Pattern Classification and Scene Analysis,* New York: John Wiley & Sons,(Preliminary Edition),1996

[5] Oja E. , " A simplified Neuron Model as a Principal Component Analyzer " *J. Math. Biol.* vol 15 , pp 267-273, 1982

[6] Chen L. H., Chang S. "An Adaptive Learning Algorithm for Principal Component Analysis "*IEEE Trans on Neural Networks,* vol 6 ,no 5 pp 1255-1263, 1995

[7] Diamantaras K. I., Kung S. Y. *Principal Component Neural Networks* New York: Wiley International Pub. , 1996

[8] Ünsalan C., Erçil E. "Classification of Rust Grades on Steel Surfaces Part 2" Boğaziçi Üniversitesi Teknik Rapor, FBE-IE-02/98-02, 1998

[9] Ünsalan C., Erçil E. “Classification of Rust Grades on Steel Surfaces Part 1” Boğaziçi Üniversitesi Teknik Rapor, FBE-IE-12/97-16, 1997