Fırat Üniversitesi Müh. Bil. Dergisi
31(2), 583-587, 2019

Alt Uzay k-NN ile Eritmato-Skuamöz Hastalık Türlerinin Sınıflandırılması

Duygu KAYA*
Fırat Üniversitesi, Mühendislik Fakültesi, Elektrik Elektronik Mühendisliği Bölümü, Elazığ, Türkiye
dgur@firat.edu.tr

(Geliş/Received: 04 /02/2019; Kabul/Accepted: 10/06/2019)

Öz: Veri analiz ve sınıflandırma tekniklerinin gelişmesinin sonucu olarak biyomedikal çalışmalarda akıllı hesaplama yöntemlerinin kullanımı oldukça önemli bir yer tutmaktadır. Eritmato-Skuamöz Hastalığı (ESD), altı çeşidi bulunan dermatoloji alanında büyük öneme sahip bir hastalıktır. Parçacık tanıma, veri madenciliği bileşenin özelliklerini tanımlamaya ve hastalığı teşhis etmeye yardımcı olur. Bu çalışmada UCI veri tabanından alınan veri setinden ESD'nin, Topluluk Öğrenim algoritmalarından alt uzay k-NN ile teşhis edilmesi amaçlanmıştır. Bu amaçla, hem klinik hem de histopatolojik özellikleri bir arada bulunduran ESD verileri ilk önce z normalizasyonu ile normalize edilmiştir. Normalize edilen veriler daha sonra alt uzay k-NN algoritması ile çapraz doğrulama uygulanarak sınıflandırılmıştır. Model başarım ölçütü için doğruluk, kesinlik, hassasiyet, F ölçütü ve Kappa katsayısı parametreleri kullanılmıştır ve sonuçlar yorumlanmıştır. Önerilen modelin doğruluğu %98.045'tir. Sonuçlar, kullanılan sınıflandırıcının ESD hastalık türlerinin sınıflandırılmasında faydalı olabileceğini göstermiştir. Ayrıca veri sayısının artırılarak derin öğrenme algoritmaları ile daha iyi sonuçlar alınabileceği öngörülmektedir

Anahtar kelimeler: Topluluk öğrenimi, Eritmato-Skuamöz Hastalığı, Alt uzay k-NN.

Classification of Erythmato-Squamous Disease Types with Subspace k-NN

Abstract: As a result of the development of data analysis and classification techniques, the use of intelligent calculation methods in biomedical studies is very important. Erythmato-Squamous Diseases (ESD) is a disease of great importance in the field of dermatology and which has six types. Particle recognition helps identify the properties of the data mining component and diagnose the disease. In this study, it is aimed to identify ESD from the data set obtained from UCI database with subspace k-NN from Ensemble Learning algorithms. For this purpose, ESD data which includes both clinical and histopathological features were first normalized by z normalization. Normalized data were then classified by applying cross validation with the subspace k-NN algorithm. Accuracy, precision, sensitivity, F score and Kappa coefficient parameters were used for the model performance criterion and the results were interpreted. The success of the proposed model is 98.045%. The results showed that the used classifier is useful in the classification of ESD disease types. In addition, it is predicted that by increasing the number of data, better results can be obtained with deep learning algorithms.

Key words: Ensemble learning, Erythmato-Squamous Diseases (ESD), Subspace k-NN.

1. Giriş

Eritmato-Skuamöz Hastalıkları (ESD) dermatoloji alanında büyük öneme sahip olup sedef hastalığı, seboreik dermatit, liken planus, pityriasis rosea, kronik dermatit ve pityriasis rubra pilaris olmak üzere bilinen altı çeşidi vardır [1, 2]. Bu hastalıklar genellikle cilt hücrelerinin kaybıyla ciltte kızarıklığa neden olur. Bu hastalık genetik veya çevresel sebeplerden dolayı ortaya çıksa da, geç çocukluk / erken ergenlik gibi yaşamın belirli dönemlerinde görülür [3]. Bu hastalığın ilerleyen aşamalarda kendine ait spesifik özellikler göstermesine rağmen, ilk aşamalarda başka bir hastalığın belirtileriyle aynı özellikler gösterebileceğinden uygun bir veri analizi yapmak gereklidir. Bu nedenle doğru tanı için bazen biyopsi gerekmektedir [4]. Literatürde, ESD hastalığının teşhisi için veri madenciliği yaygın olarak kullanılmaktadır.

Veri madenciliği alanında kullanılan çok sayıda model ve algoritma bulunmaktadır. Veri madenciliği, önceden bilinmeyen bilgilerin verilerden otomatik çıkarılması algoritmasını ifade eder. Ayrıca veri madenciliği yapay zeka, makine öğrenmesi ve istatistik gibi alanlarla ilişkili olarak değerli özelliklerin keşfedilmesini ve yorumlanmasına imkan sağlar [5].

Veri madenciliği temel olarak denetimli ve denetimsiz olmak üzere iki gruba ayrılırlar ve özelliklerin saptanması, verilerin kategorilere ve alt gruplara ayrılması amacıyla kullanılabilirler [6]. Denetimli veri madenciliğinde özellikler çıkarılır ve verileri önceden belli olan sınıflara yakınlığına göre sınıflandırılır ve

* Sorumlu Yazar: rorhan@firat.edu.tr. Yazarın ORCID Numarası: 0000-0002-6453-631X

sonuçlar yorumlanır. Literatürde sınıflandırma için denetimli algoritmalardan biri olarak düşünülen karar ağaçları, veri sınıflandırması ve tahmini için popüler bir algoritma olarak bilinmektedir [3].

Bu çalışmada hem klinik hem de histopatolojik özellikleri içeren ESD'nin verileri sınıflandırılma işlemi yapılmadan önce z normalizasyon yöntemi ile normalize edilmiştir. Normalize edilen veriler, alt uzay k-NN algoritmasına uygulanıp, elde edilen başarım ölçütleri analiz edilmiş ve yorumlanmıştır.

Makalenin geri kalanı şu şekilde düzenlenmiştir: 2. Bölümde, kullanılan ESD verilerinin klinik ve histopatolojik özellikleri ile verilerin alındığı veri tabanından bahsedilmiştir. 3. Bölümde literatürdeki ilgili çalışmalardan bahsedilmiştir. 4. bölümde, topluluk öğrenmesi ve özellikleri verilmiştir. 5. bölümde, kullanılan yöntemin etkinliğini göstermek için deneysel sonuçlar ve tartışmalar sunulmuştur. Son olarak, sonuçlar 6. bölümde sunulmuştur.

2. Kullanılan Veri Seti

Bu çalışmada kullanılan veri seti, Irvine Kaliforniya Üniversitesinde makine öğrenmesi deposundan (UCI) alınmıştır [7,8]. Ayrıca, bu veri seti ilk olarak N. Ilter (Gazi Üniversitesi) ve H.A. Guvenir (Bilkent Üniversitesi) tarafından hazırlanmıştır [1,2]. Veri seti Tablo 1'de gösterildiği gibi 12 klinik ve 22 histopatolojik olmak üzere toplam 34 özellikten, 366 örnek ve 1 hedef verisinden oluşmaktadır.

Tablo 1. Dermatoloji verisine ait klinik ve histopatolojik özellikler

Sınıflar	**Özellikler**	
	Klinik	**Histopatolojik**
sedef hastalığı	kızarıklık	melanin inkontinansı
seboreik dermatit	ölçekleme	sızıntıdaki eozinofiller
liken planus	kesin sınırlar	PNL sızması
pityriasis rosea	kaşıntı	papiller dermisin fibrozu
kronik dermatit	koebner fenomeni	ekzositoz
pityriasis rubra pilaris	poligonal papüller	akantozis
	foliküler papüller	hiperkeratoz
	oral mukozal tutulum	parakeratoz
	diz ve dirsek tutulumu	Rete Sırtı Kulübesi
	kafa derisi tutulumu	Rete sırtlarının uzaması
	aile öyküsü, (0 veya 1)	suprapapiller epidermisin incelmesi
	yaş	spongiform püstül
		munro microabcess
		fokal hipergranüloz
		granül tabakanın kaybolması
		Vacuolisation ve bazal tabaka hasarı
		spongiyoz
		retelerin testere dişi görünümü
		foliküler korna tapası
		perifoliküler parakeratoz
		inflamatuar mononükleer infiltrat
		bant benzeri sızma

Kullanılan verideki 8 örnek kaybından dolayı veri 358 örneğe düşürülüp sınıflandırma için 358x34 verisi kullanılmıştır. Kullanılan veri ve sayıları Tablo 2'de listelenmiştir

Tablo 2. Veri Dağılımı

Etiket	**Sınıflar**	**Örnek Sayısı**
1	sedef hastalığı	111
2	seboreik dermatit	61
3	liken planus	71
4	pityriasis rosea	47
5	kronik dermatit	48
6	pityriasis rubra pilaris	20
TOPLAM		358

Veri seti özelliklerinde aile üyelerinin birinde herhangi bir hastalığın olması durumunda aile öyküsü özelliği 1, aksi durumda 0 olmaktadır. Yaş özelliği sadece kişinin yaşını ifade etmektedir. Bunların dışındaki diğer tüm klinik ve histopatolojik özellikler 0 ile 3 aralığında derecelendirilmiştir. 0 özelliğin mevcut olmadığını, 3 olası en büyük miktarı, 1 ile 2 ise göreceli ara değerleri göstermektedir [1,2].

3. İlgili Çalışmalar

Veri seti ilk olarak Güvenir ve çalışma arkadaşları tarafından hazırlanmıştır [2]. Son yıllarda veri madenciliği ve topluluk öğrenimi algoritmaları ESD rahatsızlık teşhisinde kullanılmaktadır. Tablo 3'te ESD hastalığının teşhisi için literatürde kullanılan metotlar ve elde edilen doğruluk oranları verilmiştir.

Tablo 3. İlgili Çalışmalar

Çalışma	Metotlar	Doğruluk
Menai and Altayash [9]	Karar ağacı topluluğu	%96.72
Polat ve Güneş [10]	Karar ağacı sınıflandırıcısı	%86.18
K. M., M. L., L. S., M. H., M. J. Ve H. B., [4]	CART	%93.69
Bu çalışmada	Alt uzay k-NN	%98.045

4. Topluluk Öğrenimi

Denetimli sınıflandırma yöntemlerinde amaç, yeni örneklere bir sınıf etiketi atayan bir tahminleyici oluşturmaktır. Bir gözlemin sınıf etiketi, bir özellik vektörü ile tanımlanmaktadır. Gerçek dünya problemlerinde, algoritmaların sınıflandırma doğruluğunu azaltan, veri içerisinde bulunan önemli sayılamayacak özelliklerin etkisi özellik seçim veya boyut azaltma algoritmaları ile azaltılmaktadır [11].Böylece sınıflandırma performansı artacak, verinin en ayırt edici özellikleri bulunacaktır. Topluluk öğrenme teknikleri olarak bilinen çoklu sınıflandırıcıların birleştirilmesi, zayıf sınıflandırma performansını iyileştirmek için umut verici yöntemler olarak ortaya çıkmıştır. Bu teknikler birçok gerçek hayattaki uygulamalarda sınıflandırma hatasının önemli ölçüde azalmasına neden olmaktadır.

Bir topluluk modeli, bir dizi temel modeli eğiterek ve tahminlerini belirli bir toplama kuralı kullanarak birleştirir. Baz modellerin doğru ve çeşitli olması toplama sonuçlarının bireysel modellere göre daha doğru sonuçlar vermesini sağlayacaktır [12,13]. Özellikle, topluluk modellerinin başarısı temel olarak çeşitlilik özelliğine bağlıdır. Torbalama, en basit topluluk tekniğinden biri olup ilk önce Breiman tarafından kullanılmıştır [14]. Burada örnek, N boyutlu eğitim setinden muhtemel değişikliklerle B kez elde edilir. B önyükleme alt kümesinin her biri için ayrı bir CHAID modeli eğitilmiştir. Ortaya çıkan modellerin tahminleri, ağırlıklı oy çoğunluğu ile birleştirilmiştir. Tüm eğitim süreçleri paralel olarak çalışmaktadır.

Sınıflandırma için en basit ve en eski yöntemlerden biri, en yakın komşu (k-NN) sınıflandırıcısıdır. k-NN, yeni bireyin önceden k kategorisine ayrılmış bireylere yakınlığını inceler [15]. Yeni bir örnek geldiğinde, en yakın komşusuna bakar ve örneğin sınıfına karar verir. Basitliğine rağmen, kNN rekabetçi sonuçlar verir ve bazı durumlarda diğer karmaşık öğrenme algoritmalarından bile daha iyi performans gösterir. Bu çalışmada da literatürde kullanılan diğer algoritmalara göre alt uzay kNN'nin en iyi sonucu verdiği görülmüştür.

5. Deneysel Sonuçlar

Bu çalışmada 358 örnek ve 34 özelliğe sahip bir veri kullanılmıştır [8]. İstatistiksel normalizasyon, veriler arasında çok fark olduğunda veriyi bir düzende sıralar. Diğer bir deyişle, farklı sistemlerdeki veriler ortak bir sisteme taşınarak kıyaslanabilir hale getirilir. Bu veriler, klinik ve histopatolojik özellikleri birlikte bulundurmaktadır. Bu çalışmada normalizasyon tekniği olarak z normalizasyonu kullanılmıştır. Z normalizasyonu ile öncelikle verilerin Denklem 1'deki gibi standart sapması ve Denklem 2'deki gibi ortalama değeri hesaplanır. Daha sonra ortalama değer ve standart sapması kullanılarak z-normalizasyonu elde edilir Denklem 3 ile elde edilir

$$s = \sqrt{\sum_i (x_i - x_{ort})^2 / n} \quad (1)$$

$$x_{ort} = \frac{\sum_i x_i}{n} \quad (2)$$

$$z = \frac{x_i - x_{ort}}{s} \quad (3)$$

z normalizasyonu sonrası önerilen modellerle sınıflandırılan verilerin performansları, hata matrisi ve bazı başarım ölçüt parametreleri kullanılarak incelenmiş ve sonuçlar yorumlanmıştır.
z normalizasyonu sonrası 10 katlı çapraz doğrulama yapılarak sınıflandırılan verilerin hata matrisi Tablo 4'te, modele ait başarım ölçütleri ile kappa değeri de Tablo 5'te verilmiştir. Tablo 4'te verilen hata matrisinde GVS gerçek veri sayısını, BVS sınıflandırma sonrası elde edilen veri sayısını göstermektedir. Tablo 4'te görüldüğü gibi 2. ve 4. türe ait rahatsızlıkta sınıflandırma hatası mevcuttur. Sınıflandırıcının genel doğruluk oranı ise %98.045'tir.

Tablo 4. Hata Matrisi

Etiket	GVS	BVS	1	2	3	4	5	6
1	111	111	111	0	0	0	0	0
2	61	60	0	57	0	3	0	0
3	71	71	0	0	71	0	0	0
4	47	48	0	4	0	44	0	0
5	48	48	0	0	0	0	48	0
6	20	20	0	0	0	0	0	20

Ayrıca sınıflandırma doğruluğu dışında doğru tahmin edilen varların gerçek varlara oranını veren hassasiyet, doğru tahmin edilen varların toplam var tahminlere oranını veren kesinlik ve kesinlik ve duyarlılık ölçütlerini birlikte değerlendiren F ölçütü de Tablo 5'te verilmiştir.

Tablo 5. Başarım ölçüt parametrelerinin değerleri

Doğruluk		%98.045	
Kappa		0.975	
Etiket	**Hassasiyet**	**Kesinlik**	**F-skor**
1	1	1	1
2	0.93	0.95	0.94
3	1	1	1
4	0.94	0.92	0.93
5	1	1	1
6	1	1	1
Genel	0.9783	0.9783	0.9783

Bahsedilen ölçüm parametreleri dışında Kappa katsayısı da kategorik verilerin değerlendirilmesinde gözlemciler arasındaki uyumu ölçen istatistiktir. -1 ile +1 arasında bir değer alabilir. Kappa değeri +1 olduğunda gözlemciler arasında mükemmel uyum olduğunu, -1 olduğunda gözlemciler arasındaki uyumsuzluğun çok fazla olduğunu belirtir. Kappa katsayısının yorumlanmasında Tablo 6'da Landis ve Koch [16] tarafından önerilen uyum düzeyleri kullanılmaktadır. Uygulamada elde edilen Kappa katsayısı Tablo 5'te de görüldüğü gibi 0.975'tir. Bu da gözlemciler arasındaki uyumun çok yüksek olduğu göstermektedir.

Tablo 6. Kappa İstatistiği

Kappa	Uyum
<0.00	zayıf
0.00-0.20	önemsiz
0.21-0.40	düşük
0.41-0.60	Orta önemli
0.61-0.80	önemli
0.81-1.00	Çok yüksek

6. Sonuçlar ve Tartışma

Bu çalışmada dermatoloji alanında büyük öneme sahip, altı çeşidi bulunan Eritmato-Skuamöz Hastalığının sınıflandırılması için Topluluk Öğrenimi algoritmasına başvurulmuştur. Kullanılan veri seti UCI veri tabanından alınmıştır [8]. Çalışmada hem klinik hem de histopatolojik özellikleri bir arada bulunduran ESD verileri ilk önce z normalizasyonu ile normalize edilmiş daha sonra alt uzay k-NN topluluk öğrenimi algoritması ile 10 katlı çapraz doğrulama uygulanarak sınıflandırılmıştır. Model başarım ölçütü için doğruluk, kesinlik, hassasiyet ve F ölçütü değerleri ile Kappa katsayısı parametreleri elde edilmiş ve sonuçlar yorumlanmıştır. Önerilen modelin genel doğruluk oranı % 98.045 olup, % 1.955'lik bir hata oranına sahiptir. Sonuçlar, kullanılan sınıflandırıcının ESD hastalığının türlerinin sınıflandırılmasında faydalı olabileceğini göstermiştir. Ayrıca veri sayısının artırılarak derin öğrenme algoritmaları ile daha iyi sonuçlar alınabileceği öngörülmektedir.

Kaynaklar

[1] Güvenir H. A. ve Emeksiz N., "An expert system for the differential diagnosis of erythemato-squamous diseases," Expert Systems with Applications, 2000; 18: 43-49.

[2] Güvenir, H. A., Demiröz, G. ve İlter, N. "Learning differential diagnosis of erythemato-squamous diseases using voting feature intervals," Artificial Intelligence in Medicine, 1998; 13: 147-165.

[3] Elsayad, A., Dhaifullah, M., Nassef A. M., Analysis and Diagnosis of Erythemato-Squamous Diseases Using CHAID Decision Trees,15th International Multi-Conference on Systems, Signals & Devices (SSD), 2018.

[4] Maghooli, K., Langarizadeh, M., Shahmoradi. L, Habibikoolaee, M., Jebraeily, M. and Bouraghi, H., Differential Diagnosis of Erythmato Squamous Diseases Using Classification and Regression Tree, Acta Inform Med. 2016 OCT; 24(5): 338-34.

[5] Wang, L. and Sui, T. Z. "Application of Data Mining Technology Based on Neural Network in the Engineering," in International Conference on Wireless Communications, Networking and Mobile Computing, 2007, pp. 5544-5547.

[6] Baumgartner, C., Knowledge Discovery and Data Mining in Biomedicine, Thesis for Habilitation, University for Health Sciences, Medical Informatics and Technology, 2005.

[7] Bache, K. and Lichman, M.,"{UCI} Machine Learning Repository", University of California, Irvine, School of Information and Computer Sciences, 2013.

[8] http://archive.ics.uci.edu/ml/datasets/Dermatology.

[9] MEB, M. and N., Altayash, Differential Diagnosis of Erythemato-Squamous Diseases Using Ensemble of Decision Trees; Modern Advances in Applied Intelligence Springer;. 2014, pp. 369–77.

[10] Polat K. ve Güneş, S. A novel hybrid intelligent method based on C4. 5 decision tree classifier and one-against-all approach for multi-class classification problems. Expert Systems with Applications. 2009; 36(2):1587-92.

[11] Kaya, D. Biyomedikal İşaretlerin Sınıflandırılması İçin Akıllı Tekniklerin LabVIEW Ortamında Gerçeklenmesi. Fırat Üniversitesi, Fen Bilimleri Enstitüsü, Doktora Tezi, Mayıs 2018.

[12] Paleologo, G., Elisseeff, A. and Antonini, G. "Subagging for credit scoring models," European Journal of Operational Research, 2010, vol. 201, pp. 490-499.

[13] Gang, W., Jinxing, H., Jian, M. and Hongbing, J. "A comparative assessment of ensemble learning for credit scoring," Expert Syst. Appl., 2011; 38: pp. 223-230.

[14] Breiman, L., Friedman, J., Stone, C. J., ve Olshen, R. A., Classification and Regression Trees: Taylor & Francis, 1984.

[15] Kaya, D., Türk, M. ve Kaya, T., En Yakın Komşu Algoritması Kullanılarak EEG Sinyallerine Boyut Azaltmanın Etkilerinin İncelenmesi, El-Cezeri Fen ve Mühendislik Dergisi, 2018; 5(2): 591 – 595.

[16] J.R. Landis & G.G.Koch, The measurement of observer agreement for categorical data. Biometrics, 1977; 33: 159-174.