Karar ağacı tekniğinin tiroid hastalığının tanısında kullanılmasına ilişkin bir çalışma

Nilüfer Yurtay[1] nyurtay@sakarya.edu.tr
M.Fatih Adak[1] fatihadak@sakarya.edu.tr
Deniz Dural[1] ddural@sakarya.edu.tr
Soydan Serttaş[1] nyurtay@sakarya.edu.tr

[1]Bilgisayar Müh. Bölümü
Sakarya Üniversitesi,
Türkiye

Özet : Verilerin sınıflandırma yöntemlerinden biri de karar ağaçlarıdır. Karar ağaçlarının oluşturulmasında çok sayıda öğrenme yöntemi mevcuttur. Bunlardan birisi de Gini algoritmasıdır. Bu çalışmada UCI veri setlerinden tiroid veri seti üzerinde yapılan bir karar ağacı uygulaması ve bu uygulamanın tiroid hastalığının teşhisindeki performansı irdelenmiştir.

Anahtar kelimeler: Karar ağacı, Gini algoritması, Tiroid tanısı

A study on use of decision tree method in the diagnosis of thyroid disease

Abstract: Decision tree is one of the methods of data classification and contains many learning algorithms. Gini is one of these algorithms. UCI thyroid data sets were used in this study, developed a decision tree on this data sets and performance of the decision tree are investigated for diagnosis thyroid disease.

Keywords: Decision tree, Gini Algorithm, Thyroid diagnosis.

Giriş

Tıbbi veri madenciliği diğer alanlardan farklıdır çünkü veriler heterojendir. Medikal veri madenciliği metodları veri kaynaklarını hem teknik hemde sosyal sebeplerden ötürü oluşan kayıp değerlerin yaygınlaşmasını heterojen yapıda ele almalıdır(Cios and Moore, 2002). Yapılan bir çalışmada endüstriyel İki farklı medikal veri seti üzerinde ticari programlarda olmayan 3 farklı Fuzzy yöntemi ile 11 özellik seçim metodu kullanılarak bir veri madenciliği yapılmıştır. Çalışma, fuzzy yöntemin iyi sonuçlar verdiği göstermiştir(Ghazavi and Liao, 2008). Efron bootstrap yöntemi ile sınırlı sayıda veri seti olan durumlarda veri madenciliğinin güvenirliliği sağlanmıştır. Uygulanan veri madenciliği yöntemleri bootstrap yaklaşımı ile küçük veri setlerinde güvenirlilik noktaları belirlenebilmektedir (Smith et al., 2009). Bellazzi and Zupan (2008), öngörüsel veri madenciliğinin tıbbi araştırmacılar için zorunlu bir araç haline geldiğini ifade etmişlerdir. Tayvan Medikal Merkezi Acil Servis Bölümünde 501 iki aşamalı (Wand yöntemi ve K-means) kümeleme analizi ve karar ağacı analizi ile yapılan veri madenciliği sayesinde anormal teşhislerde hemşirelerin doktorlardan daha iyi olduğu görülmüştür(Lin et al., 2010). Kronik astım hastalarının ani ataklarının önlenmesi için sık sık gözlemlenmeleri gerekir. Hastaların bu durumlarını tahmin etmek için Lee et al. (2011), bio-sinyaller ve çevresel faktörleri analiz eden bir veri madenciliği yöntemi geliştirmiştir. Bu yöntemde şablon tabanlı bir karar ağacı ve şablon tabanlı sınıf birliktelik kuralı oluşturulmuştur. Sonuçlar için Tayvan'daki bir hastanete ait çocuk alerjik astım hastalarına ait veri seti kullanılmış ve kronik astım atakları tahmininde yüksek doğruluk oranları bulunmuştur. Tayvanda web tabanlı bir rapor sistemi ile 725 hasta üzerinde yapılan veri madenciliğinde ilk önce özellik seçimi uygulanmış sonra 10 kritik faktör bağımlı değişkenleri tahmin etmek için kullanılmıştır. Lee et al. (2011) tarafından yapay sinir ağları analizi, hastaların kritik durumları ile ilgili bir tahmin modeli geliştirmek için

uygulanmış ve çoklu değişken aşamalı lojistik regresyon ile karşılaştırılmıştır. Yapay sinir ağlarının diğer yönteme göre daha doğru sonuç verdiği görülmüştür. Bir diğer çalışma Delen et al. (2009) tarafından yapılmış ve bu çalışmada yapay sinir ağları ve karar ağaçları modelleri 23 değişken ve 193373 kayıt üzerinde tahmin için kullanılmıştır. Sağlıklı olan veya olmayanlar için yapılan bir sınıflandırmada bu iki popüler yöntem kesin bir doğrulukla kullanılabilir. Geniş çaplı klinik veriler geleneksel Çin tıbbi araştırmaları için temel bir deneysel bilgidir. Geleneksel Çin tıbbi için oluşturulan bir veri ambarı üzerinde yapılan veri madenciliği ile deneysel bilgiler elde edilmektedir (Zhou et al., 2010). Diyaliz hastaları sağlıklı olmayan tedavi davranışları ve uzun süren diyaliz tedavisi sebebiyle tekrar hastane kontrolüne ihtiyaç duymaktadır. Yeh et al., (2011) tarafından karar ağacı veri madenciliği yöntemi ile diyaliz hastalarının hastane kontrollerini analiz eden ve acil tedavi durumlarını hastane öncesi bildirebilen bir sistem geliştirilmiştir.

Bu çalışmada ise yukarıda tıbbi alanlardaki başarısı ifade edilmiş olan veri madenciliği yöntemlerinden karar ağacı yöntemi kullanılarak Troid hastalığının teşhisi amaçlanmıştır.

Method

Veri kaynağı

Tiroid tanısında kullanılabilecek karar desteği sağlamayı amaçlayan pek çok çalışma yapılmıştır. Bu çalışmalarda yapay zeka tekniklerinin kullanımı yaygındır. Polat et al. (2007), tiroid tanısında fuzzy ağırlıklı bir yapay bağışıklık sistemi kullanarak sınıflama yapmış ve %85 doğrukla tanı elde etmişlerdir. Kodaz et al.(2009), Tiroid tanısı için yapay bağışıklık sistemi kullanarak %95.90 oranında başarılı bir sınıflandırma yapan bir çalışmayı ortaya koymuşlardır. Temurtaş (2009), Troid tanısı için MLNN with LM (3 x FC), PNN (3 xFC), LVQ (3 xFC), MLNN with LM (10 xFC), PNN (10 xFC) ve LVQ (10 xFC) modellerinin performansını incelemiş ve çok başarılı sonuçlar elde etmiştir. Bir diğer çalışmada ESTDD(expert system for thyroid disease) ile %95.33 doğrulukla tiroid tanısı yapılmıştır (Keleş and Keleş, 2008). Doğantekin et al. (2010), ADSTG (automatic diagnosis system based on thyroid gland) çalışmasında %97.67 oranında doğru tanıya ulaşmışlardır. Doğantekin et al. (2011) bir sonraki çalışmalarında da Generalized Discriminant Analysis and Wavelet Support Vector Machine System (GDA_WSVM) method ile uzman sistem geliştirmişler ve tiroid tanısında %91.86 oranında doğru sınıflama elde etmişlerdir.

Öğrenme seti UCI veri tabanındaki Tiroid veri setidir. Tablo 1'de bu vei setinde yer alan 6 nitelik değeri gösterilmiştir. 5 giriş 1 sınıf niteliği olup sınıflar, normal, hyper ve hypo olarak belirlenmiştir (Uci,2012). Her sınıftaki veri sayıları

Sınıf 1: (normal) 150
Sınıf 2: (hyper) 35
Sınıf 3: (hypo) 30

olarak verilmiştir.

Table 1: Tiroid veri seti nitelikleri

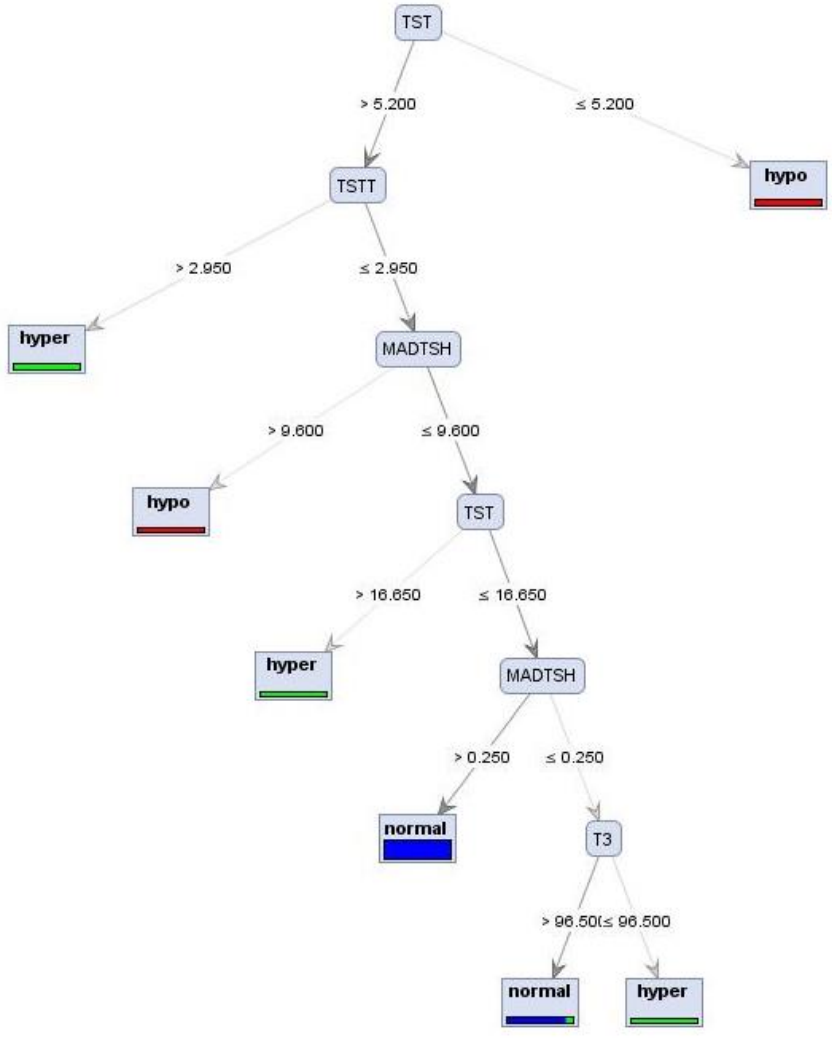
Nitelik no	Nitelik değeri
1	T3 değeri
2	Toplam serum thyroxin (TST).
3	Toplam serum triodothyronine (TSTT) .
4	Basal thyroid-stimulating hormonu (TSH)
5	200 mikro gram thyrotropin hormonu verilmesi sonrasındaki TSH maksimum değeri (MADTSH)
6	Sınıf (1 = normal, 2 = hyper, 3 = hypo)

Karar Ağacının Oluşturulması

Genel olarak veri madenciliği yöntemleri iki sınıfa ayrılabilir: Öngörü Yöntemleri (Prediction Methods) ve Tanımlayıcı Yöntemler (Description Methods). Verinin içerdiği ortak özelliklere göre ayrıştırılması işlemi de sınıflandırma olarak anılır. Sınıflandırma bir öğrenme algoritmasına dayanır. Amaç sınıflandırma modelinin oluşturulmasıdır. Sınıfı bilinmeyen herhangi bir verinin sınıfının belirlenmesi sürecidir denebilir. Karar ağaçları da sınıflandırma yöntemlerinden biridir. Bir deneme kümesi modelin doğruluğunu belirlemek için kullanılır. Genellikle verilen veri kümesi **öğrenme** ve **test kümesi** olarak ikiye ayrılır. Öğrenme kümesi modelin oluşturulmasında, deneme kümesi modelin doğrulanmasında kullanılır. Karar ağaçları oluşturmak için farklı algoritmalar geliştirilmiştir. Bunlardan biri de Gini Algoritmasıdır.

Gini Algoritması, ikili bölünmeler şeklinde gerçekleşen bir sınıflandırma yöntemi olup, ikili yinelemeli bölümleme için en iyi bilinen kurallardandır. Her bir ağaç farklı bir stil ile gelişir. Algoritma nitelik değerlerinin sol ve sağda olmak üzere ikili bölünmeler şeklinde ayrılması temeline dayanır (Özkan,2008).

Bu çalışmada 215 veri 3 bölüme ayrılmış, 2 bölüm eğitim ve 1 bölüm test olacak şekilde 3 kez Gini Algoritması kullanılmıştır. Şekil 1'de elde edilen ağaçlardan biri verilmiştir. (Diğer ağaçlar için Ek A'ya bakınız).



1-2 nolu bölümlerle eğitim için karar ağacı

Şekil 1: Troid verileri için 3 bölümlemeli eğitim verileri ile elde edilen bir karar ağacı

Karar ağacının değerlendirilmesi-ROC analizi

ROC (Receiver Operating Characteric=alıcı işletim karakteristiği) analizi, bir duyarlılık ve seçicilik değeri kullanarak tanı koymanın getirdiği sakıncaları ortadan kaldırmak için geliştirilmiş istatistik değerlendirme yöntemidir (Tomak ve Bek, 2010). Tanı testlerinde olumlu ya da olumsuz kararın doğruluk derecesi önemlidir. Pozitif ya da negatif kararların her biri için doğruluk düzeyini gösteren ölçütler vardır.

Tablo 2. Roc analiz için kullanılan parametreler

	Gerçek Durum			**Açıklama**
Test Sonucu	Pozitif	Negatif	Toplam	*DP:* Gerçek durum pozitifken test sonucu da pozitif çıkan durumlar
Pozitif	Doğru pozitif (DP)	Yanlış Pozitif (YP)	(DP+YP)	*YN:* Gerçek durum pozitifken test sonucu negatif çıkan durumlar
Negatif	Yanlış Negatif (YN)	Doğru Negatif (DN)	(YN+DN)	*YP:* Gerçek durum negatifken test sonucu pozitif çıkan durumlar
Toplam	(DP+YN)	(YP+DN)	(DP+YN+YP+DN)	*DN:* Gerçek durum negatifken test sonucu da negatif çıkan durumlar

Doğruluk(Accuracy): (DP+DN) / (DP+YP+YN+DN)

Duyarlılık (Sensitivity): DP/DP+YN

Seçicilik (Specificity): DN/DN+YP

Duyarlılık, testin, gerçek pozitif durumlar içinden pozitif olan durumları ayırma yeteneğini belirtirken, seçicilik de testin, gerçek negatif durumlar içinden negatif olan durumları ayırma yeteneği olarak ifade edilebilir. Duyarlılık ve seçicilik değerleri, testin araştırılan durumla ilgili olanlarla olmayanları birbirinden ne kadar iyi ayırt edip etmediğini tanımlar(Dirican, 2001).

Bu tanımlamalar doğrultusunda Şekil 1'de ifade edilen karar ağacı için elde edilen doğruluk, duyarlılık ve hassasiyet sonuçları tablo 4'de gösterilmiştir. Tablo 4 hazırlanırken, her bir test aşamasında tablo 3'de belirtilen durumlar göz önünde tutulmuştur. Genel doğruluk değeri %94,5, genel duyarlılık değeri %85,3 ve genel seçicilik değeri de %97,3 olarak elde edilmiştir.

Tablo 3: Doğruluk değerlerinin hesaplanmasında kullanılan kriterler

Orijinal Sonuç	Karar Ağacı Sonucu		
	Hyper	Normal	Hypo
Hyper	DP	YN	YN
Normal	YP	DN	DN
Hypo	YP	DN	DN

Tablo 4: Test sonuçlarına göre doğruluk, duyarlılık ve seçicilik değerleri

Test No	Temel değişken	Doğruluk(%)	Test Doğruluk Ortalaması (%)	Genel doğruluk Ortalaması (%)	Duyarlılık	Genel duyarlılık Ortalaması (%)	Seçicilik	Genel seçicilik Ortalaması (%)
1	Hyper	95,8	94,4	94,5	90,9	85,3	96,0	97,3
1	normal	91,6						
1	Hypo	95,8						
2	Hyper	95,8	94,4		83,3		98,0	
2	normal	91,6						
2	Hypo	95,8						
3	Hyper	94,3	94,8		81,8		98,0	
3	normal	92,9						
3	Hypo	97,1						

Sonuç

Veri madenciliği kullanımının yaygınlaşmakta olduğu alanlardan biri de tıp uygulamalarıdır. Birçok kişiden alınan tahliller ve yapılan operasyonlardan elde edilebilecek veri topluluklarından, çeşitli algoritmalar aracılığıyla karar ağaçlarına erişim sağlanabilmekte, aracı veriler neticesinde ilk etapta görünemeyen örtülü bilgilerin keşfi yapılabilmektedir. Bu çalışmada troid hastalığının tanısına ilişkin bir karar ağacı geliştirilmiş ve bu karar ağacının performansı incelenmiştir. Elde edilen karar ağaçları birbirinden performans açısından çok büyük farklılıklar göstermemektedir. Bu ağaçların her biri troid tanısı için kullanılabilir.

Kaynaklar

Ali Keleş, Aytürk Keleş (2008). ESTDD: Expert system for thyroid diseases diagnosis. Expert Systems with Applications. 34:242–246.

Chao-Hui Lee, Jessie Chia-Yu Chen, Vincent S. Tseng (2011). A novel data mining mechanism considering bio-signal and environmental data with applications on asthma monitoring . *Computer Methods and Programs in Biomedicine*. *101: 44-61*.

Dirican A. (2001). Evaluation of the diagnostic test's performance and their comparisons. *Cerrahpaşa J Med* , 32 (1): 25-30.

Dursun Delen, Christie Fuller, Charles McCann, Deepa Ray (2009). Analysis of healthcare coverage: A data mining approach. *Expert Systems with Applications*. *36: 995-1003.*

Esin Doğantekin, Akif Dogantekin, Derya Avci (2010). An automatic diagnosis system based on thyroid gland: ADSTG. Expert Systems with Applications. 37: 6368–6372

Esin Doğantekin, Akif Dogantekin, Derya Avci (2011). An expert system based on Generalized Discriminant Analysis and Wavelet Support Vector Machine for diagnosis of thyroid diseases. Expert Systems with Applications. 38:146–150.

Feyzullah Temurtaş (2009). A comparative study on thyroid disease diagnosis using neural networks. Expert Systems with Applications 36: 944–949.

Halife Kodaz, Seral Özşen, Ahmet Arslan, Salih Güneş (2009). Medical application of information gain based artificial immune recognition system (AIRS): Diagnosis of thyroid disease. Expert Systems with Applications. 36:3086–3092

Jinn-Yi Yeh, Tai-Hsi Wu, Chuan-Wei Tsao (2011). Using data mining techniques to predict hospitalization of hemodialysis patients. *Decision Support Systems. 50:439-448.*

Kemal Polat, Seral Şahan, Salih Güneş (2007). A novel hybrid method based on artificial immune recognition system (AIRS) with fuzzy weighted pre-processing for thyroid disease diagnosis. Expert Systems with Applications 32:1141–1147.

Krzysztof J. Cios, G. William Moore (2002). Uniqueness of medical data mining, Artificial Intelligence in Medicine . 26:1-24.

M.R. Smith, X. Wang, R.M. Rangayyan (2009). Evaluation of the sensitivity of a medical data-mining application to the number of elements in small databases . *Biomedical Signal Processing and Control*. *4: 262-268.*

Özkan, Y., Veri Madenciliği Yöntemleri, Papatya Yayıncılık, 2008.

Riccardo Bellazzi, Blaz Zupan (2008). Predictive data mining in clinical medicine: Current issues and guidelines. *International Journal of Medical Informatics*. *77: 81-97.*

Sean N. Ghazavi, Thunshun W. Liao (2008). Medical data mining by fuzzy modeling with selected features. *Artificial Intelligence in Medicine*. *43:195-206.*

Ting-Ting Lee, Chieh-Yu Liu, Ya-Hui Kuo, Mary Etta Mills, Jian-Guo Fong, Cheyu Hung (2011). Application of data mining to the identification of critical factors in patient falls using a web-based reporting system. *International Journal of Medical Informatics*. *80:141-150.*

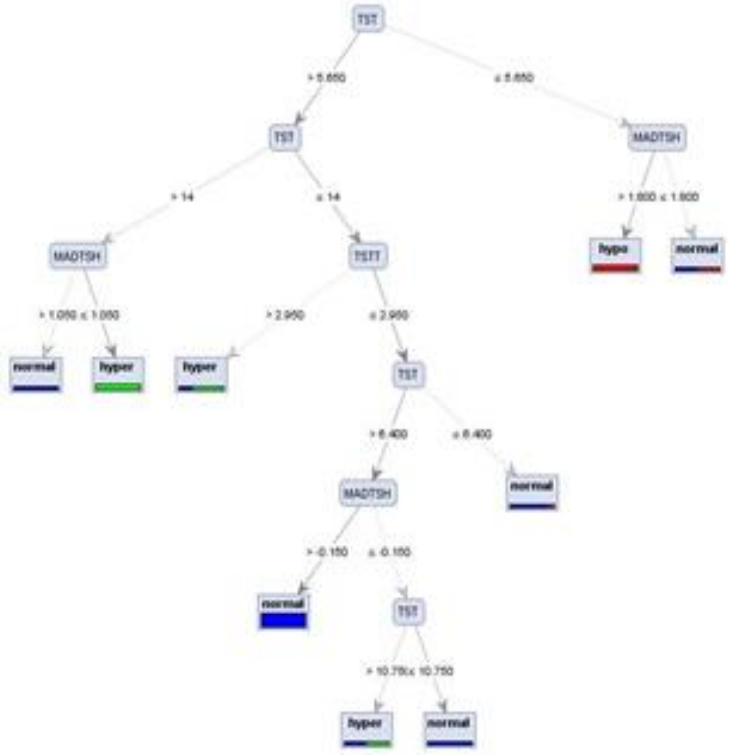
TOMAK. L., BEK.Y. (2010). İşlem Karakteristik Eğrisi Analizi Ve Eğri Altında Kalan Alanların Karşılaştırılması, Journal of Experimental and Clinical Medicine, Vol:27, no:2, s:58-65.

Uci,http://archive.ics.uci.edu/ml/machine-learning-databases/thyroid-disease/new-thyroid.names, 20-11-2012.

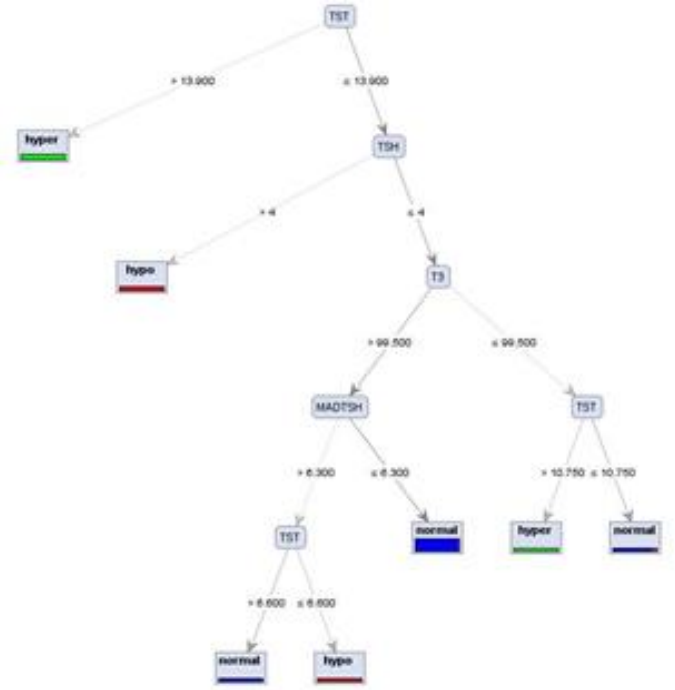
Wen-Tsann Lin, Shen-Tsu Wang, Ta-Cheng Chiang, Yu-xin Shi, Wei-yu Chen, Huei-min Chen (2010). Abnormal diagnosis of Emergency Department triage explored with data mining technology: An Emergency Department at a Medical Center in Taiwan taken as an example. *Expert Systems with Applications. 37: 2733-274.*

Xuezhong Zhou, Shibo Chen, Baoyan Liu, Runsun Zhang, Yinghui Wang, Ping Li, Yufeng Guo, Hua Zhang, Zhuye Gao, Xiufeng Yan (2010). Development of traditional Chinese medicine clinical data warehouse for medical knowledge discovery and decision support. *Artificial Intelligence in Medicine*. *48:139-152.*

Ek A



Şekil 2: 1-3 nolu bölümlerle eğitim için karar ağacı



Şekil 3: 2-3 nolu bölümlerle eğitim için karar ağacı