ISSN:1306-3111
e-Journal of New World Sciences Academy
2011, Volume: 6, Number: 1, Article Number: 1A0145

ENGINEERING SCIENCES
Received: October 2010
Accepted: January 2011
Series : 1A
ISSN : 1308-7231

Ömer Akgöbek[1]
Serkan Kaya[2]
Zirve University[1]
Harran University[2]
serkankaya@harran.edu.tr
Gaziantep-Turkey

VERİ MADENCİLİĞİ TEKNİKLERİ İLE VERİ KÜMELERİNDEN BİLGİ KEŞFİ: MEDİKAL VERİ MADENCİLİĞİ UYGULAMASI

ÖZET

Veri tabanı sistemlerinin büyük bir hızla gelişmesi, artan kullanımı ve bu sistemlerdeki bilgilerin önemi bu sistemlerden nasıl yararlanılacağı problemini de beraberinde getirmiştir. Bu sistemlerin en çok kullanıldığı alanların başında da 'tıp' gelmektedir. Günümüzde hastalara ait tüm laboratuar sonuçları, hastanın hikâyesi gibi bilgilerin yanı sıra çekilen film ve röntgen görüntüleri dahi veri tabanlarında tutulmaktadır. Bu veri tabanlarından geleneksel sorgulama metotlarıyla bilgiyi süzmek ve bu bilgileri raporlar halinde sunmak bilgiler içerisinde saklı bulunan gizli-önemli kuralların ortaya çıkmasını sağlamaz. Bundan dolayı veri tabanlarından bilgi keşfi için bu alanda kullanılan veri madenciliği tekniklerinin kullanılmasını kaçınılmaz yapmaktadır. Bu çalışmada veri madenciliği tekniklerinden *REX-1* algoritması kullanılarak medikal alanda kullanılan ve gerçek hayattan alınan Wisconsin Breast Cancer, Ljubljana Breast Cancer, Dermatology, Hepatitis ve Diabetes örnek setleri üzerinde bilgi keşfi yapılarak kural tabanı oluşturulmuştur. Elde edilen sonuçlar bu alanda yaygın olarak kullanılan C4.5, NavieBayes, PART, CN2, CORE, GA-SVM gibi algoritmalarla doğruluk oranlarına göre test edilmiştir.

Anahtar Kelimeler: Veri Madenciliği, Bilgi Keşfi, Medikal Veri Madenciliği, Sınıflandırma

KNOWLEDGE DISCOVERY FROM DATA SETS THROUGH DATA MINING TECHNIQUES: APPLICATION TO MEDICAL DATA MINING

ABSTRACT

A rapid development and growing use of database systems, and the importance of information stored in these systems raise the issue of how to make the best use of these systems. One of the leading area where the database systems are mostly used is 'medicine'. Today, all patients' laboratory results, patient history as well as x-ray images and more are kept in databases. These of the traditional database query and report methods to filter information and to present reports does not always provide the important hidden rules contained in the stored data. Therefore, the use of data mining techniques used for knowledge discovery in this area from databases is inevitable. In this study, the rules base is created thorough the knowledge discovery by employing REX-1 algorithm, a data mining technique, on the Wisconsin Breast Cancer, Ljubljana Breast Cancer, Dermatology, Hepatitis and Diabetes sample sets, which are real life data and commonly used in the medical field. In terms of the accuracy rate, the results of this study were compared to the results of the algorithms widely used in this field, such as C4.5, NavieBayes, PART, CN2, CORE, GA-SVM.

Keywords: Data Mining, Knowledge-Discovery, Medical Data Mining, Classification

1. GİRİŞ (INTRODUCTION)

Basit bir tanım ile veri madenciliği, büyük ölçekli veriler arasından bilgiye ulaşma ve bu verilerden genellenmiş bilgiler elde etmedir. Başka bir ifade ile büyük veri yığınları içerisinden gelecekle ilgili tahminde bulunabilmemizi sağlayacak bağıntıların bilgisayar programı kullanarak aranması ve ortaya çıkarılmasıdır. Veri madenciliği için farklı tanımlar yapmak da mümkündür: Eldeki verilerden üstü kapalı, çok net olmayan, önceden bilinmeyen ancak potansiyel olarak kullanışlı bilginin çıkarılmasıdır. Bu da; kümeleme, veri özetleme, değişikliklerin analizi, sapmaların tespiti gibi belirli sayıda teknik yaklaşımları içerir. Başka bir deyişle, veri madenciliği, verilerin içerisindeki desenlerin, ilişkilerin, değişimlerin, düzensizliklerin, kuralların ve istatistiksel olarak önemli olan yapıların yarı otomatik olarak keşfedilmesidir (Vahaplar ve İnceoğlu, 2001). Temel olarak veri madenciliği, veri setleri arasındaki desenlerin ya da düzenin, gizli bilgilerin, verinin analizi ve yazılım tekniklerinin kullanılarak ortaya çıkarılmasıdır. Veriler arasındaki ilişkiyi, kuralları ve özellikleri, daha önceden fark edilmemiş veri desenlerini tespit edebilmek için uygun bilgisayar yazılımlarının kullanılmasını gerektirir.

Veri madenciliği için literatürde birçok kavram kullanılmaktadır (veritabanlarında bilgi madenciliği-knowledge mining from databases), bilgi çıkarımı-knowledge extraction, veri ve örüntü analizi-data/pattern analysis, veri arkeolojisi, veritabanlarında bilgi keşfi- Knowledge Discovery From Databases). Bunların arasındaki en yaygın kullanım 'Veritabanlarında Bilgi Keşfi'dir. Alternatif olarak veri madenciliği aslında bilgi keşfi sürecinin bir parçası şeklinde kabul görmektedir. Bilgi keşfinin adımları aşağıda verilmiştir:

- Veri temizleme (veri setinde yer alan gürültülü ve/veya tutarsız verileri çıkarmak)
- Veri bütünleştirme (birden fazla veritabanında ve/veya tabloda dağınık olarak bulunan bilgileri birleştirebilmek)
- Veri seçme (yapılacak olan analizle ilgili olan verileri belirlemek, gereksiz karakteristikleri elemek)
- Veri dönüşümü (verinin veri madenciliği teknikleriyle kullanılabilecek hale dönüşümünü gerçekleştirmek)
- Veri madenciliği (veri örüntülerini yakalayabilmek için akıllı metotları/algoritmaları uygulamak)
- Örüntü değerlendirme (bazı ölçümlere göre elde edilmiş bilgiyi temsil eden ilginç örüntüleri tanımlamak)
- Bilgi sunumu (madencilik işleminden geçmiş olan bilginin kullanıcıya sunumunu gerçekleştirmek).

Veri madenciliğinin her adımında, kullanıcı ve bilgi tabanıyla sürekli olarak bir etkileşim vardır. Genelleştirilen bilgiler kural seti şeklinde kullanıcıya gösterilebilir ve daha sonra kullanılmak üzere bilgi tabanına kaydedilebilir. Buna göre veri madenciliği işlemi, gizli kalmış tüm örüntüler bulunana kadar devam eder. Bu işlemin başarısı kullanılan algoritmaya göre değişiklik gösterir. Bir veri madenciliği sistemi, aşağıdaki temel bileşenlere sahiptir:

- Veritabanı, veri ambarı ve diğer depolama teknikleri
- Veritabanı ya da veri ambarı sunucusu
- Bilgi tabanı
- Veri madenciliği motoru
- Örüntü değerlendirme
- Kullanıcı ara yüzü

Veri madenciliği; istatistik, veri tabanları, programlama teknikleri ve yüksek performanslı işlem gibi temelleri içermenin yanında, eldeki verilerden anlamlı ve kullanışlı bilgiyi çıkarmaya yarayacak tümevarım işlemlerini formüle etmeye ve uygulamaya yönelik çalışmaların bütününü de kapsar (Akgöbek, 2006).

Veri madenciliği ve bilgi keşfi (data mining & knowledge discovery), özellikle mühendislik, ticaret, tıp ve eğitim alanlarındaki uygulamalarda yeni ve temel bir araştırma sahası olarak ortaya çıkmaya başlamıştır. Veri madenciliği teknikleri, geniş veri kümelerinden anlamlı bilgileri, düzensizlikleri ve veriler arasındaki ilişkileri ortaya çıkarmakta kullanılır. Bu sayede, web üzerinde filtrelemeler, DNA sıraları içerisinde genlerin tespiti, ekonomideki eğilim ve düzensizliklerin tespiti, Internet üzerinden alışveriş yapan müşterilerin alışkanlıkları gibi karar verme mekanizmaları için önemli bulgular elde edilebilir (SAS Institute Inc., 1999, Vahaplar ve İnceoğlu, 2001).

Veri madenciliğinde kullanılacak olan veritabanları geniş, yüksek hacimli ve/veya dağıtık şekilde bulunabilir. Kullanılacak olan tekniklerin bu yapılara uygun olarak tasarlanması çok önemlidir. Bu tekniklerin bilgisayar belleğine sığmayacak kadar büyük veya farklı coğrafi konumlardaki bilgiyi işleme yeteneğine de sahip olması gerekir (Vahaplar ve İnceoğlu, 2001).

2. ÇALIŞMANIN ÖNEMİ (RESEARCH SIGNIFICANCE)

Büyük bir hızla gelişen veri tabanı yönetim sistemleri, bu sistemlerdeki bilgilerin önemini bu sistemlerden nasıl yararlanılacağı problemini de beraberinde getirmiştir. Veri tabanı yönetim sistemlerinin en çok kullanıldığı alanların başında da sağlık sektörü gelmektedir. Hastalara ait tüm laboratuar sonuçları, çekilen film ve röntgen görüntüleri dahi veri tabanlarında tutulmaktadır. Bu veri tabanlarından geleneksel sorgulama metotlarıyla bilgiyi süzmek ve bu bilgileri raporlar halinde sunmak bilgiler içerisinde saklı bulunan gizli-önemli kuralların ortaya çıkmasını sağlamaz. Bundan dolayı veri tabanlarından bilgi keşfi için bu alanda kullanılan veri madenciliği tekniklerinin kullanılması gerekmektedir. Bu çalışmada veri madenciliği teknikleri kullanılarak gerçek hayattan alınan bazı medikal veri setlerinden bilgi keşfi yapılmış ve bu bilgiler kural setine dönüştürülmüştür. Bu kural setleri kullanılarak bir uzaman sistemin alt yapısı kolaylıkla oluşturulabilir.

3. REX-1 ALGORİTMASI (REX-1 ALGORITHM)

REX-1 (REX: Rule Extraction) algoritması verilen örnek setinden genelleştirilmiş EĞER-İSE (IF-THEN) kuralları elde etmek için geliştirilmiştir (Akgöbek ve diğ., 2006). Kural çıkarma işlemini kapsama (doğrudan kural çıkarma) metodunu kullanarak yerine getirmektedir. Bu algoritmada entropi yardımıyla önem derecesi yüksek olan karakteristiklere öncelik verilerek daha genel kuralların elde edilmesi amaçlanmaktadır.

3.1. Algoritmanın Tanımı (Description of Algorithm)

REX-1 algoritması karar ağacı üreten algoritmalarda da kullanılan entropi ölçüsünü kullanır. Karar ağacı üreten algoritmaların en iyi bilinenleri ID3 ve C4.5'tir. Bu algoritmalar verilen örnek setinden genelleştirilmiş kurallar çıkarmak için öncelikle her karakteristiğe ait değerlerin ve karakteristiğin entropisini hesaplayarak karar ağacı oluştururlar. Hesaplanan bu entropi değerlerine göre ayrı ayrı bilgi kazançları hesaplanır. Bilgi kazancı en yüksek olan karakteristik ağıcın kökü olarak seçilir. Diğer karakteristikler buna göre yeniden düzenlenir. Aynı işlem diğer alt setler için de tekrarlanır. En son aşamada oluşan karar ağacına göre kurallar çıkarılır. C4.5 ve ID3 gibi algoritmalarda

olduğu gibi, her aşamada yeni bir karar ağacı oluşturmak ve örnekleri bu karar ağacının köküne göre alt setlere ayırmak çok zor bir işlemdir. Bu algoritmada işlemler daha basit bir kural çıkarma prosedürü kullanılarak yerine getirilmektedir. ID3 ve C4.5 algoritmaları karar ağacı oluşturduktan sonra, karar ağacını kural setine dönüştürürler. REX-1 algoritması entropi ölçüsünü kullanarak ve karar ağacı oluşturmadan doğrudan kural üretmektedir (Akgöbek ve diğ., 2006).

REX-1 algoritması ilk önce örnek setinde bulunan karakteristik ve bu karakteristiklere ait değerlerin entropilerini hesaplayarak, düzensizliği az olan karakteristiklere öncelik vererek kural çıkarma işlemini gerçekleştirir. Entropisi düşük olan karakteristiklerin bilgi kazancı yüksektir. Böylece örnek setinde bilgi değeri en büyük olan karakteristiklere öncelik verilir. Karakteristiklerin önem dereceleri belirlendikten sonra sırasıyla ilk karakteristikteki değerleri ve ilk örnekten başlayarak tekli kombinasyonlara göre kural üretme işlemine başlanır. Üretilen her kural tarafından sınıflandırılan örnekler seçilerek sınıflandırılmış örnek olarak ayıklanır. Tekli kombinasyonlardan sonra sırasıyla ikili, üçlü, … kombinasyonlar örnek setine uygulanır. İşlem tüm örnekler sınıflandırılıncaya kadar devam eder. Bu algoritmanın kural üretme prosedürü Şekil 1'de gösterilmiştir.

Adım-1. Verilen örnek setindeki her değerin ve her karakteristiğin entropisi hesaplanır.

Adım-2. Entropi değerler küçükten büyüğe doğru sıralanır ve örnek seti bu sıralamaya göre yeniden düzenlenir.

Adım-3. Sınıflandırılmamış ilk örnekten başlayarak, örnekteki her karakteristik değerinden birer tane almak şartıyla *n*'li kombinasyonlar oluşturulur.

Adım-4. Her kombinasyon örnek setindeki tüm örneklere uygulanır. *n* adet kombinasyondan oluşan değerlerden tek bir sınıfa karşılık gelenleri kural haline getirilir. Sınıflandırılan örnekler işaretlenir.

Adım-5. Eğer tüm örnekler sınıflandırılmış ise Adım-8'e gidilir.

Adım-6. Kombinasyon sayısı 1 artırılır. (*n=n+1)*

Adım-7. Eğer *n<Karakteristik Sayısı* ise Adım-3'e gidilir.

Adım-8. Aynı örnekleri temsil eden birden fazla kural varsa en genel kurallar seçilir.

Adım-9. Son

Şekil 1. Rex-1 Algoritmasının kural üretme prosedürü
(Figure 1. Rule generation procedure of the Rex-1 Algorithm)

4. MEDİKAL VERİ MADENCİLİĞİ (MEDICAL DATA MINING)

Medikal alanda veri madenciliği uygulamaları çeşitli konularda yapılmış ve çalışmalar yoğun bir şekilde devam etmektedir. Özellikle sağlık sektöründe doğru ve zamanında karar almanın hasta sağlığı üzerindeki etkisi önemsenmeyecek derecede çok önemlidir. Bu işlemleri kolaylaştıran veri tabanı yönetim sistemleri (VTYS) en küçük sağlık kuruluşundan en büyük sağlık kuruluşlarına kadar kullanılmaktadır.

Hastane bünyesinde VTYS'de toplanan tüm operasyonel veriler, hasta verileri, yapılan tetkikler ve sonuçları, uygulanan tedavi yöntemi ve tedavi sürecine dair veriler yöneticiler açısından incelendiğinde; hastanedeki servislerin ve programların başarısının görüntülenmesi, kaynakların maliyetlerle göreceli olarak kullanımı, kaynak kullanımı ve hasta sayıları ile ilgili trendlerin tahmini, harcamalarla ilgili normal olmayan durumların anlık tespiti ve yolsuzlukların engellenmesi, hastanede uygulanan tedavi yöntemlerinin başarısının irdelenmesi açısından önemli

bilgileri içermektedir. Bu veriler başarılı tedavi sonuçları almada etken faktörlerin belirlenmesi, ameliyatlarda yüksek risk faktörlerinin sınanması, hasta verilerinin yaş, cinsiyet, ırk ve tedavi yöntemi gibi faktörlere göre sınıflanması, hasta sağlığı açısından geriye dönük faktörlerin sınanması, tedavi yöntemi geliştirme vb. amaçlarla kullanılmaktadır (Işık, 2008).

Dünya çapında medikal alanda çok sayıda başarılı veri madenciliği uygulama örnekleri bulunmaktadır. Örneğin, San Francisco Hearth Institute; hasta sonuçlarının iyileştirilmesi, hastanın hastanede kalma süresinin azaltılması, vb amaçlarla bir çalışma başlatmış ve kurum bünyesinde toplanan verilerden hastanın geçmişine ait veriler, laboratuar verileri, kollestrol verileri, diğer medikal verileri bilgiye dönüştürmüştür. A. Kusiak ve arkadaşları tarafından akciğerdeki tümörün iyi huylu olup olmadığına dair, karar destek amaçlı bir çalışma yapılmıştır. İstatistiklere göre Amerika da 160.000 den fazla akciğer kanseri vakasının olduğu ve bunların %90'ının öldüğü belirlenmiştir. Bu bağlamda bu tümörün erken ve doğru olarak teşhisi önem kazanmaktadır. Noninvaziv testler ile elde edilen bilgi sayesinde %40-60 oranında doğru teşhis konabilmektedir. İnsanlar kanser olup olmadıklarından emin olmak için biyopsi yaptırmayı tercih etmektedirler. Ancak bu testlerin hem maliyeti yüksek hem de çeşitli riskler taşımaktadır. Faklı yerlerde ve farklı zamanlarda kliniklerde toplanan test verileri arasında yapılan veri madenciliği çalışmaları teşhiste %100 oranında doğruluk sağlamıştır (Kusiak, ve diğ. 2000, Işık, 2008, Kaya ve diğ., 2003).

Başka bir çalışma ise Kore Tıbbi Sigorta Kurumu (The Korea Medical Insurance Corporatition) tarafından hazırlanan bir veri tabanı üzerinde yapılan yüksek tansiyon ile ilgili bir çalışmadır. Bu çalışmada karar ağacı öğrenim algoritmalarından CHAD, C4.5, C5.0 kullanılmıştır. Bu çalışmalar sonucunda yüksek tansiyon tahmininde etkili değerler BMI, idrar proteini (urinary protein), kan glikozu, kolestrol değerleridir. Yaşam koşullarının (diyet, alınan tuz miktarı, alkol, tütün gibi) hiçbirinin tahminde etkili olmadığı ayrıca grafiksel değerlerde de yalnızca yaşın etkili olduğu saptanmıştır (Kaya ve diğ., 2003).

5. DENEYSEL ÇALIŞMA (EXPERIMENTAL STUDY)

Bu çalışmada gerçek dünyadan alınan ve veri madenciliği algoritmalarının başarısını test etmek amacıyla kullanılan beş adet medikal veri seti seçilmiştir: Wisconsin Breast Cancer, Ljubljang Breast Cancer, Dermatology, Hepatitis ve Diabetes. Bu veri setleri *10-cross fold* metoduna göre örnek seti ve test seti olarak iki ayrı grubu ayrılmıştır. Bu metot ile örnek seti on parçaya ayrılmakta ve bu parçaların bir tanesi (örnek setinin %10'u kadar örnek) test seti, geriye kalanları ise (örnek setinin %90'ı kadar örnek) eğitim seti olarak ayrılmaktadır. Aynı problem on defa çalıştırılır. İlk çalıştırmada birinci %10'luk kısım, ikinci çalıştırmada ikinci %10'luk kısım ve en son onuncu çalıştırmada son kısım olan onuncu %10'luk kısım test seti olarak alınır. Diğer geriye kalan örnekler ise eğitim seti olarak kullanılır. Elde edilen değerlerin doğruluk oranları 1 nolu denklem yardımıyla hesaplanır. Doğruluk oranı algoritmanın daha önce görmediği bir veriyi doğru bir şekilde sınıflandırma oranı olarak tarif edilir.

$$\text{Accuracy} = \frac{\text{No. of test examples covered by the rule set}}{\text{Total no. of test examples}} \text{x100.} \qquad (1)$$

Eğitim setlerinden REX-1 algoritması yardımıyla kural tabanı oluşturulmuş ve bu kural setleri test setlerine uygulanarak doğruluk

oranları hesaplanmıştır. Algoritmaların etkinliğini ve doğruluğunu test etmek amacıyla bu alanda yaygın olarak kullanılan beş adet medikal veri seti (UCI Machine Learning Rapository (Blake ve diğ., 1998)) seçilmiş ve elde edilen sonuçlar C4.5, NavieBayes, PART, CORE, Ant-Miner, CN2, GA-SVM gibi algoritmalarla test edilmiştir.

REX-1 algoritması için Delphi 2009 programlama dili kullanılarak bir yazılım geliştirilmiştir. Veri setleri text ortamdan DBase veritabanına aktarılmış, eksik değer içeren örnekler için özel bir kod hazırlanmış ve eksik değerlerin dikkate alınmaması sağlanmıştır. Sonuçları karşılaştırmak için program Pentium IV 3.0 GHz ve 1 GB ana belleğe sahip bir PC bilgisayarda çalıştırılmış ve sonuçlar elde edilmiştir.

Çalışmada kullanılan örnek setlerinin karakteristik özellikleri, örnek sayısı, karakteristik sayısı, sınıf sayısı ve eksik değer içerip içermediği gibi özellikler Tablo 1'de verilmiştir.

Tablo 1. Örnek setlerinin özellikleri
(Table 1. Properties of data sets)

Örnek seti	Özellik tipi	Örnek sayısı	Özellik sayısı	Sınıf sayısı	Eksik değer?
BC Wisconsin (WBCD)	Sayısal	699	10	2	Evet
BC Ljubljana (LBCD)	Sayısal,Kategorik	286	9	2	Evet
Dermatology	Kategorik	366	34	6	Evet
Hepatits	Sayısal,Kategorik	155	19	2	Evet
Diabetes	Kategorik	768	8	2	Hayır

Bu örnek setlerini karşılaştırmak için örnek setinden elde edilen ortalama doğruluk oranı, en iyi doğruluk oranı ve standart sapma değerleri kullanılmıştır.

Ljubljana Breast Cancer örnek seti için REX-1 ve diğer sınıflandırıcılar tarafından elde edilen sonuçlar Tablo 2'de gösterilmiştir. Sonuçlar, ortalama doğruluk oranlarına göre incelendiğinde Rex-1 algoritması %83.34 gibi büyük bir fark ile en yüksek doğruluk oranına ulaşmıştır.

Tablo 2. Ljubljana BC veri setini kullanarak algoritmaları karşılaştırma
(Table 2. Comparison algorithms using Ljubljana BC data set)

Algoritma	Ortalama doğruluk (%)	En yüksek doğruluk (%)	Standart sapma
GA-SVM Hybrid [a]	76.20	76.57	0.27
C4.5 [a]	71.81	78.35	3.55
PART [a]	69.32	80.41	4.33
NavieBayes [a]	72.34	94.34	3.29
CORE [a]	75.41	84.69	3.24
Ant-Miner [a,b]	75.28	--	2.24
REX-1	83.34	90.63	3.84

[a] Tan ve diğ., (2009)
[b] Parpinelli ve diğ, (2002)

Hepatitis örnek seti için Rex-1 ve diğer sınıflandırıcılar tarafından elde edilen sonuçlar Tablo 3'te gösterilmiştir. Sonuçlar, ortalama doğruluk oranlarına göre incelendiğinde tüm sonuçların birbirine yakın olduğu görülmekte ve bununla birlikte Rex-1 algoritması %83.22 ile en yüksek dördüncü sonucu üretmiştir.

Tablo 3. Hepatitis veri setini kullanarak algoritmaları karşılaştırma
(Table 3. Comparison algorithms using Hepatitis data set)

Algoritma	Ortalama doğruluk (%)	En yüksek doğruluk (%)	Standart sapma
GA-SVM Hybrid [a]	86.12	89.67	1.73
C4.5 [a,d]	78.94	90.57	4.84
PART [a, d]	80.02	94.34	4.98
NavieBayes [a,d]	83.62	94.34	4.90
CORE [a]	84.40	92.45	3.72
C4.5/GA-small [e]	79.36	--	23.4
C4.5/GA-large-SN [e]	82.52	--	7.00
REX-1	83.22	92.57	4.58

[a] Tan ve diğ., (2009)
[d] Tan ve diğ.,(2003)
[e] Carvalho, Freitas, (2004)

Diabetes örnek seti için REX-1 ve diğer sınıflandırıcılar tarafından elde edilen sonuçlar Tablo 4'te gösterilmiştir. Sonuçlar, ortalama doğruluk oranlarına göre incelendiğinde tüm sonuçların birbirine yakın olduğu görülmekte ve REX-1 algoritması GA-SVM Hybrid algoritmasından sonra %77.11 ile en yüksek ikinci sonucu üretmiştir.

Tablo 4. Diabetes veri setini kullanarak algoritmaları karşılaştırma
(Table 4. Comparison algorithms using Diabetes data set)

Algoritma	Ortalama doğruluk (%)	En yüksek doğruluk (%)	Standart sapma
GA-SVM Hybrid [a]	78.26	78.64	0.23
C4.5 [a]	73.13	77.39	2.55
PART [a]	72.78	80.08	2.53
NavieBayes [a]	75.09	81.61	2.45
CORE [a]	75.34	80.15	2.30
REX-1	77.11	81.58	2.58

[a] Tan ve diğ., (2009)

Wisconsin Breast Cancer örnek seti için REX-1 ve diğer sınıflandırıcılar tarafından elde edilen sonuçlar Tablo 5'te gösterilmiştir. Ortalama doğruluk oranlarına göre sonuçlar incelendiğinde CART algoritması hariç tüm sonuçların birbirine yakın olduğu ve bununla birlikte Rex-1 algoritması %83.22 ile en yüksek dördüncü sonucu ürettiği görülmektedir.

Tablo 5. Wisconsin BC veri setini kullanarak algoritmaları karşılaştırma
(Table 5. Comparison algorithms using Wisconsin data set)

Algoritma	Ortalama doğruluk (%)	En yüksek doğruluk (%)	Standart sapma
Navie Bayes [c,f]	97.20	100	1.71
PART [c,f]	94.70	100	2.51
C4.5 [c]	95.01	100	2.73
Ant-Miner [b,h]	96.04	--	0.93
CN2 [b,h]	94.88	--	0.88
CART [g]	77.10	--	--
REX-1	95.02	100	2.52

[b] Parpinelli ve diğ, (2002)
[c] Baykasoğlu, Özbakır, (2007)
[f] Bojarczuk, (2004)
[g] Kahramanli ve Allahverdi, (2009), [h] Su ve diğ. (2010)

Dermatology örnek seti için REX-1 ve diğer sınıflandırıcılar tarafından elde edilen sonuçlar Tablo 6'da gösterilmiştir. Ortalama doğruluk oranlarına göre sonuçlar incelendiğinde REX-1 algoritmasının %94.88 ile en yüksek sonucu ürettiği görülmektedir.

Tablo 6. Dermatology veri setini kullanarak algoritmaları karşılaştırma (Table 6. Comparison algorithms using Dermatology data set)

Algoritma	Ortalama doğruluk (%)	En yüksek doğruluk (%)	Standart sapma
BGP [g]	86.20	--	6.24
C4.5 [g]	89.10	--	0.13
Ant-Miner [b]	94.29	--	1.20
CN2 [b]	90.38	--	1.66
DE/QDE [h]	91.53	--	2.40
REX-1	94.88	100	2.81

[a] Tan ve diğ., (2009)
[b] Parpinelli ve diğ, (2002)
[g] Kahramanli ve Allahverdi, (2009)
[h] Su ve diğ. (2010)

5. SONUÇLAR (CONCLUSION)

Medikal veri madenciliği alanında çalışmalar yoğun bir şekilde devam etmektedir. Bu çalışmaların sonucunda VTYS'de bulunan veriler kullanılarak hasta ve hastalıklarla ilgili çok önemli gizli bilgi veya bilgileri keşfetmek mümkündür. Bunun için uygun veri madenciliği tekniklerinin kullanılması yeterlidir. Bu sayede geçmiş bilgileri kullanarak hastalıklarla ilgili tahminlerde bulunmak, hastalığa sebep olan faktörleri tespit etmek mümkündür. Bu çalışmada da görüldüğü gibi REX-1 algoritması medikal veri setleri üzerinde eğitim aşamasından sonra kendisine verilen test setlerinden çok yüksek oranlarda doğru sınıflandırma yapmaktadır.

KAYNAKLAR (REFERENCES)

1. Akgöbek, Ö., (2006), "Veri Madenciliğinde Otomatik Kural Üretebilen Bir Uzman Sınıflandırma Sisteminin Geliştirilmesi", Yöneylem Araştırması Dergisi, Cilt 17, Sayfa:38- 50.
2. Akgöbek, Ö., Aydın, Y.S., Öztemel, E., and Aksoy, M.S., (2006), "A New Algorithm For Automatic Knowledge Acquisition in Inductive Learning", Knowledge-Based Systems, 19, 388-395.
3. Baykasoğlu, A. and Özbakır, L., (2007), MEPAR-miner: Multi-expression programming for classification rule mining, European Journal of Operational Research 183, 767–784.
4. Blake, C.L. and MERZ, C.J., (1998), "UCI Repository of Machine Learning Databases", http://archive.ics.uci.edu/ml/datasets.html, CA: University of California, Department of Information and Computer Science, (Son erişim tarihi : 15.07.2010)
5. Bojarczuk, C.C., Lopes, H.S., Freitas, A.A., and Michalkiewicz, E.L., (2004), A constrained-syntax genetic programming system for discovering classification rules: application to medical data sets, Artificial Intelligence in Medicine 30, 27-48.
6. Carvalho, D.R. and Freitas, A.A., (2004), "A hybrid decision tree/genetic algorithm for data mining", Information Sciences 163, 13-35.
7. Işık, A., (2008). Veri Madenciliği, Sızıntı Dergisi, yıl:30, Sayı, 352.
8. Kahramanli, H. and Allahverdi, N., (2009), Rule extraction from trained adaptive neural networks using artificial immune systems, Expert Systems with Applications 36, 1513–1522.

9. Kaya, E., Bulun, M. ve Arslan, A., (2003), Tıpta Veri Ambarları Oluşturma ve Veri Madenciliği Uygulamaları, Akademik Bilişim 2003, Çukurova Üniversitesi, Adana
10. Kusiak, A., Kernstine, K.H., Kern, J.A., McLaughlin, K.A., and Tseng, T.L., (2000), Medical and Engineering Case Studies.
11. Parpinelli, R.S., Lopes, H.S., and Freitas, A.A., (2002), Data mining an ant colony optimization algorithm, IEEE transactions on Evolutionary Computation 6 (4), 321-332.
12. SAS Institute Inc., (1999), The Data Mining Challenge: Turning Raw Data Into Business Gold. www.sas.com/software/data_mining/.
13. Su, H., Yang Y., and Zhao, L., (2010), Classification rule discovery with DE/QDE algorithm, Expert Systems with Applications 37, 1216-1222.
14. Tan, K.C., Yu, Q., Heng, C.M., and Lee, T.H., (2003), Evolutionary computing for knowledge discovery in medical diagnosis, Artificial Intelligence in Medicine 27, 129-154.
15. Tan, K.C., Teoh, E.J., Yu, Q., and Goh, K.C., (2009), A hybrid evolutionary algorithm for attribute selection in data mining, Expert Systems with Applications 36, 8616-8630.
16. Vahaplar, A. ve İnceoğlu, M.M., (2001), "Veri Madenciliği ve Elektronik Ticaret", inet-tr'2001 Türkiye'de internet konferansları VII, İstanbul.