Zaman Serileri Madenciliği Kullanılarak Nüfus Artışı Tahmin Uygulaması

Zeynep Behrin Güven, Turgay Tugay Bilgin
Maltepe Üniversitesi, Yazılım Mühendisliği Bölümü, İstanbul
zeynepguven@maltepe.edu.tr, turgaybilgin@maltepe.edu.tr

Özet: Zaman serileri madenciliği, önceden bilinen bir zaman dizisi temel alınarak dizinin henüz değeri bilinmeyen elemanlarını en olası bir şekilde tahmin etmeye çalışmaktır. Bu yöntem, veri madenciliğinde sıklıkla kullanılan istatiksel analiz tekniklerinden birisidir. Bu çalışmada WEKA programı ile Türkiye İstatistik Kurumundan (TUIK) alınan veri seti üzerinde zaman serileri madenciliği algoritmaları uygulanmıştır. 2001-2010 yılları arasındaki nüfus verileri kullanılarak daha sonraki yıllar için nüfus tahmini yapılmıştır.

Anahtar Sözcükler: Veri madenciliği, zaman serileri, tahmin.

Implementation of the Population Growth Forecast Using Time Series Mining

Abstract: Time Series Mining has trying to predict the present unknown value of the array elements optimally based on the previously known time series. This method is one of the statistical analyze techniques commonly used in data mining techniques. In this study, Time Series Mining has been implemented by WEKA (Waikato Environment for Knowledge Analysis) program on population data set taken by Turkish Statistical Institute (TSI). Population forecasting for the subsequent years has been predicted using population data set between the years of 2001 to 2010.

Keywords: Date mining, time series, prediction

1. Giriş

Veri Madenciliği, istatistiksel analiz tekniklerinin, genetik algoritma yöntem-lerinin ve yapay zekâ algoritmalarının bir arada kullanılarak veri içerisindeki gizli bilgilerin açığa çıkarılması ve verinin kullanılabilir bilgiye dönüştürülmesi sürecidir [1]. Veri madenciliği yöntemlerinin hedefi, çok sayıda verinin analiz edilmesiyle bir bilgi elde etmek ve bu bilgiyi kullanarak gelecek hakkında tahminde bulunmaktır. Son yıllarda veri madenciliği endüstri, sağlık, finans gibi birçok alanda kullanılmaya başlanmıştır.

Bu çalışmada veri madenciliği yöntemlerinden biri olan zaman serileri ile WEKA uygulama aracında elimizde olan nüfus verileri analiz edilerek gelecek yıllar için tahminde bulunulmuştur.

2. Zaman Serileri

Zamanın periyodik noktalarında bir cevap değişkeni gözlemlenmesi yoluyla verilerin toplanmasına "zaman serisi " adı verilir [2]. Zaman serileri bir değişkenin değişen zamanlarda gözlenen değerlerini bildirirler.

Zaman serileri, değişkenlerin gün, hafta, ay, mevsim veya yıl gibi herhangi bir zaman dilimine göre dağılımını gösteren serilerdir. Çeşitli değişkenler için düzenlenmiş zaman serileri için özel tahmin teknikleri geliştirilmiştir. Zaman serisi analizlerinin kullanıldığı en önemli alanlar ekonomi ve işletme sektörleridir [3].

3. Uygulama

Uygulama için Türkiye İstatistik Kurumundan (TUIK) alınan 2001-2010 yılları arasındaki bölgesel nüfus verileri incelenmiş, 2011 ve daha sonraki yıllar için tahminde bulunulmuş ve tahmin sonucu ile 2011 yılı gerçek nüfus sayısı karşılaştırılmıştır.

3.1 Yazılım

Uygulamada WEKA 3.7.9 programı kullanılmıştır. WEKA java platformu üzerinde geliştirilmiş, açık kaynak kodlu bir veri madenciliği uygulama programıdır [4]. WEKA, içinde birçok sınıflandırma, regresyon, yapay sinir ağları (YSA) algoritmalarını barındıran bir veri madenciliği aracıdır. WEKA Explorer, Experimenter, Knowledge Flow, Simple CLI (Command Line) şeklinde 4 temel uygulamaya sahiptir. Desteklediği temel veri kaynak biçimleri metin tabanlı arff, csv, c45, svmlight, libsvm, Xarff formatlarıdır.

3.2 SMOreg Algoritması

Uygulamada Uygulamada WEKA içerisinde bulunan SMOreg algoritması kullanılmıştır. Bu algoritma, Destek Vektör Makineleri (DVM) üzerinde uygulayarak regresyonu gerçekleştirir [5]. SMOreg algoritması temel olarak destek vektör makineleri olarak bilinen yöntemleri kullanır.

DVM'ler 1960'lı yılların sonunda Vladimir Vapnik ve Alexey Chervonenkis tarafından geliştirilmiş temel olarak istatiksel öğrenme teorisine dayanan bir makine öğrenmesi yöntemidir. DVM'nin temel mantığı doğrusal olarak ayrıştırılabilen veri yapıları için en iyi ayırıcı düzlemin belirlenmesidir. Doğrusal olarak ayrıştırılamayan veri yapıları dönüşüm tekniği ile farklı bir boyuta taşınarak çözülür.

DVM'ler, eğitim esnasında gözlenmemiş yeni verileri de sorunsuz olarak sınıflandırabilmektedir. Bu durum destek vektör makinelerinin genelleştirebilme yeteneğini göstermektedir. Genelleştirebilme özelliği DVM'yi diğer tekniklere göre (YSA, karar ağacı vs.) iyi bir alternatif yapmaktadır. Son zamanlarda ise DVM, örüntü tanıma, yüz bulma ve tanıma, veri madenciliği gibi alanlarda kullanılmaya başlanmıştır ve birçok gerçek problemde başarılı sonuçlar vermiştir [6].

DVM'nin temelini, verilerin bir düzlem veya hiper düzlem ile ayrılarak sınıflandırılması işlemi oluşturmaktadır. DVM bu işlemi, iki sınıf arasındaki marjini maksimum yaparak gerçekleştirmektedir. Bu şekilde veri eğitildikten sonra, DVM yeni gelen veriyi doğru sınıflamayı amaçlamaktadır [7].

Doğrusal olarak ayrılabilen verileri, ait oldukları boyutta bir düzlem ile ayırabilmek mümkünken, doğrusal olarak ayrılamayan verilerin ait oldukları boyuttan daha yüksek boyutlu bir uzaya taşınarak, burada bir hiper düzlem ile ayırmak mümkün olacaktır [7]. Destek vektör makineleri, doğrusal olarak ayrılabilen veriler söz konusu olduğunda, verileri ayırabilecek sonsuz sayıdaki doğru içerisinden marjini en yüksek yapacak olan doğruyu seçmeyi hedeflemektedir[8]. Doğrusal olarak ayrılamayan verilerin olduğu durumda ise DVM, bir haritalama yöntemi ile orijinal veriyi daha yüksek boyutlu bir uzaya taşır ve burada verileri sınıflandırmak için optimum olabilecek doğrusal ayırıcı hiper düzlemi bulmaya çalışır DVM literatüründe, tahmin edici (predictor) veya bağımsız değişkene doğal nitelik (attribute), optimum hiper düzlemi belirlemek için kullanılan dönüştürülmüş doğal niteliğe belirleyici nitelik (feature) ve bir deneği (gözlemi) tanımlayan belirleyici nitelik setine ise, vektör denilmektedir [7].

4. Veri Seti

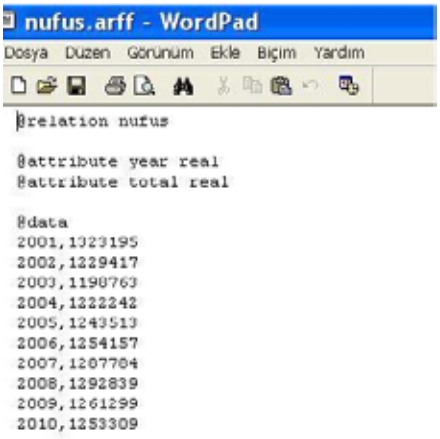
Türkiye İstatistik Kurumu tarafından her yıl güncellenen demografik veri seti birçok alanda çalışmalarda kullanılmaya açık bir kaynaktır. Veriler TUIK' in web sitesinden pdf, excel vb formatlarda indirilebileceği gibi talep formu doldurularak, mail aracılığıyla veya kütüphane

birimlerine gidilip birebir de edinilebilmektir. TUIK' den edinilen veri seti bölgeler içinde şehir bazında ve annenin yaş grubuna göre nüfus sayısının tutulduğu bir excel dosyasıdır (Şekil 1).

Yıl / Year	İBBS-SR / Nuts	Toplam / Total	Annenin yaş grubu - Age group of mother								Bilinmeyen / Unknown
			<15	15-19	20-24	25-29	30-34	35-39	40-44	45-49	50+	
(Doğum yerine göre - By the place of birth)												
2001												
	TR Türkiye - Turkey	1 323 195	2 726	154 304	436 517	373 679	197 330	113 237	29 668	8 943	2 093	4 698
	TR1-TR10-TR100 İstanbul	188 535	294	18 952	63 275	58 124	29 708	13 842	2 478	404	70	1 388
	TR2 Batı Marmara	36 984	63	4 508	13 106	11 216	5 462	2 058	358	51	11	151
	West Marmara											
	TR21 (Tekirdağ, Edirne, Kırklareli)	16 797	37	2 126	5 789	5 020	2 546	939	150	21	5	74
	TR211 Tekirdağ	8 272	19	969	3 002	2 477	1 219	443	73	11	2	37
	TR212 Edirne	4 735	13	711	1 559	1 378	726	277	35	4	2	30
	TR213 Kırklareli	3 790	5	426	1 228	1 165	601	219	42	6	1	7
	TR22 (Balıkesir, Çanakkale)	20 272	26	2 382	7 317	6 196	2 916	1 119	208	30	6	77
	TR221 Balıkesir	15 241	18	1 812	5 560	4 687	2 177	831	165	25	5	61
	TR222 Çanakkale	5 036	8	570	1 757	1 509	739	288	43	5	1	16
	TR3 Ege - Aegean	133 177	215	16 540	47 872	39 064	19 135	8 185	1 569	261	51	345
	TR31-TR310 İzmir	48 452	82	5 179	16 169	14 937	7 779	3 382	633	101	18	178
	TR32 (Aydın, Denizli, Muğla)	38 272	75	4 558	13 765	11 264	5 630	2 369	420	71	10	92
	TR321 Aydın	13 468	13	1 690	4 726	3 904	1 973	813	179	40	7	31
	TR322 Denizli	14 729	12	1 790	5 528	4 108	2 169	877	119	14	2	30
	TR323 Muğla	10 075	10	1 084	3 501	3 152	1 508	639	122	17	1	41

Şekil 1- TÜİK Veri Seti Örneği

WEKA programında işlenilecek verinin arff formatına çevrilmesi gereklidir. Veri formatı Şekil 2'deki gibidir. Veri setine verilecek olan isim @relation ile tanımlanır. Değişken isimleri ise değişkene verilecek ad ve türleri ile birlikte @attribute ile tanımlanır. Veriler ise @ data etiketinden sonra, kolonlar virgülle ayrılarak yazılır.



Şekil 2 -. arff Formatlı Veri Seti

4.1 WEKA'ya Bilgi Girişi

WEKA programında Explorer uygulaması seçilir. Daha sonra, Şekil 4'de Explorer ekranında Open File seçeneği ile arff biçimindeki veri seti WEKA'ya eklenir.

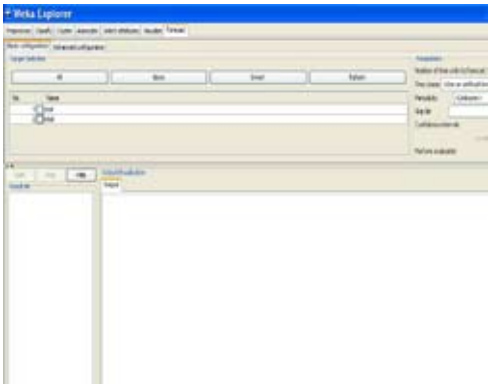


Şekil 3- WEKA Explorer Ekranı

Şekil 4'de görüldüğü gibi, tahmin algoritmasını ve grafiğini oluşturmamızı sağlayan Forecast sekmesi seçilir. Forecast sekmesinde dosyadaki attribute değerleri olan year ve total görülür. Sağ tarafta parametreler kısmındaki

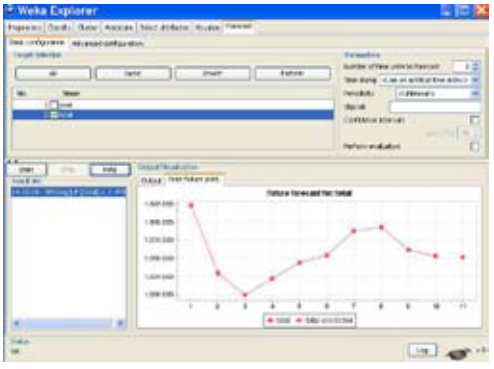
time stamp değişkeni kaç yıl sonrasını tahmin etme seçeneği sunar. Bütün bu işlemler sol taraftaki Basic Configuration bölümünün altında gerçekleştirilmektedir.

Basic Configuration 'ın hemen yanındaki Advanced Configuration kısmında ise tahmin etme işlemini gerçekleştiren algoritmalar bulunur. Tahmin işlemini gerçekleştirmeyi sağlayan SMOreg algoritmasıdır.



Şekil 4- Forecast Bileşeni

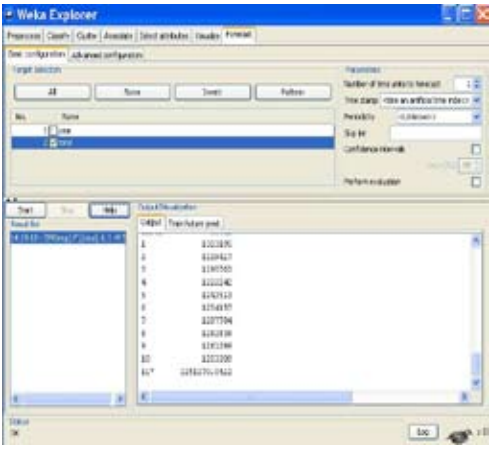
Şekil 5'de görüldüğü gibi, tahmin etmek istediğimiz total değişken seçilip, sağ taraftaki time stamp kısmında tahmin edilmesi istenen süre 1 yıl olarak belirlenerek başlatılır.



Şekil 5- Tahmin Grafiği

Şekil 5'de görülen Output kısmında Train future Prediction'da yapılan tahminin sonucu grafik olarak görülür. Grafikte görülen eğri üzerindeki kare işaretleri programa verilen sayısal verilerdir. Yuvarlak olan nokta işareti ise yapılan tahmin sonucunu verir.

Şekil 6' da Output kısmında ise 1-10 arası görünen satırlarda programa verilen 2001-2010 yılları arası gerçek sayısal nüfus verileri, 11. satıda ise 2011 yılı için tahmin edilen nüfus değeri görülmektedir.



Şekil 6- Tahmin Değerleri

Yıl	Gerçek Değer	Tahmin Değeri	Fark	Doğruluk Oranı (%)
2001	1323195	-	-	-
2002	1229417	-	-	-
2003	1198763	1061642	137121	88,56
2004	1222242	1106618	115624	90,54
2005	1243513	1202928	40585	96,74
2006	1254157	1261490	-7333	99,42
2007	1287784	1264691	23093	98,21
2008	1292839	1316454	-23615	98,17
2009	1261299	1309328	-48029	96,19
2010	1253309	1256534	-3225	99,74
2011	1237172	1251270	-14098	98,86

Tablo 1 Tahmin Edilen Değerler ve Doğruluk Oranları

5. Sonuçlar

Bu çalışmada açık kaynak kodlu veri madenciliği programı olan WEKA üzerinde zaman serileri madenciliği uygulaması gerçekleştirilmiştir. Uygulamada TUIK' den alınan 2001-2010 yılları arası sayısal nüfus verileri, uygun formata dönüştürülmüş ve programa verilmiştir. Tablo 1'de gerçek nüfus değerleri ile tahmin edilen nüfus değerleri karşılaştırılmıştır. 2011 yılı tahmininin %98,86 oranında doğru olduğu görülmektedir.

6. Kaynaklar

[1] Tekerek A., "**Veri Madenciliği Süreçleri Ve Açık Kaynak Kodlu Veri Madenciliği Araçları**"

[2] Özek Bulut M., Daş B, Akpolat H. "**Zaman Serisi Tahmininde Tip-2 Bulanık Mantık Tabanlı Veri Madenciliği Uygulaması**" (Mart 2010)

[3] Irmak S ., Köksal Deniz C., Asilkan Ö., "**Hastanelerin Gelecekteki Hasta Yoğunluklarının Veri Madenciliği Yöntemleri İle Tahmin Edilmesi**" (2012)

[4] Dener M., Dörterler M. , Orman A. , "**Açık Kaynak Kodlu Veri Madenciliği Programları: WEKA'da Örnek Uygulama**"

[5] http://weka.sourceforge.net/doc.stable/weka/classifiers/functions/SMOreg.html

[6] Karagülle F., "**Destek Vektör Makinelerini Kullanarak Yüz Bulma**" (2008)

[7] Akşehirli Ö.,Ankaralı H.,Aydın D.,Saraçlı Ö. "**Tıbbi Tahminde Alternatif Bir Yaklaşım: Destek Vektör Makineler"**

[8] Vapnik, V.N., 2000, "**The Nature of Statistical Learning Theory**", Second Edition (New York: Springer-Verlag)