BOSTON ŞEHRİNİN HAVA DURUMUNU VERİ MADENCİLİĞİ YÖNTEMLERİYLE ANALİZİ

WEATHER FORECAST WITH DATA MINING ALGORITHMS

Fırat Üniversitesi Teknoloji Fakültesi Yazılım Mühendisliği, Lisans, Nurettin Şanslı(*)

ÖZET

Bu çalışmada, geçmişteki hava verileri kullanılarak hava durumu tahmin etmek amaçlanmaktadır. *Kaggle* sitesindeki bir üyenin paylaştığı veriler kullanılmaktadır. Araştırmada sadece 2008 – 2018 yılları arasında Boston şehrindeki hava bilgileri ele alınmaktadır. Veri kümesinde 24 tane niteliğe ait toplamda 3749 tane sayısal olan veriler bulunmaktadır. Veri kümesi weka'da açıldığında hata aldığı gözlemlenmiştir. Veri kümesindeki bazı verilerde düzenlemeler yapılarak hata giderilmektedir. Sınıflandırma sonucunda J48, Random Tree, Naif Bayes, BayesNet ve Kstar algoritmalarından başarımının daha yüksek olduğu belirlenmiştir. Elde edilen sonuçların yorumlanması yapılmıştır.

Anahtar Kelimeler: *Veri madenciliği, Sınıflandırma, Boston Şehri, Hava Durumu, Karar Ağacı, J48 algoritması.*

ABSTRACT

In this study, it is aimed to estimate the weather using past weather data. The data shared by a member of the Kaggle site is used. The survey only covers weather information in Boston city from 2008 to 2018. There are totally 3749 numeric data in 24 datasets. We observed that the dataset received an error when opened in WEKA. Some data in the data set is edited to correct the error. As a result of the classification, it is determined that the performance of J48, Random Tree, Naive Bayes, BayesNet and Kstar algorithms is higher. The results obtained were interpreted.

Keywords: *Data mining, Classification, Boston City, Weather condition, Decision Tree, J48 algorithm.*

Giriş

Kısaca hava, atmosferin belirli bir durumunda sınırları belirtilmiş toprak parçasının halidir. Hava kâinat'ın varoluşundan beri tüm canlılar için hayati öneme sahiptir.[1] İklim değişikliği topluma huzursuzluk yaratması nedeniyle savaşlar çıkmaktadır. Bu araştırmada iklim değişikliğine karşı risk oranı yüksek olan Amerika Birleşik Devletleri seçilmektedir.[2] Massachusetts eyaletinde yer alan Boston şehrinin hava durumu bilgileri araştırmaya dâhil edilmektedir.

Canlılar için doğuracağı sonuçları araştırma yaparak hava durumu tahmini yapılır. Meteorolojistler gözlem ve analizlere dayanarak hava tahmini yapmaktadırlar. Bu çalışma Meteorolojistlere hava tahminlerinde yardımcı olan diğer bilim dallarından fizikçiler, kimyacılar ve matematikçiler gibi fayda sağlanmaktadır. Bu çalışmada Veri madenciliği yöntemiyle Meteoroloji biliminden farklı hava durumu tahmini yapılmaktadır.[3]

Verilerin elde edilmesi

Veri kümesi hazırlanırken *Weather Underground* sitesinden gerçek veriler elde edilmiştir. Veri kümesi *Kaggle* sitesinden hazır olarak indirilmiştir.[4]

Veri madenciliği

Kısaca veri madenciliğinin tanımını yapmak gerekirse, büyük çaplı veriler arasından bilgiye varma, bilgiyi madenleme meselesidir. Ya da bir anlamda büyük veri kümesi içerisinden gelecekle ilgili kestirimde bulunabilmemizi sağlayabilecek ilişkilerin bilgisayar yazılımı kullanarak aranmasıdır.[5]
Bu çalışma için Veri madenciliği sınıflandırma yöntemlerinden biri olan C4.5 yöntemi kullanılmıştır

C4.5 Ağacı (C4.5 Tree)

J48 algortiması weka tarafından C4.5 algoritması için geliştirilmiştir. C4.5 ağacı sayısal değerlerinde karar ağacı yaratmasına fırsat vermektedir. C4.5 ağacı normalizasyon işlemi kullanmaktadır.
C4.5 ve ID3 ağacı entropi tabanlıdır. C4.5 ağacı, ID3 ağacının geliştirilmiş durumu olmaktadır. C4.5 ağacının farkı budama işlemi yapmaktadır.

$$\text{Değer}(B) = -\sum((\text{sıklık}(S_k,B) / |B|).\log_2(\text{sıklık}(S_k,B)/|B|)$$

Değer hesaplanırken yukarıdaki formüller kullanılmaktadır.

B: Sınıftaki benzerlerin sayısı ifade etmektedir. B ile S'in oranın değerine bakılır.

Bu etapta değer parçalara bölündükten sonra parçalar ile işlem yapılmaktadır.

$$\text{Değer}_x(L) = \sum_{j=1}^{n} ((|L_i| / |L|.\text{Değer}(L_i))$$

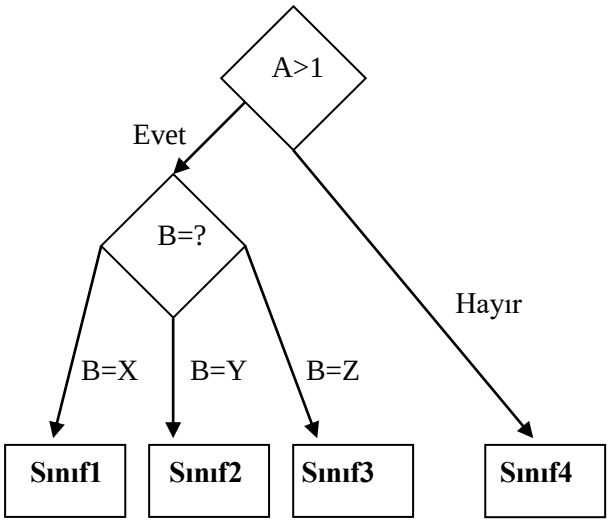
Yukarıda her j değeri için değer hesabı yapılmaktadır.

$$\text{Kazanım}(\text{Özellik X}) = \text{Değer}(L) - \text{Değer}_x(L)$$

Yukarıdaki formülde kazanım değeri, o özelliğin bağlı olduğu bütün parça ve o özelliği ilgilendiren parça arasındaki fark ile elde edilmektedir.[6][7]

Karar Ağacı

Karar ağacı ata düğümden bağlantılı bir şekilde çocuk düğümler oluşmaktadır. Karar ağacı düğümleri testteki tüm olası durumlara göre dallanır. Öğrenme sistemi yukarıdan aşağıya doğru bir yolu benimser. Yani şartlara göre hareket ederek verileri sınıflandırmaktadır. Karar ağacı, karar yapıları (if,else, else if, swtich) mantığına sahiptir.[8]
Aşağıda A ve B verilerinin sınıflandırılması için Şekil1'de örnek karar ağacı verilmiştir.



Şekil 1: Örnek karar ağacı şekli

Uygulama

WEKA, sadece Yeni Zelanda adalarında yaşayan ve uçamayan bir kuştur. Weka (Waikato Environment for knowledge analysis) yazılımı ise, 1999 yılında The University of Waikato tarafından Java ortamında geliştirilmiş açık kaynak kodlu yazılımdır. Weka yazılımında en çok kullanılan dosya uzantıları ARFF ve CSV olmaktadır. Weka, veri madenciliği görevleri için makine öğrenme algoritmaları topluluğudur. Weka; Ön işleme, Sınıflandırma, Kümeleme, Birliktelik kuralı ve Görselleştirme için araçlar içermektedir.[9][10]
Bu çalışmada Weka 3.8 sürümü kullanılmaktadır.

Sınıflandırma algoritmalarından bazılarının karşılaştırılması

Algoritmalar	Doğru sınıflandırılan örnek	Kappa İstatistiği	Ortalama Mutlak Hata	Ortalama Hata Karekök	Göreli Mutlak Hata %	Göreli Hata Karekök %	TP Oran	FP Oran	F-Ölçütü
J48	3421	0.8435	0.0564	0.1991	19.949	52.9557	0,913	0,086	0,912
RandomTree	3259	0.7685	0.0654	0.2556	23.099	67.9826	0,869	0,109	0,869
Naif Bayes	3185	0.7339	0.0777	0.2611	27.464	69.431	0,850	0,120	0,849
BayesNet	3122	0.715	0.0872	0.2647	30.830	70.8309	0,833	0,100	0,835
Kstar	2980	0.6282	0.1072	0.2953	37.899	78.5314	0,795	0,183	0,792

Tablo 1. Sınıflandırma algoritmalarının başarım oranları

Tahminlerin kalitesini ölçmek için sınıflandırma algoritmalarından 5 tanesi weka'da çalıştırılmıştır. Tablo1'e gösterilen algoritmalardan doğru sınıflandırılmış en yüksek örnek sayısıyla J48 algoritması olmuştur. Gerçek pozitif örnek sayısını veren TP (True Positive) ağırlıklı ortalama oranı 0,913, yanlış pozitif örnek sayısını veren FP (False Positive) ağırlıklı ortalama oranı 0,086 olarak yazılmaktadır. F-Ölçütünün ağırlıklı ortalama oranı ise 0,912 olarak bulunmuştur. F-ölçütü, anma ve kesinlik değerlerinin hesaplanabilmesi için aşağıdaki formüllerden yararlanılabilir.[8]

$$\text{F-Ölçütü} = \frac{2 \times \text{Anma} \times \text{Kesinlik}}{\text{Anma} + \text{Kesinlik}} \qquad \text{Anma} = \frac{TP}{TP+FN} \qquad \text{Kesinlik} = \frac{TP}{TP+FP}$$

Verilere J48 Algoritmasının Uygulanması

Örnek olarak kar yağışı sıfır küçük eşitse, Yağış sıfır küçük eşitse, Düşük görünürlük yedi küçük eşitse, Yüksek çiğ noktası otuz bir küçük eşitse, Ortalama sıcaklık otuz yedi büyükse, Yılın 8'inci ve 12'nci ayı arasındaysa hava açıktır.

Sonuç

Boston şehrine ait anlamsız veri olan hava durumu verilerini, Veri madenleciğiyle anlamlı hale getirerek sonuca vardık. Karar ağaçlarından biri olan J48 algoritması çalıştırıldıktan sonra sonucu okuma kolay olmaktadır. Çünkü karar ağacı algoritmalarının görsel özelliği olmaktadır. J48 algoritmasını seçme nedenlerimden bir diğeri niteliklerin tipi sayısal değer olmasıdır. Doğru sınıflandırılmış örnek sayısı 3421 yüzdelik olarak %91.251'dir. Başarılı sayılabilir bir yüzdelik dilimine sahibiz diyebiliriz. Veri sayısı arttıkça başarım oranı da artacaktır. Veriler şartlara göre yağmurlu, karlı ve açık hava şeklinde sınıflandırılmaktadır. Bu özelliğiyle başka şehirlere ait hava durumu verilerinin analizlerine imkân sağlamaktadır. Meteoroloji tahminlerinden daha uzun vadeli tahminler yaparak güncel hayata faydalı olabilmektedir. Bu çalışma uygulama alanı olarak Meteoroloji bilim dalında kullanılabilmektedir.

Kaynaklar

Prof. Dr. Murat Türkeş, s.3, 31 Mayıs 2013 tarihinde erişildi. [1]

https://www.climate-change-performance-index.org/, 3 Mayıs 2018 tarihinde erişildi. [2]

https://www.mgm.gov.tr/genel/meteorolojinedir.aspx?s=5i, 31 Mayıs 2018 tarihinde erişildi.[3]

https://www.kaggle.com/jqpeng/boston-weather-data-jan-2013-apr-2018, 12 Nisan 2018 tarihinde erişildi. [4]

http://www.wikizero.net/index.php?q=aHR0cHM6Ly90ci53aWtpcGVkaWEub3JnL3dpa2kvVmVyaV9tYWRlbmNpbGnEn2k, 3 Mayıs 2018 tarihinde erişildi. [5]

Doç. Dr. Murat Karabatak, Veri madenciliği dersi. [6]

Doç. Dr. Şadi Evren Şeker, http://bilgisayarkavramlari.sadievrenseker.com/2012/11/13/c4-5-agaci-c4-5-tree/, 3 Mayıs 2018 tarihinde erişildi.[7]

Mehmet Ali Alan, Karar Ağaçlarıyla Öğrenci Verilerinin Sınıflandırılması (2014). [8]

https://www.cs.waikato.ac.nz/ml/weka/index.html. 3 Mayıs 2018 tarihinde erişildi. [9]

www.wikizero.net/index.php?q=aHR0cHM6Ly9lbi53aWtpcGVkaWEub3JnL3dpa2kvV2VrYV8obWFjaGluZV9sZWFybmluZyk, 3 Mayıs 2018 tarihinde erişildi. [10]

Arş. Gör. Osman Altay, Makale hakkında bilgiler, http://www.heypasteit.com/clip/0ijcpi, 31 Mayıs 2018 tarihinde erişildi.