Veri Madenciliğinde Karar Ağacı Algoritması

Arş. Gör. Fatma TÜMİNÇİN[99]
Assoc. Prof. Dr. Alper AYTEKİN[100]
Arş. Gör. Ahmet AYAZ[101]

Özet

Karar ağacı modeli veri madenciliğinde sınıflandırma ve tahminleme algoritması olarak yaygın şekilde kullanılmaktadır. Karar ağacı modelinde bir problemin çözümü için ağaç yapısı oluşturularak tümevarım yöntemi kullanılmaktadır. Kolay yorumlanabilir ve anlaşılması kolay olduğundan dolayı literatürde karar ağacı modeline sıklıkla rastlanmaktadır.

Bu çalışmada veri madenciliği kavramı, karar ağaçlarının ne olduğu, genel yapılarının ne olduğu, avantajları ve dezavantajları, uygulamada sıkça kullanıldıkları alanlar araştırılmıştır.

Anahtar Kelimeler: Karar Ağacı, Sınıflandırma, Yönetim Bilişim Sistemleri.

Decision Tree Algorithm In Data Mining

Abstract

The decision tree model is widely used as a classification and estimation algorithm in data mining. In the decision tree model, the induction method is used by creating a tree structure to solve a problem. Since it is easy to interpret and easy to understand, decision tree model is frequently encountered in the literature.

In this study, data mining concept, decision trees, general structure, advantages and disadvantages, frequently used areas are investigated.

Keywords: Decision Tree, Classification, Management Information Systems Page.

Giriş

Son günlerde meydana gelen ve ileriki dönemlerde de artarak devam edecek olan veri sayısı, bu verilerin nerde ne zaman nasıl kullanılacağı sorunları gündeme gelmektedir. Elde edilen verileri değerlendirmede pek çok yöntem bulunmaktadır. Ancak bu yöntemler verilerin

[99] Bartın Üniversitesi İktisadi ve İdari Bilimler Fakültesi, Yönetim Bilişim Sistemleri Bölümü
[100] Bartın Üniversitesi İktisadi ve İdari Bilimler Fakültesi, Yönetim Bilişim Sistemleri Bölümü
[101] Bartın Üniversitesi İktisadi ve İdari Bilimler Fakültesi, Yönetim Bilişim Sistemleri Bölümü

büyümesiyle birlikte yetersiz kalmakta ve etkili kararlar verilmesinde kullanıcıya yardımcı olamamaktadır (Ayık vd., 2010).

İstatistiki verilerin analizleri için pek çok yöntem geliştirilmiştir. Bunlar yapa zeka teknolojileri, genetik algoritmalar, yapay sinir ağları, çok kriterli karar verme teknikleri, mantıksal programlamalar, karar ağaçları gibi modeller örnek verilebilmektedir. Ancak bu yöntemlerden pek çoğunun çalışma prensipleri bilinmemektedir. Çok iyi tahminleme yapabilmelerine karşın bu özellikleri bu tekniklerin zayıf yönlerini oluşturmaktadır. Karar ağacı modeli çalışma prensipleri olarak bilinmektedir. Diğerlerinde bulunan zayıf noktalar bu modelde bulunmamaktadır (Zorman vd., 2001).

Karar ağacı modeli adından da anlaşılacağı üzere ağaç görünümünde olan ve kullanıcıya sınıflama, kümeleme ve tahminler yapmada yardımcı bir modeldir (Ma, 1998). Karar ağaçlarının oluşturulması basit, yorumlanmasının kolay olması ve veri tabanlarına kolaylıkla bütünleşerek tahminleme yapmasından dolayı oldukça sık kullanılan bir yöntemdir. Sınıflandırma modelleri arasında en çok tercih edilen model karar ağacıdır.

Bu çalışmada kısaca karar ağaçlarının ne olduğu özellikleri, yapısı, avantajları ve dezavantajları karar ağacı algoritmaları hakkında bilgiler verilmiştir.

1. Veri Madenciliğinde Karar Ağaçları

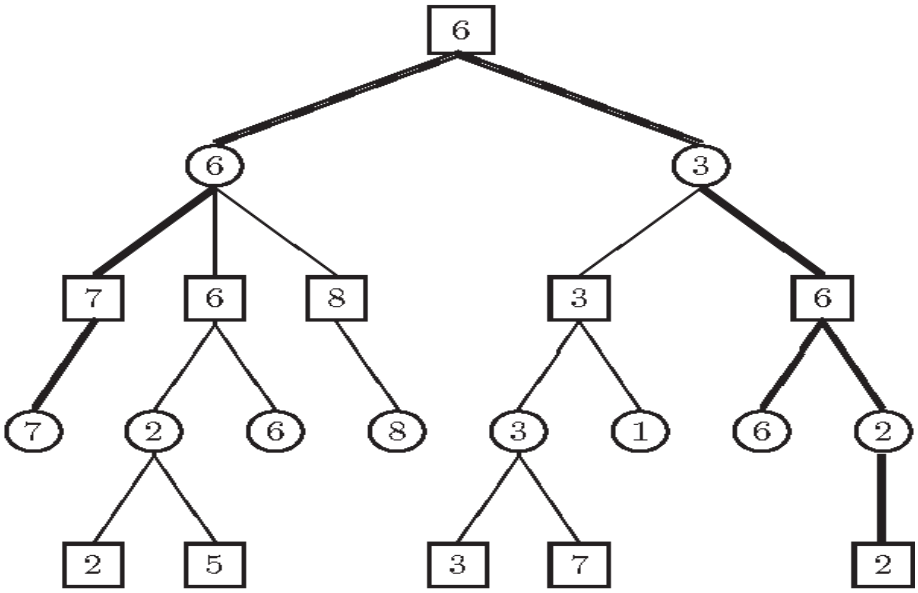
Veri madenciliği kavramı çok sayıda ve büyük yapıdaki veri ambarları ve veri tabanlarının içerisindeki verilerin anlamlandırılması ve ilişkiler kurulmasına yardımcı olan istatistiki algoritmaları ve yapay zeka teknolojilerini kullanan bir yöntemdir (Gargano ve Raggad, 1999, 81-82). Veri arama tekniği olarak da isimlendirilmektedir. Veri madenciliği sürecinde beş adımdan oluşmaktadır (Savaş vd., 2012).

1. Problemin tanımlanması
2. Verilerin hazırlanması
3. Modelin kurulması ve değerlendirilmesi
4. Modelin kullanılması
5. Modelin izlenmesi

Veri Madenciliği teknikleri, genellikle büyük ve çok sayıda verinin bulunduğu, bu verilerin eğitilmesi ve bu verilerden tahminleme işlemleri yapılmasında kullanılmaktadır (Zhang Dongsong ve Zhou Lina, 2004). Karar ağaçları önemli sınıflama araçlarından birini oluşturmaktadır. Yapının öğrenmesi kolaydır ve bilgiler anlaşılır şekilde gösterilebilme özelliğine sahip olması karar vericiler için birtakım avantaj sunmaktadır (Chien ve Chen, 2000).

- Karar ağaçlarının belki de en önemli özelliğinden biriside düşük maliyetli olmasıdır.
- Yine karar ağaçları görsel gösteriminden dolay anlaşılır, kolay yorumlanabilir ve veri tabanlarına kolay entegrasyon yapılmaktadır.
- Güvenilirlik bakımından oldukça iyi durumdadır ve bu yüzden yoğun olarak tercih edilmektedir.
- Kullanılan ağaç yapılar görselleştirilebilir.
- Veri hazırlığına çok az ihtiyaç duymaktadır.
- Hem sayısal hem de kategorik veri tipleri ile işlem yapabilmektedir.
- Çok çıktılı problemlere çözüm sunabilmektedir.
- İstatistiksel testler kullanılarak bir modelin doğrulanması mümkündür.

Karar ağacı karar vericiye birçok avantaj sağlamasına rağmen bir takım dezavantajları da bulunmaktadır. Karar ağacı yapısı veriyi açıklamaya çalışırken oldukça karmaşık bir ağaç yapısı ortaya çıkarabilir. Diğer bir dezavantajı ise budama işlemi yapılmadığında ezbere öğrenme yapabilmektedir.



Şekil 1. Karar Ağaçlarının Yapısı

Karar ağaçlarında yapı gereği tümevarım söz konusudur. Karar ağaçlarının yapısı düğümler, dallar ve yapraklar olmak üzere üç bölümden oluşmaktadır (Han and Kamber, 2000). Düğüm gerçekleştirilecek araştırmayı belirtirken ağacın her bir dalı sınıflama işlemini tanımlamaktadır. Karar ağacı modelinde her bir yaprak dallara dallar ise düğüme bağımlıdır. Karar ağaçlarında işlemler ardışık şekilde gerçekleşmektedir (Şekil 1).

2. Karar Ağacı Algoritmaları

Karar ağacının ilk yazılımı ve algoritmasını 1970'li yıllarının başlarında Morgan ve Sonquist tarafından ortaya atılmıştır. Yine 1984'te Berkeley Üniversitesi'nden Leo Breiman ve Charles J. Stone, Stanford Üniversitesi'nden Jerry Friedman ve R. Olshen "Classification And Regression Trees" adlı kitapta C&RT algoritmasından bahsetmişlerdir. Karar ağacının diğer bir algoritması olan ID3 1986 yılında J.R. Quinlan tarafından kazandırılmıştır. C4.5 algoritması 1993 yılında Quinlon "Programs For Machine Learning" adlı kitapta ortaya konulmuştur (Lee ve Keng, 2001). Daha sonra sırasıyla geliştirilen CHAID algoritması (G.V. Kass; 1980), Exhaustive CHAID algoritması (Biggs, de Ville ve Suen; 1991), MARS, QUEST, C5.0, SLIQ, SPRINT algoritmaları karar ağaçlarının yapısına eklenmiştir.

Karar ağacı tekniğini kullanarak verinin sınıflanması iki adımda gerçekleşen bir işlemdir (Han ve Kamber, 2000). İlk basamak öğrenme basamağı ikinci basamak ise sınıflandırma basamağıdır. Öğrenme adımında önceden bilinen bir eğitim verisi kullanılırken model oluşturmak işlemi için sınıflama algoritması tarafından analiz edilen sınıflama basamağında test verisi, sınıflama kurallarının veya karar ağacının doğruluğunu ortaya çıkarmak amacıyla kullanılır (Çalış Boyacı vd. 2018).

2.1. C & RT Algoritması

C&RT algoritmasında gini işlemine dayalı ikili bölme işlemi yapılmaktadır. Bu algoritmada iki uç olmayan her düğümde iki adet dal bulunmaktadır. Budama işlemi ağacın karmaşıklık yapısına göre yapılmaktadır. Yapı olarak sınıflandırma ve regresyonu tekniklerine uygundur. Sürekli hedef değişkenleri ile çalışır. Analizler için kullanılacak verinin hazırlanmasına ihtiyaç duyar (Emel ve Taşkın, 2005).

2.2. C4.5 ve C5.0 Algoritmaları

Her düğümden çıkan çoklu dallar ile ağaç oluşturur. Algoritmada dalların sayısı tahmin edicinin kategori sayısına göre belirlenir ve buna eşit sayıdadır. Sınıflama işleminde birden çok karar ağacını birleştirme özelliğine sahiptir. Ayırma işlemi için bilgi kazancını kullanır. Budama işlemi ise yapraklardaki hata oranına dayanır. C4.5 algoritması veri setinde kayıp veri bulunduğu taktirde bile düzgün çalışabilme özelliği taşımaktadır (Emel ve Taşkın, 2005).

2.3. CHAID Algoritması

Söz konusu algoritma ki-kare testleri kullanarak bölme işlemini gerçekleştirmektedir. Algoritmadaki dal sayısı tahmin edicide oluşan kategori sayısına göre değişiklik göstermektedir (Emel ve Taşkın, 2005).

2.4. SLIQ Algoritması

Karar ağacı içerisinde bulunan ve hızlı ölçeklenebilir sınıflayıcı algoritmalardan biridir. Hızlı ağaç budama algoritması mevcuttur (Çalış Boyacı vd. 2018).

2.5. SPRINT Algoritması

Büyük veri kümeleri söz konusu olduğunda ideal bir algoritmadır. Bu algoritma bölme işlemi yaparken tek bir niteliğin değerine dayanmaktadır. Tüm bellek sınırlamaları üzerinde nitelik listesi veri yapısı kullanarak işlem yapar (Çalış Boyacı vd. 2018).

3. Sonuç ve Öneriler

İstatistiki verilerin analizlerinde sıklıkla kullanılmakta olan veri madenciliği pek çok probleme çözüm üretmektedir. Veri madenciliğinin pek çok yöntemi bulunmaktadır ve etkin çözümlerde yapmaktadır bu yöntemler. Ancak bu problemler yapısı gereği çalışma prensipleri açık olarak bilinememektedir. Bu konuda karar vericilere daha açık ve net şekilde yapı oluşturup çözümler üreten karar ağaçları oldukça popüler yöntemlerden birisidir. Birçok alanda yaygın şekilde kullanılmaktadır.

KAYNAKÇA

Chien, C. F. & Chen, L. F. (2008). "Data Mining to Improve Personnel Selection and Enhance Human Capital: A Case Study in High-Technology Industry," Expert Systems with Applications, vol. 34, p. 280-290.

Çalış Boyacı, A., Durmaz, K. İ. & Gencer, C. (2018). Uçak Seferlerindeki Rötarları Etkileyen Faktörlerin Analizi, International Journal of Economic and Administrative Studies, 18. EYİ Special Issue, 179-190.

Emel, G. G. & Taşkın, Ç. (2005). Veri Madenciliğinde Karar Ağaçları ve Bir Satış Analizi Uygulaması, Eskişehir Osmangazi Üniversitesi Sosyal Bilimler Dergisi, 6 (2): 221-239.

Gargano, M. L. & Raggad, B. G. (1999). "Data mining–A Powerful Information Creating Tool", OCLC Systems & Services 2(15), s. 81–90.

Han, J. & Kamber, M. (2000). Data Mining Concepts and Techniques, Morgan Kaufmann Publishers, 1st Ed., San Francisco, USA.

Ma, Y. (1998). Data Warehousing, OLAP, And Data Mining: An Integrated Strategy For Use At FAA, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology.

Savaş, S., Topaloğlu N. & Yılmaz, M. (2012). "Veri Madenciliği ve Türkiye'deki Uygulama Örnekleri", İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi, sayı 21, s. 1-23.

Zhang, D. & Zhou, L. (2004). Discovering Golden Nuggets: Data Mining in Financial Application, IEEE Transactions on Systems, Applications and Reviews, Vol: 34, No:4, 513-515

Zorman, M., Vili, P., Kokol, P., Peterson, M., Sprogar, M. & Ojstersek, M. (2001). "Finding The Right Decision Tree's Induction Strategy For A Hard Real World Problem", International Journal Of Medical Informatics 1-2 (63), s. 109-121.