ANADOLU ÜNİVERSİTESİ

Açıköğretim Uygulamaları ve Araştırmaları Dergisi
AUAd
auad.anadolu.edu.tr

Gönderme Tarihi: 10.03.2019 Kabul Tarihi: 15.07.2019

Veri Madenciliği ve Bilgi Keşfi* (Kitap özeti)

Murat ARTSIN[a]

[a] Bahçeşehir Üniversitesi Uzaktan Eğitim Departmanı

Özet

Veri madenciliği, yapay sinir ağları, kümeleme analizi gibi konularda bilgi sahibi olmak isteyen araştırmacılar için Tuğba Şimşek Gürsoy tarafından yazılan *Veri Madenciliği ve Bilgi Keşfi* kitabı yararlı olacaktır. Kitapta veri madenciliğinin kullanımına dair uygulamalar örnekler ile sunulmuştur. Yazar kitapta veri madenciliğinde kullanılan modellerden, analizlerden bahsederek derinlemesine bilgiler sunmuştur. Nerede ise her analiz ve model için uygulama örneklerinin bulunduğu bu kitap veri madenciliğine giriş niteliğindedir. Bu bağlamda açık ve uzaktan öğrenme alanında veri madenciliği ile ilgili araştırmalar gerçekleştirecek olan araştırmacılar için titizce hazırlanmış bir kaynak olduğu belirtilebilir.

Anahtar Sözcükler: veri madenciliği, kümeleme, sınıflama, istatistik.

Abstract

Data Mining and Knowledge Discovery written by Tugba Simsek Gursoy will be useful for researchers who would like to learn about data mining, artificial neural networks, clustering analysis. In the book, applications related to the use of data mining are presented with examples. The author provides in-depth information on models and analysis used in data mining. This book is an introduction to data mining with examples of applications for each analysis and model. In this context, it can be stated that it is a meticulously prepared resource for researchers who will conduct research on data mining in the field of open and distance learning.

Keywords: data mining, clustering, classification, statistics

Kaynak Gösterme

Artsın, M. (2019). Veri Madenciliği ve Bilgi Keşfi* (Kitap özeti). *AUAd*, 5(3), 174-180.

Giriş

Şimşek Gürsoy (2009) tarafından hazırlanan bu kitap 2009 yılında yayınlanmıştır. On bölümden oluşan kitapta veri madenciliği, kümeleme analizi analizi gibi konulara değinilmiştir. Kitap bölümleri; veri ambarı, veri madenciliği, veri madenciliği modelleme yöntemlerine genel bir bakış, ilişki analizi, kümeleme analizi, sınıflandırma, istatiksel tahmin modelleri, web madenciliği, müşteri ilişkileri yönetimi ve veri madenciliği uygulama alanlarıdır.

Veri ambarı başlıklı birinci bölümde yazar etkin kararların alınabilmesi için veri ambarı hakkında bilgi sunmaktadır. Veri ambarının tanımının yanı sıra sahip olduğu bileşenlerden ve veri ambarlarının amaçlarından bahsetmiştir. Veri ambarı; iç veri kaynakları ve dış veri kaynakları olarak iki alt bileşenden oluştuğu ifade edilmiştir. İçeri veri kaynakları operasyonel verileri ve kayıtlı arşiv verileri kapsamaktadır. Dış veri kaynakları da dışarıdan sağlanan verilerdir. Bu bölümde veri ambarının sahip olması gereken özellikler verinin zamana bağlı olması, kalıcı olması, konuya yönelik olması ve entegre edilmiş olması olarak belirtmiştir. Veri ambarlarında kullanılan veri tabanlarından bahsedilmiştir. Yaygın olarak kullanılan Çevrimiçi Analitik İşleme (OLAP), Çok Boyutlu Çevrimiçi Analitik İşleme (MOLAP) ve İlişkisel Çevrimiçi Analitik İşleme (ROLAP) terimleri tanıtılmış ve veri ambarı ile veri madenciliği arasındaki ilişkiye değinilmiştir.

Veri madenciliği başlıklı ikinci bölümde yazar veri madenciliği kavramının derinlemesine tanımını gerçekleştirmiştir. Verinin anlamlı ilişkiler ve örüntüler çıkarma süreci olduğunu ve farklı isimler ile ifade edildiği belirtmiştir. Gürsoy'a göre (2009) veri madenciliği; bilgisayar programları aracılığıyla büyük miktarda var olan veriden tahminler üretme ve kurallar meydana getirilmesidir. Yazar veri madenciliğinin çok disiplinli bir alan olduğunu ve uzman sistemler, veri tabanları, görselleştirme, makine öğrenimi, istatistik gibi kavramları içerisinde barındırdığından bahsetmiştir. Genel bir veri madenciliği süreci için bir yol haritası sunulmuştur.

Veri madenciliği modelleme yöntemlerine genel bir bakış başlıklı üçüncü bölümde yazar veri madenciliğinde var olan tanımlayıcı ve tahmin edici modellerden bahsetmiştir. Tanımlayıcı modeller ilişki analizi ve kümeleme analizi olarak iki grupta incelenmiştir. Ek olarak birliktelik kuralları ve ardışık zamanlı örüntüler de ilişki analizi içerisinde yer bulmuştur. Tahmin edici modeller sınıflama ve istatiksel tahmin modelleri olarak iki başlık altında incelenmiştir. İnsan beyninin sahip olduğu öğrenme yoluyla bilgi oluşturma sürecini kullanılan bilgisayar sistemleri olan yapay sinir ağlarından bahsedilmiştir. Bunun yanı sıra

Darwin'in evrim kuramında yer alan 'doğada en iyinin yaşaması' kuralından esinlenerek oluşturulan yöntem olan genetik algoritmalardan ve istatiksel tahmin tekniklerinden bahsedilmiştir. İstatiksel tahmin modelleri genel olarak regresyon analizi, diskriminant analizi ve lojistik regresyon analizlerini kapsadığı belirtilmiştir.

İlişki analizi başlıklı dördüncü bölümde yazar birliktelik kuralları, ardışık zamanlı örüntüler ve pazar sepeti analizi veri madenciliğinde yaygın olarak kullanılan tekniklerden bahsetmiştir. Bunların yanı sıra apriori algoritması ve ardışık zamanlı örüntüler hakkında örnekler sunularak detaylı açıklamalar getirilmiştir.

Kümeleme analizi başlıklı beşinci bölümde yazar kümeleme analizinden, kümeleme analizinin geometrik gösteriminden, kümeleme analizinin amacından, benzerlik ölçüsünden ve kümeleme tekniklerinden bahsetmiştir. Sınıflandırma ve kümelemenin aralarındaki önemli farkın; sınıflandırmada sınıfların önceden belirlenmiş olmasının altı çizilmiştir. Bu bölümde kümeleme teknikleri hiyerarşik yöntemler ve hiyerarşik olmayan yöntemler olarak iki farklı yöntemden oluştuğu belirtilmektedir. Yazar bu bölümde merkezi kümeleme, tek bağlantı, tam bağlantı, ortalama bağlantı, Ward yöntemi gibi yöntemleri hiyerarşik yöntemler içerisinde sıralarken k-ortalamalar, metoid parçalama yöntemi, yığma kümeleme yöntemi, bulanık kümeleme yöntemi gibi yöntemleri de hiyerarşik olmayan yöntemler olarak sıralamaktadır. Hiyerarşik yöntemlerin kullanımı kolay, veri tipine uygun esnek algoritmalara sahip ve küme sayısı önceden bilinmiyor iken hiyerarşik olmayan yöntemlerde ise küme sayısı önceden bilinmektedir.

Sınıflandırma başlıklı altıncı bölümde yazar sınıflandırma, karar ağaçları, yapay sinir ağları, veri madenciliği ve genetik algoritmalar konusunda bilgiler sunmuştur. Sınıflandırma bir nesnenin belirli bir sınıf içinde hangi sınıfta bulunup bulunmadığını belirleyerek oluşma süreci olarak ifade edilmiştir. Karar ağaçlarında kullanılan kare, daire ve sembollerin açıklamaları belirtilmiş ve örnek karar ağacı sunulmuştur. Bu bölümde CART, CHAID, C4.5 ve QUEST gibi karar ağacı algoritmalarından bahsedilmiştir. CART sınıflandırma ve regresyon için kullanılmakta, CHAID F testi ve Ki-Kare testini kullanmaktadır. Yapay sinir ağları ile ilgili olarak; girdiler, ağırlıklar, toplama fonksiyonu, aktivasyon fonksiyonu ve çıktı gibi konu başlıkları açıklanmıştır. Yapay sinir ağlarında modeller ağın yapısına ve öğrenme türüne göre farklı sınıflandırmaların söz konusu olduğu belirtilmektedir. Bunlar; ileri beslemeli ve geri besmelidir. Son olarak da Darwin'in evrim teorisinden esinlenerek oluşturulan genetik algoritmalardan bahsedilmiştir. Genetik algoritmalar kromozon/gen, çözüm havuzu, çaprazlama, mutasyon, uygunluk fonksiyonu, yeniden üretim elemanlarından oluşmaktadır.

İstatiksel tahmin modelleri başlıklı yedinci bölümde yazar istatiksel tahmin modellerinden bahsetmiş olup regresyon analizi, basit doğrusal regresyon, çoklu doğrusal regresyon, diskriminant analizi ve lojistik regresyon analizi hakkında bilgiler sunmuştur. Regresyon analizlerinde bağımsız değişkenlerin çeşitli değerler karşılığında bağımlı değişkenin alacağı değeri tahmin edebilme olarak tanımlamaktadır. Basit doğrusal regresyon analizi bağımlı bir değişkenin tek bir bağımsız değişken ile arasındaki ilişkinin ifade edilebilmesi olarak tanımlanmaktadır. Diskriminant analizinde ise önceden sınıflandırılmış birden fazla grubu birbirinden ayıran özelliklerin belirlenmesi amacıyla kullanılmaktadır. Lojistik regresyon analizi de diskriminant analizine alternatif bir analizdir.

Web madenciliği başlıklı sekizinci bölümde yazar web madenciliği, webde saklı verilerin önemi, web logları, web madenciliği türleri, web madenciliğinin yararları ve sakıncalarından bahsedilmiştir. Web dokümanlarının harf, kelime, cümle, paragraf, bölüm ve noktalama işaretleri HTML gibi yapılarda bulunmakta olduğu belirtilmektedir. İnternette oldukça büyük miktarda veri bulunmakta olduğu ve araştırmalarda sınıflandırma ve kümeleme ile yeni çalışmaların gerçekleştirilebileceği belirtilmektedir. Web madenciliğinde dosya bazlı veriler, trafik bazlı veriler, izlenen yol bazlı veriler, ziyaretçi bazlı veriler, teknik bazlı veriler, sayfa bazlı verilerden faydalanılmaktadır. Web logları kullanıcıların web sayfalarını ziyareti sırasında sunucu tarafında kaydedilen bilgiler olarak tanımlanmaktadır. Web madenciliği kendi içerisinde web içerik madenciliği, web yapı madenciliği, web kullanım madenciliği olarak alt türlere ayrılmaktadır.

Müşteri ilişkileri yönetimi başlıklı dokuzuncu bölümde yazar müşteri ilişkileri yönetimi, müşteri ilişkilerinin ortaya çıkışı, müşteri ilişkileri kavramları, amaçları, süreçleri, mimarisi ve yararları konusuna değinmiştir. Bu bölümde veri madenciliği ile ilgili olarak kurumların müşterilerini daha iyi tanıyabilmeleri açısından veri madenciliğinin önemine vurgu yapılmaktadır. Müşterilerin hangi ürünü hangi ürün ile beraber aldıkları, hangi müşterinin hangi zaman aralığında ürün aldığı gibi konularına veri madenciliği yöntemleri kullanarak müşteri hakkında detaylı bilgilerin sağlanabilmesi için öneri bilgiler sunulmuştur. Yine bu bölümde müşteri portfolyosu anlamında geniş kitlelere sahip olan kuruluşların gerçekleştirdikleri süreçler hakkında örnekler sunmuştur.

Veri madenciliği uygulama alanları başlıklı onuncu bölümde yazar genel olarak veri madenciliğinin uygulama alanlarından bahsetmiştir. Bunlar; finans sektörü, perakende sektörü, telekomünikasyon sektörü, tıp, hilekarlığın tespiti ve diğer uygulamalar olarak listelenmiştir. Bu bölümde ilgili sektörlerde öncü niteliğinde olan büyük ve kurumsal firmaların veri

madenciliği bağlamında gerçekleştirdiği süreçler örnekler ile sunularak okuyucu bilgilendirilmiştir.

Sonuçlar

Bu kitap veri madenciliği bağlamında, modeller, karar ağaçları, yapay sinir ağları, analizler, yöntemler ve algoritmalar hakkında yol gösterici kaynak niteliği sunmakta olduğu bilgilerin yanı sıra açık ve uzaktan öğrenme alanında da kullanılabilecek olan birçok farklı yöntem ve algoritmayı sunması açısından oldukça önemlidir. Kitapta birçok analiz, yöntem örnekler ile sunulmuştur. Her ne kadar eğitimde kullanılabilecek örnekler bulunmasa da finans, telekomünikasyon ve tıp gibi alanlarda kullanılan örneklerin bulunması açısından oldukça değerlidir. Bundan dolayı bu kitapta bahsedilen algoritma ve yöntemlerin kullanılması açık ve uzaktan öğrenme alanında gerçekleştirilen çalışmalarda yol gösterici olabileceği düşünüyorum. Kitabın genel olarak temel istatistik bilgisine sahip araştırmalara kaynak olabileceği ifade edilebilir.

Öneriler

Şimşek Gürsoy (2009) tarafından oluşturulan bu eser veri madenciliğinde kullanılan birçok algoritma ve yönteme yer vermektedir. Kitap temel istatistik bilgisine sahip araştırmalar için uygun bir giriş kitabı olabilir. Bunun yanı sıra sade ve iyi örnekler ile kavramların okuyucu tarafında temellenmesinin sağlandığı belirtilebilir. Veri madenciliğinin açık ve uzaktan öğrenme alanına sağlayabileceği yeni değerler için bu bağlamda yapılacak olan yeni çalışmaların bir ihtiyaç olduğu belirtilebilir. Kitapta bahsi geçen modeller, karar ağaçları, yapay sinir ağları, analiz, yöntem ve uygulamalardan açık ve uzaktan öğrenme alanında öğrenci davranışlarının incelenmesi gibi konularda gerçekleştirilen çalışmalarda faydalanılabilir. Bunun yanı sıra tahmin modelleri ile ders tamamlama ve başarı durumlarının tespiti amacıyla gerçekleştirilecek çalışmalarda kullanılabilecek yöntemleri de barındırmaktadır. Sınıflandırma ve kümeleme analizleri ile açık ve uzaktan öğrenme alanında öğrenci davranışları, başarı notları ve benzeri öğrencilerin dersleri tamamlama noktasında etkili olduğu ifade edilen değişkenler ile dersleri tamamlayabilmeleri noktasında bu analizlerde faydalanılarak öğrencilerin profillerine göre yol haritaları sunulması konusunda çalışmalar gerçekleştirilebilir. Web madenciliğinden de öğrencilerin Öğrenme Yönetim Sistemleri (LMS) içerisindeki davranışlarının incelenmesinde faydalanılabilir. Genel olarak kitap, açık ve uzaktan öğrenme alanında yapılacak çalışmalar için başlangıç düzeyinde referans bir kaynak olabilir.

Kaynakça

Şimşek Gürsoy, U. T. (2009). *Veri Madenciliği ve Bilgi Keşfi*. Birinci Baskı, Pegem Akademi, Ankara.

Yazar Hakkında

Murat ARTSIN

Murat Artsın, Bahçeşehir Üniversitesi Uzaktan Eğitim Departmanında Eğitim Teknoloğu olarak görev yapmaktadır. Artsın, lisans eğitimini Sakarya Üniversitesi Bilgisayar ve Öğretim Teknolojileri Eğitimi Bölümünde 2016 yılında tamamlamıştır. Yüksek lisans eğitimini ise Anadolu Üniversitesi Uzaktan Eğitim Bölümünde 2018 yılında tamamlamıştır. Artsın, Bahçeşehir Üniversitesi Eğitim Teknolojileri Bölümünde Doktora eğitimine devam etmektedir. Yazarın ilgi alanları kitlesel açık çevrimiçi dersler, öz-yönetimli öğrenme, öğrenen-içerik etkileşimleridir.

Posta adresi: Bahçeşehir Üniversitesi Güney Kampüsü Beşiktaş/İstanbul
Eposta: artsinm@gmail.com