

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/328476653>

Veri Madenciliđi ve Önemi

Conference Paper · October 2018

CITATIONS

0

READS

98

2 authors:



Timuçin BARDAK

Bartın University

30 PUBLICATIONS 27 CITATIONS

SEE PROFILE



Eser Sözen

Bartın University

27 PUBLICATIONS 18 CITATIONS

SEE PROFILE

Some of the authors of this publication are also working on these related projects:



APRİORİ [View project](#)



Evaluation of Biomass Use In Terms of Energy, Environment, Health and Economy [View project](#)

Veri Madenciliği ve Önemi

Timuçin BARDAK¹, Eser SÖZEN^{2,*}

¹Bartın Üniversitesi, Bartın Meslek Yüksekokulu, Mobilya ve Dekorasyon Programı, Bartın, Türkiye

²Bartın Üniversitesi, Orman Fakültesi Orman Endüstri Mühendisliği Bölümü, Bartın Türkiye

*esozen@bartin.edu.tr

ÖZET

Veri madenciliği, büyük veri setlerinde gizli olan ilginç ve değerli bilginin keşfidir. Aynı zamanda bir sanat ve teknolojidir. Bilim ve iş dünyasında yaygın olarak veri madenciliği teknikleri kullanılmaktadır. Dünyada veri sektörüne ve veri analistlerine talep hızla artmaktadır. Firmalar veri bilimi sayesinde farklı pazarlarda büyük başarılar sağlayabilmektedir. Günümüzde kamu kaynaklarının daha verimli kullanmak için veri madenciliğinden sıklıkla faydalanılmaktadır. Biyoloji, eğitim, tıp ve bilgisayar bilimi gibi birçok araştırma alanında veri madenciliği etkili çözümler sunabilmektedir. İstatistik gibi geleneksel yöntemlerin zorlandığı noktalarda veri madenciliği yardımcı olabilmektedir. Gelişmiş ülkeler veri madenciliğinin potansiyelini ve gücünü anlamış durumdadır. Bu nedenle veri bilimine yapılan yatırımlar her geçen gün artmaktadır. Bunlara ek olarak dünyada büyük veri devri başlamıştır. Bu yeni devrimdeki gelişmeler oldukça heyecan vericidir. Bu çalışmada veri madenciliğinin tanımı ve önemi hakkında genel bilgi verilmiştir.

Anahtar Kelimeler: Veri Madenciliği, Büyük Veri, Veri İşleme, Bilgi Sistemleri

Data Mining and Its Importance

ABSTRACT

Data mining is the discovery of interesting and valuable information hidden in large data sets. It is also an art and technology. Data mining techniques are widely used in science and business. Demand for the data sector and data analysts is increasing rapidly in the world. Companies can achieve great success in different markets thanks to their knowledge. Today, data mining is frequently used to make public resources more efficient to use. Data mining can provide effective solutions in many fields of research such as biology, education, medicine and computer science. Data mining can help in places where traditional methods such as statistics are challenging. Developed countries are aware of the potential and power of data mining. For this reason, investments in data science are increasing day by day. In addition to this, large data circulation has started in the world. The developments in this new revolution are quite exciting. In this study, general information about the definition and the importance of data mining is given.

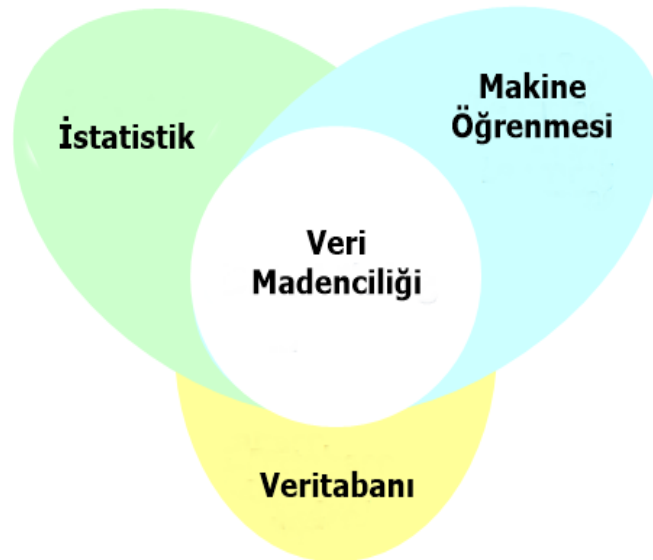
Keywords: Data Mining, Large Data, Data Processing, Information Systems

1.GİRİŞ

Bilgisayar teknolojisinin yükselmesi ve hızla gelişmesiyle teknolojik inovasyonda önemli bir aşama kat edilmiştir (Azzalini ve ark., 2012). Yaşadığımız dünyada her gün çok miktarda veri üretilmektedir. Fakat bu veriler her zaman anlamlı bilgiye dönüşmemektedir (Özçakır ve Çamurcu, 2007). Yirmi birinci yüzyılda şirketler öncelikli olarak müşteri memnuniyetini en üst düzeye çıkarmak istemektedirler. Müşteriye zamanında yüksek kaliteli ürünleri çeşitli teknikler ile firmalar hızlı bir şekilde sunabilmektedir. Bu başarıyı müşteri, satıcı ve tedarik zincirini anlamak sağlar. Genellikle bu tür bilgiler veritabanlarının incelenmesi ile elde edilir. Örneğin veri tabanında basit bir filtre ile ilk on müşteri bulunabilir (King, 2015). Fakat sadece veri tabanlarını oluşturmak yeterli değildir. Onları çeşitli veri madenciliği teknikleri ile değerli bilgilere çevirmek gerekir. Bilim dünyası açısından bakıldığında bir bilim insanının işi fiziksel dünyanın nasıl işlediğini anlamak ve onu düzenleyen kalıpları keşfetmek aynı zamanda yeni durumlarda neler olacağını tahmin etmektir. Girişimciler açısından bakıldığında ise genel olarak bir girişimcinin görevi, fırsatları bir başka deyişle herhangi bir işi karlı bir işe dönüştürülebilir davranış kalıplarını tanımlamak ve bunları kullanmaktır (Witten ve Frank 2005). Ülkelerde benzer şekilde vatandaşlarına daha iyi hizmetler sunabilmek için sorun kalıplarını anlaması gerekir. Sonuç olarak hayatın her alanında gelişmek için verilerden faydalanmaya ihtiyaç vardır. Bu çalışmada veri madenciliği ve önemine ilişkin bilgiler sunulmuştur.

2. VERİ MADENCİLİĞİ VE TANIMI

Günümüzde, bilgi sistemleri, büyük miktarlarda veriyi saklayarak, işleyerek ve analiz ederek karar verme süreçlerinde yararlı bilgiler sağladığından bir çok kuruluş için vazgeçilmezdir (Morais ve ark., 2017). Veri, madenciliği en basit tanımı ile var olan verilerden anlamlı bilgilerin elde edilmesi olarak tanımlanabilir. Veri bilimi multidisiplinerdir ve birçok bileşenden oluşur. Şekil 1'de veri madenciliğinin farklı disiplinler ile kesişimleri gösterilmiştir.



Şekil 1. Veri madenciliğinin farklı disiplinler ile kesişimleri (Sayad., 2018).

Veri madenciliği veri setlerini analiz etmek için istatistik, veritabanı yönetimi, yapay zeka, veri görselleştirme ve raporlama gibi araçları bir araya getirir. Çoğu veri madenciliği türü, belirli bireyler hakkında bilgi sahibi olmak yerine bir grup hakkında genel bilgi sağlamaya yöneliktir. Böylece daha büyük kitlelere ulaşılabilir (Clifton, 2018).

3. VERİLERİN HAZIRLANMASI

Veri madenciliğinde başarılı olmak için amaçların ve değerlendirme kriterlerinin açıkça ortaya konması gerekir. Veri biliminin önemli bir parçası da verilerin hazırlanmasıdır. Çünkü genellikle ham veriler eksiktir, işlenmemiş, gürültülüdür, çok büyük miktardadır ve birbirinden farklı türdedir. İşlenmemiş verilerden herhangi bir sınıflandırma, regresyon veya kümelendirme modeli ile anlamlı bilgi çıkarılamaz. Ham verileri temizlemenin amacı iki yönlüdür:

- Ham veri kümelerinden bazı temel bilgileri ayıklamak
- Verilerin kalitesini değerlendirmek ve denetimsiz veya denetimli öğrenim için temiz veri kümeleri oluşturmak

Veri hazırlama veri biliminin önemli bir parçasını oluşturmaktadır. Hazırlama işlemi veri temizleme ve özellik mühendisliği gibi iki ana kavramı içerir. Şekil 2’de verilerin hazırlanma işleminin temel kısımları gösterilmiştir.



Şekil 2. Verilerin hazırlanma işlemlerinin temel kısımları (Gill, 2018).

Veri ön işleme işlemleri dört gruba ayrılabilir bunlar:

- Veri temizliği: kalitesini iyileştirmek için veriden kaynaklanan hataları ve tutarsızlıkları tespit etme ve kaldırma
- Veri entegrasyonu: birden daha fazla bilgi kaynağından verileri entegre etmek
- Veri dönüşümü: verileri veri madenciliği için uygun biçimlere dönüştürmek
- Veriyi azaltmak: en fazla özellik sayısının çok büyük olduğu medikal alanda kullanılmaktadır. Bunun temel amacı, değişkenlerin, istikrarlı ve doğru bir model oluşturmak için verilerin bütünlüğünü kaybetmeksizin sayısının azaltılmasıdır (Idri ve ark., 2018).

4. VERİ MADENCİLİĞİ TEKNİKLERİ

Verileri analiz etmek ve modellemek için kullanılan yöntemler iki önemli kategoriye ayrılabilir: denetimli öğrenme ve denetimsiz öğrenme. Denetimli öğrenmede bağımsız değişkenler ve değeri tahmin edilecek bir bağımlı değişken gereklidir. Bu durumda, model bağımsız değişkenleri dikkate alarak hedef değişkenin değerini nasıl tahmin edeceğini öğrenir. Denetimsiz öğrenmede tüm değişkenleri eşit olarak ele alınır ve ana amaç, veriler arasında kalıp ve ilişkilerin bulunmasıdır (Ioniță ve Ioniță, 2018). Veri madenciliği teknikleri her türlü eski veya gelişmekte olan verilere uygulanabilir. Veri madenciliğinde birçok teknik vardır. Hangi tekniğin kullanılacağı verinin türüne, veri kümesinin boyutuna, uygulanan görevlerin kapsamına bağlıdır. Her veri kümesinin kendine ait uygun veri madenciliği

çözümü vardır. Örneğin sosyal ağlardan elde edilen devasa verilere uygulanan teknikler ile belirli bir kurumdaki sınırlı sayıda kişilerden elde edilen verilere uygulanan teknikler birbirinden farklıdır (Makhabel, 2015).

Veri madenciliği için temel teknikler yedi farklı gruba ayrılabilir bunlar aşağıda gösterilmiştir:

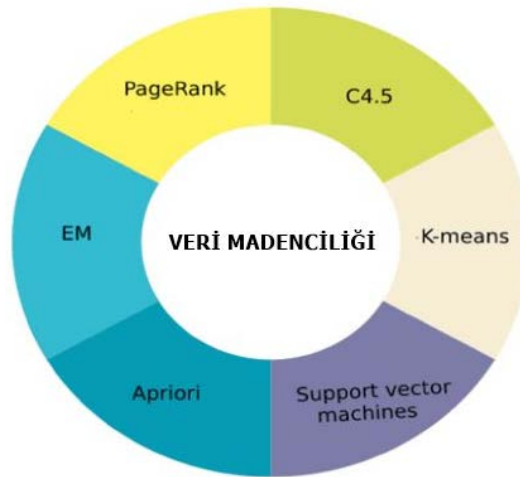
- Sınıflandırma ve tahmin,
- Kümeleme,
- İlişki kuralları,
- Aykırı değer saptama,
- Sıralı dizi analizleri,
- Zaman serisi analizi
- Metin madenciliği, web madenciliği ve sosyal ağların analizi (Zhao, 2013).

Veri madenciliği yöntemi belirli adımlardan oluşur. Şekil 3’de veri madenciliğinin genel adımları gösterilmiştir.



Şekil 3. Veri madenciliğinin temel adımları (Dontha, 2018).

Veri madenciliğinin her bir tekniği için çok sayıda algoritma bulunmaktadır. Algoritma sayısının çok fazla olması hangi durumda hangi algoritmanın daha başarılı her zaman kestirilememesi veri madenciliğinin zorluklarından biridir. Şekil 4’de veri madenciliğinde en çok kullanılan bazı algoritmalar gösterilmiştir.



Şekil 4. Veri madenciliğinde en çok kullanılan algoritmalar (Bakshi, 2018).

Verileri hazırlamak, modeli kurmak, modelin parametrelerini doğru seçmek ve çıktığı anlamak için yetenekli ve uzman kişilere ihtiyaçları vardır. Verilerin madenciliği hakkında temel sorunlar aşağıdaki gibi sıralanabilir:

- Verilerinin toplandığı kimselerin gizliliği ve güvenliğini sağlamak
- Çok büyük miktarda verilerin saklanma maliyeti
- Elde edilen anlamlı bilgilerin kötüye kullanılması
- Doğru verilerin toplanmasındaki zorluk

Tüm sayılan sorunlara rağmen veri madenciliği teknikleri ile yeni ve anlamlı bilgiler her alanda keşfedilebilmektedir.

5. BÜYÜK VERİ

Teknoloji, doğa gözlemlerinden büyük ve sürekli artan veriler üretmektedir. Bu büyük veri çalışmaları bir bilim disiplini olarak yerleşmiştir. Bu kavram günümüzde çok moda olmuştur ve herkes tarafından konuşulmaktadır. Özellikle şirketler ve hükümetler mevcut yeni bilgileri daha aktif kullanmaya çalışmaktadır (Torrecilla ve Romo, 2018). Büyük verinin yararı ve vaadiyle ilgili görüşler çok çeşitlidir. Bazıları, büyük verilerin nadir görülen hastalık süreçlerine yeni bakış açılarının sağlayacağını düşünürken, diğerleri büyük verilerin sadece verilere gürültü eklediğine söylemektedir. Son iki yılda insanoğlu tüm tarihinden daha fazla veri ürettiği tahmin edilmektedir. Büyük veri fayda ve maliyetleri ile birlikte düşünülmelidir. Çeşitli organizasyonlar şu anda bu büyük miktardaki verileri analiz edebilmekte, böylece pozitif değer yaratmakta akıllı stratejik kararlar verebilmektedir (Philip, 2018). Fakat çok büyük miktarda verinin işlenmesi ve analiz edilmesi için yeni analitik tekniklere ihtiyaç duyulmaktadır (Cobb ve ark., 2018). İlk bakışta, daha fazla veri daha az veriden daha iyi görünüyor olabilir. Fakat bu durum her şeyin eşit olması şartı ile doğrudur. Uygulamada, daha fazla veri elde etmek, çeşitli ek maliyetler içerecektir ve analizi karmaşıklaşacaktır (Faraway ve Augustin, 2018).

6. SONUÇ VE ÖNERİLER

Veri madenciliği yakın zamanda kadar istatistikçilerin bir çalışma alanı olarak düşünülürdü. Fakat veri madenciliğindeki hızlı ve büyük değişimler nedeni ile istatistikten ayrı bir bilim alanı olarak dünyada kabul görmeye başlanmıştır. Aynı zamanda bu alana gelişmiş ülkeler tarafından önemli yatırımlar yapılmakta ve konu ile ilgili araştırma sayısı her geçen gün artmaktadır. Verilerin çok hızlı artması, bilgisayarların güçlenmesi, algoritmalarındaki yeni gelişmeler ile veri bilimi birçok farklı alanda zor problemlerin çözümde yeni bir fırsat sunmuştur. Veri madenciliğinin çözdüğü problemler hayatı kolaylaştırmakta ve verimliliği arttırmaktadır. Veri madenciliğinin tüm faydaları yanında gizlilik, güvenlik ve bilginin kötüye kullanılması gibi konular endişe kaynağıdır. 2020 yılına kadar, dünyadaki her insan için her saniyede 1.7 megabayt veri oluşturulacağı düşünülmekte. Dahası, büyük veri teknolojisine yapılan harcamaların bu yıl 57 milyar dolar seviyesine ulaşması tahmin ediliyor (Gary, 2018).

KAYNAKLAR

- Azzalini, A., Walton, G., and Scarpa, B. (2012). *Data Analysis and Data Mining: An Introduction*, Oxford University Press, 263, Oxford.
- Bakshi, K., (2018). A List Of Top Data Mining Algorithms. Erişim Tarihi: 30. 08. 2018. <https://www.techleer.com/articles/438-a-list-of-top-data-mining-algorithms>
- Clifton, C., (2018). Data mining computer science Britannica.com. Erişim Tarihi: 20. 08. 2018. <https://www.britannica.com/technology/data-mining>
- Cobb, A. N., Benjamin, A. J., Huang, E. S., and Kuo, P. C. (2018). “Big data: More than big data sets,” *Surgery*, 1-3.
- Dontha, R., *Data Mining Steps - Digital Transformation for Professionals*. Erişim Tarihi: 20. 08. 2018. <https://digitaltransformationpro.com/data-mining-steps>
- Faraway, J. J., and Augustin, N. H. (2018). “When small data beats big data,” *Statistics & Probability Letters*, 136, 142–145.
- Gary, S., (2018) Why Big Data is Important to Your Business - inside BIGDATA. Erişim Tarihi:22. 08. 2018 <https://insidebigdata.com/2017/09/09/big-data-important-business>
- Gill, N. S., (2018). *Data Preparation,Preprocessing,Wragling in Deep Learning - XenonStack*. Erişim Tarihi: 22. 08. 2018. <https://www.xenonstack.com/blog/data-science/preparation-wragling-machine-learning-deep>
- Idri, A., Benhar, H., Fernández-Alemán, J. L., and Kadi, I. (2018). “A systematic map of medical data preprocessing in knowledge discovery,” *Computer Methods and Programs in Biomedicine*, 162, 69–85.
- Ioniță, I., and Ioniță, L. (2018). “Classification Algorithms of Data Mining Applied for Demographic Processes.,” *BRAIN: Broad Research in Artificial Intelligence & Neuroscience*, EduSoft SRL, 9(1), 94–100.
- King, R. S. (2015). *Cluster Analysis and Data Mining: An Introduction*, Mercury Learning & Information, 315, Dulles, Virginia.
- Makhabel, B. (2015). *Learning Data Mining with R, Community Experience Distilled*, Packt Publishing, Birmingham, 273, England.
- Morais, A., Peixoto, H., Coimbra, C., Abelha, A., and Machado, J. (2017). “Predicting the need of Neonatal Resuscitation using Data Mining,” *Procedia Computer Science*, Elsevier, 113, 571–576.
- Özçakır, F. C., and Çamurcu, A. Y. (2007). “Birliktelik Kuralı Yöntemi İçin Bir Veri Madenciliği Yazılımı Tasarımı ve Uygulaması,” *İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi*, İstanbul Ticaret Üniversitesi.
- Philip, J. (2018). “An application of the dynamic knowledge creation model in big data,” *Technology in Society*, 54, 120–127.
- Sayad, S., (2018). *Data Mining*. Erişim Tarihi: 29. 08. 2018. http://chem.eng.utoronto.ca/~datamining/dmc/data_mining.htm
- Torrecilla, J. L., and Romo, J. (2018). “Data learning from big data,” *Statistics & Probability Letters*, 136, 15–19.
- Witten, I. H., and Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques*, Second Edition, Morgan Kaufmann Series in Data Management Systems, Morgan Kaufmann, 622, Amsterdam.
- Zhao, Y. (2013). *R and Data Mining: Examples and Case Studies*, Academic Press, 225, Burlington.