Mehmet Akif Ersoy Üniversitesi
Sosyal Bilimler Enstitüsü Dergisi
ISSN: 1309-1387
Cilt: 9 Sayı: 22 2017 Aralık (s. 155-172)

BÜYÜK VERİ ANALİZİNDE YAPAY ZEKÂ VE MAKİNE ÖĞRENMESİ UYGULAMALARI

ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING APPLICATIONS IN BIG DATA ANALYSIS

Muhammet ATALAY[1], Enes ÇELİK[2]

Öz

Bilgi teknolojilerinde yüksek hızda yaşanan gelişmeler ve internet kullanımının çok yaygın hale gelmesi ile birlikte, çeşitli platformlarda biriken verinin çeşitliliği ve hacmi de artmıştır. Büyük veri kavramı ile ifade edilen bu verilerin işlenmesi ve anlamlı bilgilerin elde edilmesi, önemli sonuçlar elde edilebilmesine imkân vermektedir. Bu çalışmada, büyük veri analizinde yapay zekâ ve makine öğrenmesi tekniklerinin kullanımı tartışılmıştır. Başlıca yapay zekâ ve makine öğrenmesi teknikleri hakkında bilgiler verilerek, bu tekniklerin büyük verilerle yapılan uygulamalarından örnekler verilmiştir. Başlıca olarak; kümeleme, sınıflandırma, yapay sinir ağları, metin ve web madenciliği, fikir madenciliği ve duygu analizi alanlarında büyük verilerle yapılan çalışmalar anlatılmıştır.

Anahtar Kelimeler: Veri Madenciliği, Yapay Zekâ, Makine Öğrenmesi, Büyük Veri, Veri Analizi.

Abstract

With the rapid development of information technologies and the widespread use of the internet, the diversity and volume of data accumulated in various platforms has also increased. The processing of these data expressed in the concept of big data and the acquisition of meaningful information enable us to obtain important results. In this work, the use of artificial intelligence and machine learning techniques in big data analysis are discussed. The main artificial intelligence and machine learning techniques are given and some examples are given from the applications of these techniques. Mainly; an applications have been explained in the field of clustering, classification, artificial neural networks, text and web mining, opinion mining and sentiment analysis.

Keywords: Data Mining, Artificial Intelligence, Machine Learning, Big Data, Data Analysis.

[1] Yrd. Doç. Dr., Kırklareli Üniversitesi İktisadi ve İdari Bilimler Fakültesi İşletme Bölümü Sayısal Yöntemler Anabilim Dalı, atalay@klu.edu.tr
[2] Öğr. Gör., Kırklareli Üniversitesi Babaeski Meslek Yüksekokulu Büro Hizmetleri ve Sekreterlik Bölümü, enes.celik@klu.edu.tr

Başvuru Tarihi: 29.04.2017
Yayına Kabul Tarihi: 08.11.2017
DOI: 10.20875/makusobed.309727

1. GİRİŞ

İçeriğinden anlamlı sonuçlar çıkarmak ve gerektiğinde kullanmak üzere, devletler, kurumlar veya şahıslarca pek çok veri toplanmaktadır. Her alanda genellikle sayılar, metinler, ifadeler, şekiller, grafikler gibi malzemelerin oluşturduğu veriler, bilgisayarlarla elektronik ortamlara taşınmış bulunmaktadır. Bilgisayar, internet ve buna bağlı teknolojilerin hayatın her alanında ve daha çok yer almasıyla birlikte, bu teknolojilerin ürettiği verilerin de depolanması söz konusu olmaktadır. Bilgi teknolojilerinin gün geçtikçe daha fazla yaygınlaşması ise insanların yaşam, çalışma ve çevre şartlarını değiştirmiş; mekânlar, meslekler, çalışanlar "mobil", kullanılan cihazlar ise "mobil" ve "akıllı" hale gelmeye başlamıştır. Bununla birlikte doğan veriler ise hem çeşitlilik hem hacim bakımından çok farklı ve büyük boyutlara ulaşmış bulunmaktadır. Mobilitenin artması, sosyal ağların kullanımının yaygınlaşması, çeşitli takip sistemleri (sensörler, barkodlar, karekodlar, RFID sistemleri… vs.) teknolojilerinin gelişmesi, iletişim teknolojilerinin ulaşılabilirliğinin artması, başta ticari işlemler olmak üzere pek çok iş kolunun elektronik ortama taşınması ile birlikte hem üretilen verinin çeşitliliği hem de toplanma hızı ve miktarı da ciddi oranlarda artmıştır. Bu artış üstel olarak devam etmektedir. Öte yandan cihazlara takılan sim kart, algılayıcılar, elektronik devreler ve internet ağı sayesinde, cihazların uzaktan izlenmesini, yönetilmesini ve birbiriyle iletişim kurabilmesini sağlayan bir teknoloji olan Makineler Arası İletişim (M2M), hem bireylerin hem de şirketlerin hayatında geniş bir kullanım alanı bulmaktadır. Araç takibi, tıbbi otomasyon, akıllı ev aletleri, sayaç okuma, lojistik, güvenlik ve tarım gibi pek çok alanda bu teknolojilerin kullanılmasıyla birlikte cihazların ilettiği verilerin de analizi ihtiyaç haline gelmiştir. Kablosuz sensörlerin giderek yaygınlaşması ve Internet Protokolü Sürüm 6 (IPv6) ile adreslenebilecek nesne sayısının neredeyse sonsuz hale gelmesi ise internete bağlı olacak cihaz sayısında bir artış yaşanmasını sağlamış olup Cisco ve IBM'in öngörülerine göre 2020 yılında 50 milyar cihaz internet ağına dâhil olacaktır (Karaman vd., 2015). Bu gelişmeye paralel olarak, bir donanımın tek bir uygulamayla bağlantılandırıldığı M2M sistemleri, günümüzde neredeyse herhangi bir donanımın çeşitli uygulamalar veya cihazlarla birbirine kolayca bağlanabildiği Nesnelerin İnterneti (IoT), Her şeyin İnterneti (IoE), Nesnelerin Ağı (WoT) ve Her şeyin Ağı (WoE) gibi ortamlara evrilmiştir. Gerçek ve sanal dünyanın birbirine oldukça yakınlaştığı, hayatın her kesiminde akıllı ortamların meydana geldiği bu sistemlerde, muazzam bir veri hacmi üretilir ve üstelik bu verilerin çoğu yapılandırılmamıştır. Resim, ses, metin, video gibi pek çok türde olabilen ve ağlar üzerinden aktarılan bu veriler bulut ortamlarda da saklanmaya başlamıştır. Bu verilerle ilgili bir başka husus ise sosyal medya verileri gibi insan kaynaklı veriler başta olmak üzere, değişken ve dinamik bir diğer deyişle akan bir yapıya sahip olmalarıdır. Bir yandan sisteme cihazlardan yeni veriler dahil olmakta veya bazı veriler kesintiye uğramakta, öbür yandan mevcut verilerde değişiklik meydana gelebilmektedir. Toplanan verilerin analizi bu sebeple daha karmaşık bir hal almaktadır. "Big data" yani "büyük veri" kavramı bu sebeple de özellikle son yıllarda çokça tartışılır hale gelmiştir.

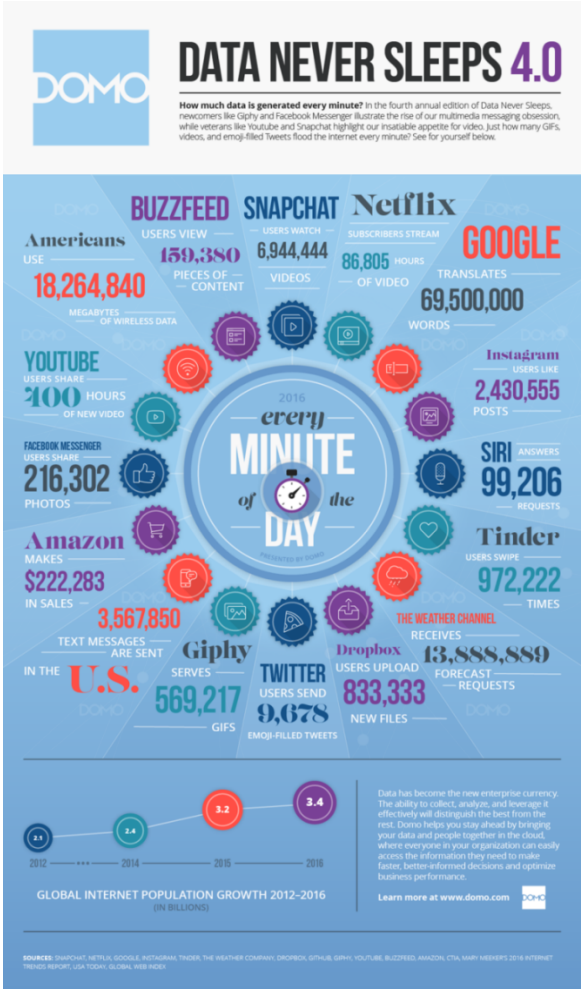
1.1. Büyük Veri Analizi

"Geleneksel veri işleme araçları ile analizi yapılamayan ve yönetilemeyecek kadar büyük miktardaki veri setleri" olarak tarif edilen (Ohlhorst, 2013: 1) büyük veri, kısaca 5V diyebileceğimiz beş kavram ile betimlenmektedir: Volume (Hacim), Velocity (Hız), Variety (Çeşitlilik), Verification

(Doğrulama) ve Value (Değer). Veri kaynakları ve veri çeşitlerindeki artış ile birlikte bu tanım; Volume, Velocity, Variety, Value, Veracity (Gerçeklik), Volatility (Oynaklık) ve Validity (Geçerlik) olmak üzere 7V (Khan vd., 2014), bunlara Vulnerability (Hassaslık), Variability (Değişkenlik) ve Visualization (Görselleştirme) ilave edilerek 10V (Firican, 2017) kavramlarını da kapsayacak şekilde genişletilebilir.

IDC tarafından yapılan "Digital Universe-Dijital Evren" çalışmasında (Turner ve Gantz, 2014), 2020 yılına kadar dijital verinin her iki yılda bir iki katına çıkacağı ve veri miktarının 44 zetabyte (44 trilyon gigabyte) olacağı tahmin edilmektedir. Öte yandan bu verilerin oluşması sürekli ve çok hızlı bir süreçtir. Domo (2016) tarafından hazırlanan "Data never sleeps 4.0" projesi kapsamında Haziran 2016 verilerine göre, yalnızca bir dakika içerisinde; Youtube video paylaşım sitesine kullanıcılar tarafından 400 saatlik video yüklemesinin gerçekleştirildiği, Twitter üzerinden 9.678 adet emoji içerikli tweet atıldığı, Google' da 69.500.000 kelime tercüme edildiği, sadece Amerikalı kullanıcıların mobil cihazlarla yaklaşık 18.000 GB veri kullandıkları, Facebook Messenger kullanıcılarının 216.302 adet fotoğraf paylaştıkları, Instagram kullanıcılarının paylaşılan görüntüler için 2.430.555 adet beğeni yaptıkları ve Amazon web sitesinden 222.283 $ satış yapıldığı tespit edilmiştir (Bkz. Şekil-1).

Şekil-1. "Data Never Sleeps 4.0" İnfografik



(Domo, 2016)

Aynı araştırmaya göre son beş yılda internet kullanıcı popülasyonu %60 artarak 3.4 milyara ulaşmış ve dünyadaki mobil cihaz sayısı insan nüfusunu geçmiş bulunmaktadır. Bu oran 2013-2015 arası ise %18.5' tur (Domo, 2015). Tüm bu rakamlar, depolanan, dolaşımdaki ve kullanımdaki verinin büyüklüğü kadar artış hızını da çarpıcı bir şekilde ortaya koymaktadır.

Büyük veriden kastedilen yalnızca hacimsel büyüklük değildir. Sosyal medya paylaşımları, ağ günlükleri, bloglar, fotoğraf, video, log dosyaları gibi farklı kaynaklardan ve farklı biçimlerde toplanan verilerin anlamlı ve işlenebilir hale getirilmesi gerekmektedir. Ayrıca bu veriler hacim ve tür yanında sürekli artan bir hızda oluşmakta ve depolanmaktadır. Şekil-1' deki rakamların yalnızca bir dakikalık sürede gerçekleşmiş olması, birbirine ve internet ağına bağlı cihaz sayısındaki artış hızıyla birlikte düşünülürse, verilerin depolanma ve değişme hızının ne noktada olduğu daha iyi anlaşılabilecektir. Öte yandan büyük verilerin çoğu zaman karmaşık, düzensiz olduğu ve yanlışlar içerebileceği (Gürsakal, 2014: 26) gerçeği, bu verilerin düzenlenmesi ve ayıklanması sorununu doğurmaktadır. 2013 yılında dijital dünyada faydalı olarak kabul edilen verilerin oranı %22 olarak gerçekleşmiş, ancak bunların analiz edilebilen kısmı %5' in altında gerçekleşmiştir. 2020 yılında nesnelerin internetine bağlı cihazların sayısına bağlı olarak faydalı verilerin oranının %35' ten fazla olacağı öngörülmektedir (TBD, 2014). Üstelik anlık alınan verilerden hemen bilginin elde edilmesi yani verinin toplandığı anda analiz edilmesi gerekmektedir. 2020 yılında dijital dünyada üretilen verilerin %10' unun makineler ve internete bağlanabilen nesnelerden kaynaklı olacağı tahmin edilmektedir (TBD, 2014). Bu katkının her geçen gün artacağı düşünülürse, bu büyük veri kümelerinin yönetilmesi, depolanması ve korunması için yeni yöntemler gerekecektir. Bu nedenle internete bağlı cihazların kaynaklık ettiği verilerin analizinde veri madenciliği yöntemleri yanında web, metin ve multimedya madenciliği teknikleri kullanılmaktadır (Gürsoy, 2017:15-17).

Burada unutulmaması gereken, tüm bu verilerin yalnızca dijital değil fotoğraf, resim, video, ses, metin, konum (GPS) bilgisi vs. gibi pek çok çeşitte ve her biri için çeşitli boyutlarda olduğudur. Böyle olunca da asıl önemli olan, bu kadar büyük, hızlı ve çeşitli olan veri topluluğundan anlamlı ve değerli bilgiyi elde etmek olmaktadır. Bu amaçla geliştirilen yöntemler için "Big Data Analysis-Büyük Veri Analizi" tabiri kullanılmaktadır.

1.2. Yapay Zeka

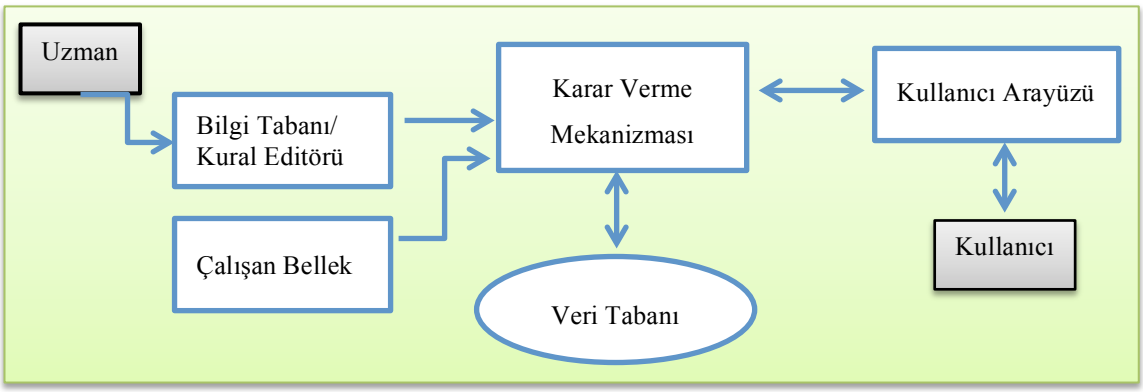
Doğadaki varlıkların akıllı davranışlarını yapay olarak üretmeyi amaçlayan (Charniak ve McDermot, 1985; Akt. Nabiyev, 2012: 25), bu meyanda işini mükemmel yapan canlı sistemlerini ve İnsan beynini model alan yapay zeka çalışmaları; günlük hayatın farklı alanlarında ürünler vermesinin yanında, tahmin, sınıflandırma, kümeleme gibi amaçlar için de kullanılmaktadır. Başlıca olarak uzman sistemler, genetik algoritmalar, bulanık mantık, yapay sinir ağları, makine öğrenmesi gibi teknikler, genel olarak yapay zeka teknolojileri olarak adlandırılmaktadır. Bu tekniklerin yanı sıra doğanın taklidi amacıyla da canlılar incelenmekte ve benzeri akıllı yöntemler önerilmektedir. Karınca kolonisi, parçacık sürü ve yapay arı gibi algoritmalar, yapay zeka optimizasyon teknikleri olarak kullanılmaktadır. Genel anlamda yapay zekadan kastedilen; insan zekasının, sinir sistemi, gen yapısı gibi fizyolojik ve nörolojik yapısının ve doğal olayların modellenerek makinelere (bilgisayar ve yazılımlara) aktarılmasıdır. Özetle yapay zeka; "insan gibi düşünen, insan gibi davranan, akılcı (rasyonel) düşünen ve akılcı davranan"

(Balaban ve Kartal, 2015: 16), canlıların zekice olarak kabul edilen davranışlarına sahip bilgisayar sistemleridir ve makine öğrenmesi bu anlamda yapay zekanın son evresi olarak kabul edilmektedir. Bu çalışmada, Çalışmanın devamında, yapay zeka tekniklerinden uzman sistemler, genetik algoritmalar, bulanık mantık, yapay sinir ağları ile makine öğrenmesi kavramı hakkında bilgi verilerek büyük veri uygulamalarına örnekler verilecektir.

1.2.1. Uzman Sistemler

Uzman sistemler, çözümü bir uzmanın bilgi ve yeteneğini gerektiren problemleri, bilgi ve mantıksal çıkarım kullanarak o uzman gibi çözebilen sistemlerdir. Yani problemi çözmede uzman kişi veya kişilerin bilgi ve mantıksal çıkarım mekanizmasının modellemesi amaçlanmaktadır (Edward Feigenbaum' dan aktaran: Harmon ve King, 1985: 5). Uzman sistemlerde, bilgiler depolanıp daha sonra bir problemle karşılaşıldığında bu bilgi üzerinden yapılan çıkarımlarla sonuçlara ulaşılmaya çalışılmaktadır. Böylelikle insan zekasının muhakeme etme sürecine, bilgisayarın kesinlik ve hızının katılması amaçlanmaktadır. Uzman sistemler şu temel ögelerden (Bkz. Şekil-2) teşekkül edilir: Bilgi tabanı (kural tabanı), veri tabanı, çalışan bellek (yardımcı yorumlama modülü), çıkarım motoru (karar verme mekanizması, mantıksal çıkarım modülü) ve kullanıcı arayüzü. Bilgi tabanı, bilgilerin tutulduğu ve tutulan bilgilerden yeni bilgiler üretilmesine imkan sağlayan birim olup uzman sistemin beyni ve yapı taşıdır. Veri tabanı ise bilgi tabanı ile sürekli ilişki halinde olmalıdır (Nabiyev, 2012: 409). Kullanıcı arayüzü; bilgi kazanma, bilgi tabanı ile hata ayıklama ve deneme, test durumlarını çalıştırma, özet sonuçlar üretme, sonuca götüren nedenleri açıklama ve sistem performansını değerlendirme gibi görevleri yerine getirmektedir. Çıkarım motoru, arama ve çıkarımın yer aldığı ögedir. Uygun bilgi için bilgi tabanını taramakta ve mevcut problem verisine dayanarak çıkarımda bulunmaktadır. Çalışan bellekte, problem ile ilgili soruların cevapları ve tanısal testlerin sonuçları gibi mevcut problem verisi saklanmaktadır (Balaban ve Kartal, 2015: 20). Çıkarım motorunda mantıksal sonuçlar elde edilmesine yardımcı olur.

Şekil-2. Uzman Sistemin Genel Yapısı



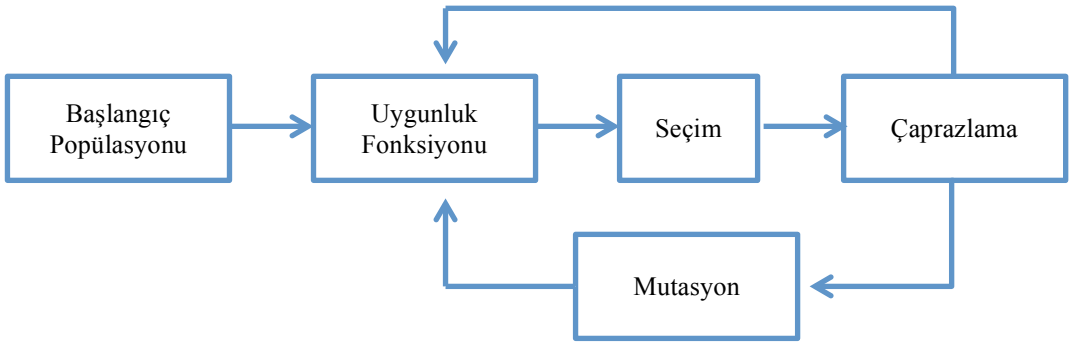
1.2.2. Genetik Algoritmalar

Genetik algoritmalar, evrim teorisinin dayandığı temel prensiplerinden olan doğal seçilim ile en iyi bireylerin hayatta kalması ilkesini taklit eden bir tekniktir. Burada yapılan, en iyi çözümün pek çok

çözüm seçeneği içinden arama yapılarak belirlenmesidir. Rassal arama teknikleri ile eldeki mevcut çözümlerden hareketle en iyi çözüme ulaşılmaya çalışılmaktadır. Basit bir genetik algoritmanın işlem adımları; problemin olası çözümlerinin dizilere (kromozomlar) kodlanarak çözüm yığınının oluşturulması, kromozomların çözüme yaklaşma başarısının uygunluk fonksiyonu ile değerlendirilmesi, genetik parametrelerin belirlenmesi, seçim stratejisi ve mekanizmaları, genetik operatörler ve durdurma kriteri olarak sıralanabilir (Elmas, 2011: 388-401). Genetik algoritmaların yapısı Şekil-3' te gösterilmiştir.

Genetik algoritmalar, bilinen yöntemlerle çözülemeyen veya çözüm süresi problemin büyüklüğüne göre oldukça fazla olan problemlerde, kesin sonuca çok yakın sonuçlar verebilen bir yöntemdir. Bu özelliği ile, NP (Nonpolynomially-Polinomal olmayan) problemler yanında gezgin satıcı, karesel atama, yerleşim, atölye çizelgeleme, mekanik öğrenme, üretim planlama, elektronik, finansman ve hücresel üretim (Elmas, 2011: 381) gibi konularda uygulanmaktadır.

Şekil-3. Genetik Algoritmanın Genel Akış Şeması



(Nabiyev, 2012: 604)

1.2.3. Bulanık Mantık

İki değerli mantıkta her şey ya doğru ya yanlıştır. Çok değerli mantıkta doğruluk derecelendirilebilir. Fakat bu iki ya da daha fazla değer arasında kalan durumlar izaha muhtaç kalmaktadır. 1965 yılında Prof. Lotfi Asker Zadeh, "Fuzzy Sets" başlıklı yazı ile bir dönüm noktası olarak yeni bir yönelim başlatmış ve bulanıklık kavramı dikkat çekmeye başlamıştır (Yang ve Liu, 2003: 305). Bulanık mantık, klasik mantıkta kullanılan kesin hatlarla birbirinden ayrılmış aralıklar yerine, tanımlanan fonksiyonlarla birbirine geçmiş çok sayıda aralıkları kullanmaktadır. Başka bir deyişle bulanık küme kuramı, klasik matematiğin standartlarına göre pek çok bakımdan belirsiz olan veya kesin olmayan karar süreçlerine matematiksel bir kesinlik kazandıran kavramlar ve yöntemler bütünüdür (Yenilmez, 2001: 2). Bu amaca matuf olmak üzere geliştirilen bulanık kümeler teorisiyle, insan gibi düşünebilen, karar verebilen ve seçim yapabilen sistemlerin oluşması amaçlanmıştır. Bulanık mantığı ve buna karşılık gelen matematiksel çatıyı kullanan sistemlere "bulanık sistemler" adı verilmektedir.

Bulanık sistemde bulanık kümelerin/bulanık mantığın kullanılması birkaç şekilde olabilir. Bunlar (Baykal ve Beyan, 2004: 190-191);

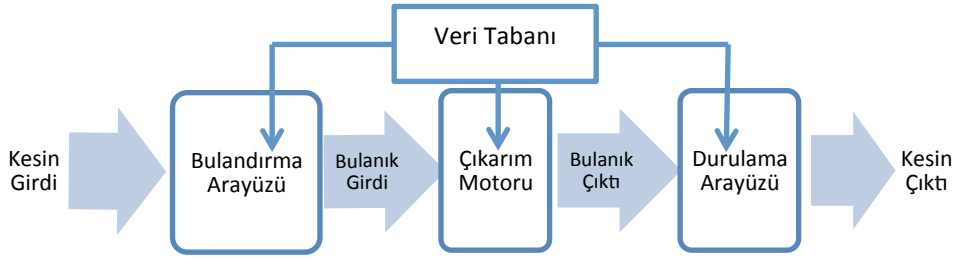
1. Sistem "eğer-o halde" şeklinde kurallarla tanımlanabilir. Bu şekilde tanımlanan sistemlere kural tabanlı bulanık sistemler adı verilir.

2. Sistem parametreleri gerçel sayılar yerine bulanık sayılar kullanılarak parametre değerlerindeki belirsizlik tanımlanabilir.

3. Sistemin girdi, çıktı ve durum değişkenleri insan algısı ile ilişkili nicelikleri ifade ediyor veya sözel bilgiyi taşıyorsa bu değişkenler bulanık küme ile tanımlanabilir.

Bu şekilde kurulan bulanık modeller, kullanım yerleri, kurulmaları sırasındaki bakış açısı ve vurgulanan boyutlarına göre; bulanık çıkarım sistemi, bulanık kural tabanlı sistem, bulanık uzman sistemler, bulanık mantık denetleyicileri olarak tanımlanırlar. Bulanık mantığa dayanan sistemler genel olarak dört bölümden oluşur: Bulandırma arayüzü, çıkarım motoru (karar verme mantığı), durulama arayüzü ve bilgi tabanı (Şekil-4).

Şekil-4. Bulanık Sistemin Genel Yapısı



(Baykal ve Beyan, 2004: 196)

1.2.4. Makine Öğrenmesi

Makine öğrenmesi, bir problemi o probleme ait veriye göre modelleyen bilgisayar algoritmalarının genel adıdır. Mevcut veri seti ve kullanılan algoritma ile oluşturulan model, en yüksek performansı vermek üzere kurulmaktadır. Bu sebeple pek çok makine öğrenmesi yöntemi geliştirilmiş olup bunlardan bazıları; k-en yakın komşu algoritması, basit (naive) Bayes sınıflandırıcı, karar ağaçları, lojistik regresyon analizi, k-ortalamalar algoritması, destek vektör makinaları ve yapay sinir ağlarıdır. Bu yaklaşımların bir kısmı tahmin ve kestirim, bir kısmı kümeleme ve bir kısmı da sınıflandırma yapabilme yeteneğine sahiptir.

Bu yöntemlerde öğrenme stratejileri; denetimli, denetimsiz ve pekiştirmeli (takviyeli) olmak üzere üç grupta incelenmektedir. Denetimli öğrenmede oluşturulan model ile, bir grup girdi değerine karşılık onlara ait hedef değerleri verilerek aralarındaki ilişkiyi öğrenmesi ve hedef değerlere en yakın çıktıların üretilmesi amaçlanır. Elde edilen en iyi model, yeni girdi değerleri için en yakın çıktıyı da verebilecektir. Denetimsiz öğrenmede ise hedef değerleri olmadan sadece girdi değerleri arasındaki ilişki ortaya çıkarılmaya çalışılır. Bu ilişki(ler) yardımı ile birbirine yakın değerler gruplandırılır yani kümeleme yapılır. Yeni bir girdi bu kümelerden hangisiyle ilişkili ise o kümeye ait olacaktır. Pekiştirmeli (takviyeli) öğrenme yönteminde, hedef çıktıyı vermek için bir danışman yerine, elde edilen çıkışın verilen girişe karşılık iyi ya da kötü olarak değerlendiren bir kriter kullanılmaktadır.

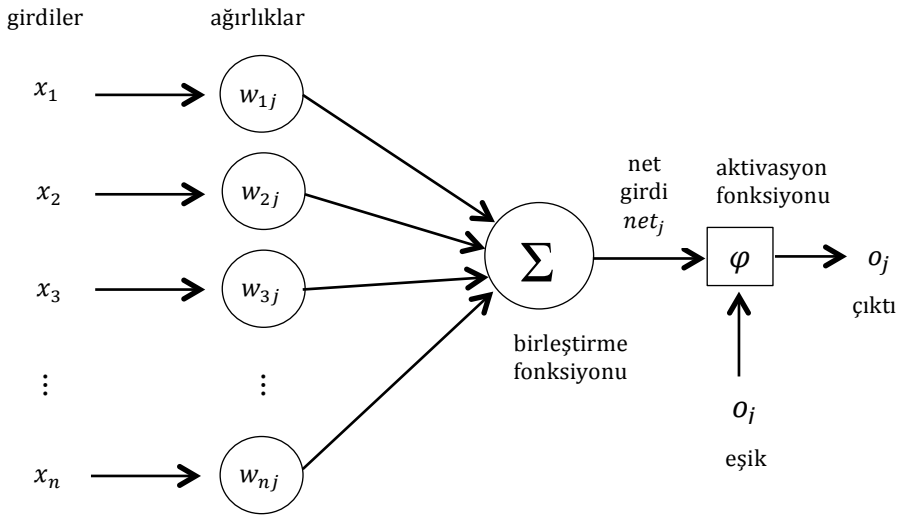
1.2.5. Yapay Sinir Ağları

İnsan beyninin temel işlem elemanı ve sinir sisteminin en basit elemanı olan nöron ve bu nöronlar arası bağlantılara şekilsel ve işlevsel olarak benzeyen bir yapay sinir ağı, bu haliyle adeta

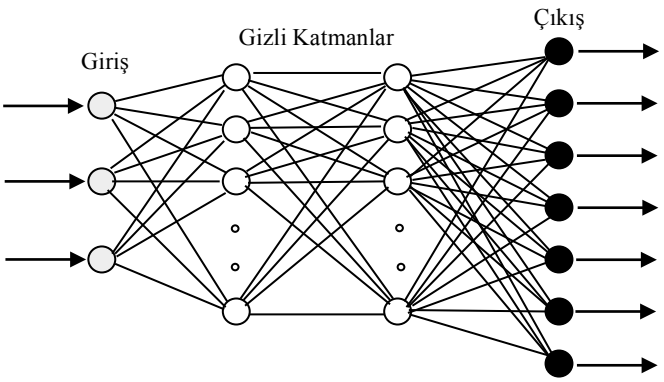
biyolojik sinir sisteminin basit bir simülasyonudur. Biyolojik sinir sisteminin matematiksel bir modeli olarak da tanımlanabilecek olan yapay sinir ağı, birbirleri ile bağlantılı yapay sinir hücrelerinin oluşturduğu bir sistem ile biyolojik sinir sisteminin bilgiyi depolama, kullanma ve işleme yeteneklerini taklit etmeyi ve insan gibi karar verebilen ve muhakeme yeteneği olan zeki sistemler elde etmeyi amaçlar.

Biyolojik sinir ağlarındaki sinir hücrelerine karşılık, yapay sinir ağında da yapay sinir hücreleri vardır. Her yapay sinir hücresinin temel elemanları; girdiler, ağırlıklar, toplama (birleştirme) fonksiyonu, transfer (aktivasyon) fonksiyonu ve hücrenin çıktısıdır (Şekil 5). Yapay sinir hücreleri katmanlar şeklinde birleşerek yapay sinir ağını oluştururlar. Yapay sinir ağında katmanlar girdi katmanı, çıktı katmanı ve bu ikisinin arasında bulunan gizli katman(lar)dır. Her bir katman nöronlardan oluşmaktadır (Şekil 6). Girdi ve çıktı katmanındaki nöron sayısı, bağımsız ve bağımlı değişkenlerinin sayısı ile belirlenmekte iken, gizli katmandaki katman sayısı ve her bir gizli katmanda bulunacak nöron sayıları, en iyi performansı verecek şekilde kullanıcı tarafından belirlenmektedir.

Şekil-5. Yapay Sinir Hücresinin Yapısı



Şekil-6. Çok Katmanlı Bir Yapay Sinir Ağının Genel Yapısı



Yapay sinir ağı, veri setindeki yapıyı öğrenerek, istenilen görevi yerine getirecek şekilde genelleştirmeler yapar. Bunun için ağ ilgili olayın örnekleri ile eğitilerek genelleme yapılabilecek

yeteneğe kavuşturulur ve benzer olaylara karşılık gelen çıktı setleri belirlenir. Ağa girilen bilgilerin kendi ağırlıkları ile çarpımlarının toplanması sonucu elde edilen net girdi bir transfer fonksiyonu ile işlenerek çıktı katmanından ağın ürettiği sonuç olarak alınmaktadır (Öztemel, 2003: 49). Ağdaki bilgi, en iyi sonucu verecek şekilde kullanıcı tarafından tespit edilen katmanlar ve bu katmanlardaki nöronlarda gizlidir. Ağ en iyi sonucu elde etmek için bu ağırlıkları güncellemektedir. Bu sebeple bu ağırlıkların anlamlandırılması veya yorumlanması söz konusu olamamakta olup bu durum yapay sinir ağlarının dezavantajı olarak bilinir. Sistem bilgisinin tüm ağa ağırlıklarla dağıtılmış olması nedeniyle, ağırlıkların anlamsal özelliklerini açıklamanın ve ağda bulunan daha önceki bilgileri birleştirmenin zor olması, ağın bir kara kutu olarak nitelendirilmesine sebep olmaktadır.

2. BÜYÜK VERİ ANALİZİNDE YAPAY ZEKÂ VE MAKİNE ÖĞRENMESİ TEKNİKLERİNİN KULLANIMI

Büyük veri kavramı ile sadece veri kümesinin olağanüstü boyutu değil, aynı zamanda yüksek veri oluşum hızı ve veri çeşitliliği de vurgulanmaktadır. Yeni bir çağın başlangıcı olarak yorumlanan büyük verinin kullanımı ile beraber bu verilerin depolanması, güvenlik ve mahremiyeti, işlenmesi ve analizi, veriye dayalı karar verme gibi problemler ortaya çıkmaktadır. Bu süreçler veri madenciliği, bilgisayar bilimi, makine öğrenmesi, veri tabanı yönetimi, matematiksel algoritmalar ve istatistiğin birlikte çalışmasını gerektirir.

Özellikle yeni teknolojilerin daha yaygın hale gelmesiyle, çok büyük miktarda veri çok hızlı bir şekilde üretilmekte ve sunucularda depolanmaktadır. Fiziki belleklerin bu hacimler arttıkça yeterli olamaması ise bulut bilişim teknolojilerinin devreye girmesini zorunlu hale getirmiştir. Öte yandan bu veriler, sağladığı avantajlar yanında güvenlik sorunlarını da beraberinde getirmektedir. Veri bilimciler ve bu alanda çalışmalar yapan diğer disiplinlerdeki araştırmacılar, sadece devasa değil aynı zamanda değişen ve çok hızlı biriken büyük veri ortamı için, şifreli ve mahremiyet korumalı veri tabanı yönetim sistemleri ve ürünleri üzerine çalışmaktadır.

Depolanan verilerin katma değere dönüşebilmesi, bu verilerin işlenerek analiz sonuçlarının elde edilmesi ile mümkündür. Günümüzde ağırlıklı olarak büyük veriler, sosyal medya, bloglar, e-postalar, sensör verileri, coğrafi işaretler, lokasyon verileri, fotoğraflar veya videolar gibi oldukça büyük boyutlarda ve çok çeşitli kaynaklardan derlenmektedir. Bu kaynakları doğru kullanabilen kurum, firma, şirket, şahıs veya devletler; müşterileri, takipçileri, taraftarları, rakipleri, düşmanları veya ortakları için daha fazla kestirimde bulunabilirler. Ayrıca böyle bir veri yığını içinden öngörülemeyen durumlara dair bilgiler de elde edilebilir. İşletmeler açısından bu durum hem gelirlerde artışı hem de iş kollarında zenginleşmeyi sağlayabilmektedir. Mesela ödeme işlemleri ile ilgili olarak MasterCard, 210 ülkedeki 1,5 milyar kart kullanıcısının yaptığı toplam 65 milyar alışveriş işlemini analiz ederek farklı tüketici alışkanlıklarını ortaya çıkarmayı başarmıştır. Bu analizlerin sonuçlarından biri de, ABD' de saat 16 civarında benzin istasyonuna gelen insanların devamındaki 1 saat içerisinde restoran veya süpermarketlerde 35 ila 50 dolar arası harcama yaptığını ortaya koymaktadır. Restoran ve marketler bu müşterilere alışveriş kuponları göndermişler ve satışlarını artırmışlardır (Schonberger ve Cukier, 2013: 127). Borsa yatırımcıları Twitter verilerini analiz ederek hisse senedi performanslarını tahmin

edebilmektedir. Amazon ve Netfix ise müşterilerinin etkileşimleri yardımı ile ürün önerilerini isabetli hale getirmektedir. Twitter, Facebook ve LinkedIn, kullanıcılarının sosyal ilişkilerini haritalandırarak ve grafikleyerek tercihlerine ilişkin öngörüler yapmaktadır (Schonberger ve Cukier, 2013: 14).

Ancak büyük verinin yukarıda da bahsedilen özellikleri nedeniyle, geleneksel veri yönetim sistemleri yerine daha zenginleştirilmiş yöntemlerle işlenmesi gerekmektedir. Çünkü yalnızca hacim olarak değil, çeşitlilik, hız, değişkenlik gibi özellikleri bakımından da büyük olan ve hem yapısal (önceden belirlenmiş bir formatta olan) hem de yarı yapısal ya da yapısal olmayan formda olan veriler her an üretilmeye devam etmektedir. Özellikle yapısal olmayan makine kaynaklı verilerin (uydu görüntüleri, bilimsel veriler, fotoğraf ve videolar, radar ve sonar verileri, sensör verileri vs.) ve insan kaynaklı verilerin (sosyal medya verileri, mobil cihazlardan elde edilen veriler, web sitelerinin içerik verileri vs.) üretilen tüm veriler içindeki oranı %80' ler (Ohlhorst, 2013: 87) civarındadır ve bu oran sürekli artmaktadır. Büyük ölçekteki verilerin saklanmasında fiziki belleklerin yeterli olmadığı durumlarda bulut bellekler devreye girmektedir. Bu verilerin işlenmesi, üzerinde çalışılması ve analizinin yapılabilmesi için yüksek hesaplama gücüne ihtiyaç duyulmaktadır. Bu sebeple artık geleneksel hesaplama yaklaşımları yerine bilgisayar kümeleri ve HDFS-Hadoop Distributed File System (Hadoop Dağıtılmış Dosya Sistmi) RDD-Resilient Distributed Datasets (Esnek Dağıtılmış Veri Setleri) gibi dağıtık dosya sistemleri; geleneksel programlar ve programlama dilleri yerine ise Hadoop, Spark, MapReduce, Pig ve Hive gibi açık kodlu yazılım çerçevesi olan platformlar yaygınlaşmaktadır (Hallaç, 2014: IV). Verinin saklanmasında ve işlenmesinde kullanılmaya başlanan bu teknolojiler verilerin analizinde kullanılan yöntemleri de etkilemiştir. Büyük veri öncesinde de istatiksel analiz tekniği olarak kullanılmakta olan yöntem ve teknikler, büyük verinin analizi için kullanılmakta ve süreç içinde büyük veri setlerine de uygulanabilir özelliklerle donatılmış olup yapay zeka ve makine öğrenmesi teknikleri de benzer şekilde buna dahildir. Büyük veri analizi ile veri madenciliği süreçleri aynı olup, birçok ortaklıkları yanında küçük farklılıkları da vardır. Ayrıca son zamanlarda büyük veri analizi için bazı teknikler geliştirilmiştir.

Büyük veri analizi uygulamaları; veri bilimcileri, öngörü modelleyiciler, istatistikçiler ve diğer analiz uzmanlarının yapılandırılmış işlem verisinin yanı sıra, yarı yapılandırılmış ve yapılandırılmamış verileri analiz etmesini sağlar. İnternet tıklama bilgisi verileri, web sunucusu günlükleri, sosyal medya içeriği, müşteri e-postalarından ve anket yanıtlarından gelen metinler, cep telefonu çağrısı detay kayıtları ve nesnelerin internetine bağlı sensörler tarafından yakalanan makine verileri bunlara örnektir. Bununla birlikte, büyük veri analizi kullanıcıları genel olarak, gelen ham veri akışları için birincil depo görevi gören bir bir Hadoop kümesinde analiz yapabilir veya Spark gibi bir işleme motoruyla çalıştırılabilirler. Veri ambarcılığında olduğu gibi, sağlam veri yönetimi büyük veri analizi sürecinde çok önemli bir ilk adımdır. Hadoop Dağıtılmış Dosya Sisteminde depolanan veriler hem ayıklanması, dönüşümü ve yüklenmesi işlerinde hem de analitik sorgularda iyi performans elde etmek için düzenlenmeli, yapılandırılmalı ve bölünmelidir. Veriler hazır olduğunda, gelişmiş analitik süreçlerde yaygın olarak kullanılan yazılımlarla analiz edilebilir. Bu yazılımlar; veri setleri arasında modeller ve ilişkiler arayan veri madenciliği, müşteri davranışını ve gelecekteki diğer gelişmeleri tahmin etmek için modeller oluşturan tahmin yöntemleri, büyük veri setlerini analiz etmek için algoritmalar kullanan makine öğrenmesi ve daha gelişmiş bir makine öğrenmesi dalı olarak derin öğrenme araçları olabilecektir. Metin

madenciliği yöntemleri ve istatistiksel analizler de hem büyük veri analizi sürecinde hem veri görselleştirme aracı olarak kullanılabilir (Rouse, 2017).

Bu bölümün devamında, veri madenciliği ve büyük veri analizinde makine öğrenmesi ve yapay zeka yöntem ve tekniklerinin kullanımı; kümeleme, sınıflandırma, yapay sinir ağları, metin ve web madenciliği, fikir madenciliği ve duygu analizi başlıkları altında gruplandırılarak, bunlara dair literatürde bulunan bazı uygulamalardan bahsedilip örnekler verilecektir.

2.1. Kümeleme

Kümeleme analizlerinde nesnelerin önceden belirlenmiş bir kritere göre gruplandırması yapılmakta olup bu sebeple denetimsiz öğrenme algoritmalarıdır. Makine öğrenme ile daha da önem kazanan kümeleme algoritmaları örüntü tanıma, konuşma tanıma, görüntü ve ses işleme, işletmelerde müşterilerin tercihlerine, coğrafi durumlarına ve demografik yapılarına göre, satın alma davranışları gibi çeşitli özelliklerine göre gruplara ayrılması, sosyal ağ analizleri, anahtar kelime aramaları, taranan anahtar kelimelerin ilişkili olduğu kavramlara göre sıralanması, trend topiklerin ortaya çıkarılması (Bayrakçı, 2015: 94-96), satış hareketleri, çağrı merkezi kayıtları (Işık, 2006: 7), ülkelerin gruplara ayrılması (Sarıman, 2011: 192) gibi konularda uygulanmaktadır. Kümeleme algoritmaları, veri madenciliği tekniklerinde de olduğu gibi, büyük veri analizinde benzer nitelikteki grupların ve alt grupların belirlenmesinde veya farklılıklarının ortaya çıkarılmasında kullanılmaktadır. Kullanılan kümeleme algoritmalarına örnek olarak, kMeans, Bulanık C-ortalamalar, Kohonen Yapay Sinir Ağları, k-medoids, Canopy, Mean Shift, MinHash, Latent Dirichlet Allocation (Peng, 2012; Özekes, 2003: 73) sayılabilir. Bu algoritmalar, ölçeklenebilirlik ve hızları artırılarak daha büyük veri kümeleri ile çalışmak üzere güçlendirilerek kullanılmaktadır. Genel olarak büyük veri kümeleme teknikleri; tek makine kümeleme teknikleri ve çoklu makine kümeleme teknikleri olmak üzere iki ana kategoriye ayrılabilir. Çoklu makine kümeleme teknikleri, ölçeklenebilirlik açısından esnek olmaları ve kullanıcılara daha hızlı yanıt vermeleri nedeniyle dikkat çekmektedir. Tek makine kümeleme teknikleri, örneklem temelli teknikler ve boyut azaltma teknikleri; çoklu makine kümeleme teknikleri ise paralel kümeleme ve MapReduce tabanlı kümeleme teknikleridir. Paralel kümeleme algoritmalarında yaşanan bellek ve işlemci dağıtımı ile ilgili karmaşıklığa alternatif olarak önemli kolaylık sağlayan MapReduce, başlangıçta Google tarafından sunulmuş olup, açık kaynak kodlu bir kütüphane olan Hadoop' da verinin işlenmesi sürecidir (Shirkhorshidi vd., 2014). Büyük veri kümeleme teknikleri için bir diğer sınıflandırma ise; bölümleme tabanlı teknikler (kMeans, K-modes, PAM, CLARA, CLARANS ve FCM gibi), hiyerarşik tabanlı teknikler (BIRCH, CURE, ROCK ve Chameleon gibi), yoğunluğa dayalı teknikler (DBSCAN, OPTICS, DBCLASD ve DENCLUE gibi), ızgara tabanlı teknikler (Wave-Cluster ve STING gibi) ve model tabanlı teknikler (MCLUST, EM karma yoğunluk modeli, COBWEB kavramsal kümeleme, kendini düzenleyen özellik haritaları gibi sinir ağı yaklaşımları gibi) şeklindedir (Fahad vd., 2014).

2.2. Sınıflandırma

Sınıflandırma, bir birimin sahip olduğu özelliklerine göre hangi gruba ait olduğunu belirlemektir. Denetimli öğrenme algoritmaları olan sınıflandırma algoritmalarında, var olan verilerden örüntü keşfedilir ve yeni eklenecek nesnelerin hangi sınıfta yer alacağı tahmin edilir. Ayrıca bu yöntemlerin geliştirilmesinde bulanık mantık da kullanılmaktadır (Bayrakçı, 2015: 98). Bu algoritmalardan başlıcaları;

lineer diskriminant analizi, karar ağaçları, yapay sinir ağları, destek vektör makineleri, lojistik regresyon, kNN (k-en yakın komşu), genetik algoritmalar, bellek temelli nedenleme ve naive Bayes algoritmasıdır. Genetik algoritmalar, sınıflandırmalarda kural tabanlı çalışmalarda kullanılmakta olup (Karr ve Freeman, 1999: 304) metin sınıflandırma, yüz tanıma, çağrı yönlendirme gibi uygulamalar bu algoritmalarla yapılabilmektedir. İnsan hallerinin algılanması sayesinde, buna uygun tepkinin verilmesinin sağlanması ve bunun otomasyona tabi tutulması için geliştirilen teknolojileri kapsayan ve ses tanıma, yüz algılama ve konuşmanın hesaplanması gibi uygulamalarla ön plana çıkan (Aksu, 2015: 117) duygusal bilişim (affective computing), sınıflandırma tekniklerinin kullanıldığı bir alan olarak ön plana çıkmaktadır. Yapay zeka tekniklerinden yapay sinir ağları ve destek vektör makineleri ile k-en yakın komşu algoritması ve C4.5 karar ağacı algoritması gibi makine öğrenmesi yöntemleri sınıflandırma için bu teknolojide yaygın olarak kullanılmaktadır (Wikipedia, 2017; ayrıca bu konudaki yayınlar için Bkz. Affective Computing Group: MIT Media Lab, http://affect.media.mit.edu/publications.php). Büyük verilerin sınıflandırılmasına biyotıp, sosyal medya, pazarlama vb. gibi çok çeşitli alanlarda ihtiyaç duyulmaktadır. Yokoyama vd. (2012), Zhang vd. (2012) ve Maillo vd. (2015)' nin çalışmalarındaki gibi, büyük veriyi büyük kümelerde işlemek üzere geliştirilmiş bir dağıtık programlama modeli olan MapReduce tabanlı uygulamalar oldukça yaygındır. Apache Spark çerçevesinde Scala programlama dili ile geliştirilen algoritma ile bir kamu üniversitesinin enerji tüketimi, çeşitli binalarda bulunan bir sensör ağından toplanan üzerine büyük veri setleri ile analiz edilmiştir (MLib, 2017). Naive Bayes sınıflandırıcı, metin sınıflandırma problemlerinde, Apache Spark MLib gibi makine öğrenmesi kütüphanelerinde kullanılmaktadır (Sakınmaz, 2017). Sosyal medyada gündemin sıcak konularının analiz edilerek kategorize edilmesine dair uygulamalar yapılmaktadır (Mishchenko, 2017). Yine milyonlarla ifade edilen sayıda film incelemesini sınıflandırmak (Liu vd., 2013), fikir madenciliği (El-Halees, 2012; Yang ve Ko, 2009) ve ölçeklenebilir duygu analizi ve sınıflandırması (Liu vd., 2013) için Naive Bayes sınıflandırıcı etkili bir araç olmaktadır. Apache Spark Streaming teknolojisi üzerine, Destek Vektör Makineleri sınıflandırma yöntemi geliştirilerek Lojistik Regresyon yöntemi ile karşılaştırılmış; Destek Vektör Makineleri yönteminin kullanılan veri kümeleri üzerinde daha başarılı sonuçlandığı gözlemlenmiştir (Akgün, 2016). Serbestçe kelimelere dökülmüş metinden üretilen yüksek boyutlu öznitelik vektörlerinin çevrimiçi işlenmesine uygun son derece etkin boyut azaltıcı tekniklerin tanıtıldığı çalışmada ise (Yar vd., 2016), tweetlerin çok sınıflı sınıflandırması incelenmiş ve tweetin ait oldugu kategoriyi belirleme işlemi olarak Destek Vektör Makineleri, K En Yakın Komşu, Karar Agaçları ve Lojistk Regresyon yöntemleri incelenmiştir.

2.3. Yapay Sinir Ağları

Yapay sinir ağlarının geleneksel analiz yöntemlerinden farkı; paralel işlem yapabilmesi yani aynı görev üzerine aynı anda birbirinden bağımsız hesaplama kaynaklarının çalışmasıdır. Yapay sinir ağ modelleri vasıtasıyla veri birbirinden bağımsız işlemcilere ayrıştırılır ve her bir işlemci bağımsız çalışır. Büyük veri analizinde kullanılan paralel işleme modellerinden en yaygın olanları MPI (Message Passing Interface), MapReduce ve Dryad modelleridir. Yapay sinir ağ uygulamaları yüz tanıma, kredi kararlarının verilmesi, el yazısı tanıma, işletmelerin finansal durumlarının derecelendirilmesi ve dolandırıcılık tespiti gibi farklı alanlarda etkin olarak kullanılmaktadır (Bayrakçı, 2015: 99-101). Yapay ve derin öğrenme

(deep learning) sinir ağları yöntemleri, görüntü tanıma, doğal dil işleme, tercüme, otomatik ses tanımaya yönelik geliştirilecek uygulamalar ile güncel hayatın içerisine girmiştir. Yapay sinir ağları esasına dayalı olarak, Microsoft ses tanıma sistemini geliştirmek, Facebook ise fotoğraf ve videolardaki yüz ve nesneleri tanımak ve reklamları doğru adreslemek üzere derin öğrenme tekniklerini kullanmaktadır (Gürsakal, 2014: 202). Google' ın Android telefonlar üzerinde ses komutlarını ve Google+ sosyal ağı üzerindeki görüntü etiketlerini tanımak, Google gözlükleri üzerinde ses ve görüntü algılamaya yönelik yaptığı çalışmalar da buna örnektir (Dal, 2014). Diğer taraftan lokasyon bazlı analizler de GPS sinyalleri (Arslan vd., 2007), haritalar, sosyal ağlar üzerinde yer bildirimleri ve trafik akışı verileri gibi yapısal olmayan ve hızla değişen büyük veriler üzerinde çalışmaktadır. Bu analizlerle, trafik akışının optimizasyonu, haritaların dijital ortamda işlenmesi, arkadaş-firma bulma, mekan bildirimi ile işletmelere viral pazarlama sağlanması gibi güncel gelişmeler ortaya çıkmaktadır.

2.4. Metin ve Web Madenciliği

Yapısal veri kavramı, bir tablodaki satır ve sütunlarla veri tabanlarında saklanabilecek verileri ifade eder. Yarı yapılandırılmış veriler, ilişkisel bir veri tabanında yer almayan, ancak analiz etmeyi kolaylaştıran bazı organizasyonel özelliklere sahip verilerdir. Bunların dışında kalan veriler ise yapısal olmayan verilerdir ve tanımlanabilir bir yapıları yoktur. Yapılandırılmamış veriler, verilerin yaklaşık% 80' ini temsil etmektedir. En çok bilinen yapısal olmayan veri türleri; uydu görüntüleri, sismik görüntüler, atmosferik veriler gibi bilimsel veriler, fotoğraf ve videolar, radar veya sonar verileri, pdf, word, text vb. formattaki belgeler, web üzerinde tutulan log dosyaları, anket sonuçları ve e-postalar gibi metinler, sosyal medya verileri, mobil veriler ve web sitesi içerikleridir (Ronk, 2014). Metin madenciliği, çok büyük belgelerin analizi, kavramlar arası ilişkilerin bulunması, anlamlı bilgilerin ortaya çıkarılması ve metin tabanlı verinin içerisindeki gizli kalıpların elde edilmesidir. Web madenciliği ise, web içerikleri, sayfa yapıları ve web bağlantı istatistiklerinin de içinde olduğu web ile ilişkili olan verinin analizini içermektedir (Tan ve Yu, 2003: 239). Çalışmalarda metin madenciliğinde çoğunlukla kullanılan algoritmalar; Naive Bayes Algoritması, Rocchio Algoritması, Karar Ağaçları, k-En Yakın Komşu Algoritması, Destek Vektör Makinesi ve K Ortalama Algoritmasıdır. Doğal dil işlemi uygulamalarına, arama motorlarında kullanıcıların hatalı yazmış olduğu sözcüklerin bulunması ve doğrusunun önerilmesi, dilden dile yapılan çeviriler en yaygın ve güncel örneklerdir (Bayrakçı, 2015: 103). Bunların dışında; müşteri ilişkileri yönetimi (müşterilerin email, işlem, çağrı merkezi ve anket gibi erişim noktalarından elde edilen metin bilgilerinden nitelikli bilgi çıkarılması), sahtekarlık tespiti (büyük çaptaki metin verilerinde kalıplar ve anormallikler), bilimsel ve medikal araştırmalar (hasta raporları, makale başlıkları, yayınlanmış araştırma sonuçları), güvenlik ve istihbarat (büyük çaptaki metin içerisinde organizasyonlar ve bireyler arasındaki bağlantıları, terörist tehlikeleri ve kriminal davranışların tahmini), pazar araştırması (pazar etkisinin ölçülmesi için basın bültenleri ve web sayfalarının izlenmesi) (Dolgun vd., 2009: 51) metin madenciliği uygulamalarıdır. Yine kurumsal finans uygulamaları, patent analizi, internetten piyasa istihbaratı, dijital kütüphanelerde doküman eşleştirme (Oğuzlar, 2011: 76-95) gibi problemlere çözümler üretilebilmektedir. Bir e-ticaret sitesi olan Amazon'a bir ürün ile ilgili bırakılan yüzlerce müşteri yorumu metin madenciliği ile işlenip, ilgili ürün için özet bir tavsiye üretilebilmektedir (Dal, 2014). Bu analizler

özel yazılımlarla yapılabildiği gibi, açık kaynak kodlu bir yazılım olan R ve R kütüphanesinde bulunan "tm" gibi paketler (Feinerer, 2017) ile de yapılabilmektedir.

Web madenciliğine ise; internet üzerinden yapılan satış verilerinin analiz edilerek müşteri profili ve kümelerinin oluşturulması, arama motorlarında aranan anahtar kelimeyi içeren web sitelerinin belirlenmesi, web sitelerinin kullanıcıların geri dönüşlerine göre düzenlenmesi (Dolgun vd., 2009: 52), sosyal medya profil verilerine göre reklam spesifikasyonu, tıklanma sayılarına göre reklam ücretlerinin belirlenmesi, lokasyon bilgilerine göre mesaj gönderilerek müşteri kazanma gibi örnekler verilebilir. Takip edilebilir ve mobil cihazların internet ve GPS bağlantıları ile hem birbirleri ile bağlantı kurması hem de sosyal ağlarla birlikte lokasyon bazlı servisler ile işlem yapması; mekân bildirimleri, arkadaş arama, mekân deneyimlerinin paylaşımı, askeri, emniyet ve istihbarat faaliyetleri, trafik akışının optimizasyonu (Aksu, 2015: 205-210), coğrafi bilgi sistemleri teknolojilerinin sosyal, fiziksel, duygusal ve coğrafi göstergeler bir arada kullanılacak şekilde genişletilerek kullanıldığı yer analitiği çözümlemeleri (Thompson, 2012) ve dahası pek çok uygulama web madenciliğinin sonuçlarıdır.

2.5. Fikir Madenciliği ve Duygu Analizi

Sosyal ağlardaki gelişmelere paralel olarak bu ağlardan elde edilen verilerin analizi ve yorumlanması konusunda yapılanlar yine metin ve web madenciliği uygulamaları olarak nitelenebilir. Ancak bu analizler özellikle kişisel hesaplarla yapıldığında, hesap sahibinin günlük faaliyetleri, düşünceleri, görüşleri, anlık duygu durumları ve ruh halini de içerdiğinden web ve metin madenciliğinin bir başka hali olan fikir madenciliği ve duygu analizine evrilecektir. Metin madenciliğinin bir uygulama alanı olarak ortaya çıkmış bir kavram olan fikir madenciliği, verilen bir konu üzerine görüş sahibinin fikrinin sınıflandırılması veya tanımlanması için istatistiksel model ve yazılımların kullanılması (Falcon, 2010) olup literatürde aynı zamanda duygu analizi/duygu madenciliği olarak da geçmektedir. Şahısların ürün, servis, kurum, olaylar ve diğer şahıslar hakkındaki duygu, fikir, görüş, yorum ve davranışlarını analiz eden (Liu, 2012: 7) fikir madenciliği, yapılandırılmamış metinlerden bilginin ve içerdiği fikrin çıkarılmasını hedefler. Yapılan bazı uygulamalara; forum, blog ve haber sitelerindeki yorumların içinde geçen karşılaştırma cümlelerinin tespiti ve bu cümlelerdeki karşılaştırma ilişkilerinin sınıflandırılması (Jindal ve Liu, 2006: 1335; Bos ve Nissim, 2006: 11; Yang ve Ko, 2009: 155; El-Halees, 2012: 268), Twitter mesajlarının sınıflandırılması (Go vd., 2009; Meral ve Diri, 2014), biyomedikal literatüründeki atıf ve özetlerin karşılaştırılması (Fiszman vd., 2007: 139), karşılaştırma cümlelerinde hangi ürün veya hizmetin diğerine göre tercih edildiğinin bulunması (Ganapathibhotla ve Liu, 2008: 243), sosyal medyada rekabet analizi (Mayda ve Aytekin, 2013: 418), sosyal medya metinlerinden duygu analizi yapılarak duygusal duruma uygun reklam gönderimi ve seçim dönemlerinde yazılan sosyal medya içerikleri ile sonuçların tahmini örnektir. Kullanılan başlıca yöntemler ise, naive Bayes, maksimum entropi, destek vektör makinesi ve K-en yakın komşu algoritmasıdır.

3. SONUÇ

Teknoloji ve yaşamın son yıllarda geldiği noktada, üretilen veriler de çağın gelişmişlik düzeyi ile birlikte katlanarak büyümektedir. Klasik yöntemlerle keşfedilemeyecek ilişkilerin keşfedilebilmesi, büyük veri analizinin parlak yönü olarak tebarüz etmektedir. Büyük veri, işletmelere yapay sinir ağları,

derin öğrenme, doğal dil işleme, görüntü tanıma ve ileriye yönelik kişiselleştirme teknolojileri ile işlem görerek çok daha fazla akıl ve öngörü verebilmektedir. Öte yandan getirdikleri yaklaşımlarla doğadaki canlıların akıllı davranışlarını taklit eden, insan gibi düşünen ve karar veren modeller oluşturmayı amaçlayan yapay zeka teknikleri, büyük verilerin üzerinde yapılan çalışmalarda da sağladığı avantajlar ile tercih edilmektedir.

Bugün artık Twitter' da tweetlerin analizi, Google' da arama yaparken olası sonuçların tahmini, Facebook' ta beğenilen sayfa, içerik veya etkileşime geçilen arkadaşların incelenerek benzer konuların önerilmesi, Apple'ın Siri ve Google'ın Google Now gibi yazılımları gibi bilgisayar ve sosyal medya analizleri yapay zeka tekniklerinin de kullanıldığı büyük veri analizleri ile yapılmaktadır. Yine tüketici tercihlerini ve rasyonel tüketiciyi analiz ederek satış optimizasyonu sağlanmakta, şirket içi ve dışı güvenlik aşamasında bireyleri tanıma ve dolandırıcı tespitinde yapay zeka ile geçmiş tecrübeleri hızlı analiz edilebildiği için tercih edilmektedir (Serim, 2015). Bir elektronik ticaret müşterisinin, Pinterest arayüzünden yüklediği resme benzer ürünleri sorgulayıp, kendi profil ve tercihlerine uygun hedefe yönelik ürün önerisi alabilmesi, popüler ses tanıma uygulaması olan Shazam' ın, sesleri dinleyerek aranan müzik parçalarını cihaza getirebilmesi, Amazon' un önceki siparişler, ürün aramaları, istek listeleri, alışveriş kartı içerikleri, iadeler ve diğer online alışveriş verilerini değerlendirerek, daha müşteri sipariş vermeden paketleme yapıp göndermesi (Dal, 2014) gibi daha nice gelişmeler farklı disiplinlerin yapay zeka ile birlikte neler yapabileceğine dair pek çok örnekten bir kaçıdır. Tüm bu gelişmeler, mobil iletişim, bulut teknolojileri ve robot teknolojisi ile birlikte gelecekte yapay zekanın çok daha önemli olacağına dair açık işaretler vermektedir.

KAYNAKÇA

Akgün, B. (2016). *Apache Spark Based Distributed Svm Algorithm For Stream Data Classification*. Yayınlanmamış Yüksek Lisans Tezi, İstanbul Teknik Üniversitesi.

Aksu, H. (2015). *Big Data Bilginin Gücü.* İstanbul: Pusula Yayıncılık.

Arslan, O., Kurt, O. ve Konak, H. (2007). Yapay Sinir Ağlarının Jeodezide Kullanımı Üzerine Öneriler. *TMMOB Harita ve Kadastro Mühendisleri Odası 11. Türkiye Harita Bilimsel ve Teknik Kurultayı*, Ankara.

Aytekin, Ç. ve Mayda, İ. (2013). Sosyal Medyada Rekabet Analizi İçin Karşılaştırma Görevine Yönelik Fikir Madenciliği Modeli. *Journal Academic Marketing Mysticism Online (JAMMO)*, Vol 7, Part 27, 414-425.

Balaban, M. E. ve Kartal, E. (2015). *Veri Madenciliği ve Makine Öğrenmesi*. İstanbul: Çağlayan Kitabevi.

Baykal N. ve Beyan T. (2004). *Bulanık Mantık Uzman Sistemler ve Denetleyiciler*. Ankara: Bıçaklar Kitabevi.

Bayrakçı, S. (2015). *Sosyal Bilimlerdeki Akademik Çalışmalarda Büyük Veri Kullanımı.* Yayınlanmamış Yüksek Lisans Tezi, Marmara Üniversitesi Sosyal Bilimler Enstitüsü.

Bos, J. ve Nissim, M. (2006). An Empirical Approach to the Interpretation of Superlatives. *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP 2006)*, pages 9–17, Sydney.

Charniak, E. ve McDermott, D. (1985). *Introduction to Artificial Intelligence.* Boston, MA, USA: Addison-Wesley Series in Computer Science.

Dal, B. (2014). Elektronik ticarette büyük veri ve yapay zekâ. Erişim 11.12.2015, http://www.retailturkiye.com/bulent-dal/elektronik-ticarette-buyuk-veri-ve-yapay-zeka

Dolgun, M. Ö., Özdemir Güzel, T. ve Oğuz, D. (2009). Veri madenciliğinde yapısal olmayan verinin analizi: Metin ve web madenciliği. *İstatistikçiler Dergisi, 2*(2009), 48-58.

Domo (2015). Data Never Sleeps 3.0. Erişim 17.10.2016, https://www.domo.com/blog/2015/08/data-never-sleeps-3-0/

Domo (2016). Data Never Sleeps 4.0. Erişim 17.10.2016, https://www.domo.com/blog/data-never-sleeps-4-0/

El-Halees, A. (2012). Opinion Mining From Arabic Comparative Sentences. *The 13th International Arab Conference on Information Technology ACIT'2012*, 265-271, Lebanon.

Elmas, Ç. (2011). *Yapay Zeka uygulamaları*. Ankara: Seçkin Yayıncılık.

Fahad, A., Alshatıı, N., Tari, Z., Alamri, A., Khalil, I., Zomaya, A. Y., Foufou, S. ve Bouras, A. (2014). A Survey of Clustering Algorithms for Big Data: Taxonomy and Empirical Analysis. *IEEE Transactions On Emerging Topics in Computing, 2*(3), 267-279.

Feinerer, I. (2017). Introduction to the tm Package Text Mining in R. Erişim 24.04.2017, https://cran.r-project.org/web/packages/tm/vignettes/tm.pdf

Fiszman, M., Demner Fushman, D., Lang, F. M., Goetz, P. ve Rindflesch, T. (2007). Interpreting Comparative Constructions in Biomedical Text. *BioNLP 2007: Biological, translational, and clinical language processing*, pages 137–144, Prague.

Firican, G. (2017). The 10 Vs of Big Data. Erişim: 24.10.2017, https://tdwi.org/articles/2017/02/08/10-vs-of-big-data.aspx

Ganapathibhotla, M. ve Liu, B. (2008). Mining Opinions in Comparative Sentences. *22nd International Conference on Computational Linguistics,* 241–248, Manchester.

Go, A., Bhayani, R. ve Huang, L. (2009). Twitter Sentiment Classification using Distant Supervision. *CS224N Project Report,* pages 1–12, Stanford.

Gürsakal, N. (2014). *Büyük Veri.* Bursa: Dora Basım Yayın.

Hallaç, İ. R. (2014). *Büyük Veri Analizinde Dağıtık Makine Öğrenmesi Algoritmalarının Kullanılması*. Yayınlanmamış Yüksek Lisans Tezi, Fırat Üniversitesi.

Harmon, P. ve King, D. (1985). *Expert Systems: Artificial Intelligence in Business*. New York: John Wiley and Sons.

Işık, M. (2006). *Bölünmeli Kümeleme Yöntemleri İle Veri Madenciliği Uygulamaları*. Yayınlanmamış Yüksek Lisans Tezi, Marmara Üniversitesi Fen Bilimleri Enstitüsü.

Opinion Mining in eDiscovery. Erişim: 29.12.2015, http://jadefalconit.com/opinion-mining/opinion-mining-in-ediscovery

Jindal, N. ve Liu, B. (2006). Mining Comparative Sentences and Relations. *21st National Conference on Artificial Intelligence*, 2, 1331-1336, Boston.

Karaman, D., Gözüaçık, N., Alagöz, M. O., İlhan, H., Çağal, U. ve Yavuz, O. (2015). Managing 6LoWPAN Sensors with CoAP on Internet. *23st Signal Processing and Communications Applications Conference (SIU), IEEE*, DOI: 10.1109/SIU.2015.7130101, Malatya, Turkey.

Karr, C. ve Freeman, L. M. (1999). *Industrial Applications of Genetic Algorithms*. Boca Raton: CRC Press.

Khan, M. A., Uddin, M. F. ve Gupta, N. (2014). Seven V's of Big Data understanding Big Data to extract value. *Conference of the American Society for Engineering Education, IEEE,* DOI: 10.1109/ASEEZone1.2014.6820689, Bridgeport, CT, USA.

Liu, B. (2012). *Sentiment Analysis and Opinion Mining***.** Williston: Morgan&Claypool Publishers.

Liu, B., Blasch, E., Chen, Y., Shen, D. ve Chen, G. (2013). Scalable Sentiment Classification for Big Data Analysis Using Naive Bayes Classifier. *IEEE International Conference on Big Data*, Santa Clara, CA, USA.

Machine Learning Library (MLlib) for Spark. Erişim: 07.11.2017, http://spark.apache.org/docs/latest/mllib-guide.html

Maillo, J., Triguero, I. ve Herrera, F. (2015). A MapReduce-based k-Nearest Neighbor Approach for Big Data Classification. *2015 IEEE International Conference on Trustcom/BigDataSE/ISPA*, Helsinki, Finland.

Mayer-Schonberger, V. ve Cukier, K. (2013). *Big Data: A Revolution That Will Transform How We Live, Work and Think*. Boston, Massachusetts: Houghton Mifflin Harcourt.

Meral, M. ve Diri, B. (2014). Sentiment Analysis on Twitter. *IEEE 22nd Signal Processing and Communications Applications Conference (SIU 2014),* page 690-693.

Mishchenko, Y. (2016). Büyük Veri Uygulamaları. Erişim: 07.11.2017, http://yumishch.me/courses/ETM562-lecture-1.pdf

Nabiyev, V. V. (2012). *Yapay Zekâ*. Ankara: Seçkin Yayıncılık.

Oğuzlar, A. (2011). *Temel Metin Madenciliği.* Bursa: Dora Basım Yayın.

Ohlhorst, F. (2013). *Big Data Analytics : Turning Big Data into Big Money*. New Jersey: Wiley Publicity.

Özekes, S. (2003). Veri Madenciliği Modelleri ve Uygulama Alanları. *İstanbul Ticaret Üniversitesi Dergisi*, Cilt 2, Sayı 3, 65-82.

Öztemel, E. (2003). *Yapay Sinir Ağları*. İstanbul: Papatya Yayıncılık.

Peng, B. (2012). Apache Mahout Algorithms. Erişim: 29.12.2015, http://lifeeditor.blogspot.com.tr/2012/11/apache-mahout-algorithms-apache-mahout.html

Ronk, J. (2014). Structured, Semi Structured and Unstructured Data. Erişim: 07.11.2017, https://jeremyronk.wordpress.com/2014/09/01/structured-semi-structured-and-unstructured-data/

Rouse, M. (2017). Big Data Analytics. Erişim: 07.11.2017, http://searchbusinessanalytics.techtarget.com/definition/big-data-analytics

Sakınmaz, S. (2017). Apache Spark – MLib ile Text Sınıflandırma (Naive Bayes). Erişim: 07.11.2017, http://www.buyukveri.co/spark-machine-learning-text-siniflandirma-naive-bayes/#more-1314

Sarıman, G. (2011). Veri Madenciliğinde Kümeleme Teknikleri Üzerine Bir Çalışma: K-Means ve K-Medoids Kümeleme Algoritmalarının Karşılaştırılması. *Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Dergisi*, *15*(3), 192-202.

Serim, M. (2015). Yapay Zeka ve Büyük Verinin Sektörlerdeki Kullanımı. Erişim: 11.12.2015, http://bigumigu.com/haber/yapay-zeka-ve-buyuk-verinin-sektorlerdeki-kullanimi/

Shirkhorshidi, A. S., Aghabozorgi, S., Wah, T. Y. ve Herawan, T. (2014). Big Data Clustering: A Review. B. Murgante vd. (Ed.) *Computational Science and Its Applications – ICCSA 2014*, *Lecture Notes in Computer Science* içinde, Switzerland: Springer International Publishing.

Şimşek Gürsoy, U. T. (2017). *Veri Madenciliğinde Güncel Yaklaşımlar*. İstanbul: Çağlayan Kitabevi.

Tan, A. H. ve Yu, P. S. (2003). Guest Editorial: Text and Web Mining. *Applied Intelligence*, Vol. 18, 239-241.

Thompson, S. (2012). Location Analytics: Bringing Geography Back. Erişim: 24.04.2017, http://sloanreview.mit.edu/article/location-analytics-bringing-geography-back/

Turner, V. ve Gantz, J. F. (2014). The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things. Erişim: 15.10.2016, http://www.emc.com/leadership/digital-universe/index.htm

Wikipedia (2017). Affective computing. Erişim: 03.04.2017, https://en.wikipedia.org/wiki/Affective_computing#Algorithms

Yang, M.-S. ve Liu, H.-H. (2003). Fuzzy least-squares algorithms for interactive fuzzy linear regression models. *Elsevier Science Fuzzy Sts and Systems*, 135, 305-316.

Yang, S. ve Ko, Y. (2009). Extracting Comparative Sentences from Korean Text Documents Using Comparative Lexical Patterns and Machine Learning Techniques. *ACL-IJCNLP 2009 Conference* Short Papers, 153-156, Singapore.

Yar, E., Delibalta, İ., Baruh, L. ve Kozat, S. S. (2016). Online Text Classification for Real Life Tweet Analysis. *24th Signal Processing and Communication Application Conference (SIU)*, Zonguldak, Turkey.

Yenilmez, K. (2001). *Bulanık Doğrusal Programlama Problemleri için Yeni Çözüm Yaklaşımları ve Duyarlılık Analizi*. Yayınlanmamış Doktora Tezi, Osmangazi Üniversitesi.

Yıldırım, P., Uludağ, M. ve Görür, A. (2008). Hastane Bilgi Sistemlerinde Veri Madenciliği. *Akademik Bilişim 2008 Konferansı*, 429-434, Çanakkale.

Yokoyama, T., Ishikawa, Y. ve Suzuki, Y. (2012). Processing all k-nearest neighbor queries in hadoop, Web-Age Information Management. H. Gao, L. Lim, W. Wang, C. Li, ve L. Chen (Ed.), *Lecture Notes in Computer Science* içinde, vol. 7418, pp. 346–351. Berlin: Springer Heidelberg.

Zhang, C. ve Li, F. ve Jestes, J. (2012). Efficient parallel knn joins for large data in mapreduce. *Proceedings of the 15th International Conference on Extending Database Technology*, *ser. EDBT '12*, New York, NY, USA: ACM, pp. 38–49.