DATA SCIENCE AND ITS APPLICATIONS

Edited by
Aakanksha Sharaff and G. R. Sinha

A **Chapman & Hall** Book

CRC Press
Taylor & Francis Group

Data Science and Its Applications

Data Science and Its Applications

Edited by
Aakanksha Sharaff
G. R. Sinha

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business
A CHAPMAN & HALL BOOK

First edition published 2021
by CRC Press
6000 Broken Sound Parkway NW, Suite 300, Boca Raton, FL 33487-2742

and by CRC Press
2 Park Square, Milton Park, Abingdon, Oxon, OX14 4RN

CRC Press is an imprint of Taylor & Francis Group, LLC

ISBN: 978-0-367-60886-6 (hbk)
ISBN: 978-0-367-60887-3 (pbk)
ISBN: 978-1-003-10238-0 (ebk)

Typeset in Palatino
by SPi Technologies India Pvt Ltd (Straive)

Dedicated to my late grandparents, my teachers, and family members.

Aakanksha Sharaff

Dedicated to my late grandparents, my teachers, and Revered Swami Vivekananda.

G. R. Sinha

Contents

Preface

Data science is interpreted as either "data with a new platform" or "data used to make market strategy." Data mining, machine learning, artificial intelligence, data analytics, deep learning, and several other related disciplines are covered under the umbrella of data science. Several multinational organizations realize that data science plays an important role in decision-making and market strategy. The revolutionary growth of digital marketing not only changes the market game, but also results in new opportunities for skilled and expert professionals. Today, technologies are rapidly changing and artificial intelligence (AI) and machine learning are contributing as game-changer technologies that are not only trending but also very popular among data scientists and data analysts. Due to widespread usage of data science concepts in almost all emerging applications, there are several challenges and issues in implementation of data analytics and big data problems. Therefore, it is essential to bring out a book that can provide a framework for data handling and management methods so that the issues related to data science applications can be mitigated. This book discusses data science–related scientific methodology, processes, and systems to extract meaningful knowledge or insight for developing concepts from different domains, including mathematics and statistical methods, operations research, computer programming, machine learning, data visualization, and pattern recognition, among others. The book also highlights data science implementation and evaluation of performance in several emerging applications such as cognitive, computer vision, social media analytics, sentiment analysis, and so on. A chapter-by-chapter description of the book follows:

Chapter 1 presents a review of how data science is applied to address various critical problems in the area of big data analytics. It also explains the opportunities and challenges that come with the increase in new computational technologies. In Chapter 2, an overview of different types of real-world recommender systems and AI, along with challenges and opportunities in the age of big data, are provided. This chapter discusses recent growth in cognitive technology, together with advancement in areas such as AI (plus ML, DL, and NLP), as well as knowledge representation, interaction, and personalization, which have resulted in substantial enhancement in the research of recommender systems. Chapter 3 presents some discussion on practical issues and how to resolve them using ensemble learning and meta-learning techniques like bagging, boosting, and stacking. In this chapter, the theory and utility of various machine learning algorithms for data science applications are discussed. The data can be distinguished based on whether it is labeled or not. The major supervised and unsupervised algorithms considered in this chapter are linear regression, decision trees, naïve Bayes, support vector machines (SVMs), and clustering techniques like K-Means. For the sake of completeness, this chapter also presents techniques for making better predictions on the new (unseen) data with performance metrics used for evaluating machine learning algorithms.

Chapter 4 proposes an advanced convolutional neural network (CNN) technique for farming by classifying and recognizing citrus disease that helps grow healthy plants. The model proposed was trained using different training epochs, batch sizes, and dropouts. The dataset includes images of unhealthy and healthy citrus leaves and fruits that can be used to prevent plant disease by using deep learning techniques. The main diseases in the datasets are canker, black spot, greening, scab, and melanose. Chapter 5 presents a detailed discussion and analysis on credibility assessment of social media data. In addition, it

discusses a deep learning–based approach for determining the credibility of user-generated healthcare–related tweets (posts on microblogging website Twitter) and the credibility of their authors by utilizing linguistic features. In particular, we focus on the superstition or misinformation that is spread by the people using such sites, which can lead to hazardous consequences in future. The presented model is based on a semi-supervised approach, where a subset of training tweets, derived from Twitter using web scraping, are labeled as true or false. The remaining tweets are labeled by the model itself. Next, BERT (bidirectional encoder representation using transformers) and CNN (Convolutional Neural Network)–based hybrid model are used for the credibility assessment of tweets. Experimental results show the efficacy of the presented work. Chapter 6 begins with linear filtering concepts of DSP, which are then correlated to time series analysis concepts in both time and spectral domain. The basic knowledge of DSP and random signal processing is assumed, and rigorous mathematical proofs for such concepts are avoided. The chapter then describes limitation of conventional linear filters; the remedy using adaptive filtering algorithms is also described. In addition, an application to forecast stock market index movement is also illustrated using a simulation exercise where the national stock exchange (NSE) index: NIFTY 50 closing values are used to train and test the time-series model. Chapter 7 describes a new framework and model, industry linked additive green curriculum using feed-backward instructional design approach, and a shift from Internet of Things to Internet of Learning Things to inform data science and learning analytics. The chapter traces the forces that shape educational systems, describes the current view of Data Science from educational system perspective including educational data mining and learning analytics, examines framework and features of smart educational systems proposed in literature, describes relevant socioeconomic and technical challenges in adopting learning analytics in educational systems. and concludes with the proposed new framework. In Chapter 8, the hypothesized computational algorithm solution has been illustrated in the method built for more practical learning.

In Chapter 9, a combination of network architectures is used, including question-answering (BERT) and image-captioning (BUTD, show-and-tell model, Captionbot, and show-attend-and-tell model) models for VQA tasks. The chapter also highlights the comparison between these four VQA models. Chapter 10 discusses the benefits of deep learning over machine learning for recommender systems by tuning and optimizing the hyperparameters of the deep neural network, and also throws light on the open issues in recommender systems for researchers. Chapter 11 represents the culmination of a major development project that can be used for the optimization of a single process or guidance during the development of similar supply chain management systems. Chapter 12 introduces various application areas where data science can be useful to analyze healthcare data. This chapter also discusses various issues and challenges faced by data scientists during the knowledge discovery process.

In Chapter 13, the authors have discussed how to solve optimization problems using Python and other tools. The intention of the authors is not to help the user become a skillful theoretician, but a skillful modeler. Therefore, little of mathematical principles related to the subject of optimization is discussed. Various aspects of optimization problems are covered in the case studies mentioned in the chapter. The chapter can be effectively used to create easy yet powerful and efficient models. In Chapter 14, the proposed model shows the numerous advantages of integrating the BioNER model. The proposed model successfully overcomes the problem of improper classification of the biomedical entity, and also improves the performance by leveraging multiple annotated datasets for various types of entities. The state-of-the-art performance of the suggested model increases the accuracy of

text-mining applications related to biomedical downstream, or to find out the relation between the biomedical-entity relationships. In Chapter 15, the performance of various machine learning (ML) classification algorithms is estimated, such as logistic regression, K-nearest neighbor, support vector classifiers, and Gaussian naïve Bayes algorithms implemented on a Crime against Women dataset. Finally, the chapter concludes with the result based on the most reliable outcome from converting the imbalanced data to balanced data by using Python. In Chapter 16, a mathematical model is used to calculte iteration and find the best score of pages. Chapter 17 explains the several deep network architectures that are utilized in the current scenario for scene text analysis. It also compares the detection, recognition, and spotting results for popular approaches on the publicly available scene text datasets.

This book provides a unique contribution to the interdisciplinary field of data science that allows readers to deal with any type of data (image, text, sound, signals, and so on) for a wide range of real-time applications

Acknowledgements

Dr. Sharaff expresses her heartfelt appreciation to her parents, husband Sanju Soni, her loving daughter Aadriti, brother Rahul, sister Shweta, and her entire family members for their wonderful support and encouragement throughout the completion of this important book. This book is an outcome of sincere efforts that could be given to the book only due to the great support of family. This book is dedicated to her parents Mr. Laxman Sharaff and Mrs. Gayatri Sharaff for their entire support and enthusiasm.

Dr. Sinha, too, expresses his gratitude and sincere thanks to his family members, wife Shubhra, daughter Samprati, parents, and teachers.

The editors would like to thank all their friends, well-wishers, and all those who keep them motivated in doing more and more, better and better. They sincerely thank all contributors for writing relevant theoretical background and real-time applications of *Data Science and Its Applications.*

They express their humble thanks to Dr. Aastha Sharma, Acquisitions Editor, and all of the editorial staff at CRC Press for great support, necessary help, appreciation, and quick responses. They also wish to thank CRC Press for giving this opportunity to contribute on a relevant topic with a reputed publisher. Finally, they want to thank everyone, in one way or another, who helped edit this book.

Dr. Sharaff especially thanks her family who encouraged her throughout the time of editing book.

Editor Biographies

Dr. Aakanksha Sharaff is a faculty member in the Department of Computer Science & Engineering at National Institute of Technology Raipur, Chhattisgarh, India, since July 2012. She has been actively involved in research activities leading to data science research and related areas. She holds Doctor of Philosophy in Computer Science & Engineering from National Institute of Technology Raipur (an Institute of National Importance) in 2017; Master of Technology from National Institute of Technology Rourkela (an Institute of National Importance) with Honours in 2012; and Bachelor of Engineering from Government Engineering College Bilaspur Chhattisgarh with Honours in 2010. She received gold medals during her graduation and post-graduation. To date, she pursues excellence and various academic success, including the Top Student in Post-Graduation Master of Technology (2012), Bachelor of Engineering (2010), and throughout her schooling. She has received the gold medal for being the Top Student in Higher Secondary School Certificate Examination (2006) and High School Certificate Examination (2004). She has completed all her degrees and schooling with Honours (Distinction) and studied from reputed national institutions. She has achieved various merit certifications, including All India Talent Search Examination, during her schooling.

She is the Vice Chair of Raipur Chapter of Computer Society of India and Secretary of IEEE Newsletter of Bombay Section. She is actively involved in various academic and research activities. She has received Young Women in Engineering Award for her contribution in the field of Computer Science and Engineering at the 3rd Annual Women's Meet AWM 2018 by Centre for Advanced Research and Design of Venus International Foundation. She has achieved the Best Paper Award for several research papers. She contributes to various conferences as session chairs, invited/keynote speakers, and has published a good number of research papers in reputed international journals and conferences. She is contributing as active technical reviewer of leading international journals for IEEE, Springer, IGI, and Elsevier. Dr. Sharaff has supervised many undergraduate and postgraduate projects. Currently she is guiding five research scholars studying for their PhDs. She has visited Singapore and Bangkok, Thailand, for professional as well as personal reasons. Her research areas focus mainly on data science, text analytics, sentiment analysis, information retrieval, soft computing, artificial intelligence, and machine and deep learning. She is editing one more book on *New Opportunities for Sentiment Analysis and Information Processing* with IGI Publisher.

G. R. Sinha is Adjunct Professor at International Institute of Information Technology Bangalore (IIITB) and is currently deputed as Professor at Myanmar Institute of Information Technology (MIIT) Mandalay Myanmar. He obtained his BE (Electronics Engineering) and MTech (Computer Technology) with Gold Medal from National Institute of Technology Raipur, India. He received his PhD in Electronics & Telecommunication Engineering from Chhattisgarh Swami Vivekanand Technical University (CSVTU) Bhilai, India. He was Visiting Professor (Honorary) in Sri Lanka Technological Campus Colombo for one year (2019–2020). He has published 254 research papers, book chapters, and books at international and national level, including *Biometrics*, published by Wiley India, a subsidiary of John Wiley; *Medical Image Processing*, published by Prentice Hall of India; and has edited five books with IOP, Elsevier, and Springer. He is an active reviewer and editorial member of more than 12 reputed international journals with IEEE, IOP, Springer, Elsevier, and others. He has teaching and research experience of 21 years. He has been Dean of Faculty and Executive Council Member of CSVTU and is currently a member of Senate of MIIT. Dr. Sinha has been delivering ACM lectures as ACM Distinguished Speaker in the field of DSP since 2017 across the world. His few more important assignments include Expert Member for Vocational Training Programme by Tata Institute of Social Sciences (TISS) for Two Years (2017–2019); Chhattisgarh Representative of IEEE MP Sub-Section Executive Council (2016–2019); Distinguished Speaker in the field of Digital Image Processing by Computer Society of India (2015). He is the recipient of many awards and recognitions, like TCS Award 2014 for Outstanding Contributions in Campus Commune of TCS, Rajaram Bapu Patil ISTE National Award 2013 for Promising Teacher in Technical Education by ISTE New Delhi, Emerging Chhattisgarh Award 2013, Engineer of the Year Award 2011, Young Engineer Award 2008, Young Scientist Award 2005, IEI Expert Engineer Award 2007, ISCA Young Scientist Award 2006 Nomination, and Deshbandhu Merit Scholarship for five years. He served as Distinguished IEEE Lecturer in IEEE India council for the Bombay section. He is a Senior Member of IEEE, Fellow of Institute of Engineers India, and Fellow of IETE India.

He has delivered more than 50 keynote and invited talks, and has chaired many technical sessions for international conferences across the world. His special session on "Deep Learning in Biometrics" was included in the IEEE International Conference on Image Processing 2017. He is also member of many national professional bodies like ISTE, CSI, ISCA, and IEI. He is a member of various committees of the university and has been Vice President of Computer Society of India for Bhilai Chapter for two consecutive years. He is a consultant for various skill development initiatives of NSDC, Government of India. He is regular referee of project grants under the DST-EMR scheme and several other schemes for Government of India. He received few important consultancy supports as grants and travel support. Dr. Sinha has supervised eight PhD scholars, 15 MTech scholars, and has been supervising one more PhD scholar. His research interest includes biometrics, cognitive science, medical image processing, computer vision, outcome-based education (OBE), and ICT tools for developing employability skills.

List of Contributors

Aakanksha Sharaff
Department of Computer Science and Engineering
National Institute of Technology
Raipur, Chhattisgarh, India

Aaryan Kapoor
Department of Computer Science and Information Systems
BITS Pilani
Pilani, Rajasthan, India

Abhilasha Chaudhuri
Department of IT
NIT Raipur, India

Ajay Dhruv
Research Scholar, Department of Information Technology
Thadomal Shahani Engineering College
Mumbai, India

Alok Negi
Department of Computer Science and Engineering
National Institute of Technology
Uttarakhand, India

Amin Beheshti
Macquarie University
Sydney, Australia

Ashutosh Kumar
National Institute of Technology Raipur
Raipur, India

Dr. Naresh Kumar Nagwani
National Institute of Technology Raipur
Raipur, India

Dr. V. Kakulapati
Sreenidhi Institue of Science and Technology
Hyderabad, India

Emir Žunić
Info Studio d.o.o. Sarajevo and Faculty of Electrical Engineering
University of Sarajevo
Bosnia and Herzegovina

Emmanuel S. Pilli
Department of CSE
MNIT Jaipur
Jaipur, India

G. R. Sinha
Myanmar Institute of Information Technology
Mandalay, Myanmar

Hari Prabhat Gupta
IIT (BHU)
Varanasi, India

Haris Hasić
Tokyo Institute of Technology, Japan and Info Studio d.o.o. Sarajevo
Bosnia and Herzegovina

J. W. Bakal
Department of Information Technology, SSJCOE
Mumbai, India

Jahangir Alam
University Women's Polytechnic, F/o. Engineering & Technology, AMU Aligarh
Alligarh, Uttar Pradesh, India

Jigarkumar H. Shah
Pandit Deendayal Petroleum University
Gandhinagar, Gujarat, India

Kerim Hodžić
Faculty of Electrical Engineering, University of Sarajevo and Info Studio d.o.o. Sarajevo
Bosnia and Herzegovina

Krishan Kumar
Department of Computer Science and Engineering
National Institute of Technology
Uttarakhand, India

Lavika Goel
Department of Computer Science and Engineering
Malaviya National Institute of Technology
Jaipur, Rajasthan, India

Mansi A. Radke
Visvesvaraya National Institute of Technology
Nagpur, Maharashtra, India

Meenakshi S Arya
Department of CSE, Faculty of E&T
SRM Institute of Science and Technology, Vadapalani Campus, Chennai, India

Meera S Datta
NIIT University
India

Mehdi Elahi
University of Bergen
Bergen, Norway

Mohit Dhawan
Department of Electrical and Electronic BITS Pilani
Pilani, Rajasthan, India

Monika Choudhary
Department of CSE, MNIT Jaipur
Jaipur, India
Department of CSE, IGDTUW
Delhi, India

P. Tamilarasi
Sri Sarada College for Women (Autonomous)
Salem, Tamil Nadu, India

Dr. R. Uma Rani
Sri Sarada College for Women (Autonomous)
Salem, Tamil Nadu, India

Rachit Rathore
Department of Computer Science and Information Systems
BITS Pilani
Pilani, Rajasthan, India

Randheer Bagi
IIT (BHU)
Varanasi, India

Ravindra B. Keskar
Visvesvaraya National Institute of Technology
Nagpur, Maharashtra, India

Robert B. Handfield
North Carolina State University and Supply Chain Resource Cooperative
NC, United States

Rutvij H. Jhaveri
Pandit Deendayal Petroleum University
Gandhinagar, Gujarat, India

Satyansh Rai
Department of Computer Science and Information Systems
BITS Pilani
Pilani, Rajasthan, India

Satyendra Singh Chouhan
Department of CSE
MNIT Jaipur
Jaipur, India

Sead Delalić
Faculty of Science
University of Sarajevo and Info Studio d.o.o. Sarajevo
Bosnia and Herzegovina

Sheri Mahender Reddy
Otto-Friedrich university of Bamberg, IsoSySc
Bamberg, Germany

Srinivasa Reddy Goluguri
Macquarie University
Sydney, Australia

Supriya Gupta
National Institute of Technology Raipur
Raipur, India

Tanima Dutta
IIT (BHU)
Varanasi, India

Tirath Prasad Sahu
Department of IT
NIT Raipur, C.G. India
Raipur, India

Ulligaddala Srinivasarao
Department of Computer Science and Engineering
National Institute of Technology
Raipur, Chhattisgarh, India

Vijay V Mandke
NIIT University
India

Yashvardhan Sharma
Department of Computer Science and Information Systems
BITS Pilani
Pilani, Rajasthan, India

1

Introduction to Data Science: Review, Challenges, and Opportunities

Ulligaddala Srinivasarao and Aakanksha Sharaff
National Institute of Technology, Raipur, Chhattisgarh, India

G. R. Sinha
Myanmar Institute of Information Technology (MIIT), Mandalay, Myanmar

CONTENTS

1.1 Introduction

Data science is a new area of research that is related to huge data and involves concepts like collecting, preparing, visualizing, managing, and preserving. Even though the term *data science* looks related to subject areas like computer science and databases, it also requires other skills, including non-mathematical ones. Data science not only combines data analysis, statistics, and other methods, but it also includes the corresponding results. Data science is intended to analyze and understand the original phenomenon related to the data by revealing the hidden features of complex social, human, and natural phenomena related to data from another point of view other than traditional methods.

Data science includes three stages: designing the data, collecting the data, and finally analyzing the data. There is an exponential increase in the applicability of data science in various areas because data science has been making enormous strides in data processing and use. Business analytics, social media, data mining, and other disciplines have benefited due to the advance in data science and have shown good results in the literature.

Data science has made remarkable advancements in the fields of ensemble machine learning, hybrid machine learning, and deep learning. Machine learning methods (ML) can learn from the data with minimum human interference. Deep learning (DL) is a subset of ML that is applicable in different areas, like self-driving cars, earthquake predictions, and so on. There are many pieces of evidence in the literature that show the superiority of DL over ML methods; DL methods include artificial neural networks, k-nearest neighbors, and support vector machine (SVM) in different disciplines, such as medical, social media, and so on.

Torabi et al. developed a hybrid model where two predictive machine learning algorithms are combined together [1]. Here, an additional optimization-based method has also been used for maximizing the prediction function. Mosavi and Edalatifar illustrated that hybrid machine learning models perform very accurately compared to single machine learning models [2].

This chapter presents a review of various data science methods and details how they are used to deal with critical challenges that arise when working with big data analytics. According to the literature, different classification, regression, clustering, and deep learning–based methods have often been used. However, there is an opportunity to improve in new areas, like temporal and frequent pattern discovery for load prediction. This chapter also discusses the future trends of data science, to explore new tools and algorithms that are capable of intelligently handling large datasets that are collected from various sources.

1.2 Data Science

Technological tools developed recently over the years have helped in many domains, including management and big data. Advancements in different areas of communications

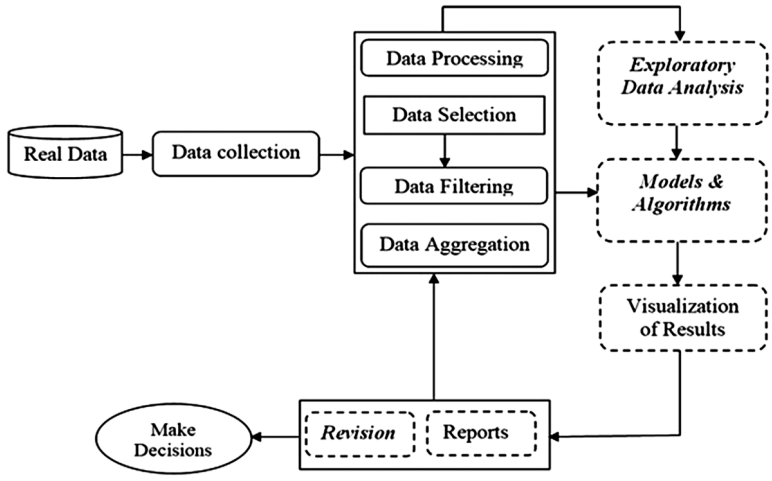

FIGURE 1.1
Data science process

and information technology—like email information privacy, market, stock data, data science, and real-time monitoring—have also been a good influence.

It is well known that data science builds algorithms and systems for discovering knowledge, detecting the patterns, and generating useful information from massive data. To do so, it encompasses an entire data analysis process that starts with the extraction of data and cleaning, and extends to data analysis, description, and summarization. Figure 1.1 depicts the complete process. It starts with data collection. Next, the data is cleaned to select the segment that has the most valuable information. To do so, the user will filter over the data or formulate queries that can erase unnecessary information. After the data is prepared, an exploratory analysis that includes visualizing tools will help decide the algorithms that are suitable to gain the required knowledge. This complete process will guide the user toward the results that will help them make suitable decisions.

Depending on the primary outcomes, the complete process should be fine-tuned to obtain improved results. This will involve changing the parameter values or making changes to the datasets. These kinds of decisions are not made automatically, so the involvement of an expert in result analysis is a crucial factor.

From a technical point of view, data science consists of a set of tools and techniques that deals with various goals corresponding to multiple situations. Some of the recent methods used are clustering, classification, deep learning, regression, association rule mining, and time-series analysis. Even though these methods are often used in text mining and other areas, anomaly detection and sequence analysis are also helpful to provide excellent results for text mining problems.

1.2.1 Classification

Wu et al. have classified a set of objects that predict the classes based on the attributes. Decision trees (DT) are used to perform and visualize that classification [3]. DTs may be generated using various algorithms, such as ID3, CLS, CART, C4.5, and C5.0. Random forest (RF) is one more classifier that will construct a set of DTs, and then predicts through the aggregation of the values generated from each DT. A classification model was developed

by using a technique known as Least Squares Support Vector Machine (LS-SVM). The classification task is performed by LS-SVM by using a hyper-plane in a multidimensional space for separating the dataset into the target classes [4].

1.2.2 Regression

Regression analysis aims for the numerical estimation of the relationship between variables. This involves the estimation of whether or not the variables are independent. If a variable is not independent, then the first step is to determine the type of dependence. Chatterjee et al. proposed a regression analysis that is often used for predicting and forecasting, and also to understand how the dependent variables will change corresponding to the fixed values of independent variables [5].

1.2.3 Deep Learning

In deep learning, many hidden layers of neural networks are used to deeply understand the information that images are attempting to predict accurately. Here, each layer will learn and detect low-level features, such as edges. Further, new layers will be merged with the features of the previous layer to represent it better. Fischer and Krauss [6] have expanded the long short-term memory (LSTM) networks for forecasting out-of-sample directional movements in the stock market. Here, a comparative study has been performed with DNN, RF, and LOG, and it demonstrates that the LSTM model outperforms the others. Tamura et al. [7] have proposed a model for predicting stock values, which is a two-dimensional approach. In this model, technical, financial indexes related to the Japanese stock market are used as input data for LSTM to predict. Using this data, the financial statements of other companies have been retrieved and are also added to the database.

1.2.4 Clustering

Jain et al. proposed a clustering-based method using the degree of similarity [8]. In clustering, the objects are separated into groups called clusters. This type of learning is called unsupervised learning, as there is no prior idea over the classes as to which group the objects belong. Based on the similarity measure criterion, cluster analysis has various models: (i) based on the connectivity distance, connectivity models are generated, i.e., hierarchical clustering; (ii) by using the nearest cluster center, the objects are assigned, centroid models are generated, i.e., k-means; (iii) by means of statistical distributions, the distributed models are generated, i.e., expectation-maximization algorithm; (iv) based on high-density areas that exist in the data, the clusters are defined in density models; (v) graphs are used for expressing the dataset in graph-based models.

1.2.5 Association Rules

Association rules are suitable tools to represent the new information that has been extracted from the raw dataset. These rules are expressed to make the decisions in terms of implication rules, per Verma et al. [9]. The respective rules indicate the frequency of occurrence of the attributes with high reliability in databases. This example represents an association rules related to the database of the supermarket. Even though the algorithms like ECLAT and FP-Growth algorithms are available for large datasets, in the Apriori algorithm, for

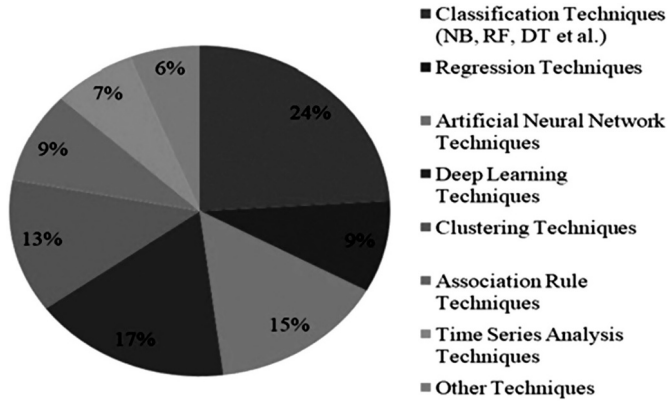

FIGURE 1.2
Data Science Techniques

example, the generalized rule induction algorithm and its adaptations are often used, per Tan et al. [10].

1.2.6 Times Series Analysis

Das provided a time-series analysis. Here the time-series data, which is collected over time, is used for modeling the data. Further, the model is used for predicting future values of the time series [11]. The often used methods are the following: (i) techniques for exploratory analysis, for example wavelets, trend analysis, autocorrelation, and so on; (ii) forecasting and prediction methods, for example signal estimation, regression methods, and so on; (iii) classification techniques that will be assigned a category to patterns related to the series; and (iv) segmentation that aims to identify a sequence of points that share particular properties. Hullermeier developed a fuzzy extension that allows for processing uncertain and imprecise data related to different domains [12]. Bezdek et al. have proposed a fuzzy k-means method. This method is similar to a type of clustering technique that has given efficient results in different scenarios, as it will permit the assignment of data elements related to single or more clusters [13]. Figure 1.2. shows different types of techniques used in data science and application.

1.3 Applications of Data Science in Various Domains

Data science is one subject that has gained popularity out of necessity, corresponding to real-world applications as a substitute to research domain. Its application began from a narrow field of analytics and statistics and has improved to be applied to different areas of industry and science. Consequently, this section explains the data science applications that can do the following: (i) economic analysis of electric consumption, (ii) stock market prediction, (iii) bioinformatics, (iv) social media analytics, (v) email mining, (vi) big data analysis, and (vii) SMS Mining, among other things!

1.3.1 Economic Analysis of Electric Consumption

Different electric companies or utilities approached data science to find out and understand when and how consumers use energy. There has been an increase in competition among companies that use data science to develop such information. Traditionally, this information has been determined via classification, clustering, and pattern analysis methods by using the association rule. Chicco et al. have grouped consumers as various classes based on their behavior and usage of electricity [14]. The comparative evaluation was made with self-organizing maps and an improved version of follow-the-leader methods. This was the first step initiated for a tariff of the electrical utilities. Figueiro et al. [15] have developed a framework for exploiting the historical data, which consists of two modules: (i) a load-profile module, which creates a set of customer classes by using unsupervised and supervised learning, and (ii) a classification module, which builds models for assigning customers to their respective classes.

1.3.2 Stock Market Prediction

An application of ML and DL techniques in the stock market is increasing compared to other areas of economics. Even though investing in the stock market gives profits, high risk is often involved along with high benefits. So, investors try to estimate and determine the value of a stock before they make an investment. The cost of the stock varies depending upon factors like local politics and economy, which causes difficulties in identifying future trends of the stock market. Fischer and Krauss [6] used LSTM to forecast future trends in the stock market. The results have been compared with LOG, DNN, and RF, and have shown improved results over the others. Tamura et al. [7] have proposed a new method for predicting the values of the stock. Here, financial data related to the stock market of Japan has been used as a prediction input in LSTMs (Long short-term memories). Further, the financial statements of the companies are recovered and then added to the database. Sharaff and Srinivasarao [16] proposed Linear Support Vector Machine (LSVM) identify the correlation among the words in content and subject of the emails.

1.3.3 Bioinformatics

Bioinformatics is a new area that uses computers to understand biological data like genomics and genetics. This helps scientists understand the cause of disease, physiological properties, and genetic properties. Baldi et al. [17] utilized various techniques to estimate the applicability and efficiency of different predictive methods in the classification task. The previous error estimation techniques are primarily focused on supervised learning using the microarray data. Michiels et al. [18] have used various random datasets to predict cancer using microarray data. Ambroise et al. [19] solved a gene selection problem based on microarrays data. Here, 10-fold validation has been used. Here, 0.632 bootstrap error estimates are used to deal with prediction rules that are overfitted. The accuracy of 0.632 bootstrap estimators for microarray classification using small datasets is proposed in Braga et al. [20]

1.3.4 Social Media Analytics

Joshi and Deshpande [21] have used Twitter data to classify the sentiments included in tweets. They have applied various machine learning methods to do so. A comparative

study has been carried out by using maximum entropy, naïve Bayes, and positive-negative word counting. Wolny [22] proposed a model to recognize the emotion in Twitter data and performed an emotion analysis study. Here, the feelings and sentiments were discussed in detail by explaining the existing methods.

The emotion and sentiment are classified based on symbols via an unsupervised classifier, and the lexicon was explained by suggesting future research. Coviello et al. [23] have analyzed the emotion contagion related to Facebook data. The instrumental variable regression technique has been used to analyze the Facebook data. Here, the emotions of the people, such as negative and positive emotions during rainy days, were detected. Roelens et al. [24] explained that the detection of the people who influence social networks is a difficult task or area of research, but one of great interest so that referral marketing and spreading information regarding products can reac the maximum possible network.

1.3.5 Email Mining

There is a threat to internet security with spam emails. Spam emails are nothing but unwanted or unsolicited emails. Mailboxes will overload with these unwanted emails, and there may be losses in storage and bandwidth, which favors quick, wrong information and malicious data. Gudkova et al. [25] conducted a study and explained that 56% of all emails are spam emails. Caruana and Li [26] illustrated that the machine learning method is successful for detecting spam data. These include learning classifier models, which map data by using features like n-gram and others into spam or ham classes. Dada et al. [27] have demonstrated that email features may be either manual or automatic. Bhowmick and Hazarika [28] demonstrated that the manually extracted rules are known as knowledge engineering, which requires expert and regular updates to maintain good accuracy. Text mining methods are used for automated feature extraction of useful information like words, enabling spam discrimination, HTML mark up, and so on. Using these features, an email is represented as Bag-of-Words (BoW) as proposed by Aggarwal [29]. Here the unstructured word tokens are used to discriminate the spam messages with the others. The BoW assumes word tokens that are not dependent that will prevent from delivering the good semantic content to represent the email. Sharaff and Nagwani [30] have identified the email threads using LDA- and NMF-based methodology.

1.3.6 Big Data Analysis Mining Methods

Big data is one of the very fast-growing technologies that is critical to handle in the present era. The information is used for analytical studies to help drive decisions for giving quick and improved services. Laney [31] proposed that big data consists of three characteristics: velocity, volume, and variety. These are also called the 3Vs.

Chen et al. [32] explained that data mining is a procedure where potentially useful, unknown, and hidden meaningful information is extracted from noisy, random, incomplete, and fuzzy data. The knowledge and information that has been extracted is used to derive new comprehensions, scientific events, and influences business scientific discovery, per Liu [33].

Two articles have aimed at improving the accuracy of data mining. Han et al. [34] have proposed a new model using the skyline algorithm. Here, a sorted positional index list (SSPL), which has low space overhead, has been used to reduce the input or output cost. Table 1.1 shows an overview of data science methods used in different applications.

TABLE 1.1

An overview of data science methods used in different applications

S.no	Applications	Methods	Source
1	Economic analysis	Follow-the-Leader Clustering (FLC) K-Means	Chicco et al. [14] Figueiredo et al. [15]
2	Stock Market	Long Short-Term Memory (LSTM)	Fischer and Krauss [6] Tamura et al. [7]
3	Bioinformatics	Gradient Descent Learning (GDL) k-nearest-neighbors (K-NN) support vector machine (SVM)	Baldi et al. [17] Michiels et al. [18] Ambroise et al. [19]
4	Social Media analytics	Naive Bayes (NB) and Maximum Entropy Algorithms (MEA) Lexicon Based Approach (LBA) Regression Methods (RM)	Joshi and Deshpande [21] Wolny [22] Coviello et al. [23]
5	Email Mining	Machine and Non-Machine Learning Methods (NMLM) Deep Leaning Methods (DLM) Machine Learning Techniques (MLT) Latent Dirichlet Allocation and Non-Negative Matrix Factorization (NNMF)	Caruana and Li [26] Dada et al. [27] Bhowmick and Hazarika [28] Sharaff and Nagwani [30]
6	Big Data Analysis	Fuzzy Clustering (FC) Data Mining Methods (DMM) Skyline Algorithm (SA)	Chen et al. [32] Liu [33] Han et al. [34]

1.4 Challenges and Opportunities

This section summarizes the key issues, challenges, and opportunities that are related to data science in different fields.

1.4.1 Challenges in Mathematical and Statistical Foundations

The main challenge in mathematical fields is to find out why theoretical foundations are not enough to solve complex problems, and then identify and obtain a helpful action plan.

1.4.2 Challenges in Social Issues

In social contexts, the challenges are to specify, respect, and identify social issues. Any domain-specific data is to be selected, and then its related concepts—like business, security, protection privacy—should be accurately handled.

1.4.3 Data-to-Decision and Actions

It is important to develope accurate decision-making systems that are data-driven. These systems should also be able to manage and govern the decision-making systems.

1.4.4 Data Storage and Management Systems

One of the challenges include designing a good storage and management system that has the capability to handle large amounts data, stream-speed in real time, and can manage such data in an Internet-based environment, including cloud.

1.4.5 Data Quality Enhancement

Another important challenge is issues of data quality like uncertainty, noise, unbalance, and so on. The level of presence of these issues will vary depending upon the data complexity.

1.4.6 Deep Analytics and Discovery

Cao [35] proposed new algorithms to deal with the deep and implicit analytics that are not able to be tackled using the existing descriptive, latent, and predictive learning. Also, how to aggregate the model based with data-driven problem-solving solutions to balance the domain-specific data complexity, intelligence-driven evidence learning, and common learning frameworks.

1.4.7 High-Performance Processing and Analytics

Systems must handle the online, real-time, Internet-based, large-scale, high-frequency, data analytics and processing with balanced resource involvement that may be local and global. This requires new array disk storage, batch, and high performance parallel processing. It is also necessary to use complex matrix calculations, data-to-knowledge management, mixed data structures, and management systems.

1.4.8 Networking, Communication, and Interoperation

The challenge involved is how to support the interoperation, communication, and networking between various data science roles like distributed and complete cycle of problem-solving in data science. Here, it is necessary to coordinate management of tasks, data, workflows, control, task scheduling, and governance.

1.5 Tools for Data Scientists

This section presents the tools required for data scientists to address the aspects discussed above. These tools are classified as data and application integration, cloud infrastructure, programming, visualization, high-performance processing, analytics, master data management, business intelligence reporting, data preparation and processing, and project management. The researcher can use any number of tools depending upon the complexity of the problem being solved.

1.5.1 Cloud Infrastructure

Like Map R, Google Cloud Platform, Amazon Web Services, Cloudera, Spark, Apache Hadoop, and other systems may be used. Most of the traditional IT vendors at present are using cloud platform.

1.5.2 Data/Application Integration

This includes Clover ETL, Information Builders, DM Express Sync sort, Oracle Data Integrator, Informatics, Including Ab Initio, and so on.

1.5.3 Master Data Management

Master data management includes SAP Net Weaver Master Data Management tool, Black Watch Data, Microsoft Master Data Services, Informatica MDM, TIBCO MDM, Teradata Warehousing, and so on.

1.5.4 Data Preparation and Processing

Stodder and Matters [36] have used some platforms and data preparation tools like Wrangler Enterprise and Wrangler, Alpine Chorus, IBM SPSS, Teradata Loom, Platfora, and so on.

1.5.5 Analytics

Analytics includes commercial tools like Rapid Miner [37], Mat Lab, IBM SPSS Modeler and SPSS Statistics, SAS Enterprise Miner, and so on, in addition to some new tools, like Google Cloud Prediction API, ML Base, Big ML [38], Data Robot, and others.

1.5.6 Visualization

Some commercial and free software listed in KDnuggets [39] to visualize include Miner3D, IRIS Explorer, Interactive Data Language, Quadrigram, Science GL, and so on.

1.5.7 Programming

Additionally, Java, Python, SQL, SAS, and R languages have been used for data analytics. Some data scientists have also included Go, Ruby, .net, and Java Script [40].

1.5.8 High-Performance Processing

Around 40 computer cluster software programs, like Platform Cluster Manager, Moab Cluster Suite, Stacki, and others, have been listed in Wikipedia [41].

1.5.9 Business Intelligence Reporting

Some of the reporting tools [42] commonly used are SAP Crystal Reports, SAS Business Intelligence, Micro Strategy, and IBM Cognos, among others.

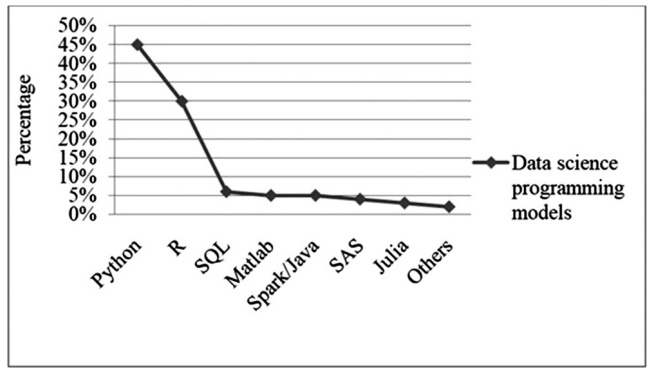

FIGURE 1.3
Data Science Programming Models

1.5.10 Social Network Analysis

Around 30 tools have been listed for social network analysis and to help visualize data. For example, Ego Net, Cuttlefish, Commetrix, Keynetiq, Node XL, and so on. [43]. Figure 1.3 shows the different types of programming languages that are used in data science.

1.6 Conclusion

This chapter has surveyed the modern advances in information technology, and the influence these advances have had on big data analytics and its applications. The effectiveness of different data science algorithms that can be applied to solve the challenges in big data has been examined. Data science algorithms will be extensively used in the future to address the problems and challenges in big data applications.

In various areas, the exploitation and discovery of meaningful insights from the dataset will be very much required. Big data applications are necessary in different fields like industry, government, and so on. This new perspective will challenge research groups to develop better solutions to manage large heterogeneous amounts of real-time data. It also deals with the uncertainty associated with it. Data science techniques reveal important tools that can extract and exploit the information and knowledge that exists in the user dataset. In the coming days, big data techniques will increase possibilities, and may also democratize them.

References

1. Torabi, M., Hashemi, S., Saybani, M. R., Shamshirband, S., & Mosavi, A. (2019). A Hybrid clustering and classification technique for forecasting short-term energy consumption. *Environmental Progress & Sustainable Energy*, 38(1), 66–76.

2. Mosavi, A., & Edalatifar, M. (2018). A hybrid neuro-fuzzy algorithm for prediction of reference evapotranspiration. In *International conference on global research and education* (pp. 235–243). Cham: Springer.
3. Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., … & Zhou, Z. H. (2008). Top 10 algorithms in data mining. *Knowledge and information systems*, 14(1), 1–37.
4. Suykens, J. A., Van Gestel, T., & De Brabanter, J (2002). *Least squares support vector machines*. World Scientific.
5. Chatterjee, S., Hadi, A. S., & Price, B. (2000). *Regression analysis by example*. New York: John Wiley & Sons Inc..
6. Fischer, T., & Krauss, C. (2018). Deep learning with long short-term memory networks for financial market predictions. *European Journal of Operational Research*, 270(2), 654–669.
7. Tamura, K., Uenoyama, K., Iitsuka, S., & Matsuo, Y. (2018). Model for evaluation of stock values by ensemble model using deep learning.
8. Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: a review. *ACM computing surveys (CSUR)*, 31(3), 264–323.
9. Verma, M., Srivastava, M., Chack, N., Diswar, A. K., & Gupta, N. (2012). A comparative study of various clustering algorithms in data mining. *International Journal of Engineering Research and Applications (IJERA)*, 2(3), 1379–1384.
10. Tan, P. N., Steinbach, M., & Kumar, V. (2016). *Introduction to data mining*. Delhi: Pearson Education India.
11. Das, S. (1994). *Time series analysis*. (Vol 10). Princeton, NJ: Princeton University Press.
12. Hüllermeier, E. (2005). Fuzzy methods in machine learning and data mining: status and prospects. *Fuzzy Sets and Systems*, 156(3), 387–406.
13. Bezdek, J. C., Ehrlich, R., & Full, W. (1984). FCM: the fuzzy c-means clustering algorithm. *Computers & Geosciences*, 10(2–3), 191–203.
14. Chicco, G., Napoli, R., Piglione, F., Postolache, P., Scutariu, M., & Toader, C. (2004). Load pattern-based classification of electricity customers. *IEEE Transactions on Power Systems*, 19(2), 1232–1239.
15. Figueiredo, V., Rodrigues, F., Vale, Z., & Gouveia, J. B. (2005). An electric energy consumer characterization framework based on data mining techniques. *IEEE Transactions on power systems*, 20(2), 596–602.
16. Sharaff, A., & Srinivasarao, U. (2020). Towards classification of email through selection of informative features. In *2020 First International Conference on Power, Control and Computing Technologies (ICPC2T)* (pp. 316–320). IEEE.
17. Baldi, P., Brunak, S., Chauvin, Y., Andersen, C. A., & Nielsen, H. (2000). Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, 16(5), 412–424.
18. Michiels, S., Koscielny, S., & Hill, C. (2005). Prediction of cancer outcome with microarrays: a multiple random validation strategy. *The Lancet*, 365(9458), 488–492.
19. Ambroise, C., & McLachlan, G. J. (2002). Selection bias in gene extraction on the basis of microarray gene-expression data. *Proceedings of the national academy of sciences*, 99(10), 6562–6566.
20. Braga-Neto, U. M., & Dougherty, E. R. (2004). Is cross-validation valid for small-sample microarray classification? *Bioinformatics*, 20(3), 374–380.
21. Joshi, S., & Deshpande, D. (2018). Twitter sentiment analysis system. *International Journal of Computer Applications*, 180(47), 0975–8887.
22. Wolny, W. (2016). Emotion analysis of twitter data that use emoticons and emoji ideograms.
23. Coviello, L., Sohn, Y., Kramer, A. D., Marlow, C., Franceschetti, M., Christakis, N. A., & Fowler, J. H. (2014). Detecting emotional contagion in massive social networks. *PloS one*, 9(3), e90315.
24. Roelens, I., Baecke, P., & Benoit, D. F. (2016). Identifying influencers in a social network: the value of real referral data. *Decision Support Systems*, 91, 25–36.
25. Gudkova, D., Vergelis, M., Demidova, N., and Shcherbakova, T. (2017). Spam and phishingin Q2 2017, Securelsit, Spam and phishing reports, https://securelist.com/spamand-phishing-in-q2-2017/81537/, 2017.

26. Caruana, G., & Li, M. (2008). A survey of emerging approaches to spam filtering. *ACM Computing Surveys (CSUR)*, 44(2), 1–27.
27. Dada, E. G., Bassi, J. S., Chiroma, H., Adetunmbi, A. O., & Ajibuwa, O. E. (2019). Machine learning for email spam filtering: review, approaches and open research problems. *Heliyon*, 5(6), e01802.
28. Bhowmick, A., & Hazarika, S. M. (2016). Machine learning for e-mail spam filtering: review, techniques and trends. *arXiv preprint arXiv:1606.01042*.
29. Aggarwal, C. C., & Zhai, C. (Eds.). (2012). *Mining text data*. Springer Science & Business Media.
30. Sharaff, A., & Nagwani, N. K. (2016). Email thread identification using latent Dirichlet allocation and non-negative matrix factorization based clustering techniques. *Journal of Information Science*, 42(2), 200–212.
31. Laney, D. (2001). 3D data management: controlling data volume, velocity and variety. *META group research note*, 6(70), 1.
32. Chen, M. M. S., & Liu, Y. (2014). Big Data: A Survey. *Mobile Networks and Applications*, 19, 171–209.
33. Liu, L. (2013). Computing infrastructure for big data processing. *Frontiers of Computer Science*, 7(2), 165–170.
34. Han, X., Li, J., Yang, D., & Wang, J. (2012). Efficient skyline computation on big data. *IEEE Transactions on Knowledge and Data Engineering*, 25(11), 2521–2535.
35. Cao, L. (2017). Data science: challenges and directions. *Communications of the ACM*, 60(8), 59–68.
36. Stodder, D., & Matters, W. D. P. (2016). Improving data preparation for business analytics. Applying technologies and methods for establishing trusted data assets for more productive users. *Best Practices Report Q*, 3(2016), 19–21.
37. RapidMiner. 2016. RapidMiner. (2016). https://rapidminer.com/.
38. BigML. 2016. BigML. Retrieved from https://bigml.com/.
39. KDnuggets. 2015. Visualization Software. Retrieved from: http://www.kdnuggets.com/software/visualization.html.
40. Davis, J. (2016). 10 Programming Languages And Tools Data Scientists Used. (2016).
41. Wikipedia. 2016. Comparison of Cluster Software. Retrieved from https://en.wikipedia.org/wiki/Comparison_of_cluster_software.
42. Capterra. 2016. Top Reporting Software Products. Retrieved from http://www.capterra.com/reporting-software/.
43. Desale, D. (2015). Top 30 Social Network Analysis and Visualization Tools. KDnuggets. https://www.kdnuggets.com/2015/06/top-30-social-network-analysis-visualization-tools.html.

2

Recommender Systems: Challenges and Opportunities in the Age of Big Data and Artificial Intelligence

Mehdi Elahi

University of Bergen, Bergen, Norway

Amin Beheshti and Srinivasa Reddy Goluguri

Macquarie University, Sydney, Australia

CONTENTS

2.1 Introduction

In the times of Big Data, choosing the right products is a challenge for consumers due to the massive *volume*, *velocity*, and *variety* of related data produced online. Because of this, users are getting more and more desperate when making choices among an unlimited set of choices. Recommender systems are support apps that can deal with this challenge by assisting shoppers to make choices on what to purchase (Jannach, Zanker, Felfernig, and Friedrich, 2010; Resnick and Varian, 1997; Ricci, Rokach, and Shapira, 2015). Recommender systems can learn from *particular* preferences and tastes of users and build personalized suggestions that tailor to users' preferences and necessities rather than offering suggestions based on mainstream taste (Elahi, 2011; Elahi, Repsys, and Ricci, 2011).

Many recommender software options and algorithms have been proposed, up to now, by the academic and industrial community. Most of these algorithms are capable of getting input data from various data types and then exploiting them to generate recommendations on top of the data. These data types can describe either the item content (e.g., category, brand, and tags) or the user preferences (e.g., ratings, likes, and clicks). The data is collected and pre-processed, cleaned, and then exploited to build a model in which the items are projected as arrays of features. Recommendation lists for a specific user is then made by filtering the items that represent alike features to the rest of the item sets that user liked/rated high.

Enhanced capabilities of recommender techniques in understanding the varied categories of user tastes and precisely tackling information burden has enabled them to become an important part of any online shop that tackles the expansion of item cataloging (Burke, 2002; Elahi, 2014). Diverse categories of recommender engines have been built in order to generate personalized selection and relevant recommendations of products and services ranging from clothing and outfits to movies and music. Such a personalized selection and suggestion is usually made based on the big data of a huge community of connected users, and by calculating the patterns and relationships among their preferences (Chao, Huiskes, Gritti, and Ciuhu, 2009; Elahi, 2011; Elahi and Qi, 2020; He and McAuley, 2016; Nguyen, Almenningen, Havig, Schistad, Kofod-Petersen, Langseth, and Ramampiaro, 2014; Quanping 2015; Tu and Dong 2010). The excellency in performance of recommender systems has been validated in the diverse range of e-commerce applications where a choice support mechanism is necessary to handle customers' needs and help them when interacting with online e-commerce. Such an assistance improves the user experiences when shopping or browsing the system catalogue (He and McAuley, 2016; Tu and Dong, 2010).

In this chapter, we will provide an outline of different types of real-world recommender systems, along with challenges and opportunities in the age of big data and AI. We will discuss the progress in cognitive technology, in addition to evolutionary development in areas such as AI (with all relevant disciplines such as ML, DL, and NLP), KR, and HCI, and how they can empower recommender systems to effectively support their users.

We discuss that modern recommendation systems require access to and the ability to understand big data, in all different forms, and that big data generated on data islands can

be used to build relevant and personalized recommendations tailored to each customer's needs and preferences. We present different application scenarios (including multimedia, fashion, tourism, banking, and education) and review potential solutions for the recommendation. The remaining parts of the chapter is organized as follows: Section 2.2 briefly describes popular methods and algorithms. Section 2.3 discusses different application scenarios, and Section 2.4 reviews real-world challenges and potential solutions. Section 2.5 extends the previous chapters by providing some advanced topics. Finally, in Section 2.6, we conclude the chapter.

2.2 Methods

2.2.1 Classical

Diverse recommendation approaches have already been developed and tested, which can be classified within a number of categories. A well-adopted category of methods is called **content-based** (Pazzani and Billsus, 2007). Methods within this category suggest items based on their descriptors (Balabanovíc and Shoham, 1997). For example, book recommender systems take terms within the text of a book as descriptors and suggest to the user other books that have descriptors similar to the book the user liked in the past. Another popular category is **collaborative filtering** (Desrosiers and Karypis, 2011; Koren and Bell, 2011). Collaborative filtering methods predict the preferences (i.e., ratings) of users by learning the preferences that a set of users provided to items and suggests to users those items with the highest predicted preferences. Methods within the **demographic** (Wang, Chan, and Ngai, 2012) category generate recommendations by identifying similar users based on the demographics of the users (Pazzani, 1999). These methods attempt to group existing users by their personal descriptors and make relevant suggestions based on their demographic descriptions. **Knowledge-based (**Felfernig and Burke, 2008) methods are another category that tries to suggest items that are inferred from the needs and constrains entered by users (Burke, 2000). Knowledge-based methods are distinguished by their knowledge about how a specific item fulfills a particular user's needs (Claypool, Gokhale, Miranda, Murnikov, Netes, and Sartin, 1999). Hence, these methods can mine inferences based on the connections within the user's need and the possible recommendation. **Hybrid** (Li and Kim, 2003) methods combine diverse individual methods among those noted earlier in order to handle the particular restrictions of an individual method.

2.2.2 Collaborative Filtering

Collaborative filtering (CF) is a recommender method used in almost all application domains. This method focuses on effective adoption of the user feedback (e.g., ratings) elicited from the users to make a profile of affinities. Such profiles are used to generate personalized recommendations. Hence, collaborative filtering relies on big data comprised of ratings acquired from typically big network of users (Desrosiers and Karypis, 2011). Using such data, collaborative filtering recommends items that a target user has not yet checked, but could probably like (Koren and Bell, 2011). Perhaps a cornerstone for these systems is to ability to estimate the feedback (or ratings) entered by users for items that they have not produced any rating for yet. Having the predicted ratings, collaborative