Book Review

Reviewed by Zhongxian Wang

Email: Wangz@mail.montclair.edu

Data Mining and Data Warehousing:
Principles and Practical Techniques
by: Parteek Bhatia
Published 2019
by Cambridge University Press
Cambridge CB2, UPH, Near Shaftesbury Road, UK, 477pp
ISBN: 978-1-108-72774-7 (Paperback)

According to the author, data is considered as the oil of this cyber world in the era of big data and business analysis. Data mining refers to extracting knowledge from large amounts of data, and to discover various types of patterns in the data that is useful for humans to interpret and use. This book intends training learners to be database mining experts in order to dig-up enormous potential to improve business bottom-line. The main strength of this textbook is the illustration of concepts with practical examples so that the learners can grasp the contents easily. Another striking feature of it is illustration of data mining algorithms with practical hands-on sessions on Waikato Environment for Knowledge Analysis (Weka) and R language. Learners can pursue a step-by-step self-practice in Weka and R.

There are 15 chapters. Chapter 1 introduces 'Beginning with machine learning'. The goal of machine learning (ML) is to continue to innovate artificial intelligence (AI), which is a machine that can function and think the same way as a human can. With the increase of web developers and computer scientists, coding and inventions are possible and easier than ever today. There are too many different application of ML in virtual personal assistants, traffic predictors, video surveillance, security, product recommendations, driverless vehicles, etc.

Chapter 2 starts with 'Introduction to data mining'. Nowadays, everything is shaped through data and technological advances. The importance of the internet and all the data being transmitted through databases is what makes up everything we consume daily. The applications of data mining are endless, ranging from credit card and loan approvals, to marketing techniques for companies, to financial risk management. Even watching a movie on the internet and using social media involves data mining. ML is the automation of a machine to act like a human. Data mining is the process of going through sums of data and breaking it down to a simpler way for humans to understand.

Chapter 3 works on 'Beginning with Weka and R language'. Weka, an open-source software, is a data mining software and it is a set of ML algorithms that can be applied to a dataset directly. It contains tools for data preprocessing, classification, regression, clustering, association rules and visualisation. R is a programming language for statistical computing and graphics. It is a well-developed, simple and effective programming

language which contains conditionals, loops, and other statistical functions. It has an effective data handling and storage facility, provides large, coherent, and integrated collection of tools for statistical functions, and provides graphical facilities for data analysis.

Chapter 4 deals with 'Data preprocessing'. It involves transforming raw data into an understandable structure which can then be used to come up with solutions. Data preprocessing goes as follows: data cleaning; data integration, data transformation and data reduction. Data cleaning is the process of cleansing raw data by filling missing values, smoothening noisy data, resolving any inconsistencies, and identifying and eliminating any outliers. Data integration is the process which combines data from several sources into a single data store. Data transformation consolidates and transforms data into a better-suited way for data mining. Finally, data reduction reduces the amount of data to interpret, while ensuring the quality and integrity of data is not diminished.

Chapter 5 focuses on 'Classification'. Classification is a classical method which is used by ML researchers and statisticians to predict the outcome of an unknown sample. It is widely used for the categorisation of objects into a given number of classes. There are two types of classification: posteriori and priori classification. Classification is a two-step process: the first step is training the model and the second step is testing the model for accuracy.

Chapter 6 presents issues in 'Implementing classification in Weka and R'. Many topics, such as building a decision tree classifier in Weka, applying naïve Bayes, creating the testing dataset, decision tree operation with R, naïve Bayes operation using R, etc. are handed out.

Chapter 7 demonstrates 'Cluster analysis'. Clustering is an unsupervised learning technique which does not require a labelled dataset. It is defined as grouping a set of related objects into classes or clusters. There are several applications that can be used with cluster analysis, such as marketing, insurance, biological studies, and web discovery, just to name a few. The preferred features of clustering are scalability, handling of noisy data, interpretability, and minimal user direction.

Chapter 8 concentrates 'Implementing clustering with Weka and R'. Handling missing values, results analysis after applying clustering, classification of unlabeled data, and clustering in R using simple k-means are the main topics.

Chapter 9 demonstrates 'Association mining'. Association rule mining can be explained as identifying recurrent patterns, correlations, associations, or causal structures among sets of objects or items in transactional databases, relational databases, and other information sources. Association rules are generally if/then statements that help in finding connections between seemingly unconnected data in a relational database or other information repository. They consist of two parts, an antecedent (if) and a consequent (then) and are often written as X-Y meaning that X and Y will coincidently appear. There are also three metrics that can evaluate the strength of association rules, which are support, confidence and lift.

Chapter 10 makes obvious 'Implementing association mining with Weka and R'. Rules generation like classifier using predictive apriori, generation of rules like classifier, application of association mining on numeric data in R have been instructed.

Chapter 11 shows 'Web mining and search engines'. Web content mining deals with extracting relevant knowledge from the contents of a web page. When you search on a web browser or search engine, the data goes through a complex process of analysing

large amounts of data from servers, and within milliseconds, sends you back the data you asked for. Search engines work by filtering keywords and relevance to websites to send you back information. Big technology companies like Google use algorithms to rank search results, they call it 'PageRank'. It helps to determine the relevance and quality of a web page.

Chapter 12 exhibits 'Data warehouse'. A data warehouse is a database and can be compared to as long-term memory for a business. They all have three fundamental components: load manager, warehouse manager and data access manager. The load manager is accountable for data collection from operational systems and functions as data conversion into operational forms. A warehouse manager is the main part of the data warehousing system as it holds the loads of amount of information from innumerable sources. Third, the query manager is that interface which links the end users with the information collected in data warehouses through specialised end-user tools.

Chapter 13 explains 'Data warehouse schema'. The term 'dimension' in data warehousing is a group of reference information about a measurable event. The events are warehoused in a fact table and are known as facts. A fact table can hold a dataset of facts at detailed or aggregated levels. They are defined as a group of supplementary data items and entails of dimensions and measurements. The star schema is one of the simplest of data warehouse schemas. Each element in the star schema represents a one-dimensional table only and the dimension table consists of a set of qualities.

Chapter 14 is 'Online analytical processing' (OLAP). Both OLAP and data mining categorise as business intelligence (BI) technologies. BI characterises the computer-based methods to extract and identify the important information from business data. Data mining is the field in computer science that handles mining important patterns from huge datasets. It is also a combination of techniques such as statistics, AI and database management. OLAP, additionally, is a collection of methods to query multiple dimensional databases. OLAP and data mining coincide with each other. OLAP can forecast values and summarise data, and data mining works at a more detailed level.

Chapter 15 expands 'Big data and NoSQL'. There are three types of data in the era of the internet boom: structured, semi-structured and unstructured data. Unstructured data has been rapidly growing and will continue to for many years to come. There is also an emergence of big data due to the internet boom, and the volume of data is rapidly changing as well. Now, it needs to have the speed to be delivered to customers, and it only takes milliseconds.

The authors emphasise that computer science will continue to be an integral part of our lives, and throughout the years it will continue to improve, especially as technology continues to innovate. This explains the importance of big data and data warehousing, and the world would not be the way it is today without it.

This extremely valuable textbook is written for undergraduate students of computer science engineering and information technology, which has brought together fundamental concepts of data mining and data warehousing in a single volume. Pedagogical features including unsolved problems and multiple-choice questions are interspersed throughout the book for better understanding. Practitioners also can benefit from its practical implementation using Weka and R language data mining tools.