Data Mining and Data Warehousing

This textbook is written to cater to the needs of undergraduate students of computer science, engineering, and information technology for a course on data mining and data warehousing. It brings together fundamental concepts of data mining and data warehousing in a single volume. Important topics including information theory, decision tree, Naïve Bayes classifier, distance metrics, partitioning clustering, associate mining, data marts and operational data store are discussed comprehensively. The text simplifies the understanding of the concepts through exercises and practical examples. Chapters such as classification, associate mining and cluster analysis are discussed in detail with their practical implementation using Weka and R language data mining tools. Advanced topics including big data analytics, relational data models, and NoSQL are discussed in detail. Unsolved problems and multiple-choice questions are interspersed throughout the book for better understanding.

Parteek Bhatia is Associate Professor in the Department of Computer Science and Engineering at the Thapar Institute of Engineering and Technology, Patiala, India. He has more than twenty years' teaching experience. His current research includes natural language processing, machine learning, and human–computer interface. He has taught courses including, data mining and data warehousing, big data analysis, and database management systems, at undergraduate and graduate levels.
Shaftesbury Road, Cambridge CB2 8EA, United Kingdom
One Liberty Plaza, 20th Floor, New York, NY 10006, USA
477 Williamstown Road, Port Melbourne, VIC 3207, Australia
314–321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre, New Delhi – 110025, India
103 Penang Road, #05–06/07, Visioncrest Commercial, Singapore 238467

Cambridge University Press is part of Cambridge University Press & Assessment, a department of the University of Cambridge.
We share the University’s mission to contribute to society through the pursuit of education, learning and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9781108727747

© Cambridge University Press & Assessment 2019

This publication is in copyright. Subject to statutory exception and to the provisions of relevant collective licensing agreements, no reproduction of any part may take place without the written permission of Cambridge University Press & Assessment.
First published 2019

A catalogue record for this publication is available from the British Library


Cambridge University Press & Assessment has no responsibility for the persistence or accuracy of URLs for external or third-party internet websites referred to in this publication and does not guarantee that any content on such websites is, or will remain, accurate or appropriate.
To

my parents, Mr Ved Kumar and Mrs Jagdish Bhatia
my supportive wife, Dr Sanmeet Kaur
loving sons, Rabat and Rishan
# Contents

*List of Figures*  
List of Tables  
Preface  
Acknowledgments  

## 1. Beginning with Machine Learning  
1.1 Introduction to Machine Learning 1  
1.2 Applications of Machine Learning 2  
1.3 Defining Machine Learning 5  
1.4 Classification of Machine Learning Algorithms  
  1.4.1 Supervised learning 5  
  1.4.2 Unsupervised learning 10  
  1.4.3 Supervised and unsupervised learning in real life scenario 12  
  1.4.4 Reinforcement learning 14  

## 2. Introduction to Data Mining  
2.1 Introduction to Data Mining 17  
2.2 Need of Data Mining 18  
2.3 What Can Data Mining Do and Not Do? 19  
2.4 Data Mining Applications 20  
2.5 Data Mining Process 21  
2.6 Data Mining Techniques  
  2.6.1 Predictive modeling 24  
  2.6.2 Database segmentation 24  
  2.6.3 Link analysis 24  
  2.6.4 Deviation detection 24  
  2.7 Difference between Data Mining and Machine Learning 25  

## 3. Beginning with Weka and R Language  
3.1 About Weka 28  
3.2 Installing Weka 29  
3.3 Understanding Fisher's Iris Flower Dataset 29  
3.4 Preparing the Dataset 31
3.5 Understanding ARFF (Attribute Relation File Format) 32
  3.5.1 ARFF header section 32
  3.5.2 ARFF data section 33
3.6 Working with a Dataset in Weka 33
  3.6.1 Removing input/output attributes 35
  3.6.2 Histogram 37
  3.6.3 Attribute statistics 39
  3.6.4 ARFF Viewer 40
  3.6.5 Visualizer 41
3.7 Introduction to R 42
  3.7.1 Features of R 42
  3.7.2 Installing R 43
3.8 Variable Assignment and Output Printing in R 44
3.9 Data Types 44
3.10 Basic Operators in R 45
  3.10.1 Arithmetic operators 46
  3.10.2 Relational operators 46
  3.10.3 Logical operators 47
  3.10.4 Assignment operators 47
3.11 Installing Packages 47
3.12 Loading of Data 49
  3.12.1 Working with the Iris dataset in R 50
4. Data Preprocessing 55
  4.1 Need for Data Preprocessing 55
  4.2 Data Preprocessing Methods 58
    4.2.1 Data cleaning 59
    4.2.2 Data integration 61
    4.2.3 Data transformation 61
    4.2.4 Data reduction 62
5. Classification 65
  5.1 Introduction to Classification 65
  5.2 Types of Classification 66
    5.2.1 Posteriori classification 66
    5.2.2 Priori classification 66
  5.3 Input and Output Attributes 66
  5.4 Working of Classification 67
  5.5 Guidelines for Size and Quality of the Training Dataset 69
  5.6 Introduction to the Decision Tree Classifier 69
    5.6.1 Building decision tree 70
    5.6.2 Concept of information theory 70
    5.6.3 Defining information in terms of probability 71
    5.6.4 Information gain 72
    5.6.5 Building a decision tree for the example dataset 73
5.6.6 Drawbacks of information gain theory 90
5.6.7 Split algorithm based on Gini Index 90
5.6.8 Building a decision tree with Gini Index 93
5.6.9 Advantages of the decision tree method 110
5.6.10 Disadvantages of the decision tree 110
5.7 Naïve Bayes Method 110
5.7.1 Applying Naïve Bayes classifier to the ‘Whether Play’ dataset 113
5.7.2 Working of Naïve Bayes classifier using the Laplace Estimator 117
5.8 Understanding Metrics to Assess the Quality of Classifiers 119
5.8.1 The boy who cried wolf 119
5.8.2 True positive 120
5.8.3 True negative 120
5.8.4 False positive 120
5.8.5 False negative 120
5.8.6 Confusion matrix 120
5.8.7 Precision 121
5.8.8 Recall 121
5.8.9 F-Measure 122
6. Implementing Classification in Weka and R 128
6.1 Building a Decision Tree Classifier in Weka 128
6.1.1 Steps to take when applying the decision tree classifier on the Iris dataset in Weka 130
6.1.2 Understanding the confusion matrix 136
6.1.3 Understanding the decision tree 136
6.1.4 Reading decision tree rules 138
6.1.5 Interpreting results 139
6.1.6 Using rules for prediction 139
6.2 Applying Naïve Bayes 139
6.3 Creating the Testing Dataset 142
6.4 Decision Tree Operation with R 148
6.5 Naïve Bayes Operation using R 151
7. Cluster Analysis 155
7.1 Introduction to Cluster Analysis 155
7.2 Applications of Cluster Analysis 156
7.3 Desired Features of Clustering 156
7.4 Distance Metrics 157
7.4.1 Euclidean distance 157
7.4.2 Manhattan distance 159
7.4.3 Chebyshev distance 160
7.5 Major Clustering Methods/Algorithms 161
7.6 Partitioning Clustering 162
7.6.1 k-means clustering 162
7.6.2 Starting values for the k-means algorithm 179
7.6.3 Issues with the k-means algorithm 179
7.6.4 Scaling and weighting 180

7.7 Hierarchical Clustering Algorithms (HCA) 181
7.7.1 Agglomerative clustering 182
7.7.2 Divisive clustering 195
7.7.3 Density-based clustering 199
7.7.4 DBSCAN algorithm 203
7.7.5 Strengths of DBSCAN algorithm 203
7.7.6 Weakness of DBSCAN algorithm 203

8. Implementing Clustering with Weka and R 206
8.1 Introduction 206
8.2 Clustering Fisher’s Iris Dataset with the Simple k-Means Algorithm 208
8.3 Handling Missing Values 209
8.4 Results Analysis after Applying Clustering 209
8.4.1 Identification of centroids for each cluster 213
8.4.2 Concept of within cluster sum of squared error 214
8.4.3 Identification of the optimum number of clusters using within cluster sum of squared error 215
8.5 Classification of Unlabeled Data 216
8.5.1 Adding clusters to dataset 216
8.5.2 Applying the classification algorithm by using added cluster attribute as class attribute 219
8.5.3 Pruning the decision tree 220
8.6 Clustering in R using Simple k-Means 221
8.6.1 Comparison of clustering results with the original dataset 224
8.6.2 Adding generated clusters to the original dataset 225
8.6.3 Apply J48 on the clustered dataset 225

9. Association Mining 229
9.1 Introduction to Association Rule Mining 229
9.2 Defining Association Rule Mining 232
9.3 Representations of Items for Association Mining 233
9.4 The Metrics to Evaluate the Strength of Association Rules 234
9.4.1 Support 234
9.4.2 Confidence 235
9.4.3 Lift 237
9.5 The Naïve Algorithm for Finding Association Rules 240
9.5.1 Working of the Naïve algorithm 240
9.5.2 Limitations of the Naïve algorithm 242
9.5.3 Improved Naïve algorithm to deal with larger datasets 242
9.6 Approaches for Transaction Database Storage 243
9.6.1 Simple transaction storage 244
9.6.2 Horizontal storage 244
9.6.3 Vertical representation 245
Contents

9.7 The Apriori Algorithm 246
  9.7.1 About the inventors of Apriori 246
  9.7.2 Working of the Apriori algorithm 247
9.8 Closed and Maximal Itemsets 280
9.9 The Apriori–TID Algorithm for Generating Association Mining Rules 282
9.10 Direct Hashing and Pruning (DHP) 285
9.11 Dynamic Itemset Counting (DIC) 297
9.12 Mining Frequent Patterns without Candidate Generation (FP Growth) 301
  9.12.1 Advantages of the FP-tree approach 314
  9.12.2 Further improvements of FP growth 314
10. Implementing Association Mining with Weka and R 319
  10.1 Association Mining with Weka 319
  10.2 Applying Predictive Apriori in Weka 321
  10.3 Rules Generation Similar to Classifier Using Predictive Apriori 325
  10.4 Comparison of Association Mining CAR Rules with J48 Classifier Rules 327
  10.5 Applying the Apriori Algorithm in Weka 330
  10.6 Applying the Apriori Algorithm in Weka on a Real World Dataset 333
  10.7 Applying the Apriori Algorithm in Weka on a Real World Larger Dataset 339
  10.8 Applying the Apriori Algorithm on a Numeric Dataset 344
  10.9 Process of Performing Manual Discretization 351
  10.10 Applying Association Mining in R 357
  10.11 Implementing Apriori Algorithm 357
  10.12 Generation of Rules Similar to Classifier 359
  10.13 Comparison of Association Mining CAR Rules with J48 Classifier Rules 360
  10.14 Application of Association Mining on Numeric Data in R 362
11. Web Mining and Search Engines 368
  11.1 Introduction 368
  11.2 Web Content Mining
    11.2.1 Web document clustering 369
    11.2.2 Suffix Tree Clustering (STC) 369
    11.2.3 Resemblance and containment 370
    11.2.4 Fingerprinting 371
  11.3 Web Usage Mining 371
  11.4 Web Structure Mining
    11.4.1 Hyperlink Induced Topic Search (HITS) algorithm 372
  11.5 Introduction to Modern Search Engines 375
  11.6 Working of a Search Engine
    11.6.1 Web crawler 377
    11.6.2 Indexer 377
    11.6.3 Query processor 378
  11.7 PageRank Algorithm 379
  11.8 Precision and Recall 385
12. Data Warehouse 388
12.1 The Need for an Operational Data Store (ODS) 388
12.2 Operational Data Store 389
12.2.1 Types of ODS 390
12.2.2 Architecture of ODS 391
12.2.3 Advantages of the ODS 393
12.3 Data Warehouse 393
12.3.1 Historical developments in data warehousing 394
12.3.2 Defining data warehousing 395
12.3.3 Data warehouse architecture 395
12.3.4 Benefits of data warehousing 397
12.4 Data Marts 398
12.5 Comparative Study of Data Warehouse with OLTP and ODS 401
12.5.1 Data warehouses versus OLTP: similarities and distinction 401

13. Data Warehouse Schema 405
13.1 Introduction to Data Warehouse Schema 405
13.1.1 Dimension 405
13.1.2 Measure 407
13.1.3 Fact Table 407
13.1.4 Multi-dimensional view of data 408
13.2 Star Schema 408
13.3 Snowflake Schema 410
13.4 Fact Constellation Schema (Galaxy Schema) 412
13.5 Comparison among Star, Snowflake and Fact Constellation Schema 413

14. Online Analytical Processing 416
14.1 Introduction to Online Analytical Processing 416
14.1.1 Defining OLAP 417
14.1.2 OLAP applications 417
14.1.3 Features of OLAP 417
14.1.4 OLAP Benefits 418
14.1.5 Strengths of OLAP 418
14.1.6 Comparison between OLTP and OLAP 418
14.1.7 Differences between OLAP and data mining 419
14.2 Representation of Multi-dimensional Data 420
14.2.1 Data Cube 421
14.3 Implementing Multi-dimensional View of Data in Oracle 423
14.4 Improving efficiency of OLAP by pre-computing the queries 427
14.5 Types of OLAP Servers 429
14.5.1 Relational OLAP 430
14.5.2 MOLAP 431
14.5.3 Comparison of ROLAP and MOLAP 432
14.6 OLAP Operations 433
14.6.1 Roll-up 433
14.6.2 Drill-down 433
14.6.3 Slice and dice 435
14.6.4 Dice 437
14.6.5 Pivot 438

15. Big Data and NoSQL 442
15.1 The Rise of Relational Databases 442
15.2 Major Issues with Relational Databases 443
15.3 Challenges from the Internet Boom 445
  15.3.1 The rapid growth of unstructured data 445
  15.3.2 Types of data in the era of the Internet boom 445
15.4 Emergence of Big Data due to the Internet Boom 448
15.5 Possible Solutions to Handle Huge Amount of Data 449
15.6 The Emergence of Technologies for Cluster Environment 451
15.7 Birth of NoSQL 452
15.8 Defining NoSQL from the Characteristics it Shares 453
15.9 Some Misconceptions about NoSQL 453
15.10 Data Models of NoSQL 453
  15.10.1 Key-value data model 454
  15.10.2 Column-family data model 456
  15.10.3 Document data model 457
  15.10.4 Graph databases 459
15.11 Consistency in a Distributed Environment 461
15.12 CAP Theorem 461
15.13 Future of NoSQL 462
15.14 Difference between NoSQL and Relational Data Models (RDBMS) 464

Index 467

Colour Plates 469
Figures

1.1 Classification of machine learning algorithms 5
1.2 Data plot for size of plot and cost 6
1.3 Estimation (prediction) of cost of house with a small dataset 6
1.4 Prediction of cost of house with large dataset 7
1.5 Data plot for tumor size and malignancy 7
1.6 Prediction about a tumor of size A 8
1.7 Considering tumor size and age as features for classification 8
1.8 Prediction for a tumor of size B 9
1.9 Prediction for tumor size B being benign 9
1.10 Google news 11
1.11 Applications of unsupervised learning 11

2.1 Per minute generation of data over the Internet according to a 2017 report 18
2.2 Data mining process 22

3.1 Downloading Weka 29
3.2 Downloading the Iris dataset 30
3.3 Sample of the Iris flower 30
3.4 Sample of Fisher’s dataset 31
3.5 Save as ‘Other Format’ 31
3.6 ARFF format of IRIS dataset 32
3.7 Weka GUI Chooser screen 34
3.8 Weka Explorer screen 34
3.9 Loading Fisher’s dataset 35
3.10 Fisher’s dataset after removal of instance number 36
3.11 Elements of the Explorer screen 36
3.12 Expansion of class designator 37
3.13 Histogram for Petal width 38
3.14 Histograms for all attributes of Iris dataset 38
3.15 Attribute statistics 39
3.16 Distinct and Unique values 40
3.17 (a) Selecting ARFF Viewer from GUI Chooser and (b) opening the file in ARFF Viewer 40
3.18 ARFF Viewer of Fisher’s dataset 41
Figures

3.19 Visualization of dataset 41
3.20 Plotting of dataset 42
3.21 Screenshot of download link for R 43
3.22 Console screen of R 43
3.23 Basic syntax in R 44
3.24 Data type of a variable 45
3.25 Screenshot of basic arithmetic operators 46
3.26 Relational operators in R 46
3.27 Working of logical operators 47
3.28 Checking of already installed packages 47
3.29 Installation of a new package 48
3.30 Console after successful installation of package 48
3.31 Attribute names of a dataset 50
3.32 Statistics of Iris dataset 50
3.33 Viewing of dataset 51
3.34 Identification of unique and missing values for Sepal width 51
3.35 Plotting the Iris dataset 52
3.36 Plotting between Petal width and Petal length 52
3.37 Histogram for Sepal width 53
4.1 Various stages of preprocessing 58
4.2 Chemical composition of wine samples 62
5.1 Input and output attributes 67
5.2 Training and testing of the classifier 68
5.3 Building a classifier to approve or reject loan applications 68
5.4 Predicting the type of customer based on trained classifier 68
5.5 Training and testing of the classifier 69
5.6 Decision tree to predict whether a customer will buy a laptop or not 70
5.7 Dataset for class C prediction based on given attribute condition 73
5.8 Data splitting based on Y attribute 75
5.9 Decision tree after splitting of attribute Y having value ‘1’ 76
5.10 Decision tree after splitting of attribute Y value ‘0’ 76
5.11 Dataset for play prediction based on given day weather conditions 77
5.12 Selection of Outlook as root attribute 80
5.13 Data splitting based on the Outlook attribute 81
5.14 Humidity attribute is selected from dataset of Sunny instances 84
5.15 Decision tree after splitting of data on Humidity attribute 85
5.16 Decision tree after analysis of Sunny and Overcast dataset 85
5.17 Decision tree after analysis of Sunny, Overcast and Rainy dataset 89
5.18 Final decision tree after analysis of Sunny, Overcast and Rainy dataset 89
5.19 Prediction of Play for an unknown instance 90
5.20 Gini Index representing perfect equality 91
5.21 Lorenz curve 92
5.22 Lorenz curves with varying income distributions 92
5.23 Dataset for class C prediction based on given attribute condition 94
5.24 Data splitting based on Y attribute 96
### Figures xvii

<table>
<thead>
<tr>
<th>Number</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.25</td>
<td>Decision tree after splitting of attribute Y having value ‘1’</td>
<td>96</td>
</tr>
<tr>
<td>5.26</td>
<td>Decision tree after splitting of attribute Y value ‘0’</td>
<td>97</td>
</tr>
<tr>
<td>5.27</td>
<td>Dataset for play prediction based on given day weather conditions</td>
<td>98</td>
</tr>
<tr>
<td>5.28</td>
<td>Selection of Outlook as root attribute</td>
<td>101</td>
</tr>
<tr>
<td>5.29</td>
<td>Data splitting based on Outlook attribute</td>
<td>101</td>
</tr>
<tr>
<td>5.30</td>
<td>Humidity attribute is selected from dataset of Sunny instances</td>
<td>104</td>
</tr>
<tr>
<td>5.31</td>
<td>Decision tree after splitting data on the Humidity attribute</td>
<td>105</td>
</tr>
<tr>
<td>5.32</td>
<td>Decision tree after analysis of Sunny and Overcast datasets</td>
<td>105</td>
</tr>
<tr>
<td>5.33</td>
<td>Decision tree after analysis of Sunny, Overcast and Rainy datasets</td>
<td>108</td>
</tr>
<tr>
<td>5.34</td>
<td>Final decision tree after analysis of Sunny, Overcast and Rainy datasets</td>
<td>109</td>
</tr>
<tr>
<td>5.35</td>
<td>Prediction of play for unknown instance</td>
<td>109</td>
</tr>
<tr>
<td>5.36</td>
<td>Dataset for play prediction based on a given day’s weather conditions</td>
<td>114</td>
</tr>
<tr>
<td>5.37</td>
<td>Probability of whether play will be held or not on a Sunny day</td>
<td>114</td>
</tr>
<tr>
<td>5.38</td>
<td>Summarization of count calculations of all input attributes</td>
<td>115</td>
</tr>
<tr>
<td>5.39</td>
<td>Probability of play held or not for each value of attribute</td>
<td>115</td>
</tr>
<tr>
<td>5.40</td>
<td>Probability for play ‘Yes’ for an unknown instance</td>
<td>116</td>
</tr>
<tr>
<td>5.41</td>
<td>Probability for play ‘No’ for an unknown instance</td>
<td>117</td>
</tr>
<tr>
<td>5.42</td>
<td>Probability of play not being held when outlook is overcast</td>
<td>117</td>
</tr>
<tr>
<td>5.43</td>
<td>Values of attributes after adding Laplace estimator</td>
<td>118</td>
</tr>
<tr>
<td>5.44</td>
<td>Probability of play held or not for each modified value of attribute</td>
<td>118</td>
</tr>
<tr>
<td>5.45</td>
<td>Attribute values for given example instance</td>
<td>118</td>
</tr>
<tr>
<td>5.46</td>
<td>Confusion matrix for bird classifier</td>
<td>119</td>
</tr>
<tr>
<td>5.47</td>
<td>Confusion matrix for tumor prediction</td>
<td>120</td>
</tr>
<tr>
<td>6.1</td>
<td>Classification using Weka's decision tree</td>
<td>128</td>
</tr>
<tr>
<td>6.2</td>
<td>Classification of an unknown sample using Weka's decision tree</td>
<td>129</td>
</tr>
<tr>
<td>6.3</td>
<td>Working of the decision tree</td>
<td>130</td>
</tr>
<tr>
<td>6.4</td>
<td>Loading the iris.arff file</td>
<td>130</td>
</tr>
<tr>
<td>6.5</td>
<td>Selecting Weka J48 algorithm</td>
<td>131</td>
</tr>
<tr>
<td>6.6</td>
<td>Selection of the Weka J48 algorithm</td>
<td>131</td>
</tr>
<tr>
<td>6.7</td>
<td>Selection of percentage split test option</td>
<td>132</td>
</tr>
<tr>
<td>6.8</td>
<td>Saving output predictions</td>
<td>132</td>
</tr>
<tr>
<td>6.9</td>
<td>Original Fisher’s Iris dataset</td>
<td>133</td>
</tr>
<tr>
<td>6.10</td>
<td>Building the decision tree</td>
<td>134</td>
</tr>
<tr>
<td>6.11</td>
<td>Decision tree accuracy statistics</td>
<td>134</td>
</tr>
<tr>
<td>6.12</td>
<td>Visualization of the tree</td>
<td>135</td>
</tr>
<tr>
<td>6.13</td>
<td>Decision tree for the Iris dataset</td>
<td>135</td>
</tr>
<tr>
<td>6.14</td>
<td>Confusion matrix</td>
<td>136</td>
</tr>
<tr>
<td>6.15</td>
<td>Decision tree showing condition for Setosa</td>
<td>136</td>
</tr>
<tr>
<td>6.16</td>
<td>Decision tree showing conditions for Virginica</td>
<td>137</td>
</tr>
<tr>
<td>6.17</td>
<td>Decision tree showing condition for Versicolor</td>
<td>137</td>
</tr>
<tr>
<td>6.18</td>
<td>Rules identified by the decision tree</td>
<td>138</td>
</tr>
<tr>
<td>6.19</td>
<td>Classification of an unknown sample according to decision tree rules</td>
<td>138</td>
</tr>
<tr>
<td>6.20</td>
<td>Size and leaves of the tree</td>
<td>139</td>
</tr>
<tr>
<td>6.21</td>
<td>Selecting dataset file</td>
<td>140</td>
</tr>
<tr>
<td>6.22</td>
<td>Selecting the classifier and setting classifier evaluation options</td>
<td>140</td>
</tr>
</tbody>
</table>
8.1 Classification of an unknown sample 207
8.2 Clustering 207
8.3 Clustering process 208
8.4 Applying the simple k-means algorithm in Weka 208
8.5 Applying the simple k-means algorithm in Weka: the next step 209
8.6 Clustering of Iris samples 210
8.7 Class to cluster evaluation and confusion matrix 210
8.8 Cluster visualization 211
8.9 Cluster visualization for Petal length vs. Petal width 212
8.10 Cluster visualization with respect to Petal length vs. Petal width 212
8.11 Cluster visualization with respect to Sepal length vs. Sepal width 213
8.12 Cluster visualization with respect to Sepal length vs. Sepal width 213
8.13 Cluster visualization with respect to Sepal length vs. Sepal width 214
8.14 Within cluster sum of squared error 214
8.15 Error vs. number of clusters 215
8.16 Classification process of unlabeled data 216
8.17 Choosing AddCluster filter 217
8.18 Configuration settings of AddCluster filter 217
8.19 Application of AddCluster filter 218
8.20 Comparison of values of the new added cluster attribute with the already existing class column 218
8.21 Prediction rules generated by J48 219
8.22 Decision tree 219
8.23 Comparison of rules of clustering with rules of the decision tree 220
8.24 Pruning the decision tree 220
8.25 Analysis of rules after increasing minNumObj 221
8.26 Iris dataset statistics 221
8.27 Iris dataframe statistics 222
8.28 Iris dataframe statistics after removal of species variable 222
8.29 Results after applying k-means clustering 223
8.30 Cluster size 223
8.31 Cluster centroids 224
8.32 Plot of Petal length vs. Petal width after clustering 224
8.33 Confusion matrix 225
8.34 Iris dataset after adding results of the clustering analysis 225
8.35 Apply decision tree on clustered results by simple k-means algorithm 226
9.1 Need for association mining 230
9.2 Association of sale of beer and diapers 230
9.3 Association of sale of beer and diapers 231
9.4 Association of sale of beer and diapers 231
9.5 Association and customer purchase bills 232
9.6 Representation of association rules 239
9.7 Process for identification of frequent itemsets 252
9.8 C1, candidate 1-itemset and their count 256
9.9 L1, frequent 1-itemset 257
**Figures**

9.10 Lattice structure of frequent itemsets 260
9.11 Generation of C2 and L2 261
9.12 Generation of C3 and L3 263
9.13 Illustration of closed and maximal frequent itemsets 280
9.14 Process by the Apriori algorithm method 292
9.15 Process by the DHP algorithm method 293
9.16 Identifying frequent item pairs and groups by Apriori algorithm 295
9.17 FP-tree for first transaction, i.e., 100 only 303
9.18 FP-tree for the first two transactions 303
9.19 FP-tree for first three transactions 304
9.20 FP-tree for first four transactions 304
9.21 The final FP-tree for the example database after five transactions 304
9.22 Illustrating the step-by-step creation of the FP-tree 308
9.23 Final FP-tree for database given in Table 9.105 308
9.24 Illustrating the step-by-step creation of the FP-tree 312
9.25 FP-tree for the example database 312
10.1 Snapshot of the ‘play-or-no-play’ dataset 319
10.2 Associations between items 320
10.3 Working of Predictive Apriori 320
10.4 Loading the weather.nominal.arff 321
10.5 Selecting the Predictive Apriori algorithm for association mining 322
10.6 Changing parameters for Predictive Apriori 322
10.7 Parameters of the Predictive Apriori algorithm 323
10.8 Association mining rules 324
10.9 Analysis of rule 2 325
10.10 Setting CAR to true for getting class association rules 326
10.11 Application of the J48 algorithm on dataset ‘play-or-no-play’ 327
10.12 Selection of use training set to build the model 328
10.13 Select the ‘Visualize tree’ to get a decision tree 328
10.14 Decision tree for dataset ‘play-or-no-play’ 329
10.15 Selection of the Apriori algorithm 330
10.16 Generic Object Editor to change the default values of the Apriori algorithm 331
10.17 Default values of the properties of the Apriori algorithm 332
10.18 Daily item dataset 334
10.19 Saving the file in CSV format 334
10.20 Weka GUI Chooser Panel 335
10.21 Dataset uploaded in Weka 335
10.22 Choosing numeric to nominal filter in Weka 336
10.23 Changing from numeric to nominal filter 336
10.24 Removing the Transaction attribute 337
10.25 Applying the Apriori algorithm 337
10.26 Opening the Generic Object Editor 338
10.27 Starting the Apriori algorithm 338
10.28 Results after running the Apriori algorithm 339
10.29 Saving the file in CSV format 340
xxii  Figures

11.12  Four web pages with hyperlinks 380
11.13  Five web pages with hyperlinks 382
11.14  Venn diagram showing relevant and retrieved results 385
12.1  Architecture of an Operation Data Store 391
12.2  Relationship between OLTP, ODS and data warehouse systems 393
12.3  Answering management queries 394
12.4  Historical developments of data warehouse 394
12.5  Architecture of a data warehouse 396
12.6  Limitations of data warehousing 399
12.7  Data mart and data warehouse 399
12.8  Relationship between data mart and data warehouse 400
13.1  (a) location dimension, (b) item dimension 406
13.2  Normalized view 406
13.3  Representation of fact and dimension tables 407
13.4  The sales fact table 407
13.5  Graphical representation of Star schema 408
13.6  Star schema for analysis of sales 409
13.7  Snowflake schema for analysis of sales 410
13.8  Snowflake schema 411
13.9  Fact constellation schema for analysis of sales 412
14.1  (a) Relational model, (b) Two dimensional view 420
14.2  (a) Relational model representation, (b) Three dimensional view 421
14.3  Two dimensional view of sale data, i.e., Item and Time 422
14.4  Three dimensional view of sale data, i.e., Item, Time and Location 422
14.5  Cubical three dimensional view of sale data 423
14.6  Emp database 423
14.7  Total number of employees in each job within each department 424
14.8  Two dimensional view of employee data 424
14.9  Use of ROLLUP for aggregation of data 425
14.10  Use of CUBE for aggregation of data 425
14.11  Employee database with third dimension state 426
14.12  Three dimensional view of the employee database 426
14.13  Cubical three dimensional representation of the employee database 427
14.14  Pre-computation of only queries of type (d, j, s) 429
14.15  No zero value facts returned by Group By query 430
14.16  ROLAP architecture 430
14.17  MOLAP Implementation 431
14.18  MOLAP architecture 432
14.19  Working of the Roll-up operation 434
14.20  Working of the Drill-down operation 434
14.21  Working of the Slice operation 435
14.22  The Slice operation 436
14.23  Working of the Dice operation 437
14.24  Dice operation 438
14.25  Workings of the Pivot operation 439
Figures xxiii

15.1 Rise of the relational model
15.2 An order, which looks like a single aggregate structure in the UI, is split into many rows from many tables in a relational database
15.3 Rise of Object databases
15.4 Relational dominance in the late 1990s and early 2000s
15.5 Generation of lots of traffic during the internet boom
15.6 Types of data
15.7 Structured data
15.8 Percentage distribution of different types of data
15.9 Info graphic of 4 V's of big data
15.10 Scaling up using a large centralized server
15.11 Handling of huge data volume through the relational model
15.12 Cluster computing emerged as a winner
15.13 SQL in cluster environment
15.14 BigTable and Dynamo for cluster environments
15.15 NoSQL Meet
15.16 Participants of NoSQL meet
15.17 Features of NoSQL
15.18 NoSQL data models
15.19 Key-value data model
15.20 Key-value model to store account information
15.21 Column-family data model
15.22 Representing customer information in a Column-family structure
15.23 Document model
15.24 Document data model
15.25 NoSQL data models
15.26 An example of the Graph structure
15.27 RDBMS versus NoSQL
15.28 CAP theorem
15.29 Consistency in a distributed environment
15.30 Future of NoSQL?
15.31 Strengths of NoSQL
15.32 The future is Polyglot persistence
15.33 Polyglot persistence in a real environment
## Tables

1.1 Fruit data for supervised learning  
1.2 Fruit data for unsupervised learning  
2.1 Tabular comparison of data mining and machine learning  
3.1 WEKA GUI applications  
3.2 Description about basic data types  
3.3 Summary about basic operators of R  
3.4 Some of the important machine learning packages  
4.1 Vendor's record extracted from the first source system  
4.2 Vendor's record extracted from the second source system by Supplier ID  
4.3 Vendor's record extracted from the third source system  
4.4 Vendor's record after pre-processing  
5.1 Information and Gini Index for a number of events  
6.1 Iris dataset sample  
7.1 Data to calculate Euclidean distances among three persons  
7.2 Database for the k-means algorithm example  
7.3 Database after first iteration  
7.4 Database after the second iteration  
7.5 Database after the second iteration  
7.6 Initial dataset for k = 3  
7.7 Final assigned cluster for k = 3  
7.8 Dataset after first iteration  
7.9 Dataset after second iteration  
7.10 Dataset after third iteration  
7.11 Dataset after fourth iteration  
7.12 Record of students' performance  
7.13 Seed records  
7.14 First iteration-allocation of each object to its nearest cluster  
7.15 Updated centroids after first iteration  
7.16 Second iteration-allocation of each object to its nearest cluster
### Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.17</td>
<td>Final allocation</td>
<td>175</td>
</tr>
<tr>
<td>7.18</td>
<td>Within (intra) cluster and between (inter) clusters distance</td>
<td>176</td>
</tr>
<tr>
<td>7.19</td>
<td>Calculations for within-cluster and between-cluster variance using Euclidean distance</td>
<td>176</td>
</tr>
<tr>
<td>7.20</td>
<td>Chemical composition of wine samples</td>
<td>180</td>
</tr>
<tr>
<td>7.21</td>
<td>Input distance matrix ($L = 0$ for all the clusters)</td>
<td>185</td>
</tr>
<tr>
<td>7.22</td>
<td>Input distance matrix, with $m$: 1</td>
<td>185</td>
</tr>
<tr>
<td>7.23</td>
<td>Input distance matrix, with $m$: 2</td>
<td>186</td>
</tr>
<tr>
<td>7.24</td>
<td>Input distance matrix, with $m$: 3</td>
<td>186</td>
</tr>
<tr>
<td>7.25</td>
<td>Input distance matrix, with $m$: 4</td>
<td>186</td>
</tr>
<tr>
<td>7.26</td>
<td>Record of students’ performance</td>
<td>187</td>
</tr>
<tr>
<td>7.27</td>
<td>Distance matrix at $m$: 0</td>
<td>188</td>
</tr>
<tr>
<td>7.28</td>
<td>Cells involved in $C_1$</td>
<td>189</td>
</tr>
<tr>
<td>7.29</td>
<td>Input distance matrix, with $m$: 2</td>
<td>189</td>
</tr>
<tr>
<td>7.30</td>
<td>Cells involved in $C_2$</td>
<td>190</td>
</tr>
<tr>
<td>7.31</td>
<td>Input distance matrix, with $m$: 3</td>
<td>190</td>
</tr>
<tr>
<td>7.32</td>
<td>Cells involved in creating $C_3$</td>
<td>191</td>
</tr>
<tr>
<td>7.33</td>
<td>Input distance matrix, with $m$: 4</td>
<td>191</td>
</tr>
<tr>
<td>7.34</td>
<td>Cells involved in creating $C_4$</td>
<td>192</td>
</tr>
<tr>
<td>7.35</td>
<td>Input distance matrix, with $m$: 5</td>
<td>192</td>
</tr>
<tr>
<td>7.36</td>
<td>Cells involved in creating $C_5$</td>
<td>192</td>
</tr>
<tr>
<td>7.37</td>
<td>Input distance matrix, with $m$: 6</td>
<td>193</td>
</tr>
<tr>
<td>7.38</td>
<td>Cells involved in creating $C_6$</td>
<td>193</td>
</tr>
<tr>
<td>7.39</td>
<td>Input distance matrix, with $m$: 7</td>
<td>193</td>
</tr>
<tr>
<td>7.40</td>
<td>Cells involved in creating $C_7$</td>
<td>194</td>
</tr>
<tr>
<td>7.41</td>
<td>Input distance matrix, with $m$: 8</td>
<td>194</td>
</tr>
<tr>
<td>7.42</td>
<td>Cells involved in creating $C_8$</td>
<td>194</td>
</tr>
<tr>
<td>7.43</td>
<td>Input distance matrix, with $m$: 9</td>
<td>194</td>
</tr>
<tr>
<td>7.44</td>
<td>Record of students’ performance</td>
<td>195</td>
</tr>
<tr>
<td>7.45</td>
<td>Distance matrix at $m$: 0</td>
<td>196</td>
</tr>
<tr>
<td>7.46</td>
<td>Distance matrix for cluster $C_1$</td>
<td>198</td>
</tr>
<tr>
<td>7.47</td>
<td>Splitting of cluster $C_1$ into two new clusters of $S_7$ and $S_8$</td>
<td>198</td>
</tr>
<tr>
<td>7.48</td>
<td>Distance matrix for cluster $C_2$</td>
<td>198</td>
</tr>
<tr>
<td>7.49</td>
<td>Splitting of cluster $C_2$ into two new clusters of $S_3$ and $S_9$</td>
<td>199</td>
</tr>
<tr>
<td>7.50</td>
<td>Distance matrix for cluster $C_4$</td>
<td>199</td>
</tr>
<tr>
<td>7.51</td>
<td>Splitting of cluster $C_4$ into two new clusters of $S_6$ and $S_8$</td>
<td>199</td>
</tr>
<tr>
<td>9.1</td>
<td>Sale database</td>
<td>233</td>
</tr>
<tr>
<td>9.2</td>
<td>Sale database</td>
<td>234</td>
</tr>
<tr>
<td>9.3</td>
<td>Example of the support measure</td>
<td>235</td>
</tr>
<tr>
<td>9.4</td>
<td>Example of the confidence measure</td>
<td>236</td>
</tr>
<tr>
<td>9.5</td>
<td>Database for identification of association rules</td>
<td>236</td>
</tr>
<tr>
<td>9.6</td>
<td>Dataset</td>
<td>238</td>
</tr>
<tr>
<td>9.7</td>
<td>Modified dataset</td>
<td>239</td>
</tr>
<tr>
<td>9.8</td>
<td>Sale record of grocery store</td>
<td>240</td>
</tr>
<tr>
<td>9.9</td>
<td>List of all itemsets and their frequencies</td>
<td>240</td>
</tr>
<tr>
<td>9.10</td>
<td>The set of all frequent items</td>
<td>241</td>
</tr>
<tr>
<td>Table</td>
<td>Description</td>
<td>Page</td>
</tr>
<tr>
<td>-------</td>
<td>-----------------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>9.11</td>
<td>All possible combinations with nonzero frequencies</td>
<td>242</td>
</tr>
<tr>
<td>9.12</td>
<td>Frequencies of all itemsets with nonzero frequencies</td>
<td>243</td>
</tr>
<tr>
<td>9.13</td>
<td>A simple representation of transactions as an item list</td>
<td>244</td>
</tr>
<tr>
<td>9.14</td>
<td>Horizontal storage representation</td>
<td>244</td>
</tr>
<tr>
<td>9.15</td>
<td>Vertical storage representation</td>
<td>245</td>
</tr>
<tr>
<td>9.16</td>
<td>Frequency of item pairs</td>
<td>245</td>
</tr>
<tr>
<td>9.17</td>
<td>Transactions database</td>
<td>247</td>
</tr>
<tr>
<td>9.18</td>
<td>Candidate one itemsets C1</td>
<td>248</td>
</tr>
<tr>
<td>9.19</td>
<td>Frequent items L1</td>
<td>248</td>
</tr>
<tr>
<td>9.20</td>
<td>Candidate item pairs C2</td>
<td>249</td>
</tr>
<tr>
<td>9.21</td>
<td>Frequent two item pairs L2</td>
<td>250</td>
</tr>
<tr>
<td>9.22</td>
<td>L1 for generation of C2 having only one element in each list</td>
<td>251</td>
</tr>
<tr>
<td>9.23</td>
<td>L2 for generation of C3 (i.e., K=3) having two elements in each list</td>
<td>251</td>
</tr>
<tr>
<td>9.24</td>
<td>L3 for generation of C4 (i.e., K = 4) having three elements in each list</td>
<td>251</td>
</tr>
<tr>
<td>9.25</td>
<td>L1</td>
<td>253</td>
</tr>
<tr>
<td>9.26</td>
<td>L2</td>
<td>253</td>
</tr>
<tr>
<td>9.27</td>
<td>L3</td>
<td>253</td>
</tr>
<tr>
<td>9.28</td>
<td>L1</td>
<td>254</td>
</tr>
<tr>
<td>9.29</td>
<td>Generation of C2</td>
<td>254</td>
</tr>
<tr>
<td>9.30</td>
<td>Generated C2</td>
<td>255</td>
</tr>
<tr>
<td>9.31</td>
<td>L2</td>
<td>255</td>
</tr>
<tr>
<td>9.32</td>
<td>C3</td>
<td>255</td>
</tr>
<tr>
<td>9.33</td>
<td>L3</td>
<td>256</td>
</tr>
<tr>
<td>9.34</td>
<td>C4</td>
<td>256</td>
</tr>
<tr>
<td>9.35</td>
<td>Transaction database</td>
<td>256</td>
</tr>
<tr>
<td>9.36</td>
<td>Generation of C2</td>
<td>257</td>
</tr>
<tr>
<td>9.37</td>
<td>Generation L2</td>
<td>257</td>
</tr>
<tr>
<td>9.38</td>
<td>Generation of C3</td>
<td>257</td>
</tr>
<tr>
<td>9.39</td>
<td>Calculation of confidence</td>
<td>259</td>
</tr>
<tr>
<td>9.40</td>
<td>Transaction database for identification of association rules</td>
<td>261</td>
</tr>
<tr>
<td>9.41</td>
<td>C1</td>
<td>261</td>
</tr>
<tr>
<td>9.42</td>
<td>Generation of C3</td>
<td>262</td>
</tr>
<tr>
<td>9.43</td>
<td>Pruning of candidate itemset C3</td>
<td>262</td>
</tr>
<tr>
<td>9.44</td>
<td>Pruned C3</td>
<td>263</td>
</tr>
<tr>
<td>9.45</td>
<td>C4</td>
<td>263</td>
</tr>
<tr>
<td>9.46</td>
<td>Pruned C4</td>
<td>263</td>
</tr>
<tr>
<td>9.47</td>
<td>Generation of association rules</td>
<td>265</td>
</tr>
<tr>
<td>9.48</td>
<td>Frequent 2-itemsets, i.e., L2</td>
<td>266</td>
</tr>
<tr>
<td>9.49</td>
<td>List of grocery items</td>
<td>267</td>
</tr>
<tr>
<td>9.50</td>
<td>Transaction data</td>
<td>268</td>
</tr>
<tr>
<td>9.51</td>
<td>Frequency count for all items</td>
<td>269</td>
</tr>
<tr>
<td>9.52</td>
<td>The frequent 1-itemset or L1</td>
<td>269</td>
</tr>
<tr>
<td>9.53</td>
<td>The 21 candidate 2-itemsets or C2</td>
<td>270</td>
</tr>
<tr>
<td>9.54</td>
<td>Frequency count of candidate 2-itemsets</td>
<td>271</td>
</tr>
<tr>
<td>9.55</td>
<td>The frequent 2-itemsets or L2</td>
<td>272</td>
</tr>
<tr>
<td>9.56</td>
<td>Candidate 3-itemsets or C3</td>
<td>272</td>
</tr>
</tbody>
</table>
xxviii Tables

9.57 Pruning of candidate itemset C3 273
9.58 Pruned candidate itemset C3 273
9.59 Candidate 3-itemsets or C3 and their frequencies 273
9.60 The frequent 3-itemsets or L3 274
9.61 Confidence of association rules from [Bournvita, Butter, Cornflakes] 274
9.62 Confidence of association rules from [Bournvita, Bread] 275
9.63 Identified rules from [Bournvita, Butter, Cornflakes] having confidence more than 70% 275
9.64 List of all possible rules from rules given in Table 9.61 275
9.65 Confidence of association rules from [Coffee, Chocolate, Eggs] 277
9.66 List of all possible rules from rules given in Table 9.65 278
9.67 All association rules for the given database 279
9.68 A transaction database to illustrate closed and maximal itemsets 280
9.69 Frequent itemsets for the database in Table 9.68 281
9.70 Transaction database 283
9.71 Transaction database T1 283
9.72 L1 283
9.73 C2 284
9.74 Transaction database T2 284
9.75 Support for C2 284
9.76 L2 285
9.77 Transaction database 287
9.78 Frequent 1-itemset L1 287
9.79 Candidate 2 itemsets C2 287
9.80 Possible 2-itemsets for each transaction 288
9.81 Frequent itemsets for a support of 50% 288
9.82 Code for each item 288
9.83 Coded representation for each item pair 288
9.84 Assigning item pairs to buckets based on hash function modulo 8 289
9.85 Pruning of C2 289
9.86 Finding L2 290
9.87 Finding three itemsets 290
9.88 Transaction database for Apriori and DHP 292
9.89 Code for each item 293
9.90 Coded representation for each item pair 293
9.91 Assigning of item pairs to buckets based on hash function modulo 7 294
9.92 Pruning of C2 294
9.93 Finding L2 294
9.94 Identifying three itemsets 295
9.95 Transaction database 295
9.96 Coded item pairs for DHP 296
9.97 Assigning of item pairs to buckets based on hash function modulo 11 296
9.98 Pruning of C2 296
9.99 Finding L2 297
9.100 Finding three itemsets 297
9.101 Working of the DIC algorithm for the example database 299
<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>9.102</td>
<td>Transaction database</td>
</tr>
<tr>
<td>9.103</td>
<td>Frequency of each item in sorted order</td>
</tr>
<tr>
<td>9.104</td>
<td>Updated database after eliminating the non-frequent items and reorganising it according to support</td>
</tr>
<tr>
<td>9.105</td>
<td>Frequent item pairs for database example given in table</td>
</tr>
<tr>
<td>9.106</td>
<td>Transaction database</td>
</tr>
<tr>
<td>9.107</td>
<td>Count for each data item</td>
</tr>
<tr>
<td>9.108</td>
<td>Frequency of each item in sorted order</td>
</tr>
<tr>
<td>9.109</td>
<td>Modified database after eliminating the non-frequent items and reorganising it according to support</td>
</tr>
<tr>
<td>9.110</td>
<td>Frequent item pairs for the example database</td>
</tr>
<tr>
<td>9.111</td>
<td>Calculation of confidence for identification of association rules</td>
</tr>
<tr>
<td>9.112</td>
<td>Transaction database</td>
</tr>
<tr>
<td>9.113</td>
<td>Frequency of each item</td>
</tr>
<tr>
<td>9.114</td>
<td>Frequency of each item in sorted order</td>
</tr>
<tr>
<td>9.115</td>
<td>Modified database after eliminating the non-frequent items and reorganizing it according to support</td>
</tr>
<tr>
<td>9.116</td>
<td>Frequent item pairs for example database</td>
</tr>
<tr>
<td>9.117</td>
<td>Association rules for database given in Table 9.76</td>
</tr>
<tr>
<td>10.1</td>
<td>Description of parameters</td>
</tr>
<tr>
<td>10.2</td>
<td>Description of available property options of the Apriori algorithm</td>
</tr>
<tr>
<td>10.3</td>
<td>Transaction database of a store</td>
</tr>
<tr>
<td>10.4</td>
<td>Sample dataset of a store</td>
</tr>
<tr>
<td>10.5</td>
<td>Performance record of students in a data warehouse and data mining course</td>
</tr>
<tr>
<td>11.1</td>
<td>Sequences of length two</td>
</tr>
<tr>
<td>11.2</td>
<td>Important parameters for web usage mining</td>
</tr>
<tr>
<td>11.3</td>
<td>Index showing keywords and related web pages</td>
</tr>
<tr>
<td>12.1</td>
<td>Generalized distinction between ODS and data warehouse</td>
</tr>
<tr>
<td>12.2</td>
<td>Comparison of OLTP systems and data warehousing systems</td>
</tr>
<tr>
<td>12.3</td>
<td>Comparing OLTP and data warehouse system</td>
</tr>
<tr>
<td>13.1</td>
<td>Comparison among Star, Snowflake and Fact Constellation Schema</td>
</tr>
<tr>
<td>14.1</td>
<td>Applications of OLAP</td>
</tr>
<tr>
<td>14.2</td>
<td>Difference between OLTP and OLAP</td>
</tr>
<tr>
<td>14.3</td>
<td>Possible number of queries on given scenario</td>
</tr>
<tr>
<td>14.4</td>
<td>Pre-computation of query analysis</td>
</tr>
<tr>
<td>14.5</td>
<td>Comparison of ROLAP and MOLAP</td>
</tr>
<tr>
<td>14.6</td>
<td>Result of the slice operation for degree = BE</td>
</tr>
<tr>
<td>15.1</td>
<td>The relational model to store account information</td>
</tr>
<tr>
<td>15.2</td>
<td>Comparison of terminologies used in Oracle and Riak</td>
</tr>
<tr>
<td>15.3</td>
<td>Comparison of terminologies used in RDBMS and Cassandra</td>
</tr>
<tr>
<td>15.4</td>
<td>Comparison of terminologies used in MongoDB and RDBMS</td>
</tr>
<tr>
<td>15.5</td>
<td>Friends database</td>
</tr>
</tbody>
</table>
Preface

In the modern age of artificial intelligence and business analytics, data is considered as the oil of this cyber world. The mining of data has huge potential to improve business outcomes, and to carry out the mining of data there is a growing demand for database mining experts. This book intends training learners to fill this gap.

This book will give learners sufficient information to acquire mastery over the subject. It covers the practical aspects of data mining, data warehousing, and machine learning in a simplified manner without compromising on the details of the subject. The main strength of the book is the illustration of concepts with practical examples so that the learners can grasp the contents easily. Another important feature of the book is illustration of data mining algorithms with practical hands-on sessions on Weka and R language (a major data mining tool and language, respectively).

In this book, every concept has been illustrated through a step-by-step approach in tutorial form for self-practice in Weka and R. This textbook includes many pedagogical features such as chapter wise summary, exercises including probable problems, question bank, and relevant references, to provide sound knowledge to learners. It provides the students a platform to obtain expertise on technology, for better placements.

Video sessions on data mining, machine learning, big data and DBMS are also available on my YouTube channel. Learners are requested to subscribe to this channel https://www.youtube.com/user/parteekbhatia to get the latest updates through video sessions on these topics.

Your suggestions for further improvements to the book are always welcome. Kindly e-mail your suggestions to parteek.bhatia@gmail.com.

I hope you enjoy learning from this book as much as I enjoyed writing it.
Writing the acknowledgments is the most emotional part of book writing. It provides an opportunity to pay gratitude to all those who matter in your life and have helped you achieve your dream and aspirations. With the grace of God and three years of effort, I have reached this stage.

I would like to express my gratitude to the many people who saw me through this book, who motivated me directly or indirectly to write this book, to all those who provided support, talked things over, read, wrote, offered comments, and assisted in the editing, proofreading, and design.

Writing a textbook is an enormous task and it requires a great deal of motivation. I appreciate the writings of great authors like Dr A. P. J. Abdul Kalam, Mr Robin Sharma, Mr Shiv Kehra and Mr Jack Canfield, who have inspired me to contribute to the education of our young generation by writing simplified content without compromising on the depth of the subject.

Writing a book is not possible without the support and motivation of one’s family. I feel blessed to have Dr Sanmeet Kaur as my wife; she has always been there to support and encourage me, despite all the time it took me, on this project. Since we both belong to the same field and same profession, having long discussions with her on different aspects of the subject is the most beautiful part of learning. These discussions helped me a long way in shaping the contents of the book. Secondly, she has always been there to take care of our whole family during my engagement with this book.

I am blessed to be born into a family of teachers. My parents, Mr Ved Kumar and Mrs Jagdish Bhatia have always provided a guiding path for excellence in life. Their life journey, in itself, is a learning path for me. I thank the almighty for providing me two loving sons, Rahat and Rishan, who filled our life with love and happiness. I thank my parents-in-laws, Mr Dalip Singh and Mrs Joginder Kaur whose daughter Sanmeet filled our home with love and affection. I thank my elder brother Mr Suneet Kumar and bhabi ji Mrs Dimple Bhatia, for always showering their love and blessings on me.

I am blessed to have mentors like M. L. Aeri, former Principal, DAV College, Amritsar; Mr O. P. Bhardwaj, former Head, Department of Computer Science, DAV College, Amritsar; Dr R. K. Sharma Professor, DCSE, TIET; Dr Seema Bawa, Professor DCSE, TIET; Dr Maninder Singh, Head CSED, TIET, and Dr Deepak Garg, former Head CSED, TIET, who groomed me as a teacher. I wish to thank my friends Dr Amardeep Gupta and Mr V. P. Singh, who always lend their ears to my thoughts and aspirations. I would like to thank my colleagues at TIET who motivate and provide an excellent environment for growth.
xxxiv  Acknowledgments

The production of this book involved a lot of help from my team of students consisting of Ms Sujata Singla, Mr Divanshu Singh, Ms Suhandhi, Mr Aditya and Ms Sawinder Kaur, who read the whole manuscript and helped me in editing and refining the text. I acknowledge the contribution of Ms Sujata in implementing the data mining algorithms in R and her assistance in finalizing the contents. There were great insights from Mr Divanshu Singh, who provided feedback and helped me refine the contents in the portions on web mining and search engine.

I would like to express my gratitude to my students at Thapar Institute of Engineering and Technology, Patiala, for their curiosity and zeal for learning which motivated me to write on this topic. I also want to thank the students at other institutes with whom I had the opportunity to interact during my 'invited talks'. They provided much-needed motivation, without which this book would have been impossible.

I want to acknowledge Dr Mark Polczynski, former Director of the MS at Marquette University, USA; Dr Saed Sayad, an Adjunct Professor at the University of Toronto; Mr Martin Fowler, ThoughtWorks, USA, for granting permission to use content from some of their published works.

I thank my publisher, Cambridge University Press, for publishing this book. I thank Mr Gauravjeet Singh of Cambridge University Press and his team for editing, refining and designing the contents, thereby providing life to the manuscript in the form of a book.