

Max Bramer

# Principles of Data Mining

Fourth Edition

 Springer

# Contents

<b>1. Introduction to Data Mining</b> .....	1
1.1 The Data Explosion .....	1
1.2 Knowledge Discovery .....	2
1.3 Applications of Data Mining .....	3
1.4 Labelled and Unlabelled Data .....	4
1.5 Supervised Learning: Classification .....	5
1.6 Supervised Learning: Numerical Prediction .....	7
1.7 Unsupervised Learning: Association Rules .....	7
1.8 Unsupervised Learning: Clustering .....	8
<b>2. Data for Data Mining</b> .....	9
2.1 Standard Formulation .....	9
2.2 Types of Variable .....	10
2.2.1 Categorical and Continuous Attributes .....	12
2.3 Data Preparation .....	12
2.3.1 Data Cleaning .....	13
2.4 Missing Values .....	15
2.4.1 Discard Instances .....	15
2.4.2 Replace by Most Frequent/Average Value .....	15
2.5 Reducing the Number of Attributes .....	16
2.6 The UCI Repository of Datasets .....	17
2.7 Chapter Summary .....	18
2.8 Self-assessment Exercises for Chapter 2 .....	18
Reference .....	19

<b>3. Introduction to Classification: Naïve Bayes and Nearest Neighbour</b> .....	21
3.1 What Is Classification? .....	21
3.2 Naïve Bayes Classifiers .....	22
3.3 Nearest Neighbour Classification .....	29
3.3.1 Distance Measures .....	32
3.3.2 Normalisation .....	35
3.3.3 Dealing with Categorical Attributes .....	36
3.4 Eager and Lazy Learning .....	36
3.5 Chapter Summary .....	37
3.6 Self-assessment Exercises for Chapter 3 .....	37
<b>4. Using Decision Trees for Classification</b> .....	39
4.1 Decision Rules and Decision Trees .....	39
4.1.1 Decision Trees: The Golf Example .....	40
4.1.2 Terminology .....	41
4.1.3 The <i>degrees</i> Dataset .....	42
4.2 The TDIDT Algorithm .....	45
4.3 Types of Reasoning .....	47
4.4 Chapter Summary .....	48
4.5 Self-assessment Exercises for Chapter 4 .....	48
References .....	48
<b>5. Decision Tree Induction: Using Entropy for Attribute Selection</b> .....	49
5.1 Attribute Selection: An Experiment .....	49
5.2 Alternative Decision Trees .....	50
5.2.1 The Football/Netball Example .....	51
5.2.2 The <i>anonymous</i> Dataset .....	53
5.3 Choosing Attributes to Split On: Using Entropy .....	54
5.3.1 The <i>lens24</i> Dataset .....	55
5.3.2 Entropy .....	57
5.3.3 Using Entropy for Attribute Selection .....	58
5.3.4 Maximising Information Gain .....	60
5.4 Chapter Summary .....	61
5.5 Self-assessment Exercises for Chapter 5 .....	61
<b>6. Decision Tree Induction: Using Frequency Tables for Attribute Selection</b> .....	63
6.1 Calculating Entropy in Practice .....	63
6.1.1 Proof of Equivalence .....	64
6.1.2 A Note on Zeros .....	66

6.2	Other Attribute Selection Criteria: Gini Index of Diversity . . . .	66
6.3	The $\chi^2$ Attribute Selection Criterion . . . . .	68
6.4	Inductive Bias . . . . .	71
6.5	Using Gain Ratio for Attribute Selection . . . . .	73
6.5.1	Properties of Split Information . . . . .	74
6.5.2	Summary . . . . .	75
6.6	Number of Rules Generated by Different Attribute Selection Criteria . . . . .	75
6.7	Missing Branches . . . . .	76
6.8	Chapter Summary . . . . .	77
6.9	Self-assessment Exercises for Chapter 6 . . . . .	77
	References . . . . .	78
<b>7.</b>	<b>Estimating the Predictive Accuracy of a Classifier . . . . .</b>	<b>79</b>
7.1	Introduction . . . . .	79
7.2	Method 1: Separate Training and Test Sets . . . . .	80
7.2.1	Standard Error . . . . .	81
7.2.2	Repeated Train and Test . . . . .	82
7.3	Method 2: $k$ -fold Cross-validation . . . . .	82
7.4	Method 3: $N$ -fold Cross-validation . . . . .	83
7.5	Experimental Results I . . . . .	84
7.6	Experimental Results II: Datasets with Missing Values . . . . .	86
7.6.1	Strategy 1: Discard Instances . . . . .	87
7.6.2	Strategy 2: Replace by Most Frequent/Average Value . . . . .	87
7.6.3	Missing Classifications . . . . .	89
7.7	Confusion Matrix . . . . .	89
7.7.1	True and False Positives . . . . .	90
7.8	Chapter Summary . . . . .	91
7.9	Self-assessment Exercises for Chapter 7 . . . . .	91
	Reference . . . . .	92
<b>8.</b>	<b>Continuous Attributes . . . . .</b>	<b>93</b>
8.1	Introduction . . . . .	93
8.2	Local versus Global Discretisation . . . . .	95
8.3	Adding Local Discretisation to TDIDT . . . . .	96
8.3.1	Calculating the Information Gain of a Set of Pseudo- attributes . . . . .	97
8.3.2	Computational Efficiency . . . . .	102
8.4	Using the ChiMerge Algorithm for Global Discretisation . . . . .	105
8.4.1	Calculating the Expected Values and $\chi^2$ . . . . .	108
8.4.2	Finding the Threshold Value . . . . .	113
8.4.3	Setting <i>minIntervals</i> and <i>maxIntervals</i> . . . . .	113

8.4.4	The ChiMerge Algorithm: Summary . . . . .	115
8.4.5	The ChiMerge Algorithm: Comments . . . . .	115
8.5	Comparing Global and Local Discretisation for Tree Induction	116
8.6	Chapter Summary . . . . .	118
8.7	Self-assessment Exercises for Chapter 8 . . . . .	118
	Reference . . . . .	119
<b>9.</b>	<b>Avoiding Overfitting of Decision Trees</b> . . . . .	<b>121</b>
9.1	Dealing with Clashes in a Training Set . . . . .	122
9.1.1	Adapting TDIDT to Deal with Clashes . . . . .	122
9.2	More About Overfitting Rules to Data . . . . .	127
9.3	Pre-pruning Decision Trees . . . . .	128
9.4	Post-pruning Decision Trees . . . . .	130
9.5	Chapter Summary . . . . .	135
9.6	Self-assessment Exercise for Chapter 9 . . . . .	136
	References . . . . .	136
<b>10.</b>	<b>More About Entropy</b> . . . . .	<b>137</b>
10.1	Introduction . . . . .	137
10.2	Coding Information Using Bits . . . . .	140
10.3	Discriminating Amongst $M$ Values ( $M$ Not a Power of 2) . . . . .	142
10.4	Encoding Values That Are Not Equally Likely . . . . .	143
10.5	Entropy of a Training Set . . . . .	146
10.6	Information Gain Must Be Positive or Zero . . . . .	147
10.7	Using Information Gain for Feature Reduction for Classification Tasks . . . . .	149
10.7.1	Example 1: The <i>genetics</i> Dataset . . . . .	150
10.7.2	Example 2: The <i>bcst96</i> Dataset . . . . .	154
10.8	Chapter Summary . . . . .	156
10.9	Self-assessment Exercises for Chapter 10 . . . . .	156
	References . . . . .	156
<b>11.</b>	<b>Inducing Modular Rules for Classification</b> . . . . .	<b>157</b>
11.1	Rule Post-pruning . . . . .	157
11.2	Conflict Resolution . . . . .	159
11.3	Problems with Decision Trees . . . . .	162
11.4	The Prism Algorithm . . . . .	164
11.4.1	Changes to the Basic Prism Algorithm . . . . .	171
11.4.2	Comparing Prism with TDIDT . . . . .	172
11.5	Chapter Summary . . . . .	173
11.6	Self-assessment Exercise for Chapter 11 . . . . .	173
	References . . . . .	174

- 12. Measuring the Performance of a Classifier** ..... 175
  - 12.1 True and False Positives and Negatives ..... 176
  - 12.2 Performance Measures ..... 178
  - 12.3 True and False Positive Rates versus Predictive Accuracy ..... 181
  - 12.4 ROC Graphs ..... 182
  - 12.5 ROC Curves ..... 184
  - 12.6 Finding the Best Classifier ..... 185
  - 12.7 Chapter Summary ..... 186
  - 12.8 Self-assessment Exercise for Chapter 12 ..... 187
  
- 13. Dealing with Large Volumes of Data** ..... 189
  - 13.1 Introduction ..... 189
  - 13.2 Distributing Data onto Multiple Processors ..... 192
  - 13.3 Case Study: PMCRI ..... 194
  - 13.4 Evaluating the Effectiveness of a Distributed System: PMCRI . 197
  - 13.5 Revising a Classifier Incrementally ..... 201
  - 13.6 Chapter Summary ..... 207
  - 13.7 Self-assessment Exercises for Chapter 13 ..... 207
  - References ..... 208
  
- 14. Ensemble Classification** ..... 209
  - 14.1 Introduction ..... 209
  - 14.2 Estimating the Performance of a Classifier ..... 212
  - 14.3 Selecting a Different Training Set for Each Classifier ..... 213
  - 14.4 Selecting a Different Set of Attributes for Each Classifier ..... 214
  - 14.5 Combining Classifications: Alternative Voting Systems ..... 215
  - 14.6 Parallel Ensemble Classifiers ..... 219
  - 14.7 Chapter Summary ..... 219
  - 14.8 Self-assessment Exercises for Chapter 14 ..... 220
  - References ..... 220
  
- 15. Comparing Classifiers** ..... 221
  - 15.1 Introduction ..... 221
  - 15.2 The Paired t-Test ..... 223
  - 15.3 Choosing Datasets for Comparative Evaluation ..... 229
    - 15.3.1 Confidence Intervals ..... 231
  - 15.4 Sampling ..... 231
  - 15.5 How Bad Is a ‘No Significant Difference’ Result? ..... 234
  - 15.6 Chapter Summary ..... 235
  - 15.7 Self-assessment Exercises for Chapter 15 ..... 235
  - References ..... 236

<b>16. Association Rule Mining I</b> .....	237
16.1 Introduction .....	237
16.2 Measures of Rule Interestingness .....	239
16.2.1 The Piatetsky-Shapiro Criteria and the RI Measure ...	241
16.2.2 Rule Interestingness Measures Applied to the <i>chess</i> Dataset .....	243
16.2.3 Using Rule Interestingness Measures for Conflict Res- olution .....	245
16.3 Association Rule Mining Tasks .....	245
16.4 Finding the Best $N$ Rules .....	246
16.4.1 The $J$ -Measure: Measuring the Information Content of a Rule .....	247
16.4.2 Search Strategy .....	248
16.5 Chapter Summary .....	251
16.6 Self-assessment Exercises for Chapter 16 .....	251
References .....	251
<b>17. Association Rule Mining II</b> .....	253
17.1 Introduction .....	253
17.2 Transactions and Itemsets .....	254
17.3 Support for an Itemset .....	255
17.4 Association Rules .....	256
17.5 Generating Association Rules .....	258
17.6 Apriori .....	259
17.7 Generating Supported Itemsets: An Example .....	262
17.8 Generating Rules for a Supported Itemset .....	264
17.9 Rule Interestingness Measures: Lift and Leverage .....	266
17.10 Chapter Summary .....	268
17.11 Self-assessment Exercises for Chapter 17 .....	269
Reference .....	269
<b>18. Association Rule Mining III: Frequent Pattern Trees</b> .....	271
18.1 Introduction: FP-Growth .....	271
18.2 Constructing the FP-tree .....	274
18.2.1 Pre-processing the Transaction Database .....	274
18.2.2 Initialisation .....	276
18.2.3 Processing Transaction 1: $f, c, a, m, p$ .....	277
18.2.4 Processing Transaction 2: $f, c, a, b, m$ .....	279
18.2.5 Processing Transaction 3: $f, b$ .....	283
18.2.6 Processing Transaction 4: $c, b, p$ .....	285
18.2.7 Processing Transaction 5: $f, c, a, m, p$ .....	287
18.3 Finding the Frequent Itemsets from the FP-tree .....	288

18.3.1	Itemsets Ending with Item $p$ . . . . .	291
18.3.2	Itemsets Ending with Item $m$ . . . . .	301
18.4	Chapter Summary . . . . .	308
18.5	Self-assessment Exercises for Chapter 18 . . . . .	309
	Reference . . . . .	309
<b>19.</b>	<b>Clustering</b> . . . . .	<b>311</b>
19.1	Introduction . . . . .	311
19.2	$k$ -Means Clustering . . . . .	314
19.2.1	Example . . . . .	315
19.2.2	Finding the Best Set of Clusters . . . . .	319
19.3	Agglomerative Hierarchical Clustering . . . . .	320
19.3.1	Recording the Distance Between Clusters . . . . .	323
19.3.2	Terminating the Clustering Process . . . . .	326
19.4	Chapter Summary . . . . .	327
19.5	Self-assessment Exercises for Chapter 19 . . . . .	327
<b>20.</b>	<b>Text Mining</b> . . . . .	<b>329</b>
20.1	Multiple Classifications . . . . .	329
20.2	Representing Text Documents for Data Mining . . . . .	330
20.3	Stop Words and Stemming . . . . .	332
20.4	Using Information Gain for Feature Reduction . . . . .	333
20.5	Representing Text Documents: Constructing a Vector Space Model . . . . .	333
20.6	Normalising the Weights . . . . .	335
20.7	Measuring the Distance Between Two Vectors . . . . .	336
20.8	Measuring the Performance of a Text Classifier . . . . .	337
20.9	Hypertext Categorisation . . . . .	338
20.9.1	Classifying Web Pages . . . . .	338
20.9.2	Hypertext Classification versus Text Classification . . . . .	339
20.10	Chapter Summary . . . . .	343
20.11	Self-assessment Exercises for Chapter 20 . . . . .	343
<b>21.</b>	<b>Classifying Streaming Data</b> . . . . .	<b>345</b>
21.1	Introduction . . . . .	345
21.1.1	Stationary v Time-dependent Data . . . . .	347
21.2	Building an H-Tree: Updating Arrays . . . . .	347
21.2.1	Array <i>currentAtts</i> . . . . .	348
21.2.2	Array <i>splitAtt</i> . . . . .	349
21.2.3	Sorting a record to the appropriate leaf node . . . . .	349
21.2.4	Array <i>hitcount</i> . . . . .	350
21.2.5	Array <i>classtotals</i> . . . . .	350



21.2.6	Array <i>acvCounts</i> . . . . .	350
21.2.7	Array <i>branch</i> . . . . .	352
21.3	Building an H-Tree: a Detailed Example . . . . .	352
21.3.1	Step (a): Initialise Root Node 0 . . . . .	352
21.3.2	Step (b): Begin Reading Records . . . . .	353
21.3.3	Step (c): Consider Splitting at Node 0 . . . . .	354
21.3.4	Step (d): Split on Root Node and Initialise New Leaf Nodes . . . . .	355
21.3.5	Step (e): Process the Next Set of Records . . . . .	357
21.3.6	Step (f): Consider Splitting at Node 2 . . . . .	358
21.3.7	Step (g): Process the Next Set of Records . . . . .	359
21.3.8	Outline of the H-Tree Algorithm . . . . .	360
21.4	Splitting on an Attribute: Using Information Gain . . . . .	363
21.5	Splitting on An Attribute: Using a Hoeffding Bound . . . . .	365
21.6	H-Tree Algorithm: Final Version . . . . .	370
21.7	Using an Evolving H-Tree to Make Predictions . . . . .	372
21.7.1	Evaluating the Performance of an H-Tree . . . . .	373
21.8	Experiments: H-Tree versus TDIDT . . . . .	374
21.8.1	The lens24 Dataset . . . . .	374
21.8.2	The vote Dataset . . . . .	376
21.9	Chapter Summary . . . . .	377
21.10	Self-assessment Exercises for Chapter 21 . . . . .	377
	References . . . . .	378
<b>22.</b>	<b>Classifying Streaming Data II: Time-Dependent Data . . . . .</b>	<b>379</b>
22.1	Stationary versus Time-dependent Data . . . . .	379
22.2	Summary of the H-Tree Algorithm . . . . .	381
22.2.1	Array <i>currentAtts</i> . . . . .	382
22.2.2	Array <i>splitAtt</i> . . . . .	383
22.2.3	Array <i>hitcount</i> . . . . .	383
22.2.4	Array <i>classtotals</i> . . . . .	383
22.2.5	Array <i>acvCounts</i> . . . . .	384
22.2.6	Array <i>branch</i> . . . . .	384
22.2.7	Pseudocode for the H-Tree Algorithm . . . . .	384
22.3	From H-Tree to CDH-Tree: Overview . . . . .	387
22.4	From H-Tree to CDH-Tree: Incrementing Counts . . . . .	387
22.5	The Sliding Window Method . . . . .	388
22.6	Resplitting at Nodes . . . . .	393
22.7	Identifying Suspect Nodes . . . . .	394
22.8	Creating Alternate Nodes . . . . .	396
22.9	Growing/Forgetting an Alternate Node and its Descendants . . . . .	400

22.10	Replacing an Internal Node by One of its Alternate Nodes . . . .	402
22.11	Experiment: Tracking Concept Drift . . . . .	410
22.11.1	<i>lens24</i> Data: Alternative Mode . . . . .	412
22.11.2	Introducing Concept Drift . . . . .	414
22.11.3	An Experiment with Alternating <i>lens24</i> Data . . . . .	415
22.11.4	Comments on Experiment . . . . .	423
22.12	Chapter Summary . . . . .	424
22.13	Self-assessment Exercises for Chapter 22 . . . . .	424
	Reference . . . . .	425
<b>23.</b>	<b>An Introduction to Neural Networks . . . . .</b>	<b>427</b>
23.1	Introduction . . . . .	427
23.2	Neural Nets Example 1 . . . . .	430
23.3	Neural Nets Example 2 . . . . .	434
23.3.1	Forward Propagating the Values of the Input Nodes . . .	437
23.3.2	Forward Propagation: Summary of Formulae . . . . .	442
23.4	Backpropagation . . . . .	443
23.4.1	Stochastic Gradient Descent . . . . .	443
23.4.2	Finding the Gradients . . . . .	445
23.4.3	Working backwards from the output layer to the hid- den layer . . . . .	447
23.4.4	Working backwards from the hidden layer to the input layer . . . . .	449
23.4.5	Updating the Weights . . . . .	452
23.5	Processing a Multi-instance Training Set . . . . .	455
23.6	Using a Neural Net for Classification: the <i>iris</i> Dataset . . . . .	456
23.7	Using a Neural Net for Classification: the <i>seeds</i> Dataset . . . . .	461
23.8	Neural Nets: A Note of Caution . . . . .	463
23.9	Chapter Summary . . . . .	465
23.10	Self-assessment Exercises for Chapter 23 . . . . .	466
<b>A.</b>	<b>Essential Mathematics . . . . .</b>	<b>467</b>
A.1	Subscript Notation . . . . .	467
A.1.1	Sigma Notation for Summation . . . . .	468
A.1.2	Double Subscript Notation . . . . .	469
A.1.3	Other Uses of Subscripts . . . . .	470
A.2	Trees . . . . .	470
A.2.1	Terminology . . . . .	471
A.2.2	Interpretation . . . . .	472
A.2.3	Subtrees . . . . .	473
A.3	The Logarithm Function $\log_2 X$ . . . . .	473
A.3.1	The Function $-X \log_2 X$ . . . . .	476

A.4	Introduction to Set Theory .....	477
A.4.1	Subsets .....	479
A.4.2	Summary of Set Notation .....	481
<b>B.</b>	<b>Datasets</b> .....	483
	References .....	504
<b>C.</b>	<b>Sources of Further Information</b> .....	505
	Websites .....	505
	Books .....	505
	Conferences .....	506
	Information About Association Rule Mining .....	507
<b>D.</b>	<b>Glossary and Notation</b> .....	509
<b>E.</b>	<b>Solutions to Self-assessment Exercises</b> .....	535
<b>Index</b>	.....	565