

Table of Contents

Chapter 1. Introduction	1
1.1. What is Data Mining and Text Mining?	1
1.2. Common steps in data mining and text mining	2
1.3. Types of Data in Data Mining/Text Mining	3
1.4. Some Data Mining Applications	9
1.5. Text Mining Application	10
1.6. Types of Mining Tasks	12
1.6.1. Classification or Categorization: Finding the class of an object	12
1.6.2. Prediction: Predicting the value for an object	14
1.6.3. Clustering/Deviation Detection: Grouping data/Detecting outliers	15
1.6.4. Association Analysis: Finding frequent co-occurrences	17
1.6.5. Characterization and Discrimination: Describing a class or concept	20
1.6.6. Meta Functionalities: Link, Outlier, and Trend/Evolutional Analysis	22
1.6.7. Visualization	23
1.7. Challenges in Data Mining and Text Mining	26
1.8. Summary	28
1.9. Historical Bibliography	29
Exercise	30
Chapter 2. Data Preprocessing	31
2.1. Basic Representation for Data: Database Viewpoint	31
2.2. Data Preprocessing in the Database Point of View	33
2.3. Data Cleaning	34
2.4. Data Integration, Transformation, and Reduction	37
2.5. Data Transformation: Attribution Construction and Normalization	42
2.5.1. Attribute Construction	43
2.5.2. Attribute Normalization	43
2.5.3. Time-dependent Attribute Transformation: (Feature Construction)	45
2.6. Data Reduction	46
2.6.1. Reduction in the number of attributes	46
2.6.2. Reduction in the number of tuples	49
2.6.3. Reduction in the number of possible values	51
2.7. Dimensionality Reduction Techniques	52
2.7.1. Discrete Wavelet Transforms (DWT)	53
2.7.2. Principal Components Analysis	55
2.8. Summary	56
2.9. Historical Bibliography	58
Exercise	59

Chapter 3.	Classification and Prediction.....	61
3.1.	Classification.....	61
3.1.1.	Fisher’s linear discriminant or centroid-based method.....	62
3.1.2.	k-nearest neighbor method.....	70
3.1.3.	Statistical Classifiers.....	74
3.1.4.	Decision Trees.....	87
3.1.5.	Classification Rules: Covering Algorithm.....	113
3.1.6.	Artificial Neural Networks.....	124
3.1.7.	Support Vector Machines (SVMs).....	127
3.2.	Numerical Prediction.....	140
3.2.1.	Regression.....	140
3.2.2.	Tree for prediction: Regression Tree and Model Tree.....	146
3.3.	Regression as Classification.....	148
3.3.1.	One-Against-the-Other Regression.....	148
3.3.2.	Pairwise Regression.....	150
3.4.	Model Ensemble Techniques.....	153
3.4.1.	Bagging: Bootstrap Aggregating.....	155
3.4.2.	Boosting: AdaBoost Algorithm.....	157
3.4.3.	Stacking.....	160
3.4.4.	Co-training.....	163
3.5.	Historical Bibliography.....	164
	Exercise.....	167
Chapter 4.	Clustering.....	171
4.1.	Cluster Analysis or Clustering.....	171
4.1.1.	Distance and similarity measurement.....	173
4.1.2.	Clustering Methods.....	177
4.1.3.	Partition-based Methods.....	179
4.1.4.	Hierarchical-based clustering.....	183
4.1.5.	Density-based clustering.....	186
4.1.6.	Grid-based clustering.....	188
4.1.7.	Model-based clustering.....	189
4.2.	Association Analysis and Frequent Pattern Mining.....	193
4.2.1.	Apriori algorithm.....	197
4.2.2.	FP-Tree algorithm.....	202
4.2.3.	CHARM algorithm.....	206
4.2.4.	Association Rules with Hierarchical Structure.....	210
4.2.5.	Efficient Association Rule Mining with Hierarchical Structure.....	216
4.3.	Historical Bibliography.....	218
	Exercise.....	221

Chapter 5. Evaluation	223
5.1. Approaches for defining the training and test sets	225
5.2. Lift chart and ROC-curve	231
5.3. Recall, precision, f-measure and accuracy	234
5.4. Evaluating numeric prediction	239
5.5. Historical Bibliography	242
Exercise	243
Chapter 6. Applications to Text Mining.....	245
6.1. Centroid-based Text Classification.....	247
6.1.1. Formulation of centroid-based text classification.....	248
6.1.2. Effect of Term distributions	251
6.1.3. Experimental Settings and Results	253
6.2. Document Relation Extraction.....	258
6.2.1. Document Relation Discovery using Frequent Itemset Mining.....	259
6.2.2. Empirical Evaluation using Citation Information.....	259
6.2.3. Experimental Settings and Results	264
6.3. Application to Automatic Thai Unknown Detection	269
6.3.1. Thai Unknown Words as Word Segmentation Problem.....	271
6.3.2. The Proposed Method.....	271
6.3.3. Experimental Settings and Results	280
Reference.....	283