Contents

1	**Introduction**	**1**
1.1	What Is Data Mining?	3
1.2	Some More Real-World Applications	3
1.3	Data Mining Methods – An Overview	6
1.3.1	Basic Problem Types	6
1.3.2	Prediction	6
1.3.2.1	Classification	6
1.3.2.2	Regression	7
1.3.3	Knowlegde Discovery	7
1.3.3.1	Deviation Detection	7
1.3.3.2	Cluster Analysis	7
1.3.3.3	Visualization	8
1.3.3.4	Association Rules	8
1.3.3.5	Segmentation	8
1.4	Data Mining Viewed from the Data Side	9
1.5	Types of Data	10
1.6	Conclusion	11
2	**Data Preparation**	**13**
2.1	Data Cleaning	13
2.2	Handling Outlier	14
2.3	Handling Noisy Data	14
2.4	Missing Values Handling	16
2.5	Coding	16
2.6	Recognition of Correlated or Redundant Attributes	16
2.7	Abstraction	17
2.7.1	Attribute Construction	17
2.7.2	Images	17
2.7.3	Time Series	18
2.7.4	Web Data	19
2.8	Conclusions	22
3	**Methods for Data Mining**	**23**
3.1	Decision Tree Induction	23
3.1.1	Basic Principle	23
3.1.2	Terminology of Decision Tree	24
3.1.3	Subtasks and Design Criteria for Decision Tree Induction	25
3.1.4	Attribute Selection Criteria	28
3.1.4.1	Information Gain Criteria and Gain Ratio	29
3.1.4.2	Gini Function	30
3.1.5	Discretization of Attribute Values	31
3.1.5.1	Binary Discretization	32
3.1.5.2	Multi-interval Discretization	34
3.1.5.3	Discretization of Categorical or Symbolical Attributes	41
3.1.6	Pruning	42
3.1.7	Overview	43
3.1.8	Cost-Complexity Pruning	43
3.1.9	Some General Remarks	44
3.1.10	Summary	46
3.2	Case-Based Reasoning	46
3.2.1	Background	47

3.2.2	The Case-Based Reasoning Process	47
3.2.3	CBR Maintenance	48
3.2.4	Knowledge Containers in a CBR System	49
3.2.5	Design Consideration	50
3.2.6	Similarity	50
3.2.6.1	Formalization of Similarity	50
3.2.6.2	Similarity Measures	51
3.2.6.3	Similarity Measures for Images	51
3.2.7	Case Description	53
3.2.8	Organization of Case Base	53
3.2.9	Learning in a CBR System	55
3.2.9.1	Learning of New Cases and Forgetting of Old Cases	56
3.2.9.2	Learning of Prototypes	56
3.2.9.3	Learning of Higher Order Constructs	56
3.2.9.4	Learning of Similarity	56
3.2.10	Conclusions	57
3.3	Clustering	57
3.3.1	Introduction	57
3.3.2	General Comments	58
3.3.3	Distance Measures for Metrical Data	59
3.3.4	Using Numerical Distance Measures for Categorical Data	60
3.3.5	Distance Measure for Nominal Data	61
3.3.6	Contrast Rule	62
3.3.7	Agglomerate Clustering Methods	62
3.3.8	Partitioning Clustering	64
3.3.9	Graphs Clustering	64
3.3.10	Similarity Measure for Graphs	65
3.3.11	Hierarchical Clustering of Graphs	69
3.3.12	Conclusion	71
3.4	Conceptual Clustering	71
3.4.1	Introduction	71
3.4.2	Concept Hierarchy and Concept Description	71
3.4.3	Category Utility Function	72
3.4.4	Algorithmic Properties	73
3.4.5	Algorithm	73
3.4.6	Conceptual Clustering of Graphs	75
3.4.6.1	Notion of a Case and Similarity Measure	75
3.4.6.2	Evaluation Function	75
3.4.6.3	Prototype Learning	76
3.4.6.4	An Example of a Learned Concept Hierarchy	76
3.4.7	Conclusion	79
3.5	Evaluation of the Model	79
3.5.1	Error Rate, Correctness, and Quality	79
3.5.2	Sensitivity and Specifity	81
3.5.3	Test-and-Train	82
3.5.4	Random Sampling	82
3.5.5	Cross Validation	82
3.5.6	Conclusion	83
3.6	Feature Subset Selection	83
3.6.1	Introduction	83
3.6.2	Feature Subset Selection Algorithms	83
3.6.2.1	The Wrapper and the Filter Model for Feature Subset Selection	84
3.6.3	Feature Selection Done by Decision Tree Induction	85
3.6.4	Feature Subset Selection Done by Clustering	86
3.6.5	Contextual Merit Algorithm	87
3.6.6	Floating Search Method	88
3.6.7	Conclusion	88

4	**Applications**	**91**
4.1	Controlling the Parameters of an Algorithm/Model by Case-Based Reasoning	91
4.1.1	Modelling Concerns	91
4.1.2	Case-Based Reasoning Unit	92
4.1.3	Management of the Case Base	93
4.1.4	Case Structure and Case Base	94
4.1.4.1	Non-image Information	95
4.1.4.2	Image Information	96
4.1.5	Image Similarity Determination	97
4.1.5.1	Image Similarity Measure 1 (ISim_1)	97
4.1.5.2	Image Similarity Measure 2 (ISIM_2)	98
4.1.5.3	Comparision of ISim_1 and ISim_2	98
4.1.6	Segmentation Algorithm and Segmentation Parameters	99
4.1.7	Similarity Determination	100
4.1.7.1	Overall Similarity	100
4.1.7.2	Similarity Measure for Non-image Information	101
4.1.7.3	Similarity Measure for Image Information	101
4.1.8	Knowledge Acquisition Aspect	101
4.1.9	Conclusion	102
4.2	Mining Images	102
4.2.1	Introduction	102
4.2.2	Preparing the Experiment	103
4.2.3	Image Mining Tool	105
4.2.4	The Application	106
4.2.5	Brainstorming and Image Catalogue	107
4.2.6	Interviewing Process	107
4.2.7	Setting Up the Automatic Image Analysis and Feature Extraction Procedure	107
4.2.7.1	Image Analysis	108
4.2.7.2	Feature Extraction	109
4.2.8	Collection of Image Descriptions into the Data Base	111
4.2.9	The Image Mining Experiment	112
4.2.10	Review	113
4.2.11	Using the Discovered Knowledge	114
4.1.12	Lessons Learned	115
4.2.13	Conclusions	116
5	**Conclusion**	**117**
	Appendix	119
	The IRIS Data Set	119
	References	121
	Index	129