

# Contents

<b>1</b>	<b>Introduction</b> .....	1
1.1	Data Mining and Knowledge Discovery .....	1
1.2	Data Mining Methods .....	2
1.3	Supervised Learning .....	6
1.4	Unsupervised Learning .....	7
1.5	Other Learning Paradigms .....	8
1.6	Introduction to Data Preprocessing .....	10
1.6.1	Data Preparation .....	10
1.6.2	Data Reduction .....	13
	References .....	16
<b>2</b>	<b>Data Sets and Proper Statistical Analysis of Data Mining Techniques</b> .....	19
2.1	Data Sets and Partitions .....	19
2.1.1	Data Set Partitioning .....	20
2.1.2	Performance Measures .....	24
2.2	Using Statistical Tests to Compare Methods .....	25
2.2.1	Conditions for the Safe Use of Parametric Tests .....	25
2.2.2	Normality Test over the Group of Data Sets and Algorithms .....	27
2.2.3	Non-Parametric Tests for Comparing Two Algorithms in Multiple Data Set Analysis .....	29
2.2.4	Non-Parametric Tests for Multiple Comparisons among More than Two Algorithms .....	32
	References .....	37
<b>3</b>	<b>Data Preparation Basic Models</b> .....	39
3.1	Overview .....	39
3.2	Data Integration .....	40
3.2.1	Finding Redundant Attributes .....	41
3.2.2	Detecting Tuple Duplication and Inconsistency .....	43
3.3	Data Cleaning .....	45
3.4	Data Normalization .....	46

3.4.1	Min-Max Normalization .....	47
3.4.2	Z-score Normalization .....	47
3.4.3	Decimal Scaling Normalization .....	48
3.5	Data Transformation .....	49
3.5.1	Linear Transformations .....	49
3.5.2	Quadratic Transformations .....	49
3.5.3	Non-Polynomial Approximations of Transformations .....	50
3.5.4	Polynomial Approximations of Transformations .....	51
3.5.5	Rank Transformations .....	52
3.5.6	Box-Cox Transformations .....	53
3.5.7	Spreading the Histogram .....	54
3.5.8	Nominal to Binary Transformation .....	54
3.5.9	Transformations via Data Reduction .....	55
	References .....	56
<b>4</b>	<b>Dealing with Missing Values</b> .....	<b>59</b>
4.1	Introduction .....	59
4.2	Assumptions and Missing Data Mechanisms .....	61
4.3	Simple Approaches to Missing Data .....	63
4.4	Maximum Likelihood Imputation Methods .....	64
4.4.1	Expectation-Maximization (EM) .....	65
4.4.2	Multiple Imputation .....	68
4.4.3	Bayesian Principal Component Analysis (BPCA) .....	72
4.5	Imputation of Missing Values. Machine Learning Based Methods ..	75
4.5.1	Imputation with K-Nearest Neighbor (KNNI) .....	76
4.5.2	Weighted imputation with K-Nearest Neighbour (WKNNI) .	77
4.5.3	K-means Clustering Imputation (KMI) .....	77
4.5.4	Imputation with Fuzzy K-means Clustering (FKMI) .....	78
4.5.5	Support Vector Machines Imputation (SVMI) .....	79
4.5.6	Event Covering (EC) .....	81
4.5.7	Singular Value Decomposition Imputation (SVDI) .....	85
4.5.8	Local Least Squares Imputation (LLSI) .....	86
4.5.9	Recent Machine Learning Approaches to Missing Values Imputation .....	88
4.6	Experimental Comparative Analysis .....	89
4.6.1	Effect of the Imputation Methods in the Attributes' Relationships .....	90
4.6.2	Best Imputation Methods for Classification Methods .....	92
4.6.3	Interesting Comments .....	97
	References .....	97
<b>5</b>	<b>Dealing with Noisy Data</b> .....	<b>103</b>
5.1	Identifying Noise .....	103
5.2	Types of Noise Data: Class Noise and Attribute Noise .....	106
5.2.1	Noise Introduction Mechanisms .....	108

5.2.2	Simulating the Noise of Real-world Data Sets	110
5.3	Noise Filtering at Data Level	111
5.3.1	Ensemble Filter	112
5.3.2	Cross-Validated Committees Filter	113
5.3.3	Iterative-Partitioning Filter	113
5.3.4	More filtering methods	114
5.4	Robust Learners Against Noise	115
5.4.1	Multiple Classifier Systems for Classification Tasks	116
5.4.2	Addressing Multi-class Classification Problems by Decomposition	119
5.5	Empirical Analysis of Noise Filters and Robust Strategies	121
5.5.1	Noise Introduction	121
5.5.2	Noise Filters for Class Noise	123
5.5.3	Noise Filtering Efficacy Prediction by Data Complexity Measures	125
5.5.4	Multiple Classifier Systems with Noise	129
5.5.5	Analysis of the OVO Decomposition with Noise	132
	References	137
<b>6</b>	<b>Data Reduction</b>	143
6.1	Overview	143
6.2	The Curse of Dimensionality	144
6.2.1	Principal Components Analysis	145
6.2.2	Factor Analysis	146
6.2.3	Multidimensional Scaling	149
6.2.4	Locally Linear Embedding	151
6.3	Data Sampling	152
6.3.1	Data Condensation	154
6.3.2	Data Squashing	155
6.3.3	Data Clustering	155
6.4	Binning and Reduction of Cardinality	157
	References	158
<b>7</b>	<b>Feature Selection</b>	159
7.1	Overview	159
7.2	Perspectives	160
7.2.1	The Search of a Subset of Features	160
7.2.2	Selection Criteria	165
7.2.3	Filter, Wrapper and Embedded Feature Selection	169
7.3	Aspects	172
7.3.1	Output of Feature Selection	172
7.3.2	Evaluation	174
7.3.3	Drawbacks	175
7.3.4	Using Decision Trees for Feature Selection	175
7.4	Description of the Most Representative Feature Selection Methods	176

7.4.1	Exhaustive Methods	177
7.4.2	Heuristic Methods	178
7.4.3	Nondeterministic Methods	179
7.4.4	Feature Weighting Methods	181
7.5	Related and Advanced Topics	182
7.5.1	Leading and Recent Feature Selection Techniques	182
7.5.2	Feature Extraction	184
7.5.3	Feature Construction	185
7.6	Experimental Comparative Analyses in Feature Selection	186
	References	187
<b>8</b>	<b>Instance Selection</b>	191
8.1	Introduction	191
8.2	Training Set Selection vs. Prototype Selection	193
8.3	Prototype Selection Taxonomy	195
8.3.1	Common Properties in Prototype Selection Methods	195
8.3.2	Prototype Selection Methods	198
8.3.3	Taxonomy of Prototype Selection Methods	200
8.4	Description of Methods	201
8.4.1	Condensation Algorithms	202
8.4.2	Edition Algorithms	206
8.4.3	Hybrid Algorithms	208
8.5	Related and Advanced Topics	216
8.5.1	Prototype Generation	217
8.5.2	Distance Metrics, Feature Weighting and Combinations with Feature Selection	217
8.5.3	Hybridizations with Other Learning Methods and Ensembles	218
8.5.4	Scaling-Up Approaches	219
8.5.5	Data Complexity	220
8.6	Experimental Comparative Analysis in Prototype Selection	220
8.6.1	Analysis and Empirical Results on Small Size Data Sets	223
8.6.2	Analysis and Empirical Results on Medium Size Data Sets	225
8.6.3	Global View of the Obtained Results	227
8.6.4	Visualization of Data Subsets: A Case Study Based on the Banana Data Set	227
	References	230
<b>9</b>	<b>Discretization</b>	239
9.1	Introduction	239
9.2	Perspectives and Background	241
9.2.1	Discretization Process	241
9.2.2	Related and Advanced Work	244
9.3	Properties and Taxonomy	245
9.3.1	Common Properties	245
9.3.2	Methods and Taxonomy	250

9.3.3	Description of the Most Representative Discretization Methods . . . . .	252
9.4	Experimental Comparative Analysis . . . . .	259
9.4.1	Experimental Set Up . . . . .	259
9.4.2	Analysis and Empirical Results . . . . .	260
	References . . . . .	267
<b>10</b>	<b>A Data Mining Software Package Including Data Preparation and Reduction: KEEL . . . . .</b>	<b>275</b>
10.1	Data Mining Softwares and Toolboxes . . . . .	275
10.2	KEEL: Knowledge Extraction based on Evolutionary Learning . . . . .	277
10.2.1	Main Features . . . . .	278
10.2.2	Data Management . . . . .	279
10.2.3	Design of Experiments: Off-Line Module . . . . .	281
10.2.4	Computer-Based Education: On-Line Module . . . . .	283
10.3	KEEL-dataset . . . . .	284
10.3.1	Data Sets Web Pages . . . . .	285
10.3.2	Experimental Study Web Pages . . . . .	287
10.4	Integration of New Algorithms into the KEEL Tool . . . . .	288
10.4.1	Introduction to the KEEL Codification Features . . . . .	289
10.5	KEEL Statistical Tests . . . . .	293
10.5.1	Case Study . . . . .	295
10.6	Summarizing Comments . . . . .	300
	References . . . . .	301
	<b>Index . . . . .</b>	<b>305</b>