

# Contents

<b>1</b>	<b>Introduction</b>	1
1.1	Data Mining and Data Compression	1
1.1.1	Data Mining Tasks	1
1.1.2	Data Compression	2
1.1.3	Compression Using Data Mining Tasks	2
1.2	Organization	3
1.2.1	Data Mining Tasks	3
1.2.2	Abstraction in Nonlossy Compression Domain	5
1.2.3	Lossy Compression Scheme and Dimensionality Reduction	6
1.2.4	Compaction Through Simultaneous Prototype and Feature Selection	6
1.2.5	Use of Domain Knowledge in Data Compaction	7
1.2.6	Compression Through Dimensionality Reduction	7
1.2.7	Big Data, Multiagent Systems, and Abstraction	8
1.3	Summary	9
1.4	Bibliographical Notes	9
	References	9
<b>2</b>	<b>Data Mining Paradigms</b>	11
2.1	Introduction	11
2.2	Clustering	12
2.2.1	Clustering Algorithms	13
2.2.2	Single-Link Algorithm	14
2.2.3	$k$ -Means Algorithm	15
2.3	Classification	17
2.4	Association Rule Mining	22
2.4.1	Frequent Itemsets	23
2.4.2	Association Rules	25
2.5	Mining Large Datasets	26

2.5.1	Possible Solutions . . . . .	27
2.5.2	Clustering . . . . .	28
2.5.3	Classification . . . . .	34
2.5.4	Frequent Itemset Mining . . . . .	39
2.6	Summary . . . . .	42
2.7	Bibliographic Notes . . . . .	43
	References . . . . .	44
<b>3</b>	<b>Run-Length-Encoded Compression Scheme . . . . .</b>	<b>47</b>
3.1	Introduction . . . . .	47
3.2	Compression Domain for Large Datasets . . . . .	48
3.3	Run-Length-Encoded Compression Scheme . . . . .	49
3.3.1	Discussion on Relevant Terms . . . . .	49
3.3.2	Important Properties and Algorithm . . . . .	50
3.4	Experimental Results . . . . .	55
3.4.1	Application to Handwritten Digit Data . . . . .	55
3.4.2	Application to Genetic Algorithms . . . . .	57
3.4.3	Some Applicable Scenarios in Data Mining . . . . .	59
3.5	Invariance of VC Dimension in the Original and the Compressed Forms . . . . .	60
3.6	Minimum Description Length . . . . .	63
3.7	Summary . . . . .	65
3.8	Bibliographic Notes . . . . .	65
	References . . . . .	66
<b>4</b>	<b>Dimensionality Reduction by Subsequence Pruning . . . . .</b>	<b>67</b>
4.1	Introduction . . . . .	67
4.2	Lossy Data Compression for Clustering and Classification . . . . .	67
4.3	Background and Terminology . . . . .	68
4.4	Preliminary Data Analysis . . . . .	73
4.4.1	Huffman Coding and Lossy Compression . . . . .	74
4.4.2	Analysis of Subsequences and Their Frequency in a Class . . . . .	79
4.5	Proposed Scheme . . . . .	81
4.5.1	Initialization . . . . .	83
4.5.2	Frequent Item Generation . . . . .	83
4.5.3	Generation of Coded Training Data . . . . .	84
4.5.4	Subsequence Identification and Frequency Computation . . . . .	84
4.5.5	Pruning of Subsequences . . . . .	85
4.5.6	Generation of Encoded Test Data . . . . .	85
4.5.7	Classification Using Dissimilarity Based on Rough Set Concept . . . . .	86
4.5.8	Classification Using $k$ -Nearest Neighbor Classifier . . . . .	87
4.6	Implementation of the Proposed Scheme . . . . .	87
4.6.1	Choice of Parameters . . . . .	87
4.6.2	Frequent Items and Subsequences . . . . .	88

4.6.3	Compressed Data and Pruning of Subsequences . . . . .	89
4.6.4	Generation of Compressed Training and Test Data . . . . .	91
4.7	Experimental Results . . . . .	91
4.8	Summary . . . . .	92
4.9	Bibliographic Notes . . . . .	93
	References . . . . .	94
<b>5</b>	<b>Data Compaction Through Simultaneous Selection of Prototypes and Features . . . . .</b>	<b>95</b>
5.1	Introduction . . . . .	95
5.2	Prototype Selection, Feature Selection, and Data Compaction . . . . .	96
5.2.1	Data Compression Through Prototype and Feature Selection . . . . .	99
5.3	Background Material . . . . .	100
5.3.1	Computation of Frequent Features . . . . .	103
5.3.2	Distinct Subsequences . . . . .	104
5.3.3	Impact of Support on Distinct Subsequences . . . . .	104
5.3.4	Computation of Leaders . . . . .	105
5.3.5	Classification of Validation Data . . . . .	105
5.4	Preliminary Analysis . . . . .	105
5.5	Proposed Approaches . . . . .	107
5.5.1	Patterns with Frequent Items Only . . . . .	107
5.5.2	Cluster Representatives Only . . . . .	108
5.5.3	Frequent Items Followed by Clustering . . . . .	109
5.5.4	Clustering Followed by Frequent Items . . . . .	109
5.6	Implementation and Experimentation . . . . .	110
5.6.1	Handwritten Digit Data . . . . .	110
5.6.2	Intrusion Detection Data . . . . .	116
5.6.3	Simultaneous Selection of Patterns and Features . . . . .	120
5.7	Summary . . . . .	122
5.8	Bibliographic Notes . . . . .	123
	References . . . . .	123
<b>6</b>	<b>Domain Knowledge-Based Compaction . . . . .</b>	<b>125</b>
6.1	Introduction . . . . .	125
6.2	Multicategory Classification . . . . .	126
6.3	Support Vector Machine (SVM) . . . . .	126
6.4	Adaptive Boosting . . . . .	128
6.4.1	Adaptive Boosting on Prototypes for Data Mining Applications . . . . .	129
6.5	Decision Trees . . . . .	130
6.6	Preliminary Analysis Leading to Domain Knowledge . . . . .	131
6.6.1	Analytical View . . . . .	132
6.6.2	Numerical Analysis . . . . .	133
6.6.3	Confusion Matrix . . . . .	134

6.7	Proposed Method . . . . .	136
6.7.1	Knowledge-Based (KB) Tree . . . . .	136
6.8	Experimentation and Results . . . . .	137
6.8.1	Experiments Using SVM . . . . .	138
6.8.2	Experiments Using AdaBoost . . . . .	140
6.8.3	Results with AdaBoost on Benchmark Data . . . . .	141
6.9	Summary . . . . .	143
6.10	Bibliographic Notes . . . . .	144
	References . . . . .	144
<b>7</b>	<b>Optimal Dimensionality Reduction . . . . .</b>	<b>147</b>
7.1	Introduction . . . . .	147
7.2	Feature Selection . . . . .	149
7.2.1	Based on Feature Ranking . . . . .	149
7.2.2	Ranking Features . . . . .	150
7.3	Feature Extraction . . . . .	152
7.3.1	Performance . . . . .	154
7.4	Efficient Approaches to Large-Scale Feature Selection Using Genetic Algorithms . . . . .	154
7.4.1	An Overview of Genetic Algorithms . . . . .	155
7.4.2	Proposed Schemes . . . . .	158
7.4.3	Preliminary Analysis . . . . .	161
7.4.4	Experimental Results . . . . .	163
7.4.5	Summary . . . . .	170
7.5	Bibliographical Notes . . . . .	171
	References . . . . .	171
<b>8</b>	<b>Big Data Abstraction Through Multiagent Systems . . . . .</b>	<b>173</b>
8.1	Introduction . . . . .	173
8.2	Big Data . . . . .	173
8.3	Conventional Massive Data Systems . . . . .	174
8.3.1	Map-Reduce . . . . .	174
8.3.2	PageRank . . . . .	176
8.4	Big Data and Data Mining . . . . .	176
8.5	Multiagent Systems . . . . .	177
8.5.1	Agent Mining Interaction . . . . .	177
8.5.2	Big Data Analytics . . . . .	178
8.6	Proposed Multiagent Systems . . . . .	178
8.6.1	Multiagent System for Data Reduction . . . . .	178
8.6.2	Multiagent System for Attribute Reduction . . . . .	179
8.6.3	Multiagent System for Heterogeneous Data Access . . . . .	180
8.6.4	Multiagent System for Agile Processing . . . . .	181
8.7	Summary . . . . .	182
8.8	Bibliographic Notes . . . . .	182
	References . . . . .	183

Contents	xiii
<b>Appendix    Intrusion Detection Dataset—Binary Representation . . . .</b>	<b>185</b>
A.1    Data Description and Preliminary Analysis . . . . .	185
A.2    Bibliographic Notes . . . . .	189
References . . . . .	189
<b>Glossary . . . . .</b>	<b>191</b>
<b>Index . . . . .</b>	<b>193</b>

Compression Schemes for Mining Large Datasets

A Machine Learning Perspective

Ravindra Babu, T.; Murty, M.N.; Subrahmanya, S.V.

2013, XVI, 197 p. 62 illus., 3 illus. in color., Hardcover

ISBN: 978-1-4471-5606-2