Contents

Preface **xiii**

1 Introduction **1**
1.1 Defining "data mining" 1
1.2 Mining science and engineering data 2
1.3 Summary 3

2 Data Mining in Science and Engineering **5**
2.1 Astronomy and astrophysics 6
2.1.1 Characteristics of astronomy data 10
2.2 Remote sensing 12
2.2.1 Characteristics of remote sensing data 15
2.3 Biological sciences 17
2.3.1 Bioinformatics 18
2.3.2 Medicine 19
2.3.3 Characteristics of biological data 21
2.4 Security and surveillance 21
2.4.1 Biometrics 22
2.4.2 Surveillance 22
2.4.3 Network intrusion detection 23
2.4.4 Automated target recognition 24
2.4.5 Characteristics of security and surveillance data 24
2.5 Computer simulations 25
2.5.1 Characteristics of simulation data 29
2.6 Experimental physics 30
2.6.1 Characteristics of experimental physics data 34
2.7 Information retrieval 34
2.7.1 Characteristics of retrieval problems 37
2.8 Other applications 37
2.8.1 Nondestructive testing 37
2.8.2 Earth, environmental, and atmospheric sciences 37
2.8.3 Chemistry and cheminformatics 38
2.8.4 Materials science and materials informatics 38
2.8.5 Manufacturing 38
2.8.6 Scientific and information visualization 38

2.9 Summary . . . 39
2.10 Suggestions for further reading . . . 39

3 Common Themes in Mining Scientific Data 41
3.1 Types of scientific data . . . 41
3.1.1 Table data . . . 42
3.1.2 Image data . . . 42
3.1.3 Mesh data . . . 43
3.2 Characteristics of scientific data . . . 46
3.2.1 Multispectral, multisensor, multimodal data . . . 46
3.2.2 Spatiotemporal data . . . 46
3.2.3 Compressed data . . . 47
3.2.4 Streaming data . . . 47
3.2.5 Massive data . . . 47
3.2.6 Distributed data . . . 48
3.2.7 Different data formats . . . 48
3.2.8 Different output schemes . . . 49
3.2.9 Noisy, missing, and uncertain data . . . 49
3.2.10 Low-level data, higher-level objects . . . 50
3.2.11 Representation of objects in the data . . . 51
3.2.12 High-dimensional data . . . 52
3.2.13 Size and quality of labeled data . . . 52
3.3 Characteristics of scientific data analysis . . . 53
3.4 Summary . . . 55
3.5 Suggestions for further reading . . . 56

4 The Scientific Data Mining Process 57
4.1 The tasks in the scientific data mining process . . . 57
4.1.1 Transforming raw data into target data . . . 58
4.1.2 Transforming target data into preprocessed data . . . 60
4.1.3 Converting preprocessed data into transformed data . . . 62
4.1.4 Converting transformed data into patterns . . . 63
4.1.5 Converting patterns into knowledge . . . 63
4.2 General observations about the scientific data mining process . . . 64
4.3 Defining scientific data mining: The rationale . . . 65
4.4 Summary . . . 66

5 Reducing the Size of the Data 67
5.1 Sampling . . . 67
5.2 Multiresolution techniques . . . 70
5.3 Summary . . . 76
5.4 Suggestions for further reading . . . 77

6 Fusing Different Data Modalities 79
6.1 The need for data fusion . . . 80
6.2 Levels of data fusion . . . 81
6.3 Sensor-level data fusion . . . 83
6.3.1 Multiple target tracking . . . 84

6.3.2 Image registration . . . 85
6.4 Feature-level data fusion . . . 90
6.5 Decision-level data fusion . . . 91
6.6 Summary . . . 92
6.7 Suggestions for further reading . . . 92

7 Enhancing Image Data 93
7.1 The need for image enhancement . . . 94
7.2 Image denoising . . . 95
7.2.1 Filter-based approaches . . . 95
7.2.2 Wavelet-based approaches . . . 99
7.2.3 Partial differential equation–based approaches . . . 102
7.2.4 Removing multiplicative noise . . . 105
7.2.5 Problem-specific denoising . . . 106
7.3 Contrast enhancement . . . 107
7.4 Morphological techniques . . . 110
7.5 Summary . . . 111
7.6 Suggestions for further reading . . . 111

8 Finding Objects in the Data 113
8.1 Edge-based techniques . . . 114
8.1.1 The Canny edge detector . . . 117
8.1.2 Active contours . . . 117
8.1.3 The USAN approach . . . 123
8.2 Region-based methods . . . 123
8.2.1 Region splitting . . . 124
8.2.2 Region merging . . . 124
8.2.3 Region splitting and merging . . . 125
8.2.4 Clustering and classification . . . 127
8.2.5 Watershed segmentation . . . 128
8.3 Salient regions . . . 128
8.3.1 Corners . . . 129
8.3.2 Scale saliency regions . . . 129
8.3.3 Scale-invariant feature transforms . . . 131
8.4 Detecting moving objects . . . 132
8.4.1 Background subtraction . . . 132
8.4.2 Block matching . . . 133
8.5 Domain-specific approaches . . . 135
8.6 Identifying unique objects . . . 136
8.7 Postprocessing for object identification . . . 138
8.8 Representation of the objects . . . 138
8.9 Summary . . . 139
8.10 Suggestions for further reading . . . 139

9 Extracting Features Describing the Objects 141
9.1 General requirements for a feature . . . 142
9.2 Simple features . . . 144

9.3 Shape features . . . 146
9.4 Texture features . . . 149
9.5 Problem-specific features . . . 153
9.6 Postprocessing the features . . . 157
9.7 Summary . . . 159
9.8 Suggestions for further reading . . . 160

10 Reducing the Dimension of the Data 161
10.1 The need for dimension reduction . . . 162
10.2 Feature transform methods . . . 164
10.2.1 Principal component analysis . . . 164
10.2.2 Extensions of principal component analysis . . . 166
10.2.3 Random projections . . . 166
10.2.4 Multidimensional scaling . . . 167
10.2.5 FastMap . . . 167
10.2.6 Self-organizing maps . . . 168
10.3 Feature subset selection methods . . . 168
10.3.1 Filters for feature selection . . . 169
10.3.2 Wrapper methods . . . 171
10.3.3 Feature selection for regression . . . 172
10.4 Domain-specific methods . . . 172
10.5 Representation of high-dimensional data . . . 174
10.6 Summary . . . 175
10.7 Suggestions for further reading . . . 175

11 Finding Patterns in the Data 177
11.1 Clustering . . . 178
11.1.1 Partitional algorithms . . . 179
11.1.2 Hierarchical clustering . . . 180
11.1.3 Graph-based clustering . . . 180
11.1.4 Observations and further reading . . . 183
11.2 Classification . . . 184
11.2.1 k-nearest neighbor classifier . . . 185
11.2.2 Naïve Bayes classifier . . . 185
11.2.3 Decision trees . . . 186
11.2.4 Neural networks . . . 189
11.2.5 Support vector machines . . . 193
11.2.6 Ensembles of classifiers . . . 194
11.2.7 Observations and further reading . . . 196
11.3 Regression . . . 199
11.3.1 Statistical techniques . . . 200
11.3.2 Machine learning techniques . . . 200
11.4 Association rules . . . 201
11.5 Tracking . . . 202
11.6 Outlier or anomaly detection . . . 204
11.7 Related topics . . . 205
11.7.1 Distance metrics . . . 205

11.7.2 Optimization techniques 206
11.8 Summary 207
11.9 Suggestions for further reading 207

12 Visualizing the Data and Validating the Results 209
12.1 Visualizing table data 210
12.1.1 Box plots 210
12.1.2 Scatter plots 211
12.1.3 Parallel plots 212
12.2 Visualizing image and mesh data 214
12.3 Validation of results 215
12.4 Summary 218
12.5 Suggestions for further reading 219

13 Scientific Data Mining Systems 221
13.1 Software for specific tasks in scientific data mining 222
13.2 Software systems for scientific data mining 222
13.2.1 Diamond Eye 223
13.2.2 Algorithm Development and Mining System 224
13.2.3 Sapphire 224
13.3 Summary 227

14 Lessons Learned, Challenges, and Opportunities 229
14.1 Guidelines for getting started 230
14.2 Challenges and opportunities 232
14.3 Concluding remarks 233

Bibliography 235

Index 279