

Contents

1	Introduction	1
2	Multivariate Classification and Pattern Recognition	4
2.1	Data Preprocessing	6
2.2	Mapping and Display	8
2.3	Clustering	14
2.4	Classification	18
2.4.1	k -Nearest Neighbor	19
2.4.2	Discriminant Analysis	20
2.4.3	Partial Least Squares	21
2.4.4	SIMCA	23
2.5	Practical Considerations	24
2.6	Applications of Pattern Recognition Techniques	26
2.6.1	Fuel Spill Identification	27
2.6.2	Sorting Plastics for Recycling	36
2.7	A Closer Look at Feature Selection	44
2.7.1	Exhaustive Search	44
2.7.2	Weight-Based Selection	45
2.8	Conclusion	48
3	Genetic Algorithms	49
3.1	The Simple Genetic Algorithm (SGA)	49

3.1.1	Schemata	51
3.1.2	The Schema Theorem	52
3.1.3	Optimization	55
3.2	Applying a GA	58
3.3	Customizing a GA	58
3.3.1	Encoding	59
3.3.2	The Fitness Function	59
3.3.3	Selection	60
3.3.4	Reproduction	60
3.3.5	Insertion	63
3.3.6	Other Operators	63
3.3.7	Controlling Parameters	64
3.4	Conclusion	64
4	A Genetic Algorithm for Feature Selection	65
4.1	Basic PCKaNN	66
4.1.1	Population	67
4.1.2	Fitness Function	68
4.1.3	Selection	71
4.1.4	Crossover	72
4.1.5	Mutation	72
4.1.6	Insertion	73
4.1.7	End Criterion	73
4.2	Advanced PCKaNN	74
4.2.1	Culling	74
4.2.2	Ordinal PCKaNN	75
4.2.3	Taking the PCA out of PCKaNN	75
4.2.4	A Clustering GA	78
4.2.5	Incorporation of Transverse Learning	81

4.3	Conclusion	83
5	Analysis of Complex Chromatographic and Spectroscopic Data	84
5.1	Chemical Communication Among Social Insects	84
5.1.1	Experimental	85
5.1.2	Results	86
5.2	Quality Control of Pharmaceutical Tablets	95
5.2.1	Un-normalized Data	97
5.2.2	Normalized Data	101
5.2.3	Conclusion	103
5.3	Raman Spectroscopy of Hard, Soft, and Tropical Woods	104
5.3.1	Results	106
5.4	Fuel Spill Identification	113
6	Extracting Information from Biological Tissue	122
6.1	DNA Microarray Data	123
6.1.1	Small Round Blue Cell Tumors	124
6.1.2	Leukemia	134
6.2	Proteomic Data	141
6.2.1	Ovarian Cancer	141
6.3	Conclusion	147
7	Extracting Information from Biological Compounds	148
7.1	Musk Odorants	148
7.1.1	Experimental	150
7.1.2	TAE Descriptors	150
7.1.3	Wavelet/PEST Descriptors	154
8	Conclusion	174
A	MATLAB Implementation	177

A.1	Data Formatting Issues	177
A.2	Data Organization within MATLAB	178
A.3	Building a Project	179
A.4	Building a Run Plan	181
A.5	Activating a Run Plan	187
A.6	Helpful Hints and Advanced Techniques	187
A.6.1	Population Parameters	190
A.6.2	Tuning the k -Value	192
A.6.3	Ordinal Fitness	195
A.6.4	Clustering	195
A.6.5	Transverse Learning	196
A.6.6	Kohonen Neural Network	197
A.6.7	Other Non-PCA Approaches	198
A.7	Investigating the Results of a Run	199
A.7.1	Using <code>projview</code>	199
A.7.2	Using <code>somproj</code>	203
A.8	Manually Editing the Project	210
A.9	Do-it-yourself Functions	212
A.9.1	<code>prune</code> and <code>process</code>	212
A.10	Conclusion	213
B	Dynamics Simulations and Molecular Binding	216
B.1	Data Formatting Issues	217
B.2	Building a Project	217
B.3	Determining Similarity from Molecular Coordinates	218
B.4	Visualization by PCA	219
B.5	Feature Selection	219
B.6	An Example: Oxygen Binding by Myoglobin	220
B.7	Conclusion	223