

Advanced Information and Knowledge Processing

Lipo Wang · Xiuju Fu

Data Mining with Computational Intelligence

With 72 Figures and 65 Tables

Lipo Wang
Nanyang Technological University
School of Electrical and Electronical Engineering
Block S1, Nanyang Avenue,
639798 Singapore, Singapore
elpwang@ntu.edu.sg

Xiuju Fu
Institute of High Performance Computing,
Software and Computing, Science Park 2,
The Capricorn
Science Park Road 01-01
117528 Singapore, Singapore
fuxj@pmail.ntu.edu.sg

Series Editors

Xindong Wu
Lakhmi Jain

Library of Congress Control Number: 200528948

ACM Computing Classification (1998): H.2.8., I.2

ISBN-10 3-540-24522-7 Springer Berlin Heidelberg New York

ISBN-13 978-3-540-24522-3 Springer Berlin Heidelberg New York

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilm or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permission for use must always be obtained from Springer. Violations are liable for prosecution under the German Copyright Law.

Springer is a part of Springer Science+Business Media
springeronline.com

© Springer-Verlag Berlin Heidelberg 2005
Printed in Germany

The use of general descriptive names, registered names, trademarks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

Cover design: KünkelLopka, Heidelberg
Typesetting: Camera ready by the authors
Production: LE-TeX Jelonek, Schmidt & Vöckler GbR, Leipzig
Printed on acid-free paper 45/3142/YL - 5 4 3 2 1 0

Preface

Nowadays data accumulate at an alarming speed in various storage devices, and so does valuable information. However, it is difficult to understand information hidden in data without the aid of data analysis techniques, which has provoked extensive interest in developing a field separate from machine learning. This new field is data mining.

Data mining has successfully provided solutions for finding information from data in bioinformatics, pharmaceuticals, banking, retail, sports and entertainment, etc. It has been one of the fastest growing fields in the computer industry. Many important problems in science and industry have been addressed by data mining methods, such as neural networks, fuzzy logic, decision trees, genetic algorithms, and statistical methods.

This book systematically presents how to utilize fuzzy neural networks, multi-layer perceptron (MLP) neural networks, radial basis function (RBF) neural networks, genetic algorithms (GAs), and support vector machines (SVMs) in data mining tasks. Fuzzy logic mimics the imprecise way of reasoning in natural languages and is capable of tolerating uncertainty and vagueness. The MLP is perhaps the most popular type of neural network used today. The RBF neural network has been attracting great interest because of its locally tuned response in RBF neurons like biological neurons and its global approximation capability. This book demonstrates the power of GAs in feature selection and rule extraction. SVMs are well known for their excellent accuracy and generalization abilities.

We will describe data mining systems which are composed of data pre-processing, knowledge-discovery models, and a data-concept description. This monograph will enable both new and experienced data miners to improve their practices at every step of data mining model design and implementation.

Specifically, the book will describe the state of the art of the following topics, including both work carried out by the authors themselves and by other researchers:

- Data mining tools, i.e., neural networks, support vector machines, and genetic algorithms with application to data mining tasks.
- Data mining tasks including data dimensionality reduction, classification, and rule extraction.

Lipo Wang wishes to sincerely thank his students, especially Feng Chu, Yakov Frayman, Guosheng Jin, Kok Keong Teo, and Wei Xie, for the great pleasure of collaboration, and for carrying out research and contributing to this book. Thanks are due to Professors Zhiping Lin, Kai-Ming Ting, Chunru Wan, Ron (Zhengrong) Yang, Xin Yao, and Jacek M. Zurada for many helpful discussions and for the opportunities to work together. Xiuju Fu wishes to express gratitude to Dr. Gih Guang Hung, Liping Goh, Professors Chongjin Ong and S. Sathiya Keerthi for their discussions and supports in the research work. We also express our appreciation for the support and encouragement from Professor L.C. Jain and Springer Editor Ralf Gerstner.

Singapore,
May 2005

Lipo Wang
Xiuju Fu

Contents

1	Introduction	1
1.1	Data Mining Tasks	2
1.1.1	Data Dimensionality Reduction	2
1.1.2	Classification and Clustering	4
1.1.3	Rule Extraction	5
1.2	Computational Intelligence Methods for Data Mining	6
1.2.1	Multi-layer Perceptron Neural Networks	6
1.2.2	Fuzzy Neural Networks	8
1.2.3	RBF Neural Networks	9
1.2.4	Support Vector Machines	14
1.2.5	Genetic Algorithms	20
1.3	How This Book is Organized	21
2	MLP Neural Networks for Time-Series Prediction and Classification	25
2.1	Wavelet MLP Neural Networks for Time-series Prediction	25
2.1.1	Introduction to Wavelet Multi-layer Neural Network	25
2.1.2	Wavelet	26
2.1.3	Wavelet MLP Neural Network	28
2.1.4	Experimental Results	29
2.2	Wavelet Packet MLP Neural Networks for Time-series Prediction	33
2.2.1	Wavelet Packet Multi-layer Perceptron Neural Networks	33
2.2.2	Weight Initialization with Clustering	33
2.2.3	Mackey-Glass Chaotic Time-Series	35
2.2.4	Sunspot and Laser Time-Series	36
2.2.5	Conclusion	37
2.3	Cost-Sensitive MLP	38
2.3.1	Standard Back-propagation	38
2.3.2	Cost-sensitive Back-propagation	40
2.3.3	Experimental Results	42

2.4	Summary	43
3	Fuzzy Neural Networks for Bioinformatics	45
3.1	Introduction	45
3.2	Fuzzy Logic	45
3.2.1	Fuzzy Systems	45
3.2.2	Issues in Fuzzy Systems	51
3.3	Fuzzy Neural Networks	52
3.3.1	Knowledge Processing in Fuzzy and Neural Systems ...	52
3.3.2	Integration of Fuzzy Systems with Neural Networks	52
3.4	A Modified Fuzzy Neural Network	53
3.4.1	The Structure of the Fuzzy Neural Network	53
3.4.2	Structure and Parameter Initialization	55
3.4.3	Parameter Training	58
3.4.4	Structure Training	60
3.4.5	Input Selection	60
3.4.6	Partition Validation	61
3.4.7	Rule Base Modification	62
3.5	Experimental Evaluation Using Synthesized Data Sets	63
3.5.1	Descriptions of the Synthesized Data Sets	64
3.5.2	Other Methods for Comparisons	66
3.5.3	Experimental Results	68
3.5.4	Discussion	70
3.6	Classifying Cancer from Microarray Data	71
3.6.1	DNA Microarrays	71
3.6.2	Gene Selection	75
3.6.3	Experimental Results	77
3.7	A Fuzzy Neural Network Dealing with the Problem of Small Disjuncts	81
3.7.1	Introduction	81
3.7.2	The Structure of the Fuzzy Neural Network Used	81
3.7.3	Experimental Results	85
3.8	Summary	85
4	An Improved RBF Neural Network Classifier	97
4.1	Introduction	97
4.2	RBF Neural Networks for Classification	98
4.2.1	The Pseudo-inverse Method	100
4.2.2	Comparison between the RBF and the MLP	101
4.3	Training a Modified RBF Neural Network	102
4.4	Experimental Results	105
4.4.1	Iris Data Set	106
4.4.2	Thyroid Data Set	106
4.4.3	Monk3 Data Set	107
4.4.4	Breast Cancer Data Set	108

4.4.5	Mushroom Data Set	108
4.5	RBF Neural Networks Dealing with Unbalanced Data	110
4.5.1	Introduction	110
4.5.2	The Standard RBF Neural Network Training Algorithm for Unbalanced Data Sets	111
4.5.3	Training RBF Neural Networks on Unbalanced Data Sets	112
4.5.4	Experimental Results	113
4.6	Summary	114
5	Attribute Importance Ranking for Data Dimensionality Reduction	117
5.1	Introduction	117
5.2	A Class-Separability Measure	119
5.3	An Attribute-Class Correlation Measure	121
5.4	The Separability-correlation Measure for Attribute Importance Ranking	121
5.5	Different Searches for Ranking Attributes	122
5.6	Data Dimensionality Reduction	123
5.6.1	Simplifying the RBF Classifier Through Data Dimensionality Reduction	124
5.7	Experimental Results	125
5.7.1	Attribute Ranking Results	125
5.7.2	Iris Data Set	126
5.7.3	Monk3 Data Set	127
5.7.4	Thyroid Data Set	127
5.7.5	Breast Cancer Data Set	128
5.7.6	Mushroom Data Set	128
5.7.7	Ionosphere Data Set	130
5.7.8	Comparisons Between Top-down and Bottom-up Searches and with Other Methods	132
5.8	Summary	137
6	Genetic Algorithms for Class-Dependent Feature Selection	145
6.1	Introduction	145
6.2	The Conventional RBF Classifier	148
6.3	Constructing an RBF with Class-Dependent Features	149
6.3.1	Architecture of a Novel RBF Classifier	149
6.4	Encoding Feature Masks Using GAs	151
6.4.1	Crossover and Mutation	152
6.4.2	Fitness Function	152
6.5	Experimental Results	152
6.5.1	Glass Data Set	153
6.5.2	Thyroid Data Set	154
6.5.3	Wine Data Set	155

6.6	Summary	155
7	Rule Extraction from RBF Neural Networks	157
7.1	Introduction	157
7.2	Rule Extraction Based on Classification Models	160
7.2.1	Rule Extraction Based on Neural Network Classifiers ..	161
7.2.2	Rule Extraction Based on Support Vector Machine Classifiers	163
7.2.3	Rule Extraction Based on Decision Trees	163
7.2.4	Rule Extraction Based on Regression Models	164
7.3	Components of Rule Extraction Systems	164
7.4	Rule Extraction Combining GAs and the RBF Neural Network	165
7.4.1	The Procedure of Rule Extraction	167
7.4.2	Simplifying Weights	168
7.4.3	Encoding Rule Premises Using GAs	168
7.4.4	Crossover and Mutation	169
7.4.5	Fitness Function	170
7.4.6	More Compact Rules	170
7.4.7	Experimental Results	170
7.4.8	Summary	174
7.5	Rule Extraction by Gradient Descent	175
7.5.1	The Method	175
7.5.2	Experimental Results	177
7.5.3	Summary	180
7.6	Rule Extraction After Data Dimensionality Reduction	180
7.6.1	Experimental Results	181
7.6.2	Summary	184
7.7	Rule Extraction Based on Class-dependent Features	185
7.7.1	The Procedure of Rule Extraction	185
7.7.2	Experimental Results	185
7.7.3	Summary	187
8	A Hybrid Neural Network For Protein Secondary Structure Prediction	189
8.1	The PSSP Basics	189
8.1.1	Basic Protein Building Unit — Amino Acid	189
8.1.2	Types of the Protein Secondary Structure	189
8.1.3	The Task of the Prediction	191
8.2	Literature Review of the PSSP problem	193
8.3	Architectural Design of the HNNP	195
8.3.1	Process Flow at the Training Phase	195
8.3.2	Process Flow at the Prediction Phase	197
8.3.3	First Stage: the Q2T Prediction	197
8.3.4	Sequence Representation	199
8.3.5	Distance Measure Method for Data — WINDist	201

8.3.6	Second Stage: the T2T Prediction	205
8.3.7	Sequence Representation	207
8.4	Experimental Results	209
8.4.1	Experimental Data set	209
8.4.2	Accuracy Measure	210
8.4.3	Experiments with the Base and Alternative Distance Measure Schemes	213
8.4.4	Experiments with the Window Size and the Cluster Purity	214
8.4.5	T2T Prediction — the Final Prediction	216
9	Support Vector Machines for Prediction	225
9.1	Multi-class SVM Classifiers	225
9.2	SVMs for Cancer Type Prediction	226
9.2.1	Gene Expression Data Sets	226
9.2.2	A T-test-Based Gene Selection Approach	226
9.3	Experimental Results	227
9.3.1	Results for the SRBCT Data Set	227
9.3.2	Results for the Lymphoma Data Set	231
9.4	SVMs for Protein Secondary Structure Prediction	233
9.4.1	Q2T prediction	233
9.4.2	T2T prediction	235
9.5	Summary	236
10	Rule Extraction from Support Vector Machines	237
10.1	Introduction	237
10.2	Rule Extraction	240
10.2.1	The Initial Phase for Generating Rules	240
10.2.2	The Tuning Phase for Rules	242
10.2.3	The Pruning Phase for Rules	243
10.3	Illustrative Examples	243
10.3.1	Example 1 — Breast Cancer Data Set	243
10.3.2	Example 2 — Iris Data Set	244
10.4	Experimental Results	245
10.5	Summary	246
A	Rules extracted for the Iris data set	251
	References	253
	Index	275