Bruno Baruque and Emilio Corchado

Fusion Methods for Unsupervised Learning Ensembles

View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Gestion del Repositorio Documental de la Universidad de...

Studies in Computational Intelligence, Volume 322

Editor-in-Chief
Prof. Janusz Kacprzyk
Systems Research Institute
Polish Academy of Sciences
ul. Newelska 6
01-447 Warsaw
Poland
E-mail: kacprzyk@ibspan.waw.pl

Further volumes of this series can be found on our homepage: springer.com

Vol. 299. Vassil Sgurev, Mincho Hadjiski, and Janusz Kacprzyk (Eds.)
Intelligent Systems: From Theory to Practice, 2010
ISBN 978-3-642-13427-2

Vol. 300. Baoding Liu (Ed.)
Uncertainty Theory, 2010
ISBN 978-3-642-13958-1

Vol. 301. Giuliano Armano, Marco de Gemmis, Giovanni Semeraro, and Eloisa Vargiu (Eds.)
Intelligent Information Access, 2010
ISBN 978-3-642-13999-4

Vol. 302. Bijaya Ketan Panigrahi, Ajith Abraham, and Swagatam Das (Eds.)
Computational Intelligence in Power Engineering, 2010
ISBN 978-3-642-14012-9

Vol. 303. Joachim Diederich, Cengiz Gunay, and James M. Hogan
Recruitment Learning, 2010
ISBN 978-3-642-14027-3

Vol. 304. Anthony Finn and Lakhmi C. Jain (Eds.)
Innovations in Defence Support Systems, 2010
ISBN 978-3-642-14083-9

Vol. 305. Stefania Montani and Lakhmi C. Jain (Eds.)
Successful Case-Based Reasoning Applications-1, 2010
ISBN 978-3-642-14077-8

Vol. 306. Tru Hoang Cao
Conceptual Graphs and Fuzzy Logic, 2010
ISBN 978-3-642-14086-0

Vol. 307. Anupam Shukla, Ritu Tiwari, and Rahul Kala
Towards Hybrid and Adaptive Computing, 2010
ISBN 978-3-642-14343-4

Vol. 308. Roger Nkambou, Jacqueline Bourdeau, and Riichiro Mizoguchi (Eds.)
Advances in Intelligent Tutoring Systems, 2010
ISBN 978-3-642-14362-5

Vol. 309. Isabelle Bichindaritz, Lakhmi C. Jain, Sachin Vaidya, and Ashlesha Jain (Eds.)
Computational Intelligence in Healthcare 4, 2010
ISBN 978-3-642-14463-9

Vol. 310. Dipti Srinivasan and Lakhmi C. Jain (Eds.)
Innovations in Multi-Agent Systems and Applications – 1, 2010
ISBN 978-3-642-14434-9

Vol. 311. Juan D. Velásquez and Lakhmi C. Jain (Eds.)
Advanced Techniques in Web Intelligence, 2010
ISBN 978-3-642-14460-8

Vol. 312. Patricia Melin, Janusz Kacprzyk, and Witold Pedrycz (Eds.)
Soft Computing for Recognition based on Biometrics, 2010
ISBN 978-3-642-15110-1

Vol. 313. Imre J. Rudas, János Fodor, and Janusz Kacprzyk (Eds.)
Computational Intelligence in Engineering, 2010
ISBN 978-3-642-15219-1

Vol. 314. Lorenzo Magnani, Walter Carnielli, and Claudio Pizzi (Eds.)
Model-Based Reasoning in Science and Technology, 2010
ISBN 978-3-642-15222-1

Vol. 315. Mohammad Essaaidi, Michele Malgeri, and Costin Badica (Eds.)
Intelligent Distributed Computing IV, 2010
ISBN 978-3-642-15210-8

Vol. 316. Philipp Wolfrum
Information Routing, Correspondence Finding, and Object Recognition in the Brain, 2010
ISBN 978-3-642-15253-5

Vol. 317. Roger Lee (Ed.)
Computer and Information Science 2010
ISBN 978-3-642-15404-1

Vol. 318. Oscar Castillo, Janusz Kacprzyk, and Witold Pedrycz (Eds.)
Soft Computing for Intelligent Control and Mobile Robotics, 2010
ISBN 978-3-642-15533-8

Vol. 319. Takayuki Ito, Minjie Zhang, Valentin Robu, Shaheen Fatima, Tokuro Matsuo, and Hirofumi Yamaki (Eds.)
Innovations in Agent-Based Complex Automated Negotiations, 2010
ISBN 978-3-642-15611-3

Vol. 320. xxx

Vol. 321. Dimitri Plemenos and Georgios Miaoulis (Eds.)
Intelligent Computer Graphics 2010
ISBN 978-3-642-15689-2

Vol. 322. Bruno Baruque and Emilio Corchado
Fusion Methods for Unsupervised Learning Ensembles, 2010
ISBN 978-3-642-16204-6

Bruno Baruque and Emilio Corchado

Fusion Methods for Unsupervised Learning Ensembles

Springer

Dr. Bruno Baruque
Departamento de Ingeniería Civil
Escuela Politécnica Superior
Universidad de Burgos
Avda. Cantabria, s/n
09006 Burgos, Spain
E-mail: bbaruque@ubu.es

Dr. Emilio Corchado
Departamento de Informática y Automática
Facultad de Ciencias
Universidad de Salamanca
Plaza de la Merced, s/n
37008 Salamanca
Spain
E-mail: escorchado@usal.es

ISBN 978-3-642-16204-6 e-ISBN 978-3-642-16205-3

DOI 10.1007/978-3-642-16205-3

Studies in Computational Intelligence ISSN 1860-949X

Library of Congress Control Number: 2010936510

© 2010 Springer-Verlag Berlin Heidelberg

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 to prosecution under the German Copyright Law.

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.

Typeset & *Cover Design:* Scientific Publishing Services Pvt. Ltd., Chennai, India.

Printed on acid-free paper

9 8 7 6 5 4 3 2 1

springer.com

Abstract

The application of a “committee of experts” or ensemble learning to artificial neural networks that apply unsupervised learning techniques is widely considered to enhance the effectiveness of such networks greatly. This book examines in one of its chapters the potential of the ensemble meta-algorithm by describing and testing a technique based on the combination of ensembles and statistical PCA that is able to determine the presence of outliers in high-dimensional data sets and to minimize outlier effects in the final results. After that, it presents its central contribution, which consists on an algorithm for the ensemble fusion of topology-preserving maps, referred to as Weighted Voting Superposition (WeVoS), which has been devised to improve data exploration by 2-D visualization over multi-dimensional data sets. This generic algorithm is applied in combination with several other models taken from the family of topology preserving maps, such as the SOM, ViSOM, SIM and Max-SIM. A range of quality measures for topology preserving maps that are proposed in the literature are used to validate and compare WeVoS with other algorithms. The experimental results demonstrate that, in the majority of cases, the WeVoS algorithm outperforms earlier map-fusion methods and the simpler versions of the algorithm with which it is compared. All the algorithms are tested in different artificial data sets and in several of the most common machine-learning data sets in order to corroborate their theoretical properties. Moreover, a real-life case-study taken from the food industry demonstrates the practical benefits of their applications to more complex problems.

Contents

1 Introduction 1

1.1 Background 1

1.2 Contributions 2

1.3 Organization 3

2 Modelling Human Learning: Artificial Neural Networks 5

2.1 The Human Learning Process 5

2.1.1 The Biological Neuron 7

2.2 Artificial Neural Networks 8

2.2.1 Learning Algorithms in Neural Networks 9

2.2.2 Reinforcement Learning 9

2.2.3 Supervised Learning 10

2.2.4 Unsupervised Learning 10

2.3 Hebbian Learning 11

2.4 Hebbian Learning and Statistics 12

2.4.1 Principal Component Analysis 13

2.4.2 Oja's Models 16

2.4.3 Negative Feedback Network 16

2.5 Competitive Learning 18

2.5.1 The Self-Organizing Map 18

2.5.2 The Visually Induced SOM 21

2.5.3 The Scale Invariant Map 23

2.5.4 Assessing Quality of Training of Topology Preserving Models 26

2.6 Conclusions 29

3 The Committee of Experts Approach: Ensemble Learning 31

3.1 The Ensemble Meta-algorithm 31

3.1.1 The Classification Problem 31

3.1.2 Ensemble General Concepts 32

3.2 Commonly Used Ensemble Models . . . 34
3.2.1 Bagging . . . 34
3.2.2 Boosting . . . 37
3.2.3 Mixture of Experts . . . 41
3.3 Combining Ensemble Results . . . 43
3.3.1 Selection . . . 43
3.3.2 Voting Combinations . . . 44
3.3.3 Linear Combinations . . . 44
3.4 Ensembles of Artificial Neural Networks . . . 45
3.4.1 Supervised ANNs . . . 46
3.4.2 Unsupervised ANNs . . . 46
3.5 Conclusions . . . 47

4 Use of Ensembles for Outlier Overcoming . . . 49
4.1 Introduction . . . 49
4.2 The Outlier Problem . . . 49
4.3 The Re-sampling PCA Algorithm . . . 51
4.3.1 Ensemble Construction . . . 51
4.3.2 Results Combination . . . 52
4.4 Experiments and Results . . . 53
4.4.1 Artificial Data Set . . . 54
4.4.2 Real Life Data Set: Liver Disorder Data Set . . . 59
4.4.3 Real Life Data Set: Food Industry Application . . . 60
4.4.4 ANNs Approach . . . 65
4.5 Conclusions . . . 66

5 Ensembles of Topology Preserving Maps . . . 67
5.1 Introduction . . . 67
5.2 Problem Statement . . . 67
5.3 Topology-Preserving Map Combination Models . . . 68
5.3.1 Previously Proposed Models for SOM Ensemble Summarization . . . 69
5.3.2 Novel Proposed Model: Superposition . . . 74
5.3.3 Discussion of the Fusion Models . . . 76
5.4 Experiments and Results . . . 77
5.4.1 Comparison between Single Model and Ensemble as Classifiers . . . 77
5.4.2 Comparison between Fusion by Distance and Fusion by Similarity Algorithms . . . 79
5.4.3 Comparison between Fusion by Distance and Superposition Algorithms . . . 83
5.4.4 Comparison between Bagging and Boosting as Ensemble Training Algorithm . . . 87
5.4.5 Food Industry Application . . . 91
5.5 Conclusions . . . 94

6 A Novel Fusion Algorithm for Topology-Preserving Maps 95
6.1 Introduction 95
6.2 Fusion by Ordered Similarity 95
6.3 The Weighted Voting Superposition Algorithm 96
6.3.1 WeVoS Algorithm 97
6.3.2 Discussion 99
6.4 Application of WeVoS to Different Models 100
6.4.1 Topology-Preserving Models 101
6.4.2 Ensemble Models 101
6.4.3 Quality Measures 102
6.5 Experiments and Results 103
6.5.1 Comparison of Fusion Algorithms over a 1-D SOM 103
6.5.2 Comparison of Fusion Algorithms over the 2-D SOM 104
6.5.3 Comparison of Fusion Algorithms over the ViSOM 111
6.5.4 Comparison of Fusion Algorithms over the SIM and Max-SIM 112
6.5.5 Comparison of Fusion Algorithms When Combined with Boosting 114
6.5.6 Food Industry Application 118
6.6 Conclusions 121
6.7 Future Work 122

7 Conclusions 123
7.1 Concluding Remarks 123
7.2 Future Research Work 124

A The Cured Ham Data Set 127
A.1 Sensory Analysis and Instruments 127
A.2 E-Nose Odour Recognition 127
A.3 The Cured Ham Data Sets 129
A.3.1 Ham Data Set 1 129
A.3.2 Ham Data Set 2 129
A.4 Analysis of the Data Set 130

B Table of Experiments 131
B.1 Chapter 4 131
B.2 Chapter 5 132
B.3 Chapter 6 134

References 137

List of Figures

2.1 Depiction of the human brain 5
2.2 Diagram of a biological neuron 7
2.3 Basic architecture of a feed-forward ANN 11
2.4 PCA of a multivariate Gaussian distribution centred at [1,3] 13
2.5 Basic architecture of a negative feedback network 18
2.6 Conceptual diagram of a 2D-SOM representation of 3D data set 19
2.7 Updating of the characteristics vectors of SOM neurons 20
2.8 Comparison of the representation of the Iris data set by a SOM and a ViSOM 22
2.9 Contraction or Expansion force for the updating of the ViSOM neurons 23
2.10 A SIM trained on uniformly distributed data 24
2.11 Results for the SIM when the learning rate is increased 25

3.1 Schematic diagram of the Bagging process 35
3.2 Mixture of Experts architecture 42

4.1 Principal components of a data set with and without outliers 50
4.2 Eigenvectors determining the direction of higher variance in the data set with and without outliers (100 samples data set) 54
4.3 Average for each of the principal components as a result of averaging the directions obtained by the Re-PCA method using 100 samples 55
4.4 Eigenvectors determining the direction of higher variance in the data set with and without outliers (30 samples data set) 57

4.5 Average for each of the principal components as a result of averaging the directions obtained by the Re-PCA method using 30 samples 59
4.6 Directions calculated by the Re-PCA ensemble over the 'BUPA' data set 60
4.7 The ham data set projected over the principal components obtained form a single statistical PCA (without outliers) 61
4.8 The ham data set projected over the principal components obtained from a single statistical PCA (including 4 outliers) 62
4.9 The ham data set projected over the principal components obtained from a Re-PCA of 80 samples (including 4 outliers) 63
4.10 The ham data set projected over the principal components obtained from a Re-PCA of 120 samples (including 4 outliers) 64

5.1 Alignment of three networks and their subsequent merging into one final map 70
5.2 Topology approximation of the SOM and the SOM Fusion by Similarity algorithm to the "doughnut" artificial dataset . . 74
5.3 Comparison of the SOM and Max-SIM bagged models when trained on a radial data set. 78
5.4 2-D representation of the compared model's grid represented over the Iris data set 80
5.5 Topographic Error calculated over the single SOM and the Fusion by Euclidean Distance 81
5.6 Ensemble of 5 SOMs trained over the Iris data set and the final Superposition from that ensemble 83
5.7 The Iris data set as represented by Single SOM, Fusion by Distance, Superposition and Superposition + Re-Labelling . . . 84
5.8 Comparative representation of the Cancer data set using SOM and ViSOM and Superposition+Re-Labelling 86
5.9 Maps obtained by a single ViSOM and the three Fusion algorithms over the Iris dataset. Ensemble trained using the AdaBoost.M2 algorithm 88
5.10 Visual Comparison of the Single ViSOM and the three presented Fusion algorithms using the ham data set 92
5.11 Same Superposition+Re-Labelling as previously shown with more detailed labelling of neurons 92

6.1 Schematic diagram of the weighted voting in WeVoS 98
6.2 Distortion measured over four different fusion algorithms for a 1-D ensemble of SOMs 103

6.3 Visual comparison of the five models -four ensemble fusion models and the single model- discussed in the book 105
6.4 The 4 quality measures obtained from the 4 different summarization algorithms and the single SOM trained on the Iris data set 106
6.5 The 4 quality measures obtained from the 4 different summarization algorithms and the single SOM trained on the Wine data set 108
6.6 The 4 quality measures obtained from the 4 different summarization algorithms and the single SOM trained on the Wisconsin Breast Cancer data set. 109
6.7 The 4 quality measures obtained from the 4 different summarization algorithms and the single ViSOM trained on the Wine data set. 111
6.8 The single SIM and the three summarizations for the same 6-network ensemble trained over the circular data set employing the SIM learning Algorithm 113
6.9 The 4 quality measures obtained from the 4 different summarization algorithms and the single SIM trained on the artificial circular data set. 115
6.10 The 4 quality measures obtained from the 4 different summarization algorithms and the single Max-SIM trained on the artificial circular data set. 116
6.11 Different maps obtained by training the ensemble of SOM maps using a different meta-algorithm -Bagging or AdaBoost- and finally applying the WeVoS algorithm 117
6.12 Results of all 4 quality measures applied to the 3 different ensemble training algorithms -single, Bagging, AdaBoost- and the single SOM trained using the Iris data set. 118
6.13 Visual comparison between a PCA analysis, a single SOM and the three fusion algorithms performed over the same ensemble of 5 SOM; all trained on the Ham dataset 119
6.14 The 4 quality measures obtained from the 3 different ensemble training algorithms and the single SOM trained on the Ham data set. 121

A.1 Human odour recognition process compared with E-Nose odour recognition process 128
A.2 E-Nose αFOX 4000 128

List of Tables

4.1 Percentage of information captured by each of the principal components (selecting 50 points but excluding outliers) 56
4.2 Percentage of information captured by each of the principal components (selecting 50 points and including outliers) 56
4.3 Percentage of information captured by each of the principal components in the first part of the experiment (30 samples without outliers) 57
4.4 Percentage of information captured by each of the principal components in the second part of the experiment (30 samples with outliers) 58
4.5 Percentage of information captured by each of the principal components in the first experiment (176 samples without outliers). Simple PCA is applied 61
4.6 Percentage of information captured by each of the principal components in the first experiment (176 samples including outliers). Simple PCA is applied 62
4.7 Percentage of information captured by each of the principal Components in the third experiment (80 samples, including outliers). Re-PCA is applied 64
4.8 Percentage of information captured by each of the principal Components in the third experiment (120 samples, including outliers). Re-PCA is applied 65

5.1 Classification accuracy of three different models applied to the radial data set (without outliers) 78
5.2 Classification accuracy of three different models applied to the radial data set (including 20 outliers) 79

5.3 Comparison of SOM and ViSOM using an ensemble of 10 maps to calculate the MQE for the average of all 10 maps, the 10-map Fusion using the Euclidean distance algorithm, and the 10-map Fusion using the Voronoi similarity algorithm (average of five different cross-validation tests) 82

5.4 Comparison of the Distortion of the Fus. By Euclidean Distance and Fus. By Voronoi Similarity of an ensemble of SOM and ViSOM 82

5.5 Classification accuracy of the different models obtained from a SOM and ViSOM ensemble for the Iris data set 85

5.6 Classification accuracy of the different models obtained from a SOM and ViSOM ensemble on the Cancer data set 87

5.7 Percentage of correct recognition of samples in the Iris data set training the ensemble with Bagging algorithm 88

5.8 Percentage of correct recognition of samples in the Iris data set training the ensemble with AdaBoost.M2 algorithm 89

5.9 Percentage of correct recognition of samples in the Wisconsin Breast Cancer data set training the ensemble with Bagging algorithm 89

5.10 Percentage of correct recognition of samples in the Wisconsin Breast Cancer data set training the ensemble with AdaBoost.M1 algorithm 90

5.11 Model classification accuracy over the Ham data set training ensembles with the Bagging meta-algorithm 93

5.12 Model classification accuracy over the Ham data set training ensembles with the AdaBoost.M2 93

List of Algorithms

1 Bagging 36
2 AdaBoost General Algorithm 39
3 PCA ensemble results combination 53
4 Map Fusion by Euclidean Distance 71
5 Map Fusion by Voronoi Polygon Similarity 73
6 Map Fusion by Superposition 75
7 Weighted Voting Summarization 99