Advances in Intelligent and Soft Computing 77

Editor-in-Chief: J. Kacprzyk

Advances in Intelligent and Soft Computing

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. 63. Á. Herrero, P. Gastaldo, R. Zunino, E. Corchado (Eds.)
Computational Intelligence in Security for Information Systems, 2009
ISBN 978-3-642-04090-0

Vol. 64. E. Tkacz, A. Kapczynski (Eds.)
Internet – Technical Development and Applications, 2009
ISBN 978-3-642-05018-3

Vol. 65. E. Kącki, M. Rudnicki, J. Stempczyńska (Eds.)
Computers in Medical Activity, 2009
ISBN 978-3-642-04461-8

Vol. 66. G.Q. Huang, K.L. Mak, P.G. Maropoulos (Eds.)
Proceedings of the 6th CIRP-Sponsored International Conference on Digital Enterprise Technology, 2009
ISBN 978-3-642-10429-9

Vol. 67. V. Snášel, P.S. Szczepaniak, A. Abraham, J. Kacprzyk (Eds.)
Advances in Intelligent Web Mastering - 2, 2010
ISBN 978-3-642-10686-6

Vol. 68. V.-N. Huynh, Y. Nakamori, J. Lawry, M. Inuiguchi (Eds.)
Integrated Uncertainty Management and Applications, 2010
ISBN 978-3-642-11959-0

Vol. 69. E. Piętka and J. Kawa (Eds.)
Information Technologies in Biomedicine, 2010
ISBN 978-3-642-13104-2

Vol. 70. Y. Demazeau, F. Dignum, J.M. Corchado, J. Bajo Pérez (Eds.)
Advances in Practical Applications of Agents and Multiagent Systems, 2010
ISBN 978-3-642-12383-2

Vol. 71. Y. Demazeau, F. Dignum, J.M. Corchado, J. Bajo, R. Corchuelo, E. Corchado, F. Fernández-Riverola, V.J. Julián, P. Pawlewski, A. Campbell (Eds.)
Trends in Practical Applications of Agents and Multiagent Systems, 2010
ISBN 978-3-642-12432-7

Vol. 72. J.C. Augusto, J.M. Corchado, P. Novais, C. Analide (Eds.)
Ambient Intelligence and Future Trends, 2010
ISBN 978-3-642-13267-4

Vol. 73. J.M. Corchado, P. Novais, C. Analide, J. Sedano (Eds.)
Soft Computing Models in Industrial and Environmental Applications, 5th International Workshop (SOCO 2010), 2010
ISBN 978-3-642-13160-8

Vol. 74. M.P. Rocha, F.F. Riverola, H. Shatkay, J.M. Corchado (Eds.)
Advances in Bioinformatics, 2010
ISBN 978-3-642-13213-1

Vol. 75. X.Z. Gao, A. Gaspar-Cunha, M. Köppen, G. Schaefer, and J. Wang (Eds.)
Soft Computing in Industrial Applications, 2010
ISBN 978-3-642-11281-2

Vol. 76. T. Bastiaens, U. Baumöl, and B.J. Krämer (Eds.)
On Collective Intelligence, 2010
ISBN 978-3-642-14480-6

Vol. 77. C. Borgelt, G. González-Rodríguez, W. Trutschnig, M.A. Lubiano, M.Á. Gil, P. Grzegorzewski, and O. Hryniewicz (Eds.)
Combining Soft Computing and Statistical Methods in Data Analysis, 2010
ISBN 978-3-642-14745-6

Christian Borgelt, Gil González-Rodríguez,
Wolfgang Trutschnig, María Asunción Lubiano,
María Ángeles Gil, Przemysław Grzegorzewski,
and Olgierd Hryniewicz (Eds.)

Combining Soft Computing and Statistical Methods in Data Analysis

Springer cajAstur

Editors

Christian Borgelt
Research Unit on Intelligent
Data Analysis and Graphical Models
European Centre for Soft Computing
Edificio Científico-Tecnológico. 3ª Planta
C/ Gonzalo Gutiérrez Quirós s/n
33600 Mieres, Spain
E-mail: christian.borgelt@softcomputing.es

Gil González-Rodríguez
Research Unit on Intelligent
Data Analysis and Graphical Models
European Centre for Soft Computing
Edificio Científico-Tecnológico. 3ª Planta
C/ Gonzalo Gutiérrez Quirós s/n
33600 Mieres, Spain
E-mail: gil.gonzalez@softcomputing.es

Wolfgang Trutschnig
Research Unit on Intelligent
Data Analysis and Graphical Models
European Centre for Soft Computing
Edificio Científico-Tecnológico. 3ª Planta
C/ Gonzalo Gutiérrez Quirós s/n
33600 Mieres, Spain
E-mail: wolfgang.trutschnig@softcomputing.es

María Asunción Lubiano
Departamento de Estadística e I.O. y D.M.
Universidad de Oviedo, Facultad de Ciencias
C/ Calvo Sotelo s/n, 33007 Oviedo, Spain
E-mail: lubiano@uniovi.es

María Ángeles Gil
Departamento de Estadística
e I.O. y D.M.
Universidad de Oviedo
Facultad de Ciencias
C/ Calvo Sotelo s/n
33007 Oviedo, Spain
E-mail: magil@uniovi.es

Dr. Przemysław Grzegorzewski
Systems Research Institute
Polish Academy of Sciences
Newelska 6, 01-447 Warsaw,
Poland

and

Faculty of Mathematics and
Information Science
Warsaw University of Technology
Plac Politechniki 1, 00-661 Warsaw
Poland
E-mail: pgrzeg@ibspan.waw.pl

Prof. Dr. Olgierd Hryniewicz
Systems Research Institute
Polish Academy of Science
Newelska 6, 01-447 Warsaw
Poland
E-mail: hryniewi@ibspan.waw.pl

ISBN 978-3-642-14745-6 e-ISBN 978-3-642-14746-3

DOI 10.1007/978-3-642-14746-3

Advances in Intelligent and Soft Computing ISSN 1867-5662

Library of Congress Control Number: 2010931605

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

Printed on acid-free paper

5 4 3 2 1 0

springer.com

Preface

> "The statistician cannot excuse himself from the duty of getting his head clear on the principles of scientific inference, but equally no other thinking man can avoid a like obligation."
>
> *Ronald A. Fisher*

Over the last forty years there has been a growing interest to extend probability theory and statistics and to allow for more flexible modelling of imprecision, uncertainty, vagueness and ignorance. The fact that in many real-life situations data uncertainty is not only present in the form of randomness (stochastic uncertainty) but also in the form of imprecision/fuzziness is but one point underlining the need for a widening of statistical tools. Most such extensions originate in a "softening" of classical methods, allowing, in particular, to work with imprecise or vague data, considering imprecise or generalized probabilities and fuzzy events, etc. The developed techniques frequently lead to more robust and interpretable models that better capture all the information contained in the given data.

About ten years ago the idea of establishing a recurrent forum for discussing new trends in the before-mentioned context was born and resulted in the first International Conference on Soft Methods in Probability and Statistics (SMPS) that was held in Warsaw in 2002. In the following years the conference took place in Oviedo (2004), in Bristol (2006) and in Toulouse (2008). In the current edition the conference returns to Oviedo. Apart from the rich number of topics already covered by the previous editions, the SMPS 2010 succeeded in incorporating statistics with censored data and robust statistics, both perfectly fitting the scope of the conference.

The wide variety of sessions taking place at the SMPS conference is reflected by the SMPS 2010' plenary talks: Peter Filzmoser from the Vienna University of Technology on "Soft Methods in Robust Statistics", Manfred Gilli from the University of Geneva on "An Introduction to Heuristic

Optimization Methods", Mario Guarracino from the High Performance Computing and Networking Institute in Naples on "Supervised Classification of Biological Data", and Enrique Ruspini from the European Centre for Soft Computing on "Ideas and Issues in Conceptual Fuzzy Clustering".

No conference can be organised without a lot of support from various people: We would like to thank all organizers of Invited Sessions, all members of the Program Committee and all additional reviewers - without their help a book like this would have been impossible. Moreover we would like to express our gratitude to "Obra social y cultural" of the main Savings Bank in Asturias, CajAstur, who kindly paid the production cost of these proceedings. The SMPS 2010 also benefited from the COST Action IC0702, this support is gratefully acknowledged. Last but not least we would like to express our gratitude to the European Centre for Soft Computing and the University of Oviedo.

Oviedo, June 2010

Christian Borgelt
Gil González-Rodríguez
Wolfgang Trutschnig
M. Asunción Lubiano
María Ángeles Gil
Przemysław Grzegorzewski
Olgierd Hryniewicz

Members of Committees

General Chairs

Christian Borgelt (Mieres, Spain)
Gil González Rodríguez (Mieres, Spain)
Wolfgang Trutschnig (Mieres, Spain)

Advisory Committee (Core SMPS Group)

María Ángeles Gil (Oviedo, Spain)
Przemysław Grzegorzewski (Warsaw, Poland)
Olgierd Hryniewicz (Warsaw, Poland)

Program Committee

Ricardo Cao (A Coruña, Spain)
Giulianella Coletti (Perugia, Italy)
Ana Colubi (Oviedo, Spain)
Renato Coppi (Rome, Italy)
Inés Couso (Gijón, Spain)
Bernard de Baets (Gent, Belgium)
Gert de Cooman (Gent, Belgium)
Thierry Denœux (Compiégne, France)
Didier Dubois (Toulouse, France)
Fabrizio Durante (Bolzano, Italy)
Juan Luis Fernández Martínez (Oviedo, Spain)
Peter Filzmoser (Vienna, Austria)
José Gámez (Albacete, Spain)
Pedro Gil (Oviedo, Spain)
Manfred Gilli (Geneve, Switzerland)
Lluis Godó (Barcelona, Spain)

Michel Grabisch (Paris, France)
Mario Guarracino (Naples, Italy)
Eyke Hüllermeier (Marburg, Germany)
Janusz Kacprzyk (Warsaw, Poland)
Etienne Kerre (Gent, Belgium)
Rudolf Kruse (Magdeburg, Germany)
Jonathan Lawry (Bristol, United Kingdom)
Shoumei Li (Beijing, China)
Uwe Ligges (Dortmund, Germany)
Miguel López (Oviedo, Spain)
Marloes Maathuis (Zurich, Switzerland)
Serafín Moral (Granada, Spain)
Domingo Morales (Elche, Spain)
Wolfgang Näther (Freiberg, Germany)
Mirko Navara (Praha, Czech Republic)
Hung T. Nguyen (Las Cruces, USA)
Dan Ralescu (Cincinnati, USA)
Marko Rojas-Medar (Campiñas, Brazil)
Enrique Ruspini (Mieres, Spain)
Antonio Salmerón (Almería, Spain)
Pedro Terán (Oviedo, Spain)
Stefan Van Aelst (Gent, Belgium)
Reinhard Viertl (Vienna, Austria)
Peter Winker (Giessen, Germany)
Marco Zaffalon (Lugano, Switzerland)

Additional Referees

Jorge Gabriel Adrover (Córdoba, Argentina)
José Alonso (Mieres, Spain)
Nicole Bäuerle (Karlsruhe, Germany)
Angela Blanco-Fernández (Oviedo, Spain)
Ulrich Bodenhofer (Linz, Austria)
Enea Bongiorno (Milano, Italy)
Ignacio Cascos (Madrid, Spain)
Etienne Côme (Paris, France)
Pierpaolo D'Urso (Rome, Italy)
Jacobo de Uña-Álvarez (Vigo, Spain)
Sébastien Destercke (Montpellier, France)
Giancarlo Diana (Padova, Italy)
Maria Brigida Ferraro (Rome, Italy)
Daan Fierens (Heverlee, Belgium)
Luis Angel García Escudero (Valladolid, Spain)
Ali Gholami (Tehran, Iran)

Paolo Giordani (Rome, Italy)
Inés González Rodríguez (Santander, Spain)
Sergio Guadarrama (Mieres, Spain)
Marc Hofmann (Neuchâtel, Switzerland)
María Amalia Jácome (A Coruña, Spain)
Osmo Kaleva (Tampere, Finland)
M. Asunción Lubiano (Oviedo, Spain)
Marek Malinowski (Zielona Góra, Poland)
Mylène Masson (Compiègne, France)
Enrique Miranda (Oviedo, Spain)
Isabel Molina (Madrid, Spain)
Susana Montes (Gijón, Spain)
Erik Quaeghebeur (Gent, Belgium)
José Juan Quesada Molina (Granada, Spain)
Ana Belén Ramos-Guajardo (Mieres, Spain)
Luis José Rodríguez Muñíz (Oviedo, Spain)
Luciano Sánchez (Gijón, Spain)
José María Sarabia (Santander, Spain)
Enrico Schumann (Geneve, Switzerland)
Fabio L. Spizzichino (Rome, Italy)
Matthias Troffaes (Durham, United Kingdom)
Paolo Vicig (Trieste, Italy)

Local Organization

Andrea Appe (Vienna, Austria)
Angela Blanco-Fernández (Oviedo, Spain)
María Rosa Casals (Oviedo, Spain)
Ana Colubi (Oviedo, Spain)
Maria Brigida Ferraro (Rome, Italy)
Marta García-Bárzana (Oviedo, Spain)
María Teresa López (Oviedo, Spain)
M. Asunción Lubiano (Oviedo, Spain)
Takehiko Nakama (Oviedo, Spain)
Antonio Palacio (Oviedo, Spain)
Ana Belén Ramos-Guajardo (Mieres, Spain)
Marc Segond (Mieres, Spain)
Beatriz Sinova (Oviedo, Spain)
Nadine Zwickl (Oviedo, Spain)

Technical Edition

M. Asunción Lubiano (Oviedo, Spain)

Contents

Prior Knowledge in the Classification of Biomedical Data

Danilo Abbate, Roberta De Asmundis, and Mario Rosario Guarracino

Abstract. Standard data analysis techniques for biomedical problems cannot take into account existing prior knowledge, and available literature results cannot be incorporated in further studies. In this work we review some techniques that incorporate prior knowledge in supervised classification algorithms as constraints to the underlying optimization and linear algebra problems. We analyze a case study, to show the advantage of such techniques in terms of prediction accuracy.

Keywords: Supervised classification, Neural Networks, Support Vector Machines, Generalized Eigenvalue Classifier.

1 Introduction

The widespread availability of biomedical data is posing new and challenging problems to standard analysis algorithms. These problems are related to the quality of data, that are often affected by errors and uncertainty. This is the case of high throughput genomic and proteomic technologies, where the signal to noise ratio is very low. Other questions raise when data produced by comparable experimental protocols are available, because there is no clear strategy to systematically take advantage of previous results and knowledge. In the case of supervised classification, where models are built from data for which the class membership is known, available labeled data is added to the training sets. This has two major drawbacks. First, enlarging the training set increases the computational time needed to elaborate the model. Then, if data are affected by errors or uncertainties, these are introduced in the new classification model, reducing its generalization capabilities.

Danilo Abbate, Roberta De Asmundis, and Mario Rosario Guarracino
High Performance Computing and Networking Institute,
National Research Council (ICAR-CNR), 80131 Naples, Italy
e-mail: mario.guarracino@cnr.it

C. Borgelt et al. (Eds.): Combining Soft Computing & Stats. Methods, AISC 77, pp. 1–8.
springerlink.com

In this paper we show how to introduce prior knowledge in Support Vector Machines (SVM) [12], Generalized Eigenvalue Proximal SVM (GPSVM) [8], and Radial Basis Functions (RBF) Neural Networks [1]. The idea is if knowledge can be expressed in terms of regions of the data space, in which all points belong to a given class, then the geometrical expression of such regions can be used to constrain the underlying mathematical programming problem. The advantage of such strategy is that, although no points are added to the training set, the model is constrained to take into account available knowledge. We provide a case study that highlights the advantages of such strategy, in terms of classification accuracy.

2 Classification Algorithms

Support Vector Machines

SVM are the state of the art supervised classification methods, widely accepted in many application areas. SVM find a plane $\mathbf{w}^T\mathbf{x} + b = 0$ with the objective to separate the elements belonging to two different classes. To this extend, we determine two parallel planes $\mathbf{w}^T\mathbf{x} + b = \pm 1$, of maximum distance, leaving all points of the two classes on different sides. Elements with the minimum distance from both classes are called *support vectors* and are the only elements needed to train the classifier.

Let us consider a data set composed of n pairs $(\mathbf{x}_i, y_i)$ where $\mathbf{x}_i \in \mathbf{R}^m$ is the feature vector of a point, and $y_i \in \{-1, 1\}$ is the class label. The optimal separating plane is the solution to a quadratic linearly constrained problem.

The advantage of this method is that a very small number of support vectors are sufficient to define the optimal separating plane. In some cases, the relationship between points and class labels can be nonlinear and it is impossible to find a separating plane. In such a case, data can be nonlinearly embedded to a higher dimensional space in which the linear separation can be found. This nonlinear mapping can be implicitly done by kernel functions, which represent the inner product of the elements in the nonlinear space.

The nonlinear classification model cannot describe the discriminating function in terms of inequalities involving linear relations among features. This can be perceived as a problem in case of medical diagnosis, in which doctors prefer to find simple correlations between the results of a clinic exams and the diagnosis or prognosis of an illness. On the other hand, it is generally accepted that results achieved by nonlinear models provide higher classification accuracy. Furthermore, the number of exams to consider for a diagnosis can be very high and cannot be correlated only with the experience. Finally, methods that provide explicit classification rules are not guaranteed to find a set of rules small enough to be easy readable.

Generalized Eigenvalue Classifier

GEPSVM is an efficient algorithm in which the binary classification problem can be formulated as a generalized eigenvalue problem.

Let us consider two matrices $A \in \mathbf{R}^{n\times m}$ and $B \in \mathbf{R}^{k\times m}$, with $m \ll n+k$, representing the two classes, each row being a point in the feature space. Mangasarian et al. [8] propose to classify these sets of points A and B using two planes in the feature space, each closest to one set of points, and furthest from the other.

Suppose that points in classes A and B are not linearly separable, then a nonlinear embedding of each point $\mathbf{x}$ can be obtained using a Radial Basis Function kernel. Each component of the transformed point is given by $K(\mathbf{x},C_i) = exp(\|\mathbf{x}-C_i\|^2/\sigma)$, where C_i is the i-th row of $C = \left[A^T, B^T\right]^T \in \mathbf{R}^{(n+k)\times m}$, and σ is a parameter.

The two planes $K(\mathbf{x},C)\mathbf{u}_1 - \gamma_1 = 0$ and $K(\mathbf{x},C)\mathbf{u}_2 - \gamma_2 = 0$ in the feature space, can be obtained solving the generalized eigenvalue problem [6]:

$$\min_{\mathbf{u},\gamma\neq 0} \frac{\|K(A,C)\mathbf{u} - \mathbf{e}\gamma\|^2 + \delta\|\tilde{K}_B\mathbf{u} - \mathbf{e}\gamma\|^2}{\|K(B,C)\mathbf{u} - \mathbf{e}\gamma\|^2 + \delta\|\tilde{K}_A\mathbf{u} - \mathbf{e}\gamma\|^2}. \tag{1}$$

Here $\tilde{K}_A$ and $\tilde{K}_B$ are diagonal matrices with the diagonal entries from the matrices $K(A,C)$ and $K(B,C)$; $\mathbf{e}$ is a vector of 1s of proper dimension, $\mathbf{u}$ is the coefficient vector of the plane, γ is the plane intercept and δ is the regularization parameter. The eigenvectors related to the minimum and the maximum eigenvalues of (1), provide the coefficients of the proximal planes P_i, $i = 1,2$. The class of a new point $\mathbf{x}$ is determined as

$$class(\mathbf{x}) = argmin_{i=-1,1}\{dist(\mathbf{x},P_i)\}, \tag{2}$$

where $dist(\mathbf{x},P_i)$ is the distance of a point $\mathbf{x}$ from plane P_i.

RBF Neural Networks

A RBF neural network is divided into two operative blocks: an inner hidden layer, and the output layer. The hidden layer creates a response localized on the input vector $\mathbf{x}$; the binary output will then be calculated as a weighted sum of these localized responses. Training a RBF network is a procedure divided into two phases: in the first one the parameters of the radial bases function are calculated using an unsupervised learning algorithm. In this phase the data set is divided in $\bar{n}+\bar{k}$ clusters. We define as $\bar{\mathbf{x}}$ the $\bar{n}+\bar{k}$ points closest to each centroid. In the second part of the training, we search for values of the weights w_i which determine the binary output:

$$h(\mathbf{x}) = \sum_{i=1}^{\bar{n}+\bar{k}} w_i K(\mathbf{x},\bar{\mathbf{x}}_i), \quad \bar{n} \ll n, \quad \bar{k} \ll k. \tag{3}$$

Such weights are calculated by minimizing the following error function, with respect to w_i:

$$E = \frac{1}{2}\sum_{i=1}^{n+k}(h(\mathbf{x}_i) - y_i)^2 \tag{4}$$

where y_i is the label of the point $\mathbf{x}_i$.

3 Prior Knowledge

SVM

We are now showing how it is possible to obtain, with a linear program [9], a nonlinear separating surface using a kernel function $K(\mathbf{x},C) : \mathbf{R}^m \times \mathbf{R}^{(n+k)\times m} \rightarrow \mathbf{R}^{n+k}$, to embed the points in a higher dimensional space. We recall that the resulting plane, projected in the feature space [11], has equation:

$$K(\mathbf{x},C)\mathbf{u} - \gamma = 0. \tag{5}$$

In standard SVM, parameters $\mathbf{u} \in \mathbf{R}^{n+k}$ and $\gamma \in \mathbf{R}$ are determined solving the following quadratic optimization problem [7], for some $\nu > 0$:

$$\begin{array}{ll} \min\limits_{\mathbf{u},\gamma,\mathbf{y}\in\mathbf{R}^{(n+k)+1+(n+k)}} & \nu\mathbf{e}^T\mathbf{y} + \frac{1}{2}\mathbf{u}^T\mathbf{u} \\ s.t. & D(K(C,C)\mathbf{u} - \mathbf{e}\gamma) + \mathbf{y} \geq \mathbf{e}, \quad \mathbf{y} \geq 0. \end{array} \tag{6}$$

where D is a diagonal matrix, with the diagonal elements equal to the labels of the corresponding element of the training set C, $\mathbf{y}$ is a vector of slack variables. Such condition places the points belonging to the two classes $+1$ and -1 on two different sides of the nonlinear separation surface (5). Problem (6) corresponds to the following linear programming problem [9]:

$$\begin{array}{lr} \underset{\mathbf{u},\gamma,\mathbf{y},\mathbf{s}}{min} & \nu\mathbf{e}^T\mathbf{y} + \mathbf{e}^T\mathbf{s} \\ s.t. & (K(C,C)\mathbf{u} - \mathbf{e}\gamma) + \mathbf{y} \geq \mathbf{e}, \\ & -\mathbf{s} \leq \mathbf{u} \leq \mathbf{s}, \\ & \mathbf{y} \geq 0, \end{array} \tag{7}$$

where $\mathbf{s} \in \mathbf{R}^{n+k}$ is a vector of non negative slack variables.

In order to improve the results obtained by a classifier solely from the training set, it is possible to impose the knowledge of an expert into the learning phase of the function (5) [10]. Such expertise is represented by the following implication, which represents a knowledge region $\Delta \subset \mathbf{R}^m$ in the input space in which all points $\mathbf{x}$ are known to belong to class $+1$:

$$g(\mathbf{x}) \leq 0 \Rightarrow K(\mathbf{x},C)\mathbf{u} - \gamma \geq \alpha, \forall \mathbf{x} \in \Delta, \alpha \in \mathbf{R}^+, \tag{8}$$

where $g(\mathbf{x}) : \Delta \subset \mathbf{R}^m \rightarrow \mathbf{R}$.