Nikunj C. Oza
Robi Polikar
Josef Kittler
Fabio Roli (Eds.)

LNCS 3541

Multiple Classifier Systems

6th International Workshop, MCS 2005
Seaside, CA, USA, June 2005
Proceedings

MCS
2005

Springer

Nikunj C. Oza Robi Polikar
Josef Kittler Fabio Roli (Eds.)

Multiple Classifier Systems

6th International Workshop, MCS 2005
Seaside, CA, USA, June 13-15, 2005
Proceedings

Springer

Volume Editors

Nikunj C. Oza
NASA Ames Research Center
Mail Stop 269-3, Moffett Field, CA 94035-1000, USA
E-mail: oza@email.arc.nasa.gov

Robi Polikar
Rowan University, Electrical and Computer Engineering
136 Rowan Hall, 201 Mullica Hill Road, Glassboro, NJ 08028, USA
E-mail: polikar@rowan.edu

Josef Kittler
University of Surrey, Center for Vision, Speech and Signal Processing
Guildford, Surrey GU2 7XH, UK
E-mail: j.kittler@eim.surrey.ac.uk

Fabio Roli
University of Cagliari, Department of Electrical and Electronic Engineering
Piazza D'Armi, 09123 Cagliari, Italy
E-mail: roli@diee.unica.it

Library of Congress Control Number: 2005927319

CR Subject Classification (1998): I.5, I.4, I.2.10, I.2, F.1

ISSN 0302-9743
ISBN-10 3-540-26306-3 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-26306-7 Springer Berlin Heidelberg New York

Springer is a part of Springer Science+Business Media

springeronline.com

Printed in Germany

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 11494683 06/3142 5 4 3 2 1 0

Lecture Notes in Computer Science 3541

Preface

The belief that a committee of people make better decisions than any individual is widely held and appreciated. We also understand that, for this to be true, the members of the committee have to be simultaneously competent and complementary. This intuitive notion holds true for committees of data sources (such as sensors) and models (such as classifiers). The substantial current research in the areas of data fusion and model fusion focuses on ensuring that the different sources provide useful information but nevertheless complement one another to yield better results than any source would on its own. During the 1990s, a variety of schemes in classifier fusion, which is the focus of this workshop, were developed under many names in different scientific communities such as machine learning, pattern recognition, neural networks, and statistics. The previous five workshops on Multiple Classifier Systems (MCS) were themselves exercises in information fusion, with the goal of bringing the different scientific communities together, providing each other with different perspectives on this fascinating topic, and aiding cross-fertilization of ideas. These five workshops achieved this goal, demonstrating significant advances in the theory, algorithms, and applications of multiple classifier systems.

Following its five predecessors published by Springer, this volume contains the proceedings of the 6th International Workshop on Multiple Classifier Systems (MCS 2005) held at the Embassy Suites in Seaside, California, USA, June 13–15, 2005. Forty-two papers were selected by the Scientific Committee, and they were organized into the following sessions: Boosting, Combination Methods, Design of Ensembles, Performance Analysis, and Applications. The workshop program was enriched by an invited talk given by Leo Breiman (University of California, Berkeley, USA).

The workshop was organized by the NASA Ames Research Center (USA), Rowan University (USA), and PureSense Environmental (USA). It was sponsored by the International Association for Pattern Recognition through its Technical Committee TC1: Statistical Techniques in Pattern Recognition. We also wish to express our gratitude to all who helped to organize MCS 2005. We thank the authors of all the submissions for their hard work and efforts toward making this workshop a true exercise in information fusion. We also thank the members of the Scientific Committee and many other reviewers for performing the difficult task of selecting the best papers from a large number of high-quality submissions. Special thanks to Terry Windeatt (University of Surrey, UK), Darren Galaviz (NASA), John Williamson (PureSense), Christopher Peri (PureSense), and Giorgio Fumera (University of Cagliari, Italy) for their substantial contributions to local organization and website management.

June 2005 Nikunj C. Oza, Robi Polikar, Josef Kittler, and Fabio Roli

Organization

Workshop Chairs

N.C. Oza (NASA Ames Research Center, USA)
R. Polikar (Rowan University, USA)
J. Kittler (University of Surrey, UK)
F. Roli (University of Cagliari, Italy)

Scientific Committee

J.A. Benediktsson (Iceland)
H. Bunke (Switzerland)
L.P. Cordella (Italy)
B.V. Dasarathy (USA)
R.P.W. Duin (The Netherlands)
C. Furlanello (Italy)
J. Ghosh (USA)
V. Govindaraju (USA)
T.K. Ho (USA)
S. Impedovo (Italy)
N. Intrator (Israel)
A.K. Jain (USA)
M. Kamel (Canada)
L.I. Kuncheva (UK)
D. Partridge (UK)
A.J.C. Sharkey (UK)
C.Y. Suen (Canada)
K. Tumer (USA)
G. Vernazza (Italy)

Organizing Committee

T. Windeatt (University of Surrey, UK)
G. Fumera (University of Cagliari, Italy)
D. Galaviz (NASA Ames Research Center, USA)
J. Williamson (Puresense Environmental, USA)
C. Peri (Puresense Environmental, USA)

Sponsoring Institutions

Intelligent Systems Division, NASA Ames Research Center
Department of Electrical and Computer Engineering, Rowan University
Puresense Environmental
International Association for Pattern Recognition
IAPR Technical Committee TC1

Lecture Notes in Computer Science

For information about Vols. 1–3442

please contact your bookseller or Springer

Vol. 3556: H. Baumeister, M. Marchesi, M. Holcombe (Eds.), Extreme Programming and Agile Processes in Software Engineering. XIV, 332 pages. 2005.

Vol. 3555: T. Vardanega, A. Wellings (Eds.), Reliable Software Technology – Ada-Europe 2005. XV, 273 pages. 2005.

Vol. 3552: H. de Meer, N. Bhatti (Eds.), Quality of Service – IWQoS 2005. XV, 273 pages. 2005.

Vol. 3543: L. Kutvonen, N. Alonistioti (Eds.), Distributed Applications and Interoperable Systems. XI, 235 pages. 2005.

Vol. 3541: N.C. Oza, R. Polikar, J. Kittler, F. Roli (Eds.), Multiple Classifier Systems. XII, 430 pages. 2005.

Vol. 3537: A. Apostolico, M. Crochemore, K. Park (Eds.), Combinatorial Pattern Matching. XI, 444 pages. 2005.

Vol. 3535: M. Steffen, G. Zavattaro (Eds.), Formal Methods for Open Object-Based Distributed Systems. X, 323 pages. 2005.

Vol. 3532: A. Gómez-Pérez, J. Euzenat (Eds.), The Semantic Web: Research and Applications. XV, 728 pages. 2005.

Vol. 3531: J. Ioannidis, A. Keromytis, M. Yung (Eds.), Applied Cryptography and Network Security. XI, 530 pages. 2005.

Vol. 3528: P.S. Szczepaniak, J. Kacprzyk, A. Niewiadomski (Eds.), Advances in Web Intelligence. XVII, 513 pages. 2005. (Subseries LNAI).

Vol. 3527: R. Morrison, F. Oquendo (Eds.), Software Architecture. XII, 263 pages. 2005.

Vol. 3526: S.B. Cooper, B. Löwe, L. Torenvliet (Eds.), New Computational Paradigms. XVII, 574 pages. 2005.

Vol. 3525: A.E. Abdallah, C.B. Jones, J.W. Sanders (Eds.), Communicating Sequential Processes. XIV, 321 pages. 2005.

Vol. 3524: R. Barták, M. Milano (Eds.), Integration of AI and OR Techniques in Constraint Programming for Combinatorial Optimization Problems. XI, 320 pages. 2005.

Vol. 3523: J.S. Marques, N.P. de la Blanca, P. Pina (Eds.), Pattern Recognition and Image Analysis, Part II. XXVI, 733 pages. 2005.

Vol. 3522: J.S. Marques, N.P. de la Blanca, P. Pina (Eds.), Pattern Recognition and Image Analysis, Part I. XXVI, 703 pages. 2005.

Vol. 3521: N. Megiddo, Y. Xu, B. Zhu (Eds.), Algorithmic Applications in Management. XIII, 484 pages. 2005.

Vol. 3520: O. Pastor, J. Falcão e Cunha (Eds.), Advanced Information Systems Engineering. XVI, 584 pages. 2005.

Vol. 3519: H. Li, P. J. Olver, G. Sommer (Eds.), Computer Algebra and Geometric Algebra with Applications. IX, 449 pages. 2005.

Vol. 3518: T.B. Ho, D. Cheung, H. Li (Eds.), Advances in Knowledge Discovery and Data Mining. XXI, 864 pages. 2005. (Subseries LNAI).

Vol. 3517: H.S. Baird, D.P. Lopresti (Eds.), Human Interactive Proofs. IX, 143 pages. 2005.

Vol. 3516: V.S. Sunderam, G.D.v. Albada, P.M.A. Sloot, J.J. Dongarra (Eds.), Computational Science – ICCS 2005, Part III. LXIII, 1143 pages. 2005.

Vol. 3515: V.S. Sunderam, G.D.v. Albada, P.M.A. Sloot, J.J. Dongarra (Eds.), Computational Science – ICCS 2005, Part II. LXIII, 1101 pages. 2005.

Vol. 3514: V.S. Sunderam, G.D.v. Albada, P.M.A. Sloot, J.J. Dongarra (Eds.), Computational Science – ICCS 2005, Part I. LXIII, 1089 pages. 2005.

Vol. 3513: A. Montoyo, R. Muñoz, E. Métais (Eds.), Natural Language Processing and Information Systems. XII, 408 pages. 2005.

Vol. 3512: J. Cabestany, A. Prieto, F. Sandoval (Eds.), Computational Intelligence and Bioinspired Systems. XXV, 1260 pages. 2005.

Vol. 3510: T. Braun, G. Carle, Y. Koucheryavy, V. Tsaoussidis (Eds.), Wired/Wireless Internet Communications. XIV, 366 pages. 2005.

Vol. 3509: M. Jünger, V. Kaibel (Eds.), Integer Programming and Combinatorial Optimization. XI, 484 pages. 2005.

Vol. 3508: P. Bresciani, P. Giorgini, B. Henderson-Sellers, G. Low, M. Winikoff (Eds.), Agent-Oriented Information Systems II. X, 227 pages. 2005. (Subseries LNAI).

Vol. 3507: F. Crestani, I. Ruthven (Eds.), Information Context: Nature, Impact, and Role. XIII, 253 pages. 2005.

Vol. 3506: C. Park, S. Chee (Eds.), Information Security and Cryptology – ICISC 2004. XIV, 490 pages. 2005.

Vol. 3505: V. Gorodetsky, J. Liu, V. A. Skormin (Eds.), Autonomous Intelligent Systems: Agents and Data Mining. XIII, 303 pages. 2005. (Subseries LNAI).

Vol. 3504: A.F. Frangi, P.I. Radeva, A. Santos, M. Hernandez (Eds.), Functional Imaging and Modeling of the Heart. XV, 489 pages. 2005.

Vol. 3503: S.E. Nikoletseas (Ed.), Experimental and Efficient Algorithms. XV, 624 pages. 2005.

Vol. 3502: F. Khendek, R. Dssouli (Eds.), Testing of Communicating Systems. X, 381 pages. 2005.

Vol. 3501: B. Kégl, G. Lapalme (Eds.), Advances in Artificial Intelligence. XV, 458 pages. 2005. (Subseries LNAI).

Vol. 3500: S. Miyano, J. Mesirov, S. Kasif, S. Istrail, P. Pevzner, M. Waterman (Eds.), Research in Computational Molecular Biology. XVII, 632 pages. 2005. (Subseries LNBI).

Vol. 3499: A. Pelc, M. Raynal (Eds.), Structural Information and Communication Complexity. X, 323 pages. 2005.

Vol. 3498: J. Wang, X. Liao, Z. Yi (Eds.), Advances in Neural Networks – ISNN 2005, Part III. L, 1077 pages. 2005.

Vol. 3497: J. Wang, X. Liao, Z. Yi (Eds.), Advances in Neural Networks – ISNN 2005, Part II. L, 947 pages. 2005.

Vol. 3496: J. Wang, X. Liao, Z. Yi (Eds.), Advances in Neural Networks – ISNN 2005, Part II. L, 1055 pages. 2005.

Vol. 3495: P. Kantor, G. Muresan, F. Roberts, D.D. Zeng, F.-Y. Wang, H. Chen, R.C. Merkle (Eds.), Intelligence and Security Informatics. XVIII, 674 pages. 2005.

Vol. 3494: R. Cramer (Ed.), Advances in Cryptology – EUROCRYPT 2005. XIV, 576 pages. 2005.

Vol. 3493: N. Fuhr, M. Lalmas, S. Malik, Z. Szlávik (Eds.), Advances in XML Information Retrieval. XI, 438 pages. 2005.

Vol. 3492: P. Blache, E. Stabler, J. Busquets, R. Moot (Eds.), Logical Aspects of Computational Linguistics. X, 363 pages. 2005. (Subseries LNAI).

Vol. 3489: G.T. Heineman, I. Crnkovic, H.W. Schmidt, J.A. Stafford, C. Szyperski, K. Wallnau (Eds.), Component-Based Software Engineering. XI, 358 pages. 2005.

Vol. 3488: M.-S. Hacid, N.V. Murray, Z.W. Raś, S. Tsumoto (Eds.), Foundations of Intelligent Systems. XIII, 700 pages. 2005. (Subseries LNAI).

Vol. 3486: T. Helleseth, D. Sarwate, H.-Y. Song, K. Yang (Eds.), Sequences and Their Applications - SETA 2004. XII, 451 pages. 2005.

Vol. 3483: O. Gervasi, M.L. Gavrilova, V. Kumar, A. Laganà, H.P. Lee, Y. Mun, D. Taniar, C.J.K. Tan (Eds.), Computational Science and Its Applications – ICCSA 2005, Part IV. XXVII, 1362 pages. 2005.

Vol. 3482: O. Gervasi, M.L. Gavrilova, V. Kumar, A. Laganà, H.P. Lee, Y. Mun, D. Taniar, C.J.K. Tan (Eds.), Computational Science and Its Applications – ICCSA 2005, Part III. LXVI, 1340 pages. 2005.

Vol. 3481: O. Gervasi, M.L. Gavrilova, V. Kumar, A. Laganà, H.P. Lee, Y. Mun, D. Taniar, C.J.K. Tan (Eds.), Computational Science and Its Applications – ICCSA 2005, Part II. LXIV, 1316 pages. 2005.

Vol. 3480: O. Gervasi, M.L. Gavrilova, V. Kumar, A. Laganà, H.P. Lee, Y. Mun, D. Taniar, C.J.K. Tan (Eds.), Computational Science and Its Applications – ICCSA 2005, Part I. LXV, 1234 pages. 2005.

Vol. 3479: T. Strang, C. Linnhoff-Popien (Eds.), Location- and Context-Awareness. XII, 378 pages. 2005.

Vol. 3478: C. Jermann, A. Neumaier, D. Sam (Eds.), Global Optimization and Constraint Satisfaction. XIII, 193 pages. 2005.

Vol. 3477: P. Herrmann, V. Issarny, S. Shiu (Eds.), Trust Management. XII, 426 pages. 2005.

Vol. 3476: J. Leite, A. Omicini, P. Torroni, P. Yolum (Eds.), Declarative Agent Languages and Technologies. XII, 289 pages. 2005.

Vol. 3475: N. Guelfi (Ed.), Rapid Integration of Software Engineering Techniques. X, 145 pages. 2005.

Vol. 3474: C. Grelck, F. Huch, G.J. Michaelson, P. Trinder (Eds.), Implementation and Application of Functional Languages. X, 227 pages. 2005.

Vol. 3468: H.W. Gellersen, R. Want, A. Schmidt (Eds.), Pervasive Computing. XIII, 347 pages. 2005.

Vol. 3467: J. Giesl (Ed.), Term Rewriting and Applications. XIII, 517 pages. 2005.

Vol. 3465: M. Bernardo, A. Bogliolo (Eds.), Formal Methods for Mobile Computing. VII, 271 pages. 2005.

Vol. 3464: S.A. Brueckner, G.D.M. Serugendo, A. Karageorgos, R. Nagpal (Eds.), Engineering Self-Organising Systems. XIII, 299 pages. 2005. (Subseries LNAI).

Vol. 3463: M. Dal Cin, M. Kaâniche, A. Pataricza (Eds.), Dependable Computing - EDCC 2005. XVI, 472 pages. 2005.

Vol. 3462: R. Boutaba, K.C. Almeroth, R. Puigjaner, S. Shen, J.P. Black (Eds.), NETWORKING 2005. XXX, 1483 pages. 2005.

Vol. 3461: P. Urzyczyn (Ed.), Typed Lambda Calculi and Applications. XI, 433 pages. 2005.

Vol. 3460: Ö. Babaoglu, M. Jelasity, A. Montresor, C. Fetzer, S. Leonardi, A. van Moorsel, M. van Steen (Eds.), Self-star Properties in Complex Information Systems. IX, 447 pages. 2005.

Vol. 3459: R. Kimmel, N.A. Sochen, J. Weickert (Eds.), Scale Space and PDE Methods in Computer Vision. XI, 634 pages. 2005.

Vol. 3458: P. Herrero, M.S. Pérez, V. Robles (Eds.), Scientific Applications of Grid Computing. X, 208 pages. 2005.

Vol. 3456: H. Rust, Operational Semantics for Timed Systems. XII, 223 pages. 2005.

Vol. 3455: H. Treharne, S. King, M. Henson, S. Schneider (Eds.), ZB 2005: Formal Specification and Development in Z and B. XV, 493 pages. 2005.

Vol. 3454: J.-M. Jacquet, G.P. Picco (Eds.), Coordination Models and Languages. X, 299 pages. 2005.

Vol. 3453: L. Zhou, B.C. Ooi, X. Meng (Eds.), Database Systems for Advanced Applications. XXVII, 929 pages. 2005.

Vol. 3452: F. Baader, A. Voronkov (Eds.), Logic for Programming, Artificial Intelligence, and Reasoning. XI, 562 pages. 2005. (Subseries LNAI).

Vol. 3450: D. Hutter, M. Ullmann (Eds.), Security in Pervasive Computing. XI, 239 pages. 2005.

Vol. 3449: F. Rothlauf, J. Branke, S. Cagnoni, D.W. Corne, R. Drechsler, Y. Jin, P. Machado, E. Marchiori, J. Romero, G.D. Smith, G. Squillero (Eds.), Applications of Evolutionary Computing. XX, 631 pages. 2005.

Vol. 3448: G.R. Raidl, J. Gottlieb (Eds.), Evolutionary Computation in Combinatorial Optimization. XI, 271 pages. 2005.

Vol. 3447: M. Keijzer, A. Tettamanzi, P. Collet, J.v. Hemert, M. Tomassini (Eds.), Genetic Programming. XIII, 382 pages. 2005.

Vol. 3444: M. Sagiv (Ed.), Programming Languages and Systems. XIII, 439 pages. 2005.

Vol. 3443: R. Bodik (Ed.), Compiler Construction. XI, 305 pages. 2005.

¥528.64元

Table of Contents

Design Methods

Semi-supervised Multiple Classifier Systems: Background and Research Directions

Fabio Roli

Dept. of Electrical and Electronic Engineering, University of Cagliari,
Piazza d'Armi, 09123 Cagliari, Italy
roli@diee.unica.it

Abstract. Multiple classifier systems have been originally proposed for supervised classification tasks. In the five editions of MCS workshop, most of the papers have dealt with design methods and applications of supervised multiple classifier systems. Recently, the use of multiple classifier systems has been extended to unsupervised classification tasks. Despite its practical relevance, semi-supervised classification has not received much attention. Few works on semi-supervised multiple classifiers appeared in the machine learning literature. This paper's goal is to review the background results that can be exploited to promote research on semi-supervised multiple classifier systems, and to outline some future research directions.

1 Introduction

During the 1990's, several classifier fusion schemes, especially the ones that operate at the so-called decision-level, have emerged under a plethora of names within various scientific communities, including information fusion, machine learning, neural networks, pattern recognition, and statistics. The initial works on multiple classifier systems (MCS) dealt almost exclusively with supervised classification, and this trend continued over the years. In the five editions of MCS workshop [1], most papers have dealt with design methods and applications of supervised multiple classifier systems. Only recently, the multiple classifiers approach has been extended to unsupervised classification, and some methods have been proposed to combine multiple clustering algorithms [2,3]. Although past MCS research was focused on supervised classification, many pattern recognition applications cannot be addressed effectively under the pure supervised approach. In fact, there are applications that are characterized by two contrasting factors: the need for large quantities of labelled data to design classifiers with high accuracy, and the difficulty of collecting such data. For example, in text classification, the common end-user is not available to manually label a large amount of data necessary to achieve reasonable classification accuracy [4]. In remote-sensing image classification, Jackson and Landgrebe pointed out that the large number of spectral bands of modern sensors and the large number of land-cover classes of interest, require a number of training examples that are too expensive or tedious to acquire [5]. Similar scenarios occur in face recognition, medical

N.C. Oza et al. (Eds.): MCS 2005, LNCS 3541, pp. 1–11, 2005.

imaging, and intrusion detection in computer networks [6,7,8]. On the other hand, in these applications, collecting unlabelled data is often easy and inexpensive. In many text classification applications (e.g., web page categorization), unlabelled examples are freely and abundantly available. In remote sensing, thanks to the high spatial resolution of modern sensors, a large number of unlabelled examples become available when new images are captured. It is easy to see that such applications demand classification methods that achieve high accuracy using only a few labelled but many unlabelled examples. Semi-supervised classification deals with the design of classifiers using both labelled (possibly few) and unlabelled training examples. Various approaches to semi-supervised classification have been proposed [9].

Despite the theoretical and practical relevance of semi-supervised classification, the proposed approaches so far dealt with only single classifiers, and, in particular, no work was clearly devoted to this topic within the MCS literature. To the best of our knowledge, few works on semi-supervised multiple classifiers have appeared in the machine learning literature [7,10,11,12,13].

This paper is aimed at reviewing the background results that can be exploited to promote research on semi-supervised multiple classifier systems, and to outline some future research directions. Section 2 provides the basic concepts and briefly reviews the main approaches to semi-supervised classification. Section 3 reviews the few works which dealt with semi-supervised multiple classifiers, and attempts to give a systematic definition of semi-supervised MCS. Some directions for future research are outlined in Section 4.

2 Semi-supervised Classification

Given a set D_l (usually, small) of labelled data, and a set D_u (usually, large) of unlabelled data, semi-supervised classification methods aim to design classifiers using both sets. In this section, we first review briefly the main methods proposed for semi-supervised classification, with a particular attention paid to the co-training method due to its strong connection to MCS. Our review is biased due to its restricted focus on MCS, and consequently is not exhaustive. We refer the reader to [9] for a broader overview on semi-supervised classification methods. Finally, we briefly review the main arguments on the usefulness of unlabelled data in increasing the classification accuracy.

2.1 Methods for Semi-supervised Classification

Decision-Directed Methods

It is easy to see that a most straightforward approach to semi-supervised classification should be based on exploiting the available labeled examples in automatically assigning class labels to unlabeled data. The basic approach works as follows. An initial classifier is designed using the labeled data set D_l. This classifier is then used to assign class labels to examples in D_u. Then the classifier is re-trained using $D_l \cup D_u$. As the convergence of this simple algorithm can not be guaranteed in general, the last two steps are usually repeated for a given number of times or until some heuristic convergence criterion is satisfied. It is worth noting that, while traditional pattern recognition systems are "open-loop" systems, this approach corresponds to the implementation of simple "closed-loop" systems [15,16]. Although encouraging results have been reported in real applications [14], it is easy to see that the performance of this approach strongly depends on the accuracy of the classifiers used to assign the "pseudo-labels". The classification accuracy can improve over the iterations only if the initial and subsequent classifiers correctly label most of the data. Unfortunately, the issue of labelling accuracy needed to guarantee low generalization error is nearly impossible to address [15]. In practical applications, unreliable class labels are disregarded using measures of classification confidence in order to limit the number of labelling errors.

This simple approach to semi-supervised classification and its related methods, have been referred under various names by different communities: self-learning methods [9,15], self-corrective recognition [15,16], adaptive methods, naïve labelling approach [17], and decision-directed methods [18]. We choose the last name to refer to this paragraph since it reflects well the fundamental mechanism of this approach and it is used in statistical parameter estimation and signal processing areas.

Expectation-Maximization Methods

Expectation-maximization (EM) is a well known class of iterative algorithms for maximum-likelihood or maximum a posteriori estimation in problems with incomplete data [19,20]. In the case of semi-supervised classification, the unlabeled data are considered incomplete because they do not have class labels. The basic EM approach first designs a probabilistic classifier (e.g., a Gaussian classifier) with the available data set D_l. Then, such classifier is used to assign probabilistically-weighted class labels to unlabeled examples by calculating the expectation of the missing class labels. Then a new classifier is trained using both the original labelled data and the formerly unlabelled data, and the process is repeated.

The main advantage of EM approach is that it allows exploiting, in a theoretically well grounded way, both labelled and unlabelled data. Therefore, it meets the main requirement of semi-supervised classification in a natural way. In addition, it is quite general, and it has been used with different probabilistic models of classifiers [10,20].

Results on different classification tasks showed that EM methods allow exploiting unlabelled data effectively. For example, Nigam et al. show that unlabelled data used with EM methods in a document categorization problem can reduce classification error by up to 30% [20]. On the other hand, Cohen et al. recently showed that EM methods can increase the classification accuracy only if the assumed probabilistic model of the classifier matches well with data distribution; otherwise, the use of unlabelled data can become counter productive [6].

Co-training

A co-training approach to semi-supervised classification was proposed by Blum and Mitchell in 1998 [21]. The basic idea can be illustrated with the web-page classification example originally used by the authors. Web pages can be characterized by two distinct types of features: features characterizing the text appearing in the web page, and features characterizing the hyperlinks pointing to the page. The key idea is to design two independent classifiers using the text and hyperlink features separately. These two classifiers are trained with the initial, small, labelled data set D_l, and it is assumed that the classifiers will exhibit a low, but better than random, accuracy (it is worth noting that this definition closely resembles the concept of "weak" classifier used in the MCS field). Each classifier is then applied to the unlabeled examples. For each classifier, the unlabelled examples that received the highest confidence by this classifier are added to labelled data, so that the two classifiers contribute to increase the data set. Both the classifiers are re-trained with this augmented data set, and the process is repeated a specified number of times. When the co-training process finishes, the two resulting classifiers can be combined by the product of the outputs. However, it is worth noting that the basic co-training algorithm does not contain this combination phase. In fact, the main goal of this approach is to increase the accuracy of the two individual classifiers by co-training. But the reported results have shown that the combination can further increase the classification accuracy.

Intuitively, co-training is expected to work because the two classifiers are assumed to be "complementary", thanks to the use of disparate features. In particular, a classifier may assign correct labels to certain examples while it may be difficult for the other classifier to do so. Therefore, each classifier can increase the training set to be used by the other classifier. It should be also noted that co-training is expected to increase the size of the training set more quickly than what each individual classifier could do using a decision-directed, self-learning, mechanism.

Several authors reported experimental results which show the effectiveness of co-training [9,22,28]. Blum and Mitchell provided some theoretical support for co-training within the PAC learning framework [21]. However, the fundamental issue about the conditions under which, and the extent at which, the use of unlabeled data with co-training can increase classification accuracy is basically unsolved.

Although, to the best of our knowledge, co-training has always been used with just two classifiers, Blum and Mitchell pointed out that this approach can be used with a larger ensemble of classifiers, particularly if are "independent". It is easy to see that this makes co-training a good candidate for the development of semi-supervised MCS. However, it should be noted that co-training was not originally meant as a method to create and combine classifiers; in fact, the basic co-training algorithm does not contain any classifier combination. Therefore, a lot of work remains to be done about the use of co-training to create good classifier ensembles, and, in particular, about the combination techniques when the independence assumption is likely to be violated.

Active Learning

This approach assumes the availability of an external "oracle" to assign class labels [9]. Basically, unlabeled examples are repeatedly selected, and the oracle (e.g., a human expert) is asked to assign class labels to such data. The goal of active learning is to select the most informative unlabeled examples, in order to effectively train the classifier with the minimum number of calls to the oracle. To this end, different selection strategies have been proposed [9]. For the purpose of this paper, the so-