José Luis Vicedo
Patricio Martínez-Barco
Rafael Muñoz
Maximiliano Saiz Noeda (Eds.)

LNAI 3230

Advances in Natural Language Processing

4th International Conference, EsTAL 2004
Alicante, Spain, October 2004
Proceedings

Springer

José Luis Vicedo
Patricio Martínez-Barco Rafael Muñoz
Maximiliano Saiz Noeda (Eds.)

Advances in Natural Language Processing

4th International Conference, EsTAL 2004
Alicante, Spain, October 20-22, 2004
Proceedings

Springer

Series Editors

Jaime G. Carbonell, Carnegie Mellon University, Pittsburgh, PA, USA
Jörg Siekmann, University of Saarland, Saarbrücken, Germany

Volume Editors

José Luis Vicedo
Patricio Martínez-Barco
Rafael Muñoz
Maximiliano Saiz Noeda
Universidad de Alicante
Departamento de Lenguajes y Sistemas Informáticos
Carretera de San Vicente del Raspeig
03690 San Vicente del Raspeig, Alicante, Spain
E-mail: {vicedo;patricio;rafael;max}@dlsi.ua.es

Library of Congress Control Number: 2004113295

CR Subject Classification (1998): I.2.7, F.4.2-3, I.2, H.3, I.7

ISSN 0302-9743
ISBN 3-540-23498-5 Springer Berlin Heidelberg New York

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 11334347 06/3142 5 4 3 2 1 0

Lecture Notes in Artificial Intelligence 3230

Edited by J. G. Carbonell and J. Siekmann

Subseries of Lecture Notes in Computer Science

Lecture Notes in Artificial Intelligence (LNAI)

Vol. 3265: R.E. Frederking, K.B. Taylor (Eds.), Machine Translation: From Real Users to Research. XI, 392 pages. 2004.

Vol. 3264: G. Paliouras, Y. Sakakibara (Eds.), Grammatical Inference: Algorithms and Applications. XI, 291 pages. 2004.

Vol. 3257: E. Motta, N.R. Shadbolt, A. Stutt, N. Gibbins (Eds.), Engineering Knowledge in the Age of the Semantic Web. XVII, 517 pages. 2004.

Vol. 3249: B. Buchberger, J.A. Campbell (Eds.), Artificial Intelligence and Symbolic Computation. X, 285 pages. 2004.

Vol. 3245: E. Suzuki, S. Arikawa (Eds.), Discovery Science. XIV, 430 pages. 2004.

Vol. 3244: S. Ben-David, J. Case, A. Maruoka (Eds.), Algorithmic Learning Theory. XIV, 505 pages. 2004.

Vol. 3238: S. Biundo, T. Frühwirth, G. Palm (Eds.), KI 2004: Advances in Artificial Intelligence. XI, 467 pages. 2004.

Vol. 3230: J.L. Vicedo, P. Martínez-Barco, R. Muñoz, M.Saiz Noeda (Eds.), Advances in Natural Language Processing. XII, 488 pages. 2004.

Vol. 3229: J.J. Alferes, J. Leite (Eds.), Logics in Artificial Intelligence. XIV, 744 pages. 2004.

Vol. 3215: M.G.. Negoita, R.J. Howlett, L.C. Jain (Eds.), Knowledge-Based Intelligent Information and Engineering Systems. LVII, 906 pages. 2004.

Vol. 3214: M.G.. Negoita, R.J. Howlett, L.C. Jain (Eds.), Knowledge-Based Intelligent Information and Engineering Systems. LVIII, 1302 pages. 2004.

Vol. 3213: M.G.. Negoita, R.J. Howlett, L.C. Jain (Eds.), Knowledge-Based Intelligent Information and Engineering Systems. LVIII, 1280 pages. 2004.

Vol. 3209: B. Berendt, A. Hotho, D. Mladenic, M. van Someren, M. Spiliopoulou, G. Stumme (Eds.), Web Mining: From Web to Semantic Web. IX, 201 pages. 2004.

Vol. 3206: P. Sojka, I. Kopecek, K. Pala (Eds.), Text, Speech and Dialogue. XIII, 667 pages. 2004.

Vol. 3202: J.-F. Boulicaut, F. Esposito, F. Giannotti, D. Pedreschi (Eds.), Knowledge Discovery in Databases: PKDD 2004. XIX, 560 pages. 2004.

Vol. 3201: J.-F. Boulicaut, F. Esposito, F. Giannotti, D. Pedreschi (Eds.), Machine Learning: ECML 2004. XVIII, 580 pages. 2004.

Vol. 3194: R. Camacho, R. King, A. Srinivasan (Eds.), Inductive Logic Programming. XI, 361 pages. 2004.

Vol. 3192: C. Bussler, D. Fensel (Eds.), Artificial Intelligence: Methodology, Systems, and Applications. XIII, 522 pages. 2004.

Vol. 3191: M. Klusch, S. Ossowski, V. Kashyap, R. Unland (Eds.), Cooperative Information Agents VIII. XI, 303 pages. 2004.

Vol. 3187: G. Lindemann, J. Denzinger, I.J. Timm, R. Unland (Eds.), Multiagent System Technologies. XIII, 341 pages. 2004.

Vol. 3176: O. Bousquet, U. von Luxburg, G. Rätsch (Eds.), Advanced Lectures on Machine Learning. IX, 241 pages. 2004.

Vol. 3171: A.L.C. Bazzan, S. Labidi (Eds.), Advances in Artificial Intelligence – SBIA 2004. XVII, 548 pages. 2004.

Vol. 3159: U. Visser, Intelligent Information Integration for the Semantic Web. XIV, 150 pages. 2004.

Vol. 3157: C. Zhang, H. W. Guesgen, W.K. Yeap (Eds.), PRICAI 2004: Trends in Artificial Intelligence. XX, 1023 pages. 2004.

Vol. 3155: P. Funk, P.A. González Calero (Eds.), Advances in Case-Based Reasoning. XIII, 822 pages. 2004.

Vol. 3139: F. Iida, R. Pfeifer, L. Steels, Y. Kuniyoshi (Eds.), Embodied Artificial Intelligence. IX, 331 pages. 2004.

Vol. 3131: V. Torra, Y. Narukawa (Eds.), Modeling Decisions for Artificial Intelligence. XI, 327 pages. 2004.

Vol. 3127: K.E. Wolff, H.D. Pfeiffer, H.S. Delugach (Eds.), Conceptual Structures at Work. XI, 403 pages. 2004.

Vol. 3123: A. Belz, R. Evans, P. Piwek (Eds.), Natural Language Generation. X, 219 pages. 2004.

Vol. 3120: J. Shawe-Taylor, Y. Singer (Eds.), Learning Theory. X, 648 pages. 2004.

Vol. 3097: D. Basin, M. Rusinowitch (Eds.), Automated Reasoning. XII, 493 pages. 2004.

Vol. 3071: A. Omicini, P. Petta, J. Pitt (Eds.), Engineering Societies in the Agents World. XIII, 409 pages. 2004.

Vol. 3070: L. Rutkowski, J. Siekmann, R. Tadeusiewicz, L.A. Zadeh (Eds.), Artificial Intelligence and Soft Computing - ICAISC 2004. XXV, 1208 pages. 2004.

Vol. 3068: E. André, L. Dybkjær, W. Minker, P. Heisterkamp (Eds.), Affective Dialogue Systems. XII, 324 pages. 2004.

Vol. 3067: M. Dastani, J. Dix, A. El Fallah-Seghrouchni (Eds.), Programming Multi-Agent Systems. X, 221 pages. 2004.

Vol. 3066: S. Tsumoto, R. Słowiński, J. Komorowski, J.W. Grzymała-Busse (Eds.), Rough Sets and Current Trends in Computing. XX, 853 pages. 2004.

Vol. 3065: A. Lomuscio, D. Nute (Eds.), Deontic Logic in Computer Science. X, 275 pages. 2004.

Vol. 3060: A.Y. Tawfik, S.D. Goodwin (Eds.), Advances in Artificial Intelligence. XIII, 582 pages. 2004.

Vol. 3056: H. Dai, R. Srikant, C. Zhang (Eds.), Advances in Knowledge Discovery and Data Mining. XIX, 713 pages. 2004.

Vol. 3055: H. Christiansen, M.-S. Hacid, T. Andreasen, H.L. Larsen (Eds.), Flexible Query Answering Systems. X, 500 pages. 2004.

Vol. 3048: P. Faratin, D.C. Parkes, J.A. Rodríguez-Aguilar, W.E. Walsh (Eds.), Agent-Mediated Electronic Commerce V. XI, 155 pages. 2004.

Vol. 3040: R. Conejo, M. Urretavizcaya, J.-L. Pérez-de-la-Cruz (Eds.), Current Topics in Artificial Intelligence. XIV, 689 pages. 2004.

Vol. 3035: M.A. Wimmer (Ed.), Knowledge Management in Electronic Government. XII, 326 pages. 2004.

Vol. 3034: J. Favela, E. Menasalvas, E. Chávez (Eds.), Advances in Web Intelligence. XIII, 227 pages. 2004.

Vol. 3030: P. Giorgini, B. Henderson-Sellers, M. Winikoff (Eds.), Agent-Oriented Information Systems. XIV, 207 pages. 2004.

Vol. 3029: B. Orchard, C. Yang, M. Ali (Eds.), Innovations in Applied Artificial Intelligence. XXI, 1272 pages. 2004.

Vol. 3025: G.A. Vouros, T. Panayiotopoulos (Eds.), Methods and Applications of Artificial Intelligence. XV, 546 pages. 2004.

Vol. 3020: D. Polani, B. Browning, A. Bonarini, K. Yoshida (Eds.), RoboCup 2003: Robot Soccer World Cup VII. XVI, 767 pages. 2004.

Vol. 3012: K. Kurumatani, S.-H. Chen, A. Ohuchi (Eds.), Multi-Agents for Mass User Support. X, 217 pages. 2004.

Vol. 3010: K.R. Apt, F. Fages, F. Rossi, P. Szeredi, J. Váncza (Eds.), Recent Advances in Constraints. VIII, 285 pages. 2004.

Vol. 2990: J. Leite, A. Omicini, L. Sterling, P. Torroni (Eds.), Declarative Agent Languages and Technologies. XII, 281 pages. 2004.

Vol. 2980: A. Blackwell, K. Marriott, A. Shimojima (Eds.), Diagrammatic Representation and Inference. XV, 448 pages. 2004.

Vol. 2977: G. Di Marzo Serugendo, A. Karageorgos, O.F. Rana, F. Zambonelli (Eds.), Engineering Self-Organising Systems. X, 299 pages. 2004.

Vol. 2972: R. Monroy, G. Arroyo-Figueroa, L.E. Sucar, H. Sossa (Eds.), MICAI 2004: Advances in Artificial Intelligence. XVII, 923 pages. 2004.

Vol. 2969: M. Nickles, M. Rovatsos, G. Weiss (Eds.), Agents and Computational Autonomy. X, 275 pages. 2004.

Vol. 2961: P. Eklund (Ed.), Concept Lattices. IX, 411 pages. 2004.

Vol. 2953: K. Konrad, Model Generation for Natural Language Interpretation and Analysis. XIII, 166 pages. 2004.

Vol. 2934: G. Lindemann, D. Moldt, M. Paolucci (Eds.), Regulated Agent-Based Social Systems. X, 301 pages. 2004.

Vol. 2930: F. Winkler (Ed.), Automated Deduction in Geometry. VII, 231 pages. 2004.

Vol. 2926: L. van Elst, V. Dignum, A. Abecker (Eds.), Agent-Mediated Knowledge Management. XI, 428 pages. 2004.

Vol. 2923: V. Lifschitz, I. Niemelä (Eds.), Logic Programming and Nonmonotonic Reasoning. IX, 365 pages. 2004.

Vol. 2915: A. Camurri, G. Volpe (Eds.), Gesture-Based Communication in Human-Computer Interaction. XIII, 558 pages. 2004.

Vol. 2913: T.M. Pinkston, V.K. Prasanna (Eds.), High Performance Computing - HiPC 2003. XX, 512 pages. 2003.

Vol. 2903: T.D. Gedeon, L.C.C. Fung (Eds.), AI 2003: Advances in Artificial Intelligence. XVI, 1075 pages. 2003.

Vol. 2902: F.M. Pires, S.P. Abreu (Eds.), Progress in Artificial Intelligence. XV, 504 pages. 2003.

Vol. 2892: F. Dau, The Logic System of Concept Graphs with Negation. XI, 213 pages. 2003.

Vol. 2891: J. Lee, M. Barley (Eds.), Intelligent Agents and Multi-Agent Systems. X, 215 pages. 2003.

Vol. 2882: D. Veit, Matchmaking in Electronic Markets. XV, 180 pages. 2003.

Vol. 2871: N. Zhong, Z.W. Raś, S. Tsumoto, E. Suzuki (Eds.), Foundations of Intelligent Systems. XV, 697 pages. 2003.

Vol. 2854: J. Hoffmann, Utilizing Problem Structure in Planing. XIII, 251 pages. 2003.

Vol. 2843: G. Grieser, Y. Tanaka, A. Yamamoto (Eds.), Discovery Science. XII, 504 pages. 2003.

Vol. 2842: R. Gavaldá, K.P. Jantke, E. Takimoto (Eds.), Algorithmic Learning Theory. XI, 313 pages. 2003.

Vol. 2838: N. Lavrač, D. Gamberger, L. Todorovski, H. Blockeel (Eds.), Knowledge Discovery in Databases: PKDD 2003. XVI, 508 pages. 2003.

Vol. 2837: N. Lavrač, D. Gamberger, L. Todorovski, H. Blockeel (Eds.), Machine Learning: ECML 2003. XVI, 504 pages. 2003.

Vol. 2835: T. Horváth, A. Yamamoto (Eds.), Inductive Logic Programming. X, 401 pages. 2003.

Vol. 2821: A. Günter, R. Kruse, B. Neumann (Eds.), KI 2003: Advances in Artificial Intelligence. XII, 662 pages. 2003.

Vol. 2807: V. Matoušek, P. Mautner (Eds.), Text, Speech and Dialogue. XIII, 426 pages. 2003.

Vol. 2801: W. Banzhaf, J. Ziegler, T. Christaller, P. Dittrich, J.T. Kim (Eds.), Advances in Artificial Life. XVI, 905 pages. 2003.

Vol. 2797: O.R. Zaïane, S.J. Simoff, C. Djeraba (Eds.), Mining Multimedia and Complex Data. XII, 281 pages. 2003.

Vol. 2792: T. Rist, R.S. Aylett, D. Ballin, J. Rickel (Eds.), Intelligent Virtual Agents. XV, 364 pages. 2003.

Vol. 2782: M. Klusch, A. Omicini, S. Ossowski, H. Laamanen (Eds.), Cooperative Information Agents VII. XI, 345 pages. 2003.

Vol. 2780: M. Dojat, E. Keravnou, P. Barahona (Eds.), Artificial Intelligence in Medicine. XIII, 388 pages. 2003.

Vol. 2777: B. Schölkopf, M.K. Warmuth (Eds.), Learning Theory and Kernel Machines. XIV, 746 pages. 2003.

Vol. 2752: G.A. Kaminka, P.U. Lima, R. Rojas (Eds.), RoboCup 2002: Robot Soccer World Cup VI. XVI, 498 pages. 2003.

Preface

EsTAL – España for Natural Language Processing – continued on from the three previous conferences: FracTAL, held at the Université de Franch-Comté, Besançon (France) in December 1997, VexTAL, held at Venice International University, Cá Foscari (Italy), in November 1999, and PorTAL, held at the Universidade do Algarve, Faro (Portugal), in June 2002. The main goals of these conferences have been: (i) to bring together the international NLP community; (ii) to strengthen the position of local NLP research in the international NLP community; and (iii) to provide a forum for discussion of new research and applications.

EsTAL contributed to achieving these goals and increasing the already high international standing of these conferences, largely due to its Program Committee, composed of renowned researchers in the field of natural language processing and its applications. This clearly contributed to the significant number of papers submitted (72) by researchers from (18) different countries.

The scope of the conference was structured around the following main topics: (i) *computational linguistics research* (spoken and written language analysis and generation; pragmatics, discourse, semantics, syntax and morphology; lexical resources; word sense disambiguation; linguistic, mathematical, and psychological models of language; knowledge acquisition and representation; corpus-based and statistical language modelling; machine translation and translation aids; computational lexicography), and (ii) *monolingual and multilingual intelligent language processing and applications* (information retrieval, extraction and question answering; automatic summarization; document categorization; natural language interfaces; dialogue systems and evaluation of systems).

Each paper was revised by three reviewers from the Program Committee or by external referees designed by them. All those who contributed are mentioned on the following pages. The review process led to the selection of 42 papers for presentation. They have been published in this volume.

We would like to express here our thanks to all the reviewers for their quick and excellent work. We extend these thanks to our invited speakers, Walter Daelemans and Rada Mihalcea for their valuable contribution, which undoubtedly increased the interest in the conference. We are also indebted to a number of individuals for taking care of specific parts of the conference program. Specially, to Miguel Angel Varó who built and maintained all Web services for the conference.

Finally, we want to thank the University of Alicante, and, specially, the Office of the Vice-President for Extracurricular Activities (*Vicerrectorado de Extensión Universitaria*) and the Department of Software and Computing Systems (*Departamento de Lenguajes y Sistemas Informáticos*) because of their support of this conference.

October 2004 José L. Vicedo and Patricio Martinez-Barco

Organization

Program and Conference Chairs

José L. Vicedo (University of Alicante, Spain)
Patricio Martinez-Barco (University of Alicante, Spain)

Organization Chairs

Rafael Muñoz (University of Alicante, Spain)
Maximiliano Saiz (University of Alicante, Spain)

Program Committee

Alfonso Ureña (University of Jaen, Spain)
Bernardo Magnini (ITC-irst, Italy)
Dan Moldovan (University of Texas at Dallas, USA)
Elisabete Ranchhod (University of Lisbon, Portugal)
German Rigau (University of the Basque Country, Spain)
Hans Uszkoreit (Saarland University at Saarbrücken, Germany)
Henri Zingle (University of Nice, France)
Horacio Rodríguez (Technical University of Catalonia, Spain)
Horacio Saggion (University of Sheffield, UK)
Igor Melcuk (University of Montreal, Canada)
Julio Gonzalo (Universidad Nacional de Educación a Distancia (UNED), Spain)
Lidia Moreno (Polytechnical University of Valencia, Spain)
Manuel de Buenaga (European University of Madrid, Spain)
Manuel Palomar (University of Alicante, Spain)
Massimo Poesio (University of Essex, UK)
Nuno Mamede (INESC-ID Lisbon, Portugal)
Peter Greenfield (Centre Lucien Tesnière, Univ. of Franche-Comté, France)
Pierre-André Buvet (University of Paris 13, France)
Rodolfo Delmonte (University of Venice, Italy)
Ruslan Mitkov (University of Wolverhampton, UK)
Sanda Harabagiu (University of Texas at Dallas, USA)
Stephane Chaudiron (Ministry of Technology and Research, France)
Sylviane Cardey (Centre Lucien Tesnière, Univ. of Franche-Comté, France)
Victor Díaz (University of Seville, Spain)
Werner Winiwarter (University of Vienna, Austria)

Referees

Alfonso Ureña
Alicia Ageno
Ana García Serrano
Anabela Barreiro
Antonio Ferrandez
Armando Suarez
Bernardo Magnini
Carlo Strapparava
Damián López
Dan Moldovan
Diamantino Caseiro
Elisabete Ranchhod
Fernando Llopis
Finley Lacatusu
Georgiana Puscasu
German Rigau
Hans Uszkoreit
Henri Zingle
Horacio Rodríguez
Horacio Saggion
Igor Melcuk
José L. Vicedo
Juan A. Pérez-Ortiz
Julio Gonzalo
Laura Hasler
Le An Ha
Lidia Moreno
Luis Oliveira
Luisa Coheur
Manuel de Buenaga
Manuel García Vega
Manuel Palomar
María Teresa Martín
Massimo Poesio
Maximiliano Saiz Noeda
Miguel Angel García
Mijail Alexandrov
Mikel Forcada
Nuno Mamede
Olivia Sanchez
Patricio Martinez-Barco
Paul Morarescu
Paula Carvalho
Peter Greenfield
Pierre-André Buvet
Rafael Muñoz
Ramón López
Ricardo Ribeiro
Richard Evans
Rodolfo Delmonte
Ruslan Mitkov
Sanda Harabagiu
Stephane Chaudiron
Sylviane Cardey
Victor Díaz
Viktor Pekar
Werner Winiwarter
Zornitsa Kozareva

Table of Contents

Adaptive Selection of Base Classifiers in One-Against-All Learning for Large Multi-labeled Collections

Arturo Montejo Ráez[1], Luís Alfonso Ureña López[2], and Ralf Steinberger[3]

[1] European Laboratory for Nuclear Research, Geneva, Switzerland
[2] Department of Computer Science, University of Jaén, Spain
[3] European Commission, Joint Research Centre, Ispra, Italy

Abstract. In this paper we present the problem found when studying an automated text categorization system for a collection of High Energy Physics (HEP) papers, which shows a very large number of possible classes (over 1,000) with highly imbalanced distribution. The collection is introduced to the scientific community and its imbalance is studied applying a new indicator: the *inner imbalance degree*. The one-against-all approach is used to perform multi-label assignment using Support Vector Machines. Over-weighting of positive samples and S-Cut thresholding is compared to an approach to automatically select a classifier for each class from a set of candidates. We also found that it is possible to reduce computational cost of the classification task by discarding classes for which classifiers cannot be trained successfully.

1 Introduction

The automatic assignment of keywords to documents using full-text data is a subtask of *Text Categorization*, a growing area where Information Retrieval techniques and Machine Learning algorithms meet offering solutions to problems with real world collections.

We can distinguish three paradigms in text categorization: the *binary* case, the *multi-class* case and the *multi-label* case. In the binary case a sample either belongs or does not belong to one of two given classes. In the multi-class case a sample belongs to just one of a set of m classes. Finally, in the multi-label case, a sample may belong to several classes at the same time, that is, classes are *overlapped*. In binary classification a classifier is trained, by means of supervised algorithms, to assign a sample document to one of two possible sets. These sets are usually referred to as belonging (positive) or not belonging (negative) samples respectively (the one-against-all approach), or to two disjoint classes (the one-against-one approach). For these two binary classification tasks we can select among a wide range of algorithms, including Naïve Bayes, Linear Regression, Support Vector Machines (SVM) [8] and LVQ [11]. SVM has been reported to outperform the other algorithms. The binary case has been set as a base case from which the two other cases are derived. In multi-class and multi-label assignment,

J. L. Vicedo et al. (Eds.): EsTAL 2004, LNAI 3230, pp. 1–12, 2004.

the traditional approach consists of training a binary classifier for each class, and then, whenever the binary base case returns a measure of confidence on the classification, assigning either the top ranked one (multi-class assignment) or a given number of the top ranked ones (multi-label assignment). More details about these three paradigms can be found in [1]). We will refer to the ranking approach as the *battery* strategy because inter-dependency is not taken into consideration.

Another approach for multi-labeling consists of returning all those classes whose binary classifiers provide a positive answer for the sample. It has the advantage of allowing different binary classifiers for each class, since inter-class scores do not need to be coherent (since there is no ranking afterwards). Better results have been reported when applying one-against-one in multi-class classification [1], but in our multi-label case this is not an option because any class could theoretically appear together with any other class, making it difficult to establish disjoint assignments. This is the reason why one-against-all deserves our attention in the present work.

Although classification is subject to intense research (see [18]), some issues demand more attention than they have been given so far. In particular, problems relating to *multi-label* classification would require more attention. However, due to the lack of available resources (mainly multi-labeled document collections), this area advances more slowly than others. Furthermore, multi-label assignment should not simply be studied as a general multi-class problem (which itself is rather different from the binary case), but it needs to be considered as a special case with additional requirements. For instance, in multi-label cases, some classes are inter-related, the degree of imbalance is usually radically different from one class to the next and, from a performance point of view, the need of comparing a sample to every single classifier is a waste of resources.

2 The Class Imbalance Problem

Usually, multi-labeled collections make use of a wide variety of classes, resulting in an unequal distribution of classes throughout the collection and a high number of rare classes. This means not only that there is a strong imbalance between positive and negative samples, but also that some classes are used much more frequently than other classes. This phenomenon, known as the *class imbalance problem*, is especially relevant for algorithms like the C4.5 classification tree [4, 3] and margin-based classifiers like SVM [16, 20, 7].

Extensive studies have been carried out on this subject as reported by Japkowicz [7], identifying three major issues in the class imbalance problem: *concept complexity*, *training set size* and *degree of imbalance*. Concept complexity refers to the degree of "sparsity" of a certain class in the feature space (the space where document vectors are represented). This means that a hypothetical clustering algorithm acting on a class with high concept complexity would establish many small clusters for the same class. Regarding the second issue, i.e. the lack of a significantly large training sets, the only possible remedy is the usage of

over-sampling when the amount of available samples is insufficient, and under-sampling techniques for classes with too many samples, e.g. just using a limited number of samples for training a SVM, by selecting those positive and negative samples that are close to each other in the feature space. The validity of these techniques is also subject to debate [4]. Finally, Japkowicz defines the degree of imbalance as an index to indicate how much a class is more represented over another, including both the degree of imbalance between classes (what we call *inter-class imbalance*) and between its positive and negative samples (what we call the *inner imbalance degree*). Unfortunately, Japkowicz defined these values for her work towards the generation of an artificial collection and rewrote them later to fit specific problems regarding fixed parameters and the C5.0 algorithm, which make them difficult to manipulate. For these reasons, we cannot reuse her equations and propose here a variant focusing on the multi-label case.

We define the *inner imbalance degree* of a certain class i as a measure of the positive samples over the total of samples:

$$i_i = |1 - 2n_i/n| \tag{1}$$

where

n is the total number of samples and
n_i is the total number of samples having the class i in their labels.

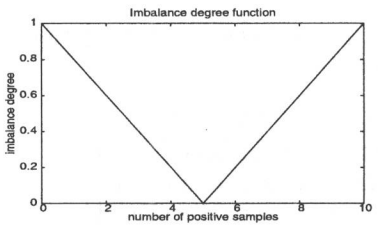


Fig. 1. The linear 'imbalance degree' function

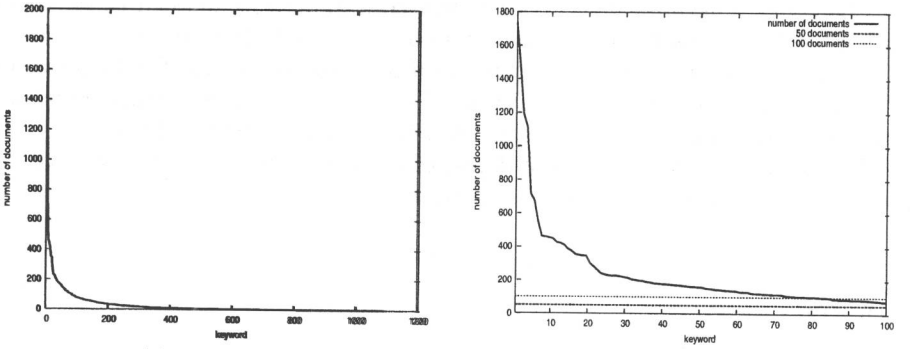
Japkowicz' definition of imbalance degree helps in the generation of artificial distributions of documents to classes. Its value does not lie within a defined range, which makes it difficult to manipulate and compare with the degree of other classes in different partitions. The value proposed in equation 1 is zero for perfectly balanced classes, i.e. when the number of positive and negative samples are the same. It has a value of 1 when all samples are either positive or negative for that class. Its linear behavior is shown in figure 1 and, as we can see, it varies within the range [0,1].

3 The HEP Collection

A very suitable document set for multi-label categorization research is the HEP collection of preprints, available from the European Laboratory for Nuclear Research. Some experiments have been carried out using this collection ([13, 12]), and its interesting distribution of classes allows us to carry out a number of experiments and to design a new approach. An analysis of the collection has shown that there is the typical high level of imbalance among classes. If a given class is rarely represented in a collection, we can intuitively foresee a biased training

that will yield classifiers with a low performance. It is clear that, if the collection were perfectly balanced, we could expect better categorization results, due to better learning.

The `hep-ex` partition of the HEP collection is composed of 2802 abstracts related to experimental high-energy physics that are indexed with 1093 main keywords (the categories).[1] Figure 2 shows the distribution of keywords across the collection.



(a) All classes (b) 100 most frequent

Fig. 2. Distribution of classes across documents in the `hep-ex` partition

As we can see, this partition is **very** imbalanced: only 84 classes are represented by more than 100 samples and only five classes by more than 1000. The uneven use is particularly noticeable for the ten most frequent keywords: In table 1 the left column shows the number of positive samples of a keyword and the right column shows the percentage over the total of samples in the collection.

Table 1. The ten most frequent main keywords in the `hep-ex` partition

No. docs.	Keyword
1898 (67%)	electron positron
1739 (62%)	experimental results
1478 (52%)	magnetic detector
1190 (42%)	quark
1113 (39%)	talk
715 (25%)	Z0
676 (24%)	anti-p p
551 (19%)	neutrino
463 (16%)	W
458 (16%)	jet

We can now study this collection applying the inner imbalance degree measure defined in equation 1. The two graphs in figures 3a and 3b show the inner imbalance degree for the main keywords in the `hep-ex` partition. We can notice how fast

[1] We did not consider the keywords related to reaction and energy because they are based on formulae and other specific data that is not easily identifiable in the plain-text version of a paper.