Advances in Intelligent Systems and Computing 403

Robert Burduk
Konrad Jackowski
Marek Kurzyński
Michał Woźniak
Andrzej Żołnierek *Editors*

# Proceedings of the 9th International Conference on Computer Recognition Systems CORES 2015

Springer

# Advances in Intelligent Systems and Computing

Volume 403

**Series editor**

Janusz Kacprzyk, Polish Academy of Sciences, Warsaw, Poland
e-mail: kacprzyk@ibspan.waw.pl

## *About this Series*

The series "Advances in Intelligent Systems and Computing" contains publications on theory, applications, and design methods of Intelligent Systems and Intelligent Computing. Virtually all disciplines such as engineering, natural sciences, computer and information science, ICT, economics, business, e-commerce, environment, healthcare, life science are covered. The list of topics spans all the areas of modern intelligent systems and computing.

The publications within "Advances in Intelligent Systems and Computing" are primarily textbooks and proceedings of important conferences, symposia and congresses. They cover significant recent developments in the field, both of a foundational and applicable character. An important characteristic feature of the series is the short publication time and world-wide distribution. This permits a rapid and broad dissemination of research results.

More information about this series at http://www.springer.com/series/11156

Robert Burduk · Konrad Jackowski
Marek Kurzyński · Michał Woźniak
Andrzej Żołnierek
Editors

# Proceedings of the 9th International Conference on Computer Recognition Systems CORES 2015

Springer

*Editors*

Robert Burduk
Department of Systems and Computer Networks
Wrocław University of Technology
Wrocław
Poland

Konrad Jackowski
Department of Systems and Computer Networks
Wrocław University of Technology
Wrocław
Poland

Marek Kurzyński
Department of Systems and Computer Networks
Wrocław University of Technology
Wrocław
Poland

Michał Woźniak
Department of Systems and Computer Networks
Wrocław University of Technology
Wrocław
Poland

Andrzej Żołnierek
Department of Systems and Computer Networks
Wrocław University of Technology
Wrocław
Poland

ISSN 2194-5357 ISSN 2194-5365 (electronic)
Advances in Intelligent Systems and Computing
ISBN 978-3-319-26225-3 ISBN 978-3-319-26227-7 (eBook)
DOI 10.1007/978-3-319-26227-7

Library of Congress Control Number: 2015959588

Printed on acid-free paper

This Springer imprint is published by SpringerNature
The registered company is Springer International Publishing AG Switzerland

# Preface

The goal of the CORES series of conferences is the development of theories, algorithms, and applications of pattern recognition methods. These conferences have always served as a very useful forum where researchers, practitioners, and students working in different areas of pattern recognition can come together and help each other keeping up with this active field of research.

This book is collection of 79 carefully selected works, which have been carefully reviewed by the experts from the domain and accepted for presentation during the 9th International Conference on Computer Recognition Systems CORES 2015.

We hope that the book can become the valuable source of information on contemporary research trends and the most popular areas of application.

The chapters are grouped into seven parts on the basis of the main topics they dealt with:

1. *Features, Learning, and Classifiers* consists of the works concerning new classification and machine learning methods;
2. *Biometrics* presents innovative theories, methodologies, and applications in the biometry;
3. *Data Stream Classification and Big Data Analytics* section concentrates on both data stream classification and massive data analytics issues;
4. *Image Processing and Computer Vision* is devoted to the problems of image processing and analysis;
5. *Medical Applications* presents chosen applications of intelligent methods into medical decision support software;
6. *Applications* describes several applications of the computer pattern recognition systems in the real decision problems;
7. *RGB-D Perception: Recent Developments and Applications* presents pattern recognition and image processing algorithms aimed specifically at applications in robotics.

Editors would like to express their deep thanks to authors for their valuable submissions and all reviewers for their hard work. Especially we would like to thank Dr. Tomasz Kornuta, Prof. Włodzimierz Kasprzak, Warsaw University of Technology, and Prof. Piotr Skrzypczyński, Poznań University of Technology, who organized a special session entitled "RGB-D Perception: Recent Developments and Applications." We would like to also thank Prof. Jerzy Stefanowski from Poznań University of Technology who helped Prof. Michał Woźniak and Bartosz Krawczyk, Wroclaw University of Technology to organize a special session on data stream classification and big data analytics.

We believe that this book could be a reference tool for scientists who deal with the problems of designing computer pattern recognition systems.

CORES 2015 enjoyed outstanding keynote speeches by distinguished guest speakers:

- Prof. Nitesh Chawla—University of Notre Dame, USA,
- Prof. Krzysztof J. Cios—Virginia Commonwealth University, USA,
- Prof. João Gama—University of Porto, Portugal,
- Prof. Francisco Herrera—University of Granada, Spain.

Last but not least, we would like to give special thanks to local organizing team (Robert Burduk, Kondrad Jackowski, Dariusz Jankowski, Bartosz Krawczyk, Maciej Krysmann, Jose Antonio Saez, Alex Savio, Paweł Trajdos, Marcin Zmyślony, and Andrzej Żołnierek) who did a great job.

This edition of the CORES was organized under the framework the ENGINE project, and thus the authors of the selected, best papers did not pay conference fee. ENGINE has received funding from the European Union's the Seventh Framework Programme for research, technological development, and demonstration under grant agreement no 316097. We would like to give our special thanks to the management of the ENGINE project—Prof. Przemysław Kazienko and Dr. Piotr Bródka—for this valuable sponsorship.

Also we would like to fully acknowledge the support from the Wrocław University of Technology, especially Prof. Andrzej Kasprzak—Chairs of Department of Systems and Computer Networks and vice Rector of the Wroclaw University of Technology, Prof. Jan Zarzycki—Dean of Faculty of Electronics, and Prof. Zdzisław Szalbierz—Dean of Faculty of Computer Science and Management, which has also supported this event.

We believe that this book could be a great reference tool for scientists who deal with the problems of designing computer pattern recognition systems.

Wrocław
July 2015

Robert Burduk
Konrad Jackowski
Marek Kurzyński
Michał Woźniak
Andrzej Żołnierek

# Contents

## Part II Biometrics

Part III Data Stream Classification and Big Data Analytics

Part IV Image Processing and Computer Vision

**Part V Medical Applications**

## Part VI Application

**Part VII RGB-D Perception: Recent Developments and Applications**

# Part I
# Features, Learning and Classifiers

# New Ordering-Based Pruning Metrics for Ensembles of Classifiers in Imbalanced Datasets

**Mikel Galar, Alberto Fernández, Edurne Barrenechea, Humberto Bustince and Francisco Herrera**

**Abstract** The task of classification with imbalanced datasets have attracted quite interest from researchers in the last years. The reason behind this fact is that many applications and real problems present this feature, causing standard learning algorithms not reaching the expected performance. Accordingly, many approaches have been designed to address this problem from different perspectives, i.e., data preprocessing, algorithmic modification, and cost-sensitive learning. The extension of the former techniques to ensembles of classifiers has shown to be very effective in terms of quality of the output models. However, the optimal value for the number of classifiers in the pool cannot be known a priori, which can alter the behaviour of the system. For this reason, ordering-based pruning techniques have been proposed to address this issue in standard classifier learning problems. The hitch is that those metrics are not designed specifically for imbalanced classification, thus hindering the performance in this context. In this work, we propose two novel adaptations for ordering-based pruning metrics in imbalanced classification, specifically the margin distance minimization and the boosting-based approach. Throughout a complete

M. Galar (✉) · E. Barrenechea · H. Bustince
Departamento de Automática y Computación, ISC (Institute of Smart Cities),
Universidad Pública de Navarra, Pamplona, Spain
e-mail: mikel.galar@unavarra.es

E. Barrenechea
e-mail: edurne.barrenechea@unavarra.es

H. Bustince
e-mail: bustince@unavarra.es

A. Fernández
Department of Computer Science, University of Jaén, Jaén, Spain
e-mail: alberto.fernandez@ujaen.es

F. Herrera
Department of Computer Science and Artificial Intelligence,
University of Granada, Granada, Spain
e-mail: herrera@decsai.ugr.es

R. Burduk et al. (eds.), *Proceedings of the 9th International Conference on Computer Recognition Systems CORES 2015*, Advances in Intelligent Systems and Computing 403, DOI 10.1007/978-3-319-26227-7_1

experimental study, our analysis shows the goodness of both schemes in contrast with the unpruned ensembles and the standard pruning metrics in Bagging-based ensembles.

**Keywords** Imbalanced datasets · Ensembles · Ordering-based pruning · Bagging

## 1 Introduction

The unequal distribution among examples of different classes in classification tasks is known as the problem of imbalanced datasets [9, 22]. The use of standard algorithms in this framework lead to undesirable solutions as the model is usually biased towards the most represented concepts of the problem [13]. Therefore, several approaches have been developed for addressing this issue, which can be divided into three large groups including preprocessing for resampling the training set [3], algorithmic adaptation of standard methods [2], and cost-sensitive learning [25]. Additionally, all these schemes can be integrated into an ensemble learning algorithm, increasing the capabilities and performance of the baseline approach [7, 8, 13]. An ensembles is a set of classifiers where its components are supposed to complement each other, so that the learning space is completely covered and the generalization capability is enhanced with respect to the single baseline learning classifier [18, 21]. When classifying a new instance, all individual members are queried and their decision is obtained in agreement. The total number of classifiers that compose an ensemble is not a synonym of its quality and performance [27], since several issues that can degrade its behavior must be taken into account: (1) the time elapsed in the learning and prediction stages; (2) the memory requirements; and (3) contradictions and/or redundancy among components of the ensemble. In accordance with the above, several proposals have been developed to carry out a pruning of classifiers within the ensemble [26]. Specifically, ordering-based pruning is based on a greedy approach that adds classifiers iteratively to the final set with respect to the maximization of a given heuristic metric, until a preestablished number of classifiers are selected [10, 16]. In this contribution, we aim at developing an adaptation of two popular metrics towards the scenario of classification with imbalanced datasets, i.e., Margin Distance Minimization (MDM) and Boosting-Based pruning (BB) [6, 15]. Specifically, we consider that the effect of each classifier in both classes must be analyzed after the construction of the classifier and not only before (for example, rebalancing the dataset). The goodness of this novel proposal is analyzed by means of a thorough experimental study, including a number of 66 different imbalanced problems. We have selected SMOTE-Bagging [23] and Under-Bagging [1] as ensemble learning schemes which, despite of being simple approaches, have shown to achieve a higher performance than many other more complex algorithms [7]. As in other related studies, we have selected the well-known C4.5 algorithm as baseline classifier [20]. Finally, our results are supported by means of non-parametric statistical tests [5]. In order to do so, this work is organized as follows. Section 2 briefly introduces the problem of imbalanced datasets. Then, Sect. 3 presents ordering-based pruning methodology, in which we describe standard metrics for performing this process and our adaptations to imbalanced classification. Next, the details about the experimental framework are provided in Sect. 4. The analysis and discussion of the experimental results are carried out in Sect. 5. Finally, Sect. 6 summarizes and concludes the work.

## 2 Basic Concepts on Classification with Imbalanced Datasets

Classification with imbalanced datasets appears when the distribution of instances between the classes of a given problem is quite different [13, 19]. Therefore, this classification task needs a special treatment in order to carry out an accurate discrimination between both concepts, independently of their representation. The presence of classes with few data can generate sub-optimal classification models, since there is a bias towards the majority class when the learning process is guided by the standard accuracy metric. Furthermore, recent studies have shown that other data intrinsic characteristics have a significant influence for the correct identification of the minority class examples [13]. Some examples are overlapping, small-disjuncts, noise, and dataset shift. Solutions developed to address this problem can be categorized into three large groups [13]: (1) *data level solutions* [3], (2) *algorithmic level solutions* [2], and (3) *cost-sensitive solutions* [25]. Additionally, when the former approaches are integrated within an ensemble of classifiers, their effectiveness is enhanced [7, 13]. Finally, in order to evaluate the performance in such a particular classification scenario, the metrics used must be designed to take into account the class distribution. One commonly considered alternative is the Area Under the ROC curve (AUC) [11]. In those cases where the used classifier outputs a single solution, this measure can be simply computed by the following formula:

$$AUC = \frac{1 + TP_{rate} - FP_{rate}}{2} \tag{1}$$

where $TP_{rate} = \frac{TP}{TP+FN}$ and $FP_{rate} = \frac{FP}{TN+FP}$.

## 3 A Proposal for Ordering-Based Pruning Scheme for Ensembles in Imbalanced Domains

Ensemble-based classifiers [18] are composed by a set of so-called *weak learners*, i.e., low changes in data produce big changes in the induced models. Diversity is quite significant in the performance of this type of approach, implying that individual classifiers must be focused on different parts of the problem space [12]. There are mainly two types of ensemble techniques: Bagging [4] and Boosting [6]. In this work, we will focus on the first scheme, due to the simplicity for the integration of data preprocessing techniques [7]. In this methodology, an ensemble of classifiers is trained with different sets of random instances from the original training data. When classifying a new sample, all individual classifiers are fired and a majority or weighted vote is used to infer the class. The first parameter to take into account when building these types of models is the number of classifiers considered in the ensemble. In this sense, pruning methods were designed to obtain the "optimal" number of classifiers by carrying out a selection from a given pool of components of the ensemble. The hypothesis is that accuracy generally increases monotonically as more elements are added to the ensemble [10, 15, 16]. Most of pruning techniques make use of an heuristic function to seek for the reduced set of classifiers. In the case of *ordering-based pruning*, a metric that measures the goodness of adding each classifier to the ensemble is defined and the classifier with the highest value is added to the final sub-ensemble. The same process is performed until the size of the sub-ensemble reaches the specified parameter value. In this work, we study two popular pruning metrics MDM and BB [6, 15]. We describe both schemes and our adaptation to imbalanced classification below:

- *MDM* is based on certain distances among the output vectors of the ensembles. These output vectors have the length equal to the training set size, and their value at the $i$th position is either 1 or $-1$ depending on whether the $i$th example is classified or misclassified by the classifier. The signature vector of a sub-ensemble is computed as the sum of the vectors of the selected classifiers. To summarize, the aim is to add those classifiers with the objective of obtaining a signature vector of the sub-ensemble where all the components are positive, i.e., all examples are correctly predicted. For a wider description please refer to [16]. This method selects the classifier to be added depending on the closest Euclidean distance between an objective point (where every components are positive) and the signature vector of the sub-ensemble after adding the corresponding classifier. As a consequence, every example has the same weight in the computation of the distance, which can bias the selection to those classifiers favoring the majority class. Therefore, we compute the distance for the majority class examples and minority class examples independently. Then, distances are normalized by the number of examples used to compute them and added afterwards. That is, the same weight is given to both classes in the distance. This new metric is noted as *MDM-Imb*.
- *BB* selects the classifier that minimizes the cost with respect to the boosting scheme. This means that boosting algorithm is applied to compute the weights (costs) for each example in each iteration, but instead of training a classifier with these weights, the one that obtains the lowest cost from those in the pool is added to the sub-ensemble and weights are updated accordingly. Hence, it makes no difference whether classifiers were already learned using a boosting scheme or not. Different from the original boosting method, when no classifier has a weighted training error below 50 %, weights are reinitialized (equal weights for all the examples) and the method continues (whereas in boosting it is stopped). Once classifiers are

selected the scores assigned to each classifier by boosting are forgotten and not taken into account in the aggregation phase. It is well-known that boosting by itself is not capable of managing class imbalance problem [7]. For this reason, we have also adapted this approach in a similar manner as in the case of *MDM*. In boosting, every example has initially the same weight and these are updated according to whether they are correctly classified or not. Even though minority class instances should get larger weights if they are misclassified, these weights can be negligible compared with those of the majority class examples. Hence, before finding the classifier that minimizes the total cost, we normalize the weights of the examples of each class by half of their sum, so that both classes has the same importance when selecting the classifier (even though each example of each class would have a different weight). This is only done before selecting the classifier, and then weights are updated according to the original (non-normalized ones). This working procedure tries to be similar to that successfully applied in several boosting models such as EUS-Boost [8]. This second weighting approach is noted as *BB-Imb*.

## 4 Experimental Framework

Table 1 shows the benchmark problems selected for our study, in which the name, number of examples, number of attributes, and IR (ratio between the majority and minority class instances) are shown. Datasets are ordered with respect to their degree of imbalance. Multi-class problems were modified to obtain two-class imbalanced problems, defining the joint of one or more classes as positive and the joint of one or more classes as negative, as defined in the name of the dataset. A wider description for these problems can be found at http://www.keel.es/datasets.php. The estimates of AUC measure are obtained by means of a Distribution Optimally Balanced Stratified Cross-Validation (DOB-SCV) [17], as suggested in the specialized literature for working in imbalanced classification [14]. Cross-validation procedure is carried out using five folds, aiming to include enough positive class instances in the different folds. In accordance with the stochastic nature of the learning methods, these five folds are generated with five different seeds, and each one of the fivefold cross-validation is run five times. Therefore, experimental results are computed with the average of 125 runs. As ensemble techniques, we will make use of SMOTE-Bagging [23] and Under-Bagging [1]. In order to apply the pruning procedure, we will learn a number of 100 *classifiers* for each ensemble, choosing a subset of only 21 *classifiers* as suggested in the specialized literature [16]. The baseline ensemble models for comparison will use 40 classifiers as recommended in [7]. For *SMOTE-Bagging*, SMOTE configuration will be the standard with a 50 % class distribution, 5 neighbors for generating the synthetic samples, and Heterogeneous Value Difference Metric for computing the distance among the examples. Finally, both learning approaches include the C4.5 decision tree [20] as baseline classifier, using a confidence level at

**Table 1** Summary of imbalanced datasets used

| Name | #Ex. | #Atts. | IR | Name | #Ex. | #Atts. | IR |
|---|---|---|---|---|---|---|---|
| Glass1 | 214 | 9 | 1.82 | Glass04vs5 | 92 | 9 | 9.22 |
| Ecoli0vs1 | 220 | 7 | 1.86 | Ecoli0346vs5 | 205 | 7 | 9.25 |
| Wisconsin | 683 | 9 | 1.86 | Ecoli0347vs56 | 257 | 7 | 9.28 |
| Pima | 768 | 8 | 1.87 | Yeast05679vs4 | 528 | 8 | 9.35 |
| Iris0 | 150 | 4 | 2.00 | Ecoli067vs5 | 220 | 6 | 10.00 |
| Glass0 | 214 | 9 | 2.06 | Vowel0 | 988 | 13 | 10.10 |
| Yeast1 | 1484 | 8 | 2.46 | Glass016vs2 | 192 | 9 | 10.29 |
| Vehicle2 | 846 | 18 | 2.52 | Glass2 | 214 | 9 | 10.39 |
| Vehicle1 | 846 | 18 | 2.52 | Ecoli0147vs2356 | 336 | 7 | 10.59 |
| Vehicle3 | 846 | 18 | 2.52 | Led7digit02456789vs1 | 443 | 7 | 10.97 |
| Haberman | 306 | 3 | 2.78 | Ecoli01vs5 | 240 | 6 | 11.00 |
| Glass0123vs456 | 214 | 9 | 3.19 | Glass06vs5 | 108 | 9 | 11.00 |
| Vehicle0 | 846 | 18 | 3.25 | Glass0146vs2 | 205 | 9 | 11.06 |
| Ecoli1 | 336 | 7 | 3.36 | Ecoli0147vs56 | 332 | 6 | 12.28 |
| Newthyroid2 | 215 | 5 | 4.92 | Cleveland0vs4 | 1771 | 13 | 12.62 |
| Newthyroid1 | 215 | 5 | 5.14 | Ecoli0146vs5 | 280 | 6 | 13.00 |
| Ecoli2 | 336 | 7 | 5.46 | Ecoli4 | 336 | 7 | 13.84 |
| Segment0 | 2308 | 19 | 6.01 | Shuttle0vs4 | 1829 | 9 | 13.87 |
| Glass6 | 214 | 9 | 6.38 | Yeast1vs7 | 459 | 8 | 13.87 |
| Yeast3 | 1484 | 8 | 8.11 | Glass4 | 214 | 9 | 15.47 |
| Ecoli3 | 336 | 7 | 8.19 | Pageblocks13vs4 | 472 | 10 | 15.85 |
| Pageblocks0 | 5472 | 10 | 8.77 | Abalone918 | 731 | 8 | 16.68 |
| Ecoli034vs5 | 200 | 7 | 9.00 | Glass016vs5 | 184 | 9 | 19.44 |
| Yeast2vs4 | 514 | 8 | 9.08 | Shuttle2vs4 | 129 | 9 | 20.50 |
| Ecoli067vs35 | 222 | 7 | 9.09 | Yeast1458vs7 | 693 | 8 | 22.10 |
| Ecoli0234vs5 | 202 | 7 | 9.10 | Glass5 | 214 | 9 | 22.81 |
| Glass015vs2 | 506 | 8 | 9.12 | Yeast2vs8 | 482 | 8 | 23.10 |
| Yeast0359vs78 | 172 | 9 | 9.12 | Yeast4 | 1484 | 8 | 28.41 |
| Yeast0256vs3789 | 1004 | 8 | 9.14 | Yeast1289vs7 | 947 | 8 | 30.56 |
| Yeast02579vs368 | 1004 | 8 | 9.14 | Yeast5 | 1484 | 8 | 32.73 |
| Ecoli046vs5 | 203 | 6 | 9.15 | Yeast6 | 1484 | 8 | 41.40 |
| Ecoli01vs235 | 244 | 7 | 9.17 | Ecoli0137vs26 | 281 | 7 | 39.15 |
| Ecoli0267vs35 | 244 | 7 | 9.18 | Abalone19 | 4174 | 8 | 129.44 |

0.25, with 2 as the minimum number of item-sets per leaf, and the application of pruning will be used to obtain the final tree. Reader may refer to [7] in order to get a thorough description of the former ensemble methods. Finally, we will make use of Wilcoxon signed-rank test [24] to find out whether significant differences exist between a pair of algorithms.

## 5 Experimental Study

Our analysis is focused on determining whether the new proposed metrics, specifically designed for dealing with class imbalance, are well-suited for this problem with respect to the original metrics, i.e., *BB* and *MDM*. Additionally, we will analyze the improvement in the performance results with respect to the original ensemble model. The average values for the experimental results are shown in Table 2, whereas full results are shown in Table 3. Regarding the comparison between the pruning schemes, in the case of *BB* and *BB-Imb* we find that for SMOTE-Bagging the metric adapted for imbalanced classification achieves a higher average performance. Regarding Under-Bagging, the relative differences are below 1 % in favour of the standard approach. On the other hand, the analysis for *MDM* and *MDM-Imb* metrics shows the need for the imbalanced approach, as it stands out looking at the experimental results. Finally, the robustness of the imbalanced metrics must be stressed in accordance with the low standard deviation shown with respect to the standard case. In order to determine statistically the best suited metric, we carry out a Wilcoxon pairwise test in Table 4. We have included a symbol for stressing whether significant differences are found at 95 % confidence degree (*) or at 90 % (+). Results of these tests agree with our previous remarks. The differences in the case of MDM are clear in favour of the imbalanced version. In the case of BB, the behaviour vary depending on the ensemble technique, where significant differences are obtained for SMOTE-Bagging whereas none are found for Under-Bagging. Finally, when we contrast these results versus the standard ensemble approach, we also observe a two-fold behaviour: in the case of SMOTE-Bagging the pruning approach enables the definition of a simpler ensemble with a low decrease of the performance, especially when the imbalanced metric is selected. On the other hand, for Under-Bagging we observe a notorious improvement of the results in all cases when the ordering-based pruning is applied,

**Table 2** Average test results for the standard ensemble approach (Base) and the ordering-based pruning with the original (BB and MDM) and imbalanced pruning metrics (BB-Imb and MDM-Imb)

| Ensemble | Base | BB | BB-Imb | MDM | MDM-Imb |
|---|---|---|---|---|---|
| SMOTE-Bagging | 0.8645 ± 0.0587 | 0.8602 ± 0.0632 | **0.8635** ± **0.0610** | 0.8596 ± 0.0629 | **0.8625** ± **0.0622** |
| Under-Bagging | 0.8647 ± 0.0516 | **0.8755** ± **0.0564** | 0.8734 ± 0.0544 | 0.8653 ± 0.0563 | **0.8699** ± **0.0558** |

**Table 3** Test results for the standard ensemble (Base) and ordering-based pruning schemes (BB, BB-Imb, MDM, and MDM-Imb) using AUC metric

| | SMOTE-Bagging | | | | | Under-Bagging | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Dataset | Base | BB | BB-Imb | MDM | MDM-Imb | Std. | BB | BB-Imb | MDM | MDM-Imb |
| Glass1 | 0.7675 | 0.8021 | 0.7925 | 0.7866 | 0.7900 | 0.7686 | 0.7979 | 0.7927 | 0.7918 | 0.7928 |
| Ecoli0vs1 | 0.9812 | 0.9750 | 0.9763 | 0.9802 | 0.9788 | 0.9805 | 0.9806 | 0.9764 | 0.9826 | 0.9809 |
| Wisconsin | 0.9707 | 0.9692 | 0.9700 | 0.9662 | 0.9666 | 0.9691 | 0.9698 | 0.9704 | 0.9678 | 0.9672 |
| Pima | 0.7568 | 0.7451 | 0.7546 | 0.7500 | 0.7558 | 0.7598 | 0.7561 | 0.7548 | 0.7532 | 0.7539 |
| Iris0 | 0.9880 | 0.9888 | 0.9880 | 0.9880 | 0.9880 | 0.9900 | 0.9900 | 0.9900 | 0.9900 | 0.9900 |
| Glass0 | 0.8347 | 0.8517 | 0.8464 | 0.8413 | 0.8430 | 0.8264 | 0.8469 | 0.8438 | 0.8399 | 0.8352 |
| Yeast1 | 0.7312 | 0.7192 | 0.7321 | 0.7301 | 0.7315 | 0.7304 | 0.7333 | 0.7310 | 0.7331 | 0.7307 |
| Vehicle2 | 0.9723 | 0.9752 | 0.9734 | 0.9686 | 0.9691 | 0.9704 | 0.9750 | 0.9744 | 0.9680 | 0.9686 |
| Vehicle1 | 0.7848 | 0.7691 | 0.7918 | 0.7898 | 0.7934 | 0.8016 | 0.8020 | 0.7983 | 0.7959 | 0.7985 |
| Vehicle3 | 0.7784 | 0.7593 | 0.7827 | 0.7795 | 0.7808 | 0.8060 | 0.7979 | 0.7976 | 0.7966 | 0.7974 |
| Haberman | 0.6627 | 0.6517 | 0.6476 | 0.6500 | 0.6498 | 0.6627 | 0.6616 | 0.6486 | 0.6488 | 0.6620 |
| Glass0123vs456 | 0.9405 | 0.9318 | 0.9357 | 0.9308 | 0.9378 | 0.9335 | 0.9432 | 0.9379 | 0.9264 | 0.9337 |
| Vehicle0 | 0.9635 | 0.9630 | 0.9636 | 0.9609 | 0.9614 | 0.9492 | 0.9558 | 0.9595 | 0.9539 | 0.9544 |
| Ecoli1 | 0.9053 | 0.8988 | 0.9067 | 0.9044 | 0.9107 | 0.8988 | 0.8981 | 0.9101 | 0.9043 | 0.9123 |
| Newthyroid2 | 0.9642 | 0.9540 | 0.9586 | 0.9567 | 0.9577 | 0.9605 | 0.9572 | 0.9696 | 0.9614 | 0.9692 |
| Newthyroid1 | 0.9558 | 0.9460 | 0.9486 | 0.9456 | 0.9467 | 0.9490 | 0.9479 | 0.9550 | 0.9594 | 0.9613 |
| Ecoli2 | 0.9145 | 0.9153 | 0.9128 | 0.9131 | 0.9099 | 0.9054 | 0.9057 | 0.8996 | 0.9017 | 0.8996 |
| Segment0 | 0.9917 | 0.9917 | 0.9924 | 0.9922 | 0.9926 | 0.9866 | 0.9881 | 0.9887 | 0.9872 | 0.9878 |
| Glass6 | 0.9291 | 0.9164 | 0.9213 | 0.9157 | 0.9203 | 0.9096 | 0.9277 | 0.9248 | 0.9228 | 0.9190 |
| Yeast3 | 0.9330 | 0.9308 | 0.9325 | 0.9315 | 0.9329 | 0.9311 | 0.9326 | 0.9305 | 0.9311 | 0.9295 |
| Ecoli3 | 0.8462 | 0.8508 | 0.8560 | 0.8506 | 0.8514 | 0.8830 | 0.8702 | 0.8670 | 0.8793 | 0.8707 |

(continued)

**Table 3** (continued)

| | SMOTE-Bagging | | | | | Under-Bagging | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Dataset | Base | BB | BB-Imb | MDM | MDM-Imb | Std. | BB | BB-Imb | MDM | MDM-Imb |
| Pageblocks0 | 0.9580 | 0.9552 | 0.9585 | 0.9572 | 0.9581 | 0.9610 | 0.9631 | 0.9626 | 0.9612 | 0.9615 |
| Ecoli034vs5 | 0.9129 | 0.9032 | 0.9018 | 0.9029 | 0.8948 | 0.8922 | 0.9148 | 0.9203 | 0.8701 | 0.9037 |
| Yeast2vs4 | 0.9277 | 0.9192 | 0.9155 | 0.9123 | 0.9223 | 0.9445 | 0.9408 | 0.9482 | 0.9383 | 0.9536 |
| Ecoli067vs35 | 0.8576 | 0.8651 | 0.8626 | 0.8653 | 0.8630 | 0.8582 | 0.8624 | 0.8578 | 0.8670 | 0.8523 |
| Ecoli0234vs5 | 0.9007 | 0.9008 | 0.9036 | 0.8935 | 0.8939 | 0.8641 | 0.9053 | 0.9027 | 0.8404 | 0.8784 |
| Glass015vs2 | 0.7041 | 0.7004 | 0.7015 | 0.7052 | 0.7025 | 0.7412 | 0.7117 | 0.7604 | 0.7553 | 0.7628 |
| Yeast0359vs78 | 0.7173 | 0.7023 | 0.7174 | 0.7016 | 0.7134 | 0.7373 | 0.7414 | 0.7386 | 0.7394 | 0.7387 |
| Yeast02579vs368 | 0.8028 | 0.7982 | 0.7995 | 0.7927 | 0.7993 | 0.8159 | 0.8090 | 0.8068 | 0.8136 | 0.8075 |
| Yeast0256vs3789 | 0.9183 | 0.9173 | 0.9176 | 0.9150 | 0.9185 | 0.9149 | 0.9136 | 0.9099 | 0.9140 | 0.9098 |
| Ecoli046vs5 | 0.9132 | 0.9086 | 0.9114 | 0.9046 | 0.9083 | 0.8869 | 0.9188 | 0.9238 | 0.8666 | 0.9123 |
| Ecoli01vs235 | 0.8988 | 0.8665 | 0.8815 | 0.8789 | 0.8883 | 0.8815 | 0.9031 | 0.9047 | 0.8893 | 0.8942 |
| Ecoli0267vs35 | 0.8617 | 0.8544 | 0.8611 | 0.8664 | 0.8642 | 0.8573 | 0.8623 | 0.8556 | 0.8662 | 0.8483 |
| Glass04vs5 | 0.9910 | 0.9836 | 0.9879 | 0.9876 | 0.9869 | 0.9940 | 0.9900 | 0.9940 | 0.9940 | 0.9940 |
| Ecoli0346vs5 | 0.8921 | 0.8888 | 0.8929 | 0.8762 | 0.8884 | 0.8799 | 0.8961 | 0.9051 | 0.8618 | 0.8956 |
| Ecoli0347vs56 | 0.8595 | 0.8701 | 0.8707 | 0.8590 | 0.8643 | 0.8762 | 0.8875 | 0.8897 | 0.9009 | 0.8800 |
| Yeast05679vs4 | 0.8177 | 0.8152 | 0.8133 | 0.8088 | 0.8124 | 0.8209 | 0.8287 | 0.8189 | 0.8018 | 0.8182 |
| Ecoli067vs5 | 0.8897 | 0.8894 | 0.8888 | 0.8909 | 0.8886 | 0.8820 | 0.8883 | 0.8888 | 0.9028 | 0.8779 |
| Vowel0 | 0.9878 | 0.9874 | 0.9880 | 0.9838 | 0.9853 | 0.9588 | 0.9671 | 0.9684 | 0.9689 | 0.9685 |
| Glass016vs2 | 0.7009 | 0.7083 | 0.7176 | 0.7168 | 0.7214 | 0.7025 | 0.7185 | 0.7291 | 0.7265 | 0.7323 |
| Glass2 | 0.7425 | 0.7390 | 0.7436 | 0.7458 | 0.7458 | 0.7569 | 0.7394 | 0.7691 | 0.7452 | 0.7702 |
| Ecoli0147vs2356 | 0.8685 | 0.8637 | 0.8719 | 0.8673 | 0.8793 | 0.8328 | 0.8625 | 0.8536 | 0.8665 | 0.8468 |
| Led7digit02456789vs1 | 0.8466 | 0.8547 | 0.8407 | 0.8500 | 0.8383 | 0.8268 | 0.8397 | 0.8322 | 0.8449 | 0.8399 |
| Ecoli01vs5 | 0.8881 | 0.8786 | 0.8782 | 0.8688 | 0.8755 | 0.8726 | 0.9142 | 0.9174 | 0.8795 | 0.8937 |

(continued)

**Table 3** (continued)

| | SMOTE-Bagging | | | | | Under-Bagging | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Dataset | Base | BB | BB-Imb | MDM | MDM-Imb | Std. | BB | BB-Imb | MDM | MDM-Imb |
| Glass06vs5 | 0.9926 | 0.9954 | 0.9954 | 0.9916 | 0.9912 | 0.9151 | 0.9910 | 0.9940 | 0.9940 | 0.9940 |
| Glass0146vs2 | 0.6961 | 0.7161 | 0.7295 | 0.7189 | 0.7254 | 0.7214 | 0.7335 | 0.7336 | 0.7323 | 0.7434 |
| Ecoli0147vs56 | 0.8703 | 0.8848 | 0.8804 | 0.8682 | 0.8750 | 0.8738 | 0.9035 | 0.8870 | 0.8819 | 0.8756 |
| Cleveland0vs4 | 0.7894 | 0.7933 | 0.8004 | 0.7815 | 0.7835 | 0.8492 | 0.8714 | 0.8305 | 0.7917 | 0.8069 |
| Ecoli0146vs5 | 0.8875 | 0.9037 | 0.9022 | 0.8828 | 0.8994 | 0.8933 | 0.9197 | 0.9273 | 0.8639 | 0.8988 |
| Ecoli4 | 0.9245 | 0.9220 | 0.9247 | 0.9094 | 0.9135 | 0.8952 | 0.9357 | 0.9349 | 0.9017 | 0.8969 |
| Shuttle0vs4 | 0.9999 | 0.9999 | 0.9999 | 0.9999 | 0.9999 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| Yeast1vs7 | 0.7458 | 0.7354 | 0.7349 | 0.7368 | 0.7303 | 0.7661 | 0.7869 | 0.7852 | 0.7463 | 0.7824 |
| Glass4 | 0.9069 | 0.8795 | 0.8788 | 0.8716 | 0.8675 | 0.9065 | 0.9182 | 0.8903 | 0.8943 | 0.8882 |
| Pageblocks13vs4 | 0.9952 | 0.9932 | 0.9964 | 0.9963 | 0.9963 | 0.9804 | 0.9937 | 0.9946 | 0.9928 | 0.9928 |
| Abalone9vs18 | 0.7120 | 0.7140 | 0.7076 | 0.7090 | 0.7085 | 0.7560 | 0.7490 | 0.7388 | 0.7222 | 0.7354 |
| Glass016vs5 | 0.9865 | 0.9493 | 0.9747 | 0.9675 | 0.9674 | 0.9429 | 0.9698 | 0.9675 | 0.9670 | 0.9663 |
| Shuttle2vs4 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| Yeast1458vs7 | 0.6330 | 0.6175 | 0.6144 | 0.6059 | 0.6153 | 0.6374 | 0.6530 | 0.6263 | 0.6009 | 0.6315 |
| Glass5 | 0.9769 | 0.9533 | 0.9619 | 0.9586 | 0.9626 | 0.9488 | 0.9596 | 0.9639 | 0.9631 | 0.9621 |
| Yeast2vs8 | 0.8064 | 0.7916 | 0.7946 | 0.8014 | 0.8068 | 0.7526 | 0.7846 | 0.7608 | 0.7579 | 0.7629 |
| Yeast4 | 0.8211 | 0.8117 | 0.8114 | 0.8046 | 0.8124 | 0.8420 | 0.8534 | 0.8537 | 0.8416 | 0.8543 |
| Yeast1289vs7 | 0.7046 | 0.6818 | 0.6905 | 0.6831 | 0.7004 | 0.7370 | 0.7194 | 0.7392 | 0.6918 | 0.7433 |
| Yeast5 | 0.9622 | 0.9536 | 0.9581 | 0.9525 | 0.9585 | 0.9593 | 0.9689 | 0.9673 | 0.9623 | 0.9625 |
| Yeast6 | 0.8375 | 0.8354 | 0.8446 | 0.8369 | 0.8431 | 0.8673 | 0.8736 | 0.8570 | 0.8706 | 0.8514 |
| Ecoli0137vs26 | 0.8347 | 0.8273 | 0.8336 | 0.8363 | 0.8400 | 0.7807 | 0.8774 | 0.7874 | 0.8060 | 0.7789 |
| Abalone19 | 0.5432 | 0.5380 | 0.5447 | 0.5375 | 0.5462 | 0.7121 | 0.7034 | 0.7251 | 0.7213 | 0.7307 |
| Average | 0.8645 | 0.8602 | 0.8635 | 0.8596 | 0.8625 | 0.8647 | 0.8755 | 0.8734 | 0.8653 | 0.8699 |

**Table 4** Wilcoxon test for pruning metrics: standard $[R^+]$ and imbalanced $[R^-]$

| Ensemble | Comparison | $R^+$ | $R^-$ | $p$-value |
|---|---|---|---|---|
| SMOTE-Bagging | BB versus BB-Imb | 540.0 | 1671.0 | 0.00028* |
| | MDM versus MDMimb | 436.0 | 1775.0 | 0.00002 |
| Under-Bagging | BB versus BB-Imb | 1277.0 | 934.0 | 0.27939 |
| | MDM versus MDMimb | 831.5 | 1379.5 | 0.07246+ |

**Table 5** Wilcoxon test to compare the standard ensemble approach (Std.) $[R^+]$ and the one with imbalanced ordering-based pruning $[R^-]$

| Ensemble | Comparison | $R^+$ | $R^-$ | $p$-value |
|---|---|---|---|---|
| SMOTE-bagging | Std. versus BB-Imb | 1261.5 | 883.5 | 0.215579 |
| | Std. versus MDMimb | 1386.5 | 758.5 | 0.039856* |
| Under-bagging | Std. versus BB-Imb | 502.0 | 1709.0 | 0.000114* |
| | Std. versus MDMimb | 637.0 | 1574.0 | 0.002735* |

showing a better behaviour for MDM-Imb and especially in BB-Imb (see Tables 2 and 3). These findings are complemented by means of a Wilcoxon test (shown in Table 5), for which we observe significant differences in favour of the ordering-based pruning for the Under-Bagging approach.

## 6 Concluding Remarks

Ordering-based pruning in ensembles of classifiers consists of carrying out a selection of those elements of the ensemble set that are expected to work with better synergy. The former process is guided by a given metric of performance which is focused on different capabilities of the ensemble. However, they have not been previously considered within been developed within the scenario of imbalanced datasets. In this work, we have proposed two adaptations of metrics for ordering-based pruning in imbalanced classification, namely BB-Imb and MDM-Imb. The experimental analysis has shown the success of these novel metrics with respect to their original definition, especially in the case of the SMOTE-Bagging approach. Additionally, we have pointed out that a significant improvement in the behaviour of the Under-Bagging ensemble is achieved by means of the application of the ordering-based pruning, outperforming the results with respect to the original model. As future work, we plan to include a larger number of pruning metrics and ensemble learning methodologies, aiming at giving additional support and strength to the findings obtained in this contribution.

**Acknowledgments** This work was supported by the Spanish Ministry of Science and Technology under projects TIN-2011-28488, TIN-2012-33856, TIN2013-40765-P; the Andalusian Research Plans P11-TIC-7765 and P10-TIC-6858; and both the University of Jaén and Caja Rural Provincial de Jaén under project UJA2014/06/15.

## References

1. Barandela, R., Valdovinos, R., Sánchez, J.: New applications of ensembles of classifiers. Pattern Anal. Appl. **6**(3), 245–256 (2003)
2. Barandela, R., Sánchez, J.S., García, V., Rangel, E.: Strategies for learning in class imbalance problems. Pattern Recognit. **36**(3), 849–851 (2003)
3. Batista, G.E.A.P.A., Prati, R.C., Monard, M.C.: A study of the behaviour of several methods for balancing machine learning training data. SIGKDD Explor. **6**(1), 20–29 (2004)
4. Breiman, L.: Bagging predictors. Mach. Learn. **24**(2), 123–140 (1996)
5. Demšar, J.: Statistical comparisons of classifiers over multiple data sets. J. Mach. Learn. Res. **7**, 1–30 (2006)
6. Freund, Y., Schapire, R.: A decision-theoretic generalization of on-line learning and an application to boosting. J. Comput. Syst. Sci. **55**(1), 119–139 (1997)
7. Galar, M., Fernández, A., Barrenechea, E., Bustince, H., Herrera, F.: A review on ensembles for class imbalance problem: bagging, boosting and hybrid based approaches. IEEE Trans. Syst., Man Cybern. Part C: Appl. Rev. **42**(4), 463–484 (2012)
8. Galar, M., Fernández, A., Barrenechea, E., Herrera, F.: Eusboost: enhancing ensembles for highly imbalanced data-sets by evolutionary undersampling. Pattern Recognit. **46**(12), 3460–3471 (2013)
9. He, H., Garcia, E.A.: Learning from imbalanced data. IEEE Trans. Knowl. Data Eng. **21**(9), 1263–1284 (2009)
10. Hernández-Lobato, D., Martínez-Muñoz, G., Suárez, A.: Statistical instance-based pruning in ensembles of independent classifiers. IEEE Trans. Pattern Anal. Mach. Intell. **31**(2), 364–369 (2009)
11. Huang, J., Ling, C.X.: Using AUC and accuracy in evaluating learning algorithms. IEEE Trans. Knowl. Data Eng. **17**(3), 299–310 (2005)
12. Kuncheva, L.I.: Diversity in multiple classifier systems. Inf. Fusion **6**(1), 3–4 (2005)
13. López, V., Fernández, A., García, S., Palade, V., Herrera, F.: An insight into classification with imbalanced data: empirical results and current trends on using data intrinsic characteristics. Inf. Sci. **250**(20), 113–141 (2013)
14. López, V., Fernández, A., Herrera, F.: On the importance of the validation technique for classification with imbalanced datasets: addressing covariate shift when data is skewed. Inf. Sci. **257**, 1–13 (2014)
15. Martínez-Muñoz, G., Suárez, A.: Using boosting to prune bagging ensembles. Pattern Recognit. Lett. **28**(1), 156–165 (2007)
16. Martínez-Muñoz, G., Hernández-Lobato, D., Suárez, A.: An analysis of ensemble pruning techniques based on ordered aggregation. IEEE Trans. Pattern Anal. Mach. Intell. **31**(2), 245–259 (2009)
17. Moreno-Torres, J.G., Sáez, J.A., Herrera, F.: Study on the impact of partition-induced dataset shift on k-fold cross-validation. IEEE Trans. Neural Netw. Learn. Syst. **23**(8), 1304–1313 (2012)

18. Polikar, R.: Ensemble based systems in decision making. IEEE Circuits Syst. Mag. **6**(3), 21–45 (2006)
19. Prati, R.C., Batista, G.E.A.P.A., Silva, D.F.: Class imbalance revisited: a new experimental setup to assess the performance of treatment methods. Knowledge and Information Systems, 1–25 (2014, in press)
20. Quinlan, J.: C4.5: Programs for Machine Learning. Morgan Kauffman, San Mateo (1993)
21. Rokach, L.: Ensemble-based classifiers. Artif. Intell. Rev. **33**(1), 1–39 (2010)
22. Sun, Y., Wong, A.K.C., Kamel, M.S.: Classification of imbalanced data: a review. Int. J. Pattern Recognit. Artif. Intell. **23**(4), 687–719 (2009)
23. Wang, S., Yao, X.: Diversity analysis on imbalanced data sets by using ensemble models. In: Proceedings of the 2009 IEEE Symposium on Computational Intelligence and Data Mining (CIDM'09), 324–331 (2009)
24. Wilcoxon, F.: Individual comparisons by ranking methods. Biom. Bull. **1**(6), 80–83 (1945)
25. Zadrozny, B., Langford, J., Abe, N.: Cost-sensitive learning by cost-proportionate example weighting. In: Proceedings of the 3rd IEEE International Conference on Data Mining (ICDM'03), 435–442 (2003)
26. Zhang, Y., Burer, S., Street, W.N.: Ensemble pruning via semi-definite programming. J. Mach. Learn. Res. **7**, 1315–1338 (2006)
27. Zhou, Z.H., Wu, J., Tang, W.: Ensembling neural networks: many could be better than all. Artif. Intell. **137**(1–2), 239–263 (2002)

# A Variant of the K-Means Clustering Algorithm for Continuous-Nominal Data

Aleksander Denisiuk and Michał Grabowski

**Abstract** The core idea of the proposed algorithm is to embed the considered dataset into a metric space. Two spaces for embedding of nominal part with the Hamming metric are considered: Euclidean space (the classical approach) and the standard unit sphere $\mathbb{S}$ (our new approach). We proved that the distortion of embedding into the unit sphere is at least 75 % better than that of the classical approach. In our model, combinations of continuous and nominal data are interpreted as points of a cylinder $\mathbb{R}^p \times \mathbb{S}$, where $p$ is the dimension of continuous data. We use a version of the gradient algorithm to compute centroids of finite sets on a cylinder. Experimental results show certain advances of the new algorithm. Specifically, it produces better clusters in tests with predefined groups.

## 1 Introduction

From the very beginning we define a dissimilarity function or a metric on combinations of continuous and nominal (categorical) data. There is a huge collection of dissimilarity functions on vectors of nominal data, used in data exploration. For example, the Hamming distance, the Jaccard distance, the distance defined after the Bayesian numerical codding of nominal values, and other concepts [6, 8]. In this paper we follow the approach with the Hamming distance. Let $(x, n)$ be a record of continuous $(x)$ and nominal $(n)$ data, where $x \in \mathbb{R}^p$. We define metric on the space of such records as $\operatorname{dist}\big((x_1, n_1), (x_2, n_2)\big) = K\big(d(x_1, x_2), H(n_1, n_2)\big)$, where

A. Denisiuk (✉)
University of Warmia and Mazury in Olsztyn, Olsztyn, Poland
e-mail: denisiuk@matman.uwm.edu.pl

M. Grabowski
Warsaw School of Computer Science, Warsaw, Poland
e-mail: mgrabowski@poczta.wwsi.edu.pl

R. Burduk et al. (eds.), *Proceedings of the 9th International Conference on Computer Recognition Systems CORES 2015*, Advances in Intelligent Systems and Computing 403, DOI 10.1007/978-3-319-26227-7_2