Addressing Covariate Shift for Genetic Fuzzy Systems Classifiers: A Case of Study with FARC-HD for Imbalanced Datasets

Victoria López
Dept. of Computer Science and A.I.
CITIC-UGR
University of Granada, Granada, Spain
Email: vlopez@decsai.ugr.es

Alberto Fernández
Dept. of Computer Science,
University of Jaén,
Jaén, Spain
Email: alberto.fernandez@ujaen.es

Francisco Herrera
Dept. of Computer Science and A.I.
CITIC-UGR
University of Granada, Granada, Spain
Email: herrera@decsai.ugr.es

***Abstract*—The estimation of the quality of the learned models in Data Mining has been traditionally carried out by means of a k-fold partition technique. However, the "random" division of the instances over the folds may results in a problem known as covariate shift, i.e. there is a different data distribution between the training and test folds.**

In classification with imbalanced datasets this problem is more severe. The misclassification of minority class instances due to an incorrect learning of the real boundaries caused by a not well defined data distribution, truly affects the measures of performance in this scenario. To avoid this harmful situation, we propose the use of a specific validation technique for the partitioning of the data, known as "Distribution optimally balanced stratified cross-validation". This methodology makes the decision of placing close-by samples on different folds, so that each partition will end up with enough representatives of every region.

In this contribution, we show the goodness of this methodology using Genetic Fuzzy Systems, as they are known to be robust approaches for all types of classification problems. Specifically, we have chosen the FARC-HD algorithm, a novel technique which has shown to obtain very accurate results. From the experimental analysis, which is carried out on a wide number of imbalanced datasets, we emphasize the necessity of using a proper validation methodology for extracting well founded conclusions.

***Index Terms*—Imbalanced Datasets, Covariate Shift, Dataset Shift, Validation Techniques, Partitioning, Genetic Fuzzy Systems**

I. Introduction

Standard learning algorithms are designed under the premise of a balanced distribution. When dealing with skewed class distributions, the classification problem becomes more difficult, specifically for correctly identifying the minority concepts within the data [1]. This issue is known as the class imbalance problem [2], [3], in which there is an under-represented class (positive) and a majority class (negative).

In order to validate the performance of a classifier, stratified cross-validation (SCV) is the most commonly employed method in the literature. It places an equal number of samples of each class on each partition to maintain class distributions equal in all partitions [4]. However, when this process is carried out in a random way, it may introduce a different data distribution between the training and test partitions, thus leading to inaccurate conclusions about the model that has been learnt from the training data. This issue is known as dataset shift [5], or more specifically covariate shift [6].

In the presence of imbalance, this problem is even more critic according to the metrics of performance. Misclassifications of the positive class instances in the test partition due to a "random clustering" of the classes, highly decreases the quality of the classifier, as their weight in the final accuracy is higher.

A more suitable validation technique needs to be employed in order to avoid inducing dataset shift issues artificially. In this paper, we suggest the use of a novel methodology called "Distribution optimally balanced SCV" (DOB-SCV) [7]. This method attempts to minimize covariate shift by keeping data distribution as similar as possible between training and test folds by maximizing diversity on each fold and trying to keep all folds as similar as possible to each other. The mechanism of this approach consists of selecting the k closest neighbors for a given instance and place them in different folds (with k being the number of total partitions), so that the data distribution between the training and test partitions remains as similar as possible.

Genetic Fuzzy Systems (GFSs) [8] have shown to be very effective techniques for classification problems in general, and for addressing imbalanced datasets in particular [9]. The significance of employing a proper validation methodology for the analysis of the results, takes a greater significance in this case, due to the stochastic character of this type of approaches. Regarding this fact, we aim to evaluate the robustness of the DOB-SCV strategy using as case of study the FARC-HD algorithm [10], a recent and accurate GFS classification technique.

Our experimental framework includes a set of sixty-six real-world problems from the KEEL dataset repository [11], [12] (http://www.keel.es/dataset.php). We measure the performance of the classifiers using the Area Under the Curve (AUC) metric [13] as suggested in imbalanced domains. Additionally, we study the significance of the results by the proper statistical tests as suggested in the literature [14], [15].

In order to do so, this contribution is arranged as follows.

First, Section II briefly introduces the problem of imbalanced data. Next, Section III contains the main concepts that are developed in this work, i.e. the basis on validation techniques and the problem of covariate/dataset shift. Then, the experimental framework is presented in Section IV, whereas all the analysis of the results is shown along Section V. Finally, Section VI summarizes and concludes the work.

II. Imbalanced datasets in Classification

In this section, we will first introduce the problem of imbalanced datasets, describing its features and why is so difficult to learn in this classification scenario. Then, we will present how to address this problem, enumerating diverse approaches that can be applied to ease the discrimination of the positive and negative classes. Next, we will discuss how to evaluate the performance of the results in this framework. Finally, we will describe the GFS used in our experimental study, the FARC-HD algorithm.

A. The problem of imbalanced datasets

The main property of this type of classification problem (in a binary context) is that the examples of one class outnumber the examples of the other one [1], [3]. The positive class is usually the most important concept to be learnt, since it might be associated with exceptional and significant cases [16] or because the data acquisition of these examples is costly [17]. Since most of the standard learning algorithms consider a balanced training set, this situation may cause the obtention of suboptimal classification models, i.e. a good coverage of the negative examples whereas the positive ones are misclassified more frequently [2], [3].

Traditionally, the imbalance ratio (IR) [18] is the main hint to identify a set of problems which need to be addressed in a special way. Additionally, other data intrinsic characteristics that are related to this concept may deteriorate even more the final performance of the models. Some of them/data intrinsic characteristics include the overlapping between classes [19], lack of representative data [20], small disjuncts [21], [22], dataset shift [23] and other issues which have interdependent effects with data distribution (imbalance).

For standard learning algorithms, obtaining a good separability between the positive and negative classes is not straightforward [3]. As they aim at obtaining the highest accuracy, this favors the covering of the negative class examples. Additionally, positive instances can be treated as noise and thus ignored by the classifier. These facts make imperative the use of specific techniques designed for addressing classification with imbalanced data.

B. Addressing the imbalanced problem: preprocessing and cost-sensitive learning

A large number of approaches have been previously proposed to deal with the class imbalance problem [24], which can be categorised in three groups:

1) Data level solutions: these are external approaches that modify the training set by oversampling the positive examples or undersampling the negative ones [25]–[27].
2) Algorithmic level solutions: in this case the inner learning procedure of the algorithm is modified in order to take into account the imbalance of the classes [28], [29].
3) Cost-sensitive solutions: these approaches aim at minimizing the cost errors, usually considering a higher cost for misclassifying the positive class examples [30], [31].

Among the aforementioned techniques, the application of data level solutions is independent of the classifier used, thus favoring the synergy with any classification algorithm. Specifically, previous analysis suggested that oversampling techniques work well in conjunction with fuzzy learning approaches [9], [27].

The simplest approach, random oversampling, makes exact copies of existing instances, and therefore several authors agree that this method can increase the likelihood of occurring overfitting [25]. According to the previous fact, more sophisticated methods have been proposed based on the generation of synthetic samples. Among them, the "Synthetic Minority Over-sampling TEchnique" (SMOTE) [26], has become one of the most significant approaches in this area.

The positive class is over-sampled by taking each minority class sample and introducing synthetic examples along the line segments joining any/all of the k minority class nearest neighbors. Depending upon the amount of over-sampling required, neighbors from the k nearest neighbors are randomly chosen. This process is illustrated in Figure 1, where x_i is the selected point, x_{i1} to x_{i4} are some selected nearest neighbors and r_1 to r_4 the synthetic data points created by the randomised interpolation.

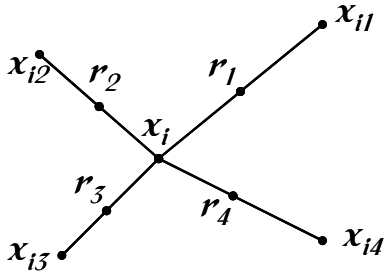


Fig. 1. An illustration of how to create the synthetic data points in the SMOTE algorithm

C. Evaluation in imbalanced domains

As stated in Section II-A, the use of the standard accuracy rate is no longer valid, since it does not provide information about the classification of both classes independently. Since in this classification scenario we intend to achieve good quality results for both classes, there is a necessity of obtaining one way to combine the individual measures of both the positive and negative classes, being none of these measures alone adequate by itself.

The AUC [32] metric (1) is a widely used evaluation criteria in imbalanced domains. It is computed over a Receiver Operating Characteristic (ROC) graphic [33], which visualizes

the trade-off between the benefits (TP_{rate}) and costs (FP_{rate}) (see Figure 2).

$$AUC = \frac{1 + TP_{rate} - FP_{rate}}{2} \quad (1)$$

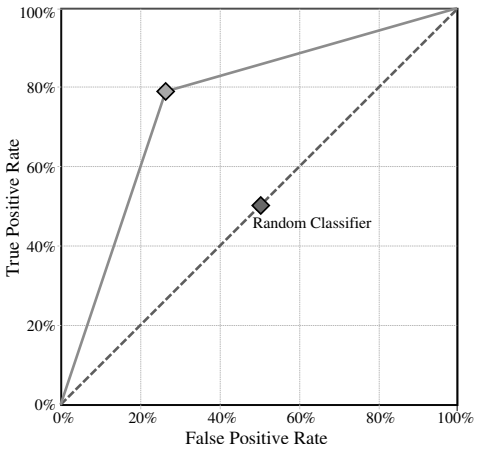


Fig. 2. Example of an ROC plot. Two classifiers' curves are depicted: the dashed line represents a random classifier, whereas the solid line is a classifier which is better than the random classifier.

D. Genetic Fuzzy Systems for Imbalanced Datasets: FARC-HD

GFSs are one of the most popular hybridizations among the Computational Intelligence areas. They are based on the combination between fuzzy logic and genetic algorithms. The final aim is to enhance the learning procedure of a fuzzy system by the application of evolutionary computation techniques [8], [34].

In this contribution, we will make use of a novel fuzzy associative classification method named FARC-HD [10]. This algorithm was shown to obtain very accurate results, and therefore it will serve as a very robust approach for validating our experimental results.

This algorithm extracts fuzzy association rules by limiting the order of the associations. The former constraint is used as a "preescrening" for high quality candidate rules during learning, which allows the achievement of more interpretable rules, i.e. a low number of rules with few antecedents. Finally, a genetic rule selection and lateral tuning procedure is applied for improving the classification accuracy of the final rule set.

Specifically, the FARC-HD model is composed of three stages, as shown in Figure 3:

1) Extracting the fuzzy association rules for classification by applying a search tree, whose depth of the branches is limited.
2) Preselecting the most interesting rules using subgroup discovery in order to decrease the computational cost of the system.
3) Optimizing the knowledge base by means of a combination between the well known tuning of the lateral position of the membership functions and a rule selection process.

III. Classifier Evaluation Techniques and the Issue of Dataset Shift

As stated in the introduction of this work, the estimation of the performance of a classifier, via partitioning in training and test folds, is a necessary procedure in order to validate the results for a given experiment. However, the conclusions extracted from the experimental analysis, actually depend on the specific procedure carried out for this task. We specifically refer to the issue of dataset shift, i.e. the space distribution of the instances in training and test may differ, thus leading to "overfitting".

In this section, we describe dataset shift in order to understand the nature of the problem we are dealing with. Next, we recall the standard and well-known SCV technique, and we identify its handicap for classification with imbalanced data. Finally, we present a recent methodology to alleviate this situation by a better organisation of the instances among the different folds.

A. Dataset shift

The problem of dataset shift [5] is defined as the case where training and test data follow different distributions. There are three potential types of dataset shift:

1) Prior Probability Shift: it refers to the differences in the class distribution between training and test partitions. In the most extreme case, the training set could not have instances for a given class. This handicap is prevented by applying a simple SCV scheme.
2) Covariate Shift: contrary to the previous case, now it is the inputs of the problem which differ in training and test sets. We focus on the impact of this type of shift for classification problems with imbalanced data.
3) Concept Drift: this problem occurs when the relationship between the input and class variables changes. This is the most difficult issue to overcome, and it is usually referred to as "Concept Drift".

The dataset shift issue is specially relevant when dealing with imbalanced classification because, in highly imbalanced domains, the positive class is particularly sensitive to singular classification errors, due to the typically low number of examples it presents [23]. In the most extreme cases, a single misclassified example of the positive class can create a significant drop in performance.

For clarity, Figures 4 and 5 present two examples of the influence of dataset shift in imbalanced classification. In the first case (Figure 4), it is easy to see a separation between classes in the training set that carries over perfectly to the test set. However, in the second case (Figure 5) it must be noted how some positive class examples in test are at the bottom and rightmost areas where there were not represented in the training set, leading to a gap between the training and test performance. These problems are represented in a two-dimensional space by means of a linear transformation of the inputs variables following the technique given in [23].

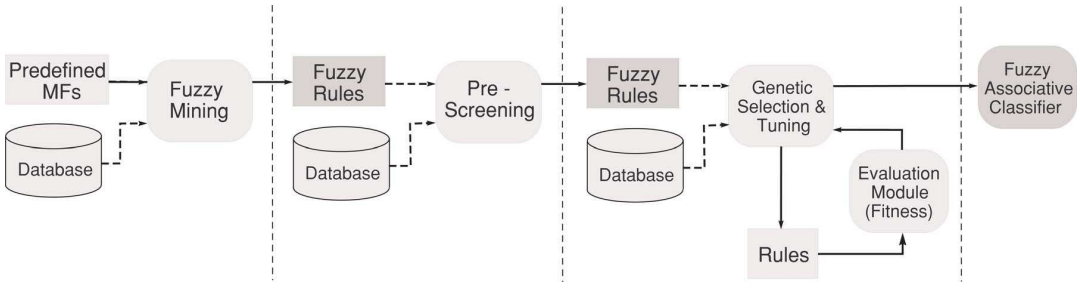
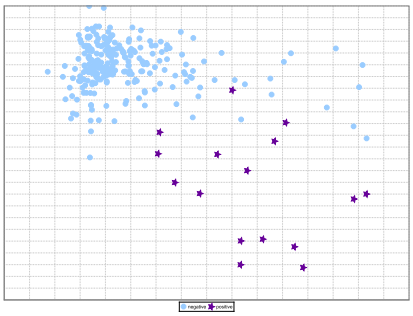
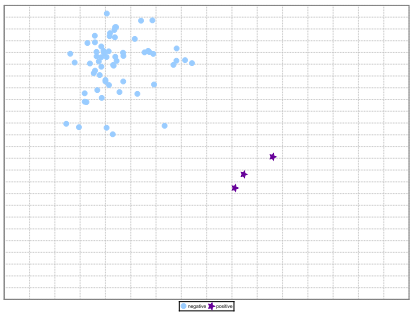


Fig. 3. Scheme of the FARC-HD method.

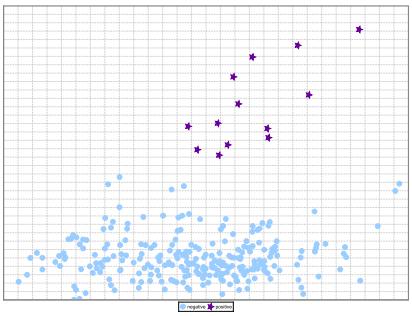


(a) Training data. AUC = .9043

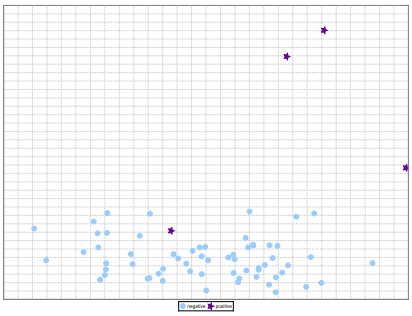


(b) Test data. AUC = 1.000

Fig. 4. Example of good behavior (no dataset shift) in imbalanced domains: ecoli4 dataset, 5th partition



(a) Training data. AUC = 1.000



(b) Test data. AUC = .8750

Fig. 5. Example of bad behavior caused by dataset shift in imbalanced domains: ecoli4 dataset, 1st partition

B. Cross-validation for classifier evaluation: Distribution optimally balanced SCV

When aiming to analyze the generalization ability of a classifier, a cross-validation technique must be employed. This methodology divides a given dataset into two different subsets with null intersection: a training set for learning the model, and a test set for checking the output performance.

Cross-validation is often carried out into k-folds, i.e. the original dataset is randomly partitioned into k subsamples. From these new sets, one of them is used for test and the remaining $k-1$ sets are joined and used for training data. This procedure is iterated k times, such as all k partitions will be used for test. Finally, the output results for the k folds must be averaged in order to give a single performance estimation.

How the instances of the dataset are placed into each fold has a severe impact in the final performance estimation for the validation stage. Traditionally, researchers in data mining have used a simple SCV procedure, which distributes the instances among folds regarding the class distribution. In this way, each fold is intended to have the same number of examples per class, thus avoiding *prior probability shift*.

However, it might induce *covariate shift* since it does not take into account the distribution of the variables of the problem. According to this fact, we consider a more sophisticated technique, known as DOB-SCV [7], which aims at preventing both the *prior probability* and *covariate shift* issues. The idea behind this procedure is quite simple: assigning close-by examples to different folds, so that representative examples for the different regions of the problem will be represented among them.

The pseudo-code for the DOB-SCV technique is depicted in Algorithm 1. For each class of the problem, it picks an unassigned example. Next, it finds its $k-1$ nearest unassigned neighbors of the same class, and then it places all of them to

a different fold. This process is iterated for each class.

Algorithm 1 DOB-SCV Partitioning Method

for each class $c_j \in C$ **do**
  while count(c_j) > 0 **do**
    $e_0 \leftarrow$ randomly select an example of class c_j from D
    $e_i \leftarrow i$th closest example to e_0 of class c_j from D $(i = 1, \ldots, k-1)$
    $F_i \leftarrow F_i \bigcup e_i (i = 0, \ldots, k-1)$
    $D \leftarrow D \setminus e_i (i = 0, \ldots, k-1)$
  end while
end for

IV. Experimental Framework

In this section we first provide details of the real-world binary-class imbalanced problems chosen for the experiments (subsection IV-A). Then, we will give the configuration parameters for the methods employed in the experimental study (subsection IV-B). Finally, we present the statistical tests applied to compare the obtained results (subsection IV-C).

A. Benchmark data

Table I shows the selected imbalanced datasets for our experimental study where we can observe, by columns, the name of the dataset, its size, number of attributes, which class(es) are considered as negative and positive ones, their percentage and the IR. According to our previous work on the topic [27] we have set a threshold for considering a dataset to be imbalanced when the ratio between the negative and positive instances is higher than 1.5, i.e. a class 60:40 distribution.

As pointed out along this paper, the estimates of the AUC measure are obtained by means of a standard SCV and the DOB-SCV. The number of folds selected in both cases is 5. This value is set up with the aim of having enough positive class instances in the different folds, hence avoiding additional problems in the data distribution, especially for highly imbalanced datasets.

Furthermore, results of this contribution can be reproduced by downloading the original dataset partitions with SCV at the KEEL dataset repository [12].

B. Parameters

In the case of the FARC-HD classifier, we have used the values suggested by the authors in [10]:

- Fuzzy Rule Based Classification System parameters:
  - Conjunction operator: product t-norm.
  - Rule weight: certainty factor.
  - Fuzzy reasoning method: additive combination [35].
  - Number of linguistic labels per variable: 5 labels
- Inner learning parameters:
  - Minimum Support = 0.05,
  - Minimum Confidence = 0.8,
  - Depth of the trees ($Depthmax$) = 3,
  - Parameter K of the prescreening = 2.
- Genetic tuning with rule selection process:
  - Maximum number of evaluations = 20000,
  - Population size = 50,
  - Parameter alpha = 0.02,
  - Bits per gen = 30

Lastly, in the case of the SMOTE preprocessing technique, we will consider the *5-nearest neighbors of the positive class* to generate the synthetic samples, and *balancing both classes to the 50% distribution*.

TABLE I
Summary description of the imbalanced datasets used in the experimental study.

Datasets	#Ex.	#Atts.	Class (-;+)	%Class(-, +)	IR
Glass1	214	9	(build-win-non_float-proc; remainder)	(35.51, 64.49)	1.82
Ecoli0vs1	220	7	(im; cp)	(35.00, 65.00)	1.86
Wisconsin	683	9	(malignant; benign)	(35.00, 65.00)	1.86
Pima	768	8	(tested-positive; tested-negative)	(34.84, 66.16)	1.90
Iris0	150	4	(Iris-Setosa; remainder)	(33.33, 66.67)	2.00
Glass0	214	9	(build-win-float-proc; remainder)	(32.71, 67.29)	2.06
Yeast1	1484	8	(nuc; remainder)	(28.91, 71.09)	2.46
Vehicle1	846	18	(Saab; remainder)	(28.37, 71.63)	2.52
Vehicle2	846	18	(Bus; remainder)	(28.37, 71.63)	2.52
Vehicle3	846	18	(Opel; remainder)	(28.37, 71.63)	2.52
Haberman	306	3	(Die; Survive)	(27.42, 73.58)	2.68
Glass0123vs456	214	9	(non-window glass; remainder)	(23.83, 76.17)	3.19
Vehicle0	846	18	(Van; remainder)	(23.64, 76.36)	3.23
Ecoli1	336	7	(im; remainder)	(22.92, 77.08)	3.36
New-thyroid2	215	5	(hypo; remainder)	(16.89, 83.11)	4.92
New-thyroid1	215	5	(hyper; remainder)	(16.28, 83.72)	5.14
Ecoli2	336	7	(pp; remainder)	(15.48, 84.52)	5.46
Segment0	2308	19	(brickface; remainder)	(14.26, 85.74)	6.01
Glass6	214	9	(headlamps; remainder)	(13.55, 86.45)	6.38
Yeast3	1484	8	(me3; remainder)	(10.98, 89.02)	8.11
Ecoli3	336	7	(imU; remainder)	(10.88, 89.12)	8.19
Page-blocks0	5472	10	(remainder; text)	(10.23, 89.77)	8.77
Ecoli034vs5	200	7	(p,imL,imU; om)	(10.00, 90.00)	9.00
Yeast2vs4	514	8	(cyt; me2)	(9.92, 90.08)	9.08
Ecoli067vs35	222	7	(cp,omL,pp; imL,om)	(9.91, 90.09)	9.09
Ecoli0234vs5	202	7	(cp,imS,imL,imU; om)	(9.90, 90.10)	9.10
Glass015vs2	172	9	(build-win-non_float-proc, tableware, build-win-float-proc; ve-win-float-proc)	(9.88, 90.12)	9.12
Yeast0359vs78	506	8	(mit,me1,me3,erl; vac,pox)	(9.88, 90.12)	9.12
Yeast02579vs368	1004	8	(mit,cyt,me3,vac,erl; me1,exc,pox)	(9.86, 90.14)	9.14
Yeast0256vs3789	1004	8	(mit,cyt,me3,exc; me1,vac,pox,erl)	(9.86, 90.14)	9.14
Ecoli046vs5	203	6	(cp,imU,omL; om)	(9.85, 90.15)	9.15
Ecoli01vs235	244	7	(cp,im; imS,imL,om)	(9.83, 90.17)	9.17
Ecoli0267vs35	224	7	(cp,imS,omL,pp; imL,om)	(9.82, 90.18)	9.18
Glass04vs5	92	9	(build-win-float-proc,containers; tableware)	(9.78, 90.22)	9.22
Ecoli0346vs5	205	7	(cp,imL,imU,omL; om)	(9.76, 90.24)	9.25
Ecoli0347vs56	257	7	(cp,imL,imU,pp; om,omL)	(9.73, 90.27)	9.28
Yeast05679vs4	528	8	(me2; mit,me3,exc,vac,erl)	(9.66, 90.34)	9.35
Ecoli067vs5	220	6	(cp,omL,pp; om)	(9.09, 90.91)	10.00
Vowel0	988	13	(hid; remainder)	(9.01, 90.99)	10.10
Glass016vs2	192	9	(ve-win-float-proc; build-win-float-proc, build-win-non_float-proc, headlamps)	(8.89, 91.11)	10.29
Glass2	214	9	(Ve-win-float-proc; remainder)	(8.78, 91.22)	10.39
Ecoli0147vs2356	336	7	(cp,im,imU,pp; imS,imL,om,omL)	(8.63, 91.37)	10.59
Led7digit02456789vs1	443	7	(0,2,4,5,6,7,8,9; 1)	(8.35, 91.65)	10.97
Glass06vs5	108	9	(build-win-float-proc,headlamps; tableware)	(8.33, 91.67)	11.00
Ecoli01vs5	240	6	(cp,im; om)	(8.33, 91.67)	11.00
Glass0146vs2	205	9	(build-win-float-proc,containers, headlamps, build-win-non_float-proc; ve-win-float-proc)	(8.29, 91.71)	11.06
Ecoli0147vs56	332	6	(cp,im,imU,pp; om,omL)	(7.53, 92.47)	12.28
Cleveland0vs4	177	13	(0; 4)	(7.34, 92.66)	12.62
Ecoli0146vs5	280	6	(cp,im,imU,omL; om)	(7.14, 92.86)	13.00
Ecoli4	336	7	(om; remainder)	(6.74, 93.26)	13.84
Yeast1vs7	459	8	(nuc; vac)	(6.72, 93.28)	13.87
Shuttle0vs4	1829	9	(Rad Flow; Bypass)	(6.72, 93.28)	13.87
Glass4	214	9	(containers; remainder)	(6.07, 93.93)	15.47
Page-blocks13vs2	472	10	(graphic; horiz.line,picture)	(5.93, 94.07)	15.85
Abalone9vs18	731	8	(18; 9)	(5.65, 94.25)	16.68
Glass016vs5	184	9	(tableware; build-win-float-proc, build-win-non_float-proc, headlamps)	(4.89, 95.11)	19.44
Shuttle2vs4	129	9	(Fpv Open; Bypass)	(4.65, 95.35)	20.5
Yeast1458vs7	693	8	(vac; nuc,me2,me3,pox)	(4.33, 95.67)	22.10
Glass5	214	9	(tableware; remainder)	(4.20, 95.80)	22.81
Yeast2vs8	482	8	(pox; cyt)	(4.15, 95.85)	23.10
Yeast4	1484	8	(me2; remainder)	(3.43, 96.57)	28.41
Yeast1289vs7	947	8	(vac; nuc,cyt,pox,erl)	(3.17, 96.83)	30.56
Yeast5	1484	8	(me1; remainder)	(2.96, 97.04)	32.78
Ecoli0137vs26	281	7	(pp,imL; cp,im,imU,imS)	(2.49, 97.51)	39.15
Yeast6	1484	8	(exc; remainder)	(2.49, 97.51)	39.15
Abalone19	4174	8	(19; remainder)	(0.77, 99.23)	128.87

C. Statistical tests for performance comparison

When developing any experimental study, it is strongly recommended to contrast the conclusions extracted from the results by means of statistical tests [14], [15]. These tests provide the support necessary for gaining credibility in the experimental analysis. However, the initial conditions that guarantee the reliability of standard parametric tests (such as the t-test) cannot always be fulfilled, leading to the use of non-parametric tests instead.

In this contribution, we will apply pairwise comparisons by means of a Wilcoxon signed-rank test [36], as the non-parametric statistical procedure analogous to the standard t-test. This test works by taking the differences in performance between two classifiers and then ranking them according to their absolute value, from the lowest to the highest one. Then, R^+ will be the sum of ranks for the datasets in which the first algorithm outperformed the second one, and R^- refers to the contrary case. Then, the p-value for the statistical distribution is computed and if it is below a specified level of significance α the null hypothesis of equality of means can be rejected.

Any interested reader can find additional information about the use of this and additional tests on the Website http://sci2s.ugr.es/sicidm/.

V. Experimental Analysis

This section is devoted to identify the possible differences regarding the estimation of the performance with the standard SCV and the suggested DOB-SCV for imbalanced datasets.

With this aim, Table II shows the average classification values obtained by FARC-HD. In this table, three values are given by rows: first the average AUC performance obtained in the test partitions for the SCV technique, then the average performance for DOB-SCV, and finally the relative difference (in percentage) between both values, i.e. $\frac{AUC_{DOB-SCV}-AUC_{SCV}}{AUC_{SCV}}$. This final value has the following mean: if the value is positive, then the estimation of the performance for DOB-SCV is more optimistic than SCV; if the value is negative it refers to the contrary case; and the higher the obtained number, the most significative the selection of the validation approach is. Additionally, we show the detailed test results for all datasets in Table III.

TABLE II
Average test results with AUC metric and percentage differences for the SCV and DOB-SCV techniques with FARC-HD.

	IR < 9		IR > 9		All	
Algorithm	AUC_{Tr}	AUC_{Tst}	AUC_{Tr}	AUC_{Tst}	AUC_{Tr}	AUC_{Tst}
SCV	0.9270	0.8674	0.9479	0.8400	0.9409	0.8491
DOB-SCV	0.9255	0.8739	0.9473	0.8471	0.9400	0.8560
%Diff	-0.1447	0.7617	-0.0511	0.8229	-0.0823	0.8026

From these tables of results we may observe that for FARC-HD, the DOB-SCV validation technique achieves a higher estimation of the performance for most datasets, therefore being more robust for analysing the quality of the models learned in imbalanced data.

TABLE III
Detailed test results with AUC metric and percentage differences for the SCV and DOB-SCV techniques with FARC-HD

		SCV		DOB-SCV		% Diff.	
Dataset	IR	AUC_{Tr}	AUC_{Tst}	AUC_{Tr}	AUC_{Tst}	Train	Test
Glass1	1.82	.8869	.7424	.8688	.7411	-.0208	-.0018
Ecoli0vs1	1.86	.9903	.9663	.9903	.9649	.0000	-.0015
Wisconsin	1.86	.9862	.9640	.9852	.9693	-.0009	.0055
Pima	1.90	.8133	.7538	.8227	.7307	.0115	-.0316
Iris0	2.00	1.000	1.000	1.000	1.000	.0000	.0000
Glass0	2.06	.9183	.7706	.9095	.7851	-.0097	.0185
Yeast1	2.46	.7677	.7173	.7717	.7247	.0052	.0101
Vehicle1	2.52	.8424	.7521	.8250	.7351	-.0212	-.0230
Vehicle2	2.52	.9887	.9560	.9875	.9700	-.0013	.0144
Vehicle3	2.52	.8322	.7520	.8059	.7210	-.0326	-.0430
Haberman	2.68	.7328	.5914	.7682	.6501	.0461	.0903
Glass0123vs456	3.19	.9741	.9141	.9795	.9434	.0055	.0310
Vehicle0	3.23	.9613	.9266	.9569	.9333	-.0046	.0072
Ecoli1	3.36	.9470	.8836	.9494	.8936	.0025	.0113
New-thyroid2	4.92	.9986	.9460	.9972	.9575	-.0014	.0120
New-thyroid1	5.14	.9986	.9917	.9903	.9750	-.0084	-.0171
Ecoli2	5.46	.9632	.8947	.9651	.9019	.0019	.0080
Segment0	6.01	.9970	.9939	.9966	.9942	-.0004	.0003
Glass6	6.38	.9716	.9219	.9716	.9338	.0000	.0127
Yeast3	8.11	.9465	.9182	.9440	.9183	-.0026	.0001
Ecoli3	8.19	.9508	.8216	.9536	.8661	.0029	.0513
Page-blocks0	8.77	.9257	.9048	.9224	.9166	-.0036	.0128
Ecoli034vs5	9.00	.9924	.9167	.9972	.9278	.0049	.0120
Yeast2vs4	9.08	.9481	.8954	.9578	.9266	.0102	.0336
Ecoli067vs35	9.09	.9788	.8425	.9750	.8600	-.0038	.0203
Ecoli0234vs5	9.10	.9938	.9085	.9979	.9085	.0041	.0000
Glass015vs2	9.12	.9516	.7462	.9395	.7411	-.0129	-.0069
Yeast0359vs78	9.12	.8609	.7512	.8626	.7490	.0020	-.0029
Yeast02579vs368	9.14	.9272	.8912	.9243	.8918	-.0032	.0007
Yeast0256vs3789	9.14	.8366	.8015	.8330	.8020	-.0044	.0006
Ecoli046vs5	9.15	.9925	.8673	.9911	.9004	-.0014	.0368
Ecoli01vs235	9.17	.9754	.8718	.9794	.8836	.0041	.0134
Ecoli0267vs35	9.18	.9753	.8729	.9672	.8730	-.0083	.0001
Glass04vs5	9.22	.9970	.9761	.9879	.9143	-.0092	-.0676
Ecoli0346vs5	9.25	.9959	.9257	.9919	.9507	-.0041	.0263
Ecoli0347vs56	9.28	.9855	.9054	.9849	.9099	-.0006	.0049
Yeast05679vs4	9.35	.8799	.7651	.9014	.8078	.0239	.0529
Ecoli067vs5	10.00	.9813	.9050	.9737	.8575	-.0077	-.0554
Vowel0	10.10	1.000	.9656	.9989	.9844	-.0011	.0192
Glass016vs2	10.29	.8895	.6100	.8971	.5936	.0086	-.0277
Glass2	10.39	.9071	.5742	.9205	.7313	.0146	.2148
Ecoli0147vs2356	10.59	.9681	.8577	.9630	.8662	-.0054	.0097
Led7digit02456789vs1	10.97	.9284	.8971	.9204	.8261	-.0087	-.0860
Glass06vs5	11.00	1.000	.9800	.9962	.9550	-.0038	-.0262
Ecoli01vs5	11.00	.9938	.8886	.9955	.9045	.0017	.0176
Glass0146vs2	11.06	.9109	.6974	.9164	.6770	.0061	-.0301
Ecoli0147vs56	12.28	.9862	.8922	.9874	.9123	.0012	.0220
Cleveland0vs4	12.62	.9906	.8392	.9977	.8469	.0070	.0091
Ecoli0146vs5	13.00	.9913	.9192	.9923	.9269	.0010	.0083
Ecoli4	13.84	.9933	.9076	.9854	.8513	-.0080	-.0662
Yeast1vs7	13.87	.8896	.6866	.8918	.7118	.0024	.0354
Shuttle0vs4	13.87	1.000	.9997	1.000	.9994	.0000	-.0003
Glass4	15.47	.9944	.8658	.9844	.9036	-.0101	.0418
Page-blocks13vs4	15.85	.9975	.9555	.9820	.9363	-.0158	-.0205
Abalone9-18	16.68	.8660	.7896	.8468	.7665	-.0227	-.0302
Glass016vs5	19.44	.9864	.8186	.9857	.8514	-.0007	.0386
Shuttle2vs4	20.50	1.000	.9960	.9980	.9960	-.0020	.0000
Yeast1458vs7	22.10	.7988	.6506	.8194	.6452	.0252	-.0084
Glass5	22.81	.9957	.7232	.9915	.8780	-.0043	.1764
Yeast2vs8	23.10	.8682	.8153	.8868	.7947	.0209	-.0258
Yeast4	28.41	.9088	.8390	.9033	.7793	-.0061	-.0766
Yeast1289vs7	30.56	.8234	.6734	.8211	.6705	-.0029	-.0043
Yeast5	32.78	.9836	.9462	.9778	.9406	-.0059	-.0059
Ecoli0137vs26	39.15	.9831	.8136	.9868	.8262	.0037	.0153
Yeast6	39.15	.9169	.8316	.9156	.8812	-.0015	.0562
Abalone19	128.87	.8628	.6846	.8548	.7110	-.0094	.0370
Average		.9409	.8491	.9400	.8560	-.0823	.8026

We can also stress that the degree of imbalance of the dataset has some influence in the obtained results, i.e. the higher the IR is, the greater the differences between the DOB-SCV and the standard SCV. This issue can be due to the fact that the lower the number of positive instances we have in a dataset with respect to the negative ones, the more significative is to maintain the data distribution to avoid the gap in performance between training and test.

In order to give statistical support to the findings previously extracted, in Table IV we carry out a Wilcoxon test to compare both validation techniques with FARC-HD. From this test, we may conclude the more optimistic estimation of DOB-SCV, as

stated by the higher sum of ranks and the low p-value obtained, which tell us about the goodness of this approach.

TABLE IV
WILCOXON TEST TO COMPARE THE RESULTS WITH THE DOB-SCV VERSUS THE STANDARD SCV. R^+ CORRESPONDS TO THE SUM OF THE RANKS FOR THE DOB-SCV PARTITIONING APPROACH AND R^- TO THE ORIGINAL SCV PARTITIONING

Comparison	R^+	R^-	Hypothesis	p-value
FARC-HD[DOB-SCV] vs FARC-HD[SCV]	1337.5	807.5	Rejected for FARC-HD[DOB-SCV]	0.0827

To summarize, we must stress that DOB-SCV is a suitable methodology for contrasting the performance of the classification algorithms in imbalanced data. When the distribution of the classes is skewed, using standard estimation models may lead to misleading conclusions on the quality of the prediction. The proposed use of this model addresses the handicap of losing the generalization ability because of the way data is distributed among the different folds.

VI. CONCLUDING REMARKS

In this contribution, we have raised up the problem of covariate shift in classification with imbalanced data. This problem is referred to those situation in which the instances in the training and test partitions follow a different distribution in the data space, thus creating a handicap in the learning and evaluation of those techniques applied in this scenario.

Specifically, taking into account the class distribution and the performance metrics for this type of problem, we have suggested the use of a novel partition-based methodology, named as DOB-SCV, in order to overcome this situation. This technique aims at carrying out an heterogeneous organization of the instances of the classes among the different folds. This validation technique turns up to be a suitable procedure in the framework of imbalanced datasets.

The stable performance estimation of DOB-SCV has been contrasted versus the classical k-fold SCV, detecting significant differences between both techniques for the FARC-HD GFS classifier. The significance of using DOB-SCV in this case of study has a two-fold view: on the one hand, GFSs have a stochastic character, which makes the use of robust validation techniques mandatory for extracting well founded conclusions. On the other hand, avoiding different data distribution inside each fold will allow researchers on imbalanced data to concentrate their efforts on designing new learning models based only on the skewed data, rather than seeking for complex solutions when trying to overcome the gaps between training and test results.

ACKNOWLEDGMENT

This work was partially supported by the Spanish Ministry of Science and Technology under project TIN2011-28488 and the Andalusian Research Plans P11-TIC-7765 and P10-TIC-6858. V. López holds a FPU scholarship from Spanish Ministry of Education.

REFERENCES

[1] N. V. Chawla, N. Japkowicz, and A. Kotcz, "Editorial: special issue on learning from imbalanced data sets," *SIGKDD Explorations*, vol. 6, no. 1, pp. 1–6, 2004.

[2] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, 2009.

[3] Y. Sun, A. K. C. Wong, and M. S. Kamel, "Classification of imbalanced data: A review," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 23, no. 4, pp. 687–719, 2009.

[4] J. R. Cano, F. Herrera, and M. Lozano, "Evolutionary stratified training set selection for extracting classification rules with trade off precision-interpretability," *Data and Knowledge Engeneering*, vol. 60, pp. 90–108, 2007.

[5] J. Q. Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence, *Dataset Shift in Machine Learning*. Cambridge, MA: MIT Press, 2009.

[6] J. G. Moreno-Torres, T. Raeder, R. Aláiz-Rodríguez, N. V. Chawla, and F. Herrera, "A unifying view on dataset shift in classification," *Pattern Recognition*, vol. 45, no. 1, pp. 521–530, 2012.

[7] J. G. Moreno-Torres, J. A. Sáez, and F. Herrera, "Study on the impact of partition-induced dataset shift on k-fold cross-validation," *IEEE Transactions On Neural Networks And Learning Systems*, vol. 23, no. 8, pp. 1304–1313, 2012.

[8] F. Herrera, "Genetic fuzzy systems: Taxonomy, current research trends and prospects," *Evolutionary Intelligence*, vol. 1, pp. 27–46, 2008.

[9] A. Fernández, M. J. del Jesus, and F. Herrera, "On the 2–tuples based genetic tuning performance for fuzzy rule based classification systems in imbalanced data–sets," *Information Sciences*, vol. 180, no. 8, pp. 1268–1291, 2010.

[10] J. Alcalá-Fdez, R. Alcalá, and F. Herrera, "A fuzzy association rule-based classification model for high-dimensional problems with genetic rule selection and lateral tuning." *IEEE Transactions on Fuzzy Systems*, vol. 19, no. 5, pp. 857–872, 2011.

[11] J. Alcalá;-Fdez, L. Sánchez, S. García, M. J. del Jesus, S. Ventura, J. M. Garrell, J. Otero, C. Romero, J. Bacardit, V. M. Rivas, J. C. Fernández, and F. Herrera, "KEEL: a software tool to assess evolutionary algorithms for data mining problems," *Soft Computing*, vol. 13, pp. 307–318, 2009.

[12] J. Alcalá-Fdez, A. Fernández, J. Luengo, J. Derrac, S. García, L. Sánchez, and F. Herrera, "KEEL data–mining software tool: Data set repository, integration of algorithms and experimental analysis framework," *Journal of Multi–Valued Logic and Soft Computing*, vol. 17, no. 2-3, pp. 255–287, 2011.

[13] Y.-M. Huang, C.-M. Hung, and H. C. Jiau, "Evaluation of neural networks and data mining methods on a credit assessment task for class imbalance problem," *Nonlinear Analysis: Real World Applications*, vol. 7, no. 4, pp. 720 – 747, 2006.

[14] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine Learning Research*, vol. 7, pp. 1–30, 2006.

[15] S. García and F. Herrera, "An extension on "statistical comparisons of classifiers over multiple data sets" for all pairwise comparisons," *Journal of Machine Learning Research*, vol. 9, pp. 2607–2624, 2008.

[16] G. M. Weiss, "Mining with rarity: a unifying framework," *SIGKDD Explorations*, vol. 6, no. 1, pp. 7–19, 2004.

[17] G. M. Weiss and Y. Tian, "Maximizing classifier utility when there are data acquisition and modeling costs," *Data Mining and Knowledge Discovery*, vol. 17, no. 2, pp. 253–282, 2008.

[18] A. Orriols-Puig and E. Bernadó-Mansilla, "Evolutionary rule–based systems for imbalanced datasets," *Soft Computing*, vol. 13, no. 3, pp. 213–225, 2009.

[19] N. Japkowicz and S. Stephen, "The class imbalance problem: a systematic study," *Intelligent Data Analysis Journal*, vol. 6, no. 5, pp. 429–450, 2002.

[20] M. Wasikowski and X.-W. Chen, "Combating the small sample class imbalance problem using feature selection," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1388–1400, 2010.

[21] A. Orriols-Puig, E. Bernadó-Mansilla, D. E. Goldberg, K. Sastry, and P. L. Lanzi, "Facetwise analysis of XCS for problems with class imbalances," *IEEE Transactions on Evolutionary Computation*, vol. 13, pp. 260–283, 2009.

[22] G. M. Weiss and F. J. Provost, "Learning when training data are costly: The effect of class distribution on tree induction," *Journal of Artificial Intelligence Research*, vol. 19, pp. 315–354, 2003.

[23] J. G. Moreno-Torres and F. Herrera, "A preliminary study on overlapping and data fracture in imbalanced domains by means of genetic programming-based feature extraction," in *Proceedings of the 10th International Conference on Intelligent Systems Design and Applications (ISDA'10)*, 2010, pp. 501–506.
[24] V. López, A. Fernández, J. G. Moreno-Torres, and F. Herrera, "Analysis of preprocessing vs. cost-sensitive learning for imbalanced classification. open problems on intrinsic data characteristics," *Expert Systems with Applications*, vol. 39, no. 7, pp. 6585–6608, 2011.
[25] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard, "A study of the behaviour of several methods for balancing machine learning training data," *SIGKDD Explorations*, vol. 6, no. 1, pp. 20–29, 2004.
[26] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over–sampling technique," *Journal of Artificial Intelligent Research*, vol. 16, pp. 321–357, 2002.
[27] A. Fernández, S. García, M. J. del Jesus, and F. Herrera, "A study of the behaviour of linguistic fuzzy rule based classification systems in the framework of imbalanced data–sets," *Fuzzy Sets and Systems*, vol. 159, no. 18, pp. 2378–2398, 2008.
[28] R. Barandela, J. S. Sánchez, V. García, and E. Rangel, "Strategies for learning in class imbalance problems," *Pattern Recognition*, vol. 36, no. 3, pp. 849–851, 2003.
[29] B. Zadrozny and C. Elkan, "Learning and making decisions when costs and probabilities are both unknown," in *Proceedings of the 7th International Conference on Knowledge Discovery and Data Mining (KDD'01)*, 2001, pp. 204–213.
[30] P. Domingos, "Metacost: A general method for making classifiers cost–sensitive," in *Proceedings of the 5th International Conference on Knowledge Discovery and Data Mining (KDD'99)*, 1999, pp. 155–164.
[31] K. M. Ting, "An instance–weighting method to induce cost–sensitive trees," *IEEE Transactions on Knowledge and Data Engineering*, vol. 14, no. 3, pp. 659–665, 2002.
[32] J. Huang and C. X. Ling, "Using AUC and accuracy in evaluating learning algorithms," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 3, pp. 299–310, 2005.
[33] A. P. Bradley, "The use of the area under the roc curve in the evaluation of machine learning algorithms," *Pattern Recognition*, vol. 30, no. 7, pp. 1145–1159, 1997.
[34] O. Cordón, F. Herrera, F. Hoffmann, and L. Magdalena, *Genetic Fuzzy Systems. Evolutionary Tuning and Learning of Fuzzy Knowledge Bases*. World Scientific, 2001.
[35] O. Cordón, M. J. del Jesus, and F. Herrera, "A proposal on reasoning methods in fuzzy rule-based classification systems," *International Journal of Approximate Reasoning*, vol. 20, no. 1, pp. 21–45, 1999.
[36] D. Sheskin, *Handbook of parametric and nonparametric statistical procedures*. Chapman & Hall/CRC, 2006.