ISSN 1870-4069

Credit Assignment: Using Resampling Methods for Dealing with the Class Imbalance Problem

Víctor D. de la Cruz-Galarza[1], Yenny Villuendas-Rey[1], Cornelio Yáñez-Márquez[2]

[1] Instituto Politécnico Nacional, Centro de Innovación y Desarrollo Tecnológico en Cómputo, Mexico City, Mexico
[2] Instituto Politécnico Nacional, Centro de Investigación en Computación, Mexico City, Mexico
mhwara@gmail.com, yenny.villuendas@gmail.com, coryanez@gmail.com

Abstract. Nowadays, credit assignment constitutes a way in which persons or entities access to money. However, bad clients can cause big distress to financial institutions. If there are appropriate data banks whose patterns contain financial information from the scope of the allocation of credits, the intelligent pattern classifiers are ideal candidates to solve the credit assignment problem. Nevertheless, working with data sets from credit environment has the disadvantage that, in most of the cases, have unbalanced classes. This situation represents a problem at the moment of work with this kind of datasets due to the fact that unbalanced classes, in general, create biased learning. The consequences of this are reflected during the testing phase because the biased learning causes the classifiers to just recognize appropriately the elements of the ruling class and therefore, give us inaccuracy results. In this paper, we tested some undersampling and oversampling algorithms, and we compared their performance, based on the Imbalance Ratio measure, over different well-known credit related datasets.

Keywords: credit assignment, sampling techniques, instance selection, imbalanced data.

1 Introduction

One of the main activities of banks is to provide loans to their clients. If a debtor does not pay the money that was loaned in the established term, it violates the trust that the creditor granted and possibly, will stop lending him money. There are companies whose purpose is to keep track of credit payment, through which it is known which people have fulfilled their obligations to pay and who have stopped doing so; these companies are known as Credit Information Companies [1]. That is why credit institutions must be very careful when granting loans, because in doing so they use the money that people have deposited in their bank accounts.

This scenario clearly demonstrates the difficulties that delinquent clients cause to financial institutions. In this sense, it is of great interest for credit companies to have the real possibility of intelligent tools that evaluate potential clients, with a certain acceptable degree of certainty; these tools should provide entities with valuable

information regarding potential clients and answer a question whose answer is crucial: is the potential client a good or bad payer?

The advantages that the correct answer of this question would bring to the financial system are translated into a scenario where, as far as possible, the expenses derived from non-payment of a bad client are avoided.

If there are appropriate data banks whose patterns contain financial information from the scope of the allocation of credits, the intelligent pattern classifiers are ideal candidates to solve the credit assignment problem. Nevertheless, working with data sets from credit environment has the disadvantage that, in most of the cases, they have unbalanced classes and mixed attribute types [2], for which is very important to choose the classification models in accordance with this situation.

A situation like unbalanced classes is present in a dataset when one of the classes has more elements than the others. This situation represents a problem while working with this kind of datasets because unbalanced classes, in general, creates biased learning. The consequences of this are reflected during the testing phase because the biased learning causes that the classifiers just recognize appropriately the elements of the ruling class and therefore, give us inaccuracy results.

In this article is an experimental work using different sampling algorithms with credit related datasets with the purpose to know which one has the best performance under the circumstances described above is presented.

The rest of the paper is organized as follows. Section 2 details some previous works and Section 3 offers a discussion about the results obtained. Finally, the paper ends with some conclusions and future research suggestions.

2 Previous Works

From the point of view of supervised classification, the problem of the assignment of credit is a problem of two classes (credit is assigned or not assigned to the requestor) and of an unbalanced nature. This imbalance occurs because, in practice, more credits are awarded than those that are rejected. However, the costs of classification are not the same for both classes, due to the very nature of the phenomenon. [3, 4].

For example, if a potential good applicant is denied credit, the financial institution loses that client. However, if a bad applicant is granted credit, the financial institution has monetary losses, and possibly expenses associated with legal actions that have to be taken to recover the money invested. That is why the class of greatest interest in this phenomenon is the detection of potential bad applicants, who should not be granted credit [5]. Paradoxically, this class of greatest interest is the minority class in this phenomenon, which adds complexity to the data banks of work that are involved in the search for solutions to the problem of credit allocation in the context of Intelligent Computing [6].

In the scientific literature of the state- of- the- art, it is possible to find research works that report attempts to solve the problem of credit allocation through the application of Intelligent Computing. In these investigations, various models of supervised classification have been used; among them, is highlighted the use of Support Vector

Machines [7], Artificial Neural Networks [8, 9] and Classifier ensembles [10, 11], among others [12, 13, 14]. The experimental comparisons made to determine the performance of the classifiers in terms of the allocation of credit [15, 16, 17], exhibit certain problems that prevent generalizing the results they have published.

On the one hand, the studies incorporate few data banks, and to complicate matters, many of the data banks used are not public, nor are they available for use; In addition, there are almost no common data banks in the different investigations. Additionally, in the documentary study of the state- of- the- art carried out in the framework of this work, it has been observed that, if a research group has used a certain supervised classifier, in other researches this is not taken into account, but rather they are used other supervised classifiers.

The No Free Lunch [18] theorems argue that there is no superiority of one classifier over others, over all data banks and all performance measures. However, recent studies point to the existence of a good performance of associative classifiers in the solution of problems of supervised classification of the financial environment [19].

It is a fact known to the scientific community that, on numerous occasions, the preprocessing of data contributes to the improvement of the performance of certain supervised classifiers; in particular, when data banks show an imbalance between classes [20, 21]. The literature reports several investigations that have been conducted in order to determine the impact of data preprocessing on improving solutions to the problem of granting credit [6, 22]. In particular, the computational problem related to the selection of instances (applicants) [4] has aroused great interest in the scientific community, so that in recent years emphasis has been placed on the study of techniques for the selection of classifiers for unbalanced data [3].

Moreover, in the comparative studies reviewed [5, 23], there is no consensus as to which are the best preprocessing techniques for the different classifiers in the allocation of credit. The previous considerations allow affirming, without a doubt, that the results of the mentioned experimental comparisons are hardly conclusive. With this investigation, it is intended to successfully attack this type of problem.

3 Results and Discussion

3.1 Datasets

In this section, we describe the data banks that will be used to evaluate the impact of the pre-processing of financial data on the performance of associative classifiers. These data banks are well known in literature, as well as being a reference, because they are widely used in many of the research works carried out so far. These data banks are known as "*Give me some credit*"[1], "*Iranian*"[2] and "*Polish bankruptcy*"[3], which are very interesting for this type of research, due to the nature of the data, because they include a wide variety of attributes, a high level of imbalance, in addition to having

[1] https://www.kaggle.com/c/GiveMeSomeCredit/data

[2] Personal shared by Hassan Sabzevari.

[3] https://archive.ics.uci.edu/ml/datasets/Polish+companies+bankruptcy+data

missing data in many of the data banks. By way of summary, a description of the data banks used in the present investigation is shown in Table 1. The abbreviation IR represents the ratio of imbalance.

Table 1. Characteristics of the datasets used in this work.

Data set	*Instances*	*Attributes*	*IR*	*Missing*
Give me credit	150000	10	13.9611	No
Iranian	1002	28	19.0400	Yes
Polish_year1	7027	64	24.9299	Yes
Polish_year2	10173	64	24.4325	Yes
Polish_year3	10503	64	20.2182	Yes
Polish_year4	9792	64	18.0136	Yes
Polish_year5	5910	64	13.4146	Yes

As shown, all data banks are very unbalanced (it is considered unbalanced from IR> 1.5 and all these data banks have IR> 13), and six contain absences of information. Note that in all cases have only two classes.

3.2 Algorithms to Compare

In a large number of works, novel methods have been proposed to address the problem of imbalance between classes. Those approaches are classified into two groups: approaches at the level of algorithms in which a new algorithm is created or one that already exists is modified and data-level approaches, in which data are modified in order to lessen the impact on the performance of classification algorithms when there is an imbalance in the distribution of classes.

In this section, the class balancing algorithms that will be evaluated in the present investigation are addressed. First, reference is made to the oversampling algorithms and, subsequently, to the sub-sampling algorithms (undersampling). In each case, its operation is detailed and a brief reference is made to its main characteristics, as well as its application or not to the financial field.

In the state- of- the- art, it is possible to find several articles [24, 25, 26, 27], where the pre-processing of data banks is addressed to reduce the impact caused by the distribution of classes. In those articles, it has been empirically demonstrated that the application of a preprocessing stage to balance the class distribution is usually a useful solution to improve the quality of the identification of new instances.

The data pre-processing techniques are divided into three groups: *Undersampling algorithms*, which are based on the elimination of instances of the majority class, *Oversampling algorithms*, which are based in the creation of instances of the minority class, by replication or modification of existing instances, and *Hybrid algorithms*, which are a combination of both over and under sampling techniques.

The oversampling algorithms seek to match the quantities of objects in each class by over-sampling the minority classes. In this way, the number of objects in these classes

will be artificially increased, ensuring that all classes have approximately the same number of objects. The main techniques of selection of instances by oversampling that will be used for the comparative analysis carried out in this work are listed and detailed below.

SMOTE: Synthetic Minority Over-sampling TEchnique [28], this method of preprocessing has become one of the most renowned in terms of oversampling techniques. The fundamental principle of this technique is based on creating synthetic instances for the minority class, through the nearest k neighbor of each of the instances of this class. The new instances are created by interpolating the sample vectors of the sample (example) of the minority class and its respective nearest neighbor. Each difference is multiplied randomly by zero or one. Then the non-zero characteristic vectors are taken as the new synthetic instances. This technique forces the decision region of the minority class to make it more general. The main disadvantage of this technique is that it can create instances that over train the classifier.

ADASYN: ADAptive SYNthetic Sampling [29], this method is based on the generation of instances adaptively for the minority class according to their distributions: the synthetic instances are generated for instances of the minority class that are more difficult to learn compared to those instances that are of the minority class and easier to learn.

ROS: Random over-sampling [24], This is a method that creates new synthetic instances in a random way. This is done until both classes contain the same number of instances.

ADOMS: Adjusting the Direction Of the synthetic Minority clasS examples [30], This method works similar to SMOTE. However, this method generates synthetic instances along the first principal component axis (PCA) of the local data of the distribution using the nearest k neighbors.

SPIDER: Selective Preprocessing of Imbalanced Data [31], this method combines local sampling of the minority class with filtering of difficult instances of the majority class. This method identifies which instances are labeled as noisy or difficult (misclassified) by the kNN classifier. Then, noisy instances can be duplicated, deleted or re-labeled depending on the option to be chosen (weak or strong).

As mentioned previously, the undersampling algorithms seek to equip the quantities of objects in each class, by sampling the major classes. Thus, the objects that are considered less relevant are eliminated, so that all classes have approximately the same number of objects. Next, the undersampling algorithms evaluated in the present investigation are explained.

TL: Tomek's modification of Condensed Nearest Neighbor [32], in this method if two instances form a Tomek link, then one of them is noise or the two instances are on the border. Prior to applying the condensed rule of the nearest neighbor (NN), this method obtains a set of objects containing only the objects near the decision boundaries.

RUS: Random under-sampling [24] is an algorithm that randomly selects instances of the majority class to be eliminated until both classes are balanced.

OSS: One Sided Selection [33], in this method, all the instances belonging to the minority class and instances of the misclassified major class are selected (by the 1-NN

classifier) in order to find Tomek links between them. Instances of the majority class participating in a Tomek link are removed.

CNNTL: Condensed Nearest Neighbor + Tomek's modification of Condensed Nearest Neighbor [24], In this algorithm the CNN and TL methods are combined. The main idea is to reduce the size of the original data set through the elimination of certain objects by applying CNN without significantly affecting the performance of the NN classification using the information provided by the Tomek link method.

NCL: Neighborhood Cleaning Rule [34], uses the ENN rule to remove objects from the majority class. ENN removes any object whose label class differs from the class of at least three of its five closest neighbors.

3.3 Discussion

Each algorithm was tested with the different datasets in the KEEL software [35] using the default parameters offered. We used a 5-fold cross validation procedure as model validation technique. Tables 2 and 3 show the results for the undersampling and oversampling algorithms, respectively. We use the Imbalance Ratio measure (IR) as performance measure.

Table 2. Imbalance Ratio for the undersampling algorithms.

Datasets	CNNTL	NCL	OSS	RUS	TL	Original
Give me credit	1.36	10.84	1.39	1.00	12.95	13.96
Iranian	1.47	14.53	2.18	1.00	17.91	19.00
Polish_year1	1.40	21.14	2.50	1.00	23.70	24.93
Polish_year2	1.41	20.31	2.56	1.00	22.97	24.43
Polish_year3	1.33	16.33	2.45	1.00	18.86	20.22
Polish_year4	1.25	14.42	2.22	1.00	16.72	18.01
Polish_year5	1.15	10.23	1.62	1.00	12.28	13.41

As can be seen in Table 2, as expected, the Random undersampling Method (RUS) obtained a perfectly balanced dataset. In addition, the CNNTL algorithms obtained very good imbalance ratios, all of them very close to one. In a similar way, one Side Selection (OSS) obtained good results, although not as good as the ones by CNNTL. Neither NLC nor TL obtained good results, failing to obtain a balanced dataset.

As shown in Table 3, all oversampling algorithms but SPIDER obtained a perfectly balanced dataset. However, it was by significantly increasing the number of instances in the dataset. Figure 2 shows the differences according to instance amount among undersampling and oversampling methods. ADASYN, ADOMS, ROS and SMOTE algorithms obtained the same number of instances; then Figure 2 only depicts the results for ADASYN.

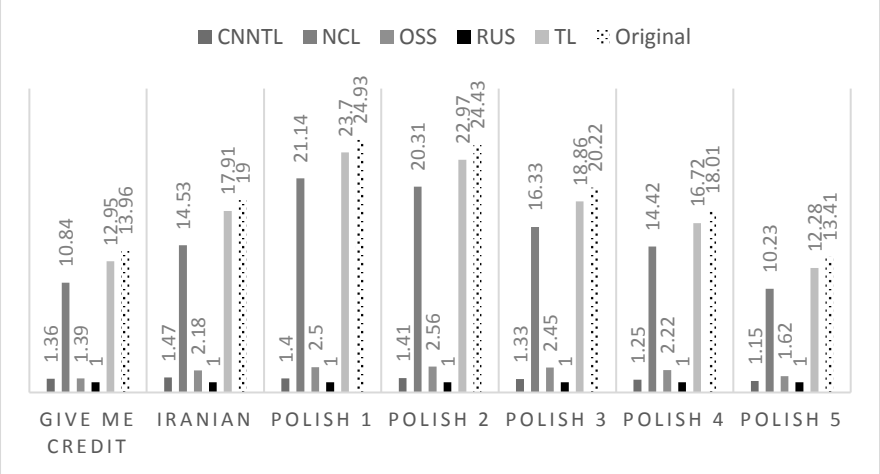

Fig. 1. Graphical representation of the Imbalance Ratio for undersampling methods.

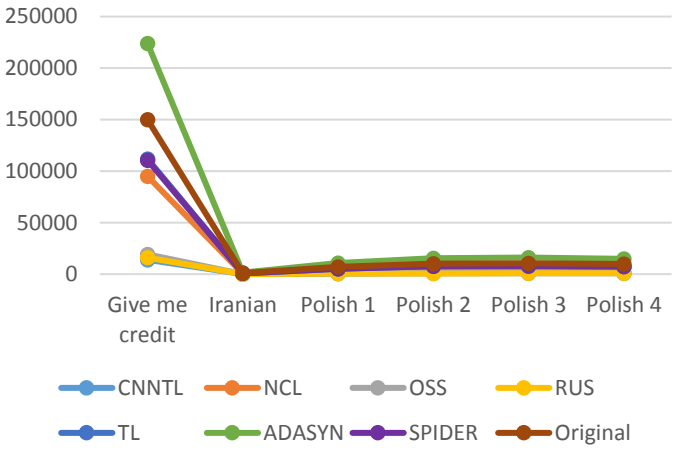

Fig. 2. Graphical representation of the amount of instances selected (X axis) for each considered dataset (Y axis).

Table 3. Imbalance Ratio for the oversampling algorithms.

Datasets	ADASYN	ADOMS	ROS	SMOTE	SPIDER	Original
Give me credit	1.00	1.00	1.00	1.00	4.22	13.96
Iranian	1.00	1.00	1.00	1.00	6.18	19.00
Polish_year1	1.00	1.00	1.00	1.00	6.57	24.93
Polish_year2	1.00	1.00	1.00	1.00	6.49	24.43
Polish_year3	1.00	1.00	1.00	1.00	5.50	20.22
Polish_year4	1.00	1.00	1.00	1.00	5.08	18.01
Polish_year5	1.00	1.00	1.00	1.00	3.98	13.41

The experiments show that oversampling methods significantly increase the amount of instances, rising both the storage and execution computational costs. On the other hand, undersampling method obtain balanced datasets, and significantly reduce the amount of instances.

4 Conclusions and Future Work

In the credit environment, there are some datasets that can be considered important to test automated decision-making systems, but in most cases, these datasets have some characteristics (such as class imbalance) that make this task more complicated. In this work, we compared 10 different sampling techniques in credit environment using the imbalance ratio measure. Our studies showed that CNNNTL and RUS models turned out to be best sampling algorithm for almost all the datasets used in this work. As future work, we would address the classification performance of associative classifier over the original and balanced datasets.

Acknowledgments. The authors would like to thank the Instituto Politécnico Nacional (Secretaría Académica, Comisión de Operación y Fomento de Actividades Académicas, Secretaría de Investigación y Posgrado, Centro de Investigación en Computación, and Centro de Innovación y Desarrollo Tecnológico en Cómputo), the CONACYT, and SNI for their economical support to develop this work.

References

1. C. D. D. D. H. C. D. LA UNIÓN: Ley para regular las sociedades de información crediticia. Diario Oficial de la Federación. Ciudad de México (2002)
2. Abdi, L., Hashemi, S.: To combat multi-class imbalanced problems by means of over-sampling and boosting techniques. Soft Comput. 19(12), pp. 3369–3385 (2015)
3. Bischl, B., Kuhn, T., Szepannek, G.: On Class Imbalance Correction for Classification Algorithms in Credit Scoring. In: Oper. Res. Proc. 2014 Sel. Pap. Annu. Int. Conf. Ger. Oper. Res. Soc. (GOR). RWTH Aachen Univ. Ger. Sept. 2-5, 2014 (2016)
4. Garcia, V., Marques, M. I., Sanchez, J. S.: On the use of data filtering techniques for credit risk prediction with instance-based models. Expert Syst. Appl. 39(18), pp. 13267–13276 (2012)
5. Marqués, A. I., García, V., Sánchez, J. S.: On the suitability of resampling techniques for the class imbalance problem in credit scoring. J. Oper. Res. Soc. 64(7), pp. 1060–1070 (2013)
6. Banasik, J., Crook, J., Thomas, L.: Sample selection bias in credit scoring models. J. Oper. Res. Soc. 54(8), pp. 822–832 (2003)
7. Danenas, P., Garsva, G.: Selection of Support Vector Machines based classifiers for credit risk domain. Expert Syst. Appl. 42(6), pp. 3194–3204 (2015)
8. Zhao, Z., Xu, S., Kang, B. H., Kabir, M. M. J., Liu, Y., Wasinger, R.: Investigation and improvement of multi-layer perceptron neural networks for credit scoring. Expert Syst. Appl. 42(7), pp. 3508–3516 (2015)
9. Khashman, A.: Neural networks for credit risk evaluation: Investigation of different neural models and learning schemes. Expert Syst. Appl. 37(9), pp. 6233–6239, Sep. (2010)

10. Xiao, H., Xiao, Z., Wang, Y.: Ensemble classification based on supervised clustering for credit scoring. Appl. Soft Comput. J. 43, pp. 73–86 (2016)
11. Wang, H., Xu, Q., Zhou, L.: Large unbalanced credit scoring using lasso-logistic regression ensemble. PLoS One 10(2) (2015)
12. Cao, V. L., Le-Khac, N. A., O'Neill, M., Nicolau, M., McDermott, J.: Improving fitness functions in genetic programming for classification on unbalanced credit card data. Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics) 9597, pp. 35–45 (2016)
13. Zhang, Z., Gao, G., Shi, Y.: Credit risk evaluation using multi-criteria optimization classifier with kernel, fuzzification and penalty factors. Eur. J. Oper. Res. 237(1), pp. 335–348 (2014)
14. Tomczak, J. M., Zięba, M.: Classification Restricted Boltzmann Machine for comprehensible credit scoring model. Expert Syst. Appl. 42(4), pp. 1789–1796 (2015)
15. Louzada, F., Ara, A., Fernandes, G. B.: Classification methods applied to credit scoring: Systematic review and overall comparison. Surv. Oper. Res. Manag. Sci. 21(2), pp. 117–134, Dec. (2016)
16. Lessmann, S., Baesens, B., Seow, H. V., Thomas, L. C.: Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. Eur. J. Oper. Res. 247(1), pp. 124–136 (2015)
17. Brown, I., Mues, C.: An experimental comparison of classification algorithms for imbalanced credit scoring data sets. Expert Syst. Appl. 39(3), pp. 3446–3453, Feb. (2012)
18. Wolpert, D. H.: The supervised learning no-free-lunch theorems. Soft Comput. Ind., pp. 25–42 (2002)
19. Villuendas-Rey, Y., Rey-Benguría, C. F., Ferreira-Santiago, Á., Camacho-Nieto, O., Yáñez-Márquez, C.: The Naïve Associative Classifier (NAC): A novel, simple, transparent, and accurate classification model evaluated on financial data. Neurocomputing. Jun. (2017)
20. 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. (Ny) 250, pp. 113–141, Nov. (2013)
21. Dal Pozzolo, A., Caelen, O., Bontempi, G.: When is Undersampling Effective in Unbalanced Classification Tasks? In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 9284, pp. 200–215 (2015)
22. Piramuthu, S.: On preprocessing data for financial credit risk evaluation. Expert Syst. Appl. 30(3), pp. 489–497, Apr. (2006)
23. Crone, S. F., Finlay, S.: Instance sampling in credit scoring: An empirical study of sample size and balancing. Int. J. Forecast. 28(1), pp. 224–238, Jan. (2012)
24. 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), pp. 20–29 (2004)
25. Batuwita, R., Palade, V.: Efficient resampling methods for training support vector machines with imbalanced datasets. In: Proceedings of the International Joint Conference on Neural Networks (2010)
26. Fernández, A., del Jesus, M. J., Herrera, F.: On the 2-tuples based genetic tuning performance for fuzzy rule based classification systems in imbalanced data-sets. Inf. Sci. (Ny). 180(8), pp. 1268–1291 (2010)
27. Fernandez, A., Garcia, S., del Jesus, M. J., Herrera, F.: A study of the behaviour of linguistic fuzzy rule based classification systems in the framework of imbalanced data-sets. Fuzzy Sets Syst. 159(18), pp. 2378–2398 (2008)

28. Chawla. N., Bowyer, K.: SMOTE: Synthetic Minority Over-sampling Technique Nitesh. J. Artif. Intell. Res. 16, pp. 321–357 (2002)
29. He, H., Bai, Y., Garcia, E. A., Li, S.: ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In: Proceedings of the International Joint Conference on Neural Networks, pp. 1322–1328 (2008)
30. Tang, S., Chen, S.P.: The generation mechanism of synthetic minority class examples. In: 5th Int. Conf. Inf. Technol. Appl. Biomed. ITAB 2008 conjunction with 2nd Int. Symp. Summer Sch. Biomed. Heal. Eng. IS3BHE 2008, pp. 444–447 (2008)
31. Stefanowski, J., Wilk, S.: Selective pre-processing of imbalanced data for improving classification performance. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 5182. LNCS, Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 283–292 (2008)
32. Tomek, I.:Two Modifications of CNN. IEEE Trans. Syst. Man Cybern. 6, pp. 769–772, (1976)
33. Kubat, M., Matwin, S.: Addressing the Curse of Imbalanced Training Sets: One Sided Selection. Icml 97, pp. 179–186 (1997)
34. Laurikkala, J.: Improving identification of difficult small classes by balancing class distribution. Proc. In: 8th Conf. AI Med. Eur. Artif. Intell. Med., pp. 63–66 (2001)
35. Alcalá-Fdez, F. H. J., Sánchez, L., García, S., del Jesus, M.J., Ventura, S., Garrell, J.M., Otero, J., Romero, C., Bacardit, C., Rivas, V.M., Fernández, J.C.: KEEL (software) (2009)