An Empirical Study of Oversampling and Undersampling for Instance Selection Methods on Imbalance Datasets

Julio Hernandez, Jesús Ariel Carrasco-Ochoa,
and José Francisco Martínez-Trinidad

Instituto Nacional de Astrofísica Óptica y Electrónica, Computer Science Department, Luis Enrique Erro No. 1, Sta. María Tonantzintla, Puebla, CP 72840, Mexico
{julio.hernandez.t,ariel,fmartine}@ccc.inaoep.mx
http://ccc.inaoep.mx

Abstract. Instance selection methods get low accuracy in problems with imbalanced databases. In the literature, the problem of imbalanced databases has been tackled applying oversampling or undersampling methods. Therefore, in this paper, we present an empirical study about the use of oversampling and undersampling methods to improve the accuracy of instance selection methods on imbalanced databases. We apply different oversampling and undersampling methods jointly with instance selectors over several public imbalanced databases. Our experimental results show that using oversampling and undersampling methods significantly improves the accuracy for the minority class.

Keywords: supervised classification, instance selection, oversampling, undersampling, imbalanced datasets.

1 Introduction

The classification process requires a training set T to create a model which will be used to assign a class to unseen examples. Nevertheless, in a training sample usually there are some redundant and/or noisy examples that are useless for the classification process and they could negatively affect the classification accuracy [1–3]. Instance selection (IS) is focused on this problem. The IS methods select a subset S of the training set T such that S allows to get an accuracy as similar as possible to the one computed using T [4].

In an ideal scenario the classes are balanced, that is, the number of instances for each class are almost the same. But, some real world databases don't have this property, i.e. their classes are imbalanced [5–7].

Instance selection algorithms have demonstrated to perform well when the classes are balanced [3], however, this is not true for imbalanced datasets, instance selection algorithms get low accuracy in this kind of problems because they tend to remove too many instances from the minority class, damaging their performance [8, 9]. For this reason, in this paper we focus on the study

J. Ruiz-Shulcloper and G. Sanniti di Baja (Eds.): CIARP 2013, Part I, LNCS 8258, pp. 262–269, 2013.

of combining oversampling and undersampling methods with instance selection algorithms, in order to get good results in imbalanced problems.

This paper is divided in the following sections: Section 2, briefly describes the instance selection algorithms and oversampling and undersampling methods that will be used in our experiments. Section 3 presents the experimental results. Finally, section 4 provides some conclusions and future work.

2 Related Work

In the literature there have been reported several instance selection algorithms. Most of them are based on the KNN rule, for example DROP3 [10], IB3 [1], ICF [2]. Another group of instance selection algorithms, clearly different from the former, are those based on evolutionary algorithms, some examples of these methods are CHC [11], GGA [12], SGA[13]. More recently some instance selection algorithms for large databases have been proposed, which can be applied in problems where conventional IS algorithms, as those previously commented, cannot produce a solution in a reasonable time. For our study we have selected DROP3 as a representative of those algorithms based on KNN; CHC as a representative of algorithms based on evolutionary algorithms; and IRB [14] as a representative of IS algorithms for large datasets.

1. **DROP3** [10]: This algorithm is based on the concept of associate. The associates of an instance P are those instances such that P is one of their k nearest neighbors. First, DROP3 applies ENN [15] for noise-filtering over the initial training set T. Then, the remaining instances in T are sorted by the distance to their nearest enemy, which is the nearest instance with different class. DROP3 iteratively removes an instance P if the majority of its associates in T would be classified correctly without P.
2. **CHC** [11]: In [16] a comprehensive study of different evolutionary algorithms applied in the instance selection field is presented. From this study the CHC algorithm was able to achieve the best overall performance among the tested evolutionary algorithms. During each generation the CHC develops the following steps: (1) CHC uses a parent population of size n to generate an intermediate population of the same size, which are used to generate n potential offsprings. (2) Then, in a competition the best n chromosomes from the parent and offsprings population are selected to form the next generation.
3. **IRB** [14]: This algorithm tries to preserve the border instances (those located in a region where there are similar instances from different classes) by computing an instance ranking for each class based on the distance of each instance to border instances. This algorithm selects a predefined % of instances having high, medium and low values, in the ranking.

The described algorithms are not able to deal with imbalance datasets by itself. However, they maintain their reduction capabilities. One way to deal with the problem of imbalance dataset is applying some oversampling or undersampling techniques. Therefore, in this paper, we present an empirical study about

the use of oversampling and undersampling methods to improve the accuracy of instance selection methods on imbalanced databases. For our study we have selected the following oversampling and undersampling methods, which are some of the most reported in the literature.

1. **Resample**: This oversampling method produces an uniform class distribution. Resample applies a random subsampling to the majority class and an oversampling, with replacement, to the minority class.
2. **Spread Subsampling**: This undersampling technique produces a random subsample of a database. The class distribution is adjusted through a random elimination of objects from the majority class.
3. **Synthetic Minority Over-sampling Technique** (SMOTE) [17]: This oversampling approach generates synthetic samples of the minority class based on nearest neighbors. The synthetic examples are generated computing the difference between feature vectors and their nearest neightbors, then this difference is multiplied by a random number between 0 and 1, and the result is added to the feature vector under consideration.

On the other hand, in [18] the One-sided method is proposed as an approach to instance selection over imbalanced datasets. The main idea of this work is to carry out instance selection only over the majority class leaving intact the minority class. Another approach to the same problem [19] involves that an instance can be choose more than once considering the number of nearest neighbours. The main idea of this work is to cover the same amount of space with fewer instances.

3 Experimental Results

For our experiments we use 18 databases taken from the KEEL repository [20]. Table 1 describes the used databases. The databases were sorted in ascending way according to their imbalance ratio (IR) computed as the ratio between the size of the majority and minority classes. As it is shown in Table 1 the imbalance ratio is very different for each database, for that reason we grouped the databases as: IR 1-3, IR 3-9 and IR $>$ 9. It will allow us to analyze the behavior of oversampling and undersampling techniques, jointly with instance selection, depending of the imbalance ratio.

For each database, we performed 10 fold cross validation averaging the classification accuracy for the minority and majority classes separately as well as the global accuracy, in our experiments we also include the F-Measure.

We used the implementations of Resample, Spread Subsample and SMOTE taken from WEKA [21] with their default parameters, except for SMOTE where we adjusted the percentage parameter (-P) according with the imbalance ratio of each database (-P (IR * 100)). We used different percentage values because each minority class needs a different percentage of oversampling, for example, in the abalone database we have an imbalance ratio of 133, the minority class have only 28 examples, if we apply SMOTE with a fix percentage of 100 the result

Table 1. Characteristics of the databases used in the experiments. IR: the imbalance ratio; N.O.: number of objects; cl+: size of the majority class; cl-: size of the minority class.

Database	IR	N.O.	cl+	cl-	Database	IR	N.O.	cl+	cl-	Database	IR	N.O.	cl+	cl-
IR 1-3					**IR 3-9**					**IR > 9**				
ionosphere	1.8	315	202	113	spliceie	3.2	2871	2180	691	ecoliom	15.8	302	284	18
pima	1.9	691	405	241	vehiclevan	3.3	761	582	179	abalone918	16.8	657	620	37
tic-tac-toe	1.9	862	564	298	ecolim	3.4	302	233	69	yeastme2	28.7	1335	1290	45
german	2.3	900	630	270	hepatitis	5.5	72	61	11	yeastme1	33.2	1335	1296	39
phoneme	2.4	4863	3436	1427	segment0	6.0	2077	1781	296	yeastexc	42.1	1335	1304	31
yeast	2.5	1335	949	386	ecolimu	8.7	302	271	31	abalone19	133.1	3756	3728	28

will be an imbalance database with an imbalance ratio very close to the original, in this case we needed an oversampling percentage of 13300% (133 * 100).

The implementations of KNN, DROP3 and CHC were taken from the KEEL software [22] and the implementation of IRB was supplied by the authors. The KNN algorithm with K = 1 is used as base line. For DROP3, CHC and KNN we used the default parameters in the KEEL software and for IRB we used the parameters suggested by the authors in [14].

For the One-sided prototype selection method [18] we followed the steps proposed by the authors. First, we apply DROP3, CHC or IRB over the whole database. The original examples of the minority class jointly with the examples selected by the IS algorithm in the majority class are used as training for the 1-NN classifier.

3.1 Experimental Comparison

Tables 2, 3 and 4 show the results of the experiments for the databases with imbalanced ratio 1-3, 3-9 and grater than 9, respectively. Each table is divided in two main columns, the right column presents the results of applying One-sided method and the left column presents the results of applying instance selection algorithms after oversampling or undersampling. In each sub-table the average accuracy for the minority and majority classes, the global accuracy, the F-Measure and the reduction percentage are reported. The numbers in bold represent the best results for the respective column and row.

The results for databases with IR in the interval 1-3 (see Table 2), show that applying an instance selection method (IS) followed by oversampling or undersampling always produces better results for minority class and global accuracy than applying the One-sided method. In terms of F-Measure it gets in most of the cases better results if an instance selection algorithm is applied after oversampling or undersampling in contrast to applying the One-sided method. These results show that IRB got the best accuracy for the minority class (and for global accuracy) with respect to CHC, DROP3 and KNN no matter if an oversampling or undersampling techniques is applied or not before applying IRB. However, SMOTE & IRB obtained the best results. On the other hand, although the accuracy for the majority class is greatly improved by One-sided, simultaneously,

Table 2. Experiment results for the databases with IR 1-3. Red.: reduction percentage, Acc+: Majority class Accuracy, Acc-: Minority class Accuracy, AccG: Global Accuracy, F-M: F-Measure

Selector	Without over or under sampling					One-sided					Resampling 300%				
	Red.	**Acc-**	**Acc+**	**AccG.**	**F-M**	**Red.**	**Acc-**	**Acc+**	**AccG.**	**F-M**	**Red.**	**Acc-**	**Acc+**	**AccG.**	**F-M**
KNN	0.00	0.66	0.79	0.74	0.54	0.00	0.66	0.79	0.74	0.54	0.00	0.61	0.81	0.75	0.59
CHC	**0.99**	**0.65**	0.79	**0.76**	**0.54**	0.68	0.37	**0.89**	0.43	0.53	**0.99**	**0.56**	0.84	**0.72**	**0.61**
DROP3	**0.84**	**0.54**	0.79	**0.71**	**0.54**	0.60	0.38	**0.85**	0.52	0.53	**0.91**	**0.55**	0.80	**0.71**	**0.55**
IRB	**0.60**	0.47	**0.92**	**0.78**	0.56	0.43	**0.54**	0.86	0.69	**0.62**	**0.59**	**0.61**	0.81	**0.75**	0.61
	SMOTE					**SMOTE and One-sided**					**Resampling 300% and One-sided**				
KNN	0.00	0.62	0.83	0.75	0.62	0.00	0.62	0.83	0.75	0.62	0.00	0.61	0.81	0.75	0.59
CHC	**0.99**	**0.55**	0.85	**0.71**	**0.61**	0.47	0.37	**0.90**	0.43	0.53	0.51	0.39	**0.95**	0.47	0.54
DROP3	**0.82**	**0.51**	0.81	**0.69**	0.56	0.38	0.45	**0.90**	0.61	**0.59**	0.47	0.39	**0.87**	0.53	0.54
IRB	**0.59**	**0.74**	0.79	**0.77**	**0.68**	0.40	0.53	**0.88**	0.67	0.60	0.45	0.50	**0.87**	0.66	**0.62**
	Spread Subsample					**Spread Subsample and One-sided**				
KNN	0.00	0.55	0.85	0.70	0.61	0.00	0.55	0.85	0.70	0.61
CHC	**0.98**	**0.56**	0.85	**0.72**	**0.62**	0.49	0.37	**0.93**	0.42	0.53
DROP3	**0.79**	**0.50**	0.81	**0.67**	0.56	0.45	0.42	**0.91**	0.56	**0.57**
IRB	**0.60**	**0.70**	0.78	**0.75**	**0.65**	0.41	0.50	**0.90**	0.63	0.62

Table 3. Experiment results for the databases with IR 3-9. Red.: reduction percentage, Acc+: Majority class Accuracy, Acc-: Minority class Accuracy, AccG: Global Accuracy, F-M: F-Measure

Selector	Without over or under sampling					One-sided					Resampling 300%				
	Red.	**Acc-**	**Acc+**	**AccG.**	**F-M**	**Red.**	**Acc-**	**Acc+**	**AccG.**	**F-M**	**Red.**	**Acc-**	**Acc+**	**AccG.**	**F-M**
KNN	0.00	0.61	0.94	0.87	0.63	0.00	0.61	0.94	0.87	0.63	0.00	0.61	0.93	0.87	0.64
CHC	**0.98**	**0.65**	0.92	**0.86**	**0.62**	0.80	0.28	**0.93**	0.49	0.42	**0.99**	**0.61**	0.95	**0.85**	**0.68**
DROP3	**0.90**	**0.60**	0.93	**0.84**	**0.65**	0.75	0.34	**0.96**	0.63	0.48	**0.95**	**0.60**	0.92	**0.84**	**0.64**
IRB	**0.60**	**0.65**	**0.96**	**0.91**	0.66	0.52	0.59	**0.96**	0.84	**0.67**	**0.59**	**0.74**	0.93	**0.90**	**0.71**
	SMOTE					**SMOTE and One-sided**					**Resampling 300% and One-sided**				
KNN	0.00	0.61	0.95	0.86	0.68	0.00	0.61	0.95	0.86	0.68	0.00	0.61	0.93	0.87	0.64
CHC	**0.99**	**0.62**	0.95	**0.86**	**0.70**	0.48	0.31	**0.99**	0.54	0.46	0.51	0.36	**0.98**	0.63	0.52
DROP3	**0.90**	**0.59**	0.91	**0.84**	**0.61**	0.42	0.42	**0.97**	0.73	0.57	0.49	0.35	**0.95**	0.64	0.50
IRB	**0.60**	**0.91**	0.90	**0.90**	**0.78**	0.43	0.52	**0.93**	0.85	0.56	0.46	0.61	**0.96**	0.86	0.68
	Spread Subsample					**Spread Subsample and One-sided**				
KNN	0.00	0.55	0.96	0.82	0.65	0.00	0.55	0.96	0.82	0.65
cline1-11 **CHC**	**0.96**	**0.57**	0.96	**0.83**	**0.67**	0.48	0.27	**0.98**	0.48	0.42
DROP3	**0.84**	**0.55**	0.93	**0.80**	**0.63**	0.40	0.37	**0.97**	0.65	0.52
IRB	**0.63**	**0.89**	0.86	**0.87**	**0.71**	0.45	0.48	**0.97**	0.77	0.61

Table 4. Experiment results for the databases with IR > 9. Red.: reduction percentage, Acc+: Majority class Accuracy, Acc-: Minority class Accuracy, AccG: Global Accuracy, F-M: F-Measure

Selector	Without over or under sampling					One-sided					Resampling 300%				
	Red.	**Acc-**	**Acc+**	**AccG.**	**F-M**	**Red.**	**Acc-**	**Acc+**	**AccG.**	**F-M**	**Red.**	**Acc-**	**Acc+**	**AccG.**	**F-M**
KNN	0.00	0.45	0.98	0.97	0.41	0.00	0.45	0.98	0.97	0.41	0.00	0.41	0.98	0.97	0.41
CHC	**0.99**	**0.41**	**0.98**	**0.97**	**0.35**	0.96	0.07	0.97	0.32	0.13	**0.99**	**0.33**	**0.99**	**0.90**	**0.41**
DROP3	**0.98**	**0.28**	**0.98**	**0.92**	**0.33**	0.95	0.07	**0.98**	0.45	0.13	**0.97**	**0.25**	**0.98**	**0.89**	**0.30**
IRB	**0.60**	0.30	**1.00**	**0.97**	0.33	0.50	**0.33**	0.98	0.92	**0.36**	**0.59**	0.39	**0.98**	**0.96**	0.38
	SMOTE					**SMOTE and One-sided**					**Resampling 300% and One-sided**				
KNN	0.00	0.35	0.99	0.93	0.42	0.00	0.35	0.99	0.93	0.42	0.00	0.41	0.98	0.97	0.41
CHC	**0.99**	**0.32**	**0.99**	**0.91**	**0.40**	0.51	0.10	**0.99**	**0.91**	**0.40**	0.51	0.14	**0.99**	0.64	0.23
DROP3	**0.94**	**0.28**	0.98	**0.91**	**0.33**	0.48	0.18	**0.99**	0.73	0.26	0.51	0.07	**0.98**	0.44	0.13
IRB	**0.59**	**0.66**	0.92	**0.91**	**0.38**	0.40	0.28	**0.99**	0.85	0.36	0.48	**0.42**	**0.98**	**0.96**	**0.44**
	Spread Subsample					**Spread Subsample and One-sided**									
KNN	0.00	0.18	0.99	0.79	0.27	0.00	0.18	0.99	0.79	0.27					
CHC	**0.95**	**0.21**	0.99	**0.82**	**0.31**	0.47	0.08	**1.00**	0.36	0.14					
DROP3	**0.82**	**0.18**	**0.99**	**0.77**	**0.27**	0.40	0.11	**0.99**	0.55	0.19					
IRB	**0.65**	**0.81**	0.79	**0.79**	**0.25**	0.35	0.12	**0.96**	0.57	0.19					

it get a worse accuracy for the minority class, therefore the overall accuracy obtained by One-sided is far outweighed by the results obtained by applying oversampling or undersampling jointly with IS.

For databases with IR in the interval 3-9 (see Table 3), the results show that applying an IS method after oversampling or undersampling always produces better results for minority class and global accuracy than applying the One-sided method. In terms of F-Measure it gets in most of the cases better results if an instance selection algorithm is applied after oversampling or undersampling in contrast to applying the One-sided method. The same as in databases with IR in the interval 1-3. The results show that although IRB gets the lower reduction percentages, it outperformed the accuracy of the minority class, global accuracy and F-Measure with respect to CHC, DROP3 and KNN no matter if an oversampling or undersampling technique is applied before applying IRB. Again, as in databases with IR in the interval 1-3, SMOTE & IRB obtained the best results. On the other hand, the One-sided technique is far outweighed by the results obtained by applying oversampling or undersampling jointly with IS methods.

The results for databases with IR greater than 9 (see Table 4), show that applying an IS method after oversampling or undersampling always produces better results for the minority class and the global accuracy than applying the One-sided method. In this type of databases Spread Subsample & IRB obtained the best accuracies for the minority class. However, the best global accuracy was obtained by Resampling & IRB. The results show that the One-sided method is far outweighed by the results obtained by applying oversampling or undersampling jointly with an IS method.

4 Conclusions and Future Work

The instance selection methods are sensitive to imbalance databases. The main problem is that the minority class always obtains lower accuracy than the majority class. Only a few works have been proposed to deal with the imbalance problem on instance selection [18], however, there are some techniques based on oversampling and undersampling that can be combined with IS methods to improve the accuracy of the minority class.

The main contribution of this work is an empirical study of combining oversampling and undersampling techniques with some instance selection methods based on nearest neighbor rule (NN), evolutionary algorithms and ranking algorithms. The results show that this combination improves the accuracy of the minority class with respect to the original dataset. For imbalanced databases with an IR in the interval 1-9 the best option is to use SMOTE & IRB, for databases with an IR greater than 9 there are two main combinations: Resample & IRB, which obtains high global accuracy, and Spread Subsample & IRB, which obtains high accuracy for the minority class.

As future work, we plan to develop an instance selection algorithm to directly deal with imbalanced datasets.

Acknowledgment. This work was partly supported by the National Council of Science and Technology of Mexico (CONACyT) through the project grants CB2008-106443 and CB2008-106366.

References

1. Aha, D.W., Kibler, D., Albert, M.K.: Instance-Based Learning Algorithms. Mach. Learn. 6, 37–66 (1991)
2. Brighton, H., Mellish, C.: Advances in Instance Selection for Instance-Based Learning Algorithms. Data Min. Knowl. Discov. 6, 153–172 (2002)
3. Garcia, S., Derrac, J., Cano, J., Herrera, F.: Prototype Selection for Nearest Neighbor Classification: Taxonomy and Empirical Study. IEEE Trans. Pattern Anal. Mach. Intell. 34, 417–435 (2012)
4. Olvera-López, J.A., Carrasco-Ochoa, J.A., Martínez-Trinidad, J.F., Kittler, J.: A review of instance selection methods. Artif. Intell. Rev. 34, 133–143 (2010)
5. Estabrooks, A., Jo, T., Japkowicz, N.: A Multiple Resampling Method for Learning from Imbalanced Data Sets. Computational Intelligence 20, 18–36 (2004)
6. Chawla, N.V., Japkowicz, N., Kotcz, A.: Editorial: special issue on learning from imbalanced data sets. SIGKDD Explor. Newsl. 6, 1–6 (2004)
7. Sun, Y.M., Wong, A.K.C., Kamel, M.S.: Classification of imbalance data: A review. International Journal of Pattern Recognition and Artificial Intelligence 4, 687–719 (2009)
8. García-Pedrajas, N., Romero del Castillo, J.A., Ortiz-Boyer, D.: A cooperative coevolutionary algorithm for instance selection for instance-based learning. Machine Learning 78, 381–420 (2010)
9. Batista, G.E.A.P.A., Prati, R.C., Monard, M.C.: A study of the behavior of several methods for balancing machine learning training data. SIGKDD Explor. Newsl. 6, 20–29 (2004)

10. Wilson, D.R., Martinez, T.R.: Reduction Techniques for Instance-Based Learning Algorithms. Mach. Learn. 30, 257–286 (2000)
11. Eshelman, L.J.: The CHC Adaptive Search Algorithm: How to Have Safe Search When Engaging in Nontraditional Genetic Recombination. In: Foundations of Genetic Algorithms, pp. 265–283. Morgan Kaufmann, San Francisco (1991)
12. Goldberg, D.E.: Genetic Algorithms in Search, Optimization and Machine Learning. Addison-Wesley Longman Publishing Co., Inc., Boston (1989)
13. Whitley, D.: The GENITOR algorithm and selection pressure: why rank-based allocation of reproductive trials is best. In: Proceedings of the Third International Conference on Genetic Algorithms, pp. 116–121. Morgan Kaufmann Publishers Inc. (1989)
14. Hernandez-Leal, P., Carrasco-Ochoa, J.A., Martínez-Trinidad, J.F., Olvera-Lopez, J.A.: InstanceRank based on borders for instance selection. Pattern Recogn. 46, 365–375 (2013)
15. Wilson, D.L.: Asymptotic Properties of Nearest Neighbor Rules Using Edited Data. IEEE Transactions on Systems, Man and Cybernetics 2, 408–421 (1972)
16. Cano, J.R., Herrera, F., Lozano, M.: Using evolutionary algorithms as instance selection for data reduction in KDD: an experimental study. Trans. Evol. Comp. 6, 561–575 (2003)
17. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: synthetic minority over-sampling technique. J. Artif. Int. Res. 16, 321–357 (2002)
18. Millán-Giraldo, M., García, V., Sánchez, J.S.: One-sided prototype selection on class imbalanced dissimilarity matrices. In: Gimel'farb, G., Hancock, E., Imiya, A., Kuijper, A., Kudo, M., Omachi, S., Windeatt, T., Yamada, K. (eds.) SSPR & SPR 2012. LNCS, vol. 7626, pp. 391–399. Springer, Heidelberg (2012)
19. Pérez-Rodríguez, J., de Haro-García, A., García-Pedrajas, N.: Instance selection for class imbalanced problems by means of selecting instances more than once. In: Lozano, J.A., Gámez, J.A., Moreno, J.A. (eds.) CAEPIA 2011. LNCS, vol. 7023, pp. 104–113. Springer, Heidelberg (2011)
20. Alcalá-Fdez, J., Fernandez, A., Luengo, J., Derrac, J., García, S., Sánchez, L., Herrera, F.: KEEL Data-Mining Software Tool: Data Set Repository, Integration of Algorithms and Experimental Analysis Framework. Journal of Multiple-Valued Logic and Soft Computing 17, 255–287 (2011)
21. Estabrooks, A., Jo, T., Japkowicz, N.: A Multiple Resampling Method for Learning from Imbalanced Data Sets. Computational Intelligence 20, 18–36 (2004)
22. Jesús, A.-F., Alberto, F., Julián, L., Joaquín, D., Salvador, G.: KEEL Data-Mining Software Tool: Data Set Repository, Integration of Algorithms and Experimental Analysis Framework. Multiple-Valued Logic and Soft Computing 17, 255–287 (2011)