Correlation of Resampling Methods for Contrast Pattern Based Classifiers

Octavio Loyola-González[1,2](✉), José Fco. Martínez-Trinidad[2],
Jesús Ariel Carrasco-Ochoa[2], and Milton García-Borroto[3]

[1] Centro de Bioplantas, Universidad de Ciego de Ávila, Carretera a Morón km 9,
69450 Ciego de Ávila, Cuba
octavioloyola@bioplantas.cu

[2] Instituto Nacional de Astrofísica, Óptica y Electrónica, Luis Enrique Erro No. 1,
Sta. María Tonanzintla, 72840 San Andrés Cholula, Puebla, México
{octavioloyola,fmartine,ariel}@inaoep.mx

[3] Instituto Superior Politécnico José Antonio Echeverría, Calle 114 No. 11901,
Marianao, 19390 La Habana, Cuba
mgarciab@ceis.cujae.edu.cu

Abstract. Applying resampling methods is an important approach for working with class imbalance problems. The main reason is that many classifiers are sensitive to class distribution, biasing their prediction towards the majority class. Contrast pattern based classifiers are sensitive to imbalanced databases because these classifiers commonly find several patterns of the majority class and only a few patterns (or none) of the minority class. In this paper, we present a correlation study among resampling methods for contrast pattern based classifiers. Our experiments performed over several imbalanced databases show that there is a high correlation among different resampling methods. Correlation results show that there are nine different groups with very high inner correlation and very low outer correlation. We show that most resampling methods allow improving the accuracy of the contrast pattern based classifiers.

Keywords: Supervised classification · Contrast patterns · Resampling methods · Imbalanced databases

1 Introduction

The main aim of a supervised classifier is to classify a query object using a model based on a representative sample of the problem classes. Sometimes, this model can be used to gain understanding of the problem domain or to make the problem easier to understand by experts in the application domain [13]. An important family of understandable classifiers is based on contrast patterns. Nevertheless, contrast pattern classifiers are sensitive to the class imbalance problems [18].

This work was partly supported by National Council of Science and Technology of Mexico under the project CB2008-106366 and scholarship grant 370272.

J.A. Carrasco-Ochoa et al. (Eds.): MCPR 2015, LNCS 9116, pp. 93–102, 2015.
DOI: 10.1007/978-3-319-19264-2_10

In some imbalanced real-world problems, the objects in a class can be underrepresented regarding the remaining problem classes. Oftentimes, the most important class contains significantly less objects because it could be associated to rare cases or because the data acquisition of these objects is costly [26]. This type of problems is known as the class imbalance problems.

Some contrast pattern based classifiers, which show good performance in problems with balanced classes, are degraded in class imbalance problems [16]. A common way to deal with the class imbalance problem is applying resampling methods. Resampling methods modify the dataset in order to produce a balanced class distribution. Resampling methods are more versatile than other approaches to deal with class imbalance problems because they do not depend on the learning algorithm [2].

Many comparative studies have been published about the application of resampling methods to improve the accuracy of several contrast pattern based classifiers [17–19,24,27]. Although, up to our knowledge, there is no correlation study among different resampling methods for contrast pattern classifiers.

In this paper, we present a correlation study about the effects of the most used resampling methods for improving the accuracy of a contrast pattern based classifier over several imbalanced databases. Our main goal is to offer an insight about which resampling methods have similar behavior for improving contrast pattern based classifiers. This knowledge would be helpful to simplify future research regarding resampling methods for contrast pattern based classifiers.

The rest of the paper has the following structure. Section 2 provides a brief introduction to contrast patterns. Section 3 reviews the most popular resampling methods. Section 4 presents our correlation study about the methods presented in Sect. 3, the experimental setup, and a discussion of the results. Finally, Sect. 5 provides conclusions and future work.

2 Contrast Patterns

A *pattern* is an expression defined in a certain language that describes a collection of objects. For example, a pattern that describes a set of sick plants can be expressed as: $[Necrosis = \text{“}Yes\text{”}] \wedge [StemHigh \in [0.6, 1.5]] \wedge [Leaves \leq 2]$. Then, a *contrast pattern* is a pattern appearing frequently in a class and infrequently in the remaining problem classes [30].

In some domains, contrast pattern based classifiers have shown to make consistently more accurate predictions than popular classification models like Naive Bayes, Nearest Neighbor, Bagging, Boosting, and even Support Vector Machines (SVM) [12,30].

Many algorithms have been proposed for mining contrast patterns but those based on decision trees gain special attention because they obtain a small collection of high quality patterns [11]. In this paper, we used Logical Complex Miner (LCMine) [12], a contrast pattern miner that extracts contrast patterns from a collection of diverse decision trees. Moreover, we used Classification by Aggregating Emerging Patterns (CAEP) [9] as a contrast pattern based classifier. LCMine jointly CAEP attains higher accuracies than other contrast pattern based classifiers (like SJEP [10]) and comparable accuracies to some state-of-the-art classifiers like SVM [12].

Contrast pattern based classifiers are sensitive to class imbalance problems [16,18]. The main reasons are the following: first, contrast pattern miners are based on patterns' frequency, therefore they are prone to generate more patterns for the majority class than for the minority class. Second, contrast patterns that predict the minority class are often highly specific and thus their support is very low, hence they are prone to be discarded in favor of more general contrast patterns that predict the majority class.

3 Resampling Methods

There are three approaches to deal with the class imbalance problem: *data level*, *algorithm level*, and *cost-sensitive* [16,27]. Resampling methods, belonging to the data level approach, are more versatile than the other two approaches since resampling methods can be applied independently of the supervised classifier, therefore most of the research has been done in this direction [2,17].

We can group resampling methods into three types: *oversampling* methods, which create new objects in the minority class, *undersampling* methods, which remove objects from the majority class, and *hybrid* methods that combine both oversampling and undersampling methods [5,16–18,21,23–25,28,29].

In this paper, we selected the most popular state-of-the-art resampling methods (see Table 1) including nine oversampling methods, three hybrid methods, and eight undersampling methods. All resampling methods with their default parameter values were executed using the KEEL Data-Mining software tool [4]. The main goal of our work is to offer researchers information regarding which resampling methods have similar behavior in order to simplify future research on resampling methods for contrast pattern based classifiers.

4 Correlation Study

This section presents the correlation study developed in this research. First, in Sect. 4.1, we describe the experimental setup. Then, in Sect. 4.2 we analyze the correlation obtained among the resampling methods and the base classifier selected in our study. Finally, in Sect. 4.3, we provide some discussion about the results.

4.1 Experimental Setup

For our experiments, we used 95 databases taken from the KEEL dataset repository[1] [3]. The databases have different characteristics regarding to the number of objects, number of features, and class imbalance ratio (see Table 2).

[1] http://www.keel.es/datasets.php.

Table 1. Summary of resamplig methods used in our study. No: the index associated to each resampling method in this paper; Abbreviation: the abbreviation name used in the literature and in this paper; Name and Reference: full name and reference; Type: the main approach used, Hybrid sampling (Hybrid), Oversampling (Over) or Undersampling (Under).

No	Abbreviation	Name and Reference	Type
1	SPIDER	Selective Preprocessing of Imbalanced Data [21]	Over
2	TL	Tomek's modification of Condensed Nearest Neighbor [5]	Under
3	ROS	Random oversampling [5]	Over
4	SPIDER2	Selective Preprocessing of Imbalanced Data 2 [21]	Over
5	NCL	Neighborhood Cleaning Rule [5]	Under
6	Borderline-SMOTE	Borderline Synthetic Minority Oversampling TEchnique [17]	Over
7	AHC	Aglomerative Hierarchical Clustering [7]	Over
8	SMOTE	Synthetic Minority Oversampling Technique [5]	Over
9	SMOTE-ENN	SMOTE + Edited Nearest Neighbor [5]	Hybrid
10	SMOTE-TL	SMOTE + Tomek's modification of Condensed Nearest Neighbor [5]	Hybrid
11	OSS	One Sided Selection [5]	Under
12	ADASYN	ADAptive SYNthetic Sampling [14]	Over
13	ADOMS	Adjusting the Direction Of the synthetic Minority classS examples [25]	Over
14	Safe Level SMOTE	Safe Level Synthetic Minority Oversampling TEchnique [17]	Over
15	CNN	Condensed Nearest Neighbor [5]	Under
16	CNNTL	CNN + Tomek's modification of Condensed Nearest Neighbor [5]	Under
17	RUS	Random undersampling [5]	Under
18	CPM	Class Purity Maximization [29]	Under
19	SMOTE-RSB	Hybrid Preprocessing using SMOTE and Rough Sets Theory [23]	Hybrid
20	SBC	Undersampling Based on Clustering [28]	Under

There are several measures to evaluate the performance of a classifier. Nevertheless the most used measure for class imbalance problems is the Area Under the Receiver Operating Characteristic curve (AUC) [15–17]. All our results are based on the AUC measure, which are averaged over 5-fold-cross-validation. Although the standard stratified cross-validation (SCV) is the most commonly employed method in the literature, we performed a Distribution optimally balanced-SCV (DOB-SCV) in order to avoid problems due to data distribution, especially for highly imbalanced databases [20]. All original dataset partitions with 5-fold-cross-validation used in this paper are available for downloading at the KEEL dataset repository.

We used Kendall's τ correlation, which is more closely related to the ranking task than correlations like Pearson's or Spearman's ρ [6]. Kendall's τ values range from -1 (perfect negative correlation) to 1 (perfect positive correlation).

We also used the Friedman test and the Bergmann-Hommel dynamic post-hoc procedure to compare all the results [8]. Post-hoc results will be shown using CD (*critical distance*) diagrams. In a CD diagram, the rightmost classifier is the best classifier, the position of the classifier within the segment represents its rank value, and if two or more classifiers share a thick line it means they have statistically similar behavior.

Table 2. Summary of the imbalanced databases used in our study. Name: the related name in the KEEL dataset repository; #Obj: number of objects; #Feat.: number of features; IR: class imbalance ratio [22].

Name	#Objects	#Feat.	IR	Name	#Objects	#Feat.	IR
glass1	214	9	1.82	ecoli0146vs5	280	6	13.00
ecoli0vs1	220	7	1.86	shuttlec0vsc4	1829	9	13.87
wisconsin	683	9	1.86	yeast1vs7	459	7	14.30
pima	768	8	1.87	glass4	214	9	15.46
iris0	150	4	2.00	ecoli4	336	7	15.80
glass0	214	9	2.06	pageblocks13vs4	472	10	15.86
yeast1	1484	8	2.46	abalone9vs18	731	8	16.40
haberman	306	3	2.78	dermatology6	358	34	16.90
vehicle2	846	18	2.88	zoo3	101	16	19.20
vehicle1	846	18	2.90	glass016vs5	184	9	19.44
vehicle3	846	18	2.99	shuttlec2vsc4	129	9	20.50
glass0123vs456	214	9	3.20	shuttle6vs23	230	9	22.00
vehicle0	846	18	3.25	yeast1458vs7	693	8	22.10
ecoli1	336	7	3.36	glass5	214	9	22.78
newthyroid1	215	5	5.14	yeast2vs8	482	8	23.10
newthyroid2	215	5	5.14	lymphography normalfibrosis	148	18	23.67
ecoli2	336	7	5.46	flareF	1066	11	23.79
segment0	2308	19	6.02	cargood	1728	6	24.04
glass6	214	9	6.38	carvgood	1728	6	25.58
yeast3	1484	8	8.10	krvskzeroonevsdraw	2901	6	26.63
ecoli3	336	7	8.60	krvskonevsfifteen	2244	6	27.77
pageblocks0	5472	10	8.79	yeast4	1484	8	28.10
ecoli034vs5	200	7	9.00	winequalityred4	1599	11	29.17
yeast2vs4	514	8	9.08	poker9vs7	244	10	29.50
ecoli067vs35	222	7	9.09	yeast1289vs7	947	8	30.57
ecoli0234vs5	202	7	9.10	abalone3vs11	502	8	32.47
glass015vs2	172	9	9.12	winequalitywhite9vs4	168	11	32.60
yeast0359vs78	506	8	9.12	yeast5	1484	8	32.73
yeast0256vs3789	1004	8	9.14	krvskthreevseleven	2935	6	35.23
yeast02579vs368	1004	8	9.14	winequalityred8vs6	656	11	35.44
ecoli046vs5	203	6	9.15	ecoli0137vs26	281	7	39.14
ecoli01vs235	244	7	9.17	abalone17vs78910	2338	8	39.31
ecoli0267vs35	224	7	9.18	abalone21vs8	581	8	40.50
glass04vs5	92	9	9.22	yeast6	1484	8	41.40
ecoli0346vs5	205	7	9.25	winequalitywhite3vs7	900	11	44.00
ecoli0347vs56	257	7	9.28	winequalityred8vs67	855	11	46.50
yeast05679vs4	528	8	9.35	abalone19vs10111213	1622	8	49.69
vowel0	988	13	9.98	krvskzerovseight	1460	6	53.07
ecoli067vs5	220	6	10.00	winequalitywhite39vs5	1482	11	58.28
glass016vs2	192	9	10.29	poker89vs6	1485	10	58.40
ecoli0147vs2356	336	7	10.59	shuttle2vs5	3316	9	66.67
led7digit02456789vs1	443	7	10.97	winequalityred3vs5	691	11	68.10
ecoli01vs5	240	6	11.00	abalone20vs8910	1916	8	72.69
glass06vs5	108	9	11.00	krvskzerovsfifteen	2193	6	80.22
glass0146vs2	205	9	11.06	poker89vs5	2075	10	82.00
glass2	214	9	11.59	poker8vs6	1477	10	85.88
ecoli0147vs56	332	6	12.28	abalone19	4174	8	129.44
cleveland0vs4	177	13	12.62				

4.2 Correlation Analysis

In this section, we analyze different levels of correlation over the AUC results obtained from LCMine+CAEP before and after applying resampling methods. We include, as *base* classifier, to LCMine+CAEP without applying resampling methods.

For the correlation analysis, we performed a Kendall's τ correlation based on the AUC results of the contrast pattern based classifier before and after applying resampling methods. Figure 1 shows, in grayscale, the correlation results regarding to the values obtained in the Kendall's τ correlation. Darker values are associated to correlations closer to one, while lighter values are associated to values closer to zero.

Then, using an agglomerative clustering [1], the resampling methods were clustered in nine different groups with very high inner correlation and very low outer correlation. In Fig. 1, squares with a thick line group those methods belonging to the same cluster. The groups are the following:

Group 1. {AHC, Base, Boderline-SMOTE, ROS, SPIDER, SPIDER2, TL, NCL}
Group 2. {SMOTE, SMOTE-ENN, SMOTE-TL}
Group 3. {ADASYN, ADOMS, Safe Level SMOTE}
Group 4. {CNN, CNNTL}
Group 5. {OSS}
Group 6. {RUS}
Group 7. {CPM}
Group 8. {SMOTE-RSB}
Group 9. {SBC}

Our analysis shows that resampling methods into Group 1 have high correlation with the base classifier. Group 2 contains three resampling methods that have a similar behavior, that can be explained because SMOTE-ENN and SMOTE-TL are extensions of SMOTE. Results in Group 3 have high correlation because ADOMS and Safe Level SMOTE are modifications of SMOTE; and ADASYN produces similar results than SMOTE [14]. Group 4 has two undersampling methods based on Condensed Nearest Neighbor (CNN) which presents a high correlation among them. The rest of the groups have only one resampling method. Group 9 has negative correlation (close to zero) regarding to the remaining groups.

Figure 2 shows a CD diagram with a statistical comparison of the AUC results obtained from LCMine before and after applying resampling methods. Note that Group 1 does not have statistical difference among the resampling methods into this group, with the exception of TL and NCL. Nevertheless, TL and NCL have high correlation with the base classifier, they always improved the AUC results regarding to the base classifier. Group 2 achieved the best AUC results regarding all resampling methods selected and the base classifier. Groups 3 and 5 have no statistical difference with the base classifier and they have a similar position into the Friedman ranking. Groups 4, 7, and 9 shown statistical difference with the

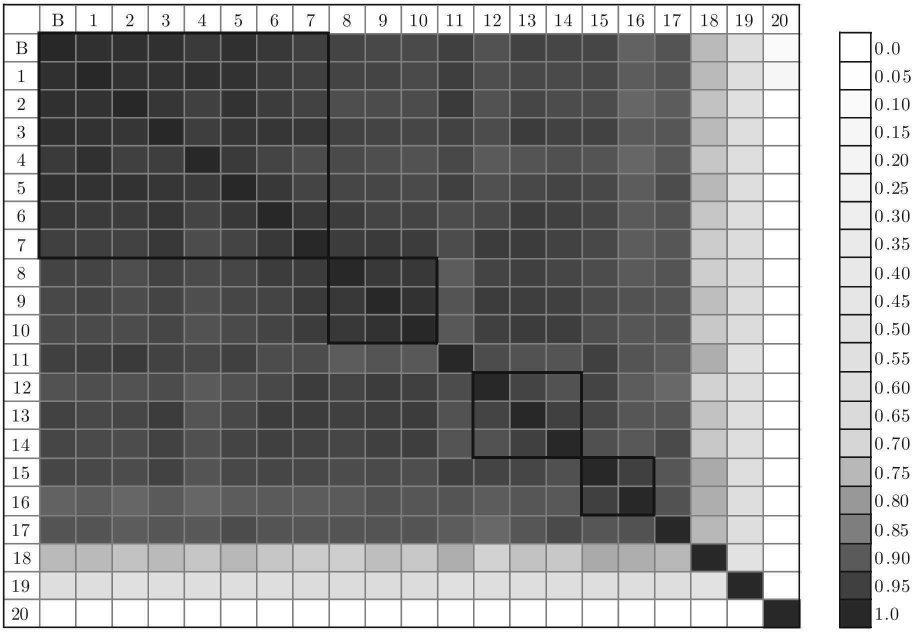

Fig. 1. Table of correlation among resampling methods and the base classifier ("B") using grayscale. The intensity of gray color is proportional to the positive correlation values.

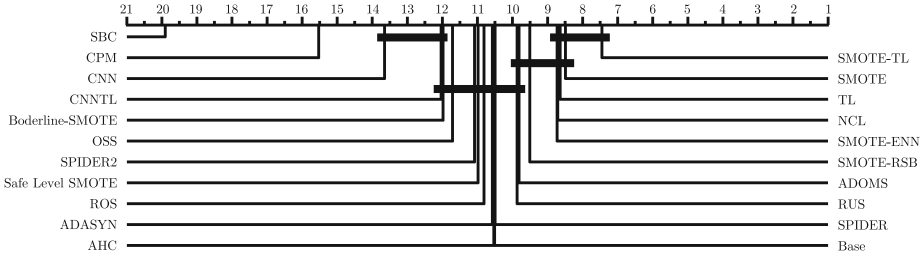

Fig. 2. CD diagram with a statistical comparison of the AUC results for the base classifier before and after using resampling methods over all the tested databases.

base classifier and they have the worst AUC results. Groups 6 and 8 have no statistical difference between them. These groups have a good position into the Friedman ranking and they have statistical difference with the base classifier.

4.3 General Concluding Remarks

The results shown in the previous section lead us to conclude that there are five resampling methods not correlated with any of the remaining 15 resampling methods or the base classifier.

Groups with more than one resampling method have high correlation among the resampling methods within each group. These groups are significant because most resampling methods contained in a group exhibit similar behavior and commonly they are extensions of the same resampling methods.

The base classifier has high correlation with resampling methods into Group 1, although only TL and NCL improved the AUC results. Groups 2 and 3 have a very high inner correlation because they contain only extensions of the SMOTE method. Resampling methods into Group 2 archived the best AUC results regarding to the remaining resampling methods. Group 4 contains only resampling methods based on Condensed Nearest Neighbor (CNN) which have bad AUC results. Groups 3, 5, and 6 have similar position into the Friedman ranking, and they have no statistical difference regarding to the base classifier. Group 8 improved the AUC results regarding the base classifier. Group 9 archived the worst AUC results regarding all resampling methods and the base classifier.

5 Conclusions and Future Work

Contrast pattern based classifiers are sensitive to the class imbalance problem. Many comparative studies have being published about resampling methods that aim to improve the accuracy in contrast pattern based classifiers. Nevertheless, no study have being published about correlations among resampling methods.

The main contribution of this paper is a correlation study among several resampling methods based on the AUC results obtained by a contrast pattern based classifier over highly imbalanced databases. This contribution would help us to simplify future research regarding resampling methods for contrast pattern based classifiers.

The experimental results show that resampling methods in Group 1 have high correlation with the base classifier, although TL and NCL improved significantly the AUC results. Group 2 archived the best AUC results regarding to the remaining groups including the base classifier. Groups 3, 5, and 6 have no statistical difference regarding to the base classifier. Groups 4, 7, and 9 have the worst AUC results. Groups 6 and 8 improved the AUC results regarding the base classifier. Finally, although the base classifier has a high correlation with some resampling methods, most of resampling methods improve the AUC results for the contrast pattern based classifier.

As future work, we plan to investigate about the influence of the imbalance ratio on these results. This way, we could suggest what resampling method would perform better for a given imbalanced dataset.

Acknowledgment. The authors want to thank Rebekah Hosse Clark for her valuable contributions improving the grammar and style of this paper.

References

1. Aggarwal, C.C., Reddy, C.K.: Data Clustering: Algorithms and Applications, 1st edn. Chapman & Hall/CRC, Boca Raton (2013)
2. Albisua, I., Arbelaitz, O., Gurrutxaga, I., Lasarguren, A., Muguerza, J., Pérez, J.: The quest for the optimal class distribution: an approach for enhancing the effectiveness of learning via resampling methods for imbalanced data sets. Prog. Artif. Intell. **2**(1), 45–63 (2013)
3. Alcalá-Fdez, J., Fernández, A., Luengo, J., Derrac, J., García, S.: KEEL data-mining software tool: data set repository, integration of algorithms and experimental analysis framework. J. Multiple-Valued Logic Soft Comput. **17**(2–3), 255–287 (2011)
4. Alcalá-Fdez, J., Sánchez, L., García, S., del Jesús, M.J., Ventura, S., Garrell, J.M., Otero, J., Romero, C., Bacardit, J., Rivas, V.M., Fernández, J.C., Herrera, F.: KEEL: a software tool to assess evolutionary algorithms for data mining problems. Soft Comput. **13**(3), 307–318 (2009)
5. 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**(1), 20–29 (2004)
6. Bruning, J.L., Kintz, B.L.: Computational Handbook of Statistics, 4th edn. Longman, New York (1997)
7. Cohen, G., Hilario, M., Sax, H., Hugonnet, S., Geissbuhler, A.: Learning from imbalanced data in surveillance of nosocomial infection. Artif. Intell. Med. **37**, 7–18 (2006)
8. Demšar, J.: Statistical comparisons of classifiers over multiple data sets. J. Mach. Learn. Res. **7**, 1–30 (2006)
9. Dong, G., Zhang, X., Wong, L., Li, J.: CAEP: classification by aggregating emerging patterns. In: Arikawa, S., Furukawa, K. (eds.) DS 1999. LNCS (LNAI), vol. 1721, pp. 30–42. Springer, Heidelberg (1999)
10. Fan, H., Ramamohanarao, K.: Fast discovery and the generalization of strong jumping emerging patterns for building compact and accurate classifiers. IEEE Trans. Knowl. Data Eng. **18**(6), 721–737 (2006)
11. García-Borroto, M., Martínez-Trinidad, J.F., Carrasco-Ochoa, J.A.: Finding the best diversity generation procedures for mining contrast patterns. Expert Syst. Appl. **42**(11), 4859–4866 (2015)
12. García-Borroto, M., Martínez-Trinidad, J.F., Carrasco-Ochoa, J.A., Medina-Pérez, M.A., Ruiz-Shulcloper, J.: LCMine: an efficient algorithm for mining discriminative regularities and its application in supervised classification. Pattern Recogn. **43**(9), 3025–3034 (2010)
13. García-Borroto, M., Martínez-Trinidad, J., Carrasco-Ochoa, J.: A survey of emerging patterns for supervised classification. Artif. Intell. Rev. **42**(4), 705–721 (2014)
14. He, H., Bai, Y., Garcia, E.A., Li, S.: ADASYN: adaptive synthetic sampling approach for imbalanced learning. In: 2008 International Joint Conference on Neural Networks (IJCNN 2008), pp. 1322–1328 (2008)
15. Huang, J., Ling, C.X.: Using AUC and accuracy in evaluating learning algorithms. IEEE Trans. Knowl. Data Eng. **17**(3), 299–310 (2005)
16. López, V., Fernández, A., García, S., Palade, V., Herrera, F.: An insight into classification with imbalanced data: empirical results and current trends on using data intrinsic characteristics. Inf. Sci. **250**, 113–141 (2013)

17. López, V., Triguero, I., Carmona, C.J., García, S., Herrera, F.: Addressing imbalanced classification with instance generation techniques: IPADE-ID. Neurocomputing **126**, 15–28 (2014)
18. Loyola-González, O., García-Borroto, M., Medina-Pérez, M.A., Martínez-Trinidad, J.F., Carrasco-Ochoa, J.A., De Ita, G.: An empirical study of oversampling and undersampling methods for LCMine an emerging pattern based classifier. In: Carrasco-Ochoa, J.A., Martínez-Trinidad, J.F., Rodríguez, J.S., di Baja, G.S. (eds.) MCPR 2012. LNCS, vol. 7914, pp. 264–273. Springer, Heidelberg (2013)
19. Menardi, G., Torelli, N.: Training and assessing classification rules with imbalanced data. Data Min. Knowl. Disc. **28**(1), 92–122 (2014)
20. Moreno-Torres, J.G., Saez, J.A., Herrera, F.: Study on the impact of partition-induced dataset shift on k-Fold cross-validation. IEEE Trans. Neural Netw. Learn. Syst. **23**(8), 1304–1312 (2012)
21. Napierała, K., Stefanowski, J., Wilk, S.: Learning from imbalanced data in presence of noisy and borderline examples. In: Szczuka, M., Kryszkiewicz, M., Ramanna, S., Jensen, R., Hu, Q. (eds.) RSCTC 2010. LNCS, vol. 6086, pp. 158–167. Springer, Heidelberg (2010)
22. Orriols-Puig, A., Bernadó-Mansilla, E.: Evolutionary rule-based systems for imbalanced data sets. Soft. Comput. **13**(3), 213–225 (2009)
23. Ramentol, E., Caballero, Y., Bello, R., Herrera, F.: SMOTE-RSB*: a hybrid preprocessing approach based on oversampling and undersampling for high imbalanced data-sets using SMOTE and rough sets theory. Knowl. Inf. Syst. **33**(2), 245–265 (2011)
24. Sáez, J.A., Luengo, J., Stefanowski, J., Herrera, F.: Managing borderline and noisy examples in imbalanced classification by combining SMOTE with ensemble filtering. In: Corchado, E., Lozano, J.A., Quintián, H., Yin, H. (eds.) IDEAL 2014. LNCS, vol. 8669, pp. 61–68. Springer, Heidelberg (2014)
25. Tang, S., Chen, S.: The Generation mechanism of synthetic minority class examples. In: 5th International Conference on Information Technology and Applications in Biomedicine (ITAB 2008), pp. 444–447 (2008)
26. Weiss, G., Tian, Y.: Maximizing classifier utility when there are data acquisition and modeling costs. Data Min. Knowl. Disc. **17**(2), 253–282 (2008)
27. Yap, B., Rani, K., Rahman, H., Fong, S., Khairudin, Z., Abdullah, N.: An application of oversampling, undersampling, bagging and boosting in handling imbalanced datasets. In: Herawan, T., Deris, M.M., Abawajy, J. (eds.) Proceedings of the First International Conference on Advanced Data and Information Engineering (DaEng 2013). LNEE, vol. 285, pp. 13–22. Springer, Heidelberg (2014)
28. Yen, S.-J., Lee, Y.-S.: Under-sampling approaches for improving prediction of the minority class in an imbalanced dataset. In: Huang, D.-S., Li, K., Irwin, K. (eds.) ICIC 2006. LNCIS, vol. 344, pp. 731–740. Springer, Heidelberg (2006)
29. Yoon, K., Kwek, S.: An unsupervised learning approach to resolving the data imbalanced issue in supervised learning problems in functional genomics. In: 5th International Conference on Hybrid Intelligent Systems (HIS 2005), pp. 303–308 (2005)
30. Zhang, X., Dong, G.: Overview and analysis of contrast pattern based classification. In: Dong, G., Bailey, J. (eds.) Contrast Data Mining: Concepts, Algorithms, and Applications. Data Mining and Knowledge Discovery Series, vol. 11, pp. 151–170. Chapman & Hall/CRC, Boca Raton (2012)