Radial-Based Approach to Imbalanced Data Oversampling

Michał Koziarski[1(✉)], Bartosz Krawczyk[2], and Michał Woźniak[1]

[1] Department of Systems and Computer Networks,
Wrocław University of Science and Technology, Wrocław, Poland
{michal.koziarski,michal.wozniak}@pwr.edu.pl
[2] Department of Computer Science, Virginia Commonwealth University,
Richmond, VA, USA
bkrawczyk@vcu.edu

Abstract. The difficulty of the many practical decision problem lies in the nature of analyzed data. One of the most important real data characteristic is imbalance among examples from different classes. Despite more than two decades of research, imbalanced data classification is still one of the vital challenges to be addressed. The traditional classification algorithms display strongly biased performance on imbalanced datasets. One of the most popular way to deal with such a problem is to modify the learning set to decrease disproportion between objects from different classes using over- or undersampling approaches. In this work a novel preprocessing technique for imbalanced datasets is presented, which takes into consideration the mutual density class distribution. The proposed approach has been evaluated on the basis of the computer experiments carried out on the benchmark datasets. Their results seem to confirm the usefulness of the proposed concept in comparison to the state-of-art methods.

Keywords: Machine learning · Classification · Imbalanced data · Oversampling · Radial basis functions

1 Introduction

Most of commonly used machine learning algorithms work under an underlying assumption that classes have roughly equal number of instances in the training set. However, in many real-life scenarios it is difficult, or even impossible, to gather representative collections of instances of similar size from all of classes [11]. We may deal with a predominant group of objects being abundant and easy to gather, and with a significantly smaller group to instances of which we have an limited access [1]. Therefore, we need to create an efficient learning system using the imperfect data at our disposal. Such an imbalanced distribution will significantly affect the training process of a classifier, as it is usually guided by predictive accuracy. This solution assumes uniform importance of all training instances, thus leading to a classifier being biased towards the majority class.

F.J. Martínez de Pisón et al. (Eds.): HAIS 2017, LNAI 10334, pp. 318–327, 2017.
DOI: 10.1007/978-3-319-59650-1_27

When concentrating on more abundant case classifier is more likely to obtain higher accuracy rates, thus making such a model preferable from the canonical point of view. However, the minority class is usually the more important one and thus we want to maximize the predictive performance on it. This has lead to development of a number of approaches for balancing the classes or alleviating the bias during training step [4]. Let us now review quickly three most important groups of methods in this domain.

Preprocessing approaches are applied directly on the training set, before a classifier is being trained [14]. They aim at manipulating instances in such a way that will lead to obtaining a balanced dataset. One may achieve this by either undersampling the majority class, or oversampling the minority one. Randomized methods are the most basic ones, characterized by a low computational complexity and ease of usage. However, they may actually have a harmful effect on the dataset. Random undersampling may lead to discarding instances that are essential to forming correct class boundary or lie in specific subregions of the target class. Random oversampling may multiply noisy or corrupted instances, thus shifting the actual class distribution. Therefore, in recent years one may see significant developments in this area that propose a more guided approach for balancing classes.

Algorithm-level approaches aim at modifying the classifier learning procedure in order to make it skew-insensitive. This requires an in-depth understanding of the modified methods, as well as of the actual learning difficulty that causes the poor performance on minority class. Here, cost-sensitive approaches are popular, as they allow to easily modify any learning method by adding a separate misclassification penalty for each class [8]. This should improve minority class recognition, as classifier will be much more penalized for misclassification of minority instance. Another potential solution include usage of one-class classifiers [3]. Here, we create a data description of the target class (one selected by the user) and treat the remaining one as outliers. While we sacrifice knowledge about one of the classes, we gain a skew-insensitive classifier that captures unique properties of its target.

Hybrid solutions use advantages of the mentioned approaches and combine them with other methodologies, mainly ensemble learners [16]. They take advantage of increased predictive power, diversity and ability to capture complex data offered by combined classifiers and augment it with tackling imbalance at the level of each classifier. Popular approaches include combination of Bagging or Boosting with preprocessing.

Despite these developments there still exists a need for introducing more efficient and robust methods for learning from imbalanced data. Especially interesting recent direction is taking into account the properties of individual instances in the minority class.

In this paper, we introduce a novel oversampling technique that uses radial functions for estimating the potential of instances. We propose to use them to model mutual class distributions and analyze the learning difficulty associated with each instance. Our solution is able to select which objects should be subject

to oversampling, instead of blindingly using all of them. By analyzing the differences in potential at a given point, we are able to predefine the nature of minority class instances use it to guide the artificial instance injection procedure. This allows for a more meaningful capturing of the minority class underlying distribution. Additionally, as our solution does not rely on neighborhood calculation, it is suitable for applications in high-dimensional datasets. Experimental study conducted on a number of benchmarks prove that the proposed radial-based oversampling is able to return satisfactory performance.

2 Radial-Based Approach to Oversampling

By far the most prevalent approach to imbalanced data oversampling is Synthetic Minority Oversampling Technique (SMOTE) [6] algorithm and its numerous extensions [5,7,9,12]. However, while widely used and empirically tested, SMOTE and its derivatives are not devoid of weaknesses. In the remainder of this section we discuss possible shortcomings of neighborhood-based oversampling strategies. Afterwards, we propose an alternative approach that aims at mitigating described issues. We describe how radial basis functions can be used to estimate mutual class density. Finally, we propose a novel algorithm, Radial-Based Oversampling, which takes advantage of this density estimation approach to guide the oversampling process in an informed manner.

2.1 Shortcomings of Neighborhood-Based Approaches

Conceptually simplest approach to imbalanced data oversampling is duplicating existing instances randomly, up to the point of achieving balanced class distributions. However, it leads to minority class distribution being highly focused in a small area, in which the original observations were present. Because of that, learning on data modified in such manner is prone to overfitting. SMOTE algorithm was designed specifically do address this issue. Instead of duplicating existing instances, SMOTE and its derivatives are based on creating new, synthetic samples. This family of methods relies on finding nearest, same-class neighbors of a given minority instance. Afterwards, new samples are being generated between the given target and one of its neighbors. This approach can be interpreted as finding the regions in which new samples can be synthesized, and these regions are lines connecting nearest minority neighbors. Since synthetic observations are spread out, SMOTE is less prone to overfitting than random oversampling. Furthermore, new objects can be synthesized in regions previously not containing minority samples. Because of that, this approach tends to move the decision border in favor of minority class, a behavior often desirable in case of highly imbalanced data.

The underlying assumption being made in SMOTE is that the regions between nearest minority neighbors are suitable for generating new instances. While often being the case, this assumption does not always hold true. An example of data distribution not meeting this requirement is presented in Fig. 1.

In presented case minority instances form several small clusters, divided by a large cluster of majority objects. Nearest minority neighborhood is therefore spread apart, which leads to generation of synthetic samples overlapping the majority cluster. This issue is so prevalent that it was addressed with several post-oversampling cleaning strategies, most notable being Tomek links [13] and Edited Nearest Neighbor Rule [15]. However, even applying such post-processing is not always sufficient to properly clean the resulting distribution. Furthermore, since sizes of minority clusters vary, it is not clear what size of neighborhood k should be chosen. Even the choice of $k = 1$ would not, however, be sufficient to fully remedy the issue of overlapping the majority cluster. To make the matters worse, it cannot be picked dynamically for different minority instances: SMOTE algorithm requires single choice of k for whole object space. Both of the mentioned issues, that is: synthetic samples overlapping existing majority instances and inability to pick number of neighbors dynamically, are deeply rooted in the fact that SMOTE does not take into the account presence of majority instances. Regions in which synthetic samples are generated are based solely on the minority class distribution, and this information is simply not sufficient in all cases.

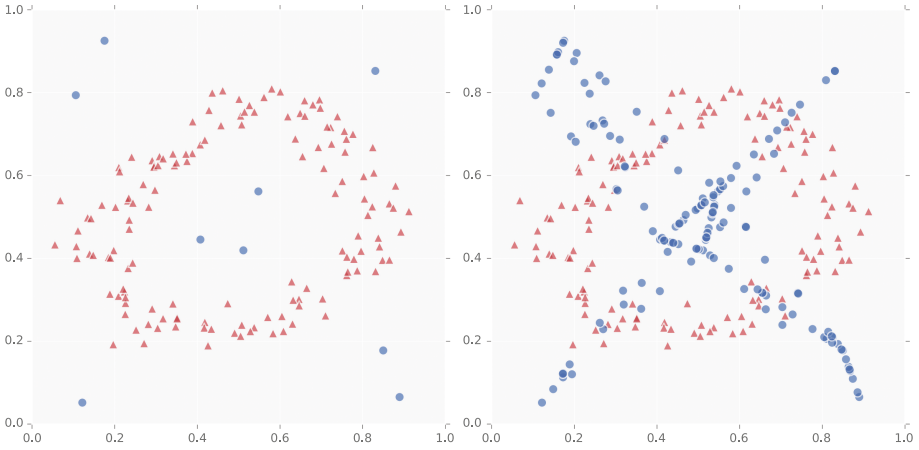

Fig. 1. Possible difficult case for SMOTE before (on the left) and after (on the right) generating synthetic samples. Due to varying sizes of minority objects clusters, it is not clear what number of neighbors k should be chosen. Even choosing $k = 1$ would lead to generating synthetic samples overlapping cluster of majority objects.

2.2 Estimating Difficulty with Radial Basis Functions

Instead of relying on nearest neighbors in process of generating new samples, in this paper we will investigate the possibility of density-based approach. Intuitively, our goal is similar to SMOTE techniques: we try to find the regions, in which generating synthetic samples is justified. However, contrary to SMOTE,

we will take into account placement of majority objects. By doing so, our hope is to reduce the amount of synthetic samples placed in regions densely packed with majority instances.

To this end we will employ radial basis functions (RBFs). RBFs are real-valued functions, value of which depends on the distance of the point from the origin. Common example of such function is Gaussian RBF. Given distance r and parameter ε, Gaussian RBF can be defined as

$$\phi(r) = e^{-(\varepsilon r)^2}. \tag{1}$$

To estimate the mutual class density in a given point in space, also referred to as a potential, we will sum the values of RBFs for all the instances, with sign determined by the class of the particular instance. Throughout this paper we will use a convention that for majority objects value of RBF will be added, whereas for the minority objects it will be subtracted. Observing high potential in a given point will therefore correspond to high confidence in the fact that it belongs to the majority class. Furthermore, observing minority objects with high potential might indicate that they will be hard to classify correctly, since it is likely to be surrounded by multiple majority instances.

2.3 Generating Synthetic Samples in a Guided Manner

Mutual class density estimated with RBFs can later be used to guide the process of synthetic samples generation. In principle, it could be used in various ways. For instance, potential could indicate difficulty associated with observation, since minority objects with high associated potential are likely to be surrounded by a large number of majority instances. Such difficult examples can be prioritized during oversampling, similar to ADASYN [10]. Instead, in this paper we will focus on finding regions, in which generation of synthetic samples should be conducted.

We will focus on regions with high potential, lying in close proximity to existing minority instances. To make the approach computationally feasible, we will employ modified hill climbing procedure to maximize the potential of the synthetic samples. Optimization will start at a position of randomly chosen, existing minority instance. Whole procedure will last limited number of steps to prevent placing new instances too deeply into the majority objects clusters. Finally, to spread synthetic samples more evenly, we will allow optimization procedure to stop early with a small probability. Pseudocode of the final algorithm has been presented in Algorithm 1. An illustration of both confidence estimation with radial basis function and the conducted oversampling has been presented in Fig. 2.

3 Experimental Study

Experimental investigations, backed up with statistical analysis of the results, were conducted to evaluate the practical usefulness of the proposed oversampling

Algorithm 1. Radial-Based Oversampling algorithm

Input: collections of majority objects M and minority objects m
Parameters: spread of radial basis function γ, optimization *step size*, number of *iterations* per synthetic sample, probability of early stopping p
Output: collection of synthetic minority objects S

function RBO(M, m, γ, *step size*, *iterations*, p):
initialize empty collection S
while $|m| + |S| < |M|$ **do**
  $point \leftarrow$ randomly chosen object from m
  for $i \leftarrow 1$ **to** *iterations* **do**
    break with probability p
    $translated \leftarrow point$ translated by *step size* in random direction
    if potential($translated$, M, m, γ) > potential($point$, M, m, γ) **then**
      $point \leftarrow translated$
    end if
  end for
  add $point$ to S
end while
return S

function potential($point$, M, m, γ):
$result \leftarrow 0$
for all majority points M_i **do**
  $result \leftarrow result + e^{-(\frac{\|M_i - point\|_1}{\gamma})^2}$
end for
for all minority points m_i **do**
  $result \leftarrow result - e^{-(\frac{\|m_i - point\|_1}{\gamma})^2}$
end for
return $result$

strategy. In the remainder of this section we describe set-up of the study, present obtained results and discuss achieved outcomes.

3.1 Set-up

Proposed strategy of dealing with data imbalance, Radial-Based Oversampling (RBO), has been compared with two state-of-the-art oversampling algorithms: SMOTE [6] and ADASYN [10]. Additionally, the baseline case was considered, in which no resampling was applied prior to classification. To assess the robustness to the choice of learner, several classification algorithms were considered, namely: k-nearest neighbors (k-NN), support vector machine with radial basis function kernel (SVM) and CART decision tree (CART).

Following parameters were used in combination with the RBO method: γ coefficient, corresponding to the spread of radial basis function, was set to 0.05. Step size used during hill climbing optimization was set to 0.001. Number of

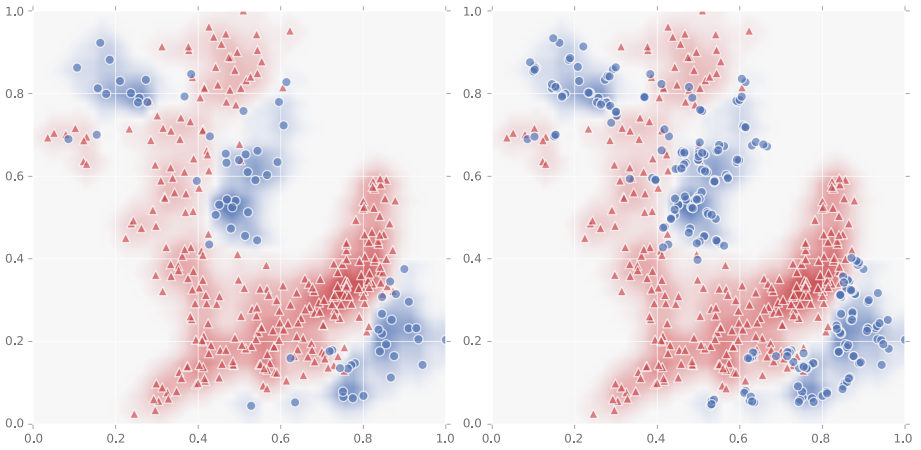

Fig. 2. On the left: confidence estimation conducted using radial basis functions. Of particular interest are minority objects (in blue) lying in regions of high confidence in majority class (in red). On the right: modified data distribution with synthetically generated samples. (Color figure online)

iterations per synthetic sample was set to 500. Finally, the probability of stopping the optimization early was set to 0.02. Meanwhile, both baseline methods, SMOTE and ADASYN, used 5 nearest neighbors to construct the synthetic samples. In all cases new samples were generated up to the point of balancing the distributions.

Evaluation was performed on 10 datasets taken from KEEL [2] repository. They were chosen to cover varying levels of imbalance and their details are presented in Table 1. During the evaluation datasets were partitioned using a 5-folds stratified cross validation. Prior to classification data was normalized to range from 0 to 1. No further preprocessing was applied.

3.2 Results

In order to properly analyze the behavior of examined methods on imbalanced data, several metrics were considered: accuracy, precision, recall, F-measure and geometric mean (G-mean). To assure statistical validity of the results Friedman test was conducted and average rankings on all the datasets were reported. They were presented in Table 2. Additionally, detailed results of F-measure for all datasets were presented in Table 3.

Oversampling strategy proposed in this paper achieved performance comparable to SMOTE method. Using ADASYN led to best results on tested datasets as far as recall was considered, at the cost of slightly lower precision. These trends were alike for all considered classifiers, suggesting robustness to the choice of base learner. Overall, performance of all three resampling algorithms was similar. This leads us to believe that further work on radial-based oversampling strategies is a promising research direction.

Table 1. Details of datasets used during the experimental study.

No	Name	IR	Features	Samples
1	glass1	1.82	9	214
2	wisconsin	1.86	7	220
3	yeast1	2.46	8	1484
4	vehicle0	3.25	18	846
5	ecoli1	3.36	7	336
6	new-thyroid1	5.14	5	215
7	segment0	6.02	19	2308
8	page-blocks0	8.79	10	5472
9	vowel0	9.98	13	988
10	abalone19	16.4	8	731

Table 2. Average rankings of various performance measures, computed for k-NN/ SVM/CART classifiers. Method proposed in this paper, Radial-Based Oversampling (RBO), was compared with SMOTE and ADASYN algorithms, as well as the baseline case in which no oversampling was applied.

Measure	None	SMOTE	ADASYN	RBO
Accuracy	**1.8**/**2.3**/**1.8**	2.6/2.6/2.8	3.1/2.4/2.7	2.4/2.6/2.6
Precision	**1.4**/**2.2**/**1.7**	2.5/2.4/2.9	3.2/2.8/2.8	2.7/2.6/2.6
Recall	4/4/3.8	2.3/2.2/2.2	**1.5**/**1.4**/**1.4**	2.1/2.3/2.5
F-measure	2.7/4/2.6	**2.4**/**2**/2.7	2.5/**2**/**2.1**	**2.4**/**2**/2.6
G-mean	4/4/3.5	2/2.3/2.4	**1.9**/**1.6**/**1.9**	2.1/2.1/2.2

Table 3. Values of F-measure achieved on specific datasets, computed for k-NN/ SVM/CART classifiers.

Dataset	None	SMOTE	ADASYN	RBO
1	0.67/0.00/0.66	0.70/0.56/0.64	**0.72**/0.53/**0.67**	**0.72**/**0.57**/**0.67**
2	**0.96**/**0.96**/0.90	**0.96**/**0.96**/0.92	**0.96**/**0.96**/**0.93**	**0.96**/**0.96**/0.92
3	0.49/0.11/**0.52**	**0.57**/0.58/**0.52**	0.56/0.58/0.50	0.54/**0.59**/**0.52**
4	**0.86**/0.00/**0.89**	0.84/0.68/0.86	0.85/**0.71**/**0.89**	0.85/0.69/0.85
5	**0.82**/0.72/0.74	0.78/0.75/0.74	0.78/**0.76**/0.74	0.77/0.75/**0.77**
6	0.89/0.62/0.90	**0.96**/**0.93**/**0.92**	0.94/**0.93**/0.90	0.95/**0.93**/0.89
7	**0.98**/0.57/**0.98**	0.97/**0.89**/**0.98**	0.97/0.66/0.97	0.97/0.88/**0.98**
8	**0.77**/0.55/**0.84**	0.76/**0.67**/0.82	0.76/0.61/0.83	**0.77**/**0.67**/0.80
9	0.97/0.00/0.92	**0.99**/0.76/0.90	**0.99**/**0.80**/**0.93**	**0.99**/0.76/0.90
10	0.00/0.00/0.00	**0.04**/**0.04**/**0.03**	**0.04**/**0.04**/**0.03**	0.03/0.03/0.02

4 Conclusions and Future Directions

In this paper we proposed a novel approach to imbalanced data oversampling. It relied on using radial basis functions to estimate classification difficulty of minority objects. An inspiration for it lied in addressing possible shortcomings of existing oversampling strategies, which we described in this paper. Results of conducted experimental evaluation seem to confirm possible usefulness of the proposed approach.

Proposed method, while capable of achieving performance comparable to other state-of-the-art oversampling algorithms, is relatively simple and can be improved upon. First of all, it is not clear whether maximization of potential is the optimal choice. Usually it corresponds to generating new minority objects in areas of the lowest certainty. In some cases it might be preferable to generate safer objects instead, especially so if preserving high precision is an important factor. Secondly, results of experimental study conducted in this paper indicate that oversampling with ADASYN leads to better results than SMOTE. This corresponds to focusing on difficult minority objects while generating synthetic samples. Incorporating such mechanism into the Radial-Based Oversampling might therefore improve performance of the method. Thirdly, in the proposed approach we considered only mutual class density, difference between the potential of majority and minority classes. In some cases it might be insufficient to describe the difficulty of classification in a particular point in space. For instance, neighborhood of an object could be densely packed with both minority and majority instances. Opposite potentials could cancel themselves out, leading to the same final value as in the case of a single object with no nearby observations. To mitigate this issue, probability distributions of individual classes could be incorporated into the oversampling procedure. Finally, in presented form radial-based approach to oversampling is computationally expensive, since at every iteration potential is computed based on all existing objects. However, since influence of far-away points is usually negligible, this operation could be significantly sped up by focusing only on nearest instances.

Acknowledgements. This work was supported by the Polish National Science Center under the grant no. UMO-2015/19/B/ST6/01597 as well as the PLGrid Infrastructure.

References

1. Ahmed, F., Samorani, M., Bellinger, C., Zaïane, O.R.: Advantage of integration in big data: Feature generation in multi-relational databases for imbalanced learning. In: 2016 IEEE International Conference on Big Data, BigData 2016, Washington DC, USA, 5–8 December 2016, pp. 532–539 (2016)
2. Alcalá, J., Fernández, 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. J. Multiple-Valued Logic Soft Comput. **17**(2–3), 255–287 (2010)

3. Bellinger, C., Sharma, S., Japkowicz, N.: One-class versus binary classification: Which and when? In: 11th International Conference on Machine Learning and Applications, ICMLA, Boca Raton, FL, USA, 12–15 December 2012, vol. 2. pp. 102–106 (2012)
4. Branco, P., Torgo, L., Ribeiro, R.P.: A survey of predictive modeling on imbalanced domains. ACM Comput. Surv. **49**(2), 31:1–31:50 (2016)
5. Bunkhumpornpat, C., Sinapiromsaran, K., Lursinsap, C.: Safe-level-SMOTE: safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem. In: Theeramunkong, T., Kijsirikul, B., Cercone, N., Ho, T.-B. (eds.) PAKDD 2009. LNCS, vol. 5476, pp. 475–482. Springer, Heidelberg (2009). doi:10.1007/978-3-642-01307-2_43
6. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: synthetic minority over-sampling technique. J. Artif. Intell. Res. **16**, 321–357 (2002)
7. Chawla, N.V., Lazarevic, A., Hall, L.O., Bowyer, K.W.: SMOTEBoost: improving prediction of the minority class in boosting. In: Lavrač, N., Gamberger, D., Todorovski, L., Blockeel, H. (eds.) PKDD 2003. LNCS, vol. 2838, pp. 107–119. Springer, Heidelberg (2003). doi:10.1007/978-3-540-39804-2_12
8. Domingos, P.M.: Metacost: a general method for making classifiers cost-sensitive. In: Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, CA, USA, 15–18 August 1999, pp. 155–164 (1999)
9. Han, H., Wang, W.-Y., Mao, B.-H.: Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. In: Huang, D.-S., Zhang, X.-P., Huang, G.-B. (eds.) ICIC 2005. LNCS, vol. 3644, pp. 878–887. Springer, Heidelberg (2005). doi:10.1007/11538059_91
10. He, H., Bai, Y., Garcia, E.A., Li, S.: ADASYN: adaptive synthetic sampling approach for imbalanced learning. In: IEEE International Joint Conference on Neural Networks, IJCNN 2008. (IEEE World Congress on Computational Intelligence), pp. 1322–1328. IEEE (2008)
11. Porwik, P., Doroz, R., Orczyk, T.: Signatures verification based on PNN classifier optimised by PSO algorithm. Pattern Recogn. **60**, 998–1014 (2016)
12. 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 (2012)
13. Tomek, I.: Two modifications of CNN. IEEE Trans. Syst. Man Cybern. **6**(11), 769–772 (1976)
14. Triguero, I., Galar, M., Merino, D., Maillo, J., Bustince, H., Herrera, F.: Evolutionary undersampling for extremely imbalanced big data classification under apache spark. In: IEEE Congress on Evolutionary Computation, CEC 2016, Vancouver, BC, Canada, 24–29 July 2016, pp. 640–647 (2016)
15. Wilson, D.L.: Asymptotic properties of nearest neighbor rules using edited data. IEEE Trans. Syst. Man Cybern. **2**(3), 408–421 (1972)
16. Wozniak, M., Graña, M., Corchado, E.: A survey of multiple classifier systems as hybrid systems. Inf. Fusion **16**, 3–17 (2014)