Diversity Analysis on Imbalanced Data Using Neighbourhood and Roughly Balanced Bagging Ensembles

Jerzy Błaszczyński and Mateusz Lango(✉)

Institute of Computing Science, Poznań University of Technology, 60-965 Poznań, Poland
{jurek.blaszczynski,mateusz.lango}@cs.put.poznan.pl

Abstract. Bagging ensembles proved to work better than boosting for class imbalanced and noisy data. We compare performance and diversity of the two best performing, in this setting, bagging ensembles: Roughly Balanced Bagging (RBBag) and Neighbourhood Balanced Bagging (NBBag). We show that NBBag makes correct prediction on a higher than RBBag number of difficult to learn minority examples. Then we detect a trade-off between correct recognition of difficult minority examples and majority examples, which makes RBBag better in some cases. We also introduce a simple but effective technique to select parameters for NBBag.

Keywords: Class imbalance · Ensembles · Roughly balanced bagging · Neighbourhood balanced bagging · Diversity · Parametrization

1 Introduction

One of the most important challenges for supervised machine learning is learning from imbalanced data [14]. The data is imbalanced when one of the classes has small number of examples (minority class) in comparison to other classes in the data set (majority classes). Such situation occurs in many important applications e.g. in fraud detection, medical problems, etc. Due to the importance of the problem, many methods to counter class imbalance has been proposed. Following [9] we divide them into two categories: data-level and algorithm-level approaches. By data-level approaches we understand techniques which apply data preprocessing methods, such as re-sampling, to improve classification of imbalanced data without changing the learning algorithm. Typically, these techniques focus on switching class distribution to a more balanced one. The other group of approaches modifies existing algorithms to better model minority class distribution. To this category we assign also specialized ensembles which are usually modifications of bagging or boosting; see their review in [3].

Experiments [6,10] have shown that bagging ensembles work better than extensions of boosting, especially on noisy data sets. Further studies [1,6] demonstrated that Roughly Balanced Bagging (RBBag), which applies specific random under-sampling to create bootstraps, achieves the best results on G-mean

L. Rutkowski et al. (Eds.): ICAISC 2016, Part I, LNAI 9692, pp. 552–562, 2016.
DOI: 10.1007/978-3-319-39378-0_47

and AUC measures among extensions of bagging. However, in the recent work Błaszczyński and Stefanowski have proposed Neighbourhood Balanced Bagging (NBBag), which modifies bootstrap sampling by weighting examples [2]. The weight of an example depends from the class label and the number of examples in the example neighbourhood which belong to the opposite class. The impact of neighbourhood on weights is controlled by parameters: size of the neighbourhood and a scaling factor. It has been shown that NBBag achieves competitive results on G-mean and better results on sensitivity measure than RBBag.

Besides results on G-mean or sensitivity metrics it is unknown how data difficulty factors impact model learned by different specialized extensions of bagging for class imbalance. Since NBBag proved to be better than RBBag on sensitivity measure, it is particularly interesting to analyze on which types of minority examples it performs better then RBBag. Another important issue when comparing two ensembles is the diversity of theirs base classifiers. To the best of our knowledge the diversity of NBBag was never investigated and experimental studies measuring diversity in the context of the minority class are very limited. Furthermore, the authors of NBBag noticed that the results of the classifier significantly depend on the values of parameters [2], which need to be selected after a careful analysis of results produced with different settings. Moreover, they advocate that the best set-up should be elected for a particular data set.

To address these issues, in this paper we propose a method intended to automatically parametrize Neighbourhood Balanced Bagging for imbalanced data sets. We also experimentally study abilities of NBBag to deal with different types of difficult distributions of the minority class and we compare this abilities to its major competitor: RBBag. Additionally, we calculate diversity measure of NBBag and compare the results to the reference algorithms.

2 Related Works

The data set is called imbalanced when one class has substantially less examples then the others. Although the problem of class imbalance relates also to multi-class classification in the majority of the research - and also in this paper - only binary classification is considered. In this case we can define statistics which measure the level of class imbalance: global imbalanced ratio $IR = \frac{N_-}{N_+}$ where N_- and N_+ are the number of majority and minority examples, respectively.

Imbalanced data is causing many problems for standard classifiers. Nevertheless, it has been noticed that the global imbalance ratio is not the only or even not the most important factor which makes learning difficult. Other data difficulty factors such as class overlapping, small disjunct or lack of representativeness significantly deteriorate the quality of induced model even on exactly balanced data. However, adding class imbalance to a data which suffers from these difficulty factors creates a real challenge for machine learning algorithms. It has been shown that in the imbalanced data the deterioration of learner's accuracy caused by other data difficulty factors affects in majority of cases only the recognition of minority class, which usually is a class of particular interest.

In [11] a method for identification of data difficulty factors in real data sets was proposed. The authors distinguish 4 types of examples (enumerated from the easiest to the hardest): safe examples (lying in the region in the feature space dominated by the same class), borderline examples (lying in the class overlapping area), rare examples (a small group of examples in the region of the opposite class) and outlier examples (lying in the area dominated by the opposite class). This types can be identified by checking the distribution of the class labels among k nearest neighbours of the example. For instance, with $k = 5$, if all examples in the neighbourhood are from the opposite class then the example is considered to be an outlier. If there is 4 opposite-class examples it is rare and if there are more than 3 examples from the same class, the example is a safe one. Finally, we assign borderline type to examples with the proportion of the same class examples and the opposite class examples equal 2:3 or 3:2.

However, extensions of bagging for imbalanced data normally do not take into account the types of examples and are just focused on construction of more balanced bootstrap. There are two ways of achieving this goal: by under-sampling majority class or by over-sampling minority class. For their review see e.g. [3].

Exactly Balanced Bagging (EBBag) [7] is the representative of the first group. It copies all minority examples to each bootstrap and then, by random sampling, it adds N_+ majority examples to construct a fully balanced bootstrap. Hido et al. [6] claimed that this sampling strategy does not reflect the true bagging philosophy and they proposed Roughly Balanced Bagging (RBBag). RBBag samples with replacement N_+ examples of the minority class and then the majority examples are sampled in the same way except that the number of examples is taken from binomial distribution ($p = 0.5$, $n = N_+$).

The most known over-sampling extension of bagging is OverBagging (OverBag) [13]. It samples with replacement N_- majority examples to each bootstrap and then the same amount of minority examples is added. This results in bootstraps having multiple copies of some minority examples.

The first bagging extension which uses knowledge of data difficulty factors is Neighbourhood Balanced Bagging (NBBag) [2]. This algorithm has two variants: over-sampling (oNBBag) and under-sampling (uNBBag) both sharing the same idea of modifying sapling probability distribution by assigning weights to examples. NBBag focuses bootstrap sampling toward difficult minority examples. Weight of minority example depends on the analysis of its k nearest neighbours. Minority example is considered the more unsafe the more it has majority examples in its neighbourhood. Hence, the formula for minority example weight is the following: $w(x) = 0.5 \cdot \left(\frac{(N'_-)^\psi}{k} + 1\right)$ where N'_- is the number of majority examples among k nearest neighbours of the example and ψ is a scaling factor. Setting $\psi = 1$ causes a linear amplification of example weight with an increase of unsafeness and setting ψ to values greater then 1 effects in an exponential amplification. Each majority example is assigned a constant weight $w(x) = 0.5 \cdot \frac{N_+}{N_-}$.

As we mentioned before, both versions of NBBag use the same sampling schema; however, they create bootstrap samples of a different size. uNBBag samples $n = 2N_+$ examples resulting in a sample which is smaller than the entire imbalanced data set. oNBBag creates a bootstrap sample consisting of $n = N_+ + N_-$ elements. Since weights of minority examples are greater then weights of majority examples this results in over-sampling of minority examples.

3 Performance of Bagging Extensions

Most of the extensions of bagging presented in Sect. 2 are non-parametric. They do not introduce any new parameters, which need to be adjusted during construction of an ensemble of classifiers. On the one hand, one can argue that bagging itself is a parametric method since the adequate size of the ensemble for a given problem is not known a priori. The size of the ensemble is an important parameter, which may influence the performance of each of the considered extensions. On the other hand, fixing this parameter enables comparison of ensembles of the same size, which should allow to distinguish ones which perform better than the others under the same conditions.

Another type of parameters are introduced in Neighbourhood Balanced Bagging (NBBag). These are two parameters that control the characteristics of neighbourhood: size of neighbourhood k, and scaling factor ψ. In the experiments comparing NBBag to other bagging extensions [2] these two parameters were carefully selected to provide the best average performance. The selection was made post-hoc, i.e., first results were obtained for a number of promising pairs of parameter values and then the best values were chosen. One down-side of this approach is additional computational cost. The second, more important, one is the robustness of the recommendation. In general, a change in the list of data sets used in experiment may lead to different recommended best values.

Selection of such a type of model parameters is a known problem in machine learning [4]. However, to our best knowledge, this problem has not been yet considered in the context of learning from imbalanced data. Data imbalance may limit application of some more advanced parameter selection techniques. To put it simply, minority class examples are to valuable to spare them for selection purposes only, while majority class examples are not. Following this observation, we investigate application of a basic technique taken from tree learning to this end. In the same way as reduced-error pruning uses training data [12], we divide training data set into two stratified samples. The first sample is used for training NBBag models and the second one to validate the trained models. After the best parameters are selected, NBBag classifier is constructed on the whole training set. Contrary to what was presented in [2], this technique, when construction of a classifier is repeated, as e.g., in cross-validation, does not allow to distinguish best values of parameters for all data sets nor even for one data set. Selection of parameters is performed independently for each constructed classifier.

In the following we present performance of two variants of Neighbourhood Balanced Bagging: under-sampling (uNBBag) and over-sampling (oNBBag) with selection of k and ψ. We consider a limited set of possible values of parameters. In case of k it is: 3, 5, 7, 11. For ψ, it is: 0.25, 0.5, 1, 1.25, 1.5, 1.75, 2, 4. During selection of best parameter phase 1/3 of the training set is used for validation.

The Performance of uNBBag and oNBBag is compared to Exactly Balanced Bagging (EBBag), Over-Bagging (OverBag), and the main competitor: Roughly Balanced Bagging (RBBag). The size of ensembles is fixed to 50 components.

Table 1. Data characteristics

data set	# examples	# attributes	minority class	IR
breast-w	699	9	malignant	1.90
abdominal-pain	723	13	positive	2.58
acl	140	6	1	2.5
new-thyroid	215	5	2	5.14
vehicle	846	18	van	3.25
car	1728	6	good	24.04
scrotal-pain	201	13	positive	2.41
ionosphere	351	34	b	1.79
pima	768	8	1	1.87
credit-g	1000	20	bad	2.33
ecoli	336	7	imU	8.60
hepatitis	155	19	1	3.84
haberman	306	4	2	2.78
breast-cancer	286	9	recurrence-events	2.36
cmc	1473	9	2	3.42
cleveland	303	13	3	7.66
hsv	122	11	4.0	7.71
abalone	4177	8	0-4 16-29	11.47
postoperative	90	8	S	2.75
solar-flare	1066	12	F	23.79
transfusion	748	4	1	3.20
yeast	1484	8	ME2	28.10
balance-scale	625	4	B	11.76

The performance of bagging ensembles is measured using: *sensitivity* of the minority class (the minority class accuracy), its *specificity* (an accuracy of recognizing majority classes), their aggregation to the *geometric mean* (G-mean). For their definitions see, e.g., [5]. These measures are estimated by a stratified 10-fold cross-validation repeated ten times to reduce the variance. The differences between classifiers average results are also analyzed using Friedman and Wilcoxon statistical tests.

The results of G-mean and sensitivity are presented in Tables 2 and 3, respectively. The last row of these tables contains average ranks calculated as in the Friedman test – the lower average rank, the better classifier. Note that, the list

Table 2. G-mean [%] of NBBag and other compared bagging ensembles

data set	EBBag	OverBag	uNBBag	oNBBag	RBBag
breast-w	96.245	96.003	96.472	96.113	96.435
abdominal-pain	79.330	79.398	81.292	80.249	80.099
acl	85.576	80.866	84.359	81.927	85.310
new-thyroid	96.515	96.497	95.867	96.634	96.308
vehicle	95.038	94.934	95.440	95.115	95.417
car	96.668	96.959	96.356	96.851	96.568
scrotal-pain	73.679	74.038	72.923	71.997	75.618
ionosphere	90.540	90.559	90.874	90.568	91.002
pima	74.849	74.358	74.852	74.068	75.626
credit-g	65.737	65.513	67.450	66.628	67.963
ecoli	88.178	83.896	88.435	85.380	88.430
hepatitis	79.137	75.816	78.035	74.762	79.457
haberman	64.144	63.329	63.742	61.779	63.533
breast-cancer	58.175	60.718	58.465	58.795	60.091
cmc	64.191	61.036	65.051	63.787	65.350
cleveland	73.628	51.629	73.260	66.754	71.130
hsv	44.080	20.501	40.957	40.155	37.494
abalone	78.845	69.230	79.517	78.706	79.035
postoperative	35.569	32.657	39.877	39.142	34.847
solar-flare	83.710	64.649	83.149	79.994	83.421
transfusion	66.607	67.748	66.449	66.476	67.143
yeast	84.018	63.167	84.475	79.557	85.016
balance-scale	2.832	23.411	43.285	59.893	54.182
average rank	2.913	4	2.478	3.435	2.174

of data sets in this comparison is the same as in [2]. Data sets in the analyzed tables are ordered from the safest one to the most unsafe one. Characteristics of these data sets are given in Table 1. Looking at both Tables 2 and 3, we can make an outright observation that uNBBag and RBBag stand out as the best performing classifiers. Another observation is that over-sampling extensions of bagging, represented by OverBag and oNBBag, provide worse performance that under-sampling extensions (the rest of classifiers). Detailed comparison on G-mean gives the best average rank to RBBag, however the difference between its rank and ranks of all other classifiers except OverBag is not significant. Friedman test on values of G-mean results in p-value around 0.0002, and according to Nemenyi post-hoc test, critical difference between ranks is around 1.272. An analogous observation is valid only for NBBag and all other classifiers except

Table 3. Sensitivity [%] of NBBag and other compared bagging ensembles

data set	EBBag	OverBag	uNBBag	oNBBag	RBBag
breast-w	96.929	95.851	97.386	96.888	96.846
abdominal-pain	82.178	75.842	84.158	80.050	79.010
acl	87	74.250	87.250	82.500	84.750
new-thyroid	95.714	95.143	95.143	96	95.143
vehicle	97.236	94.523	97.286	95.477	96.935
car	100	95.652	100	95.942	100
scrotal-pain	76.271	70.169	76.441	73.051	75.763
ionosphere	86.032	85.159	87.778	86.984	85.714
pima	80.672	74.925	81.194	79.813	78.396
credit-g	72.933	60.233	73.400	69.867	68.500
ecoli	92	76	92	84	90.571
hepatitis	83.438	67.188	79.062	69.688	77.500
haberman	56.914	59.136	63.827	66.543	55.802
breast-cancer	63.412	54	65.176	59.059	58.471
cmc	70.240	50.721	68.739	63.423	64.685
cleveland	80.286	30.571	79.143	63.429	69.143
hsv	45	7.143	40	35.714	26.429
abalone	80.925	51.224	80.776	75.851	77.045
postoperative	31.250	17.917	44.167	37.917	23.750
solar-flare	88.140	46.977	86.744	81.395	85.581
transfusion	66.517	61.236	72.697	67.753	65.674
yeast	91.765	40.980	90.392	73.529	88.431
balance-scale	99.388	7.347	94.898	79.796	66.327
average rank	1.848	4.870	1.587	3.174	3.522

OverBag. Direct comparison of RBBag and NBBag in Wilcoxon test does not show a significant difference in G-mean (p-value in this test is around 0.247).

When we move to the observed values of sensitivity in Table 3, we can notice considerably better average performance of uNBBag and EBBag than the rest of classifiers. This observation is supported by results of Friedman test (with p-value close to 0) and Nemenyi post-hoc analysis. Wilcoxon tests shows the same result in pairs of classifiers. uNBBag achieves the best average rank in this experiment. Nevertheless, direct comparison of uNBBag and EBBag in Wilcoxon test does not confirm a significant difference in sensitivity (p-value 0.677).

Experimental comparison of performance of bagging extensions leads to conclusions, which are concordant with the ones presented in [2]. RBBag and uNBBag are distinguished as two standing out alternatives. It should be noted that the results presented here are not entirely comparable with these from [2], since the set of compared classifiers has changed. We included EBBag, which proved to be a valuable extension. Another aspect of the presented comparison is the influence of parameter selection on the results. Application of a relatively simple selection technique allowed us to obtain quite satisfying results. The average performance of NBBag has not been observably improved but variability of results for unsafe data sets has decreased (e.g., `balance-scale`). We expect that a technique adapted for imbalanced data should allow to obtain even better results.

4 Measuring Diversity of Ensembles

One of the most important characteristic of an ensemble is diversity of its component classifiers. To put it simple, if all components make the same decision regarding example's classes, the construction of an ensemble is pointless. In [8] authors compare many diversity measures and recommend use of Q-statistics basing on ease of its interpretation. Q-statistics is defined for a pair of components as $Q = \frac{n_{11}n_{00} - n_{01}n_{10}}{n_{11}n_{00} + n_{01}n_{10}}$ where n_{11} is the number of examples on which both classifiers make correct decision, n_{01} and n_{10} are the numbers of examples on which one classifier is wrong and the other makes a correct decision, n_{00} is the number of examples on which both classifiers make incorrect decisions. This formula is calculated for each pair of components and then its averaged for the whole ensemble. $Q = 0$ means independence of component classifiers, positive Q means that classifiers tend to recognize the same elements correctly and negative values signify that components tend to make errors on different examples.

We calculate Q-statistic for NBBag and RBBag on all data sets from previous experiment. Due to space limits, we do not present all the results. We only briefly summarize this analysis. The most diversified classifier according to both median and average of Q-statistic is uNBBag ($Median(Q) = 0.61$). RBBag have a bit less diversified components ($Median(Q) = 0.67$) and oNBBag has the highest averaged results on Q-statistic ($Median(Q) = 0.71$). The biggest differences between algorithms is visible on `haberman` and on `balance-scale`. On these data sets the most diversified classifier has also the highest result on G-mean measure. On other data sets these two factors are not always related.

Further investigation of Q-statistic only for minority examples (Q_{min}) shows that all analyzed algorithms are more diversified on minority class. On some data sets classifiers achieve even negative values of Q_{min}. Likewise the differences between classifiers are a little higher. The ranking of most diversified classifiers remain the same as for over-all Q-statistic: uNBBag ($Median(Q_{min}) = 0.40$), RBBag ($Median(Q_{min}) = 0.47$) and oNBBag ($Median(Q_{min}) = 0.51$).

Another way of investigating diversity is analysis of votes of each component during classification of a particular example. Here, we use a margin measure defined as follows: $margin = \frac{n_{corr} - n_{incorr}}{n_{corr} + n_{incorr}}$, where n_{corr} and n_{incorr} is the number of components which vote for correct and incorrect class, respectively. The margin value equal 1 means completely certain and correct decision, margin -1 means completely certain but incorrect decision. Margin close to 0 indicates uncertainty in making final decision (the number of classifiers voting for the correct class is close to the number of classifiers voting for the opposite class).

We analyze the values of margin calculated for examples with respect to their types. Additionally, we compare margins for examples on which RBBag and uNBBag make different decisions. In Fig. 1 we present histograms of decision margin for minority class on a representative data set (`abalone`). In the first row of the plot one can see decision margins of all examples of a particular type (white bars) achieved by RBBag. Red bars of the histogram indicate margins for examples which are classified incorrectly by RBBag but they are correctly classified by uNBBag. Analogically, green bars demonstrate margin for instances which were classified correctly only by RBBag. The second row of the plot is constructed in the same way but for uNBBag.

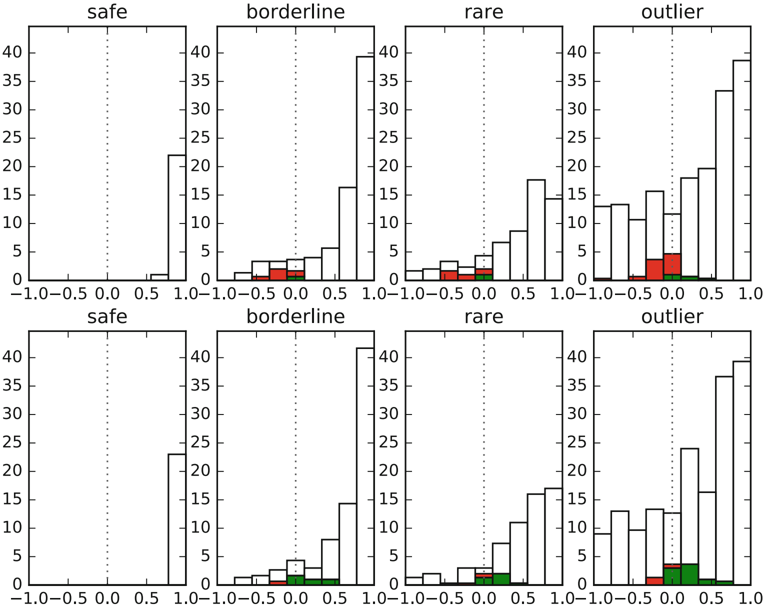

Fig. 1. Histogram of RBBag (top) and uNBBag (down) margins for `abalone` minority examples with respect to their types.

The first impression is that both classifiers work quite similar. Differences are more significant on difficult examples. uNBBag and RBBag do not have problems with correct classification of safe minority examples. Almost all of them are classified with maximal margin. However, with increase of difficulty of examples, both classifiers makes more errors and their confidence goes down. Particularly, a lot of outlier examples are classified incorrectly with high confidence.

Compared algorithms make different final predictions only on more difficult examples and it is clear that uNBBag makes correct decisions on a higher number of minority examples. Unfortunately, there seems to be some kind of trade-off between correct recognition of more difficult minority and majority examples:

this classifier makes more incorrect decisions on majority examples than RBBag. This is the reason why RBBag is sometimes better than uNBBag on G-mean measure. Furthermore, it is worth to notice that when uNBBag makes correct prediction on a minority example and RBBag makes an incorrect one, it is with a rather low confidence. It is quite unlikely to find an example correctly classified by uNBBag and classified incorrectly by RBBag with margin less then -0.5.

5 Conclusions

In this work, we have experimentally compared a number of promising bagging extensions designed to handle class imbalance problem. The best performing extensions in this comparison are: Roughly Balanced Bagging (RBBag) and Neighbourhood Balanced Bagging (NBBag). We have introduced a simple technique for automatic selection of parameters for NBBag during learning from imbalanced data. This technique proved to work well. Nevertheless, we believe that another technique better adapted for the type of learning should allow to obtain even better results. Comparative study of diversity of RBBag and NBBag have shown that NBBag is able to make correct prediction on a higher than RBBag number of difficult to learn minority examples. There is, however, a trade-off between correct recognition of difficult minority examples and majority examples, which allows RBBag to perform better in some cases.

Acknowledgement. The research was supported by NCN grant DEC-2013/11/B/ST6/00963.

References

1. Błaszczyński, J., Stefanowski, J., Idkowiak, L.: Extending bagging for imbalanced data. CORES 2013. Advances in Intelligent Systems and Computing, vol. 226, pp. 269–278. Springer, Switzerland (2013)
2. Błaszczyński, J., Stefanowski, J.: Neighbourhood sampling in bagging for imbalanced data. Neurocomputing **150 A**, 184–203 (2015)
3. Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., Herrera, F.: A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. IEEE Trans. Syst. Man Cybern. Part C Appl. Rev. **99**, 1–22 (2011)
4. Guyon, I., Saffari, A., Dror, G., Cawley, G.: Model selection : beyond the Bayesian / Frequentist divide. JMLR **11**, 61–87 (2010)
5. Japkowicz, N., Shah, M.: Evaluating Learning Algorithms: A Classification Perspective. Cambridge University Press, Cambridge (2011)
6. Hido, S., Kashima, H.: Roughly balanced bagging for imbalance data. In: Proceedings of the SIAM International Conference on Data Mining, pp. 143–152 (2008). An extended version in Statistical Analysis and Data Mining, vol. 2 (5–6), pp. 412–426 (2009)
7. Hoens, T.R., Chawla, N.V.: Generating diverse ensembles to counter the problem of class imbalance. In: Zaki, M.J., Yu, J.X., Ravindran, B., Pudi, V. (eds.) PAKDD 2010. LNCS, vol. 6119, pp. 488–499. Springer, Heidelberg (2010)

8. Kuncheva, L., Whitaker, C.: Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. Machine Learning **51**(2), 181–207 (2003)
9. Kotsiantis, S., Kanellopoulos, D., Pintelas, P.: Handling imbalanced datasets: a review. GESTS Int. Trans. Comput. Sci. Eng. **30**(1), 25–36 (2006)
10. Khoshgoftaar, T., Van Hulse, J., Napolitano, A.: Comparing boosting and bagging techniques with noisy and imbalanced data. IEEE Trans. Syst. Man Cybern. Part A **41**(3), 552–568 (2011)
11. Napierala, K., Stefanowski, J.: Identification of different types of minority class examples in imbalanced data. In: Corchado, E., Snášel, V., Abraham, A., Woźniak, M., Graña, M., Cho, S.-B. (eds.) HAIS 2012, Part II. LNCS, vol. 7209, pp. 139–150. Springer, Heidelberg (2012)
12. Quinlan, R.: C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, San Mateo (1993)
13. Wang, S., Yao, T.: Diversity analysis on imbalanced data sets by using ensemble models. In: Proceeding IEEE Symposium Computational Intelligence and Data Mining, pp. 324–331 (2009)
14. Yang, Q., Wu, X.: 10 challenging problems in data mining research. Int. J. Inf. Technol. Decis. Making **5**(04), 597–604 (2006)