Impact of Sampling on Neural Network Classification Performance in the Context of Repeat Movie Viewing

Elena Fitkov-Norris[1] and Sakinat Oluwabukonla Folorunso[2]

[1] Kingston University, Kingston Hill, Kingston-upon-Thames, KT2 7LB, UK
E.Fitkov-Norris@kingston.ac.uk
[2] Mathematical Sciences Department, Olabisi Onabanjo University (OOU) Ago – Iwoye, Ogun State, Nigeria
sakinatfolorunso@yahoo.co.uk

Abstract. This paper assesses the impact of different sampling approaches on neural network classification performance in the context of repeat movie going. The results showed that synthetic oversampling of the minority class, either on its own or combined with under-sampling and removal of noisy examples from the majority class offered the best overall performance. The identification of the best sampling approach for this data set is not trivial since the alternatives would be highly dependent on the metrics used, as the accuracy ranks of the approaches did not agree across the different accuracy measures used. In addition, the findings suggest that including examples generated as part of the over-sampling procedure in the holdout sample, leads to a significant overestimation of the accuracy of the neural network. Further research is necessary to understand the relationship between degree of synthetic over-sampling and the efficacy of the holdout sample as a neural network accuracy estimator.

Keywords: Over-sampling, under-sampling, neural network, classification.

1 Introduction

Building accurate classifiers for imbalanced data sets is challenging due to the high probability of misclassification of the underrepresented data class. The class imbalance problem corresponds to a problem domain for which one class is represented by a large number of examples, while the other is represented by just a few, the ratio of the small to large classes can be as high as 1:100, 1:1000, or 1:10,000 [1], overwhelming standard classifiers such as decision trees, neural networks and support vector machines, which exhibit a strong bias towards the majority class and ignore the smaller class [2]. This imbalance causes suboptimal classification performance since typical learning algorithms tend to maximize the overall prediction accuracy at the expense of the minority class [1], [3].

Class imbalance occurs naturally in a wide range of domains including medicine, e.g. diagnosing rare diseases, gene mutations or DNA sequencing [4], in engineering, when identifying oil spills in satellite radar images, document retrieval and classification, spam or speech patterns detection [5], [6], or banking and finance, when detecting

L. Iliadis, H. Papadopoulos, and C. Jayne (Eds.): EANN 2013, Part I, CCIS 383, pp. 213–222, 2013.

fraudulent transactions or assessing risk [7]. Imbalanced data sets also occur in areas where data for the minority class are rare e.g. space shuttle failure or in cases when cost, privacy and the effort required to obtain a representative data set create 'artificial' imbalances [1], [7].

The extent to which class imbalance affects classifier learning varies depending on the characteristic of the problem, the degree of class imbalance, the training sample size and the type of classifier used [2], [4], [8]. A large class imbalance has a significant impact on classifier performance only if a classifier is tackling a complex problem or is presented with a small training sample [2]. The type of classifier also has an impact, some classifiers, such as decision trees and support vector machines being very sensitive to class imbalance [2].

Numerous solutions to the class imbalance problem have been proposed both at data and algorithmic levels. The majority are designed for a two-class or binary problem where one class is highly under-represented but associated with a higher identification importance. Solutions at data level attempt to re-balance the class distribution by resampling the data space, while at the algorithm level solutions try to adapt existing classifier learning algorithm to strengthen learning with regards to the minority class [2]. The main advantage of data level techniques is that they are independent of the underlying classifier [3].

A number of data resampling techniques have been suggested to deal with the problem of class imbalance by balancing the distribution of the training data [9]. The most intuitive approach is either to add examples to the minority class (oversampling) or remove examples from the majority class (under-sampling) [2]. The selection of cases for under- and over-sampling could be performed at random or in a systematic manner, following a predefined rule or objective [8], [10].

A number of studies have evaluated the impact of sampling on decision tree classifiers with mixed findings and empirical evidence is emerging that the best approach could be domain specific [10]. Fewer studies have concentrated on the performance of neural network classifiers in conditions of class imbalance, possibly because they are believed to be more flexible and less likely to be affected by class imbalance problems [2]. In addition, a standard recommendation in neural network training is that duplicate observations are removed from a data set, to minimise the probability of the neural network over-fitting the data, and loosing its ability to generalise [10], [11], thus rendering the random oversampling technique irrelevant. Empirical studies have shown that a large class imbalance in the training dataset has a detrimental effect on neural network performance, in particular when the training sample size is small and random over-sampling has an advantage over under-sampling approaches [12], [13]. Furthermore, the impact of class imbalance on performance of neural networks is less pronounced compared to other types of classifier, although the classification results show significant variance [2].

This article investigates the impact of various sampling techniques on neural network classifier performance in the context of predicting repeat movie going. The problem domain of repeat movie going is naturally imbalanced as only a small proportion of movie goers are likely to see the same movie twice [14]. Identifying repeat movie going accurately is of particular interest to practitioners in the field as movie revenues are notoriously difficult to predict [15]. Neural networks have been used successfully to build accurate predictor models for movie box office success [16] although the optimal classifier type seems to be domain specific [17]. The use of neural network to predict repeat viewing has been attempted before, and it was found that, surprisingly, neural networks did not offer a significant advantage over parametric approaches such as logistic regression [18]. However, the study did not take into account the imbalanced nature of the data set and the adverse effect it may have on classifier performance. The purpose of this empirical study is to evaluate the impact of different sampling techniques on the predictive accuracy of back-propagation neural network in the context of repeat movie going.

The paper starts with a brief overview of the different over- and under-sampling techniques, followed by an introduction to the data set and research methodology. The experimental results are reported and discussed in section 4 and followed by conclusions.

2 Sampling Techniques

2.1 Under-Sampling Techniques

Random Under-Sampling (RUS). This technique removes instances from the majority class at random, until a desired class distribution is achieved. As it makes no attempt to remove examples "intelligently", it can discard potentially useful data that could be important for the learning process and make the decision boundary between minority and majority class harder to learn [13].

Condensed Nearest Neighbour Rule (CNN). This technique finds a consistent subset of the original data set which, when used as a reference for the nearest neighbour rule can classify correctly all instances in the original dataset [19]. The main problem with the CNN rule is that it is likely to include a large proportion of noisy examples which are hard to classify and are, therefore, more likely to be included in the training set [20].

Tomek Links (TL). This technique modifies the condensed nearest neighbour technique by retaining only borderline examples within the condensed subset and so reduces computational load. Let E_i, E_j, belong to different classes, and d (E_i, E_j) is the distance between them. A (E_i, E_j) pair is called a Tomek link if there is no example E_l, such that d (E_i, E_l) < d (E_i, E_j) or d (E_j, E_l) < d (E_i, E_j). Examples qualifying as Tomek links are observations that are either borderline or noisy and their removal could improve the decision boundary of the problem [21]. Combinations of Tomek Link and CNN have been suggested, with the aim of utilising the benefits of each approach [5], [8].

Wilson's Edited Nearest Neighbour Rule (ENN). Wilson proposed an edited k nearest neighbours (k-NN) rule, which consists of two steps. Firstly the k-NN rule is used to edit the set of pre-classified samples by deleting all examples whose class differs from the majority class of its k-NNs. Afterward, new examples are classified using a 1-NN rule and the reduced reference set derived in step one [22].

Neighbourhood Cleaning Rule (NCL). In this technique, the ENN rule is used to identify and remove majority class noisy examples. For each example (E_i) in the training set, if E_i belong to the majority class and is misclassified by its three Nearest Neighbours (3-NNs), then E_i is removed. If E_i belongs to the minority class and it is misclassified by its 3-NNs from the majority class, then the 3 nearest neighbours are removed. To avoid excessive reduction of small classes, the rule is modified to remove examples misclassified by 2-NN instead of 3-NN [20].

2.2 Over-Sampling Techniques

Random Over-Sampling (ROS). This is the continuous replication of the minority class at random until a more balanced or desired distribution is reached. As mentioned, this approach can increase the likelihood of classifier over-fitting and higher computational load for the classifier [10], [11].

Synthetic Minority Oversampling Technique (SMOTE). This technique generates synthetic examples by operating in feature space rather than data space. The minority class is oversampled by introducing synthetic examples along the line segments joining any/all of k minority class nearest neighbours. This technique overcomes the over-fitting problem and broadens the decision region of the minority class examples, dealing with both relative an absolute imbalance [23].

Advanced Sampling Techniques. A number of approaches combine over-sampling of the minority class, using SMOTE with under-sampling of the majority class by using ENN or TL in order to balance the training dataset and optimise classifier performance [23], [24].

3 Experimental Design

3.1 Data Set Background and Description

The data set used to test the impact of under- and over-sampling consists of the 2002 iteration of the Cinema And Video Industry Audience Research (CAVIAR) survey which identifies the demographic characteristics of cinema-goers and if they had seen a film in the cinema more than once [14]. After removing duplicate observations, the original data set consisted of 786 observations depicting whether an individual visited the cinema to see the same movie twice, their age category, social class, and preference for visiting the cinema. 33% of the entries in the data set were repeat viewers, showing a moderate imbalance ratio of 2.06 [25]. Further details of the data set can be found in [14].

3.2 Neural Networks Overview

Neural networks were developed to simulate the function of the human brain and in particular its ability to handle complex, non-linear pattern recognition tasks efficiently. Neural networks are built from simple processing units or neurons, which enable the network to learn sets of input-output mappings and thus solve classification problems. Each processing unit or neuron consist of three elements: a set of synapses or connecting links which take the input signals, an adder for summing the input signals and an activation function which limits the level of a neuron's output. In addition, each input is allocated a weight of its own, which is adjusted during training and represents the relative contribution of that input (positive or negative) to the overall neuron output. The output function of neuron *k* can be depicted in mathematical terms as:

$$y_k = \varphi\left(\sum\nolimits_{j=0}^{m} w_{kj} x_j\right) \tag{1}$$

where y_k is the output of neuron k, x_j denotes neural network inputs (from 0 to m), w_{kj} denotes the synaptic weight for input j on neuron k and $\varphi(\circ)$ is the neuron activation function. The input for neuron 0 is always +1 and it acts as an overall bias, increasing or decreasing the net output of the activation function.

Multilayer feed-forward neural networks are a subtype of neural network distinguished by the presence of hidden layers of neurons and are particularly well suited to solving complex problems by enabling the network to extract and model non-linear relationships between the input and output layers. Typically, the outputs from each layer in the network act as input signals into the subsequent layer, so the final output layer presents the response of the network to different input patterns. The optimal number of hidden layers is problem specific, and previous research has shown that a feed forward network with one hidden layer is most suited for predicting repeat viewing [18]. Back propagation, essentially a gradient-descent technique and one of the most widely used algorithms will be used to train the network and minimise the error between the target and actual classifications. The network will be simulated using Matlab (Release 2013a) with a tansig activation function and between 6 and 16 neurons in the hidden layer.

3.3 Sampled Data Sets

The under and over-sampling was carried out using the original data set and a subset was created for each of the different sampling approaches outlined above. The class distribution in each data set is shown in Table 1. Random over-sampling was not carried out to avoid neural network over-fitting and an artificially inflated accuracy. The data set consists of binary indicator variables and the HVDM rule, which uses agreement between the values of the nominal/binary variables to determine similarity between observations [11].

Table 1. Sampling Approaches and Resulting Data Sets

	Sampling Approach	Number of cases (non-repeat viewers : repeat viewers)
Under-sampling	CNN	298:229
	ENN	415:79
	NCL	205:257
	RUS	257:257
	TL	269:31
Over-sampling	SMOTE	529:514
	SMOTE + ENN	400:346
	SMOTE + TL	268:235
	SMOTE (300%)	529:1028
	Original Data	529:257

3.4 Experimental Set Up

To evaluate the performance of the neural network models across a range of different inputs, including new objects that the network has not seen before, it is common practice to use a holdout sample. The data set is split into three subsets: a training sample, a testing sample and a holdout sample [17]. The network learns pattern mappings by minimising the errors on the training set. The testing set is used as a benchmark to prevent over-fitting, while the holdout sample is used as an independent means to test the classification accuracy of the network on a sample of data that it has not seen before (out of sample accuracy). Choosing the holdout sample randomly could lead to a bias in the accuracy estimation due to random sample fluctuations but K-fold cross validation provides an alternative for testing the ability of a neural network to generalise [18]. 10-fold cross validation is well established as a reliable estimate of neural network performance [16]. The 10 subsets are derived at random for each data set and tested using 5 different random seeds. As the number of instances in each data set is different, the original data set is used as a baseline for comparison and, at the end of each training cycle, the network is also tested with the original data file. The classification performance is calculated as the average accuracy across the 10-folds for the holdout samples and the benchmark original data file.

3.5 Performance Measures

One of the most common metrics for measuring classification accuracy for categorical classifiers such as neural networks is the confusion matrix. Various measures, derived from the confusion matrix, including overall classification accuracy, sensitivity and specificity are widely used to assess classifier performance [17].

Overall classification accuracy for a particular categorical classifier is defined as the percentage of correct predictions by the classifier. A common criticism of the overall classification accuracy measure is that it does not take into account class imbalance between different categories/classes and as a result could lead to misleading results since the impact of underrepresented groups would be small [25].

Minimum sensitivity is an alternative measure, which overcomes the problem of imbalanced representation. It is the lower of the sensitivities from the different classes in the problem, in effect the worst performance of the classifier defined as:

$$MS = \min\{P(i); i = 1,\ldots,J\} \quad where \quad P(i) = n_{ii} \Big/ \sum_{j=1}^{J} n_{ij} \tag{2}$$

The problem with minimum sensitivity is that it could make direct comparison of results difficult, as the worst performing class is problem specific.

An alternative measure of classification accuracy which overcomes some of the problems of the overall and minimum sensitivity accuracy is the geometric mean (GM) which uses the concept of true positive and true negative classifier accuracy [5], [25]. It is defined as:

$$g = \sqrt{a^{+} \times a^{-}} \tag{3}$$

where a^{+} denotes the accuracy in positive examples (or true positive rate and defined as the proportion of correctly classified majority class examples), and a^{-} is the accuracy in negative examples (or true negative rate and defined as the proportion of correctly classified minority class examples). This study will use the minimum sensitivity and the geometric mean as measures of neural network performance.

4 Results and Discussion

The ranked average classification results of the neural networks using the minimum sensitivity and geometric mean measures on the holdout and benchmark datasets under different under- and over-sampling conditions are shown in Table 2. The performance of all sampling approaches were compared using ANOVA means comparison test, and the same rank was allocated to approaches without statistically significant difference in their mean accuracy performance.

SMOTE oversampling, either on its own, or combined with Tomek Link (TL) or Edited Nearest Neighbour (ENN) techniques, led to higher classification accuracy in predicting both repeat and non-repeat viewers compared to the original data and the under-sampled data. This finding is in line with other empirical studies that concluded that synthetic over-sampling has an advantage over under-sampling approaches [12], [13], in particular studies which identified SMOTE + ENN and SMOTE + TL as robust and reliable over-sampling approaches [8]. Therefore neural network performance can benefit by expanding the problem space (with SMOTE) and the removal of noisy observations using under-sampling of the majority class with ENN or TL.

Two under-sampling approaches performed better than the original data: NCL and ENN. NCL, the best performing under-sampling approach, provided the best prediction accuracy for the repeat viewer minority class (91% accuracy), but this was at the expense of shifting the error to the majority class of non-repeat viewers (26% accuracy). ENN offered one of the best accuracies in predicting the majority class in the benchmark data file (86% accuracy), although this was at the cost of predicting the minority repeat viewer class (31% accuracy). Both NCL and ENN retained the largest proportion of observations in the data sets they predicted the best (repeat viewer and non repeat viewers, respectively) in comparison to other under-sampling approaches and this suggest that observations that were identified as noisy were in fact essential dimensions of the problem space. This data set may have both absolute and relative imbalance [26] and this explains the superior performance of over-sampling.

Table 2. Neural Network Classification Results

Sampling Approach	% Non-Repeaters Correct		% Repeaters Correct		% Overall Correct		% Geometric Mean		Classifier Performance Ranking				Rank Average
									Minimum Sensitivity		**Geometric Mean**		
	Holdout	Benchmark	Holdout	Benchmark	Holdout	Benchmark	Holdout	Benchmark	Holdout	Benchmark	Holdout	Benchmark	
SMOTE + ENN	92.2	84.2	85.7	36.9	89.1	68.8	88.7	55.5	2	3.5	2	2.5	2.5
SMOTE	81.3	63.3	64.0	60.0	72.9	62.2	71.9	**61.3**	5.5	1.5	4.5	1	3.13
SMOTE + TL	95.0	90.2	90.4	24.7	93.0	68.8	92.5	46.6	1*	7.5	1	5.5	3.75
NCL	64.4	26.1	67.8	**90.5**	66.1	47.2	65.4	47.7	5.5*	3.5	4.5	5.5	4.75
ENN	83.6	**86.6**	67.2	31.4	80.6	68.5	70.6	48.3	5.5	5	4.5	5.5	5.13
ORIGINAL DATA	58.5	62.3	51.2	56.6	55.8	60.5	51.5	56.0	9*	1.5	8	2.5	5.25
SMOTE (300%)	90.7	88.0	78.4	29.9	82.6	**69.0**	83.6	50.5	3	9	4.5	5.5	5.5
TL	83.7	93.9	54.7	17.8	80.4	69.0	56.9	37.0	5.5	10	8	10	8.38
RUS	55.7	24.8	56.1	88.6	55.8	45.7	54.8	44.4	9*	7.5*	8	8.5	8.5
CNN	51.4	26.7	46.0	83.4	48.7	45.2	46.0	43.5	10	7.5*	10	8.5	9.5
Overall Average	78.6	71.5	66.9	43.2	75.7	62.2	67.0	45.6					

* Denotes minimum sensitivity on the majority class.

The classification accuracy derived using the holdout sample was consistently higher than the accuracy on the benchmark data set. This is particularly true for the predictive accuracy of the minority class. This suggests that even synthetically over-sampled data should be removed from the holdout sample to ensure that it is representative. This recommendation is generally given for random over-sampling which introduces identical examples in the data set and can lead to neural network over-fitting problems [2], but this finding is somewhat counter-intuitive for synthetic oversampling which introduces interleaved copies of the minority class. However, as the representation of the sample data is binary (0 or 1), the small data range and discrete data values give a greater likelihood of introducing examples that are identical to existing observations. This experiment suggests that, it is advisable to remove synthetically interleaved observations from the holdout sample used to test neural network in the case of discrete data with small data range.

The best sampling approach was determined using average rank of its performance; however, there is no significant agreement between the ranking of the models across the two different accuracy measures in the holdout and benchmark data sets and averaging could be masking some very poor performances for some classifiers. For example, the second best ranked sampling approach SMOTE + TL has a relatively weak overall performance and one could argue that it is not suited this data set. Although the optimal way for choosing a classifier is beyond the scope of this work, the findings suggest that the best sampling approach, in the context of repeat viewing data is dependent on the objectives of the classification (overall or minority class).

5 Conclusions and Future Work

This paper assessed the impact of different sampling approaches on neural network classification performance in the context of repeat movie going. The results showed that synthetic oversampling of the minority class, either on its own or combined with under-sampling and removal of noisy examples from the majority class using SMOTE + ENN or SMOTE + TL offered the best performance. The identification of the optimal approach for this data set is not trivial as the recommendations would be highly dependent on the accuracy measure used. The findings also suggest that including examples that were generated by the oversampling procedure in the holdout sample, leads to a significant overestimation of the accuracy of the neural network. It is hypothesised that this is a context specific problem as the data set consisted of indicator variables and so further research would be necessary to understand the relationship between synthetic oversampling and the efficacy of the holdout sample as an estimator of neural network.

References

1. Chawla, N.V., Japkowicz, N., Kotcz, A.: Editorial: Special Issue on Learning from Imbalanced Data Sets. SIGKDD Explor. Newsl. 6(1), 1–6 (2004)
2. Japkowicz, N., Stephen, S.: The Class Imbalance Problem: A Systematic Study. Intell. Data. Anal. 6(5), 429–449 (2002)
3. Fernández, A., García, S., Herrera, F.: Addressing the Classification with Imbalanced Data: Open Problems and New Challenges on Class Distribution. In: Corchado, E., Kurzyński, M., Woźniak, M. (eds.) HAIS 2011, Part I. LNCS, vol. 6678, pp. 1–10. Springer, Heidelberg (2011)
4. Pearson, R., Goney, G., Shwaber, J.: Imbalanced Clustering of Microarray Time-Series. In: Fawcett, T., Mishra, S. (eds.) 12th International Conference on Machine Learning Workshop on Learning from Imbalanced Datasets II, Washington DC, vol. 3 (2003)
5. Kubat, M., Matwin, S.: Addressing the Curse of Imbalanced Training Sets: One-Sided Selection. In: 14th International Conference on Machine Learning, Nashville, Tennessee, USA, pp. 179–186 (1997)
6. Manevitz, L.M., Yousef, M.: One-Class SVMs for Document Classification. JMLR 2, 139–154 (2002)

7. Thai-Nghe, N., Busche, A., Schmidt-Thieme, L.: Improving Academic Performance Prediction by Dealing with Class Imbalance. In: 9th IEEE International Conference on Intelligent Systems Design and Applications, Pisa, Italy, pp. 878–883 (2009)
8. 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)
9. Folorunso, S.O., Adeyemo, A.B.: Theoretical Comparison of Undersampling Techniques Against Their Underlying Data Reduction Techniques. In: EIE 2nd International Conference Computing, Energy, Networking, Robotics and Telecommunications (EIECON 2012), Lagos, Nigeria, pp. 92–97 (2012)
10. Kotsiantis, S., Kanellopoulos, D., Pintelas, P.: Handling Imbalanced Datasets: A Review. GESTS International Transactions on Computer Science and Engineering 30(1), 25–36 (2006)
11. Zhou, Z.-H., Liu, X.-Y.: Training Cost-Sensitive Neural Networks with Methods Addressing the Class Imbalance Problem. IEEE T. Knowl. Data. En. 18(1), 63–77 (2006)
12. Mazurowski, M.A., Habas, P.A., Zurada, J.M., Lo, J.Y., Baker, J.A., Tourassi, G.D.: Training Neural Network Classifiers for Medical Decision Making: The Effects of Imbalanced Datasets on Classification Performance. Neural Networks 21(2), 427–436 (2008)
13. Crone, S.F., Finlay, S.: Instance Sampling in Credit Scoring: an Empirical Study of Sample Size and Balancing. Int. J. Forecasting 28(1), 224–238 (2011)
14. Collins, A., Hand, C., Linnell, M.: Analyzing Repeat Consumption of Identical Cultural Goods: Some Exploratory Evidence from Moviegoing. J. Cult. Econ. 32(3), 187–199 (2008)
15. Sawhney, M., Eliashberg, J.: A Parsimonious Model for Forecasting Gross Box-Office Revenues of Motion Pictures. Market. Sci., 113–131 (2001)
16. Sharda, R., Delen, D.: Predicting Box-Office Success of Motion Pictures with Neural Networks. Expert Syst. Appl. 30(2), 243–254 (2006)
17. Paliwal, M., Kumar, U.A.: Neural Networks and Statistical Techniques: A Review of Applications. Expert Syst. Appl. 36(1), 2–17 (2009)
18. Fitkov-Norris, E., Vahid, S., Hand, C.: Evaluating the Impact of Categorical Data Encoding and Scaling on Neural Network Classification Performance: The Case of Repeat Consumption of Identical Cultural Goods. In: Jayne, C., Yue, S., Iliadis, L. (eds.) EANN 2012. CCIS, vol. 311, pp. 343–352. Springer, Heidelberg (2012)
19. Hart, P.E.: The Condensed Nearest Neighbor Rule. IEEE T. Inform. Theory 14(3), 515–516 (1968)
20. Laurikkala, J.: Improving Identification of Difficult Small Classes by Balancing Class Distribution. In: Quaglini, S., Barahona, P., Andreassen, S. (eds.) AIME 2001. LNCS (LNAI), vol. 2101, pp. 63–66. Springer, Heidelberg (2001)
21. Tomek, I.: Two Modifications of CNN. IEEE T. Syst. Man. Cyb. 11(6), 769–772 (1976)
22. Wilson, D.L.: Asymptotic Properties of Nearest Neighbor Rules Using Edited Data. IEEE T. Syst. Man. Cyb. SMC-2(3), 408–421 (1972)
23. 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)
24. 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)
25. García, S., Herrera, F.: Evolutionary Undersampling for Classification with Imbalanced Datasets: Proposals and Taxonomy. Evol. Comput. 17(3), 275–306 (2009)
26. Chen, S., He, H., Garcia, E.A.: RAMOBoost: Ranked Minority Oversampling in Boosting. IEEE T. Neural Networ. 21(10), 1624–1642 (2010)