Hindawi Publishing Corporation
Abstract and Applied Analysis
Volume 2014, Article ID 972786, 7 pages
http://dx.doi.org/10.1155/2014/972786

Research Article

A Hybrid Sampling SVM Approach to Imbalanced Data Classification

Qiang Wang

College of Electrical and Information Engineering, Lanzhou University of Technology, Lanzhou 730050, China

Correspondence should be addressed to Qiang Wang; wangqiang@lut.cn

Received 17 April 2014; Accepted 22 May 2014; Published 12 June 2014

Academic Editor: Fuding Xie

Imbalanced datasets are frequently found in many real applications. Resampling is one of the effective solutions due to generating a relatively balanced class distribution. In this paper, a hybrid sampling SVM approach is proposed combining an oversampling technique and an undersampling technique for addressing the imbalanced data classification problem. The proposed approach first uses an undersampling technique to delete some samples of the majority class with less classification information and then applies an oversampling technique to gradually create some new positive samples. Thus, a balanced training dataset is generated to replace the original imbalanced training dataset. Finally, through experimental results on the real-world datasets, our proposed approach has the ability to identify informative samples and deal with the imbalanced data classification problem.

1. Introduction

In the literature and in real-world problems, the scenario of imbalanced data distribution appears when the size of samples in one class is greatly larger than the size of samples in the other class. Many applications such as fraud detection, intrusion prevention, risk management, and medical research often have the imbalanced class distribution problem. Classifiers constructed based on imbalanced datasets usually perform well on the majority class data but poorly on the minority class data [1]. However, in many cases, the minority class data are the most important ones to detect, for example, in the medical field for disease diagnosis or in the industrial field for fault diagnosis.

Class imbalance has been appointed as one of the most challenging problems in the data mining field [2]. Many traditional classification methods tend to be overwhelmed by the majority class and ignore the minority class. Their classification performances on the minority class are negatively affected. Actually, these traditional classifiers, such as support vector machine (SVM), decision trees, and neural networks, are designed to optimize the overall performance on the whole dataset. In order to cope with the class imbalance problem, researchers have proposed many methods from the view of data-level approaches and algorithmic approaches. The data-level approaches balance the training dataset of the classifier by resampling techniques, while the algorithmic approaches deal with the development of new algorithms expressly designed to cope with uneven datasets. The two approaches are independent of each other and can be combined to enhance each other's performance [3].

Resampling is one of the effective approaches for balancing the training dataset of a classifier, which includes undersampling and oversampling techniques. In this paper, a new hybrid sampling approach combining oversampling and undersampling is presented to address the class imbalance problem. The proposed approach first uses undersampling to delete some samples of the majority class with less classification information and then applies oversampling to gradually create some new positive samples. Thus, a balanced training dataset is generated to replace the original imbalanced training dataset. Through experimental results on the real-world datasets, our approach has the ability to identify informative samples and deal with the imbalanced data classification problem. In addition, the proposed approach selects SVM as a base classifier. As we have known, SVM is one of the effective approaches for solving pattern recognition problems, which is an approximate implementation of the structural risk

minimization principal based on statistical learning theory (SLT) rather than the empirical risk minimization method [4].

The rest of the paper is organized as follows. Section 2 presents a comprehensive study on the class imbalance problem and discusses the existing class imbalance solutions. Section 3 gives a simple description of support vector machine and then proposes a hybrid sampling SVM approach for addressing class imbalance problem. In Section 4, we compare the performance of the proposed approach with the existing methods. Finally, Section 5 concludes this paper.

2. Related Work

Since many real applications have met the class imbalance problem, researchers have proposed several methods to solve this problem. In general, there are two kinds of approaches to cope with the class imbalance problem: data-level approaches and algorithmic approaches [2]. In this section, we will review some of the most effective methods that have been proposed within these two categories.

At the data-level approaches, resampling is one of the effective approaches which can obtain a more or less balanced class distribution. The resampling techniques try to balance out the dataset either randomly or deterministically, which include undersampling methods, oversampling methods, and hybrid methods.

Undersampling methods create a subset of the original dataset by randomly or selectively deleting some of the samples of the majority class while keeping the original population of the minority class [5, 6]. Despite the fact that this method results in information loss for the majority class, it must be noted that undersampling is generally quite successful at countering the class imbalance problem, especially when it uses sophisticated data elimination methods. EasyEnsemble and BalanceCascade proposed by Liu et al. [7] are two effective informed undersampling methods. Kim [8] proposes an undersampling method based on a self-organizing map (SOM) neural network to obtain sampling data which retains the original data characteristics. García and Herrera [9] present an evolutionary undersampling method for classification with imbalanced datasets. Yen and Lee [10] propose cluster-based undersampling approaches. Its basic idea is to select the representative data as training data, which improve the classification accuracy for minority class in the imbalanced class distribution environment.

Oversampling methods [11] generate a superset of the original dataset by replicating some of the samples of the positive class or creating new samples from the original positive class instances. A widely used oversampling technique is called SMOTE (synthetic minority oversampling technique) [11], which creates new synthetic samples to the minority class by randomly interpolating pairs of the closest neighbors in the minority class. SMOTE is effective to increase the significance of the positive class in the decision region. There exist many methods based on the SMOTE for generating more appropriate instances [12]. Borderline-SMOTE [13] is another approach based on the synthetic generation of instances proposed in SMOTE. Gao et al. [14] propose probability density function estimation based oversampling approach for two-class imbalanced classification problems. RWO-sampling [15] is a random walk oversampling approach to balance different class samples by creating synthetic samples through randomly walking from the real data. RWO-sampling also expands the minority class boundary after synthetic samples have been generated.

Hybrid methods use the oversampling technique combined with the undersampling technique to balance the class size. AdaOUBoost [16] adaptively oversamples the minority positive samples and undersamples the majority negative samples to form different subclassifiers and combines these subclassifiers according to their accuracy to create a strong classifier. Cateni et al. [17] present a new resampling approach to address the class imbalance problem, which combines a normal distribution-based oversampling technique and a similarity-based undersampling technique. Cao et al. [18] propose a hybrid probabilistic sampling combined with diverse random subspace ensemble for imbalanced data learning. Luengo et al. [19] analyze the usefulness of the data complexity measures and propose an approach based on SMOTE-based oversampling and evolutionary undersampling to deal with the class imbalance problem.

Algorithmic approach is another way to deal with the imbalanced data problem, which tries to modify the classifiers to suit the imbalanced datasets. Cost-sensitive learning is an effective solution based on algorithmic approaches, which can improve the performance of classification by setting different misclassification cost to the majority and minority datasets. In the cost-sensitive framework, the costs of misclassifying minority samples are higher with respect to other kinds of errors in order to encourage their correct classification. Cost-sensitive learning is one of the most important topics in machine learning and data mining and has attracted high attention in recent years [3, 20, 21]. Many algorithms combining resampling and cost-sensitive learning have also been proposed [22].

Many works make some modification of the classification algorithms. Several specific attempts using SVMs have been made to improve their class prediction accuracy in the case of class imbalances [23–26]. Fu and Lee [27] present a certainty-based active learning algorithm to deal with the imbalanced data classification problem. In order to improve the classification of imbalanced data, Oh [28] proposes a new error function for the error back-propagation algorithm of multilayer perceptrons.

As we have known, in recent years, an ensemble of classifiers have arisen as a possible solution to the class imbalance problem attracting great interest among researchers because of their flexible characteristics [29, 30]. Ensembles are designed to increase the accuracy of a single classifier by training several different classifiers and combining their decisions to output a single class label. Liu et al. [31] present an ensemble of SVMs to improve the prediction performance, which incorporates both oversampling and undersampling. Guo and Viktor [32] present an approach DataBoost-IM, which generated new data and classified imbalanced data by an ensemble classifier. Oh et al. [33] propose an ensemble

learning method combined with active example selection to deal with the class imbalance problem. Woźniak et al. [34] present ensemble classifiers from a new point of view including approaches to imbalanced data classification.

3. The Proposed Hybrid Sampling SVM Approach

In this section, we first give a description of support vector machine, and then we present our proposed approach.

3.1. Review of Support Vector Machine. SVM was first introduced to solve the pattern classification and regression problems by Vapnik and his colleagues [4, 35]. In recent years, SVM has drawn considerable attentions due to its high generalization ability of a wide range of applications and better performance than other traditional learning machines. The goal of the SVM learning algorithm is to find a separating hyperplane that separates these data points into two classes.

Consider a binary classification problem consisting of l training samples $\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_l, y_l)\}$, where $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{id})^T$ represents a d-dimensional data point and $y_i \in \{+1, -1\}$ denotes its class label. The decision boundary of a linear classifier can be written in the following form:

$$\mathbf{w} \cdot \mathbf{x} + b = 0, \tag{1}$$

where $\mathbf{w}$ and b are parameters of the model. To have more flexible ways to deal with nonlinear separable data, we can first transform the training samples into a high-dimensional feature space using a nonlinear mapping Φ. Therefore, (1) can be rewritten as $\mathbf{w} \cdot \Phi(\mathbf{x}) + b = 0$.

The support vector technique requires the solution of the following optimization problem:

$$\min_{\mathbf{w}} \frac{\|\mathbf{w}\|^2}{2} \tag{2}$$

subject to $y_i(\mathbf{w} \cdot \Phi(\mathbf{x}_i) + b) \geq 1$, for $i = 1, \ldots, l$.

This optimization problem can be solved by constructing a Lagrangian representation and transforming it into the following dual problem:

$$L_D = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j \Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j). \tag{3}$$

Once all α_i are found using quadratic programming techniques, we can use the KKT conditions to express the optimal primal variable $\mathbf{w}$ in terms of the optimal dual variables, according to $\mathbf{w} = \sum_{i=1}^{l} \alpha_i y_i \Phi(\mathbf{x}_i)$. Note that $\mathbf{w}$ depends only on the $\mathbf{x}_i$ for which $\alpha_i \neq 0$, which are called the *support vectors* (SVs).

When the optimal pair $(\mathbf{w}_0, b_0)$ is determined, the SVM decision function is then given by

$$f(\mathbf{x}) = \operatorname{sign}\left(\sum_{i \in \mathrm{SVs}} \alpha_i y_i \Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}) + b \right). \tag{4}$$

If $f(\mathbf{x}) = 1$, then the test sample $\mathbf{x}$ is classified as a positive class; otherwise, it is classified as a negative class.

Furthermore, the dot product $\Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j)$ in the transformed space can be expressed as the kernel function $K(\mathbf{x}_i, \mathbf{x}_j) = \Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j)$. Thus, the kernel is the key that determines the performance of the SVM. Several typical kernel functions are the lineal kernel $K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i \cdot \mathbf{x}_j$, the polynomial kernel $K(\mathbf{x}_i, \mathbf{x}_j) = (a\mathbf{x}_i^T \mathbf{x}_j + r)^d$, and the RBF kernel $K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2)$.

3.2. The Proposed Hybrid Sampling Approach. This paper proposes a hybrid sampling approach based on support vector machine to address the imbalanced data classification problem. The proposed approach first uses SVM method to generate a classification hyperplane and applies an undersampling technique to reduce negative samples which include less classification information. And then, we divide the training dataset into several subsets, in which we synthesize new positive samples using an oversampling technique. Once the majority class has been undersampled and the minority class has been oversampled, a new balanced training dataset is created and is used to train an SVM classifier. The proposed approach effectively balances the initial imbalanced dataset and improves classification precision on the basis of maximizing data balance.

The framework of our proposed approach is presented as follows. Given the training dataset $T = \{(x_1, y_1), (x_2, y_2), \ldots, (x_l, y_l)\}$, where $y_i \in \{-1, +1\}$ represents the class labels of the negative and positive samples, respectively, k is the size of the training dataset. Suppose that an imbalanced dataset contains n samples from the majority class and m samples from the minority class, where $n \gg m$ and $n + m = l$. The imbalance ratio IR is n/m.

In undersampling phase, the proposed approach first trains an SVM classifier for training dataset T and obtains a classification hyperplane $w \cdot x + b = 0$ and then deletes some negative samples with less information by undersampling. Our approach is based on the distance between sample z and the hyperplane as follows:

$$d = \|w \cdot z + b\|. \tag{5}$$

We proportionately delete some negative samples far away from the hyperplane according the calculated distances. After undersampling, the imbalance ratio is reduced; for instance, the imbalance ratio is the half of original IR. We label the training dataset after undersampling as T_0 including n_0 negative samples and m_0 positive samples.

In oversampling phase, the proposed approach first randomly divides the training dataset T_0 into k disjoint subsets $T_1, T_2, \ldots, T_k$. Each T_i contains n_i positive samples and m_i negative samples. We randomly select a subset, for instance, T_1, and oversample the positive samples using SMOTE method for subset T_1. Then, a new training dataset G_1 is generated by merging the new synthetic samples into T_1. We train an initial SVM classifier F_1 for dataset G_1.

For the ith one of the rest k–1 subsets, we select p_i negative samples near the hyperplane of classifier F_{i-1} according to

(5) and generate synthetic instances using SMOTE method. Similarly, a new training dataset G_i is generated by merging all n_i positive samples and the new synthetic samples into G_{i-1}. We train an SVM classifier F_i for dataset G_{i-1}.

In the oversampling for the positive class, the smaller the size of positive samples within a subset is, the more the instances are oversampled.

Based on the description above, the proposed hybrid sampling SVM approach is described in Algorithm 1.

4. Experiment and Analysis

In this section, we evaluate the performances for our proposed hybrid sampling approach on real datasets. In the following, we first describe several evaluation measures for class imbalanced problem and then compare F-measure and G-mean of our method with the other methods.

4.1. Evaluation Measures. In general, the performance of a classifier is evaluated based on its overall accuracy on an independent test dataset. However, the overall classification accuracy on an imbalanced dataset is mainly dominated by the majority class. Therefore, accuracy is not an appropriate evaluation measure for imbalanced data. Researchers use different metrics to evaluate the performance of imbalanced data classification methods. These metrics include the accuracy rate, F-measure, geometric mean (G-mean), and AUC [36].

The result of classification can be categorized into four cases as follows. TP (true positive) is the number of actual positives that were correctly classified as positives. FP (false positive) is the number of actual negatives that were incorrectly classified as positives. TN (true negative) is the number of actual negatives that were correctly classified as negatives. FN (false negative) is the number of actual positives that were incorrectly classified as negatives.

Accuracy is the most used evaluation metric for assessing the classification performance and guiding the classifier modeling. The overall accuracy is defined as

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}. \tag{6}$$

G-mean is the geometric mean of accuracies measured separately on each class, which is commonly utilized when performance of both classes is concerned and expected to be high simultaneously. G-mean is defined as

$$G\text{-mean} = \sqrt{S_{\text{sens}} \times S_{\text{spec}}}, \tag{7}$$

where S_{sens} and S_{spec} denote sensitivity and specificity, respectively. Sensitivity, also called the TP rate (*TPrate*) or the recall (*Recall*), shows the performance of the positive class as follows:

$$S_{\text{sens}} = TPrate = Recall = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{8}$$

Specificity, also called the TN rate (*TNrate*), shows the performance of the negative class as follows:

$$S_{\text{spec}} = TNrate = \frac{\text{TN}}{\text{TN} + \text{FP}}. \tag{9}$$

F-measure is often used in the fields of information retrieval and machine learning for measuring search, document classification, and query classification performance. F-measure considers both the *Precision* and the *Recall* to compute the score. Generally, for a classifier, if the *Precision* is high, then the *Recall* will be low; that is, the two criteria are trade-off. *Precision* and *Recall* are combined to form a criterion F-measure, which is shown in expression (10). Consider

$$F\text{-measure} = \frac{(1+\beta)^2 \times Precision \times Recall}{Precision + Recall}, \tag{10}$$

where β is set to 1 in this paper. The *Precision* for minority class is the correct-classified percentage of samples which are predicted as minority class by the classifier. It is defined as follows:

$$Precision = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{11}$$

AUC is the area under the receiver operating characteristic (ROC) curve. ROC consists of plotting the true positive rate as a function of the false positive rate along all possible threshold values for the classifier. An ROC curve depicts relative trade-offs between true positives and false positives across a range of thresholds of a classifier. However, it is difficult to compare several classification models through curves. Therefore, it is common for results to be reported with respect to AUC. AUC can be interpreted as the expected proportion of positive samples ranked before a uniformly drawn random negative sample [36].

In the following, we use the two criteria F-measure and G-mean discussed above to evaluate the performance of our approaches by comparing our methods with the other methods.

4.2. Experimental Results and Analysis. In this subsection, we will compare our proposed approach to address the class imbalance problem with several techniques. The experiments use 6 datasets which have different degrees of imbalance from KEEL [37], including *Cmc*2, *Glass*7, *Abalone*7, *Vowel*, *Yeast,* and *Letter*4. Information about these datasets is summarized in Table 1. They are very varied in their size of classes, size of attributes, size of samples, and imbalance ratio. When more than two classes exist in the dataset, the target class is considered to be positive and all the other classes are considered to be negative. For each dataset, the size of samples (number of samples), the size of attributes (number of attributes), the size of samples of each class (number of positives and number of negatives), and imbalance ratio are listed. We calculate class imbalance ratio of the size of the majority class to the size of the minority class.

We compared the performance of 4 methods, including undersampling (Under), SMOTE [11], EasyEnsemble [7], and

Input: Imbalanced training data set T_0
Output: An SVM classifier F_k

Step 1. Train an SVM classifier for training dataset T_0 and delete some negative samples using (5).
Step 2. Divide randomly T into k disjoint equivalent subsets $T_1, T_2, \ldots, T_k$.
Step 3. Select subset T_1 and over-sample the positive samples using SMOTE method; generate a new training data set G_1 by merging the new synthetic samples into T_1; train an initial SVM classifier F_1 for data set G_1.
Step 4. For each subset T_i in the rest $k-1$ subsets do
  Step 2.1. Compute the distances between negative samples and the hyperplane of classifier F_{i-1} according to (5).
  Step 2.2. Select p_i negative samples with the smallest distances; generate synthetic instances using SMOTE method.
  Step 2.3. Merge all n_i positive samples and the new synthetic samples into G_{i-1}, and obtain data set G_i.
  Step 2.4. Train an SVM classifier F_i for dataset G_i.
Step 5. Classify data set G_k using SVM method, and obtain a classifier F_k.

Algorithm 1: The hybrid sampling SVM method.

Table 1: Summary of data sets.

Data set	#Samples	#Attributes	#Positves	#Negatives	Imbalance ratio
Cmc2	1473	10	333	1140	3.4
Glass7	214	9	29	185	6.4
Abalone7	4177	8	391	3786	9.7
Vowel	990	13	90	900	10.0
Yeast	1332	8	84	1248	14.9
Letter4	20000	16	805	19195	23.8

Table 2: *F*-measure of the compared methods.

F-measure	Under	SMOTE	EasyEnsemble	Our approach
Cmc2	0.4308	0.4631	0.5267	**0.5491**
Glass7	0.8272	0.8539	**0.8965**	0.8867
Abalone7	0.2976	0.4250	0.5913	**0.5946**
Vowel	0.7632	0.7769	0.8431	**0.8752**
Yeast	0.6342	0.6281	0.7038	**0.7234**
Letter4	0.5308	0.5752	0.6076	**0.6614**

our proposed method. For undersampling, we use a random sampling method. SMOTE is used with five neighbours. EasyEnsemble selects C4.5 decision tree as the baseline classifier. In all our experiments, we perform a 10-fold cross validation.

In our experiments, *F*-measure and *G*-mean are used as metrics. Table 2 shows the average *F*-measure obtained by 4 methods. The results indicate that our proposed approach has higher *F*-measure than that of other compared methods on *Cmc2*, *Abalone7*, *Vowel*, *Yeast,* and *Letter4* datasets. EasyEnsemble outperforms other compared methods on *Glass7* dataset. The results indicate that our proposed approach can further improve the *F*-measure metric of imbalanced learning.

Table 3: *G*-mean of the compared methods.

G-mean	Under	SMOTE	EasyEnsemble	Our approach
Cmc2	0.5325	0.5461	0.5734	**0.5823**
Glass7	0.8432	0.8865	**0.9237**	0.9125
Abalone7	0.3524	0.4672	0.6613	**0.6653**
Vowel	0.8125	0.8333	0.8769	**0.9013**
Yeast	0.6631	0.6547	0.7532	**0.7845**
Letter4	0.5543	0.5945	0.6217	**0.6976**

Table 3 lists the results of the average *G*-mean of the compared methods. The results show that our proposed approach has higher *G*-mean than other compared methods on most of datasets, while EasyEnsemble is slightly higher *G*-mean than our proposed approach on *Glass7* dataset. This is consistent with our analysis in *F*-measure.

5. Conclusions

For the class imbalance problem, resampling technique is an effective approach to resolve it. This paper proposes a hybrid sampling SVM approach, which combines undersampling and oversampling techniques. The proposed approach generates a relatively balanced dataset without significant loss of information and without the addition of a great number of synthetic samples. Thus, SVM classifier employed by our proposed approach can effectively improve the classification accuracy of original imbalanced dataset. Experimental results show that the proposed approach outperforms existing oversampling and undersampling techniques.

Conflict of Interests

The author declares that there is no conflict of interests regarding the publication of this paper.

References

[1] Z.-H. Zhou and X.-Y. Liu, "Training cost-sensitive neural networks with methods addressing the class imbalance problem," *IEEE Transactions on Knowledge and Data Engineering*, vol. 18, no. 1, pp. 63–77, 2006.

[2] H. B. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, 2009.

[3] B. X. Wang and N. Japkowicz, "Boosting support vector machines for imbalanced data sets," *Knowledge and Information Systems*, vol. 25, no. 1, pp. 1–20, 2010.

[4] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer, New York, NY, USA, 1995.

[5] M. Kubat and S. Matwin, "Addressing the curse of imbalanced training sets: one-sided selection," in *Proceedings of the International Conference on Machine Learning*, pp. 179–186, Nashville, Tenn, USA, 1997.

[6] N. Japkowicz and S. Stephen, "The class imbalance problem: a systematic study," *Intelligent Data Analysis*, vol. 6, no. 5, pp. 429–450, 2002.

[7] X.-Y. Liu, J. Wu, and Z.-H. Zhou, "Exploratory undersampling for class-imbalance learning," *IEEE Transactions on Systems, Man, and Cybernetics B: Cybernetics*, vol. 39, no. 2, pp. 539–550, 2009.

[8] M. S. Kim, "An effective under-sampling method for class imbalance data problem," in *Proceedings of the 8th International Symposium on Advanced Intelligent Systems (ISIS '07)*, pp. 825–829, Sokcho-City, Republic of Korea, 2007.

[9] S. García and F. Herrera, "Evolutionary undersampling for classification with imbalanced datasets: proposals and taxonomy," *Evolutionary Computation*, vol. 17, no. 3, pp. 275–306, 2009.

[10] S.-J. Yen and Y.-S. Lee, "Cluster-based under-sampling approaches for imbalanced data distributions," *Expert Systems with Applications*, vol. 36, no. 3, pp. 5718–5727, 2009.

[11] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, no. 1, pp. 321–357, 2002.

[12] N. V. Chawla, A. Lazarevic, L. O. Hall, and K. W. Bowyer, "SMOTEBoost: improving prediction of the minority class in boosting," in *Proceedings of the 7th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD '03)*, pp. 107–119, Cavtat, Croatia, September 2003.

[13] H. Han, W. Wang, and B. Mao, "Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning," in *Proceedings of the International Conference on Intelligent Computing (ICIC '05)*, vol. 3644 of *Lecture Notes in Computer Science*, pp. 878–887, Hefei, China, August 2005.

[14] M. Gao, X. Hong, S. Chen, and C. J. Harris, "Probability density function estimation based over-sampling for imbalanced two-class problems," in *Proceedings of the International Joint Conference on Neural Networks (IJCNN '12)*, Brisbane, Australia, June 2012.

[15] H. X. Zhang and M. F. Li, "RWO-Sampling: a random walk over-sampling approach to imbalanced data classification," *Information Fusion*, vol. 20, pp. 99–116, 2014.

[16] Y. X. Peng and J. Yao, "AdaOUBoost: adaptive over-sampling and under-sampling to boost the concept learning in large scale imbalanced data sets," in *Proceedings of the ACM SIGMM International Conference on Multimedia Information Retrieval (MIR '10)*, pp. 111–118, Philadelphia, Pa, USA, March 2010.

[17] S. Cateni, V. Colla, and M. Vannucci, "A method for resampling imbalanced datasets in binary classification tasks for real-world problems," *Neurocomputing*, vol. 135, pp. 32–41, 2014.

[18] P. Cao, J. Z. Yang, W. Li, D. Z. Zhao, and O. Zaiane, "Ensemble-based hybrid probabilistic sampling for imbalanced datalearning in lung nodule CAD," *Computerized Medical Imaging and Graphics*, vol. 38, no. 3, pp. 137–150, 2014.

[19] J. Luengo, A. Fernández, S. García, and F. Herrera, "Addressing data complexity for imbalanced data sets: analysis of SMOTE-based oversampling and evolutionary undersampling," *Soft Computing*, vol. 15, no. 10, pp. 1909–1936, 2011.

[20] Y. Sun, M. S. Kamel, A. K. C. Wong, and Y. Wang, "Cost-sensitive boosting for classification of imbalanced data," *Pattern Recognition*, vol. 40, no. 12, pp. 3358–3378, 2007.

[21] J. P. Hwang, S. Park, and E. Kim, "A new weighted approach to imbalanced data classification problem via support vector machine with quadratic cost function," *Expert Systems with Applications*, vol. 38, no. 7, pp. 8580–8585, 2011.

[22] S. Z. Wang, Z. J. Li, W. H. Chao, and Q. H. Cao, "Applying adaptive over-sampling technique based on data density and cost-sensitive SVM to imbalanced learning," in *Proceedings of the International Joint Conference on Neural Networks (IJCNN '12)*, Brisbane, Australia, June 2012.

[23] R. Akbani, S. Kwek, and N. Japkowicz, "Applying support vector machines to imbalanced datasets," in *Proceedings of the 15th European Conference on Machine Learning (ECML '04)*, pp. 39–50, Pisa, Italy, September 2004.

[24] Y. Tang, Y.-Q. Zhang, and N. V. Chawla, "SVMs modeling for highly imbalanced classification," *IEEE Transactions on Systems, Man, and Cybernetics B: Cybernetics*, vol. 39, no. 1, pp. 281–288, 2009.

[25] R. Batuwita and V. Palade, "FSVM-CIL: fuzzy support vector machines for class imbalance learning," *IEEE Transactions on Fuzzy Systems*, vol. 18, no. 3, pp. 558–571, 2010.

[26] Y. H. Shao, W. J. Chen, J. J. Zhang, Z. Wang, and N. Y. Deng, "An efficient weighted Lagrangian twin support vector machine for imbalanced data classification," *Pattern Recognition*, vol. 47, no. 9, pp. 3158–3167, 2014.

[27] J. H. Fu and S. L. Lee, "Certainty-based active learning for sampling imbalanced datasets," *Neurocomputing*, vol. 119, pp. 350–358, 2013.

[28] S.-H. Oh, "Error back-propagation algorithm for classification of imbalanced data," *Neurocomputing*, vol. 74, no. 6, pp. 1058–1061, 2011.

[29] B. Krawczyk, M. Wozniak, and G. Schaefer, "Cost-sensitive decision tree ensembles for effective imbalanced classification," *Applied Soft Computing*, vol. 14, pp. 554–562, 2014.

[30] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera, "A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches," *IEEE Transactions on Systems, Man and Cybernetics C: Applications and Reviews*, vol. 42, no. 4, pp. 463–484, 2012.

[31] Y. Liu, X. H. Yu, J. X. Huang, and A. J. An, "Combining integrated sampling with SVM ensembles for learning from imbalanced datasets," *Information Processing & Management*, vol. 47, no. 4, pp. 617–631, 2011.

[32] H. Guo and H. L. Viktor, "Learning from imbalanced data sets with boosting and data generation: the DataBoost-IM approach," *SIGKDD Explorations*, vol. 6, no. 1, pp. 30–39, 2004.

[33] S. Oh, M. S. Lee, and B. T. Zhang, "Ensemble learning with active example selection for imbalanced biomedical data classification," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 8, no. 2, pp. 316–325, 2011.

[34] M. Woźniak, M. Grana, and E. Corchado, "A survey of multiple classifier systems as hybrid systems," *Information Fusion*, vol. 16, pp. 3–17, 2014.

[35] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "Training algorithm for optimal margin classifiers," in *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*, pp. 144–152, Pittsburgh, Pa, USA, July 1992.

[36] P. Flach, "ROC analysis for ranking and probability estimation," Notes from the UAI tutorial on ROC analysis, 2007.

[37] J. Alcalá-Fdez, A. Fernández, J. Luengo et al., "KEEL data-mining software tool: data set repository, integration of algorithms and experimental analysis framework," *Journal of Multiple-Valued Logic and Soft Computing*, vol. 17, no. 2-3, pp. 255–287, 2011.