Improving the Classification Quality of the SVM Classifier for the Imbalanced Datasets on the Base of Ideas the SMOTE Algorithm

Liliya Demidova [1,2,*], and *Irina* Klyueva[2]

[1]Moscow Technological Institute, 119334 Moscow, Russia
[2]State Radio Engineering University, 390005 Ryazan, Russia

Abstract. The approach to the classification problem of the imbalanced datasets has been considered. The aim of this research is to determine the effectiveness of the SMOTE algorithm, when it is necessary to improve the classification quality of the SVM classifier, which is applied for classification of the imbalanced datasets. The experimental results which demonstrate the improvement of the SVM classifier quality with application of ideas the SMOTE algorithm for the imbalanced datasets in the sphere of medical diagnostics have been given.

1 Introduction

The problem of imbalanced data is one of the main problems which must be solved before the application of some machine learning classification algorithm if we want to receive the high quality classification results.

Dataset is called imbalanced if the samples size from one class is very much smaller or larger than the other classes.

Training the classifiers for the imbalanced datasets compromises the performance of most well-known machine learning algorithms. It is fair, in particular, for the support vector machine algorithm (SVM, Support Vector Machine) [1 – 9].

The incorrect classification of objects of the minority class usually costs significantly more than the incorrect classification of the object of the majority class since the minority class instances are rare, but the most important data used in real data sets.

As shown by experimental studies, the training of the classifiers on the imbalanced datasets leads to the fact that the constructed classifier tend to classify all objects as objects of the majority class, completely ignoring the underrepresented minority class, which generally does not correspond to the actual purpose of the research [3 – 6].

There are a significant number of real-world applications that are suffering from the class imbalance problem (for example, medical and fault diagnostics, anomaly detection, face recognition, telecommunication, the web and email classification, ecology, biology and financial services. For example, in medical diagnostics the number of sick patients is usually significantly less then the number of healthy people.

Currently, the different strategies of sampling are applied to solve the problem of the imbalance datasets. In this paper the study of the aspects of the applicability of the sampling strategies to restore the balance between the classes in the problem of binary classification has been performed. In particular, the capabilities of the synthetic sampling algorithm called as the SMOTE (Synthetic Minority Oversampling Technique) [3] have been investigated.

2 The basic principles of the SVM classifier development

The SVM algorithm proposed by Vapnik [1, 4] is a modern approach for solving the pattern recognition problems. The SVM algorithm maps the sample points into a highdimensional feature space to seek for an optimal separating hyperplane through maximizing the margin between two classes.

The Support Vector Machine (SVM) algorithm is the supervised machine learning algorithm [1 – 9]. The SVM algorithm is successfully used for the different classification problems in various applications [8]. The SVM classifiers on the base of the SVM algorithm have been applied for credit risk analysis, medical diagnostics, text categorization, information extraction, etc [8].

To develop the best SVM classifier it is necessary to find correctly the kernel function type, values of the kernel function parameters and value of the regularization parameter [8]. The solution of this problem can be achieved by the grid search of the kernel function types, values of the kernel function parameters and value of the regularization parameter that demands significant computational expenses. Quality of the SVM classifier can be measured by different classification quality indicators. There are the cross validation data indicator, the accuracy indicator, the classification completeness indicator and the ROC curve analysis based indicator, etc [8].

* Corresponding author: liliya.demidova@rambler.ru

The SVM classifier with satisfactory training and testing results can be used to classify new objects.

The separating hyperplane for the objects from the training set can be represented by equation $\langle w,z\rangle + b = 0$, where w is a vector-perpendicular to the separating hyperplane; b is a parameter which corresponds to the shortest distance from the origin of coordinates to the hyperplane; $\langle w,z\rangle$ is a scalar product of vectors w and z.

The condition $-1 < \langle w,z\rangle + b < 1$ specifies a strip that separates the classes. The wider the strip, the more confidently we can classify objects. The objects closest to the separating hyperplane, are exactly on the bounders of the strip.

Finding the separating hyperplane is basically the dual problem of searching a saddle point of the Lagrange function, which reduces to the problem of quadratic programming, containing only dual variables [8].

In training of the SVM classifier it is necessary to determine the kernel function type $\kappa(z_i, z_\tau)$, values of the kernel parameters and value of the regularization parameter C, which allows finding a compromise between maximizing of the gap separating the classes and minimizing of the total error [8].

One of the approaches using for the search of the optimal values of the parameters of the SVM classifier is based on the application of the Particle Swarm Optimization algorithm (PSO algorithm) [7 – 9].

Search space in the PSO algorithm is filled with a population of particles each of which has some location and velocity in the space of the problem parameters at the concrete moment of time.

Corresponding value of the objective function is calculated for each particle location. Particle location and velocity is changed after calculation of a new value of the objective function.

After every iteration under determination of the following particle location, information on the best particle from a number of neighboring particles (particles can share information) and also information on this particle location during that iteration when the best value of the objective function corresponds to this particle (particles have "memory"), are taken into account [7].

In this research we used the canonical version of the PSO algorithm [8].

3 The basic ideas of the SMOTE algorithm

A dataset is imbalanced if the classes are not approximately equally represented.

The SMOTE algorithm [3] creates the artificial objects of the minority class based on the similarities in the feature space between the existing objects using the k-nearest neighbor algorithm (kNN alrorithm) (Fig. 1) [10].

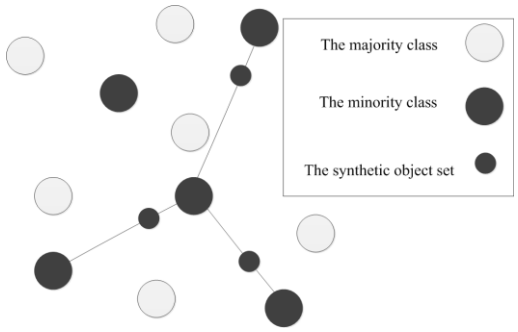

Fig. 1. The diagram of SMOTE algorithm.

Herewith, the number of the artificial objects which are "similar" to the objects of the existing minority class, but do not duplicate them is generated.

The SMOTE technique is an important approach by oversampling the minority class.

4 The experimental study

In this research, we implemented the following approach to application of the SMOTE to improve the classification quality of the SVM classifier for imbalanced datasets.

Step 1. The initial dataset is splited into the train and test datasets.

Step 2. The new synthetic objects are generated by the SMOTE algorithm for each dataset obtained in step 1.

Step 3. The SVM algorithm is implemented for the new received datasets. Herewith, the search for the optimum parameters values of the SVM classifier is implemented by the PSO algorithm (in particular, the PSO algorithm is used for the search of two optimal parameters values of the SVM classifier for the radial basis kernel function: the values of the parameter regularization C and the kernel function parameter σ) [8].

Python 2.7 was used for software implementation of the SMOTE and SVM algorithms. Herewith, the default settings were applied for the SMOTE algorithm.

We use the real medical datasets from the UCI repository of the machine learning database [4] to demonstrate the classification performance of the approach proposed in this paper.

Table 1. The characteristics of the datasets.

Dataset	Total number of the positive data	Total number of the negative data	Imbalance ratio
Heart (270 × 13)	120	150	0.2
Hepatitis (155 × 19)	123	32	0.74
Pima diabetes (768 × 8)	268	500	0.46

Table 2. The classification results on the "Heart" dataset.

Type of algorithm	The size of the training/testing set	Errors						Accuracy, (%)	Sensitivity, (%)	Specificity, (%)
		At the training			At the testing					
		positive	negative	total	positive	negative	total			
1	216/ 54	4	2	**6**	4	3	**7**	95.18	93.33	96.67
2	252/ 60	2	2	**4**	4	5	**9**	95.83	96.15	95.51
1	216/ 54	6	4	**10**	5	4	**9**	92.96	94.67	90.83
2	252/ 60	1	1	**2**	0	8	**8**	96.67	94.00	99.33

Table 3. The classification results on the "Hepatitis" dataset.

Type of algorithm	The size of the training/testing set	Errors						Accuracy, (%)	Sensitivity, (%)	Specificity, (%)
		At the training			At the testing					
		positive	negative	total	positive	negative	total			
1	124/ 31	0	17	**17**	0	11	**11**	81.94	100	12.50
2	208/ 38	0	0	**0**	2	8	**10**	95.93	98.37	93.50

Table 4. Classification results on the "Pima diabetes" dataset.

Type of algorithm	The size of the training/testing set	Errors						Accuracy, (%)	Sensitivity, (%)	Specificity, (%)
		At the training			At the testing					
		positive	negative	total	positive	negative	total			
1	614/ 154	63	20	**83**	22	15	**37**	84.38	68.28	93.00
2	800/ 200	15	34	**49**	17	19	**36**	91.50	93.60	89.40

These free medical datasets are "Heart", "Hepatitis" and "Pima diabetes". In each of them, the positive class consists of the data corresponding to the healthy, normal, or benign cases, while the negative class contains the data for the diseased, abnormal, or malignant cases. Further details of these datasets are provided in Tabl. 1.

In Tabl. 1 the value of imbalance ratio (Ratio) was calculated by the following formula:

$$Ratio = 1 - \frac{a_1}{a_2},$$

where a_1 is the number of objects in the minority class; a_2 is the number of objects in the majority class.

The results of application of the SVM-PSO algorithm and the Smote-SVM-PSO algorithm are shown in Tabl. 2, Tabl. 3, and Tabl. 4 (with the following type of algorithms: 1 is the SVM-PSO algorithm, 2 is the Smote-SVM-PSO algorithm).

For the "Heart" dataset the class imbalance is not obviously shaped, therefore, the results of the SVM-PSO algorithm is not very different from the Smote-SVMPSO algorithm. Herewith, Tabl. 1 shows the results for the "Heart" dataset for two cases depending on the choice of datasets for training and testing.

We can say that the effectiveness of the SVM classifier can be improved indeed when the structure of the data is taken into consideration.

For the other datasets the class imbalance is considerable, therefore the SVM-PSO algorithm concedes to the Smote-SVM-PSO algorithm, as the results of the SVM-PSO algorithm are characterized by the low values of accuracy, specificity and sensitivity. These results justify the fact that standard SVM algorithm are sensitive to the class imbalance problem.

The obtained results correspond to the implementation of the SMOTE algorithm with the default parameters values used in the Python library.

Also, we suggested the searching algorithm for the optimum parameters values of the SMOTE algorithm. In particular, we considered two parameters: the number k of nearest neighbours to used to construct synthetic samples; the number m of nearest neighbours to use to determine if a minority sample is in danger.

The suggested searching algorithm can be described by the following sequence of steps.

Step 1. To generate the pairs (k_i, m_j) on the base of the integer parameters values from the ranges $[k_{min}, k_{max}]$ and $[m_{min}, m_{max}]$ ($i = \overline{1, k_{max} - k_{min} + 1}$; $j = \overline{1, m_{max} - m_{min} + 1}$).

Step 2. To build for each pair (k_i, m_j) n SVM classifiers using the SMOTE algorithm for the imbalanced data (that is to apply the SMOTE algorithm for each pair (k_i, m_j) with equal probability).

Step 3. To evaluate the classification quality of the developed SVM classifiers and save the obtained SVM classifiers. To find the best SVM classifier, if the maximum value of iteration is achieved, and finish the algorithm. Otherwise, to go to step 4.

Step 4. To estimate the average classification quality of the SVM classifiers using, for example, the *F*-measure indicator for each pair (k_i, m_j). To change the probabilities of application for each pair (k_i, m_j): to increase the probability for the best pair (k_i, m_j) (with the maximum value of the average classification quality), and to decrease the probabilities for the other pairs (k_i, m_j). To build for each pair (k_i, m_j) the SVM classifiers using the SMOTE algorithm for the

imbalanced data according to the new probabilities. Go to step 3.

It is proposed to use the following ideas at the step 4 to estimate the average classification quality of the SVM classifiers for each pair (k_i, m_j):

- to find the total number N_{ij}^g of the SVM classifiers for each pair (k_i, m_j), obtained to the current number g of iteration of the suggested algorithm;
- to find the total sum S_{ij}^g of the classification quality of the SVM classifiers for each pair (k_i, m_j), obtained to the current number g of iteration of the suggested algorithm;
- to find the ratio S_{ij}^g / N_{ij}^g for each pair (k_i, m_j), and use it as the average classification quality of the SVM classifiers for pair (k_i, m_j).

It is necessary to say, that we generate the different balanced datasets, using the random number generator for each pair (k_i, m_j), therefore, the developed SVM classifiers will be differ from each other.

The offered algorithm allows minimizing the time expenditures for the search of the optimal parameters values of the SMOTE algorithm, and, hence, for development of the SVM classifier.

5 Conclusion

The experimental results show that the SMOTE algorithm improves the classification quality of the SVM classifiers for the imbalanced data. Herewith, for the datasets with the high imbalance, the effectiveness of the SMOTE algorithm is very high.

The offered algorithm allows minimizing the time expenditures for the search of the optimal parameters values of the SMOTE algorithm.

Further, we plan to consider the values of several classification quality indicators simultaneously at the choice of the optimal parameters values of the SMOTE algorithm.

References

1. V.N. Vapnik, *Statistical Learning Theory*, (Wiley, 1998)
2. L. Yu, S. Wang, K. K. Lai, L. Zhou, *Bio-Inspired Credit Risk Analysis. Computational Intelligence with Support Vector Machines* (Springer, 2008)
3. N. Chawla, K. Bowyer, L.Hall, W. Kegelmeyer, JIAR, **16**, 341-378 (2002)
4. X. Gu, T. Ni, H. Wang, Scientific World Journal, **2014**, 102-113 (2014)
5. S. Zafeiriou, A. Tefas, I. Pitas, IEEE Transactions on Image Processing, **16**(10), 2551-2564 (2007)
6. R. Batuwita, V. Palade, IEEE Transactions on Fuzzy Systems, **18**(3), 558-571 (2010)
7. L. Demidova, I. Klyueva, A. Pylkin, 5-th Mediterranean Conference on Embedded Computing (MECO'2016), 322-325 (2016)
8. L. Demidova, I. Klyueva, Y. Sokolova, N. Stepanov, N. Tyart, *Procedia Computer Science*, **103**, 222-230 (2017)
9. L. Demidova, E. Nikulchev, Y. Sokolova, International Journal of Advanced Computer Science and Applications, **7**(5), 294 (2017)
10. H. Wang, D. Bell, Computer Journal, **47**(6), 662-672 (2004)