The two-stage classification based on 1-SVM and RF classifiers

Liliya Demidova¹ and Irina Klyueva²

¹ MIREA – Russian Technological University, 119454 Moscow, Vernadsky avenue, 78, Russia
² Ryazan State Radio Engineering University, 390005 Ryazan, Gagarin street, 59/1, Russia

e-mail: liliya.demidova@rambler.ru

Abstract. The approach to the two-stage classification based on the 1-SVM classifier used as the main classifier and the RF classifier used as the auxiliary classifier has been considered. The proposed approach improves the quality of classification when using imbalanced datasets. The results of the comparative analysis of the proposed approach and the alternative approach to the two-stage classification, in which the binary SVM classifier is used as the main classifier, and the RF classifier as the auxiliary classifier, are presented.

1. Introduction

When solving the classification problem, the quality of the classifier is influenced by the level of class balance. At the same time, if most probabilistic models are weakly dependent on class balances, then when using improbability models, in particular, SVM models (Support Vector Machine) [1–10], the class imbalance problem [5–7] is relevant.

During training the SVM algorithm builds separating hyperplane so that in the case of binary classification, approximately the same number (or a comparable number) of objects of positive and negative classes is located on the dividing band. Therefore, changing the balance of classes can affect this number, and therefore the position of the border between them.

When the class imbalance level is 10:1 or more, the classification result can have high accuracy value if the minority class objects are incorrectly classified.

The following approaches can be used to solve the problem of class imbalance:
- class weighting, in which the correct classification of objects of the minority class is the most preferred;
- sampling (oversampling, undersampling, or a combination of them);
- transition from the class balancing problem to predicting the minority class using outlier detection methods.

One of the well-known algorithms for detecting anomalies is the One Class SVM algorithm (1-SVM algorithm) [11–14].

This paper presents the two-stage classification algorithm based on the implementation of the 1-SVM classifier using the auxiliary RF classifier (Random Forest) [15].
2. Class imbalance problem

Most standard machine learning algorithms use balanced training datasets with equal classification error values for all available examples when developing classifiers. The problem of training on imbalanced datasets (imbalance problem) [5-7] means the possibility of a significant decrease in the quality of the developed classifiers, since such datasets do not provide the required characteristics of data distribution during the training. The problem of learning from imbalanced datasets is a fairly common topic for research in recent years and quite often requires solutions in various areas of data analysis.

In general, imbalanced data can be called any data whose number of objects in classes is not equal (even such a ratio of the number of objects in classes as 51:49 allows to classify the corresponding dataset as the imbalanced set). However, usually when stating the fact of an imbalanced dataset, we must consider the ratio of the number of objects in classes equal to 10:1 or more, since it is precisely at these proportions that the most pronounced distortions characteristic of the learning process on imbalanced datasets are manifested.

If the objects in the dataset are not balanced, there is a high probability that solving the classification problem will lead to erroneous results. In particular, we can develop a classifier that provides high classification accuracy, such as 99% if the dataset is imbalanced in a ratio of 1:99, while ignoring 1% of objects of the minority class. However, such classifier is not able to correctly classify any objects from the dataset (in particular, the classification of objects belonging to the minority class).

The essence of imbalanced datasets, in particular in binary classification, is as follows: more objects in the initial dataset belong to one class (the "majority" class) and a much smaller number of objects belong to another class (the "minority" class).

According to basic assumptions contained in most algorithms, the goal of training is to maximize the proportion of correct decisions relative to all decisions made, herewith, the training data and the general population must have the same distribution. However, the desire to take these assumptions into account in the presence of an imbalanced dataset leads to the fact that the classifier is not able to classify data according to the algorithm better than a trivial classifier that completely ignores the less represented class and marks all objects as belonging to the majority class.

It should be noted that the costs of misclassification vary. Thus, the cost of incorrectly classifying examples of a minority class is often much more expensive than incorrectly classifying examples of a majority class, because in actual datasets, objects of a minority class are rare, but the most important examples. Thus, the correct classification of minority class examples is of greater interest.

3. Detection of anomalies

In data analysis, there are two areas which search for anomalies [11, 14]: outlier detection and novelty detection. Like an outlier, a "new object" is an object which differs in its properties from the objects in the set (training set). But unlike an outlier, it is not in the set itself yet (it will appear after a while, and the task is just to detect it when it appears). For example, if abnormally large or small temperature values are discarded when analyzing temperature measurements, this is an example of emission control. And if for a new measurement, it is estimated how similar it is to the previous ones, and anomalous ones are thrown out – this is an example of determining "novelty”.

Novelty usually results from a fundamentally new behavior of the object (for example, if the objects are the descriptions of the system, then after the virus enters it, the objects become "new"; another example is a description of the engine operation after a failure). "Newness" is defined by a completely new description of objects, which is labor-intensive or does not make sense to add to the training sample (for example, information about all sorts of virus infections or all sorts of breakdowns). In this case, you can get a fairly large sample of the normal (regular) operation of the system or mechanism.

Anomalies occur in various data analysis applications, for example, in such as:
- detection of suspicious banking transactions (credit-card fraud);
- intrusion detection (intrusion detection);
- detection of non-standard players on the exchange (insiders);
- detection of problems in mechanisms based on sensor readings;
- medical diagnosis (medical diagnosis);
- seismology.

4. One-class SVM

The idea of the 1-SVM algorithm [11-14] is to detect novelty, in other words, to detect the rare events, or anomalies. The rare events do not occur often (for example, equipment failures are relatively rare), and therefore they have few samples (examples) in the classified data, and the problem of detecting anomalies can be interpreted as the classification problem with an imbalance of classes.

The task of implementing the 1-SVM algorithm is to separate the main part of the objects of dataset which are considered to belong to the normal class from the rest of the objects of dataset which are considered abnormal in one sense or another. Those objects from the dataset are considered anomalous for which the corresponding description vectors are close to the origin in the characteristic space (Fig.1).

Let \( Z = \{ z_1, \ldots, z_s \} \) is the initial dataset, in which each object \( z_i \) corresponds to a characteristic description \( z_i = (z_{i1}, z_{i2}, \ldots, z_{in}) \) \( z_i \in Z \) \( (i = 1, s; \ s \) is the number of objects in the initial dataset) in the \( n \)-dimensional characteristic space. Let \( Z^* = \{ z_1^*, \ldots, z_s^* \} \) is the training subset taken from the objects of the initial dataset, a herewith, the objects \( Z^* = \{ z_1^*, \ldots, z_s^* \} \) belong to normal class \( (i = 1, S; \ S \) is the number of objects in the training subset, \( s > S \). Let it be required for objects which are not included in the training subset (test subset or new objects) to determine whether they belong to a normal class, or are anomalies.

In the process of building a 1-SVM classifier, the optimization problem of the following type is solved:

\[
\begin{aligned}
\frac{1}{2} ||w||^2 + \frac{1}{\nu S} \sum_{i=1}^{S} \xi_i - \rho \rightarrow \min, \\
\langle w, z_i \rangle \geq \rho - \xi_i, \quad \xi_i \geq 0, \quad i = 1, S
\end{aligned}
\]

where \( z_i \) is an object from the training subset, \( i = 1, S; \quad S \) \( (s > S) \) is the number of objects in the training subset; \( \kappa(z_i, z_j) \) is the kernel function; \( \nu \) is the maximum percentage of objects in the training subset which can be recognized as anomalies (\( \nu \) sets the value of the regularization parameter \( C \), in particular, \( C = \frac{1}{\nu S} \)); \( w \) is a vector perpendicular to the dividing hyperplane; \( \xi_i \geq 0 \) is a penalty for going beyond the dividing band, which characterizes the error value on the object \( z_i \); \( \rho \) is an algorithm parameter.

The decisive rule has the form:

\[
F(z) = \text{sign} \left( \sum_{i=1}^{S} \alpha_i \kappa(z_i, z) - \rho \right).
\]
5. **Two-stage classification based on the 1-SVM and the RF classifiers**

In [8-10], the two-stage binary classification algorithms based on the use of the classical SVM classifier are studied: first, an area containing both incorrectly and correctly classified objects (α-area) is allocated near the class-separating hyperplane constructed using the SVM classifier, and then for all objects from this α-area, their class affiliation is clarified using various auxiliary classifiers: the kNN classifier or the RF classifier.

It should be noted that in a number of cases the accuracy of classification solutions can be significantly increased by using the evolutionary algorithm [6, 16, 17] to find the optimal values of the parameters of the SVM classifier being developed. However, this is time consuming and may not provide the desired effect in the case when classification errors occur near the hyperplane separating the classes.

The effectiveness of such algorithms for object classification for a number of datasets of different origin has been experimentally proved. When refining the results of classifying objects near the hyperplane separating classes, the RF algorithm showed more confident results of improving SVM classification for arbitrary multidimensional datasets compared to the kNN classifier. However, the application of the RF-algorithm requires some increase in time.

The limitation of the applicability of previously proposed two-stage classification algorithms is as follows: since the α-area contains both incorrectly classified objects and correctly classified objects which fall into this area, if the width of the α-area is large enough or the objects inside the α-area are strongly compacted, the number of objects used for training the auxiliary classifier may not be sufficient to solve the problem of improving the quality of the binary SVM classification.

In this regard, in this paper, to solve the problem of classifying datasets with imbalanced classes, it is proposed to use the 1-SVM classifier using the auxiliary RF classifier.

Let \( Z = \{z_1, ..., z_s\} \ (i = 1, s) \) is the initial dataset with imbalanced classes. For the 1-SVM classification, we accept the majority class objects from the \( Z \) dataset as the normal class objects, and the minority class objects from the \( Z \) dataset as the anomalies.

The proposed two-stage classification algorithm can be described by the following sequence of steps.

1. To generate the training and test subsets based on the \( Z \) dataset.
2. To implement step of the 1-SVM classification: to train the 1-SVM classifier using objects of only the normal class from the training subset and make a prediction of the class of objects from the test subset (the normal class or the class of anomalies).
3. For the objects which were identified as the anomalies at the step 2, use the auxiliary RF classifier to clarify their class affiliation. To implement step of the RF classification: to train the RF classifier on all objects in the training subset and predict the class of objects-anomalies.

4. To combine the results of the 1-SVM and RF classifications: for objects assigned to the normal class at the step 2 accept the result of the 1-SVM classification; for objects-anomalies accept the RF classification result obtained at the step 3.

5. To evaluate the quality of the two-stage classification.

6. Results of experimental studies
Software implementation of the algorithms was performed using the high-level Python programming language (Python 3.7 programming environment) and the Scikit-Learn machine learning library.

The research was conducted using the binary classification datasets borrowed from the UCI machine learning library. In this work, we used the datasets shown in Table 1.

| Dataset      | Size       | Number of objects of the majority class | Number of objects of the minority class |
|--------------|------------|-----------------------------------------|----------------------------------------|
| SportsArticles | 1000×59    | 635                                     | 365                                    |
| Spam         | 4601×57    | 2788                                    | 1813                                   |
| Biodeg       | 1055×41    | 699                                     | 356                                    |
| WDBC         | 699×10     | 458                                     | 241                                    |

Table 1 shows the results of research for the simple implementation of the algorithms (1-SVM algorithm, binary SVM algorithm, RF algorithm); the previously studied implementation of the two-stage classification based on the binary SVM and RF algorithms (SVM-RF); the implementation of the two-stage classification based on the 1-SVM and RF algorithms (1-SVM-RF), considered in this paper.

In Table 2, the classification results are presented based on the obtained values of the balanced accuracy indicator (BA) for the test sample.

The formula for calculating the indicator in the case of binary classification is presented below:

\[
BA = \frac{Se + Sp}{2} = \frac{1}{2} \left( \frac{TP}{TP + FN} + \frac{TN}{TN + FP} \right)
\]

where \(TP\) is the number of true positive observations; \(TN\) is the number of true negative observations; \(FP\) is the number of false positive observations ("false detections", type I error); \(FN\) is the number of false negative observations ("false omissions", type II error).

The \(BA\) indicator differs from the overall accuracy (the total proportion of correct answers) that is considered in calculating the proportion of objects in each class. Accordingly, the \(BA\) indicator reflects the most accurate assessment of the quality of classification imbalanced data sets.

The "Number of successful runs" columns represent the number of runs out of the total number of runs \(N_{runs} = 50\) in which the two-step classification algorithms (1-SVM-RF and SVM-RF) provided the increase in the accuracy of the SVM classification. Each run of the algorithms used a different random split of the original dataset into the training and test subsets. The size of the test subset was 20% of the size of the initial dataset.

The experiments were performed for the different values of the parameter 1-SVM algorithm \(\nu\). For the binary SVM and RF algorithms, the parameters set by default in the Scikit-Learn library were used.
Table 2. The results of experiments.

| Datasets     | V  | 1-SVM, (BA,%) | Binary SVM (BA,%) | RF (BA,%) | 1-SVM-RF (BA,%) | Number of successful runs 1-SVM-RF | Number of successful runs SVM-RF |
|--------------|----|---------------|------------------|----------|----------------|-----------------------------------|----------------------------------|
|              | 0.1| 69,25         | 84,67            | 85,84    | 85,99          | 23                                | 85,60                            | 4                                |
| SportsArticles | 0.25 | 65,91       | 82,55            | 84,02    | 86,23          | 17                                |                                   |                                  |
|              | 0.5 | 63,85        | 84,11            | 85,11    | 86,68          | 24                                |                                   |                                  |
|              | 0.1 | 53,89        | 92,60            | 95,80    | 95,88          | 7                                 |                                   |                                  |
| Spam         | 0.25 | 54,77       | 91,18            | 94,59    | 94,95          | 11                                | -                                | -                                |
|              | 0.5 | 54,29        | 90,28            | 95,50    | 95,97          | 6                                 |                                   |                                  |
|              | 0.1 | 57,38        | 88,08            | 91,73    | 92,07          | 10                                |                                   |                                  |
| Biodeg       | 0.25 | 55,04       | 86,20            | 87,62    | 87,63          | 8                                 | -                                | -                                |
|              | 0.5 | 57,73        | 88,08            | 89,18    | 90,15          | 7                                 |                                   |                                  |
|              | 0.1 | 96,11        | 98,89            | 99,00    | 100,00         | 24                                |                                   |                                  |
| WDBC         | 0.25 | 89,01       | 98,90            | 98,90    | 99,45          | 22                                | 100,00                           | 30                               |
|              | 0.5 | 74,73        | 98,90            | 98,90    | 99,45          | 23                                |                                   |                                  |

The obtained results demonstrate that the combined use of the 1-SVM classifier with the auxiliary RF classifier can increase the accuracy of SVM classification when using the imbalanced datasets by an average of 2.5%.

The results of experimental studies of the SVM-RF algorithm in some cases did not show the clear improvements in the value of the balanced accuracy indicator (in Table 2, these cases correspond to the Spam and Biodeg datasets), since the number of objects used for training the auxiliary RF classifier was insufficient to improve the quality of the SVM classification of objects which fell into the $\alpha$-area. At the same time, the 1-SVM-RF algorithm presented in this paper provided the increase in the accuracy of the SVM classification for the Spam and Biodeg datasets.

7. Conclusion
The experimental studies prove that the combined use of the 1-SVM classifier with the auxiliary RF classifier can improve the accuracy of the SVM classification when using the imbalanced datasets. In some cases, the problem properties are such that using the 1-SVM - RF algorithm to get an improvement in the quality of classification of datasets with imbalanced classes is not possible. In these cases, the classification accuracy is equivalent (within 1-2%) to the classification results using a simple implementation of the SVM classifier or the RF classifier.

In contrast to the early implementation of the two-stage classification based on binary SVM and auxiliary RF classifier (SVM-RF) for cases when the number of objects used for training the auxiliary RF classifier is not enough to improve the quality of the SVM classification, the proposed 1-SVM-RF algorithm can provide the increase in the accuracy of the SVM classification.
References

[1] Vapnik V N 1998 Statistical Learning Theory (Wiley)
[2] Yu L, Wang S, Lai K K and Zhou L 2008 Bio-Inspired Credit Risk Analysis. Computational Intelligence with Support Vector Machines Springer
[3] Schölkopf B and Smola A J 2002 Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond (Adaptive Computation and Machine Learning) (MIT Press, Cambridge, MA)
[4] Klyueva I 2019 1st International Conference on Control Systems, Mathematical Modelling, Automation and Energy Efficiency (SUMMA-2019) pp 491–494
[5] Demidova L, Klyueva I and Pylkin A 2017 The european proceedings of social & behavioural sciences EPSBS: International conference on psychology and education (ICPE) pp 95–103
[6] Klyueva I and Demidova L 2017 6th Mediterranean Conference on Embedded Computing (MECO-2017) pp 472–475
[7] Demidova L and Klyueva I 2017 ITM Web of Conferences 10 pp 10–14
[8] Demidova L and Sokolova Yu 2017 6th Mediterranean Conference on Embedded Computing (MECO-2017) pp 459–462
[9] Demidova L, Klyueva I and Pylkin A 2019 Procedia Computer Science 150 pp 455–461
[10] Demidova L, Klyueva I and Tishkin R 2018 International conference on psychology and education (ICPE-2018) pp 190–199
[11] Erfani S M, Rajasegarar S, Karunasekera S and Leckie C 2016 Pattern Recognition 58 pp 121–134
[12] Alashwal H, Deris S and Othman R 2006 International Journal of Biomedical Sciences 1 (2) pp 120–127
[13] Manevitz L M and Yousef M 2001 Journal of Machine Learning Research 2 (Dec) pp 139–154
[14] Li K, Huang H, Tian S and Xu W 2003 Proceedings of the second international conference on machine learning and cybernetics pp 2–5
[15] Breiman L 2001 Machine Learning 45(1) pp 5–32
[16] Smirnov A V 2017 Russian Technological Journal 5(6) pp 20–33 (2017)
[17] Carlos F, Nuno F, Laredo Juan L J, Merelo J J and Rosa A C 2020 Swarm and Evolutionary Computation 58: 100721