SVM prediction of the attestation success on the base of the poll results

Liliya Demidova [1, *], *Maksim* Egin [1] and *Yulia* Sokolova [1]

[1]Ryazan State Radio Engineering University, 390005 Ryazan, Russia

Abstract. The problem of the data analysis in the educational sphere in the context of prediction of the passing's success of the final state attestation by the graduates of the secondary school has been considered. Such data can be imbalanced substantially. To solve this problem it is offered to use the SVM classifiers on the base of the modified PSO algorithm, which allows choosing the kernel function type, the values of the kernel function parameters and the value of the regularization parameter simultaneously. In advance, the different rebalancing strategies, based on the basic SMOTE algorithm, can be applied for rebalance the classes in the experimental datasets. The prediction results with the use of the SVM classifiers on the base of the modified PSO algorithm and the different rebalancing strategies have been presented and compared with the prediction results received on the base of the most known software packages, such as Statistica StatSoft and IBM SPSS Modeler.

1 Introduction

In the data mining problems list, there is a special place for the classification problem, the solution of which is necessary, for example, in the credit scoring area, in the field of medical diagnostics, in the area of text categorization, for the face identification etc. A good solution to this problem is also in demand in the educational sphere. In recent years, high school students and pupils (graduates) of the secondary school are actively involved in various polls and tests, including the use of the widely approved methods to review their intellectual level, individual psychological characteristics, specialization profiling, etc.

For example, to estimate the first-year students, the motivational component diagnostics can be performed [1, 2]. The cognitive component can be inspected by R. Amthauer's intelligence test [3], which allows evaluating the verbal, mathematical and spatial intelligence. Besides, the personality component diagnostics can be performed by means of the five-factor personal questionnaire, which allows evaluating the expression degrees of personal qualities on the base of five factors (introversion – extroversion, emotion stability – neuroticism, non-perception of new information – perception, non-concentration – consciousness, hostility – goodwill) [4]. Also, the problem of diagnostics of the graduates' readiness to pass the final state attestation is very relevant. To solve this problem, in particular, it is necessary to analyze the data which consolidates information on individual self-reviewing, as well as information on immediate environment and the habitat comfort. This problem can be solved as the prediction problem whether the result of the final state attestation will be high-scoring. Obviously, the results of polls and tests accumulated in large amount can be used to extract the additional hidden information, in particular, to identify the certain cause-effect relationships and interrelations in the context of the graduates' personal diagnostics and to develop the classifiers.

Nowadays, the various algorithmic tools are used to solve the problems which require to make the data analysis. The most famous algorithmic tools are the following: linear and logistic regressions, Bayesian classifier, decision trees, decision rules, neural networks, the nearest *k*-neighbors algorithm (*k*-Nearest Neighbors Algorithm), support vector machine algorithm (SVM algorithm), and so on. A herewith, from the point of view of the presented possibilities and the undeniable advantages declared in the works of the scientific community, in the context of solving the data analysis problems in the educational sphere, the most promising is the use of the SVM algorithm. It is suggested to use the SVM classifiers on the base of the modified PSO algorithm, adapted to specifics of the problem of the data analysis in the educational sphere.

The SVM algorithm is successfully used for the SVM classifiers development [5]. The SVM classifier uses the kernel function to construct a hyperplane separating the classes of data. The satisfactory quality of training and testing of the SVM classifier allows using this SVM classifier in the classification of new objects. Choosing the optimal parameters values for the SVM classifier is a relevant problem. A herewith, it is necessary to find the kernel function type, the values of the kernel function parameters and the value of the regularization parameter [5, 6]. It is impossible to provide implementing of high-accuracy data classification with the use of the SVM classifier without adequate solution to this problem. In the simplest case solution to this problem can be found by a search of the kernel function types, the values of the kernel function parameters and the value of the

* Corresponding author: demidova.liliya@gmail.com

regularization parameter that demands significant computational expenses. The gradient methods are not suitable for search of the optimum of this problem, but the stochastic optimization algorithms, such as the genetic algorithm (GA), the artificial bee colony algorithm (ABC algorithm), the particle swarm algorithm (PSO algorithm), etc., have been used.

The PSO algorithm is the simplest stochastic optimization algorithm. The traditional approach to application of the PSO algorithm for the SVM classifier development consists of the repeated applications of this algorithm for the fixed type of the kernel function to choose the optimal values of the kernel function parameters and the value of the regularization parameter with the subsequent choice of their best combination for the best kernel function type. It is suggested to use the modified PSO algorithm to find the kernel function type, the values of the kernel function parameters and the value of the regularization parameter of the SVM classifier simultaneously.

The poll results, which are applied to form the datasets for training and testing of the SVM classifier can be imbalanced substantially. It can significantly worsen the quality of the developed SVM classifier and lower the values of its quality indicators.

Currently, the different rebalancing strategies are applied to solve the problem of the imbalance datasets. It is suggested to use the synthetic sampling algorithm called as the SMOTE (Synthetic Minority Oversampling Technique) to restore the balance between the classes. In particular, it is planned to investigate the opportunities of such variations of this algorithm, as "regular", "borderline1", "borderline2" and "SVM".

It should be noted the existence of implementations of the SVM algorithm in various software packages, for example, in Statistica StatSoft and IBM SPSS Modeler. A herewith, IBM SPSS Modeler has some implementations of the SMOTE algorithm. However, all these implementations are not sufficiently flexible and do not allow changing all of their parameters as it is necessary for the developer.

Therefore, the problem of the SVM classifiers development on the base of the modified PSO algorithm and the different rebalancing strategies in the context of prediction of the passing's success of the final state attestation by the graduates of the secondary school is very relevant.

2 Theoretical part

Let the experimental data set be a set in the form of $\{(z_1,y_1),\ldots,(z_s,y_s)\}$, in which each object $z_i \in Z$ ($i=\overline{1,s}$; s is the number of objects) is assigned to a number $y_i \in Y=\{+1;\ -1\}$ having a value of +1 or −1 depending on the class of the object z_i. It is assumed that every object z_i is mapped to q-dimensional vector of numerical values of features $z_i=(z_i^1,z_i^2,\ldots,z_i^q)$, where z_i^l is the numeric value of the l-th feature for the i-th object ($i=\overline{1,s}$, $l=\overline{1,q}$) [5 – 7]. In training of the SVM classifier it is necessary to determine the kernel function type $\kappa(z_i,z_\tau)$, the values of the kernel parameters and the value of the regularization parameter C [5 – 7]. One of the following functions is used as the kernel function $\kappa(z_i,z_\tau)$: linear function; polynomial function; radial basis function; sigmoid function [5 – 7].

To build "the best" SVM classifier it is necessary to implement the numerous repeated training (for the training data set with S elements) and testing (for the test data set $s-S$ elements, $S<s$) on the different randomly generated training and test sets with following determination of the best SVM classifier in terms of the highest possible classification quality provision [6 – 8]. The quality of the SVM classifier can be measured, for example, by the overall accuracy indicator [9 – 11].

The classification function is determined in the form:

$$f(z)=\sum_{i=1}^{S}\lambda_i\cdot y_i\cdot\kappa(z_i,z)+b\,. \quad (1)$$

The classification decision is adopted in accordance with the rule [5]:

$$F(z)=sign(f(z))=sign(\sum_{i=1}^{S}\lambda_i\cdot y_i\cdot\kappa(z_i,z)+b)\,. \quad (2)$$

The training of the SVM classifier results in determining the support vectors [5].

2.1 The modified PSO algorithm

Using the modified PSO algorithm provides the better classification accuracy by choosing the kernel function type, the values of the kernel function parameters and the value of the regularization parameter. Also, the modified PSO algorithm allows reducing the time expenditures for development of the SVM classifier [7].

In the traditional PSO algorithm the n-dimensional search space is inhabited by a swarm of a particles. The position of the i-th particle is determined by vector $x_i=(x_i^1,x_i^2,\ldots,x_i^n)$. Each i-th particle ($i=\overline{1,a}$) has its own vector of speed $v_i\in R^n$ which influence i-th particle coordinates' values. The coordinates of the i-th particle in the n-dimensional search space uniquely determine the value of the objective function $u(x)$ which is a certain solution of the optimization problem. For each position of the n-dimensional search space where the i-th particle was placed, the calculation of value of the objective function $u(x_i)$ is performed. Each i-th particle remembers the best value of the objective function found personally as well as the coordinates of the position in the n-dimensional space corresponding to the value of the objective function. Each i-th particle "knows" the best position among all positions that had been "explored" by particles. After a number of iterations particles must come close to the best position.

In the classical version of the PSO algorithm correction of each j-th coordinate of the velocity vector

($j=\overline{1,n}$) of the i-th particle is made in accordance with formula [9, 10]:

$$v_i^j = v_i^j + \hat{\varphi}\cdot\hat{r}\cdot(\hat{x}_i^j - x_i^j) + \tilde{\varphi}\cdot\tilde{r}\cdot(\tilde{x}^j - x_i^j), \quad (3)$$

where v_i^j is the j-th coordinate of the velocity vector of the i-th particle; x_i^j is the j-th coordinate of vector x_i, defining the position of the i-th particle; $\hat{x}_i^j$ is the j-th coordinate of the best position vector found by the i-th particle; $\tilde{x}^j$ is the j-th coordinate of the globally best position within the swarm in which the objective function has the optimal value; $\hat{r}$ and $\tilde{r}$ are random numbers in (0, 1); $\hat{\varphi}$ and $\tilde{\varphi}$ are personal and global coefficients for particle acceleration.

The correction of each j-th coordinate of the velocity vector of the i-th particle is performed in accordance with formula [9]:

$$v_i^j = \chi\cdot[v_i^j + \hat{\varphi}\cdot\hat{r}\cdot(\hat{x}_i^j - x_i^j) + \tilde{\varphi}\cdot\tilde{r}\cdot(\tilde{x}^j - x_i^j)], \quad (4)$$

where χ is a compression ratio;

$$\chi = 2\cdot K / \left| 2 - \varphi - \sqrt{\varphi^2 - 4\cdot\varphi} \right|; \quad (5)$$

$$\varphi = \hat{\varphi} + \tilde{\varphi} \ (\varphi > 4); \quad (6)$$

K is some scaling coefficient ($K \in (0, 1)$).

The correction of the j-th coordinate of the i-th particle can be executed in accordance with the formula:

$$x_i^j = x_i^j + v_i^j. \quad (7)$$

For each i-th particle the new value of the objective function $u(x_i)$ can be calculated and the following check must be performed: whether a new position with coordinates vector x_i became the best among all positions in which the i-th particle has previously been placed. If new position of the i-th particle is recognized to be the best at the current moment the information about it must be stored in a vector $\hat{x}_i$. The value of the objective function $u(x_i)$ for this position must be remembered. Then among all new positions of the swarm particles the check of the globally best position must be carried out. If some new position is recognized as the best globally at the current moment, the information about it must be stored in vector $\tilde{x}$. Value of the objective function $u(x_i)$ for this position must be remembered.

There is a new approach to the use of the PSO algorithm that implements a simultaneous search for the best kernel function type $\tilde{T}$, the parameters' values $\tilde{x}^1$ and $\tilde{x}^2$ of the kernel function and the value of the regularization parameter $\tilde{C}$. In this case each i-th particle defined by a vector which describes particle's position in the space: (T_i, x_i^1, x_i^2, C_i), where T_i is the number of the kernel function type; the parameters x_i^1, x_i^2, C_i are the parameters of the kernel function and the regularization parameter [10]. It is possible to "regenerate" particle through changing its coordinate T_i on number of that kernel function type, for which particles show the highest quality of classification. In the case of particles' "regeneration" the parameters' values change so that they corresponded to the new type of the kernel function. Particles which didn't undergo "regeneration", carry out the movement in own space of search of some dimension. The number of particles taking part in "regeneration" must be determined before start of algorithm. The modified PSO algorithm can be presented by the following consequence of steps.

Step 1. To determine parameters of the PSO algorithm: number m of particles in a swamp, velocity coefficient K, personal and global velocity coefficients $\hat{\varphi}$ and $\tilde{\varphi}$, maximum iterations number $N_{\max}$ of the PSO algorithm. To determine types T of the kernel functions, which take part in the search and ranges boundaries of the kernel function parameters and the regularization parameter C for the chosen kernel functions' types T: $x_{\min}^{1T}$, $x_{\max}^{1T}$, $x_{\min}^{2T}$, $x_{\max}^{2T}$, $C_{\min}^{T}$, $C_{\max}^{T}$ ($x_{\min}^{2T} = 0$ and $x_{\max}^{2T} = 0$ for $T = 1$ and $T = 2$). To determine the particles' "regeneration" percentage p.

Step 2. To define equal number of particles for each kernel type function T, included in search, to initialize coordinate T_i for each i-th particle, other coordinates of the i-th particle must be generated randomly from the corresponding ranges: $x_i^1 \in [x_{\min}^{1T}, x_{\max}^{1T}]$, $x_i^2 \in [x_{\min}^{2T}, x_{\max}^{2T}]$ ($x_i^2 = 0$ under $T = 1$ and $T = 2$), $C_i \in [C_{\min}^{T}, C_{\max}^{T}]$. To initialize random velocity vector $v_i(v_i^1, v_i^2, v_i^3)$ of the i-th particle ($v_i^2 = 0$ under $T = 1$ and $T = 2$). To establish initial position of the i-th particle as its best known position $(\hat{T}_i, \hat{x}_i^1, \hat{x}_i^2, \hat{C}_i)$, to determine the best particle with coordinates' vector $(\tilde{T}, \tilde{x}^1, \tilde{x}^2, \tilde{C})$ from all the a particles, and to determine the best particle for each kernel function type T, including in a search, with coordinates' vector $(\overline{T}, \overline{x}^{1T}, \overline{x}^{2T}, \overline{C}^T)$. Number of executed iterations must be considered as 1.

Step 3. To execute while the number of iterations is less than the fixed number $N_{\max}$: "regeneration" of particles: to choose p% of particles which represent the lowest quality of classification from particles with coordinate $T_i \neq \tilde{T}$; to change coordinate on $\tilde{T}$; to change values of the parameters x_i^1, x_i^2, C_i of "regenerated" particles to let them correspond to a new kernel function type $\tilde{T}$; correction of the velocity vector v_i and position (x_i^1, x_i^2, C_i) of the i-th particle using formulas:

$$v_i^j = \begin{cases} \chi\cdot[v_i^j + \hat{\varphi}\cdot\hat{r}\cdot(\hat{x}_i^j - x_i^j) + \tilde{\varphi}\cdot\tilde{r}\cdot(\overline{x}^{jT} - x_i^j)], & j = 1, 2, \\ \chi\cdot[v_i^j + \hat{\varphi}\cdot\hat{r}\cdot(\hat{C}_i - C_i) + \tilde{\varphi}\cdot\tilde{r}\cdot(\overline{C}^T - C_i)], & j = 3, \end{cases} \quad (8)$$

$$x_i^j = x_i^j + v_i^j \text{ for } j = 1, 2; \ C_i = C_i + v_i^3,$$

where $\hat{r}$ and $\tilde{r}$ are random numbers in (0, 1), χ is a compression ratio; accuracy calculation of the SVM classifier with parameters' values (T_i, x_i^1, x_i^2, C_i) with aim to find the optimal combination $(\tilde{T}, \tilde{x}^1, \tilde{x}^2, \tilde{C})$, which will provide high quality of classification; increase

the number of iterations on 1.

The particle with the optimal combination of the parameters' values $(\widetilde{T},\widetilde{x}^1,\widetilde{x}^2,\widetilde{C})$ will be defined after execution of the offered algorithm.

2.2 The SMOTE algorithm and its variations

The dataset is imbalanced if the classes are not approximately equally represented. The SMOTE technique is an important approach by oversampling the minority class. The SMOTE algorithm [12] 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 (*k*NN algorithm). The artificial objects which are "similar" to the objects of the existing minority class, do not duplicate them. Nowadays, the following variations of the SMOTE algorithm are most known: "regular" [12], "borderline1" [13], "borderline2" [13] and "SVM" [14].

The "regular" SMOTE algorithm is the basic [12]. The "borderline1", "borderline2" and "SVM" variations of the SMOTE algorithm look for objects that could be considered noise and objects that live near the boundary between the classes. Therefore, before creating synthetic objects using the *k* neighbors in the kNN algorithm, they look for *m* nearest neighbors to decide whether or not an object is noise or near the boundary.

The "borderline1" and "borderline2" variations of the SMOTE algorithm generate synthetic objects, but only to objects that are near the border between different classes. These variations look for objects "in danger" [13]. In the "borderline1" variation initially *m* nearest neighbours for every object in the minority class are found. Minority objects that are completely surrounded by majority objects, i.e.: all *m* nearest neighbours belong to the majority class, are considered to be noise and left out of the process. Objects with at most *m*/2 NNs from the majority class are considered to be safe, and also left out of the process. Objects for which the number of NNs from the majority class is greater than *m*/2 are considered in danger (near the borderline) and used to generate synthetic objects. New minority objects are generated along the lines connecting minority objects only to their nearest minority neighbours. The "borderline2" variation is similar to the "borderline1" one. What differs the "borderline2" variation from the "borderline1" one is that synthetic objects are created both from nearest minority neighbours as well as nearest majority neighbours. However, synthetic objects created from majority neighbours are created closer to the minority objects then when created from minority neighbours. "SVM" variation of the SMOTE algorithm fits a support vector machine classifier to the data and uses the support vector to provide a notion of boundary [14]. In the "regular" variation of the SMOTE algorithm such notion relies on proportion of nearest neighbours belonging to each class. The "regular" variation of the SMOTE algorithm does not look for objects in danger, instead it creates synthetic objects directly from the k-th nearest neighbours with not filtering.

3 Experimental studies

The synthesized poll results of 546 graduates of the secondary school before passing the unified state examination (USE) and their test marks in USE on "Russian language" and "Mathematics" disciplines were used to develop the SVM classifiers on the base of software packages, such as Statistica StatSoft [15] and IBM SPSS Modeler [16], and the author's program "Intellectual Classification" (IC).

The sections of the questionnaire can be divided into 3 groups: the general questions (the questions concerning the pupil's plans after passing the exam; questions about what the pupil considers important for admission; the questions allowing to estimate the pupil's relation to the to the subjects being passed; the questions allowing to estimate the relationship of the pupil with the surrounding people, financial position of family); the questions connected with "Russian language" discipline; the questions connected with "Mathematics" discipline.

The USE system means the translation of the primary scores of exam into the test ones which are exposed according to the hundred-score scale as a result of the scaling procedure considering all statistical materials received within the USE of this year. Scaling allows to compare and estimate the level of the pupils' readiness. The results for admission to educational institutions of the secondary vocational and higher education are provided in the test scores. "Russian language" and "Mathematics" disciplines are obligatory to delivery for obtaining the certificate about the secondary education. A herewith, the minimum thresholds specified in the scores and determined in advance must be overcome. Taking into account the mentioned above, the data preparation for the training dataset on the base of the poll results of the pupils consists in performance of the following steps: the choice of discipline and installation of the threshold of division into classes in the form of test score of the USE; the formation of the characteristics' vector for each object (pupil) on the base of his poll's results in the context of the chosen discipline.

For "Russian language" and "Mathematics" disciplines two training datasets of the identical power (with 546 objects) have been created. But these datasets have the different number of characteristics: 133 characteristics for "Russian" discipline and 141 characteristics for "Mathematics" discipline that is explained by the different number of questions having direct or indirect relation to the corresponding discipline. The synthesized poll results of 546 graduates of the secondary school before passing the unified state examination (USE) and their test marks in USE on "Russian language" and "Mathematics" disciplines were used to develop the SVM classifiers on the base of software packages, such as Statistica StatSoft [15] and IBM SPSS Modeler [16], and the author's program "Intellectual Classification" (IC).

Table 1. The results of the SVM classifiers development.

Dataset *s*×*q*	Building environment otf the SVM classifier	Number of objects in the Train and Test sets (Train/Test)	Kernel function parameters		Number of support vectors	Accuracy			1-st class (class with label "+")			2-nd class (class with label "–")		
			C	σ		Train	Test	Overall	Real number of objects in the class	Number of errors	Percent from the number of objects in the class	Real number of objects in the class	Number of errors	Percent from the number of objects in the class
Rus_80 546×133	*STATISTICA*	436/110	1	0.008	102	94.954	93.636	94.689	29	29	100	517	0	0
	SPSS Modeler	436/110	10	0.1	–	100	94.55	98.90		6	20.69		0	0
	IC	437/109	9.88	7.03	217	100	99.08	99.82		1	3.45		0	0
Rus_SVM 814×133	*STATISTICA*	651/163	5	0.008	211	98.310	93.252	97.297	297	4	1.35	517	18	3.48
	SPSS Modeler	649/165	10	0.1	–	100	100	100		0	0		0	0
	IC	652/162	8.86	8.16	166	100	100	100		0	0		0	0
RUS_ regular 1034×133	*STATISTICA*	827/207	8	0.008	215	98.791	96.618	98.356	517	0	0	517	17	3.29
	SPSS Modeler	826/208	10	0.1	–	100	100	100		0	0		0	0
	IC	828/206	3.49	5.75	269	100	100	100		0	0		0	0
Rus_ borderline1 1034×133	*STATISTICA*	827/207	9	0.008	183	98.670	96.135	98.162	517	4	0.77	517	15	2.901
	SPSS Modeler	826/208	10	0.1	–	100	100	100		0	0		0	0
	IC	828/206	7.35	9.11	152	100	99.51	99.9		1	0.19		0	0
Rus_ borderline2 1033×133	*STATISTICA*	826/207	10	0.008	225	98.668	96.135	98.161	516	4	0.78	517	15	2.90
	SPSS Modeler	825/208	10	0.1	–	100	100	100		0	0		0	0
	IC	827/206	9.59	8.68	185	100	99.51	99.90		1	0.19		0	0
Math_70 546×141	*STATISTICA*	436/110	7	0.007	91	95.560	92.727	95.788	38	22	57.89	508	1	0.20
	SPSS Modeler	436/110	10	0.1	–	99.77	100	99.82		1	2.63		0	0
	IC	437/109	5.38	8.20	176	100	97.25	99.45		1	2.63		2	0.39
Math_SVM 1016×141	*STATISTICA*	812/204	5	0.007	182	98.030	95.098	97.441	508	4	0.79	508	22	4.33
	SPSS Modeler	812/204	10	0.1	–	99.88	99.51	99.8		1	0.20		1	0.20
	IC	813/203	7.22	7.90	202	100	99.51	99.9		1	0.20		0	0
Math_ regular 1016×141	*STATISTICA*	812/204	5	0.007	178	98.645	96.078	98.13	508	3	0.59	508	16	3.15
	SPSS Modeler	812/204	10	0.1	–	99.88	100	99.9		0	0		1	0.20
	IC	813/203	6.73	7.96	174	100	100	100		0	0		0	0
Math_ borderline1 1016×141	*STATISTICA*	812/204	7	0.007	149	99.138	96.078	98.524	508	2	0.39	508	13	2.56
	SPSS Modeler	812/204	10	0.1	–	99.88	100	99.9		0	0		1	0.20
	IC	813/203	9.28	9.95	144	100	99.51	99.9		1	0.20		0	0
Math_ borderline2 1016×141	*STATISTICA*	812/204	10	0.007	182	99.138	96.078	98.524	508	4	0.79	508	11	2.17
	SPSS Modeler	812/204	10	0.1	–	99.88	99.51	99.8		1	0.20		1	0.20
	IC	813/203	8.85	6.52	279	100	100	100		0	0		0	0

The sections of the questionnaire can be divided into 3 groups: the general questions (the questions concerning the pupil's plans after passing the exam; questions about what the pupil considers important for admission; the questions allowing to estimate the pupil's relation to the to the subjects being passed; the questions allowing to estimate the relationship of the pupil with the surrounding people, financial position of family); the questions connected with "Russian language" discipline; the questions connected with "Mathematics" discipline.

The USE system means the translation of the primary scores of exam into the test ones which are exposed according to the hundred-score scale as a result of the scaling procedure considering all statistical materials received within the USE of this year. Scaling allows to compare and estimate the level of the pupils' readiness. The results for admission to educational institutions of the secondary vocational and higher education are provided in the test scores. "Russian language" and "Mathematics" disciplines are obligatory to delivery for obtaining the certificate about the secondary education. A herewith, the minimum thresholds specified in the scores and determined in advance must be overcome. Taking into account the mentioned above, the data preparation for the training dataset on the base of the poll results of the pupils consists in performance of the following steps: the choice of discipline and installation of the threshold of division into classes in the form of test score of the USE; the formation of the characteristics' vector for each object (pupil) on the base of his poll's results in the context of the chosen discipline.

For "Russian language" and "Mathematics" disciplines two training datasets of the identical power (with 546 objects) have been created. But these datasets have the different number of characteristics: 133 characteristics for "Russian" discipline and 141 characteristics for "Mathematics" discipline that is explained by the different number of questions having direct or indirect relation to the corresponding discipline.

It was experimentally established that it is expedient to establish the threshold division of classes equals to 80 scores for "Russian language" discipline and 70 scores

for "Mathematics" discipline (although it is considered that the work is high-scoring, if it had more than 80 scores). The establishment of such threshold values in the made experiments can be explained with the limited size of the polls' results and the traditionally lower scores on "Mathematics" discipline (and, as a result, lack of the sufficient number of the high-scoring works).

However, at the chosen variants of the objects (pupils) division into the classes with labels "+1" and "–1" the classes are imbalanced, that is the number of objects of one class (the majority class with label "–1") considerably exceeds the number of objects of the second class (the minority class with label "+1") ("Rus_80" and "Math_70" datasets in Table 1). For example, the minority class with label "+1" in "Rus_80" dataset describe pupils with the scores equal or greater than 80. In this regard, the decision on the use of the SMOTE algorithm variations for the purpose of the imbalance decrease has been made [13]. For example, "Rus_SVM", "RUS_regular", "Rus_borderline1" and "Rus_borderline2" datasets were obtained from the "Rus_80" dataset with the use of the "SVM", "regular", "borderline1" and "borderline2" (with k=5 and m=10) variations of the SMOTE algorithm respectively. As a result, the imbalance of classes was reduced (Table 1).

Then, the experiments on the SVM classifier development were fulfilled with the use of the software packages, such as Statistica StatSoft and IBM SPSS Modeler, and the author's program IC. For all datasets the size of the test set was equal to 20% of the size of the experimental dataset. Table 1 shows the accuracy for the train and test sets, the number of objects in the train and test sets, and the overall accuracy. A herewith, the polynomial kernel function and the radial basis one were used. In all cases the radial basis kernel function showed the best result in the context of ensuring high-quality classification. These results are shown in Tabl. 1. It can be seen that, in the absence of rebalancing, in Statistica StatSoft all objects (i.e. 29 objects) of the minority class of the "Rus_80" dataset and the significant part of the objects (22 objects from 38) of the minority class of the "Math_70" dataset was classified erroneously. For the "Rus_80" dataset we received 100% and 0% of errors in the classes with labels "+1" and "–1" respectively, for the "Math_70" dataset we received 57.89% and 0.20% of errors in the classes with labels "+1" and "–1" respectively, though the overall accuracy of classification is high (94.689 % and 95.788 % respectively). Application of IBM SPSS Modeler to these datasets allow improving the overall accuracy. The program "Intellectual Classification" has a little better fulfilled. But in all cases almost all errors were in the minority class, therefore, the SVM classifier developed on the base of imbalanced dataset, will give incorrect forecasts for new objects of this class (with high-scores' works).

In Statistica StatSoft package the parameters of the kernel function (the radial basis kernel function and the polynomial one) were selected in accordance with the default settings and the value of the regularization parameter was determined using the sliding control procedure. The σ parameter of the radial basis kernel function by default is chosen according to the size of the dataset (the larger the dataset, the smaller the σ). The best results of the SVM classifier development were obtained for the "Rus_regular" dataset for "Russian language" discipline (0.78% and 3.29% errors in the classes) and for the Math_borderline1" dataset for "Mathematics" discipline (0.39% and 2.56% error in the classes) using the radial basis kernel function. In the IBM SPSS Modeler package there are no means for selecting the parameters of the SVM classifier, which provides the maximum classification accuracy, so the development of the SVM classifier with the radial basic kernel function and the polynomial one was performed using the parameters defined by default. For example, for the radial basic kernel function these default settings are the following: C=10 and σ=0.1. Also, it is impossible to estimate the number of support vectors. The best results of the SVM classifier development were obtained with the use of the radial basic kernel function. For all synthesized datasets for "Russian language" discipline the classification accuracy was equal to 100%. For "Mathematics" discipline the best results were obtained for "Math_regular" and "Math_borderline1" datasets (0 % and 0.20% errors in the classes).

The author's program IC contains the module of the search of the optimal parameters of the SVM classifier using the modified PSO algorithm. This program allows developing the SVM classifiers with the minimum number of errors for the initial imbalanced experimental datasets (1 error for the "Rus_80" dataset and 3 errors for "Math_70" dataset), and reducing to zero the number of errors for "Rus_SVM", "Rus_regular", "Math_regular" and "Math_borderline2" datasets.

4 Conclusions

The use of the SVM classifiers on the base of the modified PSO algorithm and the different rebalancing strategies in the context of prediction of the passing's success of the final state attestation by the graduates of the secondary school allows providing the high classification accuracy. The results of experimental studies confirm the expediency of further development of the offered approach to the SVM classifier development. A herewith, it is planned to use the SVM algorithm for the regression model development, that will allow forecasting the attestation's results in scores.

The reported study was funded by RFBR according to the research project № 17-29-02198.

References

1. J.W. Atkinson, American Psychologist, 36, 117-128
2. P.R. ZelickNova (eds) *Issues in the Psychology of Motivation*. (Science Publishers, Inc., 2007)
3. B. Weiss, V.C. Laties (eds) *Behavioral Toxicology*. (Springer, 1975)
4. Hege H. Bye, Gro M. Sandal, Journal of Business and Psychology, **31**, 569-582 (2016)

5. O. Chapelle, V. Vapnik, O. Bousquet, S. Mukherjee, Machine Learning, **46**, 131 (2002)
6. L. Yu, S. Wang, K. K. Lai, L. Zhou, *Bio-Inspired Credit Risk Analysis. Computational Intelligence with Support Vector Machines* (Springer, 2008)
7. L. Demidova, Yu. Sokolova, E. Nikulchev, International Review on Modelling and Simulations, **8**, 446 (2015)
8. L. Demidova, Y. Sokolova, Modification of particle swarm algorithm for the problem of the SVM classifier development, In *International Conference "Stability and Control Processes" in Memory of V.I. Zubov*, pp. 623-627 (2015)
9. L. Demidova, E. Nikulchev, Y. Sokolova, International Journal of Advanced Computer Science and Applications, **7**, 294 (2016)
10. L. Demidova, Y. Sokolova, ITM Web of Conferences, **6** 02003 (2016)
11. L. Demidova, I. Klyueva, Y.Sokolova, N. Stepanov, N. Tyart, Procedia Computer Science, **103**, 222 (2017)
12. N. Chawla, K. Bowyer, L. Hall, W. Kegelmeyer, JAIR, **16**, 341 (2002)
13. H. Han, W. Wen-Yuan, M. Bing-Huan, Advances in intelligent computing, 878-887, (2005)
14. H.M. Nguyen, E.W. Cooper, K. Kamei, Int. J. of Knowledge Engineering and Soft Data Paradigms, **3**, 4 (2001)
15. http://documentation.statsoft.com/STATISTICAHelp.aspx?path=MachineLearning/MachineLearning/SupportVectorMachine/SupportVectorMachineExample1Classification
16. https://www.ibm.com/support/knowledgecenter/SS3RA7_18.1.0/modeler_tutorial_ddita/clementine/example_svm_intro.html