ITM Web of Conferences 6, 02003 (2016)
DOI: 10.1051/itmconf/20160602003
© Owned by the authors, published by EDP Sciences, 2016

Development of the SVM classifier ensemble for the classification accuracy increase

Liliya Demidova [1, 2, a] and Yulia Sokolova [2]

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

Abstract. The problem of improving the classification accuracy using the SVM classifier ensemble has been considered. This paper defines the rules for the selection of individual SVM-classifiers used in the future for the creation of an ensemble and the strategies for the integration of ensemble members.

1 Introduction

Currently, for a wide range of classification problems in various applications the SVM algorithm (Support Vector Machines, SVM), which carries out training on precedents («supervised learning»), is successfully used and included in the group of the boundary algorithms and methods of classification [1]. It allows developing the classifiers which can be successfully used for a wide range of applications [2].

At first in the developing of the SVM classifier there is a need to implement multiple learning and testing on the basis of different randomly generated training and test sets, and then it is necessary to choose the best SVM classifier which provides the highest possible quality of classification. Assessment of classification quality can be performed using various indicators [2].

For the training of the SVM classifier it is necessary to define the internal parameters of the SVM classifier: the kernel function type, values of the kernel parameters and value of the regularization parameter. These internal parameters participate in the construction of the classifying function $f(x)$, which assign some object x to the concrete class from the set $\{-1; +1\}$ [2]. Therefore, the problem of the selection of parameters of the SVM-classifier is crucial for receiving exact results of classification.

In recent years, much attention is paid to a question of increase in accuracy of the models, based on the machine learning algorithms. Herewith questions of the opportunities' association of several classifiers and creation of ensembles of classifiers to increase quality of the solution of the applied tasks are investigated [3-5]. The learning of the ensemble of classifiers is a training procedure of a final set of the base (individual) classifiers, individual solutions of which are then combined to form the resulting classification decisions, based on the aggregated classifier. There are different approaches to choose the rules of combination of the individual classifiers in the ensemble and the strategies for creation of the resulting classification decision [2].

The purpose of this work is to improve the accuracy of classification decisions using the SVM classifier ensemble based on various strategies of integrating of the individual classifiers into an ensemble.

2 Theoretical part

Let the experimental data set be a set in the form of $\{(x_1,y_1),\ldots,(x_n,y_n)\}$, in which each object $x_i \in X$ is assigned to the number $y_i \in Y = \{-1; +1\}$ having a value of -1 or $+1$ depending on the class of the object x_i. Herewith it is assumed that every object x_i is mapped to q-dimensional vector of numerical values of characteristics $x_i = (x_i^1, x_i^2, \ldots, x_i^q)$ where x_i^l is a numeric value of the l-th characteristic for the i-th object ($i = \overline{1,n}$, $l = \overline{1,q}$) [3-5]. It is required with the use of special function $\kappa(x_i, x_\tau)$, which is called the kernel, to build the classifier $F: \; X \to Y$, comparing the class $Y = \{-1; +1\}$ with some object from the set X.

As the kernel function $\kappa(x_i, x_\tau)$ which allows separating the objects of different classes, typically one of the following functions is used [3]:

- linear function: $\kappa(x_i, x_\tau) =< x_i, x_\tau >$;
- polynomial function: $\kappa(x_i, x_\tau) = (< x_i, x_\tau > +1)^d$;
- radial basis function:

$$\kappa(x_i, x_\tau) = exp(- < x_i - x_\tau, x_i - x_\tau > /(2 \cdot \sigma^2));$$

- sigmoid function: $\kappa(x_i, x_\tau) = th(k_2 + k_1 \cdot < x_i, x_\tau >)$,

where $< x_i, x_\tau >$ is a scalar product of vectors x_i and x_τ; d [$d \in N$ (by default $d = 3$)], σ [$\sigma > 0$ (by default

[a] Corresponding author: liliya.demidova@rambler.ru

$\sigma^2 = 1$)], k_2 [$k_2 < 0$ (by default $k_2 = -1$)] and k_1 [$k_1 > 0$ (by default $k_1 = 1$)] are some parameters; th is a hyperbolic tangent.

In the training of the SVM classifier it's necessary: 1) to divide the experimental data set for training and test sets, which consist of N and $n - N$ elements respectively ($N < n$); 2) to determine the input parameters of the classifier: the kernel function type $\kappa(x_i, x_\tau)$, 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; 3) to implement multiple learning and testing on different randomly generated training and test sets, with subsequent determination of the best SVM-classifier. The test set is from 1/10 to 1/3 of the experimental data set, it is not involved in the setting of the parameters of the classifier, and used to verify its accuracy. If the quality of learning and testing is acceptable, the SVM-classifier can be used to classify new objects.

As a result of training, the classification function is determined in the following form [1-3]:

$$f(x) = \sum_{i=1}^{N} \alpha_i \cdot y_i \cdot \kappa(x_i, x) + b \tag{1}$$

The classification decision, associating the object x to the class −1 or +1, is adopted in accordance with the rule [1-3]:

$$F(x) = sign(f(x)) = sign\left(\sum_{i=1}^{N} \alpha_i \cdot y_i \cdot \kappa(x_i, x) + b\right) \tag{2}$$

In (1) and (2): $\kappa(x_i, x)$ is the kernel classifier; b is the parameter determining the shortest distance from the origin to the hyperplane that separates classes; α_i is the Lagrange multiplier; $\alpha_i \geq 0$; y_i is the classification decision (−1 or +1) [3].

The main problem with the training of the SVM classifier, is the lack of recommendations for the choice of the regularization parameter C, the function describing the kernel $\kappa(x_i, x_\tau)$, as well as the parameters of the kernel function, for which high accuracy of data classification is achieved. This problem can be solved with the use of various optimization algorithms, in particular using the PSO algorithm [6, 7].

After training, each classifier generates its own (individual) classification decisions, the same or different from the actual results of classification. Accordingly, the different individual SVM classifiers correspond to the different classification accuracy. The quality of the received classification decisions can be improved on the base of ensembles of the SVM classifiers [2-5]. In this case, the final set of individually trained classifiers must be learned. Then the classification decisions of these classifiers are combined. The resulting solution is based on the aggregated classifier. The majority vote method and the vote method based on the degree of reliability can be used as the rules (strategies) of the definition of the aggregated solutions.

The majority vote method is one of the most common and frequently used method for combining of decisions in the ensemble of classifiers. But this method does not fully use the information about the reliability of each individual SVM classifier. For example, suppose that the SVM classifier ensemble aggregates the results of five individual SVM classifiers, where values of the function $f(x)$ of the object x obtained from the three individual SVM classifiers, are negative (class –1), but very close to the neutral position, and values of the function $f(x)$ of the other two SVM classifiers are strongly positive (class +1), i.e. very far away from the neutral position. Then the result of the aggregated decision of the ensemble on the basis of «one classifier – one vote» is following: the object x belongs to the negative class (majority vote), although it is obvious, that the best and more appropriate choice for the object x is a positive class. Despite the good potential of the majority vote method for combining of the group of decisions, it is recommended to use other methods to increase the accuracy of classification.

Vote method based on the degree of reliability uses value of the function $f(x)$ for the object x obtained by each individual SVM classifier. The greater the positive value of $f(x)$ in (1), returned by the SVM classifier, the more precisely the object x is determined in class +1, and the less negative value $f(x)$, the more precisely the object x is defined in class –1. Values «–1» and «+1» for $f(x)$ indicate that the object x is situated on the boundary of the negative and positive classes, respectively.

When using an ensemble of classifiers for solving classification problems the special attention should be paid to the methods of forming a set of individual classifiers, which can later be used in the development of the final SVM classifier. It is experimentally confirmed [2-5], that the ensemble of classifiers shows greater accuracy than any of its individual members, if individual classifiers are accurate and varied. Therefore, the formation of the set of the individual SVM classifiers is required: 1) to use the various kernel functions; 2) to build classifiers in the different ranges of change of the kernel parameters and regularization parameter; 3) to use various sets of training and test data. To select the appropriate members of the ensemble in the set of the trained SVM classifiers it is recommended to use the principle of maximum decorrelation. In this case the correlation between the selected classifications should be as small as possible. After training, each private j-th classifier from the k trained classifier will correspond to a certain array of errors: $e_{ij} = | y_{ij} - \tilde{y}_{ij} |$, where e_{ij} is the error of j-th classifier at i-th row of the experimental data set ($i = \overline{1, n}$; $j = \overline{1, k}$); y_{ij} is the classification decision (−1 or +1) of j-th classifier at i-th row of the experimental data set; $\tilde{y}_{ij}$ is the real meaning of a class (−1 or +1), for which the i-th object is belong to.

The SVM classifiers, not permitting an error on the experimental data set, should be excluded from further consideration, and from the remaining quantity of the SVM classifiers, it is necessary to select an appropriate number of individual SVM classifiers with maximal variety. To solve this problem, decorrelation maximization algorithm can be used. This algorithm provides a variety of individual SVM classifiers, being used in the construction of the ensemble [2]. If the correlation between the selected classifiers is small, then the decorrelation is maximum.

Let there be an error matrix E of set of individual SVM classifiers with size $n \times k$:

$$E = \begin{bmatrix} e_{11} & e_{12} & \cdots & e_{1k} \\ e_{21} & e_{22} & \cdots & e_{2k} \\ \cdots & \cdots & \cdots & \cdots \\ e_{n1} & e_{n2} & \cdots & e_{nk} \end{bmatrix}, \tag{3}$$

where e_{ij} is the error of the j-th classifier at the i-th row of the experimental data set ($i = \overline{1,n}$; $j = \overline{1,k}$).

On the basis of the error matrix E (3) the following assessments can be calculated [2]:

• mean: $\bar{e}_j = \frac{1}{n}\sum_{i=1}^{n} e_{ij}$ ($j = \overline{1,k}$); (4)

• variance: $V_{jj} = \frac{1}{n}\sum_{i=1}^{n} (e_{ij} - \bar{e}_j)^2$ ($j = \overline{1,k}$); (5)

• covariance:

$$\mathrm{V}_{tj} = \frac{1}{n}\sum_{i=1}^{n} (e_{ij} - \bar{e}_j)\cdot(e_{it} - \bar{e}_t) \quad (j = \overline{1,k}, t = \overline{1,k}); \tag{6}$$

Then the elements r_{tj} of the correlation matrix with size $k \times k$ are calculated as:

$$r_{tj} = V_{tj} \Big/ \sqrt{V_{tt} \cdot V_{jj}}\,; \tag{7}$$

where r_{tj} is the correlation coefficient, representing the degree of correlation of t-th and j-th classifiers ($j = \overline{1,k}$; $t = \overline{1,k}$); $r_{jj} = 1$ ($j = \overline{1,k}$).

Using the correlation matrix R it is possible for each individual j-th classifier to calculate the plural-correlation coefficient ρ_j, which characterizes the degree of correlation of j-th and all other $(k-1)$ classifiers with numbers t ($t = \overline{1,k}$; $t \neq j$) [8]:

$$\rho_j = \sqrt{1 - |R| / R_{jj}} \quad (j = \overline{1,k}) \tag{8}$$

where $|R|$ is the determinant of the correlation matrix R; R_{jj} is the cofactor of the element r_{jj} of the correlation matrix R.

A quantity ρ_j^2 called the coefficient of determination. It shows the proportion of the variation of the analyzed variable, which is explained by variation of the other variables. The coefficient of determination ρ_j^2 can take values from 0 to 1. The closer the coefficient to 1, the stronger the relationship between the analyzed variables (in this case, between individual classifiers) [8]. It is believed that there is a dependency, if the coefficient of determination is not less than 0.5. If the coefficient of determination greater than 0.8, it is assumed that high dependence exists.

For selection of individual SVM classifiers for integration into the ensemble it is necessary to determine the threshold θ. Thus, the j-th individual classifier must be removed from the list of classifiers if the coefficient of determination ρ_j^2 satisfies to condition $\rho_j^2 > \theta$ ($j = \overline{1,k}$). If it is necessary to identify the most various classifiers, generating decisions with the most different arrays of errors on the experimental data set, thresholds θ, satisfying to condition $\theta < 0.7$ should be selected. Herewith the additional considerations can be also taken into account to avoid the exclusion of insufficient or excessive number of individual SVM classifiers.

The decorrelation maximization algorithm can be summarized into the following steps [2].

Step 1. To calculate the matrix V and the correlation matrix R with formulas (5), (6) and (7) respectively.

Step 2. To calculate the multiple correlation coefficients ρ_j ($j = \overline{1,k}$) with (8) for all classifiers.

Step 3. To remove classifiers, for which $\rho_j^2 > \theta$ ($j = \overline{1,k}$), from the list of classifiers.

Step 4. To repeat iteratively steps 1 – 3 for the remaining classifiers in the list until for all classifiers the condition $\rho_j^2 \leq \theta$ ($j = \overline{1,k}$) will not satisfied.

As a result, the list of classifiers used to form the ensemble will consist of m ($m \leq k$) individual classifiers.

For classifiers selected in the ensemble, it is necessary to carry out:

- the rationing of degrees of the reliability;
- the strategy search for the integration of members of the ensemble;
- the calculation of the aggregated decision of the ensemble.

Value of the reliability $f_j(x)$, which is defined for the object x by the j-th classifier, falls into the interval (-∞, + ∞). The main drawback of such values is that in the ensemble the individual classifiers with large absolute value are often dominated in the final decision of the ensemble. To overcome this drawback, the rationing is carried out: the transformation of values of degrees of reliability in the interval [0; 1] is fulfilled. In the case of binary classification in the rationalization for the object x the values of the reliability of its membership to positive class (labeled +1) $g_j^+(x)$ and to negative class $g_j^-(x)$ are determined. These values can be determined by the formulas [2]:

$$g_j^+(x) = \frac{1}{1+e^{-f_j(x)}} \quad (9)$$

$$g_j^-(x) = 1 - g_j^+(x) \quad (10)$$

The selected individual classifiers are combined into the ensemble using $g_j^+(x)$ and $g_j^-(x)$ ($j = \overline{1,m}$) in accordance with one of the following five strategies [2].

1. Maximum strategy:

$$\mathrm{A}(x) = \begin{cases} 1, & \text{if} \quad \max\limits_{j=1,m} g_j^+(x) \geq \max\limits_{j=1,m} g_j^-(x), \\ -1, & \text{otherwise.} \end{cases} \quad (11)$$

2. Minimum strategy:

$$\mathrm{A}(x) = \begin{cases} 1, & \text{if} \quad \min\limits_{j=1,m} g_j^+(x) \geq \min\limits_{j=1,m} g_j^-(x), \\ -1, & \text{otherwise.} \end{cases} \quad (12)$$

3. Median strategy:

$$A(x) = \begin{cases} 1, & \text{if} \quad \frac{1}{m}\sum\limits_{j=1}^{m} g_j^+(x) \geq \frac{1}{m}\sum\limits_{j=1}^{m} g_j^-(x), \\ -1, & \text{otherwise.} \end{cases} \quad (13)$$

4. Mean strategy:

$$A(x) = \begin{cases} 1, & if \quad \sum\limits_{j=1}^{m} g_j^+(x) \geq \sum\limits_{j=1}^{m} g_j^-(x), \\ -1, & \text{otherwise.} \end{cases} \quad (14)$$

5. Product strategy:

$$A(x) = \begin{cases} 1, & \text{if} \quad \prod\limits_{j=1}^{m} g_j^+(x) \geq \prod\limits_{j=1}^{m} g_j^-(x), \\ -1, & \text{otherwise.} \end{cases} \quad (15)$$

The value $A(x)$ is an aggregated measure of the reliability's value of the SVM classifier ensemble. It can be used to integrate the members of the ensemble.

The learning algorithm of the ensemble of the SVM classifiers can be summarized into the following steps.

Step 1. To divide the experimental data set into k training data sets: $TR_1, \ldots, TR_k$.

Step 2. To learn k individual SVM classifiers with the different training data sets $TR_1, \ldots, TR_k$ and to obtain k individual SVM classifiers (ensemble members).

Step 3. To select m ($m \leq k$) decorrelated SVM classifiers from k classifiers using the decorrelation maximization algorithm

Step 4. To determine values of m classification functions for each selected individual SVM classifier: $f_1(x), \ldots, f_m(x)$.

Step 5. To transform values of degrees of reliability, using (9) and (10), for the positive class $g_1^+(x)$, ..., $g_m^+(x)$ and for the negative class $g_1^-(x), \ldots, g_m^-(x)$.

Step 6. To determine the aggregated value $\mathrm{A}(x)$ of the reliability of the SVM classifier ensemble using (11) – (15).

This algorithm, used for the weak SVM classifiers, will provide a better quality of the classification accuracy than accuracy of any single individual classifier used for aggregation.

The problem of choosing of the threshold θ is very important. Value θ for which all five rules of classification (11) – (15) show stable improvement of quality of classification must be chosen as the threshold value θ^* ($\theta^* < 0.7$). Thus the use of each of the five rules leads to improvement of the classification quality resulting in the reduction of the number of erroneous decisions, when the smaller number of individual classifiers, corresponding to the threshold value θ^*, is applied. Herewith such stable improvement of the classification quality isn't observed for all examined values θ' (for which $\theta' > \theta^*$).

It should be noted, that the majority vote rule may be used for decisions, obtained using the classification rules (11) - (15), to determine the required threshold value θ^*.

3 Experimental studies

The feasibility of using of the SVM classifier ensembles was confirmed by test and real data.

Several individual SVM classifiers using different types of the kernel function, different values of the kernel function functions of the kernel parameters and different values of the regularization parameter were learned in the experiments for a particular data set. Herewith the different training and test sets randomly generated from the original data set used were. Then the principle of maximum decorrelation (for selection of individual classifiers, which must be included in the ensemble) and various strategies of forming of the aggregated classifier were applied for the trained classifiers. The obtained classifier should have the higher classification accuracy than the classification accuracy of any single individual classifier.

Actual data used in the experimental researches were taken from Statlog project and from UCI machine learning library. In particular, data set for credit scoring was used. The German credit data set contains 1000 instances including 700 creditworthy cases (class 1) and 300 default cases (class 2); herewith each applicant is described by 24 characteristics ($q = 24$, the source is http://archive.ics.uci.edu/ml/machine-learning-databases/statlog/german/). As a result, 18 private SVM classifiers were trained with use of various input parameters.

During testing it was found, that the individual classifiers indicate the accuracy of classification decisions from 83.1% to 93.1%, and the initial values of the determination coefficient, calculated for all 18 individual classifiers, are in the range from 0.047 to 0.563. As a result, the threshold values θ were examined from the range [0.1; 0.6] with step 0.5. Values of classification parameters corresponding to different threshold values θ are given in Table 1.

The optimal threshold value θ for the reviewed example is 0.35, since for $\theta = 0.35$ all five classification rules (11) – (15) give a stable improvement of the classification quality when the number of classifiers reduces to the number corresponding to the value $\theta = 0.35$. Herewith the final number of classifiers in the ensemble proved equal to 8. A further decrease in the number of classifiers is not feasible (due to a further sharp decrease in their number and a substantial reduction of their variety).

Table 1. Values of classification parameters at the different threshold values of the determination coefficient.

Value of classification	Strategy	The threshold value of the coefficient of determination									
		0.6	0.55	0.5	0.45	0.4	0.35	0.3	0.25	0.2/ 0.15	0.1
Overall accuracy (%)	Majority vote	96.00	96.00	96.00	96.00	96.00	96.00	96.00	96.00	96.00	96.00
	Maximum and minimum	81.10	81.50	82.10	83.00	83.40	83.90	84.60	91.50	92.20	92.20
	Median and sum	94.90	96.00	96.40	97.20	97.50	98.20	97.90	97.80	97.00	96.90
	Product	89.50	90.30	90.40	90.70	91.60	91.70	91.70	97.10	96.40	95.70
Sensitivity (%)	Majority vote	97.43	97.43	97.43	97.43	97.43	97.43	97.43	97.43	97.43	97.43
	Maximum and minimum	83.00	83.57	84.14	85.43	86.29	87.00	87.57	95.29	95.57	95.43
	Median and sum	95.86	96.86	97.29	98.00	97.57	98.57	98.43	98.86	98.57	97.86
	Product	90.57	91.29	91.29	91.29	91.57	91.86	91.86	98.86	97.71	97.00
Specificity (%)	Majority vote	92.67	92.67	92.67	92.67	92.67	92.67	92.67	92.67	92.67	92.67
	Maximum and minimum	76.67	76.67	77.33	77.33	76.67	76.67	77.67	82.67	84.33	86.33
	Median and sum	92.67	94.00	94.33	95.33	97.33	97.33	96.67	95.33	93.33	94.67
	Product	87.00	88.00	88.33	89.33	91.67	91.33	91.33	93.00	93.33	92.67
Number of errors of the 1-st type	Majority vote	22	22	22	22	22	22	22	22	22	22
	Maximum and minimum	70	70	68	68	70	70	67	52	47	41
	Median and sum	22	18	17	14	8	8	10	14	20	16
	Product	39	36	35	32	25	26	26	21	20	22
Number of errors of the 2-nd type	Majority vote	18	18	18	18	18	18	18	18	18	18
	Maximum and minimum	119	115	111	102	96	91	87	33	31	32
	Median and sum	29	22	19	14	17	10	11	8	10	15
	Product	66	61	61	61	59	57	57	8	16	21
Number of classifiers in the ensemble		18	16	15	11	9	8	7	5	4	3

Use of the median strategy with $\theta = 0.35$ allowed classifying correctly 98.2% of objects of the initial data set. At the same time, the maximum accuracy of one of individual SVM classifiers, used in the ensemble, was equal 93.1%, and the accuracy reached with use of the majority vote rule was equal 96%.

Thus, the use of the SVM classifier ensemble allowed increasing the classification accuracy more than 5% compared with the maximum accuracy of one of the individual classifiers.

4 Conclusions

The use of the SVM classifier ensembles allows reducing the accident classification decision received by one classifier, and helps to improve the classification accuracy. The shortcomings of some classifiers are compensated by strengths of others classifiers thanks to combination of their results. Classifiers counterbalance the results' accident of each other, finding on the basis of balance the most plausible output classification decision. It allows finding the best classification result with minimum classification error.

References

1. O. Chapelle, V. Vapnik, O. Bousquet and S. Mukherjee, Machine Learning, **46**, 131 (2002)
2. L. Yu, S. Wang, K.K. Lai and L. Zhou, *Bio-Inspired Credit Risk Analysis. Computational Intelligence with Support Vector Machines* (Springer, 2008)
3. L. Demidova, Yu. Sokolova and E. Nikulchev, International Review on Modelling and Simulations **8**, 446 (2015)
4. L. Demidova and Yu. Sokolova, *16th International Symposium on Advanced Intelligent Systems*, pp. 889-906 (2015)
5. L. Demidova and Yu. Sokolova, *International Conference «Stability and Control Processes» in Memory of V.I. Zubov*, pp. 619-622 (2015)
6. L. Demidova and Yu. Sokolova, *International Conference «Stability and Control Processes» in Memory of V.I. Zubov*, pp. 623-627 (2015)
7. L. Demidova, E. Nikulchev and Yu. Sokolova, International Journal of Advanced Computer Science and Applications*,* **7**, (2016)
8. S.A. Aivazian and V.S. Mkhitarian, *Applied statistics and essentials of econometrics* (Moscow, 1998)