Data classification based on the hybrid intellectual technology

Liliya Demidova [1, *], *Maksim* Egin [1]

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

Abstract. In this paper the data classification technique, implying the consistent application of the SVM and Parzen classifiers, has been suggested. The Parser classifier applies to data which can be both correctly and erroneously classified using the SVM classifier, and are located in the experimentally defined subareas near the hyperplane which separates the classes. A herewith, the SVM classifier is used with the default parameters values, and the optimal parameters values of the Parser classifier are determined using the genetic algorithm. The experimental results confirming the effectiveness of the proposed hybrid intellectual data classification technology have been presented.

1 Introduction

Nowadays, the different methods and algorithms of data mining applied to development of the data classifiers are known. The most famous data mining algorithms, such as the SVM algorithm (Support Vector Machine Algorithm) [1 – 5] and the *k*NN algorithm (*k* Nearest Neighbors Algorithm) show the high data classification quality [6]. The Parzen window algorithm is less famous in the solving the data classification problems, but it also shows the high data classification quality, and possesses by some advantages over the *k*NN algorithm. A herewith, the *k*NN classifier and the Parzen classifiers use the essentially similar approaches to the data classification.

However, as the experimental results show, currently there is no data classifier which is the best for all classification problems, as it doesn't allow providing the high classification quality for any datasets because of specifics of the tools applied in case of its development and, respectively, limitation of its opportunities. Therefore, the decision on hybridization of the SVM and Parzen classifiers for classification of the elaborate multidimensional data was accepted.

Nowadays, a small number of approaches to hybridization of the mentioned above classifiers is known. The approach, suggested in [7], uses in the *k*NN classifier the support vectors of the developed SVM classifier as the representative objects to classify the data within the separating strip. It is claimed that in this case the most useful information is utilized. The basic idea of the hybrid SVM-*k*NN classifier in the approach, offered in [8], is to find the nearest neighbors to a query sample and train a local support vector machine that preserves the distance function on the collection of neighbors. The novel SVM-*k*NN technique for data classification was suggested by one of authors of this paper in [9]. It showed the improving the accuracy of the SVM classification for the different datasets.

In this paper we describe the data classification technique, implying the consistent application of the SVM and Parzen classifiers in the found experimentally subareas of the characteristics space. This data classification technique uses some ideas of the SVM-*k*NN technique. We develop the Parzen classifier for data within the separating strip, in particular, within the Ω area containing all objects which are mistakenly classified by the SVM classifier. Also, the Ω area can contain a some number of objects classified correctly. We use data outside the Ω area as the representative objects of the corresponding classes.

We suggest to use the SVM classifier with the default parameters values, and to define the optimal parameters values of the Parser classifier using the genetic algorithm. The use of such approach to determining the parameters values of the used classifiers will allow to reduce the time costs for obtaining the hybrid classifier which ensures the high data classification accuracy.

2 SVM classification

To find the separating hyperplane in the SVM algorithm it is necessary to solve the dual problem of searching a saddle point of the Lagrange function, which can be reduced to the problem of quadratic programming [1, 2]:

$$\begin{cases} -L(\lambda) = -\sum\limits_{i=1}^{S}\lambda_i + \\ \quad + \frac{1}{2}\cdot\sum\limits_{i=1}^{S}\sum\limits_{\tau=1}^{S}\lambda_i\cdot\lambda_\tau\cdot y_i\cdot y_\tau\cdot\kappa(z_i,z_\tau)\to\min\limits_{\lambda}, \quad (1) \\ \quad \sum\limits_{i=1}^{S}\lambda_i\cdot y_i = 0, 0\le\lambda_i\le C,\ i=\overline{1,S}, \end{cases}$$

where λ_i is a dual variable; z_i is the object of the training set; y_i is a number (+1 or −1), which characterize the class of the object z_i from the experimental dataset; $\kappa(z_i,z_\tau)$ is a kernel function; C is a regularization parameter; S is a number of objects in the experimental dataset; $i=\overline{1,S}$.

* Corresponding author: demidova.liliya@gmail.com

© The Authors, published by EDP Sciences. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (http://creativecommons.org/licenses/by/4.0/).

Usually, the SVM classifier, even with the default settings, provides a high-quality data classification. To improve the accuracy of the SVM classification, it is necessary to analyse the location of the erroneously classified objects. In most cases, the mistakenly classified objects are located near the separating hyperplane. Therefore, it is necessary to use the additional tools to improve the classification accuracy for the objects within the separating strip.

3 Parzen Classification

The Parzen window method is a special case of the generalized metric classifier, defined by the following classification rule:

$$a(u, Z^{train}) = \arg\max_{y \in Y} \sum_{i=1}^{\gamma} w_i [y_z^{(i)} = y] \quad (2)$$

where z is the object, whose class affiliation should be established; Z^{train} is the training set of objects, whose class affiliation is known; w_i is the weight of the i-th object of the training set; Y is the set of classes of objects of the training set; $y_z^{(i)}$ is the class affiliation of the i-th neighbor of the object z; $[y_z^{(i)} = y]$ equals to 1, if $y_z^{(i)} = y$, else $[y_z^{(i)} = y]$ equals to 0.

In the case of the Parzen window, the weight w_i of the i-th object is given by the kernel function, which does not grow in $[0, \infty)$ [10]. Depending on the type of the Parzen window, the following ways of the weights' setting can be applied.

1. For the fixed-width Parzen window, the weights of objects are limited by the certain real number h, which defines the maximum distance at which the objects are considered:

$$w_i = K\left(\frac{d(z, z_z^{(i)})}{h}\right) \quad (3)$$

where K is the kernel function; $d(z, z_z^{(i)})$ is the distance from the object z to the i-th neighbor $z_z^{(i)}$, calculated in accordance with a certain metric.

2. For the variable-width Parzen window, the integer number k of the neighbor, whose contribution to the classification of the unknown object is taken into account instead of the window width. The number h is defined according to the distance to the k-th neighbor from the training set:

$$w_i = K\left(\frac{d(z, z_z^{(i)})}{d(z, z_k^{(i)})}\right) \quad (4)$$

Due to advantages of the variable-width Parzen window, it will be used in the future.

For the Parzen classifier development, it is necessary to determine the following parameters: the number k of neighbors, the metric and the kernel function.

To calculate distance between the objects $t = (t_1, t_2, \ldots, t_q)$ and $p = (p_1, p_2, \ldots, p_q)$ in the q-dimensional space, the following metrics can be used:

• the Euclidean distance:

$$d(t, p) = \sqrt{\sum_{i=1}^{q} (t_i - p_i)^2}, \quad (5)$$

• the square of the Euclidean distance:

$$d(t, p) = \sum_{i=1}^{q} (t_i - p_i)^2, \quad (6)$$

• the Manhattan distance:

$$d(t, p) = \sum_{i=1}^{q} |t_i - p_i|, \quad (7)$$

• the Chebyshev distance:

$$d(t, p) = \max_i |t_i - p_i|. \quad (8)$$

The following kernel functions can be used in the Parzen classifier:

• the Epanechnikov kernel function:

$$K(r) = \frac{3}{4} \cdot (1 - r^2); \quad (9)$$

• the quadratic kernel function:

$$K(r) = \frac{15}{16} \cdot (1 - r^2)^2; \quad (10)$$

• the triangular kernel function:

$$K(r) = 1 - |r|; \quad (11)$$

• the Gaussian kernel function:

$$K(r) = (2\pi)^{-\frac{1}{2}} \cdot e^{-\frac{1}{2} \cdot r^2}, \quad (12)$$

where r is the real number; $r \in [0; 1]$.

4 Data Classification Technique

Improving the accuracy of the SVM classification can be achieved using the following data classification technique.

1. To develop the SVM classifier on the basis of the initial dataset. To estimate the data classification quality with application of the various classification quality indicators.

2. To form the dataset, which will contain all objects which are mistakenly classified by the SVM classifier and are located in the Ω area near the separating hyperplane. It is possible to use the symmetric or asymmetric Ω area. The asymmetric Ω area can be defined as $\Omega = \Omega^- \cup \Omega^+$, where Ω^- and Ω^+ are the subareas with the mistakenly classified objects which belong to the classes with the labels "–1" and "+1" in the initial dataset correspondently. The symmetric Ω area can be defined as the area containing the objects located at the distance which isn't exceeding $\Delta = max\{d_{\Omega^-}, d_{\Omega^+}\}$, where d_{Ω^-} and d_{Ω^+} are the distances from the separating hyperplane to the furthest erroneously classified objects which belong to the classes with the labels "–1" and "+1" in the initial dataset correspondently. Also, some objects which are correctly classified by means of the SVM classifier can be in the Ω area.

3. To form the new dataset that will consist only of those objects of the initial dataset with their corresponding class labels, for which the class affiliation using the SVM classifier was defined correctly.

4. To develop the Parzen classifier based on the dataset obtained at the step 3, using the genetic algorithm applied to find the optimal parameters values of the Parzen classifier such as the number k of the nearest neighbors, the kernel function type and the distance metric type when classifying the objects from the dataset obtained at the step 2.

6. To estimate the final data classification quality with application of the quality indicators mentioned at the step 2.

5 Genetic algorithm

The search for the optimal parameters values of the Parzen classifier can be realized by looking through all possible combinations of the nearest neighbors k, the kernel function type and the distance metric type. However, such search can be accompanied by the tangible time expenditures with an increase in the number of the kernel functions types, the number of the distance metrics and, especially, the number of the possible neighbors. To reduce the time expenditures, it is suggested to use the genetic algorithm that operates with the integer parameters values.

Let the number k of the possible neighbors satisfies the condition: $1 \le k \le N/3$, where k is the integer odd number (that allow to avoid problems with voting); N is the of objects in the training dataset, formed at the step 3 of the hybrid intellectual classification technology; the distance metric types (5)–(8) are coded by the integer number from 1 to 4; the kernel function types (9)–(12) are coded by the integer number from 1 to 4. Then, the m-th chromosome in the genetic algorithm can be coded as:

$$v = (v_1, v_2, v_3), \qquad (13)$$

where v_1 is the number of the nearest neighbors; v_2 is the number of the kernel function type; v_3 is the number of the distance metric type.

The classification quality indicator such as the overall accuracy, sensitivity, specificity, F-measure can be used as the fitness function of the genetic algorithm. The fitness function should be maximized.

In this research the standard scheme of realization of the genetic algorithm has been used. It can be described by the following sequence of steps.

1. To initialize the chromosome population size P.

2. To calculate the "success" of each chromosome using the fitness function. To check the condition for stopping the genetic algorithm. If the break condition is satisfied, then exit from the genetic algorithm, otherwise go to step 3.

3. To perform the genetic selection operator to choose the most successful chromosomes for the role of parents to create the children generation on the basis of the parents genes.

4. To create the children generation with the use of the genetic operators of crossing and mutation. Depending on the operator used, the children can be produced from one parent by accidentally changing its genes (in the case of the mutation operator) or from two parents by combining their genes (in the case of the crossing operator).

5. To form the resulting generation of chromosomes by size P, using the generation of parents and the generation of children in accordance with the principle of elitism.

The following criteria can be used as the stopping criterion for the genetic algorithm: the criterion of achievement of the maximum number of generations, the criterion of achievement of the maximum execution time of the algorithm, the criterion of achievement of the maximum number of stagnant generations. The generation of chromosomes can be considered as the stagnant generation, if the best value of the fitness function does not differ by some small constant ε ($\varepsilon > 0$) from the best value of the fitness function of the previous generation.

The options for specifying the genetic operators such as selection, crossing and mutation (for example, the options for setting the parameters values of the crossing and mutation operators – the probabilities values of the crossing and mutation) are of essential interest.

Within the scope of the problem being solved, there is no need for excessively complex and self-adjusting crossing and mutation operators in view of the discreteness of the parameters values of the fitness function. In this connection, the standard options of specifying for the genetic operators are used: in the case of the mutation operator, the random numbers from the Gauss distribution are added to the values of the random chromosome genes of the parent, in the case of the crossing operator, the child chromosome randomly inherits the gene of one of the parents.

The considered genetic algorithm can be used after appropriate adaptation for the simultaneous search of the best variant of the auxiliary classifier from the set of the potentially possible classifiers and the parameters values of of the best auxiliary classifier.

For example, the kNN classifier and Parzen classifier can be used as the potential auxiliary classifiers, because the effectiveness of these classifiers has already been proven separately in the implementation of the hybrid intellectual technology. In the future, the set of the potentially auxiliary classifiers can be extended.

6 Experimental studies

Approbation of the offered data classification technique was made on the real datasets taken from the Statlog project and the UCI Machine Learning Repository. Particularly, we used experimental datasets of medical diagnostics [WDBC (http://archive.ics. uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/); Heart (http:// archive.ics.uci.edu/ml/machine-learning-databases/statlog/ heart/)], credit scoring (German, http://archive.ics.uci.edu/ ml/machine-learning-databases/statlog/german/), the artificially created data MOTP12 (MOTP12; http://machine-learning.ru/ wiki/images/b/b2/MOTP12_svm_ example.rar) (Table 1). The case of binary classification takes place in all experimental datasets.

Table 1. The values of the classifiers parameters and the values of the classification quality indicators.

Dataset	SVM classification			Classification type	Ω		Parzen classification			Number of errors	
	Number of support vectors	Number of errros			d_Ω	N_Ω	h	Metric	Kernel	Er_{train}	Er_{test}
		Er_{train}	Er_{test}								
1	2	3	4	5	6	7	8	9	10	11	12
WDBC (569 × 30)	455	0 of 456	42 of 113	SVM +aSim +Sim	— 0.373 0.745	— 42 42	— 3 3	— Euclid. Euclid.	— Epanechn. Epanechn.	0 0 0	42 1 1
German (1000 × 24)	796	0 of 800	60 of 200	SVM +aSim +Sim	— 0.784 1.567	— 199 434	— 137 -	— Chebyshev —	— Quadr. —	0 1 —	60 48 —
Heart (270 × 13)	216	0 of 216	20 of 54	SVM +aSim +Sim	— 0.138 0.276	— 37 38	— 23 33	— Chebyshev Chebyshev	— Epanechn. Gaussian	0 0 0	20 7 7
MOTP12 (400 × 2)	132	32 of 360	1 of 40	SVM +aSim +Sim	— 2.000 2.000	— 97 97	— 5 5	— Chebyshev Chebyshev	— Epanechn. Epanechn.	32 26 26	1 2 2

Each dataset was repeatedly divided into training and test sets with the subsequent training of the SVM classifier. The SVM classifiers were trained with the default parameters: the regularization parameter C is equal to 1, the kernel function is the radial basis, the kernel function parameter σ is equal to 1. Then, after selecting the partition with the smallest number of errors at the training and test sets, the main classification quality indicators were calculated. Further, the decision on the possibility of using of the Parzen classifier should be adopted: the majority of mistakenly classified objects should be on the separating strip, the number of representatives of each class should be at least 20% of the new reduced training set.

If the use of the Parzen classifier is approved, it is necessary to find the optimal values of the following parameters: the number of neighbors k, the metric and the kernel function. We use the genetic algorithm to solve this problem.

In this research, the size of the genetic algorithm population is equal to 20, the number of the stagnant generations is equal to 20, the minimum value deviation of the objective function for determining the stagnant generation is equal to 1e-10, the maximum generations number is equal to 200, the crossing probability is equal to 0.8, the mutation probability is equal to 0.2.

As the justification for the genetic algorithm's preference for the full search of the parameters values during the Parzen classifier development, the following example can be cited. For the MOTP12 dataset, consisting of 400 objects with two characteristics, the full search of the parameters values of the Parzen classifier requires the development of 864 Parzen classifiers, which demands 12 seconds, if 4 kernel functions, 4 distance metrics, and 67 variants of the number k of nearest neighbors (the number k of the nearest neighbors satisfies the condition: $1 \le k \le [320/3] = 107$, where k is the odd number) are used. A herewith, the training and test sets contained 320 and 80 objects, respectively.

The development of the Parzen classifier implies: the calculating of the distance from the object with the unknown class membership to each object of the training set, sorting of the objects of the training set in accordance with the calculated distance, determining of the class membership of the unknown object, taking into account the weights of the neighbors.

If the genetic algorithm is used to find the optimal parameters values of the Parzen classifier for 50 runs of the genetic algorithm, the average search time is 3 seconds for the population with 20 chromosomes and the number of stagnant generations which is equal to 5. The average number of the developed Parzen classifiers is equal to 185. Thus, the search time for the optimal parameters values of the Parzen classifier decreased by 4 times.

Column 1 of Table I contains the name of the experimental dataset, the number of objects and the number of object's features. Columns 2 – 4 combine information about the results of the SVM classifier developed at the first step: the number of the support vectors and the number of errors for each set. Column 5 indicates the algorithm type SVM classifier (SVM), SVM classifier and Parzen classifier with the asymmetric area (+aSim), SVM classifier and Parzen classifier with the symmetric area (+Sim). Columns 6 and 7 show the width of the Ω area and the number of objects lying in this area. The parameters values found by the genetic algorithm are shown in columns 8 – 10. Number of errors at the training and test sets are shown in columns 11 and 12.

For the WDBC dataset the SVM classifier gives 0 and 42 errors at the training and test sets respectively. The mistakenly classified objects are at the distance of 0.373 from the separating strip on the one side (the supports vectors are at distance of 1 from the separating hyperplane). The symmetric and asymmetric areas contain equal number of objects. Therefore, the Parzen classifier will give the identical results for both areas. The areas with the mistakenly classified objects contain 42 objects. The search of the optimum parameters' values of

the Parzen classifier with application of the genetic algorithm has defined the number of neighbors equal to 3 at the Euclidean metrics and the Epanechnikov kernel function. The number of errors at the use of the offered classification technology equal to 0 and 1 at the training and test sets respectively.

For the German data the SVM classifier gives 0 and 60 errors at the training and test sets respectively. The use of the symmetric area is inexpedient as the number of the excluded objects makes the most part of initial dataset and owing to the removal of objects the classes become considerably unbalanced. At the removal of objects of the symmetric area, the number of errors at the training and test sets equal to 1 and 48 respectively. Increase in the number of errors at the training set can be explained by a possible removal of the support vectors. The search of the optimum parameters' values of the Parzen classifier with application of the genetic algorithm has defined the number of neighbors equal to 137 at the Chebyshev's metrics and the Epanechnikov kernel function.

For the Heart dataset the SVM classifier gives 0 and 20 errors at the training and test sets respectively. The asymmetric and symmetric areas contain 37 and 38 objects respectively. The removal of objects for both variants of areas and the subsequent use of the Parzen classifier has lowered the number of errors at test set to 7. The search of the optimum parameters' values of the Parzen classifier with application of the genetic algorithm has defined the number of neighbors equal to 33 at the Chebyshev's metrics and the Gaussian kernel function.

For the MOTP12 dataset the SVM classifier gives 32 and 1 errors at the training and test sets respectively. As this dataset has the difficult division into classes, the large number of the mistakenly classified objects lies out of the separating strip. In the case, when the mistakenly classified objects lie out of the dividing strip, the distance from the dividing strip to these objects is big than 1. In order to avoid the excess removal of objects, the decision on the exception only of objects in the strip is made. Therefore, the width of the symmetric and asymmetric area is equal to 2. In both cases the area includes 97 objects. The search of the optimum parameters' values of the Parzen classifier with application of the genetic algorithm has defined the number of neighbors equal to 5 at the Chebyshev's metrics and the Epanechnikov kernel function. The number of errors at the use of the offered classification technology equal to 26 and 2 at the training and test sets respectively.

7 Conclusions

By results of the conducted experimental researches, it is possible to draw a conclusion that the use of the offered technique increases the classification quality, as the application of the Parzen classifier to the objects located near the hyperplane dividing the classes and determined by the SVM classifier reduces the number of the mistakenly classified objects. The offered classification technique allows making the high-precision decisions on classification of the elaborate multidimensional data.

The use of the Parzen classifier as the auxiliary classifier instead of the *k*NN classifier can be justified by the fact that the *k*NN classifier uses the distance to the object to determine the nearest neighbors, but does not use this distance for making decisions on the class affiliation of the object. The definition of the weights functions as the dependence of the quantitative nearest neighbor number in some modifications of the *k*NN classifier is impractical, since it is not possible to evaluate the actual location of objects in the feature space. The Parzen classifier makes the decision on the class affiliation of object on the basis of the distance to the neighbor by means of the kernel function.

The purpose of further research is the adaptation of the genetic algorithm to realization of the simultaneous search for the best variant of the auxiliary classifier from the set of the potentially possible classifiers and the parameters values of the best auxiliary classifier.

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

References

1. V. Vapnik, *Statistical Learning Theory* (New York: John Wiley & Sons, 1998)
2. O. Chapelle, V. Vapnik, O. Bousquet, S. Mukherjee, Machine Learning, **46**, 131 (2002)
3. L. Yu, S. Wang, K. K. Lai, L. Zhou, *Bio-Inspired Credit Risk Analysis. Computational Intelligence with Support Vector Machines* (Springer, 2008)
4. L. Demidova, Yu. Sokolova, *International Conference "Stability and Control Processes" in Memory of V.I. Zubov* (2015)
5. L. Demidova, E. Nikulchev, Y. Sokolova, International Journal of Advanced Computer Science and Applications, 7, 294 (2016)
6. H. Wang, D. Bell, The Computer Journal, **47**, 11 (2004)
7. R. Li, H.-N. Wang, H. He, Y.-M. Cui, Zh.-L. Du, Chinese Journal of Astronomy and Astrophysics, **7**, 7 (2007)
8. H. Zhang, A.C. Berg, M. Maire, J. Malik, *Computer Society Conference on Computer Vision and Pattern Recognition*, **2**, 11 (2006)
9. L. Demidova, Yu. Sokolova, *6-th Mediterranean Conference on Embedded Computing*, 4 (2017)
10. E. Parzen, *The Annals of Mathematical Statistics*, **33**, 12 (1962)