International Journal of Smart Home
Vol. 10, No. 5 (2016), pp. 119-130
http://dx.doi.org/10.14257/ijsh.2016.10.5.12

Study on a Novel Data Classification Method Based on Improved GA and SVM Model

Jing Huo and Yuxiang Zhao

School of Electronic Information and Electronical Engineering, Tianshui Normal University, Tianshui 741001, Gansu, China

Abstract

Support vector machine(SVM) can effectively solve the classification problem with small samples, nonlinear and high dimensions, but it exits the weak generalization ability and low classification accuracy. So an improved genetic algorithm(IGA) is introduced in order to propose a new classification(IGASVM) method based on combining improved GA and SVM model. In the proposed IGASVM method, the self-adaptive control parameter strategy and improving convergence speed strategy are introduced into the GA to keep the diversity of the population, promptly reflect the premature convergence of the individual and escape from the local optimal solution for improving the search performance. Then the improved GA is used to optimize and determine the parameters of the SVM model in order to improve the learning ability and generalization ability of the SVM model for obtaining new classification (IGASVM) method. Finally, the experiment data is selected to test the effectiveness of the proposed IGASVM method. The experiment results show that the improved GA can effectively optimize and determine the parameters of the SVM model, and the IGASVM method takes on the better learning ability, generalization ability and classification accuracy.

***Keywords**: Support vector machine; genetic algorithm; classification; optimization; generalization ability; learning ability*

1. Introduction

The human beings have entered the period of information explosion in the 21st century. people in various fields of scientific research and application, and even daily life can not take a large amount of data as a information source. However, now people have been unable to use the simple and intuitive methods to process data for extracting accurate information. So how to effectively mine practical value information from a large amount of data has become a research problem. The data classification is an important data analysis method. In order to solve the classification problem of information data, researchers have been studying and proposed a variety of data classification methods[1], such as neural network, support vector machine(SVM), k nearest neighbor method, Bayes, Adaboost and so on.

From the use and result of these data analysis methods, the classification result of SVM is best. The SVM classifier is a classification method based on statistical learning theory[2]. It is the principle of structural risk minimization. And it can better solve these classification methods with small sample, nonlinear, high dimension and so on, take on good generalization ability, overcome these problems of local optimal solution, large sample, slow convergence speed and so on in the neural network. It also has strong practicability. However, the SVM is a new method of machine learning, it exits the biggest problem that is the selection of kernel function and the optimization of parameters. For some data sets, the different kernel functions have similar effects on the final classification results, but for the other data sets, the selection of kernel functions will have an important impact on the

ISSN: 1975-4094 IJSH

classification results. So a lot of researchers proposed many methods for improving the SVM model. Fung and Mangasarian[3] proposed a concave minimization approach for classifying unlabeled data based on the small representative percentage and linear support vector machine. Park and Zhang[4] proposed an approach for classifying large scale unstructured documents by incorporating both the lexical and the syntactic information of documents. Liu[5] proposed an active learning algorithm with support vector machine for performing active learning with support vector machine and applied the algorithm to gene expression profiles of colon cancer, lung cancer, and prostate cancer samples. Jayadeva *et al.*[6] proposed a multi-category extension of fuzzy proximal support vector machines, where a fuzzy membership is assigned to each data point. Jin *et al.*[7] proposed a genetic fuzzy feature transformation method for support vector machines (SVMs) to do more accurate data classification. Yang *et al.*[8] proposed a weighted support vector machine (WSVM) to improve the outlier sensitivity problem of standard support vector machine (SVM) for two-class data classification. The basic idea is to assign different weights to different data points such that the WSVM training algorithm learns the decision surface according to the relative importance of data points in the training data set. Li *et al.*[9] proposed a clustering algorithm for efficient learning. The method mainly categorizes data into clusters, and finds critical data in clusters as a substitute for the original data to reduce the computational complexity. Cervantes[10] proposed a novel SVM classification approach for large data sets by using minimum enclosing ball clustering. Mathur and Foody[11] proposed a crop classification method based support vector machine with intelligently selected training data for an operational application. Essam[12] proposed a new accurate classifier based on Signal-to-Noise, support vector machine, Bayesian neural network and AdaBoost for data mining and classification. Li and Liu[13] proposed a new kernel generating method dependent on classifying related properties of the data structure itself. The new kernel concentrates on the similarity of paired data in classes, where the calculation of similarity is based on fuzzy theories. Ji *et al.*[14] introduced the support vector machine which the training examples are fuzzy input, and give some solving procedure of the Support vector machine with fuzzy training data. Jordi *et al.*[15] proposed two semisupervised one-class support vector machine (OC-SVM) classifiers for remote sensing applications. The first proposed algorithm is based on modifying the OC-SVM kernel by modeling the data marginal distribution with the graph Laplacian built with both labeled and unlabeled samples. The second one is based on a simple modification of the standard SVM cost function which penalizes more the errors made when classifying samples of the target class. Al-Ataby *et al.*[16] proposed several multi-resolution approaches employing the wavelet transform and texture analysis for de-noising and enhancing the quality of data to help in the automatic detection and classification of defects. Hwang *et al.*[17] proposed a new weighted approach on Lagrangian support vector machine for imbalanced data classification problem. The weight parameters are embedded in the Lagrangian SVM formulation. Jan *et al.*[18] proposed a mixed effects least squares support vector machine model to extend the standard LS-SVM classifier for handling longitudinal data. Tian *et al.*[19] proposed a new method based on support vector machine (SVM) and genetic algorithm (GA) to analyze signals of wound infection detection. Ahmed *et al.*[20] proposed to apply an ensemble of SVMs coupled with feature-subset selection methods to alleviate the curse of dimensionality associated with expression-based classification of DNA data in order to achieve stable and reliable results. Li *et al.*[21] proposed a novel SVM classification approach based on the random selection and de-clustering technique for large data sets. Chau *et al.*[22] proposed a novel method for SVM classification based on convex-concave hull and support vector machine, called convex-concave hull SVM (CCH-SVM). Li *et al.*[23] proposed a probabilistic support vector machine (PSVM) to capture the probabilistic information of the separating margin and formulate the decision function within such a noisy environment. Wei *et al.*[24] proposed a least squares support vector machine with L1 norm (LS-SVM-L1) to deal with above shortcomings. This method is equivalent to solve a linear equation set with deficient rank just like the over complete problem in independent component analysis (ICA). Xu *et al.*[25] proposed a model based on particle swarm optimization (PSO) and support vector machine (SVM) for using in the classification of a large, imbalanced dataset. This model is referred to as the PSO-SVM (particle swarm optimization-based support vector machine) model. Zhang *et al.*[26] proposed a scaling kernel-based support vector machine (SVM) approach to deal with the multi-class imbalanced data classification problem and a scaling kernel function and calculate its parameters using the chi-square test and weighting factors. Bordoloi and Tiwari[27] proposed a multi-fault classification of gears based on support vector machine (SVM) learning techniques with the help of time-frequency (wavelet) vibration data. Cervantes *et al.*[28] proposed a genetic algorithm (GA)-based classification method. A draft hyperplane and support vectors are first generated by SVMs. Then, GA is applied to compensate the imbalanced data. Shao *et al.*[29] proposed an efficient weighted Lagrangian twin support vector machine (WLTSVM) for the imbalanced data classification based on using different training points for constructing the two proximal hyperplanes. Maldonado and López[30] proposed a novel second-order cone programming (SOCP) formulation, based on the LP-SVM formulation principle: the bound of the VC dimension is loosened properly using the l-norm, and the margin is directly maximized using two margin variables associated with each class. Peng and Xu[31] proposed a structural regularized PTSVM (SRPTSVM) classifier for binary classification. This proposed SRPTSVM focuses on the cluster-based structural information of the corresponding class in each optimization problem, which is vital for designing a good classifier in different real-world problems. Rebentrost *et al.*[32] proposed an optimized binary classifier, called quantum support vector machine for big data classification. Jorge *et al.*[33] proposed a novel contextual classifier based on a Support Vector Machine (SVM) and an Evolutionary Majority Voting (SVM-EMV) to develop thematic maps from LiDAR and imagery data. Subsequently, the performance of SVM-EMV is compared to that achieved by a pixel-based SVM as well as to a contextual classified based on SVM and MRF. Tomar and Agarwal[34] proposed an effective Weighted Multi-class Least Squares Twin Support Vector Machine (WMLSTSVM) approach to address the problem of imbalanced data classification for multi class. This research work employs appropriate weight setting in loss function. Hejazi *et al.*[35] proposed an experimental study on multiclass Support Vector Machine (SVM) methods over a cardiac arrhythmia dataset that has missing attribute values for electrocardiogram (ECG) diagnostic application. Hüseyin *et al.*[36] proposed an effective classification approach based on k-means and least square support vector machine to classify harmonic data. Mohd Pozi *et al.*[37] proposed a new optimization task based on genetic programming, built on top of support vector machine, in order to improve the classification rate for minority class without significant reduction on accuracy metric. Shounak and Swagatam[38] proposed a Near-Bayesian Support Vector Machine (NBSVM) for such imbalanced classification problems, by combining the philosophies of decision boundary shift and unequal regularization costs.

The genetic algorithm(GA)is a population-based stochastic search technique. It is an ideal method for solving optimization problem, and takes on high optimization efficiency and is easy to jump out the local sub optimal solution. But in the process of solving optimization problem, it exits poor local search ability, which leads to the time-consuming and lower search efficiency in the later stage of evolution. So a self-adaptive control parameter strategy and improving convergence speed strategy are introduced into the GA in order to keep the diversity of the population, promptly reflect the premature convergence of the individual and escape from the local optimal solution for improving the search performance. Then the improved GA is used to optimize the parameters of kernel function and penalty factor in order to propose a new classification(IGASVM) method for improving the operation speed and classification precision.

The rest of this paper is organized as follows. Section 2 briefly introduces the GA and SVM model. Section 3 proposed an improved GA. Section 4 presents a new classification(IGASVM) method based on improved GA and SVM model. In this section, the idea of the IGASVM method is introduced in detail. Section 5 is experiment and simulation analysis. Finally, the conclusions are discussed in Section 6.

2. GA and SVM

2.1. Genetic Algorithm

The GA[39] is a class of population-based stochastic search technique that is used to solve the complex problems by imitating processes observed during natural evolution. The GA is based on the principle of the survival and reproduction of the fitness. It is a parallel iterative algorithm with certain learning ability, which repeats the evaluation, selection, crossover and mutation operator until the stopping criteria is reached. A fitness function is used to evaluate the fitness value of each chromosome. A real-coded GA is a genetic representation that uses a vector of floating-point numbers instead of 0 and 1 for implementing chromosome encoding. The crossover operator of a real-coded GA is performed by using the concept of convex combination. The random mutation operator is used to change the gene with one random number. Assuming that we employ GA to search for the largest fitness value with the given fitness function, shown in Figure 1.

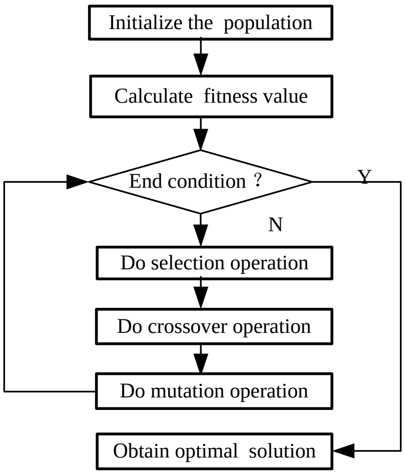

Figure 1. Searching of the GA

2.2. Support Vector Machine(SVM)

Support vector machine (SVM) [2], introduced by Vapnik is one of the popular tools for a supervised machine learning method based on structural risk minimization. The basic characteristic of SVM is to map the original nonlinear data into a higher-dimensional feature space where a hyperplane is constructed to bisect two classes of data and maximize the margin of separation between itself and those points lying nearest to it (the support vectors). The hyperplane should be used as the basis for classifying unknown data. The LS-SVM model is to use the least square linear system as the loss function, the inequality constraints is revised the equality constraints in the standard SVM.

The Given the training sample is $S = \{(x_i, y_i) \mid i = 1,2,3,\cdots,m\}$, m is the number of samples, the set $\{x_i\} \in R_n$ represents the input vector, $y \in \{-1,1\}$ indicates the corresponding desired output vector, the input data is mapped into the high dimensional feature space by using nonlinear mapping function $\phi(\bullet)$. Then the existed optimal classification hyper plane must meet the following conditions:

$$\begin{cases} \omega^T x_i + b \geq 1, & y_i = 1 \\ \omega^T x_i + b \leq -1, & y_i = -1 \end{cases} \tag{1}$$

where ω is Omega vector of super plane, b is offset quantity. Then the classification decision function is described as follow:

$$f(x_i) = \mathrm{sgn}(\omega^T x_i + b) \tag{2}$$

The classification model of SVM model is described by he optimization function $\min\limits_{\omega,\xi,b} J(\omega,\xi_i)$:

$$\min_{\omega,\xi,b} J(\omega,\xi_i) = \frac{1}{2}\omega^T\omega + \frac{1}{2}\gamma\sum_{i=1}^{m}\xi_i^2 \tag{6}$$

$$s.t. \quad y_i[\omega^T\phi(x_i) + b] = 1 - \xi_i, i = 1,2,3,\cdots,m \tag{3}$$

where ξ_i is slack variable, b is offset, ω is support vector, $\xi = (\xi_1,\xi_2,\cdots,\xi_m)$, γ is classification parameter for balancing the fitting error and model complexity.

The optimization problem transforms into its dual space. Lagrange function is introduced to solve it. The corresponding optimization problem of the SVM model with Lagrange function is:

$$L(\omega,b,\xi,\alpha) = \frac{1}{2}\omega^T\omega + \frac{1}{2}\gamma\sum_{i=1}^{m}\xi_i^2 - \sum_{k=1}^{m}\alpha_i\{y_i[\omega^T\phi(x_k) + b] - 1 + \xi_i\} \tag{4}$$

where α_i is the Lagrange multiplier, and $\alpha_i \geq 0 (i = 1,2,3.\cdots,m)$. The optimal conditions are described:

$$\begin{cases} \dfrac{\partial L}{\partial \omega} = 0 \Rightarrow \omega = \sum\limits_{i=1}^{m}\alpha_i y_i \phi(x_i) \\ \dfrac{\partial L}{\partial b} = 0 \Rightarrow \sum\limits_{i=1}^{m}\alpha_i y_i = 0 \\ \dfrac{\partial L}{\partial \xi_i} = 0 \Rightarrow \alpha_i = \gamma\xi_i \\ \dfrac{\partial L}{\partial \alpha_i} = 0 \Rightarrow y_i(\omega^T\phi(x_i) + b) - 1 + \xi = 0_i \end{cases} \tag{5}$$

The following linear equation is obtained:

$$\begin{bmatrix} 0 & L^T \\ L & \Omega + \gamma^{-1}I \end{bmatrix}\begin{bmatrix} b \\ \alpha \end{bmatrix} = \begin{bmatrix} 0 \\ Y \end{bmatrix} \tag{6}$$

where $Y = [y_1, y_2, \cdots, y_m]^T \in R^m$, $L \in R^m$ is vector of the element m, $y^T = [y_1, y_2, \cdots, y_m]$, I is unit matrix, $I_m = [1,1,\cdots,1]^T$, $\alpha = [\alpha_1, \alpha_2, \cdots, \alpha_m]^T$, $\Omega = [\Omega_{ij}]_{m \times m}$, $\Omega_{ij} = y_i y_j K(x_i, x_j)$. Then the classification decision function is described as follow:

$$f(x_i) = \mathrm{sgn}(\sum_{i=1}^{m} \alpha_i y_i K(x, x_i) + b) \tag{7}$$

There are several kernel functions, such as linear kernel function, polynomial kernel function, radial basis kernel function(RBF), Sigmoid kernel function and Fourier kernel function and so on. For the data classification of different system, there has the corresponding optimal kernel function. Because the RBF has the advantages of simple form, symmetry radial, good smoothness and analyticity and so on. So the RBF is selected to be regarded as kernel function of the SVM model. The specific express of the RBF is shown as follows:

$$K(x, x_i) = \exp[\,-(x - x_i)^2 / 2\sigma^2\,] \tag{8}$$

The regularization parameter(γ) is an important parameters in the SVM regression model. Their selection directly influences the learning ability and generalization performance.

3. An Improved GA

3.1. Self-Adaptive Control Parameter Strategy

In order to balance the search range and search ability of the GA, a self-adaptive control parameter strategy is introduced into the GA in order to keep the diversity of the population, promptly reflect the premature convergence of the individual and escape from the local optimal solution for improving the search performance. $f(t)$ is the fitness value of the individual, $\overline{f(t)}$ is the average fitness value of the population, $\overline{\overline{f(t)}}$ is the fitness value of these the individuals with redoing the mean. And the premature convergence extent of the population is $\Phi = f_{GA}^{\max}(t) - \overline{\overline{f(t)}}$. Then the values of $P_c(t)$ and $P_m(t)$ are calculated by using the following expressions:

$$P_c(t) = \begin{cases} k_1 & , f_{GA}'(t) \le \overline{\overline{f(t)}} \\ \dfrac{k_2(f_{GA}^{\max}(t) - f_{GA}'(t))}{f_{GA}^{\max}(t) - \overline{\overline{f(t)}}} & , f_{GA}'(t) \ge \overline{\overline{f(t)}} \end{cases} \tag{9}$$

$$P_m(t) = \begin{cases} k_3 & , f_{GA}(t) \le \overline{\overline{f(t)}} \\ \dfrac{k_4(f_{GA}^{\max}(t) - f_{GA}(t))}{f_{GA}^{\max}(t) - \overline{\overline{f(t)}}} & , f_{GA}(t) \ge \overline{\overline{f(t)}} \end{cases} \tag{10}$$

where the values of the k_1, k_2, k_3 and k_4 are [0,1]. The $f_{GA}'(t)$ is the larger value of fitness function in two crossover individuals. The $f_{GA}(t)$ is the used function value of mutation individual.

3.2. Improving Convergence Speed Strategy

The fitness degree is adjusted to avoid the premature convergence and stagnation phenomenon in the late stage of the search in this paper. The average value of the population $\overline{f(t)}$ and variance δ are defined as follows:

$$\overline{f(t)} = \frac{\sum_{i=1}^{N} f_i}{N} \tag{11}$$

$$\delta = \sqrt{\frac{\sum_{i=1}^{N} (f_i - \overline{f(t)})}{N}} \tag{12}$$

where f_i is fitness value of the i^{th} individual, N is the size of the population.

The fitness value in the GA is a criterion for evaluating individual quality. The selection of the individual with a certain probability is according to the fitness of the individual. If the fitness value of the individual is larger, the individual is more likely to be repeated selection, and the fitness of the individual is smaller, the individual is not selected. This will realize the survival of the fittest.

4. A New Classification(IGASVM) Method

The kernel function of support vector machine is to convert the nonlinear separable sample into linear separable feature space. The selected different kernel generates the different classification superplane of SVM. So the change of kernel function can make the SVM with larger difference, it directly affects the performance of SVM. And the change of the kernel function parameter γ actually alters the parameter of mapping function and function relation, and also alters the complexity of sample mapping feature space. So the performance of SVM is also influenced by the kernel function parameter γ . At the same time, the penalty factor C is to realize a compromise between the proportion of error sample and the complexity of algorithm. In the determined feature subspace, the proportion of the learning machine's confidence range and the empirical risk is adjusted to control the generalization ability of SVM. Genetic algorithm is a random search algorithm by referencing the natural selection and natural genetic mechanism in the biological community. It is widely used to solve the optimization problem. Its superiority is mainly described: the characteristics of population search and intrinsic heuristic random search, the inherent parallelism and the ability of parallel computing. But in the process of solving optimization problem, it exits poor local search ability. So a self-adaptive control parameter strategy and improving convergence speed strategy are introduced into the GA in order to keep the diversity of the population, promptly reflect the premature convergence of the individual and escape from the local optimal solution for improving the search performance. Then the improved GA is used to select and optimize the parameters of kernel function and penalty factor in order to propose a new classification(IGASVM) method for improving the speed and efficiency of parameter selection and the learning ability and generalization ability of SVM.

5. Experiment and Simulation Analysis

In order to verify the performance of the proposed IGASVM method, Wine, Pittsburgh Bridges, Balance Scale, Libras Movement, Arrhythmia, Low Resolution, Spectrometer in UCI database are selected as the experiment data. UCI database is a database, which is used in machine learning. The composition and dimension of data set sample are shown Table 1. The experiment works on Intel(R) Core i5,2G RAM with Windows XP and Matlab 2012. The basic SVM and GA-SVM are selected for comparing the classification ability with the proposed IGASVM method. Because the initial values of parameters in SVM, GA-SVM and IGASVM could seriously affect the experiment result, the most reasonable initial values are obtained by continuously testing. The obtained initial values of parameters in SVM, GA-SVM and IGASVM are shown: popsize m =40, crossover probability P_c =0.05, mutation probability P_m =0.5, maximum iteration times $T_{\max}$ =1000. the penalty parameter $C = 100$, the value range of the radial basis kernel width is σ =[0,10], the value range of the regularization parameter is γ =[1,1000]. The RBF kernel function is selected as the kernel function in SVM model.

Table 1. The Data Sets

Index	Data	Sample number	Attribute
1	Wine	178	13
2	Pittsburgh Bridges	108	13
3	Balance Scale	625	4
4	Libras Movement	360	91
5	Arrhythmia	452	279
6	Spectrometer	531	102

The SVM and GA-SVM are select to compare with the proposed IGASVM algorithm. The classification results are shown in Table 2.

Table 2. The Classification Results

Index	Data	Sample number	Classification accuracy(%)		
			SVM	GA-SVM	IGASVM
1	Wine	178	81.73	88.62	**93.65**
2	Pittsburgh Bridges	108	82.95	85.36	**94.79**
3	Balance Scale	625	83.68	89.17	**90.34**
4	Libras Movement	360	76.81	83.72	**91.48**
5	Arrhythmia	452	74.21	78.63	**90.65**
6	Spectrometer	531	82.36	85.94	**91.13**

Table 2. shows that the results are obtained using SVM,GA-SVM and IGASVM algorithm and the optimal results are represented by using black body. As can be seen

from Table 2, the proposed IGASVM algorithm can obtain best classification results in the given data from the UCI database. The average classification accuracy respectively is 93.65%, 94.79%, 90.34%, 91.48%, 90.65% and 91.13% for Wine, Pittsburgh Bridges, Balance Scale, Libras Movement, Arrhythmia and Spectrometer. And the improved GA can better obtain the parameters of kernel function and penalty factor of SVM for improving the speed and efficiency of parameter selection and the learning ability and generalization ability. In general, the classification results of the IGASVM algorithm are more better and has higher optimization ability and classification accuracy.

6. Conclusion

In this paper, an improved GA based on self-adaptive control parameter strategy and improving convergence speed strategy is proposed. Then a new classification(IGASVM) method based on improved GA and SVM model is proposed. Firstly, the self-adaptive control parameter strategy and improving convergence speed strategy are introduced into the GA in order to keep the diversity of the population, promptly reflect the premature convergence of the individual and escape from the local optimal solution for improving the search performance. Then the improved GA is used to select and optimize the parameters of kernel function and penalty factor of SVM for improving the speed and efficiency of parameter selection and the learning ability and generalization ability. Finally, the proposed IGASVM method is applied to solve the classification of Wine, Pittsburgh Bridges, Balance Scale, Libras Movement, Arrhythmia, Low Resolution, Spectrometer from UCI database. The experiment results show that the improved GA can effectively optimize and determine the parameters of the SVM model, and the IGASVM method takes on the better learning ability, generalization ability and classification accuracy.

Acknowledgements

This research was supported by Colleges' and Universities' Scientific Research Projects of Gansu Province (2015B-104), Science and Technology Projects of Gansu Province (1506RJZE115). The program for the study, initialization, training, and simulation of the proposed algorithm in this article was written with the tool-box of MATLAB 2010b produced by the Math-Works.

References

[1] M. Boucekine, A. Loundou and K. Baumstarck, “Using the random forest method to detect a response shift in the quality of life of multiple sclerosis patients: a cohort study”, BMC medical research methodology, vol. 13, no. 1,(**2013**), pp.1-8.

[2] V. Cortesc, “Support vector networks”, Machine Learning, vol. 20, no. 3, (**1995**), pp. 273 -297.

[3] G. Fung and O. L. Mangasarian, “Semi-supervised support vector machines for unlabeled data classification”, Optimization Methods and Software, vol. 15, no. 1, (**2002**), pp. 29-44.

[4] S.B. Park and B.T. Zhang, “Co-trained support vector machines for large scale unstructured document classification using unlabeled data and syntactic information”, Information Processing and Management, vol. 40, no. 3, (**2004**), pp. 421-439.

[5] Y. Liu, “Active learning with support vector machine applied to gene expression data for cancer classification”, Journal of Chemical Information and Computer Sciences, vol. 44, no. 6, (**2004**), pp. 1936-1941.

[6] D. Jayadeva, R. Khemchandani and S. Chandra, “Fuzzy linear proximal support vector machines for multi-category data classification”, Neurocomputing, vol. 67, no. 1-4, (**2005**), pp.426-435.

[7] B. Jin, Y.C. Tang and Y. Q. Zhang, “Support vector machines with genetic fuzzy feature transformation for biomedical data classification”, Information Sciences, vol.177, no. 2, (**2007**), pp. 476-489.

[8] X. L. Yang, Q. Song and Y. Wang, “A weighted support vector machine for data classification”, International Journal of Pattern Recognition and Artificial Intelligence, vol. 21, no. 5, (**2007**), pp. 961-976.

[9] D.C. Li and Y. H. Fang, "An algorithm to cluster data for efficient classification of support vector machines", Expert Systems with Applications"0, vol. 34, no. 3, (**2008**), pp. 2013-2018.

[10] J. Cervantes, X. O. Li, W. Yu and K. Li, "Support vector machine classification for large data sets via minimum enclosing ball clustering.", Neurocomputing, vol. 71, no. 4-6, (**2008**), pp. 611-619.

[11] A. Mathur and G. M. Foody, "Crop classification by support vector machine with intelligently selected training data for an operational application", International Journal of Remote Sensing, vol. 29, no. 8, (**2008**), pp.2227-2240.

[12] A.D. Essam, "Integration of support vector machine and Bayesian neural network for data mining and classification", World Academy of Science, Engineering and Technology, vol. 64, (**2010**), pp. 202-207.

[13] D. C. Li and C.W. Liu, "A class possibility based kernel to increase classification accuracy for small data sets using support vector machines", Expert Systems with Applications, vol. 37, no. 4, (**2010**), pp. 3104-3110.

[14] A. B. Ji, J.H. Pang and H. J. Qiu, "Support vector machine for classification based on fuzzy training data", Expert Systems with Applications, vol. 37, no. 4, (**2010**), pp. 3495-3498.

[15] M. M. Jordi, B. Francesca, G. C. Luis, B. Lorenzo and C.V. Gustavo, "Semisupervised one-class support vector machines for classification of remote sensing data', IEEE Transactions on Geoscience and Remote Sensing, vol. 48, no. 8, (**2010**), pp. 3188-3197.

[16] A. A.-Ataby, W. A.-Nuaimy, C.R. Brett and O. Zahran, "Automatic detection and classification of weld flaws in TOFD data using wavelet transform and support vector machines", Insight: Non-Destructive Testing and Condition Monitoring, vol. 52, no. 11, (**2010**), pp.597-602.

[17] J.P. Hwang, S. Park and E. Kim, "A new weighted approach to imbalanced data classification problem via support vector machine with quadratic cost function", Expert Systems with Applications, vol. 38, no. 7, (**2011**), pp. 8580-8585.

[18] L. Jan , M. Geert, V. Geert, V.H. Sabine and J. A. K. Suykens, "A mixed effects least squares support vector machine model for classification of longitudinal data", Computational Statistics and Data Analysis, vol. 56, no. 3, (**2012**), pp.611-628.

[19] F.C. Tian, J.X. Yan, S. Xu, J.W. Feng, Q.H. He, Y. Shen, P.F. Jia and C.B. Kadri, "Classification of electronic nose data on wound infection detection using support vector machine combined GA", Journal of Computational Information Systems, vol. 8, no. 8, (**2012**), pp.3349-3357.

[20] E. Ahmed, E. G. Neamat and I.A.E. Azab, "Support vector machine ensembles using feature-subset selection for enhancing microarray data classification", International Journal of Applied Mathematics and Statistics, vol. 28, no. 4, (**2012**), pp. 1-11.

[21] X.O. Li, J. Cervantes and W. Yu, "Fast classification for large data sets via random selection clustering and Support Vector Machines", Intelligent Data Analysis, vol.16, no. 6, (**2012**), pp. 897-914.

[22] A. L. Chau, X. O. Li and W. Yu, "Large data sets classification using convex-concave hull and support vector machine", Soft Computing, vol.17, no. 5, (**2013**), pp. 793-804.

[23] H.X. Li, J.L. Yang,G. Zhanga and B. Fan, "Probabilistic support vector machines for classification of noise affected data", Information Sciences, vol. 221, (**2013**), pp. 60-71.

[24] L.W. Wei, H. Yu and J.H. Liu, "Data classification using sparse and robust model: Least squares support vector machine with L1 norm", Computer Modelling and New Technologies, vol. 18, no. 12, (**2014**), pp. 686-691.

[25] Z.Y. Xu, J. Watada, M.N. Wu, Z. Ibrahim and M. Khalid, "Solving the imbalanced data classification problem with the particle swarm optimization based support vector machine", IEEJ Transactions on Electronics, Information and Systems, vol. 134, no. 6, (**2014**), pp. 788-795.

[26] Y. Zhang, P. P. Fu, W. Z. Liu and G. L. Chen, "Imbalanced data classification based on scaling kernel-based support vector machine", Neural Computing and Applications, vol. 25, no. 3-4, (**2014**), pp. 927-935.

[27] D. J. Bordoloi and R. Tiwari, "Support vector machine based optimization of multi-fault classification of gears with evolutionary algorithms from time-frequency vibration data", Measurement: Journal of the International Measurement Confederation, vol. 55, no. 9, (**2014**), pp. 1-14.

[28] J. Cervantes, X. O. Li and W. Yu, "Imbalanced data classification via support vector machines and genetic algorithms", Connection Science, vol. 26, no. 4, (**2014**), pp. 335-348.

[29] Y. H. Shao, W. J. Chen, J. J. Zhang, Z. Wang, N. Y. Deng, "An efficient weighted Lagrangian twin support vector machine for imbalanced data classification", Pattern Recognition, vol. 47, no. 9, (**2014**), pp. 3158-3167.

[30] S. Maldonado and J. López, "Imbalanced data classification using second-order cone programming support vector machines", Pattern Recognition, vol. 47, no. 5, (**2014**), pp. 2070-2079.

[31] X. J. Peng and D. Xu, "Structural regularized projection twin support vector machine for data classification", Information Sciences, vol. 279, no. 9, (**2014**), pp. 416-432.

[32] P. Rebentrost, M. Mohseni and S. Lloyd, "Quantum support vector machine for big data classification", Physical Review Letters, vol. 113, no. 13, (**2014**), pp. 1-8.

[33] G.G. Jorge , M,G, Daniel, G. Mariano and R.S. José, "An evolutionary-weighted majority voting and support vector machines applied to contextual classification of LiDAR and imagery data fusion", Neurocomputing, vol. 163, no. 9, (**2015**), pp.17-24.

[34] D. Tomar and S. Agarwal, “An effective Weighted Multi-class Least Squares Twin Support Vector Machine for Imbalanced data classification”, International Journal of Computational Intelligence Systems, vol. 8, no. 4, (**2015**), pp. 761-778.

[35] M. Hejazi, S.A.R. Al-Haddad,Y. P. Singh, S. J. Hashim and F. A. Aziz, “Multiclass Support Vector Machines for Classification of ECG Data with Missing Values”, Applied Artificial Intelligence, vol. 29, no. 7, (**2015**), pp. 660-674.

[36] E. Hüseyin, T. Vedat, Y. Özal, E. Belkis and D. Yakup, “Automatic classification of harmonic data using k-means and least square support vector machine”, Turkish Journal of Electrical Engineering and Computer Sciences, vol. 23, no. 5, (**2015**), pp. 1312-1325.

[37] M. S. Mohd Pozi, M. N. Sulaiman, N. Mustapha and T. Perumal, “A new classification model for a class imbalanced data set using genetic programming and support vector machines: Case study for wilt disease classification”, Remote Sensing Letters, vol. 6. no. 7, (**2015**), pp. 568-577.

[38] D. Shounak and D, Swagatam, “Near-Bayesian Support Vector Machines for imbalanced data classification with equal or unequal misclassification costs”, Neural Networks, vol. 70, (**2015**), pp. 39-52.

[39] K.Y. Lee and F.F. Yang, “Optimal reactive power planning using evolutionary algorithms: A comparative study for evolutionary programming”, evolutionary strategy, genetic algorithm, and linear programming. IEEE Transactions On Power Systems, vol. 12, no. 1, (**1998**), pp. 101-108.

Authors

Jing Huo, Engineer, received the Master degree in control engineering from Qingdao University of Science & Technology, China in 2008. Her research interests include machine learning and intelligent algorithm.

Yuxiang Zhao, Engineer, received the Master degree in atmosphere physcics from Cold and Arid Regions Environmental and Engineering Research Institute, Chinese Academy of Sciences in 2008. His research interests is atmosphere sounding.