Evolutionary Intelligence (2019) 12:341–347
https://doi.org/10.1007/s12065-018-0182-0

SPECIAL ISSUE

Imbalanced data classification algorithm with support vector machine kernel extensions

Feng Wang[1] · Shaojiang Liu[3] · Weichuan Ni[3] · Zhiming Xu[1] · Zemin Qiu[1] · Zhiping Wan[1,2] · Zhihong Pan[1]

Received: 14 September 2018 / Accepted: 3 October 2018 / Published online: 10 October 2018

Abstract
Learning from the imbalanced data samples so as to achieve accurate classification is an important research content in data mining field. It is very difficult for classification algorithm to achieve a higher accuracy because the uneven distribution of data samples makes some categories have few samples. A imbalanced data classification algorithm of support vector machines (KE-SVM) is proposed in this article, this algorithm achieve the initial classification of data samples by training the maximum margin classification SVM model, and then obtaining a new kernel extension function. based on Chi square test and weight coefficient calculation, through training the samples again by the new vector machine with kernel function to improve the classification accuracy. Through the simulation experiments of real data sets of artificial data set, it shows that the proposed method has higher classification accuracy and faster convergence for the uneven distribution data.

Keywords Support vector machine · Classification algorithm · Imbalanced data · Kernel extension function

1 Introduction

Data classification is an important research direction in the field of data mining, classifying the new objects based on the observed similar characteristics is one of the typical tasks in the field of data mining. Especially when there is a rare class, we usually get a problem of establishing a accurate classification model [1, 2], among the different classes, the imbalance of this particular class contains the data samples uneven distribution problem. The accurate classification of imbalanced data sample contributes to many practical problems of many fields [3, 4], for example, it can diagnose the disease with rare symptoms in medicine and correctly classify the pictures of forest fire, marine oil spill zone and other sudden accident in satellite radar image [5–7]. With the classification problem of imbalanced data in data mining field is becoming more and more common, the method which can classify a few special samples correctly is began to receive more attention.

Gu et al. [8] proposed a new fast cross-disciplinary data classification algorithm based on multi-source, this algorithm combined with the CD dual algorithm on the basis of MSCC, building several source domain classifiers and guiding the data classification of target domain synthetically. This algorithm has high classification accuracy for large sample data set, and it also has certain adaptability for less samples data set but its accuracy is not high. Wang et al. [9] proposed a network data classification method based on probability generation model, this method took the edges in network as the observation variables, made the categories of unknown category nodes as the variables, and set up a probabilistic generation model to describe the network, to solve the model problem by using the Gibbs sampling process and get the category of unknown nodes by the value of latent variables. Shao et al. [10] proposed an imbalanced data classification Rumsfeld double weighted support vector machine algorithm, based on using the different training points, to build two proximal hyperplane. The introduction of the sampling strategy keep the data samples in proximity information and this classification algorithm overcomes the bias phenomenon of TWSVM algorithm for the

Feng Wang, Shaojiang Liu, Weichuan Ni, Zhiming Xu contribute equally to this paper.

✉ Zhihong Pan
zhpan@xhsysu.edu.cn

[1] Department of Information Science, Xinhua College of Sun Yat-Sen University, Guangzhou, China

[2] Sun Yat-Sen University, Guangzhou, China

[3] Department of Equipment and Laboratory Management, Xinhua College of Sun Yat-Sen University, Guangzhou, China

imbalanced data classification. But this algorithm has longer training time, and the convergence speed is slow. Peng et al. [11] put forward a double support vector machine for imbalanced data classification structure normalized projection, to seek two projection directions for the two classes by solving the two smaller quadratic programming problems (QPPs), made a kind of projection samples were completely separated to their sub space from other classes, and the experimental results demonstrated that this algorithm is better than the other algorithms in generalization performance, but in the small data samples there is a greater influence on the classification accuracy of the algorithm. Zhang et al. [12] proposed a random traversal over-sampling method of imbalanced data classification, to create and compose the samples to balance the different classes through the actual data random traversal, when the synthetic samples generated, random traversal over sampling-method can enlarge the minority class boundary and reduce the standard deviation of predicted results and actual results in a few samples classification, but it still has some requirement on the samples number of minority classes, otherwise it will cause some troubles in synthetic samples.

The proposed method modify the support vector machine by presenting a new kernel function, firstly after the initial training of standard support vector machine and then the new vector machine training again. Since the Chi square test is a better hypothesis testing methods, and the weighting factor generate a significant relationship for the optimal classification hyperplane, so in the article, it can use the Chi square test and the weighting coefficients to extend and transform the kernel functions, making the new vector machine has greater adaptability, and it also can play a better classification performance for the minority class.

2 Maximum interval classification SVM model

Support vector machine (Support Vector Machine, SVM) is a method for pattern classification and regression, and it is proposed by Vapnik and his co authors [13, 14]. Its basic idea is to use the kernel function to map the original data into a higher dimensional space, making the data separated into two classes of data linearly as possible. The given data set $\{(x_1,y_1),(x_2,y_2),\cdots(x_m,y_m)\}$ represents the binary classification problems, (x_i,y_i) represents a training sample, among them $x_i(i\in m)$ indicates a data sample, $y_i$stands for the results tab which is class for a data sample belong. The used results tab is $y_i=-1$, $y_i=1$, and we the defined the function interval is as follow [15]:

$$\hat{\gamma}=y_i(\omega x_i+b). \tag{1}$$

A maximum interval separating hyperplane can be expressed as:

$$\omega x_i+b=0. \tag{2}$$

Once obtained the optimal ω and b, you can get the support vector machine classification model, and its function could be expressed as [16]:

$$f(x)=\sum_{i=1}^{m}\alpha_i y_i\langle x\cdot x_i\rangle+b. \tag{3}$$

The non zero number α_i which is corresponding to the data sample represented a support vector.

For a high dimensional feature space, the inner product $\langle x\cdot x_i\rangle$ is replaced by a kernel function $K(x,x_i)$, and it is expressed as:

$$K(x,x_i)=(x\cdot x_i+c)^d. \tag{4}$$

The radial basis function (RBF) could map the original feature to infinite dimension, and it is expressed as [17]:

$$K(x,x_i)=\exp\left(-\frac{\|x-x_i\|^2}{2\sigma^2}\right). \tag{5}$$

For the application of SVM, there are generally two types of classification methods, the one is building a multistage classifier by the establishment and combination of a plurality of binary classifiers, another is to construct the multistage classifier directly. For the imbalanced data classification problem, although there are some technologies can control the impact of classification results caused by individual specific data sample, but the classification performance obtained is still not ideal.

3 The kernel function expansion method

The parameters of the kernel function allows the geometry operation of feature space to a certain point, but in order to obtain greater adaptability and meet the demand of the imbalanced data classification, and making a few classes samples can also effectively enlarge their border, so it needs to extend and transform the kernel function. The isogonal extension transformation of geometric space is defined as a function maps the space to a new space, and the angle between the curves is retained at the same time. The general Extension transformation form is expressed as:

$$\hat{K}(x,x_i)=\lambda(x)K(x,x_i)\lambda(x_i). \tag{6}$$

$\lambda(x)$ represents a chosen positive function:

$$\lambda(x)=\sum_{x_j\in m}^{m}e^{-k\|x-x_j\|^2}, \tag{7}$$

Wherein k represents a positive constant, and you can get a new kernel function $\hat{K}\left(x, x_i\right)$ after the expansion and transform. There is a kernel extension method which makes declined along with the mapping distance, and using the traditional SVM training to get prior knowledge, readjusting the initial kernel function to make two classes of data been effectively amplified in the separation process. And the function expression of $\lambda(x)$ is as follow:

$$\lambda(x) = e^{-k\left(\sum_{i=1}^{m} \alpha_i y_i \langle x \cdot x_i \rangle + b\right)^2}, \tag{8}$$

When $\sum\limits_{i=1}^{m} \alpha_i y_i \langle x \cdot x_i \rangle + b = 0$, $\lambda(x)$ reaches a maximum value at the boundary surface of space, and when $\sum\limits_{i=1}^{m} \alpha_i y_i \langle x \cdot x_i \rangle + b = \pm 1$, $\lambda(x)$ will be declined to e^{-k}.

4 Imbalanced data classification algorithm

For the imbalanced data classification algorithm proposed in this article, firstly we can use the standard support vector machine algorithm to get the approximate hyperplane, put forward a extension transformation kernel function and calculate its parameters by using the Chi square test and the weighting factor.

In Sect. 4.1, we can use the selected function $\lambda(x)$ to get an extension transformation kernel function. In Sect. 4.2, we calculate the parameters by Chi square test and weight coefficient, obtained the best classification results of kernel function. And then there are flow diagram and pseudo-code for describing the algorithm in Sect. 4.2.

4.1 kernel function selection

For Vgiven imbalanced data sets, firstly we carry out the formula (2) and the formula (3) to get the approximate boundary position of data samples.

According to the predicted results after the classification we marked the data sample as $W^1, W^2, \cdots, W^V$. According to the formula (6), the effect of extension transformation depends on function, in order to make the two classes of data been effectively amplified in the separation process, The selected $\lambda(x)$ uses the kernel extension method from formula (8):

$$\lambda(x) = \begin{cases} e^{-k_1\left(\sum\limits_{i=1}^{m} \alpha_i y_i \langle x \cdot x_i \rangle + b\right)^2}, \ if \ x \in W^1 \\ e^{-k_2\left(\sum\limits_{i=1}^{m} \alpha_i y_i \langle x \cdot x_i \rangle + b\right)^2}, \ if \ x \in W^2 \\ \vdots \\ e^{-k_V\left(\sum\limits_{i=1}^{m} \alpha_i y_i \langle x \cdot x_i \rangle + b\right)^2}, \ if \ x \in W^V \end{cases}, \tag{9}$$

wherein k_i is the parameter calculated by Chi square test. We use W^i represent the ith sample set after the initial training.

4.2 The parameter based on the calculation of chi square test and weight coefficient

Chi square test (χ^2 test) is one of the statistical hypothesis testing approach. Testing the sampling distribution of the statistic is Chi square distribution or not when the null hypothesis was established. For one or more categories, Chi square test was used to determine whether there is a big difference between the expected frequency and the observed frequency. The formula of χ^2 is:

$$\chi^2 = \sum \left[\frac{\left(f_0 - f_c\right)^2}{f_c} \right], \tag{10}$$

f_0 represents the observed frequency in each category, f_c stands for expected frequency in each category. As can be seen from the expression of the formula, χ^2 is the sum that the square difference between observed data and predicted data dividing by predicted data. Due to the classification results of k_i and the data sample have inevitable connection, and we obtained the parameters k_i by Chi square test after getting the weight coefficient:

For the imbalanced data classification problem, taking weighted method to set up the proper weight coefficient ω is a key problem, the optimal weight has an important link on the generation of optimal classification hyper plane, in the vector machine, the weights is depended on the amount of training data ratio,, so the weights must satisfy two conditions, firstly the majority class of the data sample weights will be lower than the minority class of data sample weights, then the weights should meet the condition $\omega \in (0, 1]$.

In the proposed KE-SVM algorithm, we compensate uneven data distribution by setting the weight factor for the major class imbalanced data distribution problem. Assuming that the size of training sample is M, and class number is V, $L_i (i \in V)$ represent the sample size of each class, the weighting coefficient is defined as:

$$\omega_i = \frac{M}{L_i} \left(\frac{1}{\sum\limits_{i=1}^{V} \frac{M}{L_i}} \right), \tag{11}$$

Wherein the weight factor meet the condition $\sum_{i=1}^{V} \omega_i = 1$, the sparse distribution degree of each category can been seen by ω_i, and each kind of data sample is less, the weighted coefficient is bigger.

After obtaining the weighting coefficient ω_i, then we can calculate the parameters k_i by combining the Chi square test. When each class of data samples is in the optimal distribution, the Chi square value of χ^2 it is:

$$\chi^2 = \sum_{i=1}^{V} \left(L_i - \frac{M}{V}\right)^2 \cdot \left(\frac{V}{M}\right). \tag{12}$$

In order to get the optimal parameters $k_i$of the each class, we make $Y_i = \left(L_i - \frac{M}{V}\right)^2 \cdot \left(\frac{V}{M}\right)$, the calculation formula for parameter k_i is:

$$k_i = \omega_i \frac{Y_i}{\chi^2} = \frac{MY_i}{\sum_{i=1}^{V} Y_i L_i} \left(\frac{1}{\sum_{i=1}^{V} \frac{M}{L_i}} \right). \tag{13}$$

4.3 The flow chart of the algorithm and pseudo code

For each support vector based on the Chi square distribution, the algorithm perform the initial operation by standard SVM model firstly in each iteration, calculating ω_i and k_i, and then obtain the kernel spread function $\hat{K}\left(x, x_i\right)$, finally training the data samples through the new vector machine (KE-SVM) with $\hat{K}\left(x, x_i\right)$. Flow chart of the algorithm is as follows (See Fig. 1).

The algorithm pseudo code is as follows (See Fig. 2).

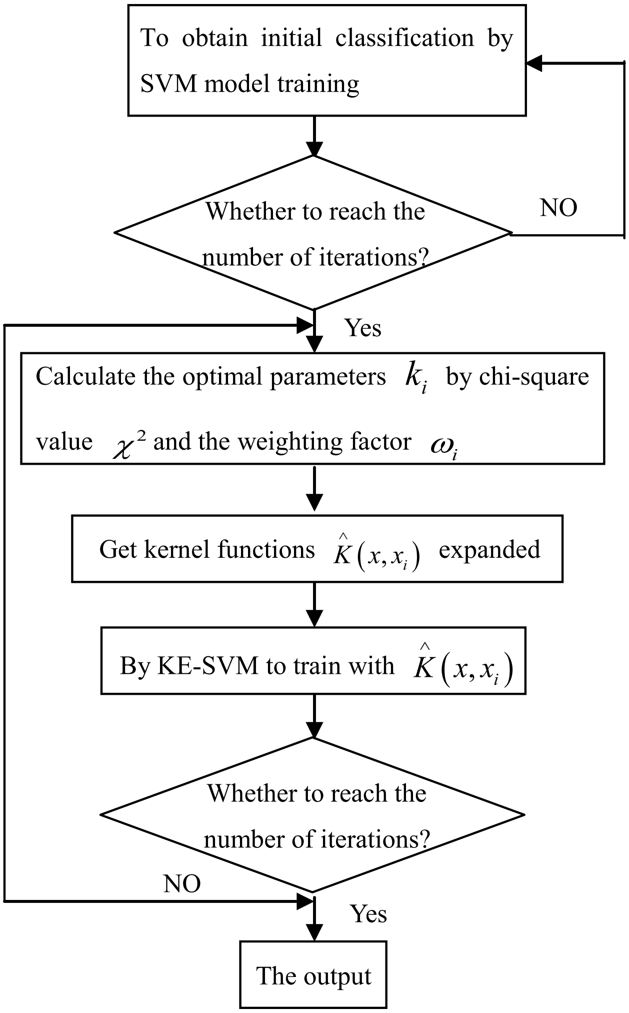

To obtain initial classification by SVM model training
Whether to reach the number of iterations?
NO
Yes
Calculate the optimal parameters k_i by chi-square value χ^2 and the weighting factor ω_i
Get kernel functions $\hat{K}(x, x_i)$ expanded
By KE-SVM to train with $\hat{K}(x, x_i)$
Whether to reach the number of iterations?
NO
Yes
The output

Fig. 1 Flow chart of algorithm

5 The experiment results

In this section, the capability of the proposed classification algorithm will be evaluated and compared with other classification algorithm. There is the artificial data in first experimental group and the text type of real data sets in second experimental group, the experiment mainly prove that the algorithm has classification accuracy for imbalanced data and fast convergence for sample data.

5.1 Performance metric of imbalanced data classification algorithm

Assume that the data samples can be divided into the positive class and the negative class, there is most of samples belong to the positive class, and few samples belong to negative class. The following representation shows Vector machine successfully predicted positive class data and negative class data and prediction error in the classification process:

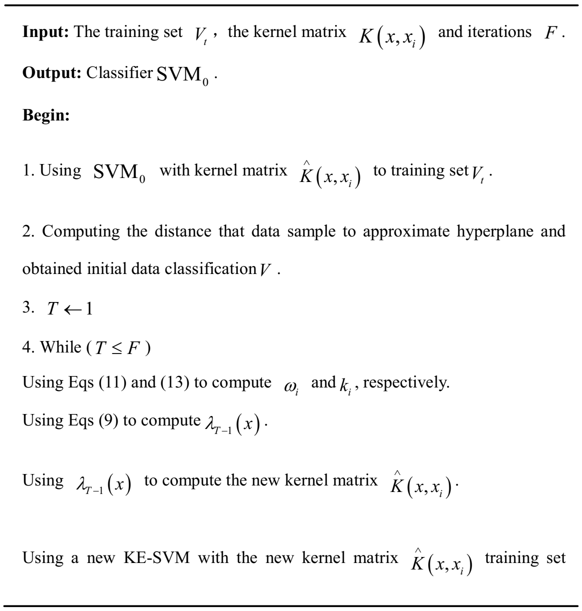
According to Table 1, the expression formula of algorithm classification accuracy is:

$$P_z = \frac{\mathrm{M}_p + \mathrm{M}_n{'}}{\left(\mathrm{M}_p + \mathrm{M}_P{'}\right) + \left(\mathrm{M}_n + \mathrm{M}_n{'}\right)}, \tag{14}$$

When value is greater it illustrate that the algorithm classification result is more accurate. Aiming at the imbalanced data classification problem, the overall classification accuracy of algorithm is not the only assessment standard. Thus in the experiment we also used the F-measure method and the G-mean method to evaluate the performance of imbalanced classification algorithm, these two ways are able to properly test the sensitivity and specificity of imbalanced data classification algorithm. The sensitivity is reflect the performance of the algorithm in predicting the positive class, and the specificity is reflect the performance of algorithms in predicting negative class. The

Fig. 2 Pseudo code of the algorithm

The algorithm pseudo code is as follows:



```
Input: The training set $V_t$, the kernel matrix $K(x,x_i)$ and iterations $F$.
Output: Classifier $\mathrm{SVM}_0$.
Begin:
1. Using $\mathrm{SVM}_0$ with kernel matrix $\hat{K}(x,x_i)$ to training set $V_t$.
2. Computing the distance that data sample to approximate hyperplane and
   obtained initial data classification $V$.
3. $T \leftarrow 1$
4. While ( $T \le F$ )
   Using Eqs (11) and (13) to compute $\omega_i$ and $k_i$, respectively.
   Using Eqs (9) to compute $\lambda_{T-1}(x)$.
   Using $\lambda_{T-1}(x)$ to compute the new kernel matrix $\hat{K}(x,x_i)$.
   Using a new KE-SVM with the new kernel matrix $\hat{K}(x,x_i)$ training set
```

Table 1 The forecast result and the actual result of the algorithm

	The number of samples that predicted to be positive class	The number of samples that predicted to be negative class
Really is positive class	M_p	$\mathrm{M}_P{}'$
Really is negative class	M_n	$\mathrm{M}_n{}'$

Table 2 Artificial selection data set

Data sets	Samples	Attributes	classes
Wdbc	569	7	4
Heart_c	303	4	5
Sonar	208	3	3
Dis	972	9	5

F-measure method is often used to test the information retrieval and the measurement of machine learning field, the performance of document classification and query classification. The calculation formula is:

$$F = \frac{2P_{Se}P_D}{P_{Se} + P_D} = \left(\frac{2}{\frac{1}{P_D} + \frac{1}{P_{Se}}} \right). \tag{15}$$

P_{Se} reflects the sensitivity of classification algorithm, and it is expressed as:

$$P_{Se} = \frac{\mathrm{M}_p}{\mathrm{M}_p + \mathrm{M}_P{}'}. \tag{16}$$

P_D represents the ratio of prediction accuracy of the majority class accounted for:

$$P_D = \frac{\mathrm{M}_p}{\mathrm{M}_p + \mathrm{M}_n}. \tag{17}$$

The G-mean method provides a fairer comparison method on the algorithm performance in the classification of majority class and minority class, and it is not limited to the number of samples. The computational formula is:

$$G = \left(P_{Se} \cdot P_{pe}\right)^{\frac{1}{2}}. \tag{18}$$

P_{pe} represents the specificity of classification algorithm, the representation formula is:

$$P_{pe} = \frac{\mathrm{M}_n{}'}{\mathrm{M}_n + \mathrm{M}_n{}'}. \tag{19}$$

5.2 Simulation results and analysis

The sample size, attributes and class number of the first group of data sets are as shown in Table 2.

The sample size, attributes, class number, the number of each class samples, and imbalanced ratio of second group are as shown in Table 3.

The control groups of experiment are the imbalanced data classification Lagrangian dual weighted support vector machine algorithm proposed by Shao et al. in document

Table 3 Real data sets

Data sets	Thyroid	Glass	lymph	leukocyte
Samples	367	245	232	314
Attributes	3	5	8	6
Classes	2	3	6	7
Samples of each class	223, 144	146, 66, 33	89, 42, 23, 8, 37, 33	178, 77, 2, 3, 17, 21, 16
Imbalance ratio	4.65	4.65	8.12	67.45

[10] and normal structure projection double support vector machine algorithm of imbalanced data classification proposed by Peng et al. in document [11]. In experimental process, Matlab7.1 is used to the programming simulation, and the iteration number is 100 times. All the algorithms perform the simulation by using the same data set and working in PC machine Window platform. In Table 4, there is the algorithm overall classification accuracy, sensitivity value, F-measure value and value of G-mean after simulation.

As can be seen from the graph, whether artificial selection data set and a real data set, the overall classification accuracy P_z of the proposed algorithm is higher than the algorithm of Refs. [10, 11]. It can be seen from formula (15), When value P_D is stay unchanged, value F will be grow by the increase of value P_{Se}, and when value P_{Se} is stay unchained, the bigger value P_D, the bigger value F. As can be seen from formula (17) the bigger value P_D, the smaller value M_n will be, and the lower prediction accuracy on minority class. Look at the results in Table 4, value P_{Se} of this algorithm is higher than the reference [10, 11], indicating that value P_D of this algorithm is smaller than literature [10, 11], and value M_n is higher, the prediction accuracy of the minority class is higher than the method proposed in reference [10, 11]. The greater G-mean value is represented the better comprehensive performance of the algorithm in the majority and minority class classification. So from the comparison of the value in Table 4 we could verify the performance advantages of algorithm plays on the data sample classification.

Figure 3 shows the change in algorithm classification accuracy of the KE-SVM algorithm proposed in this paper and other two kinds of algorithms with the running time. As can be seen from the graph, the classification accuracy will be gradually promoted in the operation process and then reach a stable value, this is because the vector machine requires a certain amount of time in training data samples, with the growing of iterations of algorithm, the classification accuracy will be improved. From the distribution curve in the figure, the proposed KE-SVM algorithm takes less time in achieving a stable classification accuracy, that it has a faster convergence in classification of data samples.

6 Conclusions

In this article, we propose a vector machine kernel extensions support imbalanced data classification algorithm by the study of imbalanced data classification problem. In this algorithm, maximum margin classification of SVM mode is regarded as the prototype to obtain the approximate hyperplane, and then put forward a expansion transformation kernel function, calculating its parameters by using Chi square test and weight coefficient, finally using new vector machine (SVM) model to carry on the training of data samples. In the experiment, we evaluate the performance of imbalanced classification algorithm by overall classification accuracy, F-measure method and G-mean method. From the experimental contrast it can be seen that the algorithm has played out the better data classification performance.

Table 4 The comparison after algorithm simulation

	Proposed method (%)	Shao' method (%)	Peng' method (%)
Artificial datasets			
Overall classification accuracy (P_z)	95.82	90.45	91.76
Sensitivity (P_{Se})	95.31	89.10	91.34
F-measure (F)	68.10	68.77	73.12
G-mean (G)	95.10	90.14	90.72
Real datasets			
Overall classification accuracy (P_z)	95.67	90.03	90.66
Sensitivity (P_{Se})	95.91	89.87	91.15
F-measure (F)	68.03	69.78	73.64
G-mean (G)	95.45	89.79	90.45

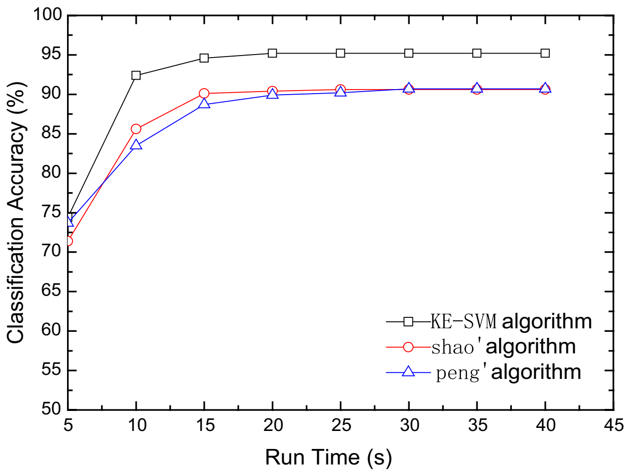


Fig. 3 The change in algorithm classification accuracy over time

References

1. Buczak AL, Guven E (2017) A survey of data mining and machine learning methods for cyber security intrusion detection. IEEE Commun Surv Tutor 18(2):1153–1176
2. Papalexakis EE, Faloutsos C, Sidiropoulos ND (2016) Tensors for data mining and data fusion: models, applications, and scalable algorithms. Acm Trans Intell Syst Technol 8(2):1–44
3. Adeniyi DA, Wei Z, Yongquan Y (2016) Automated web usage data mining and recommendation system using K-nearest neighbor (KNN) classification method. Appl Comput Inf 12(1):90–108
4. Deng Y, Ren Z, Kong Y et al (2017) A hierarchical fused fuzzy deep neural network for data classification. IEEE Trans Fuzzy Syst 25(4):1006–1012
5. Gu Y, Wang Q, Xie B (2017) Multiple kernel sparse representation for airborne LiDAR data classification. IEEE Trans Geosci Remote Sens 55(99):1–21
6. Pourpanah F, Lim CP, Saleh JM (2016) A hybrid model of fuzzy ARTMAP and genetic algorithm for data classification and rule extraction. Expert Syst Appl 49:74–85
7. Zhang J, Wang S, Chen L et al (2017) Multiple Bayesian discriminant functions for high-dimensional massive data classification. Data Min Knowl Discov 31(2):1–37
8. Gu X, Wang S-T, Xu M (2014) A new cross-multidomain classification algorithm and its fast version for large datasets. Acta Autom Sin 40(3):531–547
9. Wang Z-W, Xiao W-D, Tan W-T (2013) Classification in networked data based on the probability generative mode. J Comput Res Dev 50(12):2642–2650
10. Shao YH, Chen WJ, Zhang JJ et al (2014) “An efficient weighted Lagrangian twin support vector machine for imbalanced data classification. Pattern Recognit 47(9):3158–3167
11. Peng X, Xu D (2014) “Structural regularized projection twin support vector machine for data classification. Inf Sci 279(279):416–432
12. Zhang H, Li M (2014) RWO-sampling: a random walk over-sampling approach to imbalanced data classification. Inf Fusion 20(1):99–116
13. Yin Y, Xu D, Wang X et al (2017) Online state-based structured SVM combined with incremental PCA for robust visual tracking. IEEE Trans Cybern 45(9):1988–2000
14. He H, Kong F, Tan J (2017) DietCam: multi-view food recognition using a multi-kernel SVM. IEEE J Biomed Health Inf 20(3):848–855
15. Yoon H, Park CS, Kim JS et al (2013) Algorithm learning based neural network integrating feature selection and classification. Expert Syst Appl 40(1):231–241
16. Chen Y, Nasrabadi NM, Tran TD (2013) Hyperspectral image classification via kernel sparse representation. IEEE Trans Geosci Remote Sens 51(1):217–231
17. Zhun M, Li X-L, Li X-L (2012) A two-stage support vector machine algorithm based on meta learning and stacking generalization. Pattern Recognit Artif Intell 25:943–949