Simultaneous Feature with Support Vector Selection and Parameters Optimization Using GA-Based SVM Solve the Binary Classification

Ye Fei

School of Information Science and Technology
Southwest Jiaotong University
Chengdu, Sichuan, China
e-mail: 122404504@qq.com

Han Min

School of Software
Chengdu Information Engineering University
Chengdu, Sichuan, China
e-mail: hanmin@cuit.edu.cn

Abstract—**Feature selection and parameters optimization is an important step in the using of SVM. In recent years, more researchers are mainly focus in feature selection and parameters optimization. However, the number of support vectors with the selected support vector subset also has an effect on classification performance of SVM. Few researchers concentrate on this area. This paper proposed a novel optimization approach which aim to select the support vector subset and feature subset simultaneously based on genetic algorithms, in optimization of while, also constantly to search the best penalty parameter C and kernel function parameters. We conduct the experiments on real-world dataset from the openly UCI Machine Learning Repository using the proposed approach and the GA-based FS technology. The experimental results show that the proposed approach can efficiently choose the optimal input features with SVM parameters and also achieve the best classification performance. Moreover, it turns out that the proposed optimization method generates a less complex SVM model with fewer support vectors.**

Keywords-support vector machine; kernel function; binary classification; genetic algorithm

I. INTRODUCTION

Classification and prediction are two of the most important tasks in data mining, the binary classification problems in many of these problems are the most common and the most typical problems. Binary classification has already been applied in many areas. For example in the field of sentiment analysis, classifying the texts into the positive category or the negative category belongs to binary classification problem. In the field of knowledge graph, predicating the triples' classification also can be viewed as a binary classification problem. An early work is that Panget [1] used a standard Word bags and a variety of machine learning methods, which is used to solve the sentiment classification problem, and the experimental results show that in the supervised learning method, SVM and Native Bayes have a high accuracy rate. This result also demonstrates that SVM has an excellent performance in dealing with binary classification problem. Bao Y [2] use a different kernel function to create a different classifier to the binary comparison experiment. Experimental results show that the RBF kernel has the best performance.

Support vector machine in 1995 quickly became the new study field of machine learning. It is the biggest difference from neural networks. Because the support vector machine is based on statistical principles, and seek structural risk minimization. Neural network are seeking empirical structural risk minimization. In recent years, support vector machine has gotten a lot of attention, and the main reason is that it has a mature mathematical theory and computational advantages. The SVM able to classify different types of data is mainly through the support vectors of training sets in the feature vectors space to construct a hyper-plane. On one side of the hyper-plane y=1, on other side of the hyper-plane y=1.

When optimizing SVM classifier, there have two problems must be solved. The first is to choose the optimal feature subset for SVM classifier. The second is to adjust the penalty parameter C and kernel parameters. These two problems are playing an important role in the optimization of SVM classifier. Furthermore, choice of kernel function will have a big impact on the performance of the classifier. Therefore, how to choose appropriate kernel functions for different classification tasks and parameters adjustment has been the focus of research. There have no relevant theory to guide the selection of kernel functions and parameters adjustment. Hence, simultaneous to select the optimal feature subset and to optimize the kernel parameters are crucial.

Support vector machine used in the training process creates a discriminant function that minimize the training mistake with maximize the interval of hyper-plane. The maximization of the interval of hyper-plane is a optimization regularized process which can generates a light model with few parameters (Support vectors). Support vectors are parts of the training set which were used as decision boundary. Therefore, the number of support vectors does not rely on the feature dimension of the input space. This preponderance of SVM can be avoided dimension disaster. Furthermore, reducing the number of support vectors and the obtaining the optimal support subset are available in the optimization of SVM classifier.

In addition to the selection of support vector subset and feature subset, proper parameters setting of the SVM can improve the classification performance of SVM. Generally we need to optimize the parameters that include the penalty parameter C and the kernel function parameters. For example, radial basis function (RBF) need to optimize the parameter G. When we use the SVM classifier, we need to choose a kernel function, to set the kernel parameters and to select a best soft

margin constant C. To do this work, the gird algorithm is an alternative to find the best parameter C and kernel parameters. However, this approach usually needs the high large computational overhead and does not simultaneously perform support vector subset selection. Therefore, this approach cannot be effective in multiple parameters optimization problems. Genetic algorithm is a typical evolutionary algorithm. In recent years, genetic algorithm has been greatly promoted due to it can be used to encode the solution of problems, thereby reducing the complexity of problems. The proposed approach performs the optimization process in an evolutionary way. In the previous literature, only a few algorithms have been proposed for SVM support vector selection [12, 13]. There have a more algorithms have been proposed for feature subset selection and optimization of SVM parameters [14, 15, 16, 17]. However, these algorithms only focused on feature selection or parameters optimization for SVM classifier and did not consider that the number of support vectors and support vectors selection.

This paper first proposed simultaneously the support vector selection, feature selection, and SVM parameters optimization scheme. Experimental results show that it can be feasibility in binary classification tasks. Then, the proposed approach use an improved algorithm presented in this paper based on the standard genetic algorithm. It is used to cope with multi-parameter optimization problem of the proposed approach. The experiment results show that the proposed approach has better classification performance than GA-based feature selection technology.

This paper is organized as follows. In Section 2 we review some basic work about support vector machine classifiers and Genetic algorithms. In Section 3 we describe the proposed method for support vectors with feature subset selection and parameters optimization for SVM classifier. In Section 4 we present the experimental results using the proposed approach with GA-based feature selection technology on the real-world datasets. A summary of this paper are found in section 5 where we draw a main conclusion and address future developments.

II. RELATED WPRK

A. Support Vector Machines

Let's define a data set $\{x_i, y_i\}$ i=1,2,..n To represent the samples. x_i represents an n-dimensional vector, y_i represents the class of x_i. In the binary classification, $y_i \in \{-1,1\}$, i=1,2,3...n. According to the classification function $f(x) = w^T x + b$. If $f(x) = 0$, then x is the point located on the Super square meters. If $f(x) < 0$ then representation corresponding to y = -1, If $f(x) > 0$ indicates that y = 1. as shown in the figure, the SVM principle is to find a hyper-plane that maximize space. This interval is positive and negative categories from hyper-plane closest to the border point. So the problem can be transformed to minimize $\|w\|^2 / 2$. By Lagrange's come up with the following formula:

$$L(w, b, a) = \frac{1}{2} \|w\|^2 - \sum_{i=1}^{n} a_i(y_i(w^T x_i + b) - 1)$$

$$\text{s.t } a_i \geq 0, i = 1,2, ., n. \qquad (1)$$

When faced with the nonlinear data, require the kernel function was applied. It is based on the low-dimensional input vector mapped to higher-dimensional feature space. Several classical kernel functions are shown in the table below.

TABLE I. THE EXPRESSION OF SERVERAL CLASS KERNELS

Name	calculate
Linear	$K(x, y) = x \times y$
Sigmoid	$K(x, y) = (ax \times y + b)^d$
RBF	$K(x, y) = \exp(-\mid x - y \mid^2) / \delta^2$
Polynomi	$K(x, y) = a(\exp(\frac{r^2}{\mid x - y \mid^2 + \delta^2}) - 1)$

Choose a suitable kernel function is an important step of SVM optimization. The using sigmaid as a kernel function of SVM model has been proved to the equivalent of two-layer feed forward neural networks [8]. In recent years, the using Polynomial kernel and RBF kernel is most popular. RBF kernel is used most frequently. Because the low-dimensional vectors is can mapped to a higher dimensional spaces. And it contains only one parameter. And other kernel functions often contain 2 or more parameters. Seen from the RBF image that parameter controls the width of the image, if the parameter is smaller, the Support vector will gather near the vertex of the image, you will get a complex decision boundary. If the parameter is large, you will get a smooth decision boundary. Both situations are adversely affected to prediction performance of SVM. Therefore appropriate setting of parameter C and kernel parameters is very important.

B. Genetic Algorithm

Genetic algorithms is a powerful optimization tool, it is usually looking for optimal solution as a goal. Genetic algorithms have been used in dealing with real-world problems. In some cases, the genetic algorithm compared to traditional optimization algorithms have better search efficiency, and easier to find the global optimal solution. Genetic algorithms are initially through the observation of natural systems evolved and developed by genetic mechanisms. Its basic principle is to maintain the population representing solutions, and obtained better populations through continuous evolution of the species. Because genetic algorithm can used a string expressed for the problem, so reduce the complexity of the problem, and through the

crossover, duplicate, mutation operation can easily change the solution expression. In recent years, genetic algorithms is usually combination other algorithm or model to solve the problems that need to be optimized. [5] De et al presents a new model of combination of genetic algorithm and neural network (NEA).The experiment show that it has high accuracy in classification tasks than traditional neural network algorithm and required less time. Cantarella [6] use genetic algorithm to solve the optimization problem of signal nodes in signal setting. Wang Y [7] use genetic algorithm to optimize the BP fuzzy neural network which apply to the elevator group control system.

III. CREATION OF OPTIMIZATION MODEL

This section focuses on the chromosome design, the fitness function design, the genetic algorithm optimization and the model creation process.
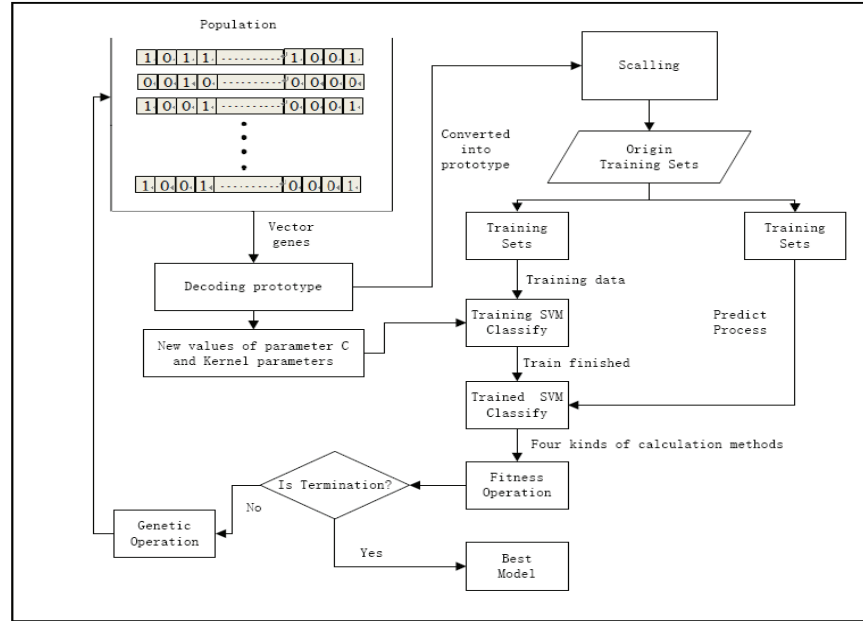

Figure 1.   The proposed model framework.

A. Chromosome Design

To encode the solution of the problem is the first step of genetic algorithm. In this paper, we are mainly using RBF Kernel as a kernel function. Because low-dimensional data can through RBF kernel function mapped to higher-dimensional feature space, and it has only a single parameter G. And other kernel functions have two or more parameters, which mean that adds complexity to the SVM parameters optimization. Several experiment results also show that the SVM to select RBF kernel function better than other kernel functions to enhance the performance of SVM. When the RBF kernel as a kernel function is used, we need to optimize penalty parameter C and kernel parameter G, as well as to select feature subset and support vector subset. Therefore, chromosome is mainly consists of four parts, respectively for the penalty parameter C and kernel parameter G, feature mask, support vectors binary strings. In this paper, we use binary strings representing chromosomes. Of course, you can also use a decimal or floating-point encoding. Figure 2 shows a chromosome encoding design.
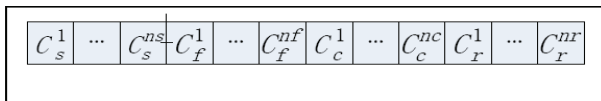

Figure 2.   Chromosome design.

where $c_f^1 - c_f^{n_f}$ represents the vector of feature mask. n_f is the length of feature mask. Where $c_i^1 - c_i^{n_i}$ represent the vector of the input variables. n_i is the length of input variables. Where $c_c^1 - c_c^{n_c}$ represent the value of parameter C and n_c is the representing calculation precision of parameter C. Where $c_g^1 - c_g^{n_g}$ represent the value of kernel parameter G and n_g is the representing calculation precision of kernel parameter G. For binary strings representing the genotype of feature mask, the bit with value "1" represents the feature is selected, and "0" indicates feature is ignored. For binary strings representing the genotype of input variables, the bit with value "1" represents the input variables is learning set in the training process, and "0" indicates the input variables is validation set in the testing process. The binary strings representing the value of parameter C and kernel parameter G must be converted to a decimal value. Note that determining the calculation precision of the decimal value depends on the length of the binary string. The equation (2) is used to calculating the binary strings of the parameter C and the kernel parameter G.

For parameter c and g represented the chromosomes string convert into the prototype. We need to design the

428

transform formula 2:

$$C = Min_c + \frac{D_c}{2^{lc} - 1} \times \left(Max_c - Min_c\right) \quad (2)$$

The C represents the value of binary strings representing parameter. The Min_c and Max_c respectively is representative of the minimum and maximum value of the variable. It is usually setting by the user. The lc value is length of chromosome of the parameter C, and D_c represents a decimal number of the binary strings representing parameter C. The obviously $D_c / (2^{lc} - 1)$ values are between 0 and 1, then adjust the values only depend on the bits of binary strings representing parameter C. Therefore, we using the formula convert chromosome to a real number.

B. Fitness Function Design

Fitness is as probability measure for the individual selection and crossover operation. So well designed fitness function can find a better individual, in this paper, we will mainly consider the four indicators to design a fitness function, They are forecasting accuracy, The difference value between the positive accuracy and negative accuracy, the capacity of feature set and the number of support vectors. To solve the multiple criteria problem by constructing a single objective fitness function that combines the four indicators into a unitized fitness function is used to represent the individual fitness. The detailed of indicators is as the following table:

TABLE II. FITNESS VALUE CALCULATE

Name	Calculation
fitnessA	$f1 = Total.accuracy$
fitnessB	$f2 = 1 - abs(Positive.accuracy - Negative.accuracy)$
fitnessC	$f3 = 1 - \left(\sum_{i-1}^{nf} F_i \times C_i\right) / nf$
fitnessD	$f4 = 1 - \left(\sum_{i=1}^{ns} G_i \times S_i\right) / ns$

FitnessA and FitnessB are the internal fitness function, where F1 is the value of forecast accuracy of the individual. F2 is to first calculate the different value between positive accuracy and negativity accuracy, and followed by 1 minus the absolute the difference value. Benefits of this calculation is that the value is always between 0 and 1, if the different value is too large, then F2 became smaller. If the different value is too small, then F2 became bigger. It is also to meet the demand of the difference value as small as possible. F3 is to first calculate the value of loss function of all selected

features. The C_i represents the loss value of i-th feature, here we set it value of 1. The $F_i = 1$ represents i-th feature are selected. If $F_i = 0$, it illustrate that i-th feature are ignored. When the loss value is too large, then F3 value is smaller, indicating more selected features, and the value of F3 are larger, indicating less number of features is selected. F4 calculation is the same as r3, G_i represents i-th loss function of support vector, S_i represents I-th selected support vector as a training set. Four fitness calculation formula values are between 0 and 1. Finally using a unified variable was obtained through the multiplied by four kinds of fitness values and they weights. Formula is as follows.

$$fitness = w1 \times f1 + w2 \times f2 + w3 \times f3 + w4 \times f4 \quad (3)$$

C. Genetic Algorithm Optimization

A chromosome are consists of four parts. It in the processing of the crossover and mutation may lead to large difference between the individuals representing prototype. In the experiment, we also found that standard genetic algorithms in solving multi-parameter optimization of our model can easily lead to obtained local solutions and the evolution process can easily becoming a random search process. Such method is not conducive to the search for the best solution.

When a point of the cross operator appears in the part of support vector of chromosome, by comparing the new individual and it parent body, we can see that a new individual simultaneously changing three parameters. It is obviously not the result we wanted. Mutate operation have the same problem. Therefore, we needed to improve the standard genetic algorithm for model in order to meet the requirements of multi-parameter optimization model. The detailed explanation of improved algorithm is as follows:

Crossover: First, using the roulette wheel method chose one of four parts of the chromosome. The proportion of every part can be determined by the user according to the importance of each part. The proportions of each part can be assigned based on experimental observations. First generate a random number, which falls into the part as a string of crossover, and follow the single point cross way to produce a new string and insert it back to the originally parent body.

Mutation: Its procedure is the same as crossover operation, the first to select chromosome subset by roulette method, and then performs the mutation operation.

Selection method: because the proposed approach needs multi-parameter optimization, after genetic manipulation cannot guarantee to create the best individual into next generation. Therefore, in each iteration stage, to select the best individual directly access to the next generation without performed crossover and mutation.

D. Construction of the Proposed Model

This section describes the construction of the proposed model. The construction of the model is starting with of 3 major aspects, which are the support vector selection, the

feature subset selection and parameters optimization. In the process of construction, we can divide the stage into internal computing layer and external computing layer. The internal computing layer is mainly using SVM for the training and forecasting on the binary strings representing prototype. External computing layer is mainly genetic algorithm operation of the chromosomes. The following main steps is shown in figure1 and the detailed explanation of procedures is as follows:

(1) Data processing: The values that all features represented scale between 0 and 1. Because if the value of the feature is larger than the value of other features. Then it may leads to an unbalanced feature set, and does not facilitate the calculation. Each feature value is usually within a fixed range of values. In this model we set the feature value between 0 and 1, and see the following formula 4. The Min_f represents the minimum value of all the features values. The Max_f represents the maximum value of all the features values. Through using the formula we can get the scale value.

$$f' = \frac{f - Min_f}{Max_f - Min_f} \qquad (4)$$

(2) Initialize groups: the first step is to initialize the group, the commonly way is the randomly initialized. We recognize that individuals can keep big differences during the initialization procedure. Therefore, It can search a wider solution space of the genetic algorithm. We need to establish several conditions for initialization operation. The value of parameter C is generally considering range from 10 to 1000. When the randomly generated multiple parameter C is difference between they values as large as possible.

In addition, we also need to take into account the distribution of the training set and validation set. We setting the initial number of training set range is the 60 and 80 of per cent of the capacity of training set. By generating a random number used to indicate the capacity of training set number, the rest of the samples as a validation set. To initialize the number of selected features, we set to generate a random number between the 70 and 80 of per cent of the total number of samples as the capacity of selected feature subset.

(3) Transformed to the prototype: this step is the binary code string converted into the corresponding prototype. The detailed explanation of transformation is in the section 3.2.

(4) Individual fitness calculation: Every chromosome represents the prototype contained the parameter C, the parameter G, the selected feature subset and the selected support vector subset. With support vector machine training and testing, you can get the values of these indicators. The fitness value of each individual is obtained through the fitness formula. The specific formulas are seen in table 2.

(5) Objective function: the objective function as a terminate conditions of genetic algorithms. It stops when generally found the optimal solution. But since the model's optimal solution cannot be determined. Therefore, we set the objective function of the stop condition is when a number of iterations has been reached a certain value. Furthermore, if generating fitness value of the individuals was not increase compared to the previous generation, then stop iteration.

IV. EXPERIMENT

A. Experiment Descriptions

To evaluate the classification performance of the proposed approach in binary classification tasks, we use the real-world datasets from the openly UCI database. The UCI database has been widely used as experiment data to measure the classifier performance of different classification tasks. Because of our study focus is mainly on the binary classification task. Therefore, we select the "Diabetic" data as the testing dataset. It contains 1151 instance and 20 attributes.

In the experiment, we use SVM as classifier, and the experimental code is extensions based on the Java version of the libsvm libraries. It is development and maintenance by Taiwan University Cheng and Lin. Our genetic algorithm is developed based on the standard genetic algorithm to suit the requirement of multi-parameter optimization of the proposed model.

When using UCI dataset, in face of the requirements of experiment need to harmonize data formats, thus the originally data formats is converted to a format that SVM classifier can recognize. For example, "1 2:1.3 3:25 ...", where 1 means that category, "3:0.2" means 3rd attribute value is 0.2. If it is a binary classification task, the category tag can take 1 and -1. After this step, we also need to standardize the data for 0-1. Because when a feature value is large than other feature value which can lead to an unbalanced dataset, and affecting the performance of the classifier.

In order to ensure that the testing results are not affected by other factors, we randomly assigned data set into k parts, each part is completely independent. Each part is as a test set, remaining k-1 part of dataset as the training set. We use k-fold cross-validation to experiment. The advantages of cross validation are that you can get more objective results. One of the most classic examples is the k=10, which dataset was divided into 10 equal parts, each of which will act as a test set, remaining 9 parts as a training set. This experiment sets the k=10.

Parameter settings: SVM model parameters setting include penalty parameter C, kernel parameter G, range of support vector count and range of feature count. For penalty parameter C, we set the value range between 0.1 and 1024, its calculation precision of 2^20 (20 bytes). We set the parameter G range between 0.1 and 10, its calculation precision of 2 ^20. We set the number of training support vectors range between 30% and 90% of the total number of samples, because too little or too much number of selected support vectors can lead to low performance of SVM classifier. We set the number of selected feature vectors range between 50% and 100% of the total number of the features.

B. Evaluation Prediction Results

Evaluation prediction results in the binary classification task can use the positive hit rate, the negative hit rate and total hit rate. For multi-classification tasks, assessments rely solely on average precision of the prediction results. The several statistical variables are shown in Table 3. The True Positive (TP) represents all positive samples being correctly split into positive categories. The False Negative (FN) represents all positive samples being incorrectly classified as a negative category. In contrast, The True Negative (TN) represents all negative samples being correctly split into negative categories. The False Positive (FP) represents all negative samples being incorrectly classified in the positive category. These statistics variables can be used as an important performance indicator of binary classification task.

TABLE III.   CONFUSION MATRIX IN BINADY CLASSIFICATION

Classification	Positive	Negative
Is True	True Positive(TP)	False Positive(FP)
Is False	False Negative(FN)	True Negative(TN)

The Positive hit rate and the Negative hit rate can be usually used to represent distinguish ability of classifiers between positive samples and negative samples, especially in the face of some unbalanced distribution of the test sample. The Positive hit rate shows the proportion of positive samples were correctly classified into the positive category, its formula is $P1=TP (TP + FN)$. The Negative hit rate that the proportion of negative samples were correctly classified into the negative category. Its formula is $P2 = TP / (TP + FN)$. The total hit rate represents the accuracy rate in all samples, its formula is $(TP + TN) / (TN + FP + FN + FP)$.

C. Experimental Results and Comparison

In the proposed approach, accuracy indicator weight of fitness function and other indicators weights will affect the classification performance. As shown in the Fig. 3, when accuracy indicator weight gradually increases, the model will have a higher degree of accuracy. When accuracy indicator weight gradually decreases, the number of the selected features increased. The numbers of support vectors are less stable and not affected by the accuracy weight adjustment. Therefore, we set the accurate weight of 0.7, the support vectors quantity weight is 0.5, the features quantity weight is 0.5, the weight of different value between the positive accuracy and negative accuracy is 0.5.
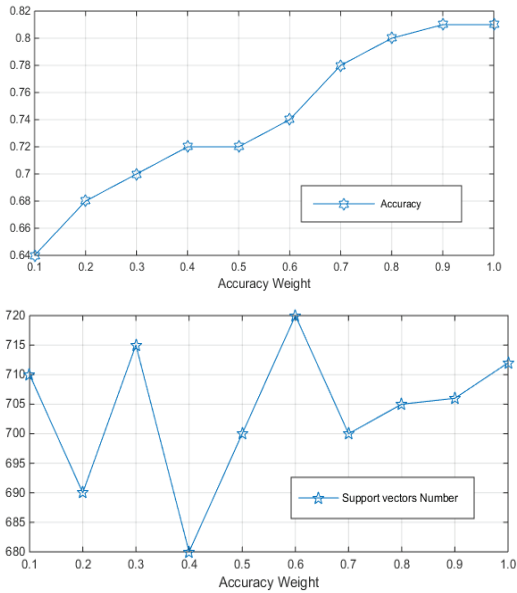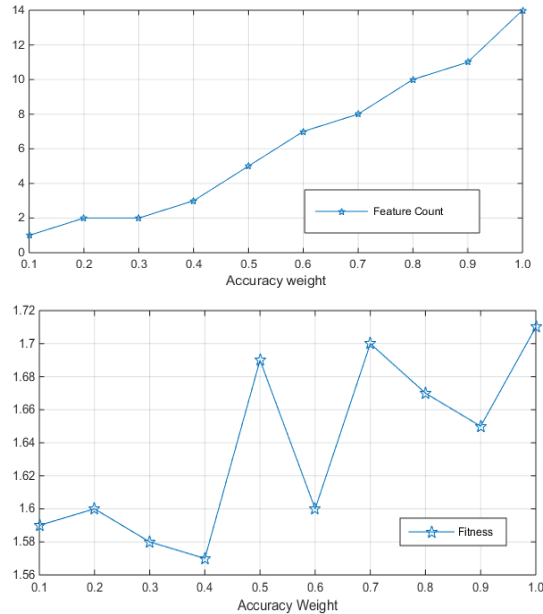

Figure 3.   Illustration of the classification accuracy versus the accuracy weight in fold #2 datase

To test diabetic dataset, we record the total accuracy, the positive hit rate, the negative hit rate, the number of selected features, the number of selected support vectors, the values of parameters C and G. Each fold using the proposed approach and the GA-based feature selection method perform training and prediction. Experiment results are displayed in Table 4 and Table 5. For the proposed approach, the average total hit rate is 0.7989. The average positive hit rate is 0.7464. The average negative hit rate is 0.7685. The number of selected support vectors is 987. The number of selected features is 12.5. For the GA-based feature selection method, the average total hit rate is 0.7355. The average positive hit rate is 0.7764. The average negative hit rate is 0.7125. The number of selected features is 13.

By comparing the prediction performance was pro-duced by the proposed model and the classical model in the binary classification task. We use the average total hit rate as measuring standard. Table 4 shows that the proposed models in 10 separate testing dataset have a higher accuracy than the classical model, and with less number of support vectors.

By comparing the prediction performance was produced by the proposed approach and the GA-based feature selection method in the binary classification task. We use the average total hit rate as measuring standard. Table 4 shows that the proposed approach in 10 separate testing dataset have a higher accuracy than the GA-based feature selection method, and with less number of support vectors.

Assess the distinguish ability between positive samples and negative samples. This paper is using Receiver operator characteristic (ROC) curve analysis as evaluation indicator. The ROC curve is used to evaluate classifier performance. In Figure 2, the x axis represents positive hit rate. The y axis represents 1-Negative hit rate. If ROC curve is closer to the top edge of the image and close to the left edge of the image, then the classifier has a higher performance. The AUC is also used be indicator to measure performance of the classifier. It is the integral of the ROC curve. In Figure 2, we show that ROC curve of two methods in the fold #2 testing. By selecting the number of breaking point generated a pair of positive hit rate and negative hit rate. We can see that the AUC values of the proposed approach and the GA-based feature selection method respectively are 0.81, 0.74. By calculating the mean AUC values for the 10 folds of testing dataset respectively are 0.8124 0.7276. The proposed approach can be improved prediction performance than the GA-based feature selection method

TABLE IV. EXPERIMENTAL RESULTS FOR BINARY CLASSIFICATION DATASET USING THE PROPOSED APPROACH AND C.ASSIC METHOD

| Fold#	The Proposed Model						The Classic Model					
	Total hit rate	Positive hit rate	Negative hit rate	Optimized C	Optimized G	Selected features	Total hit rate	Positive hit rate	Negative hit rate	Optimized C	Optimized G	Selected features
1	0.7739	0.8627	0.7031	226.4354	0.3681	16	0.7576	0.8627	0.6718	509.2337	0.2418	16
2	0.8152	0.8820	0.7492	249.6175	0.7518	12	0.7031	0.8235	0.7031	536.3040	0.4592	15
3	0.7843	0.7752	0.7931	766.3938	0.5584	13	0.7652	0.8627	0.6875	597.4171	0.8430	15
4	0.7913	0.8620	0.7192	249.6170	0.7500	12	0.7130	0.7647	0.6718	909.7031	0.1472	12
5	0.8320	0.8950	0.7862	818.6679	0.1359	15	0.7217	0.7647	0.6875	659.4121	0.4251	12
6	0.8172	0.8743	0.7685	279.9942	0.1524	12	0.7478	0.7647	0.7343	651.6701	0.1752	11
7	0.8072	0.8549	0.7963	386.9669	0.6425	11	0.7304	0.8235	0.6562	785.9299	0.6795	11
8	0.7954	0.8432	0.8064	232.9374	0.1832	12	0.7217	0.6078	0.8125	94.990	0.1479	11
9	0.8015	0.8324	0.7854	328.1938	0.7789	11	0.7562	0.7450	0.7656	657.5550	0.4793	13
10	0.7719	0.7819	0.7758	715.3472	0.1752	11	0.7391	0.7450	0.7343	335.8990	0.1824	14
Average	0.7989	0.7464	0.7685	N/A	N/A	12.5	0.7355	0.7764	0.7125	N/A	N/A	13
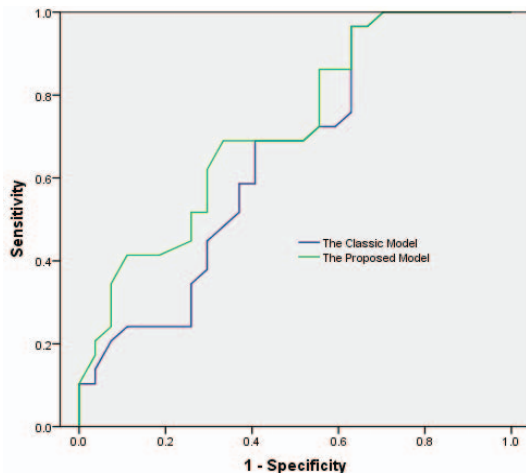
Figure 4. The ROC curve for fold #2.

V. CONCLUSION AND FUTURE WORK

Support vector subset with feature subset and SVM parameters were optimized simultaneously in the proposed approach due to the selected support subset with feature subset have motivate the kernel parameters adjustment for SVM classifier. In this paper, the number of support vectors and the optimal support vectors selection is mainly focus in our study. Because it can combine the other optimization tricks to offers the following advantages:

(1) A few number of support vectors can decrease SVM model's complexity which can enhance classification performance.

(2) A low-dimension feature vectors with a few support vectors requires less computational cost.

(3) A few support vectors with a low-dimension feature vectors can reduce the risk of overfitting of SVM classifier.

For the fitness function of genetic algorithm, The proposed approach not only considers the accuracy as the indicator of individual fitness and also to think through the other indicators that include the feature dimension, the number of support vectors and the balance of prediction results. Finally, four indicators of the individual fitness were together with theirs weight to form a unified fitness Function. To adjust weights of four fitness indicators in order to obtain the best parameters configuration in fitness function. Experimental results also demonstrate that different parameters configuration have an effect on classification performance of SVM classifier.

We presented an optimization approach for parameters optimization, support vector subset selection and feature subset selection and it is feasible in the binary classification task. This study constructed binary classification experiments to evaluate the classification performance of the proposed approach combined with the RBF kernel function and the GA-based feature selection method on the real-world datasets from openly UCI database. The experimental results demonstrate that the proposed approach has higher classification performance with lowest support vectors.

Results also show that although our research focus is on the binary classification tasks and obtained a good prediction performance, but also it can be used to on multiple classification tasks.

In this study we are only to select RBF kernel as kernel function of SVM in binary classification experiment, however, other kernel functions can also be applied to this proposed approach.

There have several directions can be done in future work: first, it would be interesting to extend the proposed approach which can apply to the multiple classification tasks, such as text classification. Second, it would be interesting to using the proposed approach with other kernel functions like liner kernel or polynomial kernel to construct several experiments to find which kernel function combined with the proposed approach that can obtain the best performance.

REFERENCES

[1] Pang B, Lee L, Vaithyanathan S. Thumbs up?: sentiment classification using machine learning techniques[C]//Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10. Association for Computational Linguistics, 2002: 79-86.

[2] Bao Y, Wang T, Qiu G. Research on Applicability of SVM Kernel Functions Used in Binary Classification[C]//Proceedings of International Conference on Computer Science and Information Technology. Springer India, 2014: 833-844.

[3] Goh K S, Chang E, Cheng K T. SVM binary classifier ensembles for image classification[C]//Proceedings of the tenth international conference on Information and knowledge management. ACM, 2001: 395-402.

[4] Müller K R, Mika S, Rätsch G, et al. An introduction to kernel-based learning algorithms[J]. Neural Networks, IEEE Transactions on, 2001, 12(2): 181-201.

[5] de Campos L M L, de Oliveira R C L, Roisenberg M. Optimization of neural networks through grammatical evolution and a genetic algorithm[J]. Expert Systems with Applications, 2016, 56: 368-384.

[6] Cantarella G E, de Luca S, Di Pace R, et al. Signal setting design at a single junction through the application of genetic algorithms[M]//Computer-based Modelling and Optimization in Transportation. Springer International Publishing, 2014: 321-331.

[7] Wang Y, Zhang J, Zhao Y, et al. Application of elevator group control system based on genetic algorithm optimize BP fuzzy neural network[C]//Intelligent Control and Automation, 2008. WCICA 2008. 7th World Congress on. IEEE, 2008: 8702-8705.

[8] Huang, Y. Advances in artifical neural networks-methodological development and application[J]. Algorithms, 2009, 2(3): 973-1007

[9] 1.Cheng-Lung, Huanga, Wang, Chieh-Jen. A GA-based feature selection and parameters optimizationfor support vector machines[J]. Expert Systems with Applications, 2005, 31(2): 231-240

[10] Dong Wang, Xiangbin Wu, "The Optimization of SVM Parameters Based on PSO", Computer Applications, 2008,Vol. 28, No. 1, pp. 134-139.