Expert Systems with Applications 38 (2011) 5197–5204

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Feature selection and parameter optimization for support vector machines: A new approach based on genetic algorithm with feature chromosomes

Mingyuan Zhao *, Chong Fu, Luping Ji, Ke Tang, Mingtian Zhou

School of Computer Science and Technology, University of Electronic Science and Technology of China, Chengdu 610054, China

ARTICLE INFO

Keywords:
Feature chromosomes
Genetic algorithm
Feature selection
Parameters optimization
Support vector machines

ABSTRACT

Support vector machines (SVM) are an emerging data classification technique with many diverse applications. The feature subset selection, along with the parameter setting in the SVM training procedure significantly influences the classification accuracy. In this paper, the asymptotic behaviors of support vector machines are fused with genetic algorithm (GA) and the feature chromosomes are generated, which thereby directs the search of genetic algorithm to the straight line of optimal generalization error in the superparameter space. On this basis, a new approach based on genetic algorithm with feature chromosomes, termed GA with feature chromosomes, is proposed to simultaneously optimize the feature subset and the parameters for SVM.

To evaluate the proposed approach, the experiment adopts several real world datasets from the UCI database and from the Benchmark database. Compared with the GA without feature chromosomes, the grid search, and other approaches, the proposed approach not only has higher classification accuracy and smaller feature subsets, but also has fewer processing time.

1. Introduction

Support vector machines (SVM), one of the emerging data classification technique, were proposed by Vapnik (1995), and have recently been widely adopted in various fields of classification problems including pattern recognition, bioinformatics and finance (Kotsia & Pitas, 2007; Lee, 2009; Ma, Nguyen, & Rajapakse, 2009; Melgani & Bazi, 2008; Yu, Chen, Wang, & Lai, 2009). In the study of improving learning ability and generalization ability for SVM, a crucial problem is to optimize feature subset and parameters for SVM simultaneously, so as to improve classification accuracy of SVM.

Feature selection is used to identify a powerfully predictive subset of fields within a database and reduce the number of fields presented to the computational process. Feature selection affects several pattern classification aspects, including the accuracy of the learned classification algorithm, the time needed for learning a classification function, the number of examples needed for learning, and the cost associated with the features (Yang & Honavar, 1998). In a specific application problem, not all of these features are equally important. Better performance may be achieved by discarding some features. Thus, we imply eliminating noisy, irrelevant and redundant data, while maintaining the discriminating power of the data by feature selection.

Besides feature selection, parameters setting of SVM have an important influence on its classification accuracy. Inappropriate parameter settings lead to poor classification results (Keerthi & Lin, 2003). The optimal classification accuracy of SVM is obtained by searching optimal parameters setting. The parameters that should be optimized include the error penalty parameter C and the kernel function parameters such as the Gaussian kernel parameter γ for the Gaussian kernel function. The grid search is an alternative and straightforward search approach. However, search ability of this approach is low (Hsu, Chang, & Lin, 2003). Moreover, the grid search cannot perform the feature selection. The two one-dimensional searches try to improve the grid search (Keerthi & Lin, 2003), but the improvement of search effect is low, and cannot perform the feature selection.

In the literature, some feature selection methods based on genetic algorithm (GA) were proposed (Raymer, Punch, Goodman, Kuhn, & Jain, 2000; Yang & Honavar, 1998). Some other methods for SVM feature selection have been proposed (Guyon, Weston, Barnhill, & Bapnik, 2002; Mao, 2004). However, these papers did not deal with parameters optimization for SVM.

The trend in recent years is to simultaneously optimize feature subset and parameter for SVM, so as to increase classification accuracy for SVM. Genetic algorithm has the potential to generate both the optimal feature subset and SVM parameters at the same time. Huang and Wang (2006) proposed feature selection and parameters optimization of support vector machines based GA.

In these literatures mentioned above except Keerthi's paper, all proposed methods cannot fuse with the asymptotic behaviors of

* Corresponding author.
E-mail addresses: zmgyn@mail.sc.cninfo.net, myzhao@uestc.edu.cn (M. Zhao).

0957-4174/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2010.10.041

SVM. Keerthi et al. proposed the two one-dimensional searches based on the asymptotic behaviors of SVM, but cannot combine search function of genetic algorithm, also cannot perform the feature selection. In this paper, the asymptotic behaviors of SVM are fused with genetic algorithm and the feature chromosomes are generated, which thereby directs the search of genetic algorithm to the straight line of optimal generalization error in the superparameter space. On this basis, a new approach based on genetic algorithm with feature chromosomes is proposed to simultaneously optimize the feature subset and the parameters for SVM.

The remainder of this paper is organized as follows: a brief introduction to support vector machines and its asymptotic behaviors is given in Section 2. Section 3 presents basic genetic algorithm concepts. Section 4 then describes genetic algorithm with feature chromosomes. Section 5 presents the system architectures of feature selection and parameter optimization for SVM based on genetic algorithm with feature chromosomes. Section 6 presents the experimental results from using the proposed method to classify several real world datasets. Section 7 summarizes this paper.

2. Support vector machines and its asymptotic behaviors

2.1. Support vector machines

Given the training sample of instance-label pairs $(\mathbf{x}_i, y_i)$, $i = 1,\ldots,l$, $\mathbf{x}_i \in \mathbf{R}^n$, $y_i \in \{1,-1\}$, support vector machines require the solution of the following (primal) problem (Keerthi & Lin, 2003):

$$\begin{aligned} \min_{w,b,\xi} \quad & \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{l}\xi_i \\ \text{subject to} \quad & y_i(\mathbf{w}^T\mathbf{z}_i + b) \geqslant 1 - \xi_i, \\ & \xi_i \geqslant 0, \quad i = 1,\ldots,l, \end{aligned} \tag{1}$$

where the training vector $\mathbf{x}_i$ has been mapped onto a high dimensions space by mapping function φ as $\mathbf{z}_i = \varphi(\mathbf{x}_i)$. $C > 0$ is the penalty parameter of the error term.

Usually we solve the Eq. (1) by solving the following dual problem:

$$\begin{aligned} \min_{\alpha} \quad & F(\alpha) = \frac{1}{2}\alpha^T Q\alpha - e^T\alpha \\ \text{subject to} \quad & 0 \leqslant \alpha_i \leqslant C, \quad i = 1,\ldots,l, \\ & y^T\alpha = 0, \end{aligned} \tag{2}$$

where e is the vector of all ones and Q is an l by l positive semidefinite matrix. The (i,j)th element of Q is given by $Q_{i,j} \equiv y_i y_j K(\mathbf{x}_i\mathbf{x}_j)$, $K(\mathbf{x}_i,\mathbf{x}_j) \equiv \varphi^T(\mathbf{x}_i)\varphi(\mathbf{x}_j)$ is called the kernel function, $\{\alpha_i\}_{i=1}^{l}$ is Lagrange multipliers, and $\mathbf{w} = \sum_{i=1}^{l}\alpha_i y_i \varphi(\mathbf{x}_i)$ is the weight vector.

The classification decision function is

$$\operatorname{sgn}(\mathbf{w}^T\phi(\mathbf{x}) + b) = \operatorname{sgn}\left(\sum_{i=1}^{l}\alpha_i y_i K(\mathbf{x}_i,\mathbf{x}) + b\right). \tag{3}$$

The kernel function $K(\mathbf{x}_i,\mathbf{x}_j)$ has manifold forms. This paper discussion object is Gaussian kernel function that applies widely among them. Gaussian kernel function is expressed as follows:

$$K(\mathbf{x},\mathbf{x}_i) = \exp\left(-\gamma\|\mathbf{x} - \mathbf{x}_i\|^2\right), \tag{4}$$

or

$$K(\mathbf{x},\mathbf{x}_i) = \exp\left(-\frac{1}{\sigma^2}\|\mathbf{x} - \mathbf{x}_i\|^2\right). \tag{5}$$

The two equations are in essence the same, which can transform parameter γ and parameter σ^2 by $\gamma = \frac{1}{\sigma^2}$. The Gaussian kernel parameter γ is determined by the user.

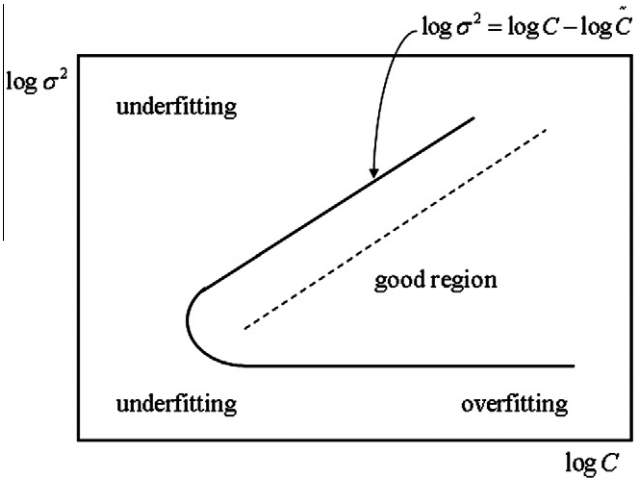

Fig. 1. Hyperparameter space.

The parameters of support vector machines with Gaussian kernel refer to the error penalty parameter C and the Gaussian kernel parameter γ, namely which is parameters (C,γ).

2.2. Asymptotic behaviors of support vector machines

Asymptotic behaviors of support vector machines with Gaussian kernel were suggested by Keerthi and Lin (2003). Taking $\log C$ and $\log\sigma^2$ as the parameters of the hyperparameter space, in the asymptotic (outer) regions of the $(\log C, \log\sigma^2)$ space, there exists a contour of generalization error (or an estimate such as LOO error or k-fold cross validation error) and helps separate the hyperparameter space into two regions: an overfitting/underfitting region and a good region (which most likely has the hyperparameter set with the best generalization error), which looks like that shown in Fig. 1.

For each fixed $\tilde{C}$, the equation

$$\log\sigma^2 = \log C - \log\tilde{C}, \tag{6}$$

defines a straight line of unit slope. As $\sigma^2 \to \infty$ along this line, the SVM classifier converges to the linear SVM classifier with penalty parameter $\tilde{C}$. The dotted line corresponds to the choice of $\tilde{C}$ that gives the optimal generalization error for the linear SVM.

Keerthi et al. have conducted theoretical analysis and mathematical reasoning for performance feature of SVM with Gaussian kernel. Asymptotic behaviors of SVM with Gaussian kernel, which are proposed by them, are a research achievement about performance feature of SVM. But compared with classification accuracy of the grid search, there is little improvement made by the two one-dimensional searches on this basis. Thus it is necessary to propose a new algorithm based on asymptotic behaviors of SVM to further improve classification accuracy of SVM.

3. Genetic algorithm

Genetic algorithm (GA) (Davis, 1991; Goldberg, 1989; Holland, 1975), a general adaptive optimization search methodology based on analogy to Darwinian natural selection and genetic in biology systems, uses fitness function and probability transform rule to guide search direction. In GA, a population is a set of candidate solutions. The population is composed of the chromosomes. Each chromosome is a candidate solution, and each individual is a chromosome. According to Darwinian principle of 'survival of the fittest', GA obtains the optimal solution after a series iterative computation.

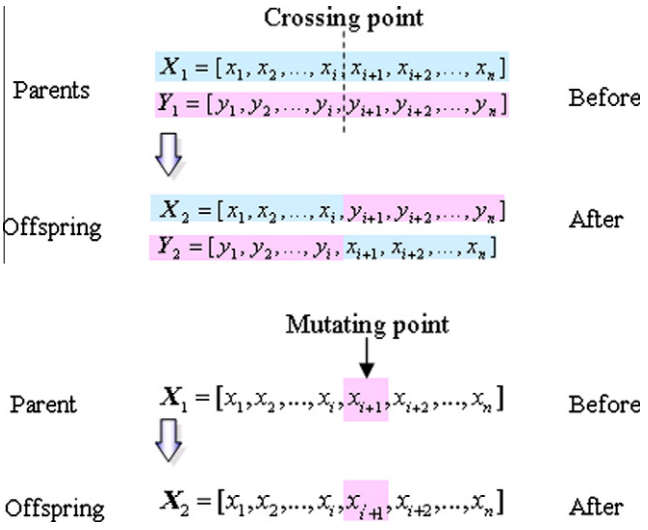

Fig. 2. Crossover and mutation operation.

The fitness function expresses adaptive ability of each chromosome for its living environment. A fitness function estimates the quality of a solution in the evolutionary process. The genetic operator, which includes the crossover operator, the mutation operator and the selection operator, randomly impact the fitness value. After many generations, the solution of the chromosome that obtains best fitness value is optimal solution.

The crossover is most important genetic operator, which is a random mechanism for exchanging genes between two chromosomes using the one point crossover, two point crossover, or homologue crossover. The mutation is the genes in the chromosome to alter occasionally, namely genes code from 1 to 0 or vice versa. It is shown in Fig. 2. In selection, the chromosomes of highest fitness value have higher probability to be selected into the recombination pool by the roulette wheel or the tournament selection methods.

GA uses evolutionary strategy, which includes elitist selection GA, father-offspring combined selection GA and so on, to guarantee the convergence of GA.

Genetic algorithm applies to feature selection and parameters optimization for support vector machines, so that obtained result is optimal parameters and feature subset.

The basic process of Genetic algorithm is follows:

Step 1: Randomly generate initial population.
Step 2: Estimate the fitness value of each chromosome in the population.
Step 3: Perform the crossover, the mutation and the selection etc. genetic operations.
Step 4: Stop the algorithm if termination criterion is satisfied; return to Step 2 otherwise.

The termination criterion is the pre-determined maximum number of iteration.

4. Genetic algorithm with feature chromosomes

Genetic algorithm is the basis of proposed genetic algorithm with feature chromosomes. Fundamental difference between them, when the latter is applied to feature selection and parameters optimization for SVM, is fused with asymptotic behaviors of SVM and generating feature chromosome. The flow chart of genetic algorithm with feature chromosomes is shown in Fig. 3.

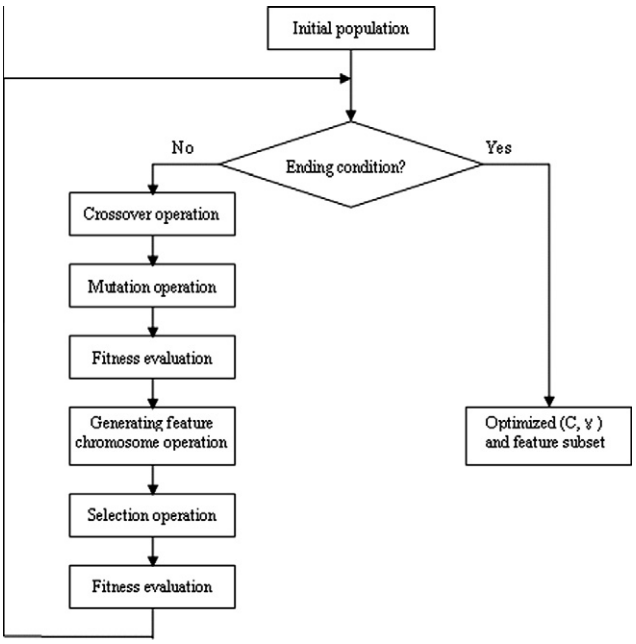

Fig. 3. The flow chart of genetic algorithm with feature chromosome.

The detailed description of the genetic algorithm with feature chromosomes is as follow.

4.1. Chromosome coding

The coding is an important step of the genetic algorithm with feature chromosome. In the step, variable value of parameters (C,γ) and feature subset selection f for SVM are transformed to binary coding. The chromosome coding of the genetic algorithm with feature comprises two parts, the coding of parameters and the coding of feature subset selection. It is shown in Fig. 4.

In the coding of parameters (C,γ), $b_C^1 \sim b_C^{n_C}$ is binary coding of C, $b_\gamma^1 \sim b_\gamma^{n_\gamma}$ is binary coding of γ, n_C represents the number of bits in binary coding of C, n_γ represents the number of bits in binary coding of γ, and selection of n_C and n_γ are according as computational precision.

In the coding of feature subset selection f, '1' represents that feature is selected; '0' represents that feature is not selected, $b_f^1 \sim b_f^{n_f}$ is binary coding of f, n_f represents the number of bit in binary coding of f, and n_f equals the number of feature in the dataset.

4.2. Selection of initial population

Initial population $\boldsymbol{X}(0)$, which consists of N pairs of parent, is generated randomly. Set to $t := 0$.

If the size of the population is too large, then the complexity of the algorithm is too high, and the computation quantity is too large. If the size of the population is too small, then optimal performance of the algorithm is reduced, and the algorithm is plunged into local optimal solution easily. The size of the training samples

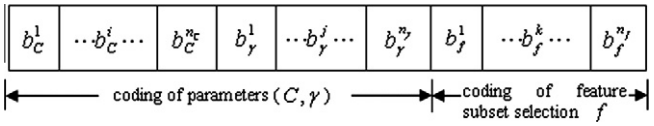

Fig. 4. The chromosome coding comprises two parts: the coding of parameters (C,γ) and the coding of feature subset selection f.

should be considered generally. The size range from 40 to 200 is suitable.

4.3. Crossover operation

Crossover operation generates intermediate population $\boldsymbol{C}(t)$ by performing crossover for N pairs of parent in current population $\boldsymbol{X}(t)$ independently. Crossover operator $T_c : \boldsymbol{X}(t) \to \boldsymbol{C}(t)$ acts on the subspace of the individual space.

4.4. Mutation operation

Mutation operation generates filial generation population $\boldsymbol{M}(t)$ by performing mutation for intermediate individual in intermediate population $\boldsymbol{C}(t)$ independently. Mutation operator $T_m : \boldsymbol{C}(t) \to \boldsymbol{M}(t)$ changes the subspace constantly, which has search ability of full space of the individual space.

4.5. Fitness function

The classification accuracy of SVM, the number of selected features, and the feature cost are used to construct a fitness function. As Huang Cheng-Lung et al. decides on the strategy of the fitness function, a high fitness value is decided by high classification accuracy, a small number of features, and low total feature cost. Further, considering to avoid that denominator reaches zero, the fitness function is

$$fit = W_A \times A + W_F \times \left(P + \left(\sum_{i=1}^{n_f} C_i \times F_i \right) \right)^{-1}, \tag{7}$$

where A is classification accuracy, W_A is weight of classification accuracy, F_i is feature value, W_F is feature weight, C_i is feature cost, P is setting constant of avoiding that denominator reaches zero.

W_A can be set from 75% to 100%, and $W_F = 1 - W_A$. In F_i, '1' represents that feature i is selected; '0' represents that feature i is not selected. C_i references to feature cost in the dataset from the UCI, if we do not have, also can be set from 1 to 8 or another number. P sets from 1 to 10 generally.

4.6. Generating feature chromosome operation

In the good region of the $(\log C, \log \sigma^2)$ superparameter space, by choosing appropriate $\widetilde{C}$ values, we can get the straight line of optimal generalization error (the dotted line in Fig. 1), namely the optimal straight line of optimal accuracy rate, thus finding optimal parameter values. The problem is how to choose appropriate $\widetilde{C}$ values.

Using search function of genetic algorithm, after crossover and mutation, we select r chromosomes of highest fitness value in the same generation chromosomes and conduct appropriate choice of $\widetilde{C}$ values.

When environment changes from the hyperparameter space to genetic algorithm, we transform Eq. (6) into Eq. (8)

$$\widetilde{C} = \frac{C}{\sigma^2}. \tag{8}$$

To state conveniently, after replacing $\widetilde{C}$ by K and replacing σ^2 by γ in Eq. (8), we obtain

$$K = C\gamma. \tag{9}$$

Eq. (9) is an essentially equation to describe the asymptotic behaviors of support vector machines applied to genetic algorithm.

First, the binary coding of parameters (C, γ) is extracted in every chromosome from choosing r chromosomes of highest fitness value and transformed into corresponding variable values, respectively. Next, K value of Every choice chromosome is calculated by Eq. (9). Then, we discretize searching range of variable values of C to d values, and d values of γ are calculated by $\gamma = \frac{K}{C}$ correspondingly. The variable values of $r \times d$ new parameters (C, γ) are generated. Finally, these parameters are transformed into binary coding and connected with corresponding binary coding of the feature subset selection f in former choice chromosomes. Thus, $r \times d$ new chromosomes are generated:

$$\begin{aligned}
&\underbrace{(C_{11}, \gamma_{11}, f_1), (C_{12}, \gamma_{12}, f_1), \ldots, (C_{1d}, \gamma_{1d}, f_1)}_{\text{generating new chromosomes by first choice chromosome}}, \\
&\underbrace{(C_{21}, \gamma_{21}, f_2), (C_{22}, \gamma_{22}, f_2), \ldots, (C_{2d}, \gamma_{2d}, f_2)}_{\text{generating new chromosomes by second choice chromosome}}, \\
&\ldots \underbrace{(C_{r1}, \gamma_{r1}, f_r), (C_{r2}, \gamma_{r2}, f_r), \ldots, (C_{rd}, \gamma_{rd}, f_r)}_{\text{generating new chromosomes by } r\text{th choice chromosome}}.
\end{aligned}$$

Because new generated chromosomes have selected appropriate $\widetilde{C}$ values (namely$K$ values) and contained feature of asymptotic behaviors of support vector machines, they are known as feature chromosomes.

The steps of generating feature chromosome operation are as follow.

(1) Select parent of generating feature chromosome.
The chromosomes after crossover and mutation in the same generation are arranged in fitness values order from high to low. Top r chromosomes are selected as parent of generating feature chromosome. The coding of parameters (C, γ) in the parent chromosome is extracted and transformed into corresponding variable values.

(2) Compute K values of every selected chromosome

$$K_j = C_j \gamma_j, \quad j = 1, 2, \ldots, r. \tag{10}$$

(3) Generate new parameters.
Discretize searching range of variable values of C to d values. Then d values of C_j are $C_{j1}, C_{j2}, \ldots, C_{jd}$.
Compute d values of $\gamma_j : \gamma_{j1}, \gamma_{j2}, \ldots, \gamma_{jd}$, by K_j, C_j and Eq. (11)

$$\gamma_j = \frac{K_j}{C_j}, \quad j = 1, 2, \ldots, r. \tag{11}$$

The variable values of $r \times d$ new parameters (C, γ) are generated.

(4) Generating feature chromosomes.
The variable values of $r \times d$ new parameters (C, γ) are transformed into binary coding and connected with corresponding binary coding of the feature subset selection f in former choice chromosomes. $r \times d$ feature chromosomes are generated. For example, top 12 highest fitness value chromosomes are selected as parent of generating feature chromosome, to discretize searching range of variable values of C to 10 values, and every parent generates 10 feature chromosomes, so that 120 feature chromosomes are generated in total.

In generating feature chromosomes operation, the asymptotic behaviors of SVM are fused with genetic algorithm by computing $\widetilde{C}$ value of the chromosome of high fitness value in filial generation population $\boldsymbol{M}(t)$, and feature chromosomes population $\boldsymbol{F}(t)$ is generated. Generating feature chromosomes operator $T_f : \boldsymbol{M}(t) \to \boldsymbol{F}(t)$ strengthens searching dynamics of genetic algorithm and improves classification accuracy of SVM.

4.7. Selection operation

Selection operation adopts father-offspring combined selection of genetic algorithm. In parent generation population $\boldsymbol{X}(t)$, and filial generation population $\boldsymbol{M}(t)$ and feature chromosomes population $\boldsymbol{F}(t)$, N parent pairs are selected as new generation parent population $\boldsymbol{X}(t+1)$. Selection operator is T_s : $\boldsymbol{M}(t) \cup \boldsymbol{F}(t) \cup \boldsymbol{X}(t) \rightarrow \boldsymbol{X}(t+1)$.

5. System architectures of feature selection and parameters optimization for SVM based on genetic algorithm with feature chromosomes

System architectures of feature selection and parameters optimization for SVM based on genetic algorithm with feature chromosomes are shown in Fig. 5. Its main steps are described as follows:

(1) Input dataset.
Input dataset includes training dataset and testing dataset.
(2) Data preprocess.
Data preprocess adopts linear scaling. Each feature of the dataset can be linearly scaled to the range [0,1] by Eq. (12)

$$a' = \frac{a - \min}{\max - \min}, \tag{12}$$

where a is original value, a' is scaled value, max is upper bound of the feature value, and min is low bound of the feature value.
Each feature also can be scaled to the range $[-1,+1]$.
The advantage of linear scaling is to avoid attributes in greater numeric ranges dominating those in smaller numeric ranges, and to avoid numerical difficulties during the calculation Feature value, and help to get higher classification accuracy.
(3) Select feature subset.
Relational feature subsets are chosen and unrelated feature subsets are discarded by feature subset selection. After training dataset and testing dataset discard unrelated feature subsets, respectively, they become training dataset of selected feature subset and testing dataset of selected feature subset.
(4) Train SVM classifier.
SVM classifier is trained by training set with selected feature subset and variable value of parameters (C,γ). The weight vector and the bias value of SVM classifier are acquired.
(5) Calculate accuracy rate.
Classification accuracy rate of SVM is calculated by testing set with selected feature subset and variable value of parameters (C,γ), and the acquired weight vector and bias value of SVM classifier.
(6) Fitness evaluation.
Classification accuracy of SVM, the number of selected features, and the feature cost are used to construct a fitness function. Every chromosome is evaluated by fitness function equation (7).
(7) Ending condition.
When the ending condition is satisfied, the operation ends; otherwise, we proceed with the next generation operation.
(8) Genetic operation.
The system searches for better solutions by genetic operations, including crossover, mutation and selection.
(9) Generating feature chromosome operation.
Generating feature chromosome by fusing with asymptotic behaviors of support vector machines to strengthen searching dynamics of genetic algorithm and improve classification accuracy of support vector machines.
(10) Convert genotype to phenotype.
Converting genotype to phenotype refers to convert coding of parameters (C,γ) and coding of feature subset selection f to variable values of parameters (C,γ) and feature subset selection.

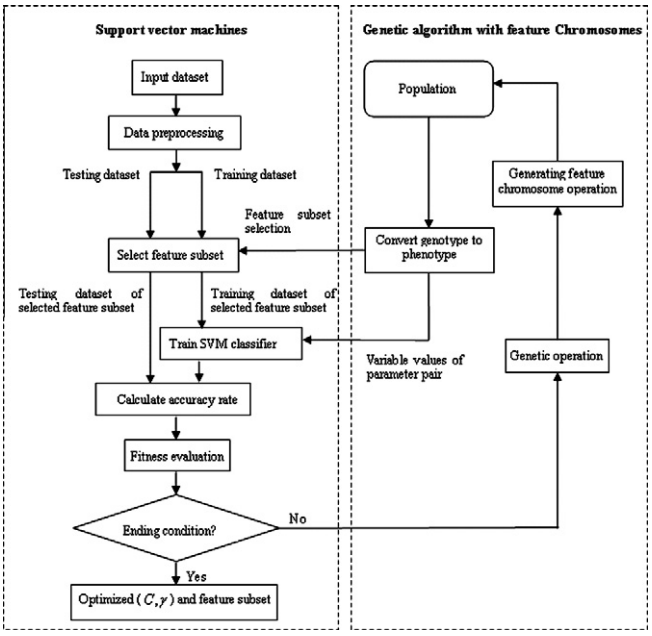

Fig. 5. System architectures of feature selection and parameters optimization for SVM based on genetic algorithm with feature chromosomes.

6. Experiment

The used platform is Intel Pentium IV 2.2 GHz CPU, 512 MB RAM, Windows XP operating system. The development environment is MATLAB 7.3. The software of SVM is LIBSVM (Chang & Lin, 2001).

In experiment, searching range of parameter C is [0.01,35,000], searching range of parameter γ is [0.0001,10] (Hsu et al., 2003).

To evaluate the classification accuracy of the proposed approach, the experiment adopts several real world datasets. One is datasets from the UCI machine learning repository (Hettich, Blake, & Merz, 1998): Australian, Breast cancer, German, Heart Disease, Ionosphere, Iris, Liver disorders, Pima, Sonar, Vehicle, Wine and Vowel. Their number of classes, number of instances, and number of features are shown in Table 1. Another is datasets from the Benchmarks database (Rätsch, 1999): Banana, Diabetes, Image, Splice, Ringnorm, Twonorm and Waveform. Their number of classes, number of training examples, number of testing examples and number of features are shown in Table 2. These data sets have been used as benchmarks to compare the performance of different classification methods in the paper frequently.

In the experiment of UCI datasets, we adopts k-fold cross validation. The value of k is set to 10. Thus, the dataset was randomly split into 10 independent subsets. The size of each subset is approximately equality. One subset in them is used as independent testing dataset, and the rest of nine subsets are used as training dataset (Han & Kamber, 2003). The program was run 10 times to enable each slice of data to take a turn as the testing dataset. The rate of accuracy in classification of the experiment was computed by summing the individual accuracy rate for each run of testing, and then dividing the total by 10. The advantages of cross validation are that all of the test sets were independent and the reliabil-

Table 1
Dataset from the UCI repository.

Number	Dataset	Number of classes	Number of instances	Number of features
1	Australian (Statlog Project)	2	690	14
2	Breast cancer (Wisconsin)	2	699	10
3	German (Statlog Project)	2	1000	24
4	Heart Disease (Statlog Project)	2	270	13
5	Ionosphere	2	351	34
6	Iris	3	150	4
7	Liver disorders	2	345	6
8	Pima (Indian diabetes)	2	768	8
9	Sonar	2	208	60
10	Vehicle (Statlog Project)	4	846	18
11	Wine	3	178	13
12	Vowel	11	990	13

Table 2
Dataset from the benchmarks.

Number	Dataset	Number of classes	Number of training examples	Number of testing examples	Number of features
1	Banana	2	400	4900	2
2	Diabetes	2	468	300	8
3	Image	2	1300	1010	18
4	Splice	2	1000	2175	60
5	Ringnorm	2	400	7000	20
6	Twonorm	2	400	7000	20
7	Waveform	2	400	4600	21

ity of the results could be improved. In all of the experiments, 10-fold cross validation was used to estimate the classification accuracy of the proposed approach.

In the experiment of Benchmarks datasets, the SVM classifier is trained by training dataset, and the testing accuracy is calculated by testing dataset. The testing accuracy is as classification accuracy of this dataset.

The results obtained by the proposed approach based on genetic algorithm with feature chromosome and the GA-based approach by Huang and Wang (2006). Table 3 shows the results comparison of classification accuracy. 10-fold cross validation was used to estimate the classification accuracy of each approach. The Obtained classification accuracy is illustrated with the form of 'average ± standard deviation'. The proposed approach yielded the higher classification accuracy rate in 9 datasets, while in 1 dataset the proposed approach and the GA-based approach yielded the same classification accuracy. Thus, the proposed approach yielded higher classification accuracy rate on different datasets.

The experiment is 7 datasets from the Benchmarks database: Banana, Diabetes, Image, Splice, Ringnorm, Twonorm and Waveform. The comparison results obtained by the proposed approach based on genetic algorithm with feature chromosome and the two one-dimensional searches by Keerthi and Lin (2003). Table 4 shows the comparison results of classification accuracy. In the experiment, the SVM classifier is trained by training dataset, and the testing accuracy is calculated by testing dataset which is used as classification accuracy of this dataset. To facilitate comparison, the testing error in Keerthi's paper is converted into classification accuracy. The proposed approach yielded the higher classification accuracy rate on all datasets. Thus, the proposed approach obtained higher classification accuracy rate.

To estimate the validity of our proposed approach, Table 5 gives the classification accuracies of our approach and previous approaches in Breast cancer dataset. As we can see from the results, our proposed approach obtains the highest classification accuracy so far.

Comparison of our study with other studies in literature is an indispensable part to evaluate our proposed approach. The experimental results of our proposed approach and approaches in literature in Heart disease dataset are shown in Table 6. As can be seen, the highest classification accuracy was obtained with our approach.

Table 7 gives the experiment results of our proposed approach based on genetic algorithm with feature chromosome and the grid search in the 12 UCI dataset. 10-fold cross validation was used to estimate the classification accuracy of each approach. The Obtained classification accuracy is illustrated with the form of 'average ± standard deviation.' As we can see, the proposed approach produces small feature subset, and grid search uses all features. To compare classification accuracy of the proposed approach with the Grid search, we used the nonparametric Wilconxon signed rank test for all of the datasets. As shown in Table 7, the p-value for wine is larger than the prescribed statistical significance level of 0.005, but other p-values are smaller than the significance level of 0.005. Generally, compared with the grid search, the proposed approach has higher classification accuracy and fewer features.

To identify any differences in the classification accuracy rates of our proposed approach based on genetic algorithm with feature

Table 3
Comparison between the GA with feature chromosome and the GA-based approach.

Dataset	GA with feature chromosome	GA-based approach
Australian	91.59 ± 2.14	88.10 ± 2.25
Breast cancer	99.00 ± 1.66	96.19 ± 1.24
German	86.10 ± 1.97	85.60 ± 1.96
Heart disease	95.56 ± 2.34	94.80 ± 3.32
Ionosphere	99.43 ± 1.21	98.56 ± 2.03
Iris	100.00 ± 0.00	100.00 ± 0.00
Pima	83.84 ± 5.14	81.50 ± 7.13
Sonar	99.00 ± 2.11	98.00 ± 3.50
Vehicle	88.24 ± 1.47	84.06 ± 3.54
Vowel	99.60 ± 0.71	99.30 ± 0.82

Table 4
Comparison between the GA with feature chromosome and the two one-dimensional searches.

Dataset	GA with feature chromosome	Two one-dimensional searches
Banana	89.57	88.22
Diabetes	78.33	75.67
Image	98.12	97.53
Splice	90.53	89.89
Ringnorm	98.34	98.20
Twonorm	97.26	97.09
Waveform	89.70	89.22

Table 5
Comparison accuracies among the GA with feature chromosome and other approaches from literature in Breast cancer dataset.

Author (year)	Method	Accuracy (%)
Our study (2009)	GA with feature chromosome	99.00
Lin et al. (2008)	PSO + SVM	97.95
Polat et al. (2007)	Fuzzy-AIRS	98.51
Polat and Gunes (2007)	LS-SVM	98.53
Huang and Wang (2006)	GA-based	96.19
Abonyi and Szeifert (2003)	Supervised fuzzy clustering	95.57
Goodman et al. (2002)	AIRS	97.20
Goodman et al. (2002)	Big LVQ	96.80
Goodman et al. (2002)	Optimize-LVQ	96.70

Table 6
Comparison accuracies among the GA with feature chromosome and other approaches from literature in Heart disease dataset.

Author (year)	Method	Accuracy (%)
Our study (2009)	GA with feature chromosome	95.56
Huang and Wang (2006)	GA-based	94.80
Ozsen and Gunes (2009)	GA-AWAIS	87.43
Sahan et al. (2005)	AWAIS	82.59
Cheung (2001)	Naïve-Bayes	81.48
Cheung (2001)	BNND	81.11
Cheung (2001)	C4.5	81.11
Cheung (2001)	Logistic regression	80.96

chromosome and the approach based on genetic algorithm without feature chromosome, we used 10-fold cross validation to estimate the classification accuracy of each approach in the 12 UCI dataset, and used the nonparametric Wilconxon signed rank test for all of the datasets. As shown in the Table 8, the *p*-value for Ionosphere, Pima, Wine and Vowel are larger than the prescribed statistical significance level of 0.005, but other *p*-values are smaller than the significance level of 0.005. Generally, compared with the genetic algorithm, the proposed approach has higher classification accuracy and less feature subset.

To conduct performance comparison among the proposed approach based on genetic algorithm with feature chromosome, the approach based on genetic algorithm without feature chromosome and the grid search, and 4 datasets (Breast cancer, Heart Disease, Ionosphere and Liver in the UCI machine learning repository) are adopted. 10-fold cross validation was used to estimate the classification accuracy of each approach. All obtained classification accu-

Table 7
Experiment results of the proposed GA with feature chromosome and the grid search.

Dataset	Number of original features	GA with feature chromosome		Grid search	*p*-Value for Wilcoxon testing
		Classification accuracy (%)	Number of selected features	Classification accuracy (%)	
Australian	14	91.59 ± 2.14	5.2 ± 2.15	84.74 ± 4.52	<0.005
Breast cancer	10	99.00 ± 1.66	2.5 ± 0.88	95.43 ± 2.50	<0.005
German	24	86.10 ± 1.97	10.3 ± 1.76	78.9 ± 1.73	<0.005
Heart disease	13	95.56 ± 2.34	6.2 ± 1.12	88.15 ± 5.18	<0.005
Ionosphere	34	99.43 ± 1.21	13.9 ± 3.45	94.29 ± 3.56	<0.005
Iris	4	100.00 ± 0.00	1.2 ± 0.28	94.09 ± 4.77	<0.005
Liver	6	83.14 ± 7.19	2.8 ± 0.99	72.89 ± 5.60	<0.005
Pima	8	83.84 ± 5.14	3.7 ± 1.26	76.58 ± 5.14	<0.005
Sonar	60	99.00 ± 2.11	26.4 ± 3.20	90.5 ± 8.32	<0.005
Vehicle	18	88.24 ± 1.47	9.2 ± 1.71	83.94 ± 2.74	<0.005
Wine	13	100.00 ± 0.00	4.2 ± 0.5	97.16 ± 2.88	0.0054
Vowel	13	99.60 ± 0.71	5.5 ± 1.58	95.36 ± 2.24	<0.005

Table 8
Experiment results of the proposed GA with feature chromosome and the GA without feature chromosome.

Dataset	Number of features	GA with feature chromosome		GA without feature chromosome		*p*-Value for Wilcoxon testing
		Classification accuracy (%)	Number of selected features	Classification accuracy (%)	Number of selected features	
Australian	14	91.59 ± 2.14	5.2 ± 2.15	86.81 ± 3.64	6.7 ± 3.16	<0.005
Breast cancer	10	99.00 ± 1.66	2.5 ± 0.88	96.04 ± 2.18	2.9 ± 0.99	<0.005
German	24	86.10 ± 1.97	10.3 ± 1.76	80.80 ± 2.10	11.8 ± 3.33	<0.005
Heart disease	13	95.56 ± 2.34	6.2 ± 1.12	91.11 ± 2.58	7.0 ± 1.05	<0.005
Ionosphere	34	99.43 ± 1.21	13.9 ± 3.45	98.57 ± 2.02	15.4 ± 3.32	0.3222
Iris	4	100.00 ± 0.00	1.2 ± 0.28	96.00 ± 3.44	1.8 ± 0.38	<0.005
Liver	6	83.14 ± 7.19	2.8 ± 0.99	81.43 ± 7.29	3.2 ± 1.14	<0.005
Pima	8	83.84 ± 5.14	3.7 ± 1.26	81.97 ± 5.34	5.1 ± 1.63	0.4427
Sonar	60	99.00 ± 2.11	26.4 ± 3.20	95.00 ± 2.36	28.7 ± 4.00	<0.005
Vehicle	18	88.24 ± 1.47	9.2 ± 1.71	84.74 ± 2.32	10.3 ± 2.72	<0.005
Wine	13	100.00 ± 0.00	4.2 ± 0.50	99.44 ± 1.76	4.6 ± 0.72	0.3681
Vowel	13	99.60 ± 0.71	5.5 ± 1.58	98.79 ± 1.70	6.9 ± 1.60	0.3980

Table 9
Performance comparison among the GA with feature chromosome, the GA without feature chromosome and the grid search.

Dataset	Classification accuracy (%)			Number of selected features		Average processing time (s)		
	Proposed approach	GA without feature chromosomes	Grid search	Proposed approach	GA without feature chromosomes	Proposed approach	GA without feature chromosomes	Grid search
Breast cancer	99.00	96.04	95.43	2.5	2.9	13.28.	14.49	15.90
Heart disease	95.56	91.11	88.15	6.2	7.0	14.53	11.41	12.08
Ionosphere	99.43	98.57	94.29	13.9	15.4	11.84	12.15	16.32
Liver	83.14	81.43	72.89	2.8	3.2	6.37	10.45	11.90

racy, number of selected features and average processing time use mean value of 10 experiments to denote. It is shown in Table 9. In Breast cancer dataset, the proposed approach obtained the classification accuracy of 99.00%, number of selected features of 2.5, and using 13.28 s. However, the grid search only obtained the classification accuracy of 95.43% and using 15.90 s, and cannot perform the feature selection. Similarly, the GA without feature chromosome only obtained the classification accuracy of 96.04%, number of selected features of 2.9, using 14.49 s. The classification accuracy of the proposed approach is 3.57% high than that of the grid search and 0.41% higher than that of the GA without feature chromosome, the average processing time of the proposed approach is 2.62 s less than that of the grid search and 1.21 s less than that of the GA without feature chromosome. In the rest of 3 datasets, the proposed approach obtained highest classification accuracy, least number of selected features and least average processing time except the average processing time in Heart Disease dataset. To sum up, the proposed approach achieved higher classification accuracy, less number of selected features and less average processing time, and had better performance.

7. Conclusions

In the paper, asymptotic behaviors of support vector machines are fused with genetic algorithm and feature chromosomes are generated, which thereby directs the search of genetic algorithm to the straight line of optimal generalization error in the superparameter space. On this basis, a new approach based on genetic algorithm with feature chromosomes is proposed. The proposed approach has following characters:

(1) When feature chromosomes is generated, it chooses appropriate $\widetilde{C}$ values in the superparameter space, which thereby directs the search of genetic algorithm to the straight line of optimal generalization error in the superparameter space. Therefore, it strengthens searching dynamics of genetic algorithm and improves classification accuracy of SVM.
(2) Feature chromosomes are one kind of new generating chromosomes. In order to resolve how to introduce new chromosomes of high fitness value in initialization later, middle period and later period of evolution and avoid premature convergence, a new method and thinking are proposed.
(3) The asymptotic behaviors of support vector machines are fused with genetic algorithm and the feature chromosomes are generated. Its essence is to make the performance feature of the application problem fuse with the algorithm, so as to increase efficiency of the algorithm and better solve the application problem.

Because the proposed approach has above characters and is proved by the experiments, it not only simultaneously optimizes the feature subset and the parameters for SVM, but improves classification accuracy and reduces processing time.

The experimental results are obtained by UCI datasets and Benchmarks database. However, other public datasets and relational question of the real world can also used to test, validate and extend the proposed approach.

Acknowledgements

The work was supported partially by the National Science Foundation (NSF) of China under the contract number 60671033, and the Doctoral Research Foundation of the Ministry of Education of China under the contract number 20090185120009.

References

Abonyi, J., & Szeifert, F. (2003). Supervised fuzzy clustering for the identification of fuzzy classifiers. *Pattern Recognition Letters, 14*(24), 2195–2207.
Chang, C. C., & Lin, C. J. (2001). *LIBSVM: A library for support vector machines*. Available from http://www.csie.ntu.edu.tw/~cjlin/libsvm.
Cheung, N. (2001). Machine learning techniques for medical analysis. School of Information Technology and Electrical Engineering. B.Sc. Thesis, University of Queenland.
Davis, L. (1991). *Handbook of genetic algorithms*. New York: Van Nostrand Reinhold.
Goldberg, D. E. (1989). *Genetic algorithms in search optimization and machine learning*. Reading, MA: Addison-Wesley.
Goodman, D. E., Boggess, L., & Watkins, A. (2002). Artificial immune system classification of multiple-class problems. In *Proceedings of the artificial neural networks in engineering* (pp. 179–183).
Guyon, I., Weston, J., Barnhill, S., & Bapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning, 46*(1–3), 389–422.
Han, J., & Kamber, M. (2003). *Data mining: Concepts and techniques*. San Francisco: Morgan Kaufmann.
Hettich, S., Blake, C. L., & Merz, C. J. (1998). *UCI repository of machine learning databases*. Department of Information and Computer Science, University of California, Irvine, CA. Available from http//www.ics.uci.edu/~mlearn/MLRepository.html.
Holland, J. H. (1975). *Adaptation in neural and artificial systems*. MIT Press.
Hsu, C. W., Chang, C. C., & Lin, C. J. (2003). A practical guide to support vector classification. Available from http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf.
Huang, C.-L., & Wang, C.-J. (2006). A GA-based feature selection and parameters optimization for support vector machines. *Expert Systems with Applications, 31*, 231–240.
Keerthi, S. S., & Lin, C.-J. (2003). Asymptotic behaviors of support vector machines with Gaussian kernel. *Neural Computation, 15*, 1667–1689.
Kotsia, I., & Pitas, I. (2007). Facial expression recognition in image sequences using geometric deformation features and support vector machines. *IEEE Transactions on Image Processing, 16*(1), 172–187.
Lee, M.-C. (2009). Using support vector machine with a hybrid feature selection method to the stock trend prediction. *Expert Systems with Applications, 36*, 10896–10904.
Lin, S.-W., Ying, K.-C., Chen, S.-C., & Lee, Z.-J. (2008). Particle swarm optimization for parameter determination and feature selection of support vector machines. *Expert Systems with Applications, 35*, 1817–1824.
Ma, J., Nguyen, M. N., & Rajapakse, J. C. (2009). Gene classification using codon usage and support vector machines. *IEEE/ACM Transactions on Computational Biology and Bioinformatics, 6*(1), 134–143.
Mao, K. Z. (2004). Feature subset selection for support vector machines through discriminative function pruning analysis. *IEEE Transactions on Systems, Man, and Cybernetics, 34*(1), 60–67.
Melgani, F., & Bazi, Y. (2008). Classification of electrocardiogram signals with support vector machines and swarm optimization. *IEEE Transactions on Information Technology in Biomedicine, 12*(5), 667–677.
Ozsen, S., & Gunes, S. (2009). Attribute weighting via genetic algorithms for attribute weighted artificial immune system (AWAIS) and its application to heart disease and liver disorders problems. *Expert Systems with Applications, 36*, 386–392.
Polat, K., & Gunes, S. (2007). Breast cancer diagnosis using least square support vector machine. *Digital Signal Processing, 17*(4), 694–701.
Polat, K., Sahan, S., Kodaz, H., & Gunes, S. (2007). Breast cancer and liver disorders classification using AIRS with performance evaluation by fuzzy resource allocation mechanism. *Expert Systems with Applications, 32*, 172–183.
Rätsch, G. (1999). *Benchmark data sets*. Available from http://ida.first.gmd.de/~raetsch/data/benchmarks.htm.
Raymer, M. L., Punch, W. F., Goodman, E. D., Kuhn, L. A., & Jain, A. K. (2000). Dimensionality reduction using genetic algorithms. *IEEE Transactions on Evolutionary Computation, 4*(2), 164–171.
Sahan, S., Kodaz, H., Gunes, S., & Polat, K. (2005). The medical applications of attribute weighted artificial immune system (AWAIS): Diagnosis of heart and diabetes diseases. In *Proceedings of fourth international conference of on artificial immune systems-ICARIS 2005, Canada* (pp. 456–468).
Vapnik, V. N. (1995). *The nature of statistical learning theory*. New York: Springer.
Yang, J., & Honavar, V. (1998). Feature subset selection using a genetic algorithm. *IEEE Intelligent Systems, 13*(2), 44–49.
Yu, L., Chen, H., Wang, S., & Lai, K. K. (2009). Evolving least squares support vector machines for stock market trend mining. *IEEE Transactions on Evolutionary Computation, 13*(1), 87–102.