Parameter determination of support vector machine and feature selection using simulated annealing approach

Shih-Wei Lin\textsuperscript{a,b,*}, Zne-Jung Leeb, Shih-Chieh Chen\textsuperscript{c}, Tsung-Yuan Tseng\textsuperscript{b}

\textsuperscript{a} Department of Information Management, Chang Gung University, No. 259 Wen-Hwa 1st Road, Kwei-Shan Tao-Yuan 333, Taiwan, ROC
\textsuperscript{b} Department of Information Management, Huafan University, No. 1 Huafan Road, Taipei, Taiwan, ROC
\textsuperscript{c} Department of Industrial Management, National Taiwan University of Science and Technology, No. 43 Keelung Road, Sec. 4, Taipei, Taiwan, ROC

Received 31 January 2007; received in revised form 6 October 2007; accepted 21 October 2007

Available online 26 October 2007

Abstract

Support vector machine (SVM) is a novel pattern classification method that is valuable in many applications. Kernel parameter setting in the SVM training process, along with the feature selection, significantly affects classification accuracy. The objective of this study is to obtain the better parameter values while also finding a subset of features that does not degrade the SVM classification accuracy. This study develops a simulated annealing (SA) approach for parameter determination and feature selection in the SVM, termed SA-SVM.

To measure the proposed SA-SVM approach, several datasets in UCI machine learning repository are adopted to calculate the classification accuracy rate. The proposed approach was compared with grid search which is a conventional method of performing parameter setting, and various other methods. Experimental results indicate that the classification accuracy rates of the proposed approach exceed those of grid search and other approaches. The SA-SVM is thus useful for parameter determination and feature selection in the SVM.

© 2007 Elsevier B.V. All rights reserved.

Keywords: Support vector machines; Simulated annealing; Parameter determination; Feature selection

1. Introduction

Classification problems have been extensively studied. Numerous factors, from incomplete data to the choice of parameter values for a given model, may influence classification outcomes. Classification problems have previously typically been tackled by statistical methods, such as logistic regression or discriminate analysis. Advances in technology have led to new techniques for solving classification problems, including decision trees, back-propagation neural networks, rough set theory and support vector machines (SVM). SVM is an emerging data classification technique first developed by Vapnik \cite{1}, and has been widely adopted in various fields of classification problems recently \cite{2–9}.

In the SVM, the model for classification is generated from the training stage using the sampling data. Classification is then performed based on the trained model. The biggest difficulties in setting up the SVM model are choosing the kernel function and its parameter values. If the parameter values are not set properly, then the classification outcomes will be less than optimal \cite{10}.

The bearing conditions are classified from the statistical features of both the original data and the data with some pre-processing, using differentiation and integration, low- and high-pass filtering, and spectral data of the database. In complex classification domains, some features may contain false correlations, which impede data processing. Moreover, some features may be redundant, since the information that they add is contained in other features. Redundant features can lengthen the computational time, influencing the classification accuracy. Hence, the classification process must be fast and accurate using the minimum number of features, which is a goal attainable through the use of feature selection. Feature selection has been applied to enhance classification performance, and to reduce data noise \cite{11–13}.

If the SVM is adopted without feature selection, then the dimension of the input space is large and non-clean, lowering...
the performance of the SVM. Thus, the SVM requires an efficient and robust feature selection method that discards noisy, irrelevant and redundant data, while still retaining the discriminating power of the data. Features extracted from the original data are adopted as inputs to the classifiers in the SVM.

This study integrates a simulated annealing (SA)-based approach with SVM to identify the optimal classification outcomes, termed SA-SVM. The proposed SA-SVM approach not only obtain the better parameter values of SVM, but also finds a subset of features for specific problems, thus maximizing the classification accuracy rate of SVM. This makes the optimal separating hyper-plane obtainable in both linear and non-linear classification problems.

The remainder of this paper is organized as follows. Section 2 reviews relevant literature on SVM and feature selection. Section 3 then introduces the proposed SA-SVM approach for improving SVM parameters with and without feature selection. Experimental results are compared with those of existing approaches in Section 4. Conclusions are finally drawn in Section 5, along with recommendations for future research.

2. Literature review

2.1. Support vector machine

SVM can be briefly described as follows [14–16]. Let \((x_1, y_1), \ldots, (x_m, y_m) \in X \times \{\pm 1\}\) denote a set of training data, where \(X\) represents some nonempty set from which the pattern \(x_i; y_i\) are called target \(y_i \in \{\pm 1\}\).

SVM attempts to identify a hyper-plane in a multi-dimensional space. This hyper-plane functions as a separating plane for data classification. Consider the class of hyper-planes in some dot product space \(H\). The parameters \(w\) and \(b\) are described as follows:

\[
(w \cdot x) + b = 0, \quad \text{where} \quad w \in H, \; b \in R. \quad (1)
\]

If a hyper-plane that satisfies Eq. (1) exists, then linear separation is obtained. In this case, \(w\) and \(b\) can be rewritten as follows: Eq. (1) becomes

\[
\max_{w \in H, b \in R} \min_{1 \leq i \leq N} \{||x_i - x|| \mid x \in H, (w \cdot x) + b = 0, i = 1, \ldots, m\}. \quad (2)
\]

Let the distance from the data point to the hyper-plane be \(1/||w||\).

The separating hyper-planes include one optimal separating hyper-plane (OSH), which has the largest distance between two support vector points on its two sides. Because the distance between two support vector points is \(1/||w||^2\), the minimal distance to OSH, \(||w||^2\), can be derived from Eq. (2).

The margin \(2/||w||\) of a separating hyper-plane can be regarded as the hyper-plane’s generalization ability, and the OSH has the maximal margin among separating hyper-planes. \(||w||^2\) can be minimized with Eq. (2) and Lagrange’s polynomial. Let \(\alpha\) denote \((\alpha_1, \ldots, \alpha_m)\). Combining Lagrange’s polynomial (in the order of \(m\)) with Eq. (2) produces the following maximization equations:

\[
W(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle \quad (3)
\]

subject to \(\alpha_i \geq 0\) for all \(i = 1, \ldots, m\) and \(\sum_{i=1}^{m} \alpha_i y_i = 0\). \(\quad (4)
\]

Quadratic programming can be applied to solve this maximization problem.

Given a vector satisfies Eq. (3) in maximization, the OSH expressed in terms of \((w, b)\) can be written as follows:

\[
w = \sum_{i=1}^{m} \alpha_i y_i x_i \quad (5)
\]

where the support vector points must satisfy \(\alpha_i \geq 0\) and Eq. (2). When considering expanding in constraint Eq. (4), the determinant function of hyper-plane is written as follows:

\[
f(x) = \text{sgn} \left( \sum_{i=1}^{m} y_i \alpha_i \langle x, x_i \rangle + b \right) = 0 \quad (6)
\]

In most cases, the data is not linearly separable, and is thus mapped to higher dimensional feature space. This means that if the data cannot be classified explicitly in the current dimensional space, then the SVM maps the data to a higher dimensional space for classification.

The input data are mapped to a higher dimensional feature space by plotting a nonlinear curve. The OSH is built in the feature space. The feature space vectors \(x, x'\) are constructed in terms of the kernel \(k(x, x') = \langle x, x' \rangle\).

The kernel function can be applied because all feature vectors occur only in dot products. The weight vector then becomes an expansion in feature space, and therefore typically no longer corresponds to the \(\Phi\)-image of a single input space vector.

The decision function is expressed as follows:

\[
f(x) = \text{sgn} \left( \sum_{i=1}^{m} \alpha_i y_i \langle \Phi(x), \Phi(x_i) \rangle + b \right) = \text{sgn} \left( \sum_{i=1}^{m} \alpha_i y_i k(x_i, x) + b \right), \quad (7)
\]

The quadratic program is written as

\[
\max_{\alpha \in R^m} \quad W(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle \quad (8)
\]

subject to \(\alpha_i \geq 0\) for all \(i = 1, \ldots, m\) and \(\sum_{i=1}^{m} \alpha_i y_i = 0\). \(\quad (9)
\]

Several kernel functions help the SVM in finding the optimal solution. The most frequently used such functions are the polynomial kernel, sigmoid kernel and radial basis kernel function (RBF) [17–19]. The RBF is used most often in general cases, since it can classify multi-dimensional data, unlike a
linear kernel function. Furthermore, the RBF has fewer parameters than a polynomial kernel. Overall, performance of an RBF and other kernel functions is not significantly different. Consequently, the RBF is an effective choice for the kernel function. Therefore, this study employs an RBF kernel function in the SVM to discover the optimal solution.

Two major parameters of the RBF applied in SVM, C and $\gamma$, have to be set appropriately. Parameter C is the cost of the penalty. The choice of value for parameter C influences the classification outcome. If C is too large, then the classification accuracy rate is very high in the training stage, but very low in the testing stage. If C is too small, then the classification accuracy rate is unsatisfactory, making the model useless. Parameter $\gamma$ has a much stronger impact than parameter C on classification outcomes, because its value influences the partitioning outcome in the feature space. An excessive value for parameter $\gamma$ leads to over-fitting, while a disproportionately small value results in under-fitting [20].

Grid search [13,21] is the simplest way to determine the values for parameters C and $\gamma$. Sets of values for parameters C and $\gamma$ that produce the highest classification accuracy rate in this interval are found by setting the upper and lower limits (search interval) for parameters C and $\gamma$ and the jumping interval in the search. However, this type of search is a local search and prone to a local optimality. Additionally, setting the search interval is a problem. Too large a search interval wastes computing power, while too small a search interval might render a satisfactory outcome impossible. Methods for obtaining the optimal parameters in the SVM are currently still under development.

In addition to the commonly adopted grid search technique, other techniques are used in SVM to improve the possibility of an appropriate choice of parameter values. The $F$-score approach utilizes the statistical concepts of type I and type II errors and Random forest (RF) [22]. Pai and Hong [23] presented an approach that combine genetic algorithms (GA) and the SVM. Their model imitates chromosome coding in their genetic algorithms to generate a set of parameter values for SVM. Pai and Hong [24,25] also presented an SA approach to obtain parameter values for SVM and applied it to real data. However, their approach does not address feature selection, and therefore may exclude the best result.

2.2. Feature selection

The classification accuracy rate of SVM is influenced by not only the two parameters C and $\gamma$, but also other factors including the quality of the feature’s dataset. For instance, the correlation between features influences the classification result. Accidental elimination of important features might decrease the accuracy rate of classification. Additionally, some features of the dataset may have no effect at all, or contain a high level of noise. Removal of such features can increase the search speed and the accuracy rate.

Feature selection methods can be categorized as filter models and wrapper models [26]. Filter models [26] utilize statistical techniques, such as principal component analysis (PCA), factor analysis (FA), independent component analysis (ICA) and discriminate analysis (DA) in the investigation of other indirect performance measures, mostly based on distance and information measures. Chen and Hsieh [27] presented latent semantic analysis (LSA) and web page feature selection (WPFA), which are combined with the SVM technique to screen features. Gold et al. [28] developed a Bayesian viewpoint of SVM classifiers to tune hyper-parameter values in order to determine useful criteria for pruning irrelevant features. Chapelle et al. [29] developed an automatically tuning multiple parameters and applied principal components to obtain features for SVM. SVM was used as a method of feature selection in gene selection for cancer classification, which is called SVM recursive feature elimination (SVM-RFE) by Guyon et al. [30]. Xie et al. [31] pointed out the problem that SVM-RFE cannot delete the redundant features. Therefore, they proposed an approach using correlation to delete the redundant features to enhance SVM-RFE. The proposed method can not only reduce the number of features, but also keep the classification accuracy. Even though the filter model is fast, it may not produce the optimal feature subset [26].

The simplest form of feature selection, the wrapper models [32], adopt the accuracy rate of the classifier as the performance measure. Some researchers argue that if the highest predictive accuracy is obtained by minimizing the classifier error rate and equalizing the measurement cost for all features, wrapper models are more suitable. A classifier is constructed with the aim of maximizing the predictive accuracy. The features utilized by the classifier are then selected as the optimal features. The wrapper models often apply meta-heuristic approaches to help in searching for the optimal feature subset. Although meta-heuristic approaches are slow, they produce the (near) optimal feature subset.

Neumaan et al. [33] observed that feature selection is a significant combinatorial optimization problem in supervised pattern classification, and proposed an additional regularization and embedded nonlinear feature selection method based on the difference of convex functions programming, which is a general framework for non-convex continuous optimization. Jack and Nandi [34] and Shon et al. [35], adopted GA in screening the features of a dataset. The selected subset of features is then fed into the SVM for classification testing. Zhang et al. [16] presented a GA-based method to obtain a beneficial subset of features for SVM in machine condition monitoring. However, their methods do not address the parameter values setting for SVM, and therefore may exclude the optimal result. Samanta et al. [36] developed a GA approach to change the RBF width parameter of SVM with feature selection. Nevertheless, since their approach only searches for the RBF width parameter of the SVM, it may exclude the optimal parameter setting. To the best of our knowledge, few studies have simultaneously considered the feature selection and the optimal parameter setting for SVM.

3. The proposed SA-SVM approach

Simulated annealing (SA) is a global search algorithm, which was first proposed by Metropolis et al. [37], and was later
popularized by Kirkpatrick et al. [38]. SA borrows its basic insight from metallurgy. Molecules in a metal gradually crystallize to a low-energy state as their temperature slowly decreases. All crystal grains will eventually attain the lowest energy state as long as the metal is heated to a sufficiently high initial temperature, and the cooling rate is slow enough. The algorithm proposed by Metropolis allows SA not only to improve the result during search iterations, but also to escape from local optimality. Moreover, the “cooling” process, which is analogous to the cooling of a metal, enables SA to converge gradually to the search outcomes to accomplish global optimality.

This study proposes adopting a Hide-and-Seek SA approach, first developed by Romeijn et al. [39,40], to search for the optimal parameter values for SVM and determine the appropriate feature subset simultaneously. Hide-and-Seek SA can handle continuous variables, enabling it to make feasible solutions within the constrained or bounded ranges converge to the optimal solution. Hide-and-Seek SA uses fewer parameters than traditional SA. The constraints and the objective function of problem can be non-differentiable, and the feasible region can be non-convex or even disconnected. The algorithm starts by taking any one of the feasible solutions as its current solution. It then picks a point from feasible regions as the solution in the next iteration of the search process by following a random vector. Hide-and-Seek SA differs from the traditional SA approach in that it generates the next solution in all feasible regions, whereas traditional SA generates the next solution in the neighboring regions. Hide-and-Seek SA requires the Metropolis acceptance rule and a suitable annealing procedure to pick the next solution. The current solution always has a possibility of converging to the (near) global optimum, regardless of how quickly the “temperature” falls to 0. The biggest difference between Hide-and-Seek SA and traditional SA is that the former performs an immediate annealing whenever a better solution is found, rather than waiting for the current annealing cycle to end. In the case of traditional SA, the algorithm decreases the jumping interval along with the slow decline in temperature in each annealing cycle. Convergence to the global optimum is attained by gradually reducing the jumping interval. Thus, the Hide-and-Seek SA can converge to the global optimum much more quickly than the traditional SA [39]. Fig. 1 illustrates the search concept of the Hide-and-Seek SA.

This study proposed an SA-based approach for parameter determination and feature selection in the SVM, termed SA-SVM. The objective function of the SA-SVM when searching for the optimal parameter values for SVM, is to maximize the accuracy rate in classifying the testing dataset. This is equivalent to an optimization problem seeking a maximum solution. In solving a maximization problem, if the next feasible solution raises the objective function value, then it is accepted as the current feasible solution, and as the starting point for a new search for the next feasible solution. If the next feasible solution has a lower accuracy rate of classification than the current solution, then the Metropolis acceptance rule is adopted to decide whether to replace the current solution with the next feasible solution.

Two basic decision variables, designated $C$ and $Y$, are necessary in the case of no feature selection. For the feature selection, if $n$ features are required to determine which features are selected, then $2 + n$ decision variables must be established. The value of any parameter is either 0 or 1. If the value of a variable is equal to 0, then its corresponding feature is not selected. Conversely, if the value of a variable is 1, its corresponding feature is selected. Fig. 2 depicts the solution representation.

As shown in Fig. 3, the procedure of the proposed SA-SVM approach can be briefly described as follows. First, the current temperature $T$ is set to $T_o$, which is set to be a very large number. An initial solution $X$ is randomly generated. In each iteration, taking $X$ as the starting point, a random vector is followed to select the next feasible solution $Y$. Let obj($X$) be the calculation of the objective function value (i.e., the calculation of classification accuracy rate of SVM given $X$) of $X$, and $\Delta E$ denote the difference between obj($X$) and obj($Y$); that is, $\Delta E = \text{obj}(Y) - \text{obj}(X)$. The probability of replacing $X$ with $Y$, where $X$ denotes the current solution, and $Y$ denotes the next solution, given that $\Delta E < 0$, is given by $e^{\Delta E/T}$. This probability is obtained by generating a random number $r \in [0,1]$, and replacing the solution $X$ with $Y$ if $r < e^{\Delta E/T}$. Meanwhile, if $\Delta E \geq 0$, then the probability of replacing $X$ with $Y$ is 1. If obj($X$) $\geq$ obj($X_{\text{opt}}$), then $X_{\text{opt}} = X$, and the temperature $T$ cools once. The SA-SVM terminates when the number of iterations $I_{\text{iter}}$ is reached. The optimal parameter values for SVM and the beneficial subset of features can be determined from $X_{\text{opt}}$.

The SA-SVM approach requires an initial feasible solution. If an initial feasible solution is generated in a random fashion, then many search iterations may be necessary to improve the solution, because of the poor initial solution. To avoid this
problem, this study adopts the concept of the greedy method. The value spaces of parameters $C$ and $U$ are divided into six regions, respectively. The boundaries of the regions are employed as possible solutions for determining the initial solution, producing 49 possible solutions to be tested as shown in Fig. 4. The best solution among these 49 is adopted as the initial feasible solution in the SA-SVM. If different solutions have identical classification accuracy rates, then the one with the lowest $C$ value is chosen. If the $C$ values are identical, then the one with the smaller $U$ value is chosen.

4. Experiment results

The proposed SA-SVM approach was implemented on a PC with an Intel Pentium IV 3.0 GHz CPU, 512 MB RAM, the Windows XP operating system and the Visual C++ 6.0 development environment. The following datasets, taken from the UCI machine learning repository, were used to evaluate the performance of the proposed SA-SVM approach: Bupa live, Cleveland heart, Pima, Ionosphere, Breast cancer, Boston housing, Australian, Sonar, Vowel, Vehicle and Glass [41]. Table 1 describes the characteristics of these datasets.

Scaling is employed to prevent feature values in greater numeric ranges from dominating those in smaller numeric ranges, as well as to avoid numerical difficulties during the calculation. Experimental results demonstrate that scaling the feature value facilitates SVM accuracy. The range of each feature value can generally be linearly scaled to the range $[0,1]$ using formula (10), where $V$ denotes the original value; $V_0$ denotes the scaled value; $\text{MAX}_a$ represents the upper bound of the feature value, and $\text{MIN}_a$ represents the lower bound of the feature value. This study scaled feature values to the range $[0,1]$:

$$V_0 = \frac{V - \text{MIN}_a}{\text{MAX}_a - \text{MIN}_a} \times 1$$

The $k$-fold method presented by Salzberg was applied in the experiments, with $k = 10$ [42,43]. Fig. 5 illustrates the system architectures. Because the proposed SA-SVM approach is non-deterministic, the solution obtained may not be equal for the same data. Thus, the proposed SA-SVM approach is executed five times for each dataset to calculate the average classification accuracy rate.

Through initial experiment without feature selection, the parameter values of the proposed SA-SVM approach were set as follows: $I_{\text{iter}} = 300; T_0$ was set to a sufficiently large number; $C = [1,35000]$, while $\gamma = [0.0001,32]$. With feature selection, the number of features selected for use can be obtained by SA-SVM approach. Since the chosen features may be different, the accuracy rate obtained is not the same, even though the $C$ and $\gamma$ value are unchanged. Therefore, the greedy method cannot be adopted in searching for the initial solution, and the initial solution has to be generated randomly. Because the solution space, as determined by the number of features, is larger than that of without feature selection, the number of search iterations must also be increased in order to explore more solutions. Therefore, $I_{\text{iter}}$ is increased to 1000.

The results obtained by the proposed SA-SVM approach without feature selection were compared with those of Fung

![Fig. 3. The procedure of proposed SA-SVM.](image)

![Fig. 4. The greedy method.](image)

<table>
<thead>
<tr>
<th>Dataset from the UCI repository</th>
<th>Number of classes</th>
<th>Number of instances</th>
<th>Number of features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bupa live</td>
<td>2</td>
<td>345</td>
<td>6</td>
</tr>
<tr>
<td>Cleveland heart</td>
<td>2</td>
<td>296</td>
<td>13</td>
</tr>
<tr>
<td>Pima</td>
<td>2</td>
<td>768</td>
<td>8</td>
</tr>
<tr>
<td>Ionosphere</td>
<td>2</td>
<td>351</td>
<td>34</td>
</tr>
<tr>
<td>Breast cancer</td>
<td>2</td>
<td>683</td>
<td>10</td>
</tr>
<tr>
<td>Boston housing</td>
<td>2</td>
<td>1012</td>
<td>13</td>
</tr>
<tr>
<td>Australian</td>
<td>2</td>
<td>653</td>
<td>15</td>
</tr>
<tr>
<td>Sonar</td>
<td>2</td>
<td>208</td>
<td>60</td>
</tr>
<tr>
<td>Vowel</td>
<td>11</td>
<td>528</td>
<td>10</td>
</tr>
<tr>
<td>Vehicle</td>
<td>4</td>
<td>846</td>
<td>18</td>
</tr>
<tr>
<td>Glass</td>
<td>6</td>
<td>214</td>
<td>9</td>
</tr>
</tbody>
</table>

Table 1. Dataset from the UCI repository.
and Mangasarian [44], who adopted several datasets from UCI. They tested datasets using Newton SVM (NSVM), conventional SVM and Lagrangian SVM (LSVM) without feature selection. Table 2 presents a comparison of their results. Five of the average accuracy rates of the proposed SA-SVM approach exceeded those obtained with Fung and Mangasarian. The proposed SA-SVM approach appeared to obtain the most appropriate C and U values, producing the highest classification accuracy rate across different datasets.

Furthermore, the results obtained by the proposed SA-SVM approach without feature selection were compared with those of the approach developed by Liao et al. [17], who applied three kernel functions, a Gaussian kernel, a polynomial kernel, and a sigmoid kernel, in the SVM to test several datasets from UCI. Table 3 compares these results, revealing that results obtained by the proposed SA-SVM approach were superior to those of Liao et al.

The performance of the proposed SA-SVM approach with and without feature selection was then compared with eleven datasets from UCI. The results of this experiment were compared with those of the grid search [13,21], as shown in Table 4. The proposed SA-SVM approach with and without feature selection performed better than grid search in all cases examined. The average classification accuracy rate for each dataset improved significantly after feature selection. This result reveals that good results are also obtainable when fewer features are included in the model, which means that some features are redundant or insignificant relative to particular classification problems. Clearly, the SA-SVM approach can simultaneously obtain optimal parameter values and find a subset of features without decreasing the SVM classification accuracy.

To verify the effectiveness of the proposed SA-SVM approach, a paired $t$-test on the average classification accuracy

![Fig. 5. System architecture of the proposed SA-based parameters determination and feature selection for SVM.](image)

Table 2

<table>
<thead>
<tr>
<th>Dataset</th>
<th>SA-SVM</th>
<th>NSVM</th>
<th>SVM</th>
<th>LSVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ionosphere</td>
<td>97.50</td>
<td>89.80</td>
<td>88.30</td>
<td>89.80</td>
</tr>
<tr>
<td>Bupa liver</td>
<td>80.81</td>
<td>70.20</td>
<td>69.30</td>
<td>70.20</td>
</tr>
<tr>
<td>Cleveland heart</td>
<td>87.97</td>
<td>86.30</td>
<td>85.90</td>
<td>86.30</td>
</tr>
<tr>
<td>Pima</td>
<td>80.19</td>
<td>77.00</td>
<td>77.10</td>
<td>77.00</td>
</tr>
<tr>
<td>Boston housing</td>
<td>99.90</td>
<td>86.60</td>
<td>85.80</td>
<td>86.60</td>
</tr>
</tbody>
</table>

* The highest classification accuracy rate among approaches.

Table 3

<table>
<thead>
<tr>
<th>Dataset</th>
<th>SA-SVM</th>
<th>Gaussian kernel</th>
<th>Polynomial kernel</th>
<th>Sigmoid kernel</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ionosphere</td>
<td>97.50</td>
<td>93.12</td>
<td>92.15</td>
<td>94.37</td>
</tr>
<tr>
<td>Bupa liver</td>
<td>80.81</td>
<td>71.35</td>
<td>72.85</td>
<td>73.17</td>
</tr>
<tr>
<td>Cleveland heart</td>
<td>87.97</td>
<td>85.11</td>
<td>84.67</td>
<td>85.17</td>
</tr>
<tr>
<td>Breast cancer</td>
<td>97.95</td>
<td>96.37</td>
<td>96.37</td>
<td>96.23</td>
</tr>
</tbody>
</table>

* The highest classification accuracy rate among approaches.

Table 4

<table>
<thead>
<tr>
<th>Dataset</th>
<th>(1) SA-SVM with feature selection</th>
<th>(2) SA-SVM without feature selection</th>
<th>(3) Grid search</th>
<th>$P$-value (1) vs. (3)</th>
<th>$P$-value (2) vs. (3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Breast cancer</td>
<td>99.38 a</td>
<td>97.95</td>
<td>96.64</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Cleveland heart</td>
<td>95.33 a</td>
<td>87.97</td>
<td>81.37</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Boston housing</td>
<td>100.00 a</td>
<td>99.90</td>
<td>99.80</td>
<td>&lt;0.001</td>
<td>0.100</td>
</tr>
<tr>
<td>Ionosphere</td>
<td>99.07 a</td>
<td>97.50</td>
<td>93.08</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Australian</td>
<td>92.19 a</td>
<td>88.34</td>
<td>84.54</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Bupa live</td>
<td>83.13 a</td>
<td>80.81</td>
<td>71.83</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Pima</td>
<td>82.22 a</td>
<td>80.19</td>
<td>76.69</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Sonar</td>
<td>95.99 a</td>
<td>91.85</td>
<td>87.90</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Vowel</td>
<td>100.00 a</td>
<td>99.27</td>
<td>98.91</td>
<td>&lt;0.001</td>
<td>0.013</td>
</tr>
<tr>
<td>Vehicle</td>
<td>90.14 a</td>
<td>88.76</td>
<td>84.28</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Glass</td>
<td>87.93 a</td>
<td>78.38</td>
<td>70.61</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

* The highest classification accuracy rate among approaches. Confidence level $a = 0.05$. 

---

features were selected for use in the SVM classification model and 10, respectively. Analytical results indicate that not all of 10 folds for the Bupa live and Vowel datasets. The original SVM with and without feature selection. Because the approach with feature selection achieved significantly better accuracy rate than grid search. The proposed SA-SVM approach without feature selection was not significantly better than of the grid approach without feature selection. The computational time required for the former is larger than that of SA-SVM without feature selection, the computation time required for the later. The search iteration $I_{iter}$ for SA-SVM with feature selection is larger than that of SA-SVM without feature selection, the computation time required for the former is larger than that of the latter.

To show how many features are selected with SA-SVM, and which features are selected. Table 6 shows the selected features of 10 folds for the Bupa live and Vowel datasets. The original numbers of features in the Bupa live and Vowel datasets are 6 and 10, respectively. Analytical results indicate that not all features were selected for use in the SVM classification model after the feature selection. Furthermore, feature selection increased the classification accuracy rates, as demonstrated in Table 5.

5. Conclusions and future research

This study proposes an SA-based approach, which can search for continuous decision variables, to optimize the parameter values for SVM, and to obtain a subset of beneficial features. This optimal subset of features is then adopted in both training and testing to obtain the optimal outcomes in classification. A comparison of the obtained results with those of other approaches demonstrates that the proposed SA-SVM approach improves the classification accuracy rates. After using feature selection in the experiment, the SA-SVM approach was applied to remove trivial or insignificant features and effectively find better parameter values. This process, in turn, enhances the overall outcomes of the classification.

Experimental results of this study were obtained with an RBF kernel function. However, other kernel parameters can also be optimized by the same approach. Experimental results obtained from UCI datasets, other public datasets and real world problems should be tested in the future to verify and extend this approach.

Acknowledgement

The authors would like to thank the National Science Council of the Republic of China, Taiwan for financially supporting this research under Contract No. NSC96-2416-H-211-002.

Table 5

<table>
<thead>
<tr>
<th>Dataset</th>
<th>With feature selection</th>
<th>Without feature selection</th>
<th>$P$-value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Number of original features</td>
<td>Number of selected features</td>
<td>Average accuracy rate (%)</td>
</tr>
<tr>
<td>Breast cancer</td>
<td>10</td>
<td>4.90 ± 1.233</td>
<td>99.38 ± 0.520</td>
</tr>
<tr>
<td>Cleveland heart</td>
<td>13</td>
<td>9.34 ± 2.626</td>
<td>93.33 ± 4.674</td>
</tr>
<tr>
<td>Boston housing</td>
<td>13</td>
<td>6.62 ± 0.835</td>
<td>100.00 ± 0.000</td>
</tr>
<tr>
<td>Ionosphere</td>
<td>34</td>
<td>22.36 ± 5.017</td>
<td>99.07 ± 0.730</td>
</tr>
<tr>
<td>Australian</td>
<td>15</td>
<td>10.60 ± 2.457</td>
<td>92.19 ± 3.229</td>
</tr>
<tr>
<td>Bupa live</td>
<td>6</td>
<td>4.50 ± 1.333</td>
<td>83.13 ± 6.502</td>
</tr>
<tr>
<td>Pima</td>
<td>8</td>
<td>5.86 ± 1.170</td>
<td>82.22 ± 3.547</td>
</tr>
<tr>
<td>Sonar</td>
<td>60</td>
<td>48.18 ± 5.534</td>
<td>95.99 ± 3.899</td>
</tr>
<tr>
<td>Vowel</td>
<td>10</td>
<td>6.40 ± 0.990</td>
<td>100.00 ± 0.000</td>
</tr>
<tr>
<td>Vehicle</td>
<td>18</td>
<td>15.56 ± 2.167</td>
<td>90.14 ± 2.210</td>
</tr>
<tr>
<td>Glass</td>
<td>9</td>
<td>5.76 ± 1.118</td>
<td>87.93 ± 7.309</td>
</tr>
</tbody>
</table>

Table 6

<table>
<thead>
<tr>
<th>No.</th>
<th>Selected features for Bupa live dataset</th>
<th>Selected features for Vowel dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1,2,3,4,5</td>
<td>1,2,3,5,6,8,9</td>
</tr>
<tr>
<td>2</td>
<td>1,2,3,4,5</td>
<td>1,2,3,5,6,9</td>
</tr>
<tr>
<td>3</td>
<td>2,3,4,5</td>
<td>1,2,4,5,6,9</td>
</tr>
<tr>
<td>4</td>
<td>1,2,3,4,5</td>
<td>1,2,3,4,7,8,9</td>
</tr>
<tr>
<td>5</td>
<td>1,2,3,4,5</td>
<td>1,2,3,4,5,6,7,8,9</td>
</tr>
<tr>
<td>6</td>
<td>1,2,3,4,5</td>
<td>1,2,4,5,6,7,8,9</td>
</tr>
<tr>
<td>7</td>
<td>1,2,3,4,5</td>
<td>1,3,4,5,6,7,8,9</td>
</tr>
<tr>
<td>8</td>
<td>1,2,3,4,5</td>
<td>1,2,3,4,6,8</td>
</tr>
<tr>
<td>9</td>
<td>1,3,4,5</td>
<td>1,3,4,5,6,7,8,9</td>
</tr>
<tr>
<td>10</td>
<td>1,2,3,4,5</td>
<td>1,2,3,4,5,6,8,9</td>
</tr>
</tbody>
</table>

Table 5 compares the classification accuracy rates of SA-SVM with and without feature selection. Because the $P$-values of all datasets were lower than 0.05, significant differences were observed in classification accuracy rates. Clearly, the SA-SVM approach with feature selection outperforms the SA-SVM approach without feature selection. The computational time required for SA-SVM with and without feature selection is also shown in Table 5. Since the search iteration $I_{iter}$ for SA-SVM with feature selection is larger than that of SA-SVM without feature selection, the computation time required for the former is larger than that of the latter.

Confidence level $\alpha = 0.05$. ± denotes the standard deviation.

References


