Available online at www.sciencedirect.com
ScienceDirect
Expert Systems with Applications 34 (2008) 754–763
Expert Systems with Applications
www.elsevier.com/locate/eswa

Feature selection for the SVM: An application to hypertension diagnosis

Chao-Ton Su [a,*], Chien-Hsin Yang [b]

[a] Department of Industrial Engineering and Engineering Management, National Tsing Hua University, 101, Section 2, Kuang Fu Road, Hsinchu, Taiwan
[b] Department of Industrial Engineering and Management, National Chiao Tung University, Hsinchu, Taiwan

Abstract

A support vector machine (SVM) is a novel classifier based on the statistical learning theory. To increase the performance of classification, the approach of SVM with kernel is usually used in classification tasks. In this study, we first attempted to investigate the performance of SVM with kernel. Several kernel functions, polynomial, RBF, summation, and multiplication were employed in the SVM and the feature selection approach developed [Hermes, L., & Buhmann, J. M. (2000). Feature selection for support vector machines. In *Proceedings of the international conference on pattern recognition (ICPR'00)* (Vol. 2, pp. 716–719)] was utilized to determine the important features. Then, a hypertension diagnosis case was implemented and 13 anthropometrical factors related to hypertension were selected. Implementation results show that the performance of combined kernel approach is better than the single kernel approach. Compared with backpropagation neural network method, SVM based method was found to have a better performance based on two epidemiological indices such as sensitivity and specificity.

Keywords: Support vector machine; Kernel; Polynomial; RBF; Classification; Hypertension; Diagnosis

1. Introduction

The support vector machine (SVM) is a promising classification technique proposed by Vapnik and his group at AT&T Bell Laboratories (Cortes & Vapnik, 1995). SVM is a good tool for the two classifications. It can separate the classes with a particular hyperplane which maximizes a quantity called the *margin*. The margin is the distance from a hyperplane separating the classes to the nearest point in the dataset. The advantage of maximum margin criterion is its robust characteristic against noise in data, and making a solution unique for linearly separable problems. In addition, it is important that the SVM with a theoretically strong support is based on the statistical learning theory framework. An important finding of the statistical learning theory is that the generalization error can be bound by the sum of the empirical error and term, which depends on the VC dimension that characterizes the complexity of the approximating function class (Pardo & Sberveglieri, 2005; Vapnik, 1998).

SVM has been extensively used as a classification tool with a great deal of success in a variety of area from object recognition (Oren, Papageorgiou, Sinha, Osuna, & Poggio, 1997) to classification of cancer morphologies (Mukherjee et al., 1999). In addition, it has also been successfully applied to a number of real-world problems such as handwritten characters and digit recognition (Cortes & Vapnik, 1995; Scholkopf, 1997; Vapnik, 1995), face detection (Osuna et al., 1997) and speaker identification (Schmidt, 1996).

Features are called attributes, properties, variables, or characteristics. Feature selection (the so-called variable selection) has become the focus of much research in the area of application for which datasets with tens or hundred of thousands of variables are available. Feature selection problems are found in many machine learning tasks including classification, regression, time series prediction, etc. An appropriate feature selection can enhance the effectiveness

* Corresponding author.
E-mail address: ctsu@mx.nthu.edu.tw (C.-T. Su).

0957-4174/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2006.10.010

and domain interpretability of an inference model. Liu and Motoda (1998) indicated that the effect of feature selection are (1) to improve performance (speed of learning, predictive accuracy, or simplicity of rules); (2) to visualize the data for model selection; and (3) to reduce dimensionality and remove noise. The universal algorithms of feature selection are often divided into three lines: filters, wrappers, and embedded (Guyon & Elisseeff, 2003; Kohavi & John, 1997). Both wrappers and filters have encountered some success with induction tasks, but they can be very computationally expensive for tasks with a larger number of variables. Although the embedded method has lower computational cost compared with the above, the exhaustive search is not good for dealing with the large features. Simply stated, these three methods may suffer from a block of wasting computational cost when variables are too large.

Although the approach of the SVM with kernel function is useful for classification, its performance must be improved, especially for complex data. This is particularly important for people who want to obtain a high level of accuracy in advanced areas such as precision engineering and medical diagnosis. In addition to accuracy, feature selection is another substantial issue for classification. Although feature selection offers many advantages, it may face the risks of accuracy decreasing or over-fitting. Thus, how to achieve/keep the expected classification performance and avoid the risks in feature selection is an important consideration.

In this study, we attempt to investigate the theory and application of classifier support vector machine. First the SVM with polynomial kernel, Gaussian Radius Base Function kernel (the so-called RBF kernel) and combined kernels are experimented in this study, respectively. Next, feature selection for SVM is also discussed. We apply the idea of Hermes and Buhmann (2000) to develop our method. This is a feature selection strategy which defines scores for available features on the basis of a single training run, and provides users ease in computation. Finally, a case study – hypertension diagnosis is implemented by the proposed approach and the relevant epidemiological discussion is also provided.

The remainder of this study is organized as follows. In Section 2, we describe the related researches including relevant kernel functions and feature selection approaches. The methodology utilized in this study is presented in Section 3. In Section 4, we illustrate the effectiveness of the proposed approach using various real-world datasets. Next, a case-study (hypertension diagnosis) is described in Section 5. Finally, we provide conclusions and some future works in Section 6.

2. Related works

2.1. Kernels and its properties

In the last few years, Kernel function from kernel methods (Aronszajn, 1950) have become one of the most popular approaches to learning from examples with many potential applications in science and engineering (Cristianini & Taylor, 2000; Scholkopf, 1997; Scholkopf, Smola, & Muller, 1998; Vapnik, 1998). Kernel functions of the form $k(\mathbf{x}_1, \mathbf{x}_2) = \varphi(\mathbf{x}_1) \cdot \varphi(\mathbf{x}_2)$, where $\cdot$ is an inner product and φ is in general, a nonlinear mapping from input space X onto feature space Z. In fact, the kernel function k is directly defined. φ and the feature space Z are simply derived from its definition. Kernel substitution of the inner product can be applied for generating SVM for classification based on margin maximization (Sanchez, 2003). In other words, SVM find a hyperplane in a space different from that of the input data $\mathbf{x}$. It is a hyperplane in a feature space induced by a kernel k. Through the kernel k, the hypothesis space is defined as a set of ''hyperplanes'' in the feature space induced by k. We can also say that the fundamental concept of the kernel method is deformation of the vector (lower) space itself to higher dimensional space. Often, a higher dimension is clearer to classify than a low dimension.

Definition of positive definite kernel (Scholkopf & Smola, 2002).

Let X be a nonempty set. A function k on $X \times X$ which for all $m \in N$ and all $x_1 \ldots, x_m \in \chi$ gives rise to a positive definite Gram matrix is called a positive definite (pd) kernel. Often, we shall refer to it simply as a kernel.

The use of a kernel function is an attractive computational shortcut. To use this approach, we first need to create a complicated feature space, and then work out what the inner product in that space would be, and finally find a direct method for computing that value in terms of the original inputs. In practice, the approach taken is to define a kernel function directly, hence implicitly defining the feature space. In this way, we avoid the feature space not only in the computation of inner products, but also in the design of the learning machine itself (Cristianini & Taylor, 2000). We argue that defining a kernel function for an input space is frequently more natural than creating a complicated feature space. However, we must first determine what properties of a function $k(\mathbf{x}, \mathbf{x}')$ are necessary to ensure that it is a kernel for some feature space (Cristianini & Taylor, 2000). Clearly, the function must be symmetric,

$$k(\mathbf{x}, \mathbf{x}') = \langle \phi(\mathbf{x}) \cdot \phi(\mathbf{x}') \rangle = \langle \phi(\mathbf{x}') \cdot \phi(\mathbf{x}) \rangle = k(\mathbf{x}', \mathbf{x}) \tag{2.1}$$

and satisfy the inequalities following the Cauchy–Schwarz inequality,

$$\begin{aligned} k(\mathbf{x}, \mathbf{x}')^2 &= \langle \phi(\mathbf{x}) \cdot \phi(\mathbf{x}') \rangle^2 \leqslant \|\phi(\mathbf{x})\|^2 \|\phi(\mathbf{x}')\|^2 \\ &= \langle \phi(\mathbf{x}) \cdot \phi(\mathbf{x}) \rangle \langle \phi(\mathbf{x}') \cdot \phi(\mathbf{x}') \rangle \\ &= k(\mathbf{x}, \mathbf{x})(\mathbf{x}', \mathbf{x}') \end{aligned} \tag{2.2}$$

However, these conditions are not sufficient to guarantee the existence of a feature space. In practice, it should provide a characterization of Mercer's theorem of when a function $k(\mathbf{x}, \mathbf{x}')$ is a kernel (Cristianini & Taylor, 2000).

The Mercer's theorem. (Mercer, 1909) *Let* $K(\mathbf{x}, \mathbf{x}')$ *be a continuous symmetric kernel that is defined in the closed interval* $\mathbf{a} \leqslant \mathbf{x} \leqslant \mathbf{b}$ *and likewise for* $\mathbf{x}'$. *The kernel* $K(\mathbf{x},\mathbf{x}')$ *can be expanded in the series*

$$K(\mathbf{x}, \mathbf{x}') = \sum_{i=1}^{\infty} \lambda_i \phi_i(\mathbf{x}) \phi_i(\mathbf{x}')$$

With positive coefficients, $\lambda_i > 0$ *for all i. For this expansion to be valid and for it to converge absolutely and uniformly, it is necessary and sufficient that the condition*

$$\int_b^a \int_b^a K(\mathbf{x}, \mathbf{x}') \psi(\mathbf{x}) \psi(\mathbf{x}') \, d\mathbf{x} \, d\mathbf{x}' \geqslant 0$$

holds for all $\psi(\cdot)$ *for which*

$$\int_b^a \psi^2(\mathbf{x}) \, d\mathbf{x} < \infty$$

The functions $\phi_i(\mathbf{x})$ are called eigenfunctions and the λ_i are called eigenvalues. The fact that all of the eigenvalues are positive means that the kernel $K(\mathbf{x}, \mathbf{x}')$ is a positive definite (Haykin, 1999).

2.2. Feature selection

Features are called attributes, properties, variables, or characteristics. Feature selection is a process by which a sample in the measurement space is described by a finite and usually smaller set of number classed features. The features become components of the pattern space. Feature selection is regarded as a procedure to determine which variables (attributes) are to be measured either first or last. Guyon and Elisseeff (2003) indicated that there are many potential benefits of feature selection: facilitating data visualization and data understanding, reducing the measurement and storage requirements, reducing training and utilization times, and defying the curse of dimensionality to improve prediction performance.

The universal algorithms of feature selection are often divided into three lines: wrappers, filters and embedded (Guyon & Elisseeff, 2003; Kohavi & John, 1997). Both wrappers and filters do this work by selected subsets of variables. Wrappers approach is one of the subset selection methods. It assesses subsets of variables according to their usefulness to a given predictor. The filters approach is a preprocessing step, independent of the choice of the predictor. Obviously, the exhaustive search in these two approaches can conceivably be performed, if the number of variables is not too large. However, the problem is known to be NP-hard (Amaldi & Kann, 1998) and the search becomes quickly computationally intractable. They may suffer from a block of wasting computational cost when variables are too large. As for embedded method, it is a machine learning algorithm that returns a model using a limited number of features. Similar to wrappers, it measures feature subset usefulness, but it has a lower computational cost. In addition to these algorithms of feature selection, variable ranking is as a principal or auxiliary selection mechanism because of its simplicity, scalability, and good empirical success. Several papers (Caruana & de Sa, 2003; Weston, Elisseff, Schoelkopf, & Tipping, 2003) in this issue use variable ranking as a baseline method. Furthermore, the information theoretic ranking criterion is also a common approach for variable classification (Dhillon, Mallea, & Kumar, 2003; Torkkola, 2003).

2.2.1. Wrappers approach

A wrappers model consists of two phases (Liu & Motoda, 1998):

Phase 1 – feature subset selection, which selects the best subset using a classifier's accuracy (on the training data) as a criterion.

Phase 2 – learning and testing, a classifier is learned from the training data with the best feature subset, and is tested on the test data.

The wrappers approach consists of using the prediction performance of a given learning machine to assess the relative usefulness of subsets of variable. When feature subsets are generated for each subset of feature, a classifier is generated from the data with chosen features. If the number of variables is not too large, an exhaustive search can conceivably be performed. However, the problem is to make it NP-hard (Amaldi & Kann, 1998).

Although wrappers approach are often criticized to be a "brute force" method which causes a required massive amount of computation, some of the researchers have different opinions. In this regard, Reunanen (2003) indicated that coarse search strategies may alleviate the problem of over-fitting. In addition, Greedy search strategies are good against over-fitting. *Forward selection* and *backward elimination* are two usages in these strategies. In forward selection, variables are progressively incorporated into larger and larger subsets, whereas in backward elimination, one starts with the set of all the variables and progressively eliminates the least promising ones. However, either forward selection or backward elimination requires time-consuming computation when the variable is very large (Kohavi & John, 1997).

2.2.2. Filters approach

Filters approach is built on the intrinsic properties of the data, not on a bias of particular classifier. The essence of filters is to seek the relevant features and to eliminate the irrelevant ones. According the classification guideline of Kohavi and John (1997), the preprocessing step of filters approach is to determine the independent of the choice of the predictor. Still, it may be optimal under certain independence of orthogonality assumption, with respect to a given predictor.

A filters model of feature selection also consists of two phases (Liu & Motoda, 1998):

Phase 1 – feature selection using measures such as information, distance, dependence, or consistency, and no classifier is engaged in this phase.

Phase 2 – is the same as in the wrappers model, a classifier is learned on the training data with the selected features, and tested on the test data.

In addition to the characteristic that is built on the intrinsic properties of the data, the filters approach has other characteristics as follows:

1. Measuring information gains, distance, dependence, or consistency is usually cheaper than measuring accuracy of a classifier, so a filters method can produce a subset faster, with other things being equal.
2. Because of the simplicity of the measures and low time complexity, a filters method can handle larger sized data than a classifier can; so in the case where a classifier cannot directly be learned from the large data, it can be used to reduce data dimensionality so that the classifier can be learned from the data with reduced dimensionality. However, there is a danger that the features selected by a filters model cannot allow a learning algorithm to fully exploit its bias.

Compared with wrappers, filters are faster. In addition, some filters provide a generic selection of variables that are not tuned for a given learning machine. It is an advantage to note that the filters approach can be used as a preprocessing step to reduce space dimensionality and over-fitting.

2.3. SVM-based feature selection and variable ranking

The SVM-RFE (recursive feature elimination) algorithm has been recently proposed by Guyon, Weston, Barnhill, and Vapnik (2002) for selecting genes that are relevant for a cancer classification problem. The object of this method is to find a subset of size r among d variables ($r < d$), which maximizes the performance of the predictor. This method is a backward sequential selection approach. One starts with all the features and removes one feature at a time until only r features are left. The removed variable is the one whose removal minimizes the variation of $\|\mathbf{w}\|^2$. In other words, this method is similar to neural networks in the sense that the ranking criterion is the sensitivity of $\|\mathbf{w}\|^2$ with respect to a variable (Rakotomamonjy, 2003).

In addition to SVM-RFE, several algorithms for feature selection based on SVM have been developed. For instance, users can minimize the $R^2\mathbf{w}^2$ bound to find the best variable subset (Weston et al., 2001). For this criterion, the method differs from that of Rakotomamonjy's (2003) in the variable space search algorithm. In fact, instead of using a greedy algorithm, they use a gradient descent to minimize the bound with respect to a scaling vector associated to variables.

In fact, variable ranking or feature ranking is a necessary process for feature selection. Hence, the ranking criterion always affects the result of feature selection. Most researches agree in connection with the bound L error (Weston et al., 2001) and used the radius/margin bound for feature selection using a gradient descent algorithm.

2.4. L–J method

In 2000, Lothar Hermes and Joachim M. Buhmann proposed a strategy to rank individual components according to their influence on the decision hyperplane. The L–J method is an appellation gathered by authors' first name. The L–J method is an approach to select suitable subsets of features to replace original attributes. It is a selection approach that defines scores for the available features at training. The strategy ranks the features according to their influence on the decision hyperplane. The influence of the jth feature is evaluated by the angle between $\nabla f(\mathbf{x})$ and e_j. In other words, this method provides a mechanism of feature ranking by α_j that is the angle between $\nabla f(\mathbf{x})$ and e_j. The α_j is defined as follows:

$$\alpha_j(\mathbf{x}_i) = \min_{\beta\in\{0,1\}} \left\{ \beta\pi + (-1)^{\beta} \arccos\left(\frac{(\nabla f(\mathbf{x}))^{\mathrm{T}} e_j}{\|\nabla f(\mathbf{x})\|} \right) \right\} \tag{2.3}$$

Values $\alpha_j(\mathbf{x}_i) \approx \pi$ represent that the feature j has a weak influence on the assignment $f(\mathbf{x})$ of $\mathbf{x}$. Small values (which are close to zero) indicate that feature j is important. In addition to α_j, Hermes and Buhmann provide another index $\tilde{\alpha}_j$ to rank the features. In this part, they added the number of points in order to permit some error on support vectors. They included all vectors within a δ-region around the support vectors, choosing the points x_i satisfies $|f(\mathbf{x}_i) - 1| < \delta$. Let I_δ be the indices set of all training vectors which match the condition abbreviate $\alpha_j(\mathbf{x}_i)$ by α_{ij}. To evaluate the influence of the jth component of all support vectors, Hermes and Buhmann define $\tilde{\alpha}_j \in [0,1]$ which averages the angles α_{ij} over $i \in I_\varepsilon$:

$$\tilde{\alpha}_j = 1 - \frac{2}{\pi} \cdot \frac{\sum_{i\in I_\varepsilon} \alpha_{ij}}{|I_\varepsilon|} \tag{2.4}$$

In summary of the above, users can use an index $\tilde{\alpha}_j$ to rank the features, supposing that the number of margin vectors is quite small. Users can drop the unimportant feature if its $\tilde{\alpha}_j$ is smaller (Hermes & Buhmann, 2000).

Compared with wrappers and filters approaches, L–J method is a feature selection strategy which defines scores for available features on the basis of a single training run, and is easy for users to compute. In addition, this method may avoid general problems such as over-fitting and decrease in accuracy when the feature selection is implemented.

3. Methodology

3.1. Selected kernel functions

Several investigations (Wang, Li, & Xu, 2004; Yao et al., 2005) indicate that the kernel functions are useful for

classification. For example, some researches show that SVM with polynomial kernel provide a good performance for prediction and classification (Wang et al., 2004); some researches indicate that SVM with RBF kernel has stronger ability for classification (Dong, Cao, & Lee, 2005; Hammer & Gersmann, 2003; Lukas et al., 2004; Yao et al., 2005). These kernel functions were selected in this study to generate SVM for classification. They are described as follows:

(1) Polynomial Kernel

$$k_P = k(\mathbf{x}_i, \mathbf{x}') = (a + b\langle \mathbf{x}_i \cdot \mathbf{x}' \rangle)^d \tag{3.1}$$

where a and b are constants and its degree is d. For this kernel, there are $\binom{n+d-1}{d}$ distinct feature, being all monomials up to and including degree d and the number of attributes n in an instance of the data set. A special case of this kernel $a = 0$ and $b = d = 1$ forms a linear kernel.

(2) Gaussian Radius Base Function (RBF) kernel

$$k_G = k(\mathbf{x}_i, \mathbf{x}') = \exp\left(-\frac{1}{\gamma}\|\mathbf{x}_i - \mathbf{x}'\|^2\right) \tag{3.2}$$

where γ is kernel width and it is $2\sigma^2$. The kernel width common to all the kernels is specified a priori by the user.

(3) Combined kernel
As the polynomial kernel function owns the advantage of changing the degree d in the feature space (see Eq. (3.1)) and Gaussian RBF kernel is in itself a normalized kernel (see Eq. (3.2)), the kernel k_P and k_G can be employed to develop the combined kernels, k_{P+G} and $k_{P\cdot G}$.

First, to simplify the tasks of the classification process, parameter a was ignored and parameter b was set at 1. We can rewrite Eqs. (3.1) and (3.2) as follows:

$$k_P(\mathbf{x}_i, \mathbf{x}) = (\langle \mathbf{x}_i \cdot \mathbf{x} \rangle)^d \tag{3.3}$$

where d is its degree and is adjustable.

$$k_G(\mathbf{x}_i, \mathbf{x}) = \exp\left(-\frac{1}{2\gamma}\|\mathbf{x}_i - \mathbf{x}\|^2\right) \tag{3.4}$$

where γ is kernel width and γ is adjustable.

Consequently, the kernel function k_{P+G} is defined as follows:

$$k_{P+G}(\mathbf{x}_i, \mathbf{x}) = \left((\langle \mathbf{x}_i \cdot \mathbf{x} \rangle)^d + \exp\left(-\frac{1}{2\gamma}\|\mathbf{x}_i - \mathbf{x}\|^2\right)\right) \tag{3.5}$$

The kernel function $k_{P\cdot G}$ is defined as follows.

$$k_{P\cdot G}(\mathbf{x}_i, \mathbf{x}) = \left((\langle \mathbf{x}_i \cdot \mathbf{x} \rangle)^d . \exp\left(-\frac{1}{2\gamma}\|\mathbf{x}_i - \mathbf{x}\|^2\right)\right) \tag{3.6}$$

As a result, the SVM decision functions using kernels, k_{P+G} and $k_{P\cdot G}$, can be rewritten in the following:

$$f(\mathbf{x}) = \text{sign}\left(\sum_{i=1}^{M_s} \alpha_i y_i k_{P+G}(\mathbf{x}_i, \mathbf{x}) + b\right) \tag{3.7}$$

$$f(\mathbf{x}) = \text{sign}\left(\sum_{i=1}^{M_s} \alpha_i y_i k_{P\cdot G}(\mathbf{x}_i, \mathbf{x}) + b\right) \tag{3.8}$$

Either the kernel k_{P+G} or $k_{P\cdot G}$ is satisfied with the characterization of Mercer's theorem.

3.2. Feature selection for the SVM using the L–J Method

We apply the L–J method to our selected kernels. The feature selection procedure is described as follows:

Step 1: Construct a support vector machine structure $g(\mathbf{x})$ with a given training set by using complete data components.
Step 2: Compute the gradient of $g(\mathbf{x})$ at position $\mathbf{x}$, and the so-called $\nabla g(\mathbf{x})$.
Step 3: Select the kernel functions into the gradient of $g(\mathbf{x})$ at position $\mathbf{x}$.
Step 4: Implement the task of project of the unit vector ε_j on $\nabla g(\mathbf{x})$.
Step 5: Estimate the importance of separate feature components to $g(\mathbf{x})$ by the influence.
If the margin has quite small support vectors, we should proceed with step 6. Otherwise, the procedure of feature selection ends here.
Step 6: Rank the features by $\tilde{\alpha}_j = 1 - \frac{2}{\pi} \cdot \frac{\sum_{i \in I_\varepsilon} \alpha_{ij}}{|I_\varepsilon|}$, $\tilde{\alpha}_j \in [0, 1]$.

4. Illustration

We have discussed the underlying principle in the design of kernel-based support vector machine. However, to make the algorithm practical for various problems, we have to describe by several datasets. Therefore, we devote this section to a description of the implementation details.

4.1. Datasets

A total of three datasets, hyperlipidemia, liver disease and renal disease were collected from the Department of Health Examination from those seeking an annual physical health check-up at Chang Gung Memorial Hospital in Tao-Yuan, Taiwan. Thirty-two anthropometrical data were measured by the whole body scanner employing the independent variables. The dependent variable was that subjects suffer or do not suffer from the disease in each set of the disease data. In addition to the medical data, nine data sets from the UCI repository (Blake & Merz, 1998) were used. These datasets were census income, shuttle, mushroom, letter, ionosphere, vehicle silhouettes, spambase, vowel, and sonar.

Among the 12 datasets, seven were considered as the larger ones, as each contained more than 5000 samples (Oyang, Hwang, Ou, Chen, & Chen, 2005). The remaining five datasets were considered as the smaller ones. Before our experiment, we had worked some data preprocess including eliminating of missing value and normalizing the row data. The meta-data including the number of features, classes, cases and feature style, are represented in Tables 1 and 2.

Table 1
The meta-data of the larger data sets

	Features	Data style	Class	Cases
Shuttle	9	c	7	14,500
Census_income	14	c, d	2	32,561
Mushroom	22	d	2	8124
Letter	16	c	26	15,000
HyperLipidemia	33	c	2	6000
Liver disease	33	c	2	6000
Renal disease	33	c	2	6000

c: Continuous; d: discrete.

Table 2
The meta-data of the smaller data sets

	Features	Data style	Class	Cases
Sonar	60	c	2	208
Ionosphere	34	c	2	351
Vehicle	18	c	4	846
Spambase	57	c	2	4601
Vowel	13	c, d	11	990

c: Continuous; d: discrete.

4.2. Implementation results

In this study, four kernels, polynomial (k_P), RBF (k_G), summation (k_{P+G}), and multiplication ($k_{P\cdot G}$) were employed to construct the SVM models. Firstly, the full SVM models were established. In order to simplify the process of classification, the parameter a was set at 0, and b was set at 1 in the polynomial kernel. We only changed the degree d. As for the RBF kernel, it remained in its original form, i.e. kernel width γ could be changed. In our experiment, parameter d was set between 2 and 10. Parameter γ was set at 10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}, respectively. Secondly, the kernels were applied to L–J method for feature selection. The optimal parameter settings were employed in SVM modeling for L–J feature selection. Note that as the data were of two classes, we can use the original SVM technique. If the data are more than two classes, the SVM model is worked by multi-class, one against one process.

The average accuracies of the classification for the seven larger data sets and five smaller data sets are shown in Tables 3 and 4. Their standard deviations are listed in the brackets. The two tables indicated that the combined kernel $k_{P\cdot G}$ (polynomial multiplies RBF kernel) has better performance than the other approaches. After feature selection (from 75% to 25%), the kernel $k_{P\cdot G}$ also showed a better performance both in larger and smaller data. In the larger data, the combined kernel k_{P+G} showed a better performance than the polynomial and RBF kernel. The result in the smaller data was the same as that in the larger one. Furthermore, the kernel $k_{P\cdot G}$ almost had the lowest standard deviation among the four approaches in the larger

Table 3
The accuracy of feature selection for the SVM using the L–J method (larger data sets)

Dataset	Kernel							
	Full (100%)				Reduced (75%)			
	k_P	k_G	k_{P+G}	$k_{P\cdot G}$	k_P	k_G	k_{P+G}	$k_{P\cdot G}$
Shuttle	98.2	95.5	96.8	95.5	98.4	98.2	99.8	98.8
Census_income	75	71.5	74	75.8	69.6	71.6	71.6	76
Mushroom	100	99.73	99.73	100	96.8	98.2	98	98.8
Letter	86.75	90.75	87.5	90.75	81.25	81	83	83
HyperLipidemia	51.92	68.06	69	69	50.83	67.83	70	69.5
Liver disease	73.5	73.5	77	77	71.13	71.13	75.5	76.5
Renal disease	78.58	78.5	82	82	76.25	72.13	81.75	80.38
Average	80.56	82.51	83.72	84.29	77.75	80.01	82.81	83.28
(Standard deviation)	(16.49)	(12.65)	(11.55)	(11.39)	(16.53)	(13.06)	(12)	(11.4)
	Reduced (50%)				Reduced (25%)			
Shuttle	91.5	91.41	91.5	95.5	81.4	82.6	81.4	82.8
Census_income	68.5	69.5	78	72.5	72	73.6	72	74
Mushroom	98.82	99.73	99.73	99.73	100	99.89	100	99.92
Letter	67.5	73	73.5	79	46.5	52.5	52	57.5
HyperLipidemia	50.25	67	68.5	68.86	48.5	62.5	61.83	62.38
Liver disease	72	72	73	75	62.5	71.25	60.25	71.75
Renal disease	75	70	78	78.5	69.4	63.94	74.75	72.88
Average	74.8	77.52	80.32	81.3	68.61	72.33	71.75	74.46
(Standard deviation)	(16.12)	(12.71)	(11.2)	(11.74)	(18.67)	(15.43)	(15.92)	(13.91)

Table 4
The accuracy of feature selection for the SVM using the L–J method (smaller data sets)

Dataset	Kernel							
	Full				Reduced (75%)			
	k_P	k_G	k_{P+G}	$k_{P\cdot G}$	k_P	k_G	k_{P+G}	$k_{P\cdot G}$
Sonar	88.1	88.1	92.86	95.23	85.71	88.1	88.1	95.23
Ionosphere	84.29	84.29	91.43	91.43	74.28	75.71	88.57	91.43
Vehicle	79.3	82.84	79.3	85.8	78.85	77.51	78.25	83.25
Spambase	94.5	94.5	94.5	95	92.5	94	91.5	94
Vowel	90.91	98.98	99.49	99.49	88.43	94.47	92.13	95
Average	87.42	89.74	91.52	93.39	83.95	85.96	87.71	91.78
(Standard deviation)	(5.88)	(6.86)	(7.48)	(5.11)	(7.34)	(8.92)	(5.57)	(5)
	Reduced (50%)				Reduced (25%)			
Sonar	76.19	78.57	85.71	88.1	78.57	76.2	85.71	85.71
Ionosphere	71.42	78.57	85.71	90	72.86	77.14	82.56	88.57
Vehicle	78.1	72.19	78.1	80.47	69.29	69.41	72.92	79.51
Spambase	90.8	93.4	92	94.4	87.5	86.8	88	89.75
Vowel	84.48	93.93	89.39	94.94	44.44	39.71	44.95	45.51
Average	80.2	83.33	86.18	89.58	70.53	69.85	74.83	77.81
(Standard deviation)	(7.55)	(9.79)	(5.24)	(5.86)	(16.13)	(17.95)	(17.66)	(18.49)

data. In the smaller data set, the kernel $k_{P\cdot G}$ performed well.

4.3. Discussion

In our experiment, it seems that the kernel $k_{P\cdot G}$ is superior to the other approaches. The reason for this may be because the kernel $k_{P\cdot G}$ has some functions by changing degree and adjusting width at the same time, which seems to increase the classification performance. However, the influences of these functions are not significant in the other kernels. Next, we show the results with 100%, 75%, 50%, and 25% features after feature selection by twelve data sets. Obviously, the performance of classification decrease follows the number of features reduced. It is interesting to note that the more the number of classes there was, the larger the decreasing percentage of classification was noted.

As for feature selection process, many investigators consider that the most straightforward idea is to use a leave-one-out procedure or a cross-validation set to assess the generalization error with regard to the number of features and choose the number of attributes which minimizes the test error. It was deemed to be unfavorable for the computation. Compared with this process, L–J method just selects variables by index influence (α_j) and avoids this predicament. However, kernel selection in L–J method plays an important role and greatly affects the performance of classification.

5. Application

In this section, a real-case from medical diagnosis is presented. We will show that the L–J method using SVM with the selected kernel function can be applied to reduce the attributes by a hypertension diagnosis via anthropometrical data. Further explanation and discussion will likewise be provided.

5.1. Problem description

Hypertension is a major disease and is a significant cause of death all over the world. The relevant researches show that the cardiovascular disease is an important risk causing hypertension (Jeppesen, Hein, Suadicani, & Gyntelberg, 2000; Mykkane et al., 1997). As defined by the National High Blood Pressure Education Program (NHPEP), hypertension can be summarized as shown in Table 5.

Recently, syndrome X has been investigated more and more (Chen et al., 2000a). In fact, there is a significant relationship between body size and syndrome X (Lin, Chiou, Weng, Tsai, & Liu, 2002). Hence, it is feasible to explore the relation between hypertension and body size via syndrome X indirectly.

In the past, the human body size is measured by the worker with his experience. The drawback of this approach is that it is not accurate and time consuming. Hence, 3D anthropometrical measure prevails in this area. There are many advantages related to this measure, such as

Table 5
Classification of blood pressure for adults aged 18 and older (NHPEP, 2002)

Category	Systolic (mm Hg)		Diastolic (mm Hg)
Optimal	<120	and	<80
Normal	<130	and	<85
High-normal	130–139	or	85–89
Hypertension			
Stage 1	140–159	or	90–99
Stage 2	160–179	or	100–109
Stage 3	⩾180	or	⩾110

convenience and time saving. In addition, this technique can be employed to medical diagnosis.

A memorial hospital in Taiwan has dealt with disease diagnosis for several years. Recently, they provide a whole body 3D scanning technique for patients in their Department of Health Examination. The purpose of the techniques is to explore the relationship between the body size and some chronic disease by some three-dimensional body surface anthropometrical scanning data. In fact, too many anthropometrical data collected from this equipment and as listed on the diagnosis make the more difficulty of explanation for the physicians. Hence, how to reduce the unimportant or noisy features is necessary. Here, we implement a hypertension diagnosis using the proposed approach for feature selection.

5.2. Implementation

A total of 32 anthropometrical items were collected from the hospital's 3D whole body data bank. These data included height, weight, head circumference, breast circumference, waist circumference, hip circumference, left upper arm circumference, right upper arm circumference, left forearm circumference, right forearm circumference, right thigh circumference, left thigh circumference, right leg circumference, left leg circumference, breast width, waist width, hip width, breast profile area, hip profile area, volume of head, surface area of head, volume of trunk, surface area of trunk, volume of left arm, surface area of left arm, volume of right arm, surface area of right arm, volume of left leg, surface of left leg, volume of right leg, and surface area of right leg. In addition to these measurements, the subjects' age and gender were collected as well. Furthermore, the patients who suffered from hypertension were noted.

A total of 6000 datasets were selected randomly from the original database via data pre-processing. Four kernel functions including k_P, k_G, k_{P+G}, and $k_{P\cdot G}$ were employed to construct the SVM models. The relevant parameter of polynomial kernel d was set between 2 and 10 and the parameter of RBF kernel γ was set at 10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}, respectively. The result shows that the combined kernel, $k_{G\cdot P}$ has a better performance than the other approaches. Next, these kernels were applied to L–J method for feature selection. In addition to accuracy, the important feature will be selected.

By using the kernel function to L–J method, we selected the important features using the influence index α_j. For instance, when the $k_{P\cdot G}$ was employed, a total of 13 anthropometrical attributes were selected, including age, weight, waist circumference, right thigh circumference, left thigh circumference, right leg circumference, left leg circumference, breast width, volume of trunk, surface area of trunk, volume of left arm, volume of right arm and volume of right leg.

5.3. Comparisons

In order to explain the effectiveness of the proposed approach, the collected data were also analyzed by the backpropagation neural network. In this study, *Professional II Plus* software was used to perform the computation. The result showed that the structure 33-12-1 provided a better performance when the learning rate was 0.15 and the momentum was 0.75. After that, we pruned the network based on index P_i. P_i is the priority index of the input nodes in backpropagation neural network structure. It can be defined as follows (Su, Hsu, & Tsai, 2002):

$$P_i = \sum_{i=1}^{n} \sum_{j=1}^{m} \sum_{k=1}^{s} |W_{ij} \times V_{jk}| \tag{5.1}$$

where W_{ij} is the weight between the ith input node and the jth hidden node, V_{jk} is the weight between the jth hidden node and the kth output node, and P_i is the sum of absolute multiplication values of the weights W_{ij} and V_{jk}.

Based on the definition, the input nodes with $P_i < 1.65$ from the network were removed. Finally, 14 anthropometrical factors were determined, including weight, waist circumference, left forearm circumference, right forearm circumference, right thigh circumference, left thigh circumference, right leg circumference, breast width, hip profile area, volume of trunk, surface area of trunk, volume of left arm, volume of right arm and volume of left leg.

SVM based and neural network based models were assessed by the epidemiology based indices, namely, sensitivity and specificity. In addition, accuracy was employed to evaluate their performance. We selected 13 and 14 features from SVM based and neural network based method, respectively. As shown in Table 6, we found that the neural network based model is inferior to SVM based approach in terms of the three indices. We also found that the sensitivity was decreased; however, the specificity was increased in

Table 6
A comparison of performance of feature selection

Methods	Number of features	Sensitivity		Specificity		Accuracy	
		Full	Reduced	Full	Reduced	Full	Reduced
Neural network	14	0.4478	0.3963	0.7186	0.7289	0.6883	0.6233
SVM (L–J based)							
k_P	13	0.4689	0.4655	0.7356	0.7639	0.7033	0.67
k_G	13	0.4929	0.4805	0.7368	0.7693	0.7083	0.6767
k_{P+G}	13	0.5143	0.4830	0.7396	0.7699	0.7133	0.6783
$k_{P\cdot G}$	13	0.5373	0.4987	0.743	0.7790	0.72	0.69

both neural network based and SVM based method. This means that the ability of testing TN (true negative) improved but it deteriorated on test TP (true positive). Indeed, this is not favorable for diagnosing. Fortunately, the decrease range observed was small. Also the specificity increase would be beneficial in minimizing the cost of developing new medicines for hypertension. Furthermore, the accuracy of SVM based model is better than neural network based approach. SVM based method also showed fewer decreasing percentage after feature selection.

5.4. Discussion

After the feature selection, the common anthropometric factors including weight, waist circumference, right thigh circumference, left thigh circumference, breast width, volume of trunk, surface area of trunk and volume of right arm were collected by SVM based and neural based method, separately. A number of researches are concerned about X syndrome or cardiovascular disease, and the indices BMI and WHR are often employed in their investigations (Chen, Lin, Tsai, & Chou, 2000b; Kim et al., 2001). However, some researches indicate that BMI and WHR without the significant position could be disapproving. For instance, it makes the BMI imprecise because the pure height and/or weight measure varies significantly across ethnic groups. In clinical research, most of the syndromes and cardiovascular diseases such as hypertension are derived from abnormal diet behavior aside from environmental and psychological factors. The behaviors, in particular, preferring greasy food, have been found to bring many changes to the human body size. The findings of this study also specify that larger trunk and weight are significant factors that cause hypertension. In addition, similar to other studies previously conducted, our study considered waist circumference as a predictor of hypertension (Lin et al., 2002). Moreover, as for thigh circumference, the accumulation of fat, especially viscera fat was noted to result in a wider thigh circumference. In clinical practice, our method presents a good way to predict cardiovascular disease, such as hypertension.

6. Conclusions

In this study, we represent a SVM based procedure of feature selection by using L–J method. First, we used four kernels including polynomial kernel, RBF kernel, multiplication kernel ($k_{G\cdot P}$) and summation kernel (k_{G+P}) to construct the SVM model. Our experiments show that the multiplication kernel has the best performance both in larger and smaller data sets; summation kernel is next and RBF kernel is last. Next, these kernels were applied to L–J method for feature selection. The result shows that the L–J with multiplication kernel generally has a better performance than other approaches. Finally, a case study on hypertension diagnosis was investigated in this paper. We selected 13 anthropometrical factors that need to be considered by people suffering from hypertension. Except for the indices BMI and WHR, we also found that some anthropometrical factors like wider thigh circumference will bring the risk of hypertension. The result provides a new guide in preventive medicine for hypertension diagnosis. In addition to our proposed approaches, the backpropagation neural network was employed for feature selection and compared with our approach. Three indices, such as sensitivity, specificity and accuracy were used to evaluate the performance of these two methods. Implementation results show that our method is better than the neural network based approach. After feature selection, sensitivity and accuracy are reduced and specificity is increased. Although a decreased sensitivity is not good at diagnosing, fortunately, the decreased range is small. Naturally, the ability of explanation is decreased due to fewer features noted in the prediction model. Next, the ability of testing TN is good at saving the cost of developing new medicines for hypertension. In summary of the above, SVM with combined kernel functions by using L–J method seems to be a feasible approach for feature selection.

As compared with neural network based method, L–J approach with combined kernel functions was observed to have a better performance. In addition, L–J method has the advantage on the basis of a single training run and is easier to compute for feature selection as compared with other SVM based methods. However, the computation speed is relatively slow when the kernel functions are complicated. Hence, this subject is worth investigating in the future.

Acknowledgement

This work was supported in part by National Science Council of Taiwan (Grant No. 95-2221-E-007-181-MY3).

References

Amaldi, E., & Kann, V. (1998). On the approximation of minimizing non zero variables or unsatisfied relations in linear systems. *Theoretical Computer Science, 209*, 237–260.

Aronszajn, N. (1950). Theory of reproducing kernels. *Transactions of the American Mathematical Society, 68*(3), 337–404.

Blake, C. L., & Merz, C. J. (1998). UCI repository of machine learning databases. Dept. Infor. Comput. Sci., Univ. California, Irvine, CA.

Caruana, R., & de Sa, V. (2003). Benefitting from the variables that variable selection discards. *Journal of Machine Learning Research, 3*, 1245–1264.

Chen, W., Bao, W., Begum, S., Elkasabany, A., Srinivasan, S. R., & Berenson, G. S. (2000a). Age-related patterns of the clustering of cardiovascular risk variables of syndrome X from childhood to young adulthood in population made up of black and white subjects: the Bagalusa Heart Study. *Diabetes, 49*, 1042–1048.

Chen, C. H., Lin, K. C., Tsai, S. T., & Chou, P. (2000b). Different association of hypertension and insulin-related metabolic syndrome between man and women in 8437 nondiabetic Chinese. *American Journal of Hypertension, 13*(7), 846–853.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning, 20*(3), 273–297.

Cristianini, N., & Taylor, J. S. (2000). *An introduction to SVMs and other kernel-based learning methods*. Cambridge University Press.

Dhillon, I., Mallea, S., & Kumar, R. (2003). A divisive information-theoretic feature clustering algorithm for text classification. *Journal of Machine Learning Research, 3*, 1265–1287.

Dong, B., Cao, C., & Lee, S. E. (2005). Applying support vector machines to predict building energy consumption in tropical region. *Energy and Buildings, 37*, 545–553.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research, 3*, 1157–1182.

Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning, 46*(1–3), 389–442.

Hammer, B., & Gersmann, K. (2003). A note on the universal approximation capability of support vector machines. *Neural Processing Letters, 17*, 43–53.

Haykin, S. (1999). *Neural networks: a comprehensive foundation* (2nd ed.). New Jersey: Prentice Hall.

Hermes, L., & Buhmann, J. M. (2000). Feature selection for support vector machines. In *Proceedings of the international conference on pattern recognition (ICPR'00)* (Vol. 2, pp. 716–719).

Jeppesen, J., Hein, H. O., Suadicani, P., & Gyntelberg, F. (2000). High triglycerides and low HDL cholesterol and blood pressure and risk of ischemic heart disease. *Hypertension, 36*, 226–232.

Kim, Y. I., Kim, C. H., Choi, C. S., Chung, Y. E., Lee, M. S., Lee, S. I., et al. (2001). Microalbuminuria is associated with the insulin resistance syndrome independent of hypertension and type 2 diabetes in the Korean population. *Diabetes Research and Clinical Practice, 52*, 145–152.

Kohavi, R., & John, G. (1997). Wapper for feature selection. *Artificial Intelligence, 97*(1-2), 273–324.

Lin, J. D., Chiou, W. K., Weng, H. F., Tsai, Y. H., & Liu, T. H. (2002). Comparison of three-dimensional anthropometric body surface scanning to waist-hip ratio and body mass index in correlation with metabolic risk factors. *Journal of Clinical Epidemiology, 55*, 757–766.

Liu, H., & Motoda, H. (1998). *Feature selection for knowledge discovery and data mining*. Norwell, MA: Kluwer Academic.

Lukas, L., Devos, A., Suykens, J. A. K., Vanhamme, L., Howe, F. A., Majos, C., et al. (2004). Brain tumor classification based on long echo proton MRS signals. *Artificial Intelligence in Medicine, 31*, 73–89.

Mercer, J. (1909). Functions of positive and negative type and their connection with the theory of integral equations. *Philosophical transactions of the Royal Society, London, A, 209*, 415–446.

Mukherjee, S., Tamayo, T., Slonim, D., Verri, A., Golub, T., Mesirov, J., et al. (1999). Support vector machine classification of microarray data. AI Memo 1677, Massachuetts Institute of Technology.

Mykkane, L., Haffner, S. M., Ronnemaa, T., Bennemaa, T., Bergman, R. N., & Laakso, M. (1997). Low insulin sensitive is associated with clustering of cardiovascular disease risk factors. *American Journal of Epidemiology, 146*, 315–321.

Oren, M., Papageorgiou, C., Sinha, P., Osuna, E., & Poggio, T. (1997). Pedestrain detection using wavelet templates. In *Proceedings of the computer vision and pattern recognition* (pp. 193–199). Puerto Rico, June 16–20.

Osuna E., Freund R., & Girosi F. (1997). Training support vector machines: an application to face detection. In *Proceedings of the computer vision and pattern recognition '97* (pp. 130–136).

Oyang, Y. J., Hwang, S. C., Ou, Y. Y., Chen, C. Y., & Chen, Z. W. (2005). Data classification with radial basis function networks based on a novel kernel density estimation algorithm. *IEEE Transactions on Neural Networks, 16*(1), 225–236.

Pardo, M., & Sberveglieri, G. (2005). Classification of electronic nose data with support vector machines. *Sensors and Actuators B, 107*, 730–737.

Rakotomamonjy, A. (2003). Variable selection using SVM-based criteria. *Journal of Machine Learning Research, 3*, 1357–1370.

Reunanen, J. (2003). Overfitting in making comparisons between variable selection methods. *Journal of Machine Learning Research, 3*, 1371–1382.

Sanchez, V. D. A. (2003). Advanced support vector machines and kernel methods. *Neurocomputing, 55*, 5–20.

Schmidt, M. (1996). Identifying speakers with support vector networks. In *Interface '96 Proceedings*. Sydney.

Scholkopf, B. (1997). Support Vector Learning, PhD theiss, R. Oldenbourg Verlag, Munich.

Scholkopf, B., & Smola, A. J. (2002). *Learning with kernels: support vector machines, regularization, optimization, and beyond*. Cambridge, Mass: MIT Press.

Scholkopf, B., Smola, A. J., & Muller, K. R. (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation, 10*(5), 1299–1319.

Su, C. T., Hsu, H. H., & Tsai, C. H. (2002). Knowledge mining from trained neural networks. *Journal of Computer Information Systems, 42*(4), 61–70.

Torkkola, K. (2003). Feature extraction by non-parametric mutual information maximization. *Journal of Machine Learning Research, 3*, 1415–1438.

Vapnik, V. (1995). *The nature of statistical learning theory*. New York: Springer-Verlag.

Vapnik, V. (1998). *Statistical learning theory*. New York: Wiely.

Wang, M. L., Li, W. J., & Xu, W. B. (2004). Support vector machines for prediction of peptidyl prolyl *cis*/*trans* isomerization. *Journal of Peptide Research, 63*, 23–28.

Weston, J., Elisseff, A., Schoelkopf, B., & Tipping, M. (2003). Use of the zero norm with linear models and kernel methods. *Journal of Machine Learning Research, 3*, 1439–1461.

Weston, J., Mukherjee, S., Chapelle, O., Pontil, M., Poggio, T., & Vapnik, V. (2001). Feature selection for svms. In *Proceedings of the advances in neural information processing systems* (Vol. 13).

Yao, X. J., Panaye, A., Doucet, J. P., Chen, H. F., Zhang, R. S., Fan, B. T., et al. (2005). Comparative classification study of toxicity mechanisms using support vector machines and radial basis function neural networks. *Analytica Chimica Acta, 535*, 259–273.