

國立交通大學

工業工程與管理學系

博士論文

核心函數為基礎的支持向量分類器：理論與應用



Kernel-Based SVM: Theory and Application

研究生：楊健炘

指導教授：蘇朝墩

陳文智

中華民國九十五年十一月

核心函數為基礎的支持向量分類器：理論與應用

Kernel-Based SVM: Theory and Application

研究生：楊健焯

Student : Yang, Chien-Hsin

指導教授：蘇朝墩

Advisor : Su, Chao-Ton

陳文智

Chen, Wen-Chih



A Dissertation

Submitted to Department of Industrial Engineering and Management

College of Management

National Chiao Tung University

in partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

in Industrial Engineering and Management

2006, 11

Hsinchu, Taiwan, Republic of China

核心函數為基礎的支持向量分類器：理論與應用

研究生：楊健焯

指導教授：蘇朝墩
陳文智

國立交通大學

工業工程與管理學系

在兩種類別的分類作業上，支持向量分類器（support vector machine, SVM）是一個好的資料探勘工具，使用者可以透過簡單的計算，以超平面（hyperplane）和邊界（boundary）完成資料分類。為了解決許多非線性的問題，數學家們建議使用 SVM 結合核心函數（kernel function）的方式來分析。這樣的做法雖有利於分類，但是在處理大量和複雜資料上，SVM 仍受到限制。實際上，不同領域的龐大資料庫中往往隱含許多資訊和知識，而特徵選取是擷取資訊和知識的其中一種程序。因此如何能夠快速地刪減不重要的屬性，進而獲得正確的特徵屬性，是一重要的議題。

本研究首先提出一個新的以 SVM 為基礎的分類方法，透過常用的核心函數（Polynomial 和 RBF）建構分類器，並提出參數設定與核心函數選擇的指引。接著，將所提出的 SVM 分類器與 Hermes 和 Buhmman (2000) 所提的屬性選擇方法作結合，構建一特徵選取程序。最後，本研究以高血壓檢測為例，透過所提出之程序進行一個案研究：包含模型建構與刪減不重要的屬性，並與倒傳遞類神經網路、決策樹、粗略集合等方法比較。結果顯示，無論是正確率和精確度的評估上，所提出方法的績效優於其他方法。

關鍵詞：特徵選取、支持向量分類器、核心函數、倒傳遞類神經網路、決策樹、粗略集合、高血壓。

Kernel-Based SVM: Theory and Application

Student: Chien-Hsin Yang

Adviser: Chao-Ton Su
Wen-Chih Chen

Department of Industrial Engineering and Management
National Chiao Tung University

SVM is a good data mining tool for the two class classification. The classification task is worked by the hyperplane and boundary. In order to solve nonlinear classification problems, mathematicians provided related kernel functions to deal with them. Although the approach of the SVM with kernel function is useful for classification, its performance must be improved especially for some data, such as large and complex data. In practice, large data sets often connote information and knowledge in many fields. Feature selection is one of the procedures to gather the information. Thus, it is an important issue that how to reduce attributes and select correct features in this field.

In this dissertation, we attempt to investigate the theory and application of classifier support vector machine. We hope to increase the performance of classification through the new classifier. Two popular kernel functions, polynomial kernel and Gaussian Radius Base Function kernel are used. The relevant strategies, including the setting of parameters and selecting of kernels will be provided. Next, we apply Hermes and Buhmann's (2000) idea to our proposed new classifier. Also, we construct a procedure of feature selection based on it.

Finally, we demonstrate a case study of feature selection for hypertension detection. This study will construct prediction model by the developed approaches.

Implementation results show that the performance of the developed approach is better than those of backpropagation neural network, decision tree (DT) and rough sets (RS) methods based on accuracy and specificity. In addition, this paper provides some medical discussions of the position of anthropometric factors after feature selection.

Keywords: Feature selection, support vector machine, kernel function, backpropagation neural network, decision tree, rough sets, hypertension.



CONTENTS

摘要	i
ABSTRACT	ii
CONTENTS	iv
LIST OF TABLES	v
LIST OF FIGURES	vi
CHAPTER 1 INTRODUCTION	1
1.1 Overview	1
1.2 Motivations	2
1.3 Objectives	3
1.4 Organization	4
CHAPTER 2 RELATED WORKS	6
2.1 Support Vector Machine	6
2.2 Kernel Function	8
2.3 Properties of the Kernel Function	11
2.4 Feature Selection	12
2.4.1 Wrappers Approach	15
2.4.2 Filters Approach	16
2.4.3 Information Theoretic Ranking Criterion	17
2.4.4 Embedded Approach	18
2.5 L-J Method	18
2.6 Data Complexity	20
CHAPTER 3 PROPOSED APPROACHES	22
3.1 SVM with Combined Kernel Functions	22
3.2 Feature Selection for the SVM by Using the L-J Method	25
CHAPTER 4 ILLUSTRATION	27
4.1 Data Sets	27
4.2 Implementation Results	28
4.3 Discussions	34
CHAPTER 5 A CASE STUDY: HYPERTENSION DETECTION	39
5.1 Problem Description	39

5.2	Implementation	40
5.3	Comparisons	41
5.4	Discussion	45
CHAPTER 6 CONCLUSIONS		48
6.1	Summary	48
6.2	Further Research	49
REFERENCES		50



LIST OF TABLES

Table 4.1	Data sets used in this study	30
Table 4.2	Comparison of classification accuracy with the larger and smaller data sets	31
Table 4.3	The accuracy of feature selection for the SVM using the L-J method (larger data sets)	32
Table 4.4	The accuracy of feature selection for the SVM using the L-J method (smaller data sets)	33
Table 4.5	The strategies of parameter setting of polynomial and RBF kernels	37
Table 5.1	Classification of blood pressure for adults aged 18 and older	39
Table 5.2	A comparison of performance of feature selection	45



LIST OF FIGURES

Figure 1.1	Research framework	4
Figure 2.1	Hyperplane with the maximal margin by a linear SVM	6
Figure 2.2	Original space (input space)	9
Figure 2.3	Transformed space (feature space)	9
Figure 2.4	A wrappers model of feature selection	15
Figure 2.5	A filter model of feature selection	17
Figure 4.1	The relationship between parameters and accuracy for the larger data set	34
Figure 4.2	The relationship between parameters and accuracy for the smaller data set	35



誌謝

本論文得以完成，我要感謝恩師蘇朝墩教授，在這五十二個月的時間裡，從他身上學得論文撰寫的技巧，以及為人處事的道理。在論文口試上，我要感謝駱景堯教授、洪瑞雲教授、邱文科教授、陳穆臻教授、陳文智教授們的指教，得以讓整體論文更為完善。另外，能與研究室的成員：志華、俊欽、隆昇、敬森、宗銘、家任、宇翔相互切磋、鼓勵，謝謝你們！

這些年來，家人的支持是我求學階段的最大動力，在此我要將完成博士學位的榮耀和喜悅獻給我親愛的父母和家人，謝謝他們讓我沒有後顧之憂，完成學業。

最後，我要向曾經幫助我的貴人，表達最深的謝意！



CHAPTER 1

INTRODUCTION

1.1 Overview

Nowadays, Knowledge Discovery in Databases (KDD) is concerned with extracting useful information from databases (Fayyad *et al.*, 1996). Data mining is a set of techniques used in an automated approach to exhaustively explore and bring to the surface complex relationships in very large datasets (Liu and Motoda, 1998). Two objectives in the data mining areas are gathering the model accuracy and important information. Recently, many algorithms or tools were developed to construct a more precise model to explain the relationship between input and output variables in the data mining areas. Support Vector Machine (SVM) is one of the tools spring up among the classification applications.

The SVM is a promising classification technique proposed by Vapnik and his group at AT&T Bell Laboratories (Cortes and Vapnik, 1995). It is a universal approximator that can be used to learn a variety of representations from training samples and regression tasks. It has also been successfully applied to a number of real-world problems such as handwritten characters and digit recognition (Schoelkopf, 1997; Cortes and Vapnik, 1995; LeCun *et al.*, 1995; Vapnik, 1995), face detection (Osuna, 1997) and speaker identification (Schmidt, 1996). SVM is a good tool for the two class classification. 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. This maximum margin criterion has the advantage of being robust 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 Vapnik Chervonenkis (VC) dimension, which characterizes the complexity of the approximating function class (Vapnik, 1998; Pardo and Sberveglieri, 2005).

However, not all of the cases are linear and separable for classification. In fact data that is both vague and overlapping is common in many cases. Thus, many interactions occur, particularly at the input spaces. Based on available studies (Oyang *et al.*, 2005; Scholkopf *et al.*, 1995), it seems that the original SVM did not perform well for these cases. In order to solve this problem, mathematicians provided related kernel functions to deal with nonlinear classification problems on the basis of the above limitations (Muller *et al.*, 2001). There are several types of kernels being used for all kinds of problems. Each kernel may be suitable for some of the problems. For instance, some well-known special problems, such as text classification (Joachims, 2000) and DNA problems (Yeang *et al.*, 2001) are reported to be classified more correctly using the linear kernel.

1.2 Motivations

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. Owing to the available kernel function own the advantages by themselves, it seems that users can get a better accuracy on classification tasks using a combination of different kernel functions. Therefore, further study is highly desired.

In addition to accuracy, feature selection is another substantial issue for classification. Feature selection can avoid any unnecessary computation for

classification process. Limiting the number of feature can sometime be helpful because it cuts down the model capacity and thus reduces the risk of over-fitting. However, we should note that reducing the features always bears the danger of reducing the expected classification performance. Thus, how to achieve/keep the expected classification performance and to avoid decreasing the accuracy after feature selection is an important problem. In the disease diagnosis, diagnosticians or physicians need to discovery some information or knowledge from the data set based on fewer features or subsets. Owing to most of the data sets with a large number of variables. They need a good tool/algorithm to implement feature selection quickly and precisely.

1.3 Objectives

In this study, we attempt to investigate the theory and application of classifier SVM. First, a kernel-based SVM will be developed. We hope to increase the performance of classification through the new classifier. Two popular kernel functions, polynomial kernel and Gaussian Radius Base Function kernel (so-called RBF kernel) transform the row data from low dimension to high dimension. SVM with single and combined kernels are experimented in this study respectively. Furthermore, the relevant strategies, including the setting of parameters and selecting of kernels will be tested by using the data sets collected from the UCI data bank and the hospital. Next, feature selection for SVM will be discussed. We investigate the feature selection problem of our proposed new classifier. We apply Hermes and Buhmann's (2000) idea to our method. Finally, we demonstrate a case study of feature selection for hypertension detection. This study constructs a prediction model for hypertension using anthropometrical body surface scanning data. In addition to our proposed approaches, some feature selection methods: backpropagation neural network,

decision tree and rough sets are in the benchmark and used to predict hypertension. The relevant indices on epidemiology such as sensitivity and specificity are used to evaluate the position of anthropometric factors after feature selection. Finally, technical and medical discussions are provided. In summary, the framework of the kernel-based SVM discussed in this dissertation is showed in figure 1.1.

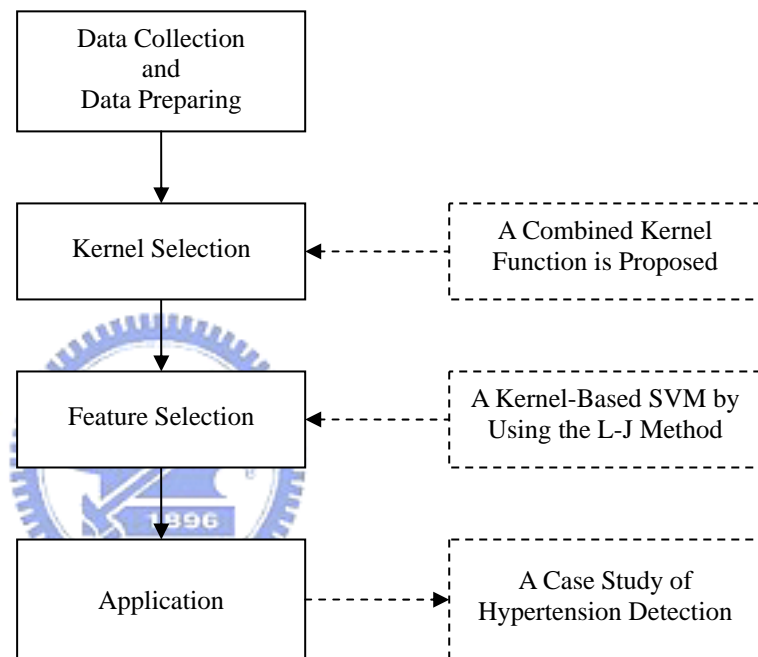


Figure 1.1 Research framework

1.4 Organization

This remainder of this dissertation is organized as follows. Chapter 2 describes related research, including a brief introduction to SVM, relevant kernel functions, and feature selection approaches. In addition, an indicator to evaluate the kernel selection criterion, mess level, is briefly introduced in this chapter. Our proposed approaches including combined kernel method and feature selection method are described in Chapter 3. In the Chapter 4, we illustrate proposed approach's effectiveness using

various real-world datasets. A case-study (hypertension detection) is described in Chapter 5. Finally, the conclusions and the direction of further research are given in Chapter 6.



CHAPTER 2

RELATED WORKS

2.1 Support Vector Machine

SVM recently gained popularity in the learning community. In its simplest linear form, an SVM is a hyperplane that separates a set of positive elements from a set of negative elements with maximum interclass distance, so-called the margin. Figure 2.1 shows such a hyperplane with the associated margin.

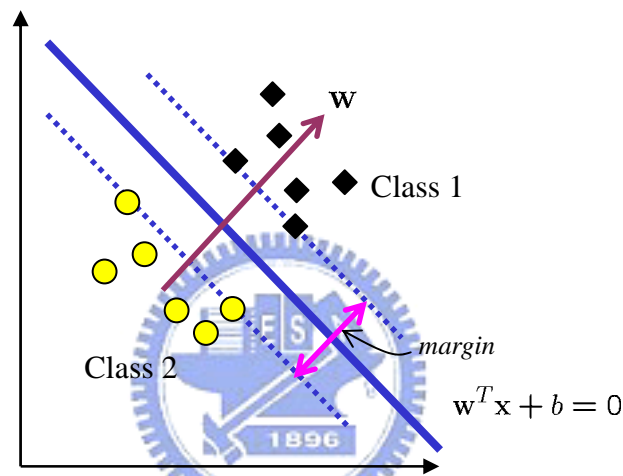


Figure 2.1 Hyperplane with the maximal margin by a linear SVM.

The formula for output of linear SVM is

$$u = \mathbf{w}^T \cdot \mathbf{x}_i + b \quad (2.1)$$

where \mathbf{w} is normal vector (weight coefficient vector), \mathbf{x}_i is input vector and b is bias term. Based on that, we can get the class u which is 1 or -1. The distance between a training vector \mathbf{x}_i and the boundary, called margin, is expressed as follows:

$$\frac{|\mathbf{w}^T \cdot \mathbf{x}_i + b|}{\|\mathbf{w}\|}$$

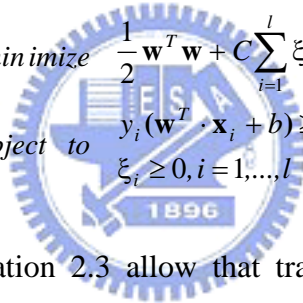
According to original theory by Vapnik (1995), we want to find the margin m that $\mathbf{w}^T \cdot \mathbf{x}_i + b > 1$ and $\mathbf{w}^T \cdot \mathbf{x}_i + b < -1$ to separate the elements which are in positive or

negative class. In order to compute the boundary, we need to maximize m , i.e. minimize $\frac{1}{2}\|\mathbf{w}\|^2$. Consequently, we can draw the optimization formulation as

$$\begin{aligned} & \text{minimize } \mathbf{w}^T \mathbf{w} \\ & \text{subject to } y_i (\mathbf{w}^T \cdot \mathbf{x}_i + b) \geq 1 \end{aligned} \quad (2.2)$$

where x_i is the i^{th} training element and $y_i \in \{-1, 1\}$ is the correct output of the SVM for the i^{th} training element. Note that the hyperplane is determined by the training elements x_i on the margin, so-called *support vectors*. As seen in figure 1, they are “physically supporting” the final hyperplane $\mathbf{w}^T \mathbf{x} + b = 0$.

However, practically some of the problems are with the nonseparable patterns. Hence Cortes and Vapnik (1995) introduced a penalty term $C \sum_{i=1}^l \xi_i$ in the objective function and allowed training errors:



$$\begin{aligned} & \text{minimize } \frac{1}{2} \mathbf{w}^T \mathbf{w} + C \sum_{i=1}^l \xi_i \\ & \text{subject to } y_i (\mathbf{w}^T \cdot \mathbf{x}_i + b) \geq 1 - \xi_i \\ & \quad \xi_i \geq 0, i = 1, \dots, l \end{aligned} \quad (2.3)$$

That is, constraints equation 2.3 allow that training data may not be on the correct side of the separating hyperplane $\mathbf{w}^T \mathbf{x} + b = 0$ while we minimize the training error $C \sum_{i=1}^l \xi_i$ in the objective function. Hence if the penalty parameter C is large enough and the data is linear separable, problem equation 2.3 goes back to equation 2.2 as all ξ_i will be zero (Lin, 2001). In addition to linear cases, most of the problems are with nonlinear patterns. In order to solve the nonlinear cases, the kernel function was often used in these cases. As for nonlinear cases, the plane is found by solving the following constrained quadratic programming problem:

$$\text{maximize } \mathbf{W}(\alpha) = \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j k(\mathbf{x}, \mathbf{x}') \quad (2.4)$$

under the constraints $\sum_{i=1}^n \alpha_i y_i = 0$ and $0 \leq \alpha_i \leq C$ for $i = 1, 2, \dots, n$ where $x_i \in R$ are the training sample vectors, and $y_i \in \{-1, +1\}$ the corresponding class labels

(Cortes and Vapnik, 1995). The kernel function $k(\mathbf{x}, \mathbf{x}')$ will be detailed in the next section.

2.2 Kernel Function

Kernel function from kernel methods (Aronszajn, 1950) have become in the last few years one of the most popular approaches to learning from examples with many potential applications in science and engineering (Cristianini and Taylor, 2000; Vapnik, 1998; Scholkopf, 1997; Scholkopf *et al.*, 1998; Roth and Steinhage, 2000). Kernel functions of the form $k(\mathbf{x}_1, \mathbf{x}_2) = \varphi(\mathbf{x}_1) \cdot \varphi(\mathbf{x}_2)$, \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 (the kernel defines an inner product in that space). Through the kernel k the hypothesis space is defined as a set of “hyperplanes” in the feature space induced by k . We also can say that the fundamental concept of the kernel method is deformation of the vector (lower) space itself to higher dimensional space. Often the higher dimension is clearer to classify than low dimension. See a linearly non-separable example presented in the follows. A total of six points x_1, x_2, x_3, x_4, x_5 and x_6 are vector \mathbf{x} showed in figure 2.2. Significantly, six points are nonlinearly separable in two dimensions. Therefore, we need a kernel function Φ to transform to a higher dimensional space to solve it.

$\Phi: R^2 \rightarrow R^3$, i.e. $\Phi(x) = \begin{bmatrix} x_1^2 \\ \sqrt{2}x_1x_2 \\ x_2^2 \end{bmatrix}$. After kernel function transformation, we can get

vector \mathbf{t} , $\mathbf{t} \in R^3$, which is a linearly separable case showed in figure 2.3.

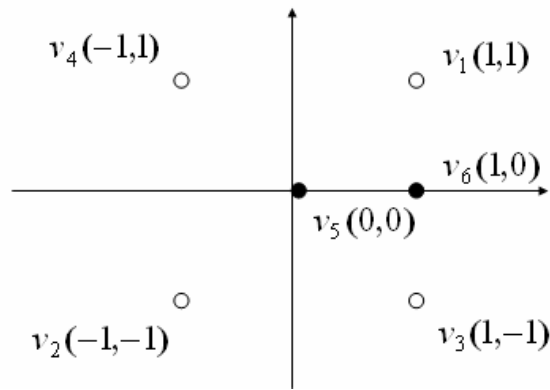


Figure 2.2 Original space (input space).

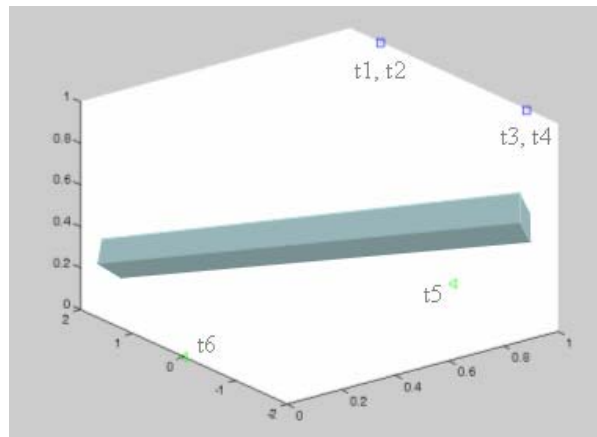


Figure 2.3 Transformed space (feature space).

In order to solve more complex problems, the kernel function was used to generate SVM for classification is a popular approach described as above. The kernel function is usually presented as $k(\mathbf{x}, \mathbf{x}')$ and introduced for satisfying the distance is defined in transformed space and it has a relationship to the distance in the original space. Several examples of such kernel functions are known, as follows:

(1) Polynomial kernel

$$k(\mathbf{x}_i, \mathbf{x}') = (a + b \langle \mathbf{x}_i \cdot \mathbf{x}' \rangle)^d \quad (2.5)$$

where a and b are constants. Its degree is d . For this kernel there are $\binom{n+d-1}{d}$ distinct features, being all the 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. Some simple cases using linear kernel are good enough for SVM-based classification (Zien *et al.*, 2000).

(2) RBF kernel

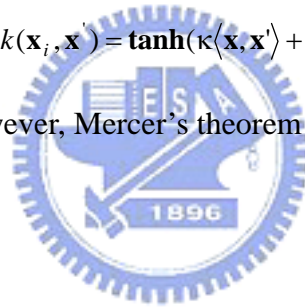
$$k(\mathbf{x}_i, \mathbf{x}') = \exp\left(-\frac{1}{\gamma} \|\mathbf{x}_i - \mathbf{x}'\|^2\right) \quad (2.6)$$

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) Signomid kernel

$$k(\mathbf{x}_i, \mathbf{x}') = \tanh(\kappa \langle \mathbf{x}, \mathbf{x}' \rangle + \vartheta) \quad (2.7)$$

where $\kappa > 0$ and $\vartheta < 0$. However, Mercer's theorem is satisfied only for some value of κ and ϑ .



Summary of the statement concerning with kernels, we need to make a mathematical definition of kernel function in proposition 2.2.1 (Scholkopf and Smola, 2002).

Proposition 2.2.1 Definition positive definite kernel

Let X be a nonempty set. A function k on $X \times X$ which for all $m \in \mathbb{N}$ and all $x_1, \dots, x_m \in X$ 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 selection of the kernel function is very important for the performance of the classifier (Papadopoulos *et al.*, 2005). Wahba (2000) has suggested using the kernel

function to increase the dimensionality, and then it is easier to classify the data in the higher dimensional space by hyperplane. Although it is well known that the choice of kernel affects SVM's performance, only a few kernels have been used in practice, because it is difficult to choose proper turning of parameters (Dudoit *et al.*, 2002). As for the parameters selection for SVM, it is an other important issue to SVM's performance. Scholkopf and Smola (2002) indicated that it is suitable of smaller C for SVM classification. They also present that both the kernel parameters and the SVM parameter (value of C) are often chosen using cross validation. Zhu and Zhang (2004) considered that too larger parameters will bring very time consuming. Still, the parameters selection lacks the consistency for researchers.

2.3 Properties of the Kernel Function

The use of a kernel function is an attractive computational short-cut. If we wish to use this approach, there appears to be a need to first create a complicated feature space, and then work out what the inner product in that space would be, and finally find a direct method of 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 and Taylor, 2000). We will argue that defining a kernel function for in input space is frequently more natural than creating a complicated feature space. Before we can follow this route, 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 and Taylor, 2000). Clearly, the function must be symmetric,

$$k(\mathbf{x}, \mathbf{x}') = \langle \phi(\mathbf{x}) \cdot \phi(\mathbf{x}') \rangle = \langle \phi(\mathbf{x}') \cdot \phi(\mathbf{x}) \rangle = k(\mathbf{x}', \mathbf{x}) \quad (2.8)$$

and satisfy the inequalities that follow from the Cauchy-Schwarz inequality,

$$\begin{aligned}
k(\mathbf{x}, \mathbf{x}')^2 &= \langle \phi(\mathbf{x}) \cdot \phi(\mathbf{x}') \rangle^2 \leq \|\phi(\mathbf{x})\|^2 \|\phi(\mathbf{x}')\|^2 \\
&= \langle \phi(\mathbf{x}) \cdot \phi(\mathbf{x}) \rangle \langle \phi(\mathbf{x}') \cdot \phi(\mathbf{x}') \rangle = k(\mathbf{x}, \mathbf{x})(\mathbf{x}', \mathbf{x}')
\end{aligned} \tag{2.9}$$

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 and Taylor, 2000). The Mercer's theorem can be formally stated as (Mercer, 1908; Courant and Hilbert, 1970):

Proposition 2.3.1 *Let $K(\mathbf{x}, \mathbf{x}')$ be a continuous symmetric kernel that is defined in the closed interval $\mathbf{a} \leq \mathbf{x} \leq \mathbf{b}$ 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

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

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 positive definite (Haykin, 1999).

2.4 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 numbers classed features, say x_1, x_2, \dots, x_n . The features become components of the pattern space. The feature selection is regarded as

a procedure to determine that which variables (attributes) are to be measured first or last. Liu and Montoda (1998) defined that feature selection is a process that chooses an optimal subset of features according to certain criterion. Feature selection may be multistage process to enhance the accuracy or performance of classification (Chiang, 2002).

Feature selection (so-called variable selection) has become the focus of much research in 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 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. 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, defying the curse of dimensionality to improve prediction performance.

Liu and Motoda (1998) provided a detailed survey and overview of the existing methods for feature selection. They suggest a feature selection process that includes four parts: feature generation, feature evaluation, stopping criteria and testing. In addition to the classic evaluation measures (accuracy, information, distance, and dependence) used for removing irrelevant features, they provided consistency measures (inconsistency rate) to determine a minimum set of relevant features.

Two methods, feature selection and feature extraction were usually used to do the work of dimensionality reduction of data sets (Jain *et al.*, 2000). As described as

above, feature selection is the way of selecting the sub-features in the measurement space. But, feature extraction method determines an appropriate subspace of dimensionality m (either in a linear or a nonlinear way) in the original feature space of dimensionality d . It should be noted that the features or attributes are not changed in the feature selection process; however, new attributes were established after feature extraction. It is obvious to find that the feature selection is superior to feature extraction in the interdisciplinary applications cause of the consistency of attribute.

The universal algorithms of feature selection are often divided along three lines: wrappers, filters and embedded (Kohavi and John., 1997; Guyon and Elisseeff, 2003). Both wrappers and filters do this work by select 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. Still, under certain dependence or orthogonality assumptions, it may be optimal with respect to a give predictor. Obviously, an exhaustive search can conceivably be performed, if the number of variables is not too large. But, the problem is known to be NP-hard (Amaldi and 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, the disadvantages are as similar as theirs. 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 (Bekkerman *et al.*, 2003; Caruana and de Sa, 2003; Weston *et al.*, 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 (Bekkerman *et al.*, 2003; Dhillon *et al.*, 2003; Torkkola, 2003).

2.4.1 Wrappers Approach

Regardless of any approach what used in the feature selection tasks, people always want to provide the classifier with the data of better quality and to improve the classification performance. If they can select the relevant features and remove noise, they can achieve their objective possibly. However, the essence of the wrappers approach owns the function to do that.

A wrappers model (see figure 2.4) consists of two phases (Liu and 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 tested on the test data.

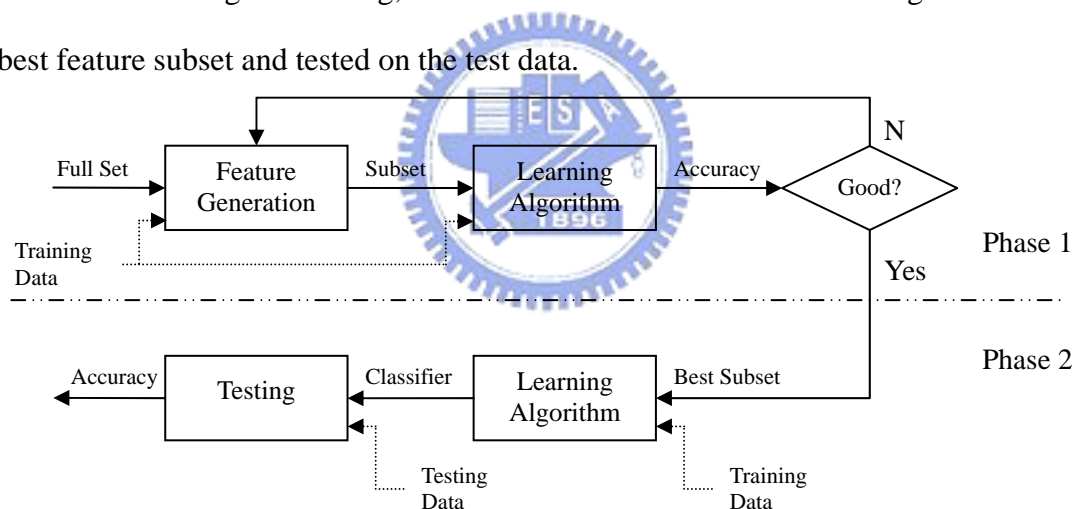


Figure 2.4 A wrappers model of feature selection (Liu and Motoda, 1998).

The wrappers approach consists in using the prediction performance of 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 be NP-hard (Amaldi and Kann, 1998).

Although wrappers approach often criticized that it seems to be a “brute force”

method cause of required massive amount of computation, some of the researchers own different opinions. Such as Reunanen (2003) indicated that coarse search strategies may alleviate the problem of overfitting. In addition, Greedy search strategies are good at for against overfitting. *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 variables and progressively eliminates the least promising ones. However, either forward selection or backward elimination, it seems that they do not avoid the time-consuming computation when the number of variables is very large. To solve this limitation, researchers often use heuristic learning method like Naïve Basesian Classifiers or Decision Tree Induction (Kohavi and John, 1997).



2.4.2 Filters Approach

Filters approach 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 Kohavi and John (1997) classification guideline, the preprocessing step of filters approach is to determine the independence of the choice of the predictor. Still under certain independence of orthogonality assumption, it may be optimal with respect to a given predictor.

A filters model of feature selection (see figure 2.5) also consists of two phases (Liu and 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.

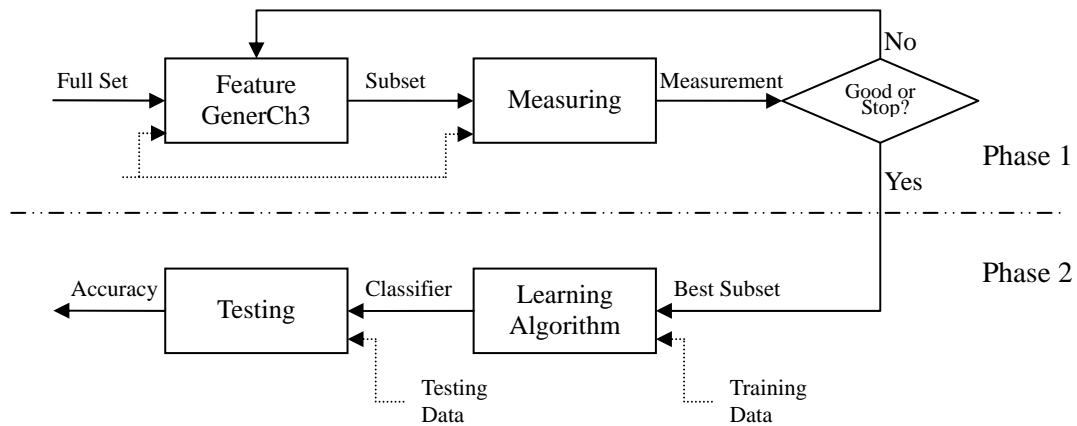


Figure 2.5 A filters model of feature selection (Liu and Motoda, 1998).

In addition to the characteristic that built on the intrinsic properties of the data, the filters approach has the 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, 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. Therefore 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. Still, recently proposed efficient embedded methods are competitive. In addition, some filters provide a generic selection of variables, 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 overfitting.

2.4.3 Information Theoretic Ranking Criterion

Information Theoretic Ranking Criteria is to gather the empirical estimates of the

mutual information between each variable and the target:

$$I(i) = \int_{x_i} \int_y (p(x_i, y) \log \frac{p(x_i, y)}{p(x_i)p(y)}) dx dy \quad (2.10)$$

where $p(x_i)$ and $p(y)$ are the probability densities of x_i and y , and $p(x_i, y)$ is the joint density. The criterion $I(i)$ is a measure of dependency between the density of variable x_i and the density of the target y . Supposed that the variable is discrete or nominal, the $I(i)$ has to describe as follows.

$$I(i) = \sum_{x_i} \sum_y P(X = x_i, Y = y) \log \frac{P(X = x_i, Y = y)}{P(X = x_i)P(Y = y)} \quad (2.11)$$

It is mentioned to note that the estimation obviously becomes harder with larger numbers of classes and variable values.

2.4.4 Embedded Approach

Guyon and Elisseeff (2003) gave the definition of Embedded approach is to perform variable selection in the process of training and this approach is usually specific to given learning machines. In other words, embedded approach is based on the built-in mechanism to perform variable selection, such as Classification and Regression Tree (CART) (Breiman *et al.*, 1984).

2.5 L-J Method

The L-J method is a feature selection approach that defines scores for the available feature at training. It was developed for two authors, Lothar Hermes and Joachim M. Buhmann in 2000. They use the influence to determine the important features. The influence means that the ability of affecting the decision of hyperplane in SVM structure. Compared with wrappers and filters approaches, L-J methods is a feature selection strategy which defines scores for available features on the basis of a single training run (Hermes and Buhmann, 2000) and is easy to compute for users. Next, the brief introduction of L-J method is described as follows.

The initial tasks of the L-J method should construct a SVM structure with a given training sets by using complete data components. After constructing the classifier $f(\mathbf{x})$ (equation 2.12), they estimated the importance of separate feature components to $f(\mathbf{x})$, λ_i is the Lagrange multipliers. To rank the components of a given vector \mathbf{x} according to their influence on the classification, users then should compute the gradient of $f(\mathbf{x})$ at position \mathbf{x} (equation 2.13).

$$f(\mathbf{x}) = \sum_{i=1}^l \lambda_i y_i \mathbf{x}_i^T \mathbf{x} + b \quad (2.12)$$

$$\nabla f(\mathbf{x}) = \sum_{i \in SV} \lambda_i y_i \nabla_{\mathbf{x}} K(\mathbf{x}_i, \mathbf{x}) \quad (2.13)$$

where $\nabla f(\mathbf{x})$ is the gradient of $f(\mathbf{x})$, and it is perpendicular to the optimal hyperplane. Next, user should implement the task of project of the unit vector e_j on $\nabla f(\mathbf{x})$. If the projection of the unit vector e_j on $\nabla f(\mathbf{x})$ is small, it represent that the feature j is not important at position \mathbf{x} . In other word, the $\nabla f(\mathbf{x})$ should be roughly orthogonal to e_j , the feature j should not influence the distance to the decision hyperplane. Summary of above, we can compute the angle $\alpha_j(\mathbf{x}_i)$ between $\nabla f(\mathbf{x})$ and e_j , $j=1,2,\dots,n$, representing the indices of the individual feature, to measure what are the important factors as follows:

$$\alpha_j(\mathbf{x}_i) = \min_{\beta \in \{0,1\}} \left\{ \beta \pi + (-1)^\beta \arccos \left(\frac{(\nabla f(\mathbf{x}))^T e_j}{\|\nabla f(\mathbf{x})\|} \right) \right\} \quad (2.14)$$

Values $\alpha_j(\mathbf{x}_i) \approx \pi$ represent that the feature j has only weak influence on the assignment $f(\mathbf{x})$ of \mathbf{x} . Small values (is closed to 0) indicate important features. Finally, we must compute the $\tilde{\alpha}_j$ to rank the features. The index $\tilde{\alpha}_j$ is defined as follows:

$$\tilde{\alpha}_j = 1 - \frac{2}{\pi} \cdot \frac{\sum_{i \in I_\varepsilon} \alpha_{ij}}{|I_\varepsilon|} \quad (2.15)$$

$\tilde{\alpha}_j$ is an index for ranking features. The features with smaller $\tilde{\alpha}_j$ can be dropped.

甲、 Data Complexity

In order to show the effectiveness of our proposed approach, the data complexity was utilized to evaluate its corresponding performance. As for the complexity of the data, many indicators were measured in different fields. For instance, entropy can help us to see the complexity of the input data. In this study we explain the data complexity by using mess level (Wang, 2003). Before calculating the mess level, related symbols are defined as follows.

A_{ij} is the value of the j^{th} attribute in an instance.

\bar{A}_j^+ is the mean of the attribute while $y_i = 1$ for the instance.

\bar{A}_j^- is the mean of the attribute while $y_i = -1$ for the instance.

A_{jmax} is the maximum value of the j^{th} attribute.

A_{jmin} is the minimum value of the j^{th} attribute.

k is the number of attributes in an instance, and $k \geq 2$.

x_i^+ is the instance with $y_i = 1$.

x_i^- is the instance with $y_i = -1$.

n is the number of instances in the data set, and $n \geq 2$.

Now, we define $M_a(\bullet)$ and $M_b(\bullet)$ in the following:

$$M_a(x_i^+) = \sqrt{\frac{\sum_{j=1}^k \left(\frac{A_{ij} - \bar{A}_j^+}{A_{jmax} - A_{jmin}} \right)^2}{k-1}} \quad (2.16)$$

$$M_a(x_i^-) = \sqrt{\frac{\sum_{j=1}^k \left(\frac{A_{ij} - \bar{A}_j^+}{A_{jmax} - A_{jmin}} \right)^2}{k-1}} \quad (2.17)$$

$$M_a(S) = \sum_{i=1}^n (M_a(x_i^+) + M_a(x_i^-)) \quad (2.18)$$

$$M_b(x_i^+) = \sqrt{\frac{\sum_{j=1}^k \left(\frac{A_{ij} - \bar{A}_j^-}{A_{jmax} - A_{jmin}} \right)^2}{k-1}} \quad (2.19)$$

$$M_b(x_i^-) = \sqrt{\frac{\sum_{j=1}^k \left(\frac{A_{ij} - \bar{A}_j^+}{A_{jmax} - A_{jmin}} \right)^2}{k-1}} \quad (2.20)$$

$$M_b(S) = \sum_{i=1}^n (M_b(x_i^+) + M_b(x_i^-)) \quad (2.21)$$

If all the elements belong to the positive class, $M_a(x_i^-)$ approaches to 0. On the contrary, $M_a(x_i^+)$ approaches to 0 if all the elements belong to the negative class. Therefore, most of the elements approach to the positive class if $M_a(S)$ is larger. The elements are also concentrated in their space. $M_b(x_i^+)$ approaches to 0 if all the elements belong to the negative class; else, if all the elements belong to the positive class, $M_b(x_i^-)$ approaches to 0. According to the definition, we know that if $M_b(S)$ is larger then positive and negative elements are further. Dividing equation 2.18 by equation 2.21, we get equation 2.22 called mess level (ML) which is shown as:

$$\text{Mess level} = \frac{M_b(S)}{M_a(S)} \quad (2.22)$$

When the ML is close to 1, the data set is complex and is not easy to classify.

CHAPTER 3

PROPOSED APPROACHES

3.1 SVM with Combined Kernel Functions

Several investigations (Yao *et al.*, 2005; Wang *et al.*, 2004) indicated that the kernel functions are useful for classification. For examples, 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 (Hammer and Gersmann, 2003; Dong *et al.*, 2005; Lukas *et al.*, 2004; Yao *et al.*, 2005). The polynomial and RBF kernels own the advantages by themselves and this encourages us to propose a combined approach to pursue a much better classification accuracy.

Suppose given a training set of M samples or input vectors $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_i, \dots, \mathbf{x}_M\}$ with known class labels $\{y_1, y_2, \dots, y_i, \dots, y_M\}$, $y_i \in \{+1, -1\}$, a new data point \mathbf{x} is assigned a label by the SVM according to the decision function

$$f(\mathbf{x}) = \text{sign} \left(\sum_{i=1}^{M_s} \alpha_i y_i k(\mathbf{x}_i, \mathbf{x}) + b \right) \quad (3.1)$$

where

$$k(\mathbf{x}_i, \mathbf{x}) = \langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}) \rangle \quad (3.2)$$

is the kernel function that defines feature space, $\Phi(\mathbf{x})$ is a nonlinear mapping function from input space to feature space, $\langle \cdot, \cdot \rangle$ denotes an inner product, b is a bias value, and α_i are positive real numbers obtained by solving a Quadratic Programming (QP) problem that yield the maximal margin hyperplane (Vapnik, 1998).

Owing to the polynomial kernel function owns the advantage of changing the degree d in the feature space (see equation 2.5) and Gaussian RBF kernel is itself a

normalized kernel (see equation 2.6), the kernel k_P and k_G are employed in this study to develop new kernels, k_{P+G} and $k_{P.G}$.

First, to simplify the tasks of the classification process, parameter a was ignored and parameter b was set at 1. We can rewrite the equations 2.5 and 2.6 as follows:

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

where d is its degree and is adjustable.

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

where γ is kernel width; γ is adjustable.

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

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

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

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

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

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

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

Next, we need to provide the relevant proof of our new kernels, i.e. they should be symmetric and satisfied with the characterization of Mercer's Theorem. Here, kernel k_{P+G} is a representative case to proof.

Lemma. Let k_P and k_G be kernels over $X \times X$, $X \subseteq R^n$, then the kernel k_{P+G} in

equation 3.5 is symmetric and satisfied with the characterization of Mercer's theorem.

Proof

Let $k_P = \phi_P = \langle \langle \mathbf{x} \cdot \mathbf{z} \rangle \rangle^d$, $k_G = \phi_G = \exp(-\frac{1}{2\gamma} \|\mathbf{x} - \mathbf{z}\|^2)$.

Then

$$\begin{aligned}
 & k_{P+G}(\mathbf{x}, \mathbf{z}) \\
 &= k_P(\mathbf{x}, \mathbf{z}) + k_G(\mathbf{x}, \mathbf{z}) \\
 &= \langle \phi_P(\mathbf{x}), \phi(\mathbf{z}) \rangle^2 + \langle \phi_G(\mathbf{x}), \phi(\mathbf{z}) \rangle^2 \\
 &= \langle \phi_P(\mathbf{z}), \phi(\mathbf{x}) \rangle^2 + \langle \phi_G(\mathbf{z}), \phi(\mathbf{x}) \rangle^2 \\
 &= k_P(\mathbf{z}, \mathbf{x}) + k_G(\mathbf{z}, \mathbf{x}) \\
 &= k_{P+G}(\mathbf{z}, \mathbf{x})
 \end{aligned}$$

Hence, kernel k_{P+G} is symmetric.

From the Cauchy-Schwarz inequality it follows that:

$$\begin{aligned}
 & k_{P+G}(\mathbf{x}, \mathbf{z}) \\
 &= k_P(\mathbf{x}, \mathbf{z}) + k_G(\mathbf{x}, \mathbf{z}) \\
 &\leq k_P(\mathbf{x}, \mathbf{x})k_P(\mathbf{z}, \mathbf{z}) + k_G(\mathbf{x}, \mathbf{x})k_G(\mathbf{z}, \mathbf{z}) \\
 &\leq k_P(\mathbf{x}, \mathbf{x})k_P(\mathbf{z}, \mathbf{z}) + k_G(\mathbf{x}, \mathbf{x})k_P(\mathbf{z}, \mathbf{z}) + k_P(\mathbf{x}, \mathbf{x})k_G(\mathbf{z}, \mathbf{z}) + k_G(\mathbf{x}, \mathbf{x})k_G(\mathbf{z}, \mathbf{z}) \\
 &= (k_P(\mathbf{x}, \mathbf{x}) + k_G(\mathbf{x}, \mathbf{x}))(k_P(\mathbf{z}, \mathbf{z}) + k_G(\mathbf{z}, \mathbf{z})) \\
 &= k_{P+G}(\mathbf{x}, \mathbf{x})K_{P+G}(\mathbf{z}, \mathbf{z})
 \end{aligned}$$

Next, let k_{P+G} be defined in the closed interval $\mathbf{a} \leq \mathbf{x} \leq \mathbf{b}$ and likewise for \mathbf{z} .

The kernel $k_{P+G}(\mathbf{x}, \mathbf{z})$ can be expanded in the series

$$k_{P+G}(\mathbf{x}, \mathbf{z}) = \sum_{i=1}^{\infty} \lambda_i \phi_i(\mathbf{x})\phi_i(\mathbf{z})$$

with positive coefficients, $\lambda_i > 0$ for all i . According to the following condition,

$$\begin{aligned}
& \int_b^a \int_b^a K(x, z) \psi(x) \psi(z) dx dz \\
&= \int_b^a \int_b^a K_P(x, z) \psi(x) \psi(z) dx dz + \int_a^b \int_a^b K_G(x, z) \psi(x) \psi(z) dx dz \\
&\geq 0 + 0 = 0
\end{aligned}$$

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

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

The kernel k_{P+G} is satisfied with the characterization of Mercer's theorem. It is existence in the feature space. Thus, k_{P+G} is a kernel.

3.2 Feature Selection for the SVM by Using the L-J Method

The L-J method enlighten us a good idea for feature selection by using the influence of j^{th} feature. First, the L-J method needs to compute the angle $\alpha_j(\mathbf{x}_i)$ between $\nabla f(\mathbf{x})$ and e_j . Next, the feature is ranked by the index $\tilde{\alpha}_j$. This process can be constructed by the SVM classifier. We hope that applying the L-J method to our combined kernels (developed in the section 3.1) may obtain a good performance for classification. Similar algorithm can be illustrated as follows:

Step 1: Construct a SVM structure $g(\mathbf{x})$ with a given training sets by using complete data components.

Step 2: Select the combined kernel function into the SVM classification for classification. The combined kernel functions are k_{P+G} and $k_{P.G}$ defined in (3.7) and (3.8).

Step 3: Compute the gradient of $g(\mathbf{x})$ at position \mathbf{x} , so-called $\nabla g(\mathbf{x})$.

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 $\alpha_j(\mathbf{x}_i)$.

Step 6: Rank the features by $\tilde{\alpha}_j = 1 - \frac{2}{\pi} \cdot \frac{\sum_{i \in I_\varepsilon} \alpha_{ij}}{|I_\varepsilon|}$, $\tilde{\alpha}_j \in [0,1]$

Drop several unimportant features with smaller $\tilde{\alpha}_j$ values.



CHAPTER 4

ILLUSTRATION

In order to illustrate the proposed approaches' effectiveness, we use twelve datasets to implement the classification tasks. In addition, the relevant strategies of kernel selection and parameter setting are investigated.

4.1 Data Sets

A total of three data sets, 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-one 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 disease data. In addition to the medical data, nine data sets from the UCI repository (Blake and Merz, 1998) were used. These data sets were census income, shuttle, mushroom, letter, ionosphere, vehicle silhouettes, spambase, vowel, and sonar.

Among the twelve data sets, seven were considered as the larger ones, as each contained more than 5,000 samples (Oyang *et al.*, 2005). The remaining five data sets were considered as the smaller ones. Before our experiment, we had worked some data preprocess. Due to the fact that some anthropometrical data tend to be incomplete, we deleted these data. In order to reduce the differences among the features, we normalized the data prior to implementing the SVM classifier. All the normalized data transferred to x_{new} were scaled to the [-1, 1] interval via equation 4.1. The meta-data including the number of features, classes, cases and feature style, are represented in Table 4.1. In addition, the data complexity computed by ML and the ratio of positive to negative were also appended.

$$x_{new} = \frac{x - \bar{x}}{\max(x) - \min(x)} \quad (4.1)$$

4.2 Implementation Results

Five approaches including linear kernel, two popular kernels (polynomial and RBF), and two proposed kernels (polynomial plus RBF, k_{P+G} ; polynomial multiplies RBF, $k_{P \cdot G}$) were implemented for the classification tasks. We use the one-against-one procedure to calculate the accuracy of classification in the multi-class SVM model, else the general procedure is employed to acquire that. Furthermore, a popular classifier, K nearest neighbor (KNN) was employed as the benchmark in our experiment. In order to simplify the process of classification, the parameter a was set at 0, 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 to 10. Parameter γ was set at 10^{-3} , 10^{-2} , 10^{-1} , 10^0 , 10^1 , 10^2 , respectively.

A total of twelve data sets were separated into large (more than 5000 samples) and small ones (less than 5000 samples) as mentioned before. The imbalanced data sets is shuttle. Table 4.2 compares the accuracy of the classification with the larger and smaller data sets respectively. In the larger data sets, in general the accuracy of the classification of the SVM based approaches is better than that of the KNN approach. Among them, the proposed kernel $k_{P \cdot G}$ (polynomial multiplies RBF kernel) has the best performance, and the next one is another proposed kernel k_{P+G} . As for smaller data sets, the results are as similar as the larger data sets. In general, the performance of the SVM based approaches is better than that of the KNN approach. Among them, the combined kernel $k_{P \cdot G}$ also has the best performance. In addition we found that the performance of the proposed kernels is not so good for the

classification of imbalanced data. Based on the ML, our proposed combined kernels have a better performance when the ML is small.

The implementation result of feature selection is showed in Tables 4.3 and 4.4. In this procedure, the kernels were applied to L-J method for feature selection. The optimal parameter settings were employed in SVM model for L-J feature selection. As same as above, we use the original SVM technique if the data were of two classes. Else the SVM model is worked by one-against-one process if the data are more than two classes.

The average accuracies of the classification for the seven larger data sets and five smaller data sets are shown in Tables 4.3 and 4.4. Their standard deviations are listed in the brackets. The two tables indicated that the combined kernel $k_{P.G}$ has better performance than the other approaches. After feature selection (from 75% to 25%), the kernel $k_{P.G}$ 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 $k_{P.G}$ almost had the lowest standard deviation among the four approaches in the larger data. In the smaller data set, the kernel $k_{P.G}$ performed well.

Table 4.1 Data sets used in this study.

No	Data set	# of samples	# of features	# of classes	Data Style	Data complexity (ML)	Ratio of positive to negative
1	Hyperlipidemia	6000	33	2	c	1.00	1:2.08
2	Liver disease	6000	33	2	c	1.46	1:2.81
3	Renal disease	6000	33	2	c	1.06	1:3.60
4	Census income	32561	14	2	c, d	1.41	1:3.15
5	Shuttle*	14500	9	7	c	3.63	1:14.20
6	Mushroom	8124	22	2	d	1.01	1:1.07
7	Letter	15000	16	26	c	1.00	1:1.13
8	Sonar	208	60	2	c	1.01	1:1.14
9	Ionosphere	351	34	2	c	1.08	1:1.79
10	Vehicle silhouettes	846	18	4	c	1.09	1:1.40
11	Spambase	4601	57	2	c	1.04	1:1.54
12	Vowel	990	13	11	c, d	1.05	1:1.00



c: continuous; d: discrete.

Table 4.2

Comparison of classification accuracy with the larger and smaller data sets.

No.	Data sets	Data classification algorithms						KNN $k=1$	KNN $k=3$
		SVM linear	SVM Polynomial	SVM RBF	SVM Poly + RBF	SVM Poly \times RBF			
1	Hyperlipidemia	68.06	51.92	68.06	69	69	59	68	
2	Liver disease	73.75	73.5	73.5	77	77	60	68	
3	Renal disease	78.5	78.58	78.5	82	82	67	81	
4	Census income	70.5	75	71.5	74	75.8	74	73	
5	Shuttle*	95.5	98.2	95.5	96.8	95.5	100	100	
6	Mushroom	97.33	100	99.73	99.73	100	91.33	94	
7	Letter	82.25	86.75	90.75	87.5	90.75	81	78.65	
	AVERAGE-large	80.84(11.65)	80.56(16.49)	82.51(12.65)	83.72(11.55)	84.29(11.39)	76.05(15.63)	80.38(12.48)	
8	Sonar	85.33	88.1	88.1	92.86	95.23	83.33	85.71	
9	Ionosphere	81.43	84.29	84.29	91.43	91.43	81.43	78.57	
10	Vehicle silhouettes	75	79.3	82.84	79.3	85.8	75	83	
11	Spambase	94.5	94.5	94.5	94.5	95	88	87.5	
12	Vowel	98.89	90.91	98.89	99.49	99.49	97.22	99.72	
	AVERAGE-smaller	87.03(9.69)	87.42(5.88)	89.72(6.83)	91.52(7.48)	93.39(5.11)	85(8.27)	86.9(7.92)	

*: imbalanced data set

(): standard deviation

No1-7: larger data sets.

No8-12: smaller data sets.

Table 4.3 The accuracy of feature selection for the SVM using the L-J method (larger data sets)

Kernel \ Dataset	Full (100%)				Reduced (75%)			
	k_P	k_G	k_{P+G}	$k_{P,G}$	k_P	k_G	k_{P+G}	$k_{P,G}$
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
Census_income	75	71.5	74	75.8	69.6	71.6	71.6	76
Shuttle*	98.2	95.5	96.8	95.5	98.4	98.2	99.8	98.8
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
AVERAGE	80.56	82.51	83.72	84.29	77.75	80.01	82.81	83.28
(Std. dev.)	(16.49)	(12.65)	(11.55)	(11.39)	(16.53)	(13.06)	(12)	(11.4)
Kernel \ Dataset	Reduced (50%)				Reduced (25%)			
	k_P	k_G	k_{P+G}	$k_{P,G}$	k_P	k_G	k_{P+G}	$k_{P,G}$
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
Census_income	68.5	69.5	78	72.5	72	73.6	72	74
Shuttle*	91.5	91.41	91.5	95.5	81.4	82.6	81.4	82.8
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
AVERAGE	74.8	77.52	80.32	81.3	68.61	72.33	71.75	74.46
(Std. dev.)	(16.12)	(12.71)	(11.2)	(11.74)	(18.67)	(15.43)	(15.92)	(13.91)

*: imbalanced data set

(): standard deviation

Table 4.4 The accuracy of feature selection for the SVM using the L-J method (smaller data sets)

Kernel \ Dataset	Full				Reduced (75%)			
	k_P	k_G	k_{P+G}	$k_{P,G}$	k_P	k_G	k_{P+G}	$k_{P,G}$
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
(Std. dev.)	(5.88)	(6.86)	(7.48)	(5.11)	(7.34)	(8.92)	(5.57)	(5)

Kernel \ Dataset	Reduced (50%)				Reduced (25%)			
	k_P	k_G	k_{P+G}	$k_{P,G}$	k_P	k_G	k_{P+G}	$k_{P,G}$
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
(Std. dev.)	(7.55)	(9.79)	(5.24)	(5.86)	(16.13)	(17.95)	(17.66)	(18.49)

(): standard deviation

4.3 Discussions

In the experiment, we found that parameters d and γ heavily influenced the classification accuracy. These two parameters have a different impact on larger and smaller data sets. In larger data sets, degree d should be higher, and γ should be lower. On the other hand, degree d should be lower and γ should be higher in the smaller data sets. Figures 4.1 and 4.2 show the relationship between parameters and accuracy for a large data set (Renal disease) and a smaller data set (Vowel), respectively.

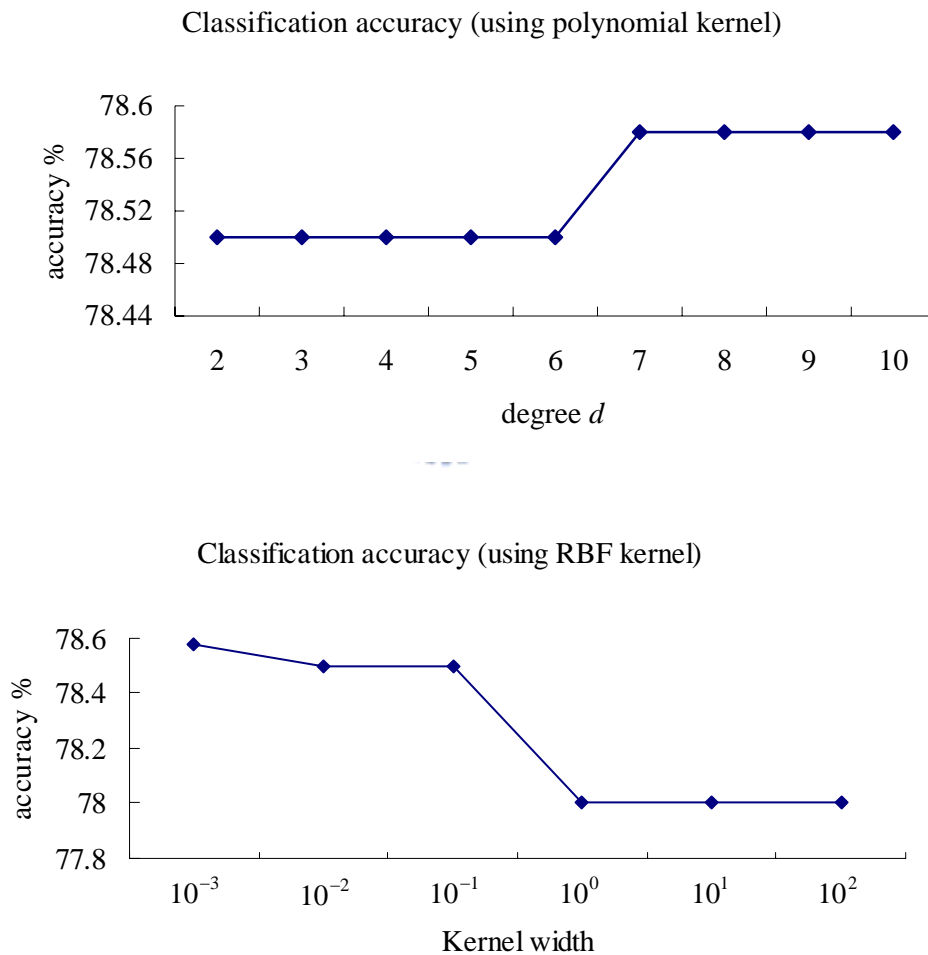


Figure 4.1 The relationship between parameters and accuracy for the larger data set.

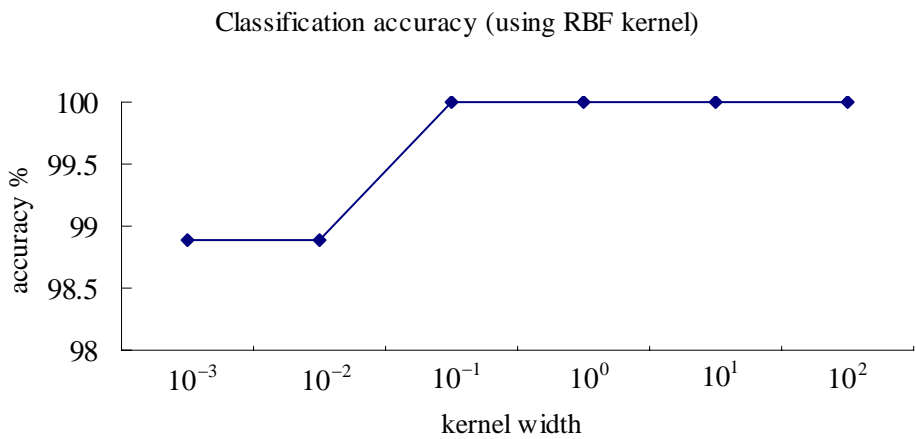
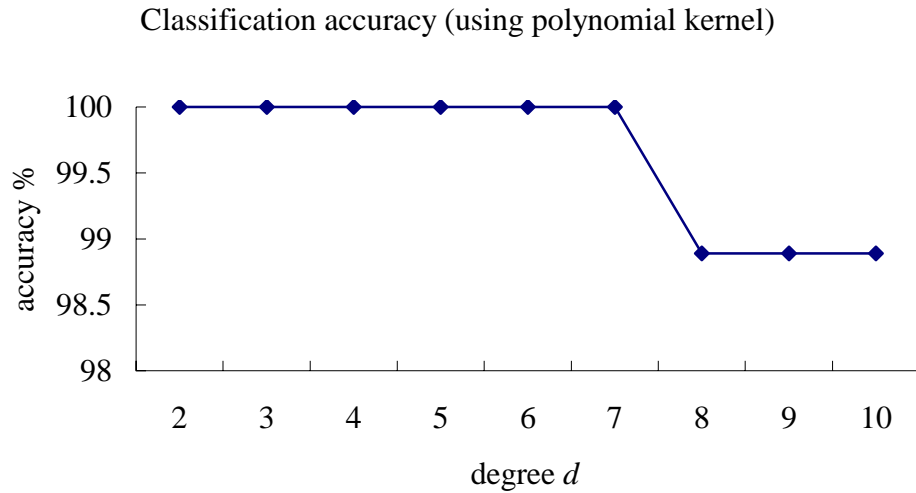


Figure 4.2 The relationship between parameters and accuracy for the smaller data set.

Some research indicated that the SVM with the kernel method provided a better performance for classification than the linear methods (Tefas *et al.*, 2001). In the present study, our experiment showed similar results (see Table 4.2). Although the linear kernel is not the best of the kernel based approaches for large data sets, it is acceptable compared with the KNN approach. The other two popular kernels, polynomial and RBF, also provided an acceptable performance in both larger and

smaller data sets. We found that the performance of the RBF is better than that of the polynomial, both in the larger and smaller data sets.

In the setting of the parameters for the polynomial and RBF kernels, Pardo and Sberveglieri (2005) consider that larger values for the polynomial kernel parameter d mean more complex classification functions (higher order polynomials). These functions are useful for solving classification problems. At the same time however, a smaller value for the RBF kernel parameter γ is also good at solving classification problems. In this study, the results of our experiment are similar to those of Pardo and Sberveglieri (2005) (see Figures 6 and 7). It is evident that larger d is good at complex data because it can obtain a greater probability of classification. Hence, we feel that the input space, with a lower dimension transformation to the feature space with a higher dimension, seems to make it easier to classify a separable bound.

In the following we discuss the effect of parameters d and γ on classification accuracy. First, for the polynomial kernel, we set $a=b=1$, if d is adjusted from 3 to 5, and the terms of the polynomial will be expanded from 4 to 6. As a result of the terms being expanded, the number of boundaries is also increased. Although the larger the d value, the poorer the performance of the classification, we can slightly adjust the d value based on the complexity of the data.

Next, suppose the width γ of the RBF kernel is adjusted from 10^0 to 10^1 , then the increment of the RBF kernel is positive. On the contrary, the increment of the RBF kernel is negative when the kernel width is decreased. Thus the user can change the kernel width until the kernel is satisfied with his need. From the mathematical viewpoint, when the smaller data sets are in the lower space, a larger width is useful to easily and quickly achieve the optimal solution. However, when the larger data sets are in the higher space and when there are many local optimal solutions, then it is easy to fall into the trap of larger kernel width. Thus, the small

width is best for larger data sets. Our experiment only shows the classification accuracy difference for larger and smaller data sets using different kernel widths; however, we could not find a significant difference in the classification accuracy for the data sets with a different data complexity.

Based on the above discussion, some useful strategies for determining parameters d and γ are summarized in Table 4.5. In the polynomial kernel, a larger parameter d is suitable for larger data sets; and a smaller d is suitable for smaller data sets. In the RBF kernel, a smaller parameter γ is suitable for larger data sets; and a larger γ is suitable for smaller data sets.

Table 4.5 The strategies of parameter setting of polynomial and RBF kernels.

Kernel type	Polynomial	RBF
Data set size	(d)	(γ)
Larger data set	larger	smaller
Small data set	smaller	larger

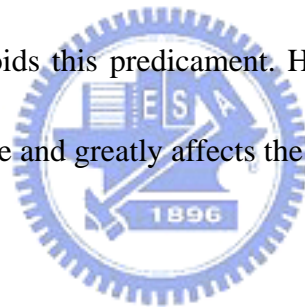
In our experiment, it seems that the multiplication kernel ($k_{p,G}$) is superior to the summation one (k_{p+G}). The reason for this may be that the multiplication kernel has some functions by changing degree and adjusting width at the same time, which seems to increase the classification performance. However, the influences of these functions are not significant in the summation kernels.

In addition, ML was used to evaluate the data complexity. As expected, the combined kernel, $k_{p,G}$ provides a better performance for classification when the ML approaches to 1. However, it seems that the combined kernels are not superior to the other approaches when the ML is greater than approximately 1.5. A possible explanation may be that, for simple problems using the SVM with the original kernel is good enough for classification. The combined kernels are not recommended for

addressing simple problems because they will complicate the data space.

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.



CHAPTER 5

A CASE STUDY: HYPERTENSION DETECTION

In this chapter, 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 detection 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 (Mykkanen *et al.*, 1997; Jeppesen *et al.*, 2000). As defined by the National High Blood Pressure Education Program (NHPEP), hypertension can be summarized as shown in Table 5.1.

Table 5.1 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

Recently, syndrome X has been investigated more and more (Chen *et al.*, 2000^a). In fact, there is a significant relationship between body size and syndrome X (Lin *et al.*, 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 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 3D 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 detection using the proposed approach for feature selection.



5.2 Implementation

A total of thirty-one 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 6,000 data sets were selected randomly from the original database via data pre-processing. Four kernel functions including k_p , k_G , k_{p+G} , and $k_{p.G}$ were employed to construct the SVM models. The relevant parameter of polynomial kernel d was set between 2 to 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, $k_{p.G}$ 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 features were 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 $k_{p.G}$ was employed, a total of thirteen 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 three approaches. They are backpropagation neural network (BPNN), rough sets and decision tree. In this study, *Professional II Plus* software was used to perform BPNN 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 the trained backpropagation neural network structure. It can be defined as follows (Su *et al.*, 2002):

$$P_i = \sum_{j=1}^n \sum_{k=1}^m \sum_{l=1}^s |W_{ij} \times V_{jk}| \quad (5.1)$$

Where:

W_{ij} is the weight between the i^{th} input node and the j^{th} hidden node;

V_{jk} is the weight between the j^{th} hidden node and the k^{th} output node;

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 trained network 33-12-1 were removed. Finally, fourteen 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.

The Rough Sets theory proposed by Pawlak (1982) provides a mathematical tool for representing and reasoning about vagueness and uncertainty. It can be approached as an extension of the Classical Set Theory and can be considered sets with fuzzy boundaries – sets that cannot be precisely characterized using the available of attributes. The basic concept of the rough sets is the notion of approximation space, which is an ordered pair $A = (U, R)$, where U is nonempty set of subjects, called universe; R is equivalence relation on U , called *indiscernibility relation*. If $x, y \in U$ and xRy then x and y are *indistinguishable* in A . Each equivalence class induced by R , i.e. each element of the quotient set $\tilde{R} = \frac{U}{R}$, is called an *elementary set* in A . An approximation space can be alternatively note by $A = (U, \tilde{R})$. It is assumed that the empty set is also elementary for every approximation space A . A definable set in A is any finite union of elementary sets in A . For $x \in U$ let $[x]_R$ denote the equivalence class of R , containing x . For each $X \subseteq U$, X is characterized in A by pair of sets – its lower and upper approximation in A ,

defined respectively as:

$$A_{low}(X) = \{x \in U \mid [x]_R \subseteq X\}$$
$$A_{upp}(X) = \{x \in U \mid [x]_R \cap X \neq \emptyset\}$$

A rough sets in A is the family of all subsets of U having the same lower and upper approximations. After the lower and the upper approximation have been found, the rough sets theory can be used to derive both certain and uncertain information, and induce certain and possible rules from them. In this case study, the important anthropometric factors selected from the rough sets are as similar as BPNN approach except for the breast width.

A decision tree is another feature selection approach. It is a popular classifier in machine learning applications and is also used as a diagnostic model in medicine. Decision tree is connected via nodes and branches. The tree construction process is heuristically guided by choosing the ‘most informative’ attribute at each step, aimed at minimizing the expected number of tests needed for classification. Let E be the entire initial set of training examples, and C_1, \dots, C_N be the decision classes. A decision tree is constructed by repeatedly calling a tree construction algorithm in each generated node of the tree. Tree construction stops when all examples in a node are of the same class, or if some other stopping criteria are satisfied. In brief, a decision tree is a flow-chart-like tree structure, in which each internal node denotes a test on an attribute, each branch represents an outcome of the test, and leaf nodes represent classes or a class distribution.

C4.5 and CART are two typical decision trees. C4.5 is an entropy-based algorithm; however, CART is a binary tree based on the Gini Index (GI) to determine the condition for constructing the tree. In this study, the entropy-based tree (Quinlan, 1986) has been chosen to analyze and induct this diagnostic tree owing to it has the a flow-chart-like structure and makes more user-friendly. In our hypertension detection

case, by running the C4.5, we select thirteen anthropometric factors. They are age, waist circumference, right thigh circumference, left thigh circumference, right leg circumference, left leg circumference, breast width, hip width, volume of trunk, surface area of trunk, volume of left arm, volume of right arm and volume of right leg.

For medical applications, two measures, sensitivity and specificity, are frequently used to discuss the performance. The four elements of them are defined as True positives (TP): True positive answers of a classifier denoting correct classifications of positive cases; True negatives (TN): True negative answers denoting correct classifications of negative cases; False positives (FP): False positive answers denoting incorrect classifications of negative cases into class positive; False negatives (FN): False negative answers denoting incorrect classifications of positive cases into class negative. Sensitivity measures the fraction of positive cases that are classified as positive. Specificity measures the fraction of negative cases classified as negative. The two epidemiological measures can be described as follows. In addition, accuracy was also described.

$$\text{Sensitivity} = \frac{TP}{TP + FN} \quad (5.2)$$

$$\text{Sepcificity} = \frac{TN}{TN + FP} \quad (5.3)$$

$$\text{Accuracy} = \frac{TP + FN}{TP + FN + TN + FP} \quad (5.4)$$

All of the feature selection approaches are assessed by the epidemiology based indices, namely, sensitivity and specificity. In addition, accuracy was employed to

evaluate their performance. There are 13 and 14 features selected by the implemented approaches. As shown in table 5.2, we found that the neural network based model is the worst in terms of the three indices among the various approaches. We also found that the sensitivity was decreased and the specificity was increased in SVM based approaches. This means that the ability of testing TN improved but it deteriorated on test TP. Indeed, this is not favorable for diagnosing. Fortunately, the decreased range observed was small. Also as the specificity increases, it would be beneficial in minimizing the cost of developing new medicines for hypertension. Furthermore, the accuracy of SVM based model is better than those of the neural network based, decision tree and rough sets approaches. In addition, although the results showed that the decision tree and rough sets is better on sensitivity, the SVM based methods have the fewer decreased range after feature selection. In other word, SVM based methods are still better than the other approaches. Hence, we consider that the SVM based method has the advantage of optimization computation and prevails over all other methods.

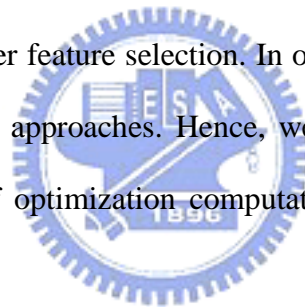


Table 5.2 A comparison of performance of feature selection

Methods	Features*		Sensitivity		Specificity		Accuracy	
	Full	Reduced	Full	Reduced	Full	Reduced	Full	Reduced
Neural network	33	14	0.4478	0.3963	0.7186	0.7289	0.6883	0.6233
k_P	33	13	0.4689	0.4655	0.7356	0.7639	0.7033	0.6700
SVM k_G	33	13	0.4929	0.4805	0.7368	0.7693	0.7083	0.6767
(L-J based) k_{P+G}	33	13	0.5143	0.4830	0.7396	0.7699	0.7133	0.6783
k_{P-G}	33	13	0.5373	0.4987	0.743	0.7790	0.7200	0.6900
DT	33	13	0.5970	0.5300	0.7178	0.7082	0.7040	0.6785
Rough Sets	33	14	0.5996	0.5538	0.7189	0.7160	0.6876	0.6593

*: number of features

5.4 Discussion

The aim of this study is to investigate the relationship between anthropometrical

factors and hypertension. In addition, some significant anthropometrical factors are selected by our approaches. After the feature selection, the common anthropometric factors including waist circumference, right thigh circumference, left thigh circumference, volume of trunk, surface area of trunk and volume of right arm were collected by these methods.

A number of researches are concerned about X syndrome or cardiovascular disease, and the indices BMI and WHR are often employed in their investigations (Kim *et al.*, 2001; Chen *et al.*, 2000^b). However, some researches indicate that BMI and WHR without the significant position could be disapproving (McNeely *et al.*, 2001). For instance, it makes the BMI imprecise because the pure height and/or weight measure varies significantly across ethnic groups. In other words, two people have different body sizes (ex. one is apple type but another is pear type), yet their WHR is the same. Hence, it is not suitable using these indices. 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 general, the people have the larger trunk, the function of their heart-lung have the heavy loading naturally.

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. For adults, the fat usually disperses uniformly to viscera and subcutaneous tissue. Erwin *et al.* (2000) consider that the subcutaneous adipose tissue in people with syndrome X, especially in the lower trunk is greater than healthy people. Furthermore, they indicate that the visceral fat is a risk factor for

syndrome X and infer to hypertension. However, we obtained a crude index when we based on indices BMI and WHR. It is relatively easy to do, even though advanced medical techniques such as computer topography (CT) and magnetic resonance imaging (MRI) are available to evaluate visceral fat. These techniques however are much too expensive for screening all patients, and it could reduce the wish to be examined for hypertension for some patients. Thus, hypertension detection via a simple and accurate approach is worthy of performing. Summary of the case study, 3D whole body scanner is useful for anthropometrics collection; our proposed combined kernel is good at the performance of hypertension detection.



CHAPTER 6

CONCLUSIONS

6.1 Summary

In this study, we discuss the theory and application of the kernel-based SVM. First, we used four kernels including polynomial kernel, RBF kernel, multiplication kernel ($k_{P,G}$) and summation kernel (k_{P+G}) 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 detection was investigated in this study. We selected thirteen 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 detection; such as more sport and/or nutrition intervention/control may decrease the risk of hypertension. In addition to our proposed approaches, the backpropagation neural network, decision tree and rough sets were 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, decision tree and rough sets approaches. After feature selection, sensitivity and accuracy are reduced and specificity is increased in our approaches. 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.

6.2 Further Research

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.

In addition to computation speed problems, the rules generalization is another important issue. Obviously, SVM based methods do not provide the ability of rule generalization although they have the strong foundation in statistical learning theory. Most of the engineering or medical problems with regard to the rule generalization are still more exigent than feature selection for operators or non-experts. Hence, how to provide a function of rule generalization extend the SVM based method deemed an important issue in the interdisciplinary applications.

REFERENCES

1. Amaldi, E. and Kann, V., On the approximation of minimizing non zero variables or unsatisfied relations in linear systems, *Theoretical Computer Science*, 209, 237-260, 1998.
2. Aronszajn, N., theory of reproducing kernels, *Transactions of the American Mathematical Society*, 68(3), 337-404, 1950.
3. Bekkerman, R., El-Yaniv, R., Tishby, N. and Winter, Y., Distributional word cluster vs. words for text categorization, *Journal of Machine Learning Research*, 3, 1183-1208, 2003.
4. Bi, J., Bennett, K., Embrechts, M., Breneman, C. and Song, M., Dimensionality reduction via sparse support vector machines, *Journal of Machine Learning Research*, 3, 1229-1243, 2003.
5. Blake C.L. and Merz, C.J., UCI repository of machine learning databases, Dept. Infor. Comput. Sci., Univ. California, Irvine, CA, 1998.
6. Breiman, L., Friedman, J.H., Olshen, R.A. and Stone, C.J., *Classification and Regression Trees*, Wadsworth and Brooks, 1984.
7. Caruana, R. and de Sa, V., Benefitting from the variables that variable selection discards, *Journal of Machine Learning Research*, 3, 1245-1264, 2003.
8. Chen, W., Bao, W., Begum, S., Elkasabany, A., Srinivasan, S.R. and Berenson, G.S., 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, 2000^a.
9. Chen, C.H., Lin, K.C., Tsai, S.T. and Chou, P., 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, 2000^b.
10. Chiang, C.F., A Feature Selection for Support Vector Machines, M.S. thesis, Tung Hai University, 2002.
11. Cortes, C. and Vapnik, V., Support-vector networks, *Machine learning*, 20(3), 273-297, 1995.
12. Cristianini, N. and Taylor J.S., An introduction to SVMs and other kernel-based learning methods. *Combridge University Press*, 2000.
13. Dhillon. I., Mallea, S. and Kumar, R., A divisive information-theoretic feature clustering algorithm for text classification, *Journal of Machine Learning Research*, 3, 1265-1287, 2003.
14. Dong, B., Cao, C. and Lee, S.E., Applying support vector machines to predict building energy consumption in tropical region, *Energy and Buildings*, 37, 545-553, 2005.

15. Dudoit, S., Fridlyand, J. and Speed, T., Comparison of discrimination methods for the classification of tumors using gene expression data, *Journal of American Statistical Association*, 97(457), 77-87, 2002.
16. Erwin, T., Reinhard, M., Karl, S. and Gilbert, R., Artificial neural networks compared to factor analysis for low-dimensional classification of high-dimensional body fat topography data of healthy and diabetic subjects, *Computers and Biomedical Research*, 33, 365-374, 2000.
17. Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P., From data mining to knowledge discovery: An overview. In Fayyad, U., Piatetsky-Shapiro, G., Smyth, P. and Uthurusamy, R., editors, *Advances in Knowledge Discovery and Data Mining*, pages 495-515, AAAI Press / The MIT Press, 1996.
18. Guyon, I. and Elisseeff A., An introduction to variable and feature selection, *Journal of Machine Learning Research*, 3, 1157-1182, 2003.
19. Hammer, B. and Gersmann, K., A note on the universal approximation capability of support vector machines, *Neural Processing Letters*, 17, 43-53, 2003.
20. Haykin, S., *Neural networks: a comprehensive foundation*, New Jersey, 2nd ed., Prentice Hall, 1999.
21. Hermes, L. and Buhmann, J.L., Feature selection for support vector machines, *Proc. of the International Conference on Pattern Recognition*, 2, 716-719, 2000.
22. Jain, A.K., Duin, R.P.W. and Mao, J., Statistical pattern recognition: a review, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1), 4-37, 2000.
23. Jeppesen, J., Hein, H.O., Suadicani, P. and Gyntelberg, F. High triglycerides and low HDL cholesterol and blood pressure and risk of ischemic heart disease. *Hypertension*, 36, 226-232, 2000.
24. Joachims, T., *The Maximum-Margin Approach to Learning Text Classifiers: Methods, Theory, and Algorithms*. PhD thesis, Universitaet Dortmund, 2000.
25. Kim, Y.I., Kim, C.H., Choi, C.S., Chung, Y.E., Lee, M.S., Lee, S.I., Park, J.Y., Hong, S.K. and Lee, K.U., 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, 2001.
26. Kohavi, R. and John, G., Wrapper for feature selection, *Artificial Intelligence*, 97(1-2), 273-324, 1997.
27. LeCun, Y., Jackel, L. D., Bottou, L., Brunot, A., Cortes, C., Denker, J. S., Drucker, H., Guyon, I., Muller, U. A., Sackinger, E., Simard, P. and Vapnik, V., Comparison of learning algorithms for handwritten digit recognition, *International Conference on Artificial Neural Networks*, Fogelman, F. and Gallinari, P. (Ed.), 53-60, 1995.
28. Lin, C.J., Formulations of support vector machines: a note from an optimization

- point of view, *Neural Computation*, 13(2), 307-317, 2001.
29. Lin, J.D., Chiou, W.K., Weng, H.F., Tsai, Y.H., & Liu, T.H. 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, 2002.
 30. Liu, H. and Motoda, H., *Feature selection for knowledge discovery and data mining*, Norwell, MA: Kluwer Academic, 1998.
 31. Lukas, L., Devos, A., Suykens, J.A.K., Vanhamme, L., Howe, F.A., Majos, C., Moreno-Torres, A., Van Der Graff, M., Tate, A.R., Arus, C. and Van Huffel, S., Brain tumor classification based on long echo proton MRS signals, *Artificial Intelligence in Medicine*, 31, 73-89, 2004.
 32. Muller, K. R., Mika, S., Ratsch, G., Tsuda, K. and Scholkopf, B., An introduction to kernel-based learning algorithms, *IEEE Transactions on Neural Networks*, 12(2), 181-201, 2001.
 33. Mykkanen, L., Haffner, S.M., Ronnema, T., Bennema, T., Bergman, R.N. and Laakso, M., Low insulin sensitive is associated with clustering of cardiovascular disease risk factors. *American Journal of Epidemiology*, 146, 315-321, 1997.
 34. McNeely, M.J., Boyko, E.J., Shofer, J.B., Newell-Morris, L., Leonetti, D.L. and Fujimoto, W.Y., Standard definitions of overweight and central adiposity for determining diabetes risk in Japanese American, *American Journal of clinical Nutrition*, 74, 101-107, 2001.
 35. Osuna E., Freund, R. and Girosi, F., Training support vector machines: an application to face detection, *Proc. Computer Vision and Pattern Recognition '97*, 130-136, 1997.
 36. Oyang, Y.J., Hwang, S.C., Ou, Y.Y., Chen, C.Y. and Chen, Z.W., Data classification with radial basis function networks based on a novel kernel density estimation algorithm, *IEEE Transactions on Neural Networks*, 16(1), 225-236, 2005.
 37. Papadopoulos, A., Fotiadis, D.I. and Likas, A., Characterization of clustered microcalcifications in digitized mammograms using neural networks and support vector machines, *Artificial Intelligence in Medicine*, 34, 141-150, 2005.
 38. Pardo, M. and Sberveglieri G., Classification of electronic nose data with support vector machines, *Sensors and Actuators B*, 107, 730-737, 2005.
 39. Quinlan, J.R., Induction of decision trees, *Machine Learning*, 1(1), 81-106, 1986.
 40. Rakotomamonjy, A., Variable selection using SVM-based criteria, *Journal of Machine Learning Research*, 3, 1357-1370, 2003.
 41. Reunanen, J., Overfitting in making comparisons between variable selection methods, *Journal of Machine Learning Research*, 3, 1371-1382, 2003.

42. Roth, V. and Steinhage., Nonlinear discriminant analysis using kernel functions. In: Advances in neural information processing system, 12, 568-574, 2000.
43. Sanchez, V.D.A., Advanced support vector machines and kernel methods, *Neurocomputing*, 55, 5-20, 2003.
44. Schmidt, M., *Identifying Speakers with Support Vector Networks*, in Interface '96 Proceedings, Sydney, 1996.
45. Scholkopf, B. and Smola, A.J., *Learning with kernels: support vector machines, regularization, optimization, and beyond*, Cambridge, Mass: MIT Press, 2002.
46. Scholkopf, B., Smola, A.J. and Muller, K.R., Nonlinear component analysis as a kernel eigenvalue problem, *Neural Computation*, 10(5), 1299-1319, 1998.
47. Scholkopf, B., *Support Vector Learning*, PhD Theiss, R. Oldenbourg Verlag, Munich, 1997.
48. Scholkopf, B., Burges, C.J.C. and Smola, A.J., Extracting support data for a given task. In U.M. Fayyad and R. Uthursamy, editors, Proceedings, First International Conference on Knowledge Discovery and Data mining, Menlo Park, 1995. AAAI Press.
49. Su, C.T., Hsu, H.H. and Tsai, C.H., Knowledge mining from trained neural networks. *Journal of Computer Information Systems*, 42(4), 61-70, 2002.
50. Tefas, A., Kotropoulos, C. and Pitas, I., Using support vector machines to enhance the performance of elastic graph matching for frontal face authentication, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(7), 2001.
51. Torkkola, K., Feature extraction by non-parametric mutual information maximization, *Journal of Machine Learning Research*, 3, 1415-1438, 2003.
52. Vapnik, V., *The Nature of Statistical Learning Theory*, Springer-Verlag, New York, 1995.
53. Vapnik, V., *Statistical Learning Theory*, Wiely: New York, 1998.
54. Wang, C.W., *Index of Kernel Functions for Support Vector Machine*, M.S. Thesis, National Cheng Kung University, 2003.
55. Wang, M.L., Li, W.J. and Xu, W.B., Support vector machines for prediction of peptidyl prolyl cis/trans isomerization, *Journal of Peptide Research*, 63, 23-28, 2004.
56. Wahba, G., An introduction to model building with reproducing kernel Hilbert space, *Computing Science and Statistics*, 32, /I2000Proceedings/GWahba/wahba_short.pdf.
57. Weston, J., Elisseeff, A., Schoelkopf, B. and Tipping, M., Use of the zero norm with linear models and kernel methods, *Journal of Machine Learning Research*, 3, 1439-1461, 2003.
58. Yao, X.J., Panaye, A., Doucet, J.P., Chen, H.F., Zhang, R.S., Fan B.T., Liu M.C.

- and Hu, Z.D., Comparative classification study of toxicity mechanisms using support vector machines and radial basis function neural networks, *Analytica Chimica Acta*, 535, 259-273, 2005.
59. Yeang, C.H., Ramaswamy, S., Tamayo, P., Mukherjee, S., Rifkin, R.M., Angelo, M., Reich, M., Lander, E., Mesirov, J. and Golub, T., Molecular classification of multiple tumor types, *Bioinformatics: Discovery Note*, 1(1), 1-7, 2001.
60. Zhu, Y., Li, C. and Zhang, Y., A practical parameters selection method for SVM, Springer-Verlag Berlin Heidelberg, 2004.
61. Zien, A., Ratsch, G., Mika, S., Scholkopf, B., Lengauer, T. and Muller, K.R., Engineering support vector machine kernels that recognize translation initiation sites, *Bioinformatics*, 16, 851-824, 2000.

