Support Vector Machines Classification for Discriminating Coronary Heart Disease Patients from Non-coronary Heart Disease

S Hongzong1, W Tao2, Y Xiaojun3, L Huanxiang3, H Zhide3, L Mancang3, F BoTao4

ABSTRACT

Objective: The present contribution concentrates on the application of support vector machines (SVM) for coronary heart disease and non-coronary heart disease classification.

Methods: We conducted many experiments with support vector machine and different variables of low-density lipoprotein cholesterol (LDLC), high-density lipoprotein cholesterol (HDLC), total cholesterol (TC), triglycerides (TG), glucose and age (dataset 346 patients with completed diagnostic procedures). Linear and non-linear classifiers were compared: linear discriminant analysis (LDA) and SVM with a radial basis function (RBF) kernel as a non-linear technique.

Results: The prediction accuracy of training and test sets of SVM were 96.86% and 78.18% respectively, while the prediction accuracy of training and test sets of LDA were 90.57% and 72.73% respectively. The cross-validated prediction accuracy of SVM and LDA were 92.67% and 85.4%.

Conclusion: Support vector machine can be used as a valid way for assisting diagnosis of coronary heart disease.

INTRODUCTION

Coronary heart disease (CHD) is one of the leading human diseases of high mortality in industrialized countries (1, 2). However, it is equally prevalent as a cause of death in developing countries. A risk factor is accepted as a causal factor if the results of observational studies and randomized controlled trials (RCT, interventional studies) are also supported by results from basic research (biological plausibility) (3).
Epidemiological studies have established that low-density lipoprotein cholesterol (LDL-C), high-density lipoprotein cholesterol (HDL-C), total cholesterol (TC) and triglycerides (TG) are major risk factors for developing CHD (4–7). Scientific consensus exists on the importance of the three main risk factors for the development of CHD: hypercholesterolaemia, diabetes mellitus and age. Cardiovascular disease accounts for approximately 70% of all deaths in people with diabetes mellitus (8), and the risk of cardiovascular mortality is two to three times higher in men and three to five times higher in women with diabetes than in those without diabetes (9–11). These factors play an important role in the development of atherosclerosis and CHD in those without diabetes (9–11). Three to five times higher in women with diabetes than in those without diabetes (9–11). These factors play an increasingly important role in the development world but also play an increasingly important role in the developing world (12).

The three risk factors (mentioned above) only aid to diagnose CHD by traditional methods. There is no publication that shows that the diagnosis of CHD only depends on them. Machine learning methods may be capable of objective interpretation of all available results for the same patient and in this way increase the diagnostic accuracy. In the usual setting, the machine learning algorithms are tuned to maximize classification accuracy. Pattern recognition methods which can develop models with maximal generalization ability from large and generally noisy data sets are proposed.

Methods of artificial intelligence were gradually introduced into clinical decision-making research from 1970 to 1974. There are many pattern recognition methods suitable for classification: two of the most commonly used are linear discriminant analysis (LDA) and support vector machines (SVM) (13). The technique of SVM, developed by Vapnik, was proposed essentially for classification problems of two classes. Support vector machines use geometric properties to exactly calculate the optima separating hyperplane directly from the training data (14–16).

Due to its remarkable generalization performance and small number of learning parameters, the SVM has attracted attention and gained extensive applications. Support vector machines have been effective in disease diagnosis (17–19), DNA sequence analysis, protein structure prediction and gene expression pattern discovery (20–25). They are particularly suitable for CHD prediction because of their ability to build effective predictive models when the dimensionality of the data is high and the number of observations is limited. They are also based on a strong theoretical foundation for avoiding over-fitting training data.

Based on the laboratory tests (TG, TC, LDL-C, HDL-C and glucose) and age, we proposed SVM for the classification of CHD and non-CHD controls, as results show that SVM is a superior method in diagnosis of CHD and it can be extended for classification of other diseases.

SUBJECTS AND METHODS
Serum Samples
In this study, cases were chosen from the First Hospital of Lanzhou University. It is one of top-ranking hospitals in China and the major hospital in Gansu Province. One-hundred and seventy-two patients with CHD were diagnosed by coronary angiography in the hospital over two years. One-hundred and seventy-four persons without CHD comprised the control group selected from persons who came for routine medical examination. Serum samples from 172 patients with CHD and 174 persons with non-CHD were obtained. The ages of the patients ranged from 24 to 83 years (mean: 53.9 years). Of the control group, the ages ranged from 30 to 76 years (mean: 49.4 years). Test samples must be collected in the manner routinely used for analysis. Freshly drawn serum from a fasting individual is preferred. Plasma or serum samples should be physically separated from contact with cells as soon as possible within two hours. Tubes of blood are to be kept closed at all times in a vertical position. Serum samples were stored at +2°C to +8°C and assayed within eight hours.

Clinical Chemistry
Glucose reagent is used to measure the triglyceride concentration by timed endpoint method. In the reaction, hexokinase catalyses the transfer of the phosphate group from adenosine triphosphate to glucose to form adenosine diphosphate and glucose-6-phosphate. The glucose-6-phosphate is then oxidized to 6-phosphogluconate with the concomitant reduction of β-nicotinamide adenine dinucleotide to reduced β-nico-tinamide adenine dinucleotide by the catalytic action of glucose-6-phosphate dehydrogenase.

Triglyceride reagent is used to measure the triglyceride concentration by timed endpoint method. Triglycerides in the sample are hydrolyzed to glycerol and free fatty acids by the action of lipase. A sequence of three coupled enzymatic steps using glycerol kinase (GK), glycerophosphate oxidase (GPO) and horseradish peroxidase (HPO) cause the oxidative coupling of 3,5-dichloro-2-hydroxybenzenesul-fonic acid (DHBS) with 4-aminoantipyrine to form a red quinoneimine dye.

High-density lipoprotein cholesterol reagent is used to measure the cholesterol concentration by a time-endpoint method. In the reaction, the cholesterol esterase (CE) hydrolyzes cholesterol esters to free cholesterol and fatty acids. The free cholesterol is oxidized to cholesterol-3-one and hydrogen peroxide by cholesterol oxidase (CO). Peroxidase (HPO) catalyzes the reaction of hydrogen peroxide with 4-aminoantipyrine (4-AAP) and phenol to produce a coloured quinoneimine product.

The Synchron LX System automatically proportions the appropriate HDLC, TG and glucose samples and reagent volume into a cuvette. The ratio used is one part sample to 60 parts reagent of HDLC and to 100 parts reagent of TG and
glucose. The system monitors the change in absorbance at 520, 520 to 340 nanometers of HDLC, TG and glucose respectively. This change in absorbance is directly proportional to the concentration of cholesterol in the sample and is used to calculated and express the corresponding concentration.

At the same time, the concentration of TC and LDLC were calculated according to the concentration of TG and HDLC.

Low-density lipoprotein cholesterol was considered high if the level exceeded 3.7 mmol/L or the individual was on LDL lowering therapy. High density lipoprotein cholesterol was considered low when under 0.6 mmol/L, TG, TC and glucose level over 1.8 mmol/L, 5.7 mmol/L and 6.11 mmol/L respectively were considered high.

**Choice of Parameters**

Data on age, gender, smoking, alcohol, family history, blood pressure, chest pain symptoms, ECG changes and serum indices (TC, TG, LDLC, HDLC, glucose, calcium, potassium, phosphorus, myocardial enzymes) were collected from the hospital files of 346 persons (172 CHD patients and 174 non-CHD patients). All indices were analyzed with a stepwise method of linear discriminant analysis. Finally, only 6 indices including serum lipids (TC, TG, LDLC, HDLC) glucose and age entered the model. Although SVM can easily tolerate more parameters, most of the features are usually irrelevant for the classification task and only introduce noise. The precise order of features might change from iteration to iteration. Because of the multivariate properties of the SVM algorithm, each feature ranking takes into account (at least to some extent) correlations between single variables. Evaluating the classification performance at each step makes it possible not only to identify a suitable subset of descriptors but also to determine how many of them are actually needed for a reliable classification.

**Linear Discriminant Analysis**

Linear discriminant analysis is useful in building a predictive model of group membership based on observed characteristics of each case. The procedure generates a discriminant function (or, for more than two groups, a set of discriminant functions) based on linear combinations of the predictor variables that provide the best discrimination between the groups. The functions are generated from a sample of cases for which group membership is known; the functions can then be applied to new cases with measurements for the predictor variables but with unknown group membership.

All of these approaches are analogous discriminant function analysis used to determine which variables discriminate between two or more naturally occurring groups. If we code the two groups in the analysis as Group 1 CHD and Group 2 non-CHD patients and use that variable as the dependent variable in a multiple regression analysis, then we would get results that are analogous to those we would obtain via linear discriminant analysis.

**Support Vector Machine**

What follows is a brief description of the SVM algorithm. Overfitting of data can be avoided by limiting the complexity of the models that the method can possibly generate. A specific approach for controlling the complexity of the models is given by the Vapnik-Chervonenkis (VC) theory and the structural risk minimization principle (26). This is applied to the training of a classification SVM by fitting of a hyperplane such that the largest margin is formed between two classes of chemicals while minimizing the classification errors. Non-linearity in a data set is accounted for with kernel functions, which map the input vectors to some higher dimensional space such that a hyperplane with reduced classification errors can be found (27). A major advantage is that optimization problems resulting from SVMs have a global minimum and can be solved with standard quadratic programming tools.

Support vector machine is a learning system that uses a hypothesis space of linear functions in a high dimensional feature space, trained with a learning algorithm from the optimization theory. It attempts to minimize the upper bound on the generalization error based on the principle of structural risk minimization (SRM) (28). The decision function implemented by SVM can be written as:

$$ f(x) = \text{sign} \left( \sum_{i=1}^{N} y_i a_i k(x_i, x) + b \right) $$

Two typical kernel functions are listed below:

Polynomial function $k(x_i, x_j) = (x_i \bullet x_j + 1)^d$

Radial basis function (RBF) $k(x_i, x_j) = \exp(\|x_i - x_j\|^2)$

Training parameters $\gamma$ and $C$ were optimized using a gradient descent-like algorithm to achieve maximum accuracy of prediction for the validation set. Parameter $C$ is an internal parameter that is set prior to SVM training. It defines the trade-off between the separating margin and the penalty for incorrect predictions (27).

**Training and Test Data Sets**

The data set was split randomly into a 242-member training set and an external prediction set of 104 cases. Of the training set, there are 120 CHD cases, 122 non-CHD cases. Of the test set, there are 52 CHD and non-CHD cases respectively. The training set was used to adjust the parameters of the models and the test set was used to evaluate its prediction ability. Leave-one-out (LOO) cross-validation was used to prevent the network from overfitting.

**Accuracy of Diagnostic Tests**

Accuracy of a diagnostic test can be expressed through sensitivity and specificity. Sensitivity refers to the ability of a certain diagnostic test to detect a particular disease. It is expressed as the probability of testing positive if the particular disease is truly present ie the probability of having both a
positive test and a positive diagnosis. Hence a test with 96% sensitivity means that 96% of those with the disease will test positive. Specificity, on the other hand, refers to the probability of testing negative if the disease is truly absent. In other words, 96% specificity means that 96% of those who are truly negative for the disease or problem will have a negative test while 4% of them will have a false positive test.

RESULTS
Since disease diagnosis is of great concern, positive predictive value was used to evaluate the models. Table 1 shows the performance of SVM for the training and test sets. The prediction accuracy of training and test sets of SVM are 96.86% and 78.18% respectively while the prediction accuracy of training and test sets of LDA are 90.57% and 72.73% respectively. With the cross-validation, the prediction accuracy of SVM and LDA was 92.67% and 85.4%.

In general, in the two-group case we fit a linear equation with LDA:

Group 1 = -28.97 + 0.57 age + 1.36 × 10⁻³ TC + 1.17 TG + 6.57 HDLC + 2.21 LDLC + 0.56 glucose
Group 2 = -19.58 + 0.4 age + 5.54 × 10⁻³ TC + 0.81 TG + 6.18 HDLC + 2.69 LDLC + 0.3 glucose

Each variable has a different contribution to the above equation (Table 2). The interpretation of the results of a two-group problem is straightforward and closely follows the logic of multiple regression. Those variables with the largest regression coefficients are the ones that contribute most to the prediction of group membership.

The observation was confirmed in the results obtained using SVM and LDA as shown in Table 3. Of the 172 CHD samples and 174 normal samples, applying SVM with radial basis function (RBF) yields a sensitivity of 0.95 and specificity of 0.90 with 163 true positives, 158 true negatives, 16 false positives and 9 false negatives; applying LDA obtain a sensitivity of 0.91 and specificity of 0.77, with 156 true positives, 134 true negatives, 40 false positives and 16 false negatives.

DISCUSSION
In general, LDA is a very useful tool for detecting the variables that allow the researcher to discriminate between different groups and for classifying cases into different groups with a better than chance accuracy. However, the prediction ability of the LDA method is much lower than SVM. The quality of the SVM models depends on the kernel type and various parameters that control the kernel shape. Using a quadratic programming algorithm, SVM offers a unique maximal separation hyperplane. As other multivariate statistical models used in chemometrics, there are no clear guidelines for selecting the optimum set of theoretical parameters and decision function (kernel type and associated parameters). Therefore, the only practical way of finding an optimally predictive SVM model is through extensive experiments. In this work, SVM training included the selection of capacity parameter C, the corresponding parameters of the kernel function. Parameter C is a regularization parameter that controls the trade-off between maximizing the margin and minimizing the training error. If C is too small, then...
insufficient stress will be placed on fitting the training data. If C is too large, then the algorithm will overfit the training data. But, literature indicated that prediction error was scarcely influenced by C (29). To make the learning process stable, a large value should be set up for C first. The kernel type is another important one. Because the use of SVM models in chemometrics is only in the beginning, there are no clear guidelines on selecting the most effective kernel for a certain classification problem. But for classification tasks, you will most likely use C-classification with the RBF kernel, because of its good general performance and the few number of parameters (only two: C and \( \gamma \)) (30). To select the type of kernel function, which determines the sample distribution in the mapping space, many studies indicated that the radial basis function is commonly used because of its good general performance and few parameters to be adjusted (30). Therefore, in this work, the RBF was used, the form is as follows:

\[
\exp (-\gamma \| u - v \|^2)
\]

Where \( \gamma \) is a parameter of the kernel and \( u \) and \( v \) are two independent variables. The \( \gamma \) of kernel function greatly affects the number of support vectors which has a close relation with the performance of the SVM and training time. Too many support vectors can produce overfitting and make the training time longer. The \( \gamma \) also controls the amplitude of the RBF function and, therefore, controls the generalization ability of SVM. Thus, to find the optimal parameter \( \gamma \), experiments were performed using a different value of \( \gamma \) with the leave one out (LOO) procedure of the same training set and the testing set. For the training data set, the first group of models, parameter \( \gamma \) was set in the range of 0.01 to 0.15 with 0.01 increment and \( C = 100 \). The curve of training accuracy and \( \gamma \) versus training accuracy is shown in Figure 1. The low number of support vectors prompted the selection of 0.06 as the optimal value of the gamma. In addition, to test the effect of C, the second group of models using the same training data set were obtained with capacity parameter C from 10 to 150, every 10 and a certain \( \gamma = 0.06 \). The curve of training accuracy and C value is shown in Figure 2. It can be seen from it that the selection of parameter C has some influence on the performance. The optimal C was found as 100 with a highest training accuracy of 92.67%. The best choices for C and \( \gamma \) of the SVM were 90 and 0.06 with the support vector number of 48. The test set was presented using the SVM model. The mechanism of risk factors that has been used in this study is a key step on development of CHD.

Dyslipidaemia is common in Type 2 diabetic patients and is characterized by elevated TG and reduced HDLC (31). Studies have indicated the role of high TG and low HDLC as cardiovascular risk factors (32, 33). The uptake of cholesterol by macrophages in the arterial wall and the development of foam cells are facilitated by oxidation of LDL which increases the affinity of LDL particles for scavenger receptors on these cells (34). Unlike LDL receptors, scavenger receptors on macrophages are not down-regulated by increased cellular cholesterol and rapid accumulation of lipid may therefore occur, producing lipid-laden foam cells and leading to the formation of atherosclerotic plaques (35). High levels of HDLC are thought to have protective effects against the development of atherosclerotic plaques and a low HDLC level is associated with increased risk of CHD (36, 37). Triglycerides level is often inversely related to the level of HDLC (38, 39). Although TG does not accumulate in atherosclerotic plaques, hypertriglyceridaemia is associated with an increased risk of CHD and this may be a secondary consequence of the atherogenic effects of low HDLC levels. Several studies have highlighted the importance of elevated TG and low HDLC levels in predicting coronary events in asymptomatic subjects. Levels of HDLC and TG, in addition to LDLC levels and age were found to be independent risk factors for myocardial infarction (40). This gradient in risk is greater
than that which could be predicted by analysis limited to LDLc alone and demonstrates the importance of including HDLc and TG levels in the assessment of CHD risk (41). Observations may reflect that, as the absolute risk of CHD is higher in patients with diabetes than in non-diabetic cohorts (42) diabetes is associated with important quantitative and qualitative changes in lipid and lipoprotein metabolism that are likely to contribute appreciably to the excess CHD risk allied with this condition. In particular, the coexistence of elevated plasma triglycerides, small, dense LDL and low HDL cholesterol represents a lipid 'triad' that is highly atherogenic. While these lipid abnormalities are responsive to therapeutic intervention, the majority of patients with diabetic dyslipidaemia are under-diagnosed and this necessitates an efficient diagnostic approach. If the serum glucose is above the normal level for CHD patients, it will accelerate atherosclerosis. Diabetic glucose levels and impaired glucose tolerance can be maintained if β-cell numbers are reduced to < 20% of normal (43).

From the above discussion, we can see that serum lipids, glucose and age play key roles on CHD development. SVM as a machine learning method has strong performance. So we use SVM to classify CHD with several main risk factors of CHD and gain good result.

CONCLUSION
In the present work, age, serum lipid and glucose concentrations were used to build predictive models for the diagnosis of CHD by the use of LDA and SVM. Compared with the results obtained by LDA, the model using SVM exhibited a better predictive ability with the minimal misclassified number. It showed that the SVM method based on selected features can be used as a valid way for the diagnosis of CHD. More importantly, SVM was shown to be a very promising tool for classification due to the embodying of the structural risk minimization principle which minimizes an upper bound of the generalization error rather than minimizes the training error. This eventually leads to better generalization. In addition, there are fewer free parameters to be adjusted in the SVM, which made the model-selecting process easy to be controlled. Therefore, the SVM is a very effective machine learning technique for the diagnosis of many diseases.

ACKNOWLEDGEMENT
The authors thank the R Development Core Team for affording the free R1.7.1 software.

REFERENCES


41. Assmann G. Pro and con: High-density lipoprotein, triglycerides, and other lipid subfractions are the future of lipid management. Am J Cardiol 2001; 87: (Suppl) 2B–7B.
