A support vector machine approach for predicting heart conditions

Article · January 2010

CITATIONS
0

READS
13

4 authors:

Pooya Tabesh
University of St. Thomas - Houston
10 PUBLICATIONS **17** CITATIONS
SEE PROFILE

Gino Lim
University of Houston
131 PUBLICATIONS **809** CITATIONS
SEE PROFILE

Suresh Khator
University of Houston
50 PUBLICATIONS **862** CITATIONS
SEE PROFILE

Clifford Dacso
Baylor College of Medicine
76 PUBLICATIONS **793** CITATIONS
SEE PROFILE

Proceedings of the 2010 Industrial Engineering Research Conference
A. Johnson and J. Miller, eds.

A Support Vector Machine Approach for Predicting Heart Conditions

Pooya Tabesh, Gino Lim, Suresh Khator
University of Houston
Houston, TX, 77204

Cliff Dacso
Methodist Hospital
Houston, TX, 77030

Abstract

Early diagnosis of heart-related problems can potentially reduce mortality rate and help patients maintain a better quality of life. Recently, advanced mathematical and statistical models enable us to analyze the continuous flow of data and to predict hazardous heart conditions. In this paper, we present a categorical feature-based classification method using support vector machines (SVMs) to predict the heart condition. Existing medical and demographic data of the patients will be used as input for classifying the heart condition as normal or abnormal. Using the introduced categorical feature-based classification method, we identified the significance of the categorical features in classification accuracy. Also, SVM classification accuracy was increased from 65.8% to 84% by modifying the kernel width by using trial-and-error approach.

Keywords
Classification, support vector machine, categorical feature-based classification

1. Introduction

1.1 Background
In recent years, there is an increase in heart-related illnesses that are among the leading causes of hospitalizations and deaths in the United States, Canada and other countries, killing one person every 34 seconds in the United States alone [1]. Monitoring the heart regularly and consequently taking appropriate therapeutic actions may reduce the risk of hospitalization for the patients with heart problems and will potentially reduce the mortality rate. Due to hospitalization costs, patients with heart-related disease may not choose to be hospitalized for monitoring for a long period of time. Therefore, a distance monitoring system integrated with a decision making tool for detecting abnormalities can help improve the diagnostic accuracy of the heart problems without hospitalizing them.

1.2 Literature Review
There are different techniques for monitoring and detecting abnormal conditions of heart. Time series analysis, wavelet analysis, and data mining are some of the most common approaches for classification and prediction. Time series is defined as a sequence of values of a variable at equally spaced time intervals and obtains an understanding of the underlying forces and patterns in the observed data. There are not many recent specific studies on applications of time series in dealing with forecasting and prediction of abnormalities in human-related clinical data. Leong et al. [2] utilized time series analysis to determine the variations and to estimate the strength of the seasonality in age-related (senile) cataract hospitalizations and phacoemulsification surgeries (surgical removal of the lens of the eye by liquefying it and removing it by suction).
Data mining techniques such as artificial neural networks (ANN) and support vector machine (SVM) are mainly used for solving classification problems. Data mining is the methodology of sorting through large amounts of data and extracting relevant understandable correlations and patterns in data. It is a technique to extract the previously unknown but potentially useful information from data [3]. SVM is one of the data mining techniques for dealing

with classification problems. Laufer et al. [4] used an SVM classifier to distinguish between heart and kidney tissues using specific electrical measurements acquired around the tissues. The electrical measurements were performed at a remote site. Then, the raw data were transmitted to another computational site for performing the classification. The results of the tissue analysis were sent back to the original data measurement site. The tissue type was correctly determined with a specificity of over 90 percent.

Babaoglu et al. [5] used Support vector machine with k-fold cross-validation method to diagnose the existence of coronary artery disease. They mainly performed feature selection using genetic algorithm (GA) and binary particle swarm optimization (BPSO) in order to increase the SVM classification accuracy and compare it with the results of simple SVM model without feature selection. The results of their experiments show that the classification accuracy increased from 76.67 in simple SVM method to 81.46 in the one using BPSO as the feature selection method.

Hu et al. [6] utilized ensemble classification methods such as bagging and boosting along with the main SVM classifier and evaluated the performance on breast cancer and heart disease datasets. The outcome of their experiments shows that the ensemble methods may reach a higher accuracy than single SVM. The highest accuracy on the heart disease database was 83.5% which was obtained by bagging.

SVM was selected among numerous data mining methods because we are dealing with a binary classification (separating normal and abnormal) of heart conditions.

2. Methodology

We use SVM as the main classifier in this paper. Therefore, we first discuss our SVM model. Then a categorical-feature-based classification method will be introduced which enables us achieve a higher accuracy of classification.

2.1 Support Vector Machines

SVM was developed by Vladimir Vapnik at AT&T Bell Labs [7]. It is based on the concept of decision planes that define decision boundaries. A decision plane is a hyperplane that separates the objects having different class memberships. SVM classifiers separate the observations into two or more classes in such a way that maximum separation is achieved [8]. A hypothetical hyperplane is the separator in SVM classification problems. In other words, SVM constructs a hyperplane that separates the two sets so as to minimize the number of misclassified points. Generally, there are two types of SVM models: linear and nonlinear. Linear SVM works better on linearly separable datasets but nonlinear SVM model works well even on hardly separable datasets. Since we are dealing with hardly separable data in our experiments we use nonlinear SVM. The dual formulation of the nonlinear SVM function can be formulated as

$$MaxW(\alpha) = \sum_{i=1}^{m} \alpha_i - 0.5 \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j K(x_i \cdot x_j) \tag{1}$$

Subject to:

$$\sum_{i=1}^{m} \alpha_i y_i = 0,$$

$$0 \le \alpha_i \le C.$$

Input vectors $x_i \in R^m$, $i = 1, 2, 3, \ldots, m$, which are called features or attributes are extracted from the database. Associated with every particular input we have a corresponding label ($y_i = \pm 1$) which is called the target value or output in the database. The variable α_i is the Lagrange multiplier in the dual formulation and C is a user-specified parameter representing the penalty for misclassification $K(x_i, x_j)$ is the kernel function and maps the original data points to another space. One of the popular choices for the kernel is Gaussian kernel which is also known as Radial Basis Function (RBF) in the literature. The formulation for this kernel is

$$K(x_i, x_j) = e^{-\|x_i - x_j\|^2 / 2\sigma}, \tag{2}$$

where parameter σ is known as the kernel width.

Gaussian kernel will be used as in the experiments of this paper. Using nonlinear SVM, the condition of the heart will be predicted on the basis of existing patient data.

2.2 Categorical feature-based Classification

Some databases contain both categorical and continuous values in their features. Categorical features may play different roles in classification accuracy. In this method, we partition the data based on the existing values of the

categorical features. Each categorical feature has 2 or more values in the database. We construct multiple subsets equal to the total number of existing values for all the categorical features. The following algorithm shows the procedure for partitioning the data based on the categorical features:

i = index for categorical feature. $i = 1, 2, \ldots, f$ (f is number of categorical features)
$v(i)$ = index for value of categorical feature i. $v(i)$: $1, 2, \ldots, s$ (s is the number of possible values for a categorical feature)
k = index for data instances. $k = 1, 2, \ldots, d$ (d is data size or number of instances in the database)
V_{ik} = value of categorical feature i for k^{th} instance of data
VC_{ij} = j^{th} value of categorical feature i
S_{ij} =subset of instances that have j^{th} value for categorical feature i

Loop i
  Loop $v(i)$
    $j \leftarrow v(i)$
    Loop k
      If $V_{ik} = VC_{ij}$ then
      Add the k^{th} row of data to the corresponding subset (S_{ij})
    End loop k
  End loop $v(i)$
End loop i

For each of the categorical features (i) we have certain number of values in the database ($v(i)$). The database is constructed of d rows (data instances) and c columns (features) from which f features contain categorical values. In this procedure we construct the subsets of data. Suppose a case in which we have 10 categorical features ($f=10$). Also suppose that each of the categorical features has 2 possible values ($s=2$ for all of the features). By running the procedure on this case we construct 20 ($2*10=20$) subsets. Each of the subsets contains the data rows that have a certain value for a specific categorical feature. In each of the iterations, the instances which share a certain value for a certain categorical feature ($V_{ik} = VC_{ij}$) will be added to subset S_{ij}.

After constructing these subsets, we train and run the SVM classifier on each of the subsets (S_{ij}). If the classification accuracy of a subset is higher than the overall classification accuracy of the database we can conclude that the categorical feature used for constructing that certain subset has a significant influence on the classification accuracy of that database.

3. Numerical Experiments

3.1 UCI Heart Disease Dataset

University of California at Irvine (UCI) machine learning repository comprises of 107 useful datasets on various fields of science. One of the datasets on this repository is heart disease dataset that contains 4 databases on heart disease diagnosis. All attributes have numbers as the values. The data were collected from the following four locations: Cleveland Clinic Foundation, Hungarian Institute of Cardiology, Budapest, V.A. Medical Center, Long Beach, CA and University Hospital, Zurich, Switzerland. The Cleveland heart disease database [9] contains 303 observations of which 297 are complete observations and six are observations with missing values [10]. Originally, this database has 76 raw attributes. However, only 14 attributes have been used in most of the published experiments. The fourteenth attribute is the goal field or output variable (or class).

3.2 Initial Experiments

Using nonlinear SVM, we ran the initial experiments using a binary classification method on the Cleveland data. Our experiments on this database are concentrated on separating the cases with heart disease (values 1, 2, 3, 4) from no heart disease cases (value 0). The accuracy of 65.8% was obtained.

3.3 Experiments on Different Subsets of Data

The features in the Cleveland heart disease dataset are both continuous and categorical. In order to understand the nature of the data, two datasets were constructed from the original data. The first one consists of the categorical

features and the second one consists of the continuous features only. Then, two separate experiments were made on each of the subsets using the nonlinear SVM model. The result is tabulated in Table 1.

Table 1: Accuracy of the classification on continuous vs. categorical subsets

Subset	Accuracy
Observations on categorical features only	82%
Observations on continuous features only	55%

The first experiment, which was implemented on categorical features, gives accuracy of 82% which is much higher than 55% on the continuous features only. Higher accuracy for the first classifier is an indicator for the importance and significance of categorical features in the classification and diagnosis of the heart disease. In other words, the prediction of the heart conditions on the basis of existing categorical features is more accurate than that of the continuous features.

We utilized our categorical feature-based classification approach for doing extensive experiments on each of the categorical features. These categorical features have discrete integer values showing the state or attribute of the feature. As it is summarized in Table 2, we performed 23 different experiments on 23 subsets of data. For each of the categorical features, we partition the dataset into several subsets. The number of subsets for each feature is equal to the number of existing values for that specific feature. For instance, the first categorical feature in the dataset is sex and assigns values 0 and 1 for male and female patients, respectively. Therefore, we partition the dataset into two subsets ($v(i)=2$). The first subset is the male patient data while the other subset contains the female patient data. Then we run the nonlinear SVM model on each of these subsets. The results of the experiments on the subsets indicate that the accuracy for each of the subsets is higher than the overall accuracy (65.8%). The only exception is for the subset consisting of the patients with the "thal = 3". The size of this subset is rather small and the SVM model did not perform well. The results are compelling evidence that the categorical features play a key role in predicting the heart conditions.

Table 2: Model Accuracy for the defined subsets

Feature	Subset Description	Accuracy
Sex	Male patients	71.9%
	Female patients	77%
Chest pain type	Patients with chest pain type = 1	69.6%
	Patients with chest pain type = 2	82%
	Patients with chest pain type = 3	79%
	Patients with chest pain type = 4	68%
Fasting blood sugar	Patients with fasting blood sugar = 0	64%
	Patients with fasting blood sugar = 1	68%
ECG result	Patients with ECG result = 0	70%
	Patients with ECG result = 1	68%
	Patients with ECG result = 3	72%
Exercise Induced Angina	Patients with exercise induced angina = 0	68%
	Patients with exercise induced angina = 1	74%
Slope of ST	Patients with slope of ST = 1	72.5%
	Patients with slope of ST = 2	66.2%
	Patients with slope of ST = 3	69%
Ca(vessels colored by fluoroscopy)	Patients with Ca = 0	71%
	Patients with Ca = 1	69.5%
	Patients with Ca = 2	76%
	Patients with Ca = 3	80%
Thal	Patients with Thal = 3	50.3%
	Patients with Thal = 6	75.3%
	Patients with Thal = 7	66.6%

3.4 Improvements on Initial Results

By modifying the kernel width (σ) we could improve the overall classification accuracy from 65.8% to 84%. This was done by a trial-and-error-based method by running the SVM model using different kernel widths and recording the corresponding accuracy for each σ . More experiments will be necessary to understand how to select an appropriate value of σ and this is beyond the scope of this paper.

4. Summary and Future Work

The results of experiments using the proposed categorical feature-based classification method show that the categorical features in the Cleveland heart disease dataset have significant effect in predicting the heart conditions. The obtained classification accuracy after modifications of the kernel width is 84% which is the highest value reported in the literature that adopted the support vector machines approach for the Cleveland heart disease dataset [5, 6]. The performance of the SVM model highly depends on the kernel and its parameters [11]. In our experiments we used Gaussian kernel and utilized the trial-and-error method to define the kernel width. There are more robust methods in the literature to define the parameter for Gaussian kernel. Chaotic adaptive particle swarm optimization (CAPSO) [12], Nedler-Mead simplex method (N-M) [13], and maximum likelihood method are some of the existing algorithms to define this parameter.

The significance of the categorical features encourages us to design a tree-based clustering algorithm to cluster the data in order to further increase the accuracy. The observations can be clustered into several groups according to specific rules. Then, the classification can be performed using the SVM model. Suppose we have a dataset with *r* rows and *c* columns. Each row corresponds to an observation from a patient. The first *c-1* columns correspond to features about the patient and the last column contains a pre-recorded value (or class) that indicates whether the patient is in a normal heart condition (class=-1) or an abnormal condition (class=+1). Suppose also that the features are categorical and have discrete numbers as their values. Some of them have binary values and others may have two, three or four possible integer numbers as their values in the dataset. In a tree structure, we keep all observations on the source node. We begin the tree construction from one of the categorical features in the dataset, say, "sex". Since the feature has two possible branches (male or female), we branch out to two new nodes according to the values of the selected feature. One node includes the observations that have the value 0 (female patients) and the other node includes all the observations with the value 1 (male patients). Then, we select another feature and keep on branching for each of the two existing nodes on the basis of the newly selected feature. This continues until all the features are selected. Finally, we will have all the observations gathered on the leaf nodes. This way, the observations with identical pattern of the feature values will be located at the same leaf. At this point, we also bring the class values (normal=-1, abnormal=+1) related to each observation and put it on the leaf nodes. Using these groups of data, we train the SVM model on each of the datasets to build multiple classifiers. Hence, if any of them gives better prediction accuracy for a specific observation, we can use that classifier for that certain type of observation. Suppose that the prediction accuracy of the k^{th} group is much higher than those of other groups. Also, assume that a specific data pattern *p* belongs to k^{th} group (G^k). Now, if a new observation *q* is collected such that its pattern is identical to *p*, then the prediction for *q* will be made based on the k^{th} classifier.

References

1. Minino, A.M., Heron, M.P., Murphy, S.L., and Kochanek, K.D, 2007, "Deaths: Final Data for 2004," National Vital Statistics Reports, 55(19):1–119.
2. Leong, A.M., Crighton, E.J., Moineddin R., Mamdani, M. and Upshur, R.E.G.,2006, "Time Series Analysis of Age Related Cataract Hospitalizations and Phacoemulsification," BMC Ophthalmology, 6(1):2.
3. Frawley, W.J., Piatetsky-Shapiro, G., and. Matheus, C.J., 1992, "Knowledge Discovery in Databases: An Overview," AI Magazine, 13(3):57–70.
4. Laufer, S., and Rubinsky, B., 2009, "Cellular Phone Enabled Non-Invasive Tissue Classifier" Plos One, 4(4).
5. Babaoglu, I., Findik, O., Ulker, E., 2010, "A Comparison of Feature Selection Models Utilizing Binary Particle Swarm Optimization and Genetic Algorithm in Determining Coronary Artery Disease Using Support Vector Machine," Expert Systems with Applications, 37(4): 3177–3183.
6. Hu, Z., Li, Y., Cai Y., Xu, X., "An Empirical Comparison of Ensemble Classification Algorithms with Support Vector Machines," Proc. of the Third International Conference on Machine Learning and Cybernetics, Shanghai, 26-29 August 2004, 3520–3523.

7. H. Drucker, C.J.C., Burges, L. Kaufman, A., Vapnik, V., 1997, “Support Vector Regression Machines,” Advances in Neural Information Processing Systems, 155–161.
8. Vapnik, V., 1998. Statistical Learning Theory. Wiley, New York.
9. Asuncion, A., Newman, D.J., 2007, UCI Machine Learning Repository, http://www.ics.uci.edu/ ~ mlearn/MLRepository.html.
10. Das, R., Turkoglu, I., and Sengur, A., 2009, “Effective Diagnosis of Heart Disease through Neural Networks Ensembles,” Expert Systems with Applications, 36(4):7675–7680.
11. Smola, A.J., Schölkopf, B. and Müller, K.R., 1998, “The Connection between Regularization Operators and Support Vector Kernels,” Neural Networks, 11: 637–649.
12. Yang, H., Luo, Q., Zhou, Y., "Using Chaotic Adaptive PSO-SVM for Heart Disease Diagnosis," 2nd IEEE International Conference on Computer Science and Information Technology, Beijing, 8-11 August 2009, 156-160.
13. Nelder, J.A., Mead, R., 1964, “A Simplex Method for Function Minimization,” Computer Journal, 308–313.