Feature Selection based Sequential Minimal Optimization (SMO) Classifier for Heart Disease Classification

Aung Nway Oo
Faculty of Computer Science
University of Information Technology
aungnwayoo78@gmail.com

Khin Thuzar Win
Department of Information Technology Supporting and Maintenance
University of Computer Studies Hinthada
khinthuzarwin87@gmail.com

Abstract: Feature Selection is one of the pre-processing steps in machine learning. Feature Selection effectively reduced the dimensionality of dataset, removing irrelevant and redundant feature. In this paper, we proposed a Correlation based Feature subset Selection (CFS) based Sequential Minimal Optimization (SMO) classifier for heart disease classification. Experimental results of CFS-SMO and SMO are compared by using heart disease dataset from UCI. Comparative results show that the proposed CFS-SMO classifier is better than SMO classifier.

Keywords: Feature selection, Classification, Correlation and SMO

1. INTRODUCTION

Heart disease is a general term that means that the heart is not working normally. Babies can be born with heart disease. This is called congenital heart disease. If people get heart disease later, it is called acquired heart disease. According to the statistic report from WHO, heart disease takes the lives of 17.9 million people every year, 31% of all global death. Heart disease is found to be the leading cause of death globally. The main causes of heart disease include diabetes, family history of Heart disease, smoking, obesity, high LDL cholesterol and low HDL cholesterol. The diagnosis of heart disease in most cases depends on a complex combination of clinical and pathological data; this complexity leads to the excessive medical costs affecting the quality of the medical care.

Automated heart disease prediction system is needed for healthcare sector to reduce the number of tests to be taken by a patient, to save cost and also to save time for both, Doctors and patients.

The technique of Computational biology can be applied for translating biological knowledge into clinical practice, as well as in the understanding of biological phenomena from the clinical data. Prediction model is needed for this process. Predictive modeling is a process that uses data mining and statistical analysis to forecast outcomes.

Data Mining is defined as extracting information from huge sets of data. In other words, we can say that data mining is the procedure of mining knowledge from data. Data Mining can be considered to be an inter-disciplinary field involving concepts from Machine Learning, Database Technology, Statistics, Mathematics, Clustering and Visualisation among others. There are several data mining techniques such as tracking patterns, association, clustering, classification, etc.

Classification is a classic data mining technique based on machine learning. Basically, classification is used to classify each item in a set of data into one of a predefined set of classes or groups. Classification method makes use of mathematical techniques such as decision trees, linear programming, neural network, and statistics. One of the major goals of a classification algorithm is to maximize the predictive accuracy obtained by the classification model when classifying samples in the test set unseen during training.

Medical data mining is the new area for exploring hidden data pattern from huge amount of data. Furthermore, feature selection is also important for medical dataset classification. Many irrelevant attributes may be present in data to be mined. So they need to be removed. Also many mining algorithms don't perform well with large amounts of features or attributes. Therefore, feature selection techniques needs to be applied before any kind of mining algorithm is applied. The main objectives of feature selection are to avoid over fitting and improve model performance and to provide faster and more cost-effective models.

Therefore, before the classification of medical dataset we need to use the feature selection algorithm to remove the unimportant features. By removing unrelated features, not only can reduce the dimension of data but also improve the accuracy of classification.

The rest of the paper is organized as follows. Section 2 provides the related work and section 3 presents the overview of correlation based feature subset selection approach. General description of sequential minimal optimization is presented in section 4. The experimental results are discussed in section 5. Finally, conclusion of this study was provided in section 6.

if they are actually no more informative Furthermore the correlations in Equation should be normalized to ensure they are comparable and have the same effect Symmetrical uncertainty compensates for information gain's bias toward attributes with more values and normalises its value to the range[0, 1].

$$\text{symmetrical uncertainty} = 2.0 \times \left[\frac{gain}{H(Y)+H(X)}\right] \qquad (5)$$

3.2 Searching the Feature Subset Space

The purpose of feature selection is to decide which of the initial features to include in the final subset and which to ignore. If there are n possible features initially then there are 2^n possible subsets. The only way to find the best subset would be to try them all.

CFS starts from the empty set of features and uses a forward best first search with a stopping criterion of live consecutive fully expanded non improving subsets.

4. SEQUENTIAL MINIMAL OPTIMIZATION (SMO)

The SMO algorithm was proposed by John C. Platt[11] in 1998 and became the fastest quadratic programming optimization algorithm. Sequential Minimal Optimization (SMO) is used for training a support vector classifier using polynomial or RBF kernels. It replaces all missing the values and transforms nominal attributes into binary ones. A single hidden layer neural network uses exactly the same form of model as an SVM.

SMO is an iterative algorithm for solving the optimization problem described above. SMO breaks this problem into a series of smallest possible sub-problems, which are then solved analytically. Because of the linear equality constraint involving the Lagrange multipliers αi, the smallest possible problem involves two such multipliers. Then, for any two multipliers α1 and α2, the constraints are reduced to:

$$0 \leq \alpha1 \text{ , } \alpha2 \leq C \text{ ,} \\ y1\alpha1 + y2\alpha2 = k,$$

and this reduced problem can be solved analytically: one needs to find a minimum of a one-dimensional quadratic function. k is the negative of the sum over the rest of terms in the equality constraint, which is fixed in each iteration. The algorithm proceeds as follows:

1. Find a Lagrange multiplier α1.
2. Pick a second multiplier α2 and optimize the pair (α1, α2).
3. Repeat steps 1 and 2 until convergence.

When all the Lagrange multipliers satisfy the KKT conditions (within a user-defined tolerance), the problem has been solved. Although this algorithm is guaranteed to converge, heuristics are used to choose the pair of multipliers so as to accelerate the rate of convergence. This is critical for large data sets since there are n*(n − 1) / 2 possible choices for αi and αj.

5. EXPERIMENTAL RESULTS

Heart disease dataset is collected from UCI Repository in the website www.ucirepository.com. This dataset contains 270 observations and 2 classes: the presence and absence of heart disease. There are 150 patient records without suffer heart disease and 120 records for patient with heart disease. Table 1 shows the features and description of heart disease dataset.

Table 1. Features and description

No.	Features	Description
1.	Age	Instance age in years
2.	Sex	Instance gender
3.	Cp	Chest pain type
4.	Trestbps (mmHg)	Resting blood pressure
5.	Chol (mg/dl)	Serum cholesterol
6.	Fbs	Fasting blood sugar
7.	Restecg	Resting electrocardiographic results
8.	Thalach	Maximum heart rate achieved
9.	Exang	Exercise induced angina
10.	Oldpeak	ST depression induced by exercise relative to rest
11.	Slope	The slope of the peak exercise ST segment
12.	Ca	Number of major vessels (0-3) colored by flourosopy
13.	Thal	3 = normal; 6 = fixed defect; 7 = reversible defect

Firstly, find the important features of datasets by using correlation based feature subset selection (CFS) methods. The following table shows the features extracted from heart disease dataset.

Table 2. Top extracted features

Selected attributes: 3,7,8,9,10,12,13 :
chest
resting_electrocardiographic_results
maximum_heart_rate_achieved
exercise_induced_angina
oldpeak

	number of_major_vessels thal

After the important features were extracted 66 % of dataset is used for training and remaining 34 % is used for testing. The accuracy of classifiers and error rate of classifiers are described in the following table 3 and table 4.

Table 3. Accuracy of classifiers

Classifier	Accuracy	TP Rate	FP Rate	Class
CFS-SMO	86.96%	0.837	0.093	Patient with heart disease
		0.907	0.163	Healthy person
SMO	82.61%	0.796	0.140	Patient with heart disease
		0.860	0.204	Healthy person

Table 4. Errot rate of the classifiers

	CFS-SMO	SMO
Mean absolute error	0.1304	0.1739
Root mean squared error	0.3612	0.417
Relative absolute error	25.7511 %	34.3348 %
Root relative squared error	69.9569 %	80.7792 %

6. CONCLUSIONS

The paper proposed the CFS based SMO classifier for heart disease classification. The experimental results showed that accuracy of CFS-SMO classifier is 86.96 % which is better than SMO only approach. All error rates of proposed classification method is smaller than the SMO classifier. Therefore, we can conclude that feature selection based heart disease classification is more efficient than normal classifier.

7. ACKNOWLEDGMENTS

We would like to express our gratitude to all Professors and Teachers from Faculty of Computer Science, University of Information Technology for their suggestion and supports throughout this research work. I would like to express my thanks to Center for Machine Learning and Intelligent Systems at the University of California, Irvine for supporting free dataset for testing of my proposed method. Finally, I also thanks all of my colleague for their participation and contribution for this study.

8. REFERENCES

[1] Atul Kumar Pandey, Prabhat Pandey, K.L. Jaiswal, Ashish Kumar Sen. A Heart Disease Prediction Model using Decision Tree, IOSR Journal of Computer Engineering (IOSR-JCE), Volume 12, Issue 6 (Jul. - Aug. 2013), PP 83-86

[2] P. Suresh, etal Study and Analysis of Prediction Model for Heart Disease: An Optimization Approach using Genetic Algorithm, International Journal of Pure and Applied Mathematics, Volume 119 No. 16 2018, PP 5323-5336

[3] I. Frederix, S. Sankaran, K. Coninx, and P. Dendale, "MobileHeart, a mobile smartphone-based application that supports and monitors coronary artery disease patients during rehabilitation"

[4] KaliaOrphanou, Athena Stassopoulou, and ElpidaKeravnou,"DBN-Extended: A Dynamic Bayesian Network Model Extended With Temporal Abstractions for Coronary Heart Disease

Prognosis", IEEE Journal of Biomedical and Health Informatic, VOL. 20, NO. 3, MAY 2016

[5] C.Sowmiya, Dr.P. Sumithra, A Comparative Study of heart disease prediction using Data Mining Techniques, International Journal of Scientific & Engineering Research, Volume 7, Issue 12, December-2016

[6] R. Sugnya, etal. Novel Feature Selection Method for Predicting Heart Disease with Data Mining Techniques, Asia Journal of Information Technology, 1314-1321, 2016

[7] Oreski, D., & Novosel, T. Comparison of Feature Selection Techniques in Knowledge Discovery Process, 2014.

[8] Kaur, R., Kumar, G., & Kumar, K. A Comparative Study of Feature Selection Techniques for Intrusion Detection, 2015.

[9] Sheena etal., Analysis of Feature Selection Techniques: A Data Mining Approach, International Journal of Computer Applications (0975 –8887)

[10] Blum & Langley, 1997; Kohavi &John, 1997

[11] Platt, John (1998), Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines

[12] I.H. Witten, E. Frank, M.A. Hall " Data Mining Practical Machine Learning Tools & Techniques" Third edition, Pub. – Morgan kouffman.