Available online at www.elixirpublishers.com (Elixir International Journal)

Mechanical Engineering

Elixir Mech. Engg. 94 (2016) 40473-40475

ISSN: 2229-712X

Support Vector Machine in the Prediction of Heart Disease Based on Simple K-Means Clustering

Prabhat Pandey, K.L.Jaiswal and Atul Kumar Pandey

ARTICLE INFO

Article history:
Received: 26 May 2015;
Received in revised form: 16 May 2016;
Accepted: 21 May 2016;

Keywords
Medical Diagnosis,
Heart Disease,
Classification via Clustering,
Sequential Minimal Optimization (SMO),
Simple K-Means Clustering.

ABSTRACT

The Healthcare industry is generally "information rich", but unfortunately not all the data are mined which is required for discovering hidden patterns & effective decision making. We are evaluating the performance of Simple K-Means algorithm Clustering using the mode of classes to clusters evaluation with the prediction attribute nom. The performance of these techniques is compared, based on accuracy. As per our results accuracy of Simple K-Means Clustering, Sequential Minimal Optimization and Sequential Minimal Optimization via Simple K-Means Clustering are 80.85%, 83.82% and 96.69% respectively. In our studies 10-fold cross validation method was used to measure the unbiased estimate of prediction model. The model uses medical terms such as sex, blood pressure, cholesterol like 13 attributes to predict the likelihood of patient getting a Heart disease. Until now, 13 attributes are used for prediction. Our analysis shows that classification model SVM via Simple K-Means Clustering predicts cardiovascular disease with least error rate and highest accuracy of 96.69%.

Introduction

Cardiovascular diseases entail a large number of deaths in the world annually. The most common type of them is CAD which is the reason of about 1/3 of deaths [16]. SMO is an algorithm for training support vector machines. Training a support vector machine requires the solution of a very large quadratic programming (QP) optimization problem. SMO breaks this large QP problem into a series of smallest possible QP problems. These small QP problems are solved analytically, which avoids using a time-consuming numerical QP optimization as an inner loop. SMO is able to handle very large training sets [14].

An SVM is a general algorithm based on guaranteed risk bounds of statistical learning theory i.e. the so called structural risk minimization principle. Recent advances in statistics, generalization theory, computational learning theory, machine learning and complexity have provided new guidelines and deep insights into the general characteristics and nature of the model building/learning/fitting process [1]. Some researchers have pointed out that statistical and machine learning models are not all that different conceptually [2,3].

Material and method

Description of dataset

The data was collected from the Cleveland clinic foundation, and it is available at UCI Repository. The dataset has 14 attributes and 303 records. Missing values have been replaced with the mean value using the Replace Missing Values unsupervised attribute filter available on Weka.

Dr. Prabhat Pandey OSD, Additional Directorate, Higher Education, Division Rewa (M.P.)-India, Mobile No-09425183334, e-mail: prabhatpandey51@gmail.com.

Dr. K.L. Jaiswal Assistant Professor and In charge of BCA, DCA & PGDCA, Department of Physics, Govt. PG Science College, Rewa(M.P.)-India-486001, Mobile No-09424746167, e-mail: drkanhaiyalaljaiswal@gmail.com.

Mr. Atul Kumar Pandey Assistant Professor of Computer Science, Department of Physics, Govt. PG Science College, Rewa (M.P.)-India, Mobile No-09424944538, e-mail: atul.pandey.it2009@gmail.com.

Support Vector Machine

Support vector machine (SVM) is a novel learning machine introduced first by Vapnik [4]. It is based on the Structural Risk Minimization principle from computational learning theory. Hearst et al. [5] positioned the SVM algorithm at the intersection of learning theory and practice: ‘‘it contains a large class of neural nets, radial basis function (RBF) nets, and polynomial classifiers as special cases.

SVM has yielded excellent generalization performance on a wide range of problems including bioinformatics [6,7,8], text categorization [9], image detection [10], etc. The SVM approach has been applied in several financial applications recently, mainly in the area of time series prediction and classification [11,12]. They reported that SVM was competitive and outperformed other classifiers (including neural networks and linear discriminant classifier) in terms of generalization performance [13]. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on [15].

Simple K-Means Clustering

The k-means is the simplest, most commonly and good behavior clustering algorithm used in many applications [17, 21]. The K mean algorithm works on the Euclidian Distance Method, is initialized from some random or approximate solution. Several researchers have identified that age, blood pressure and cholesterol are critical risk factors associated with heart disease [18, 19-20].

Performance Measure

To measure the stability of the performance of the proposed model the data is divided into training and testing data with 10-fold cross validation. The confusion matrix shows the

Tele:
E-mail address: atul.pandey.it2009@gmail.com

number of samples which have been correctly/falsely classified into the two classes of C1 and C2. The entries of this matrix are used to explain the performance measures.

Table 1. Confusion Matrix of Simple k-means, SVM, SVM via Simple K-means and Accuracy of algorithms

Predicted Class	Simple K-Means	
	1	0
Cluster 1 <-- 0	27	138
Cluster 0 <-- 1	107	31

Predicted Class	SVM via Simple K-Means Clustering	
	b	a
b=cluster 1	129	5
a=cluster 0	5	164

Predicted Class	Support Vector Machine	
	1	0
1	148	17
0	32	106

Classification Techniques	Accuracy %
Simple K-Means Clustering	80.8581
Support Vector Machine	83.8284
SVM via Simple K-Means	96.6997

We are evaluating the performance of Simple K-Means algorithm Clustering using the mode of classes to clusters evaluation with the prediction attribute nom. Table 1 illustrates the confusion matrix of Simple k-means, SVM, SVM via Simple K-means (Classification via Clustering) and Accuracy of algorithm respectively.

The resulted Clustered Instances have cluster 0 is 169 (56%) instances and cluster 1 is 134 (44%) on Classes to Clusters evaluation mode. Figure 1 represents Weka Clusterer Visualizer of Simple K-Means Clustering. Figure 2 and 3 illustrates the cost/benefit analysis of function SMO for the class cluster 0 and 1.

The Kernel used for the SMO function is polynomial kernel: $K(x, y) = <x, y>^p$ or $K(x, y) = (<x, y>+1)^p$ and filter type is normalize training data.

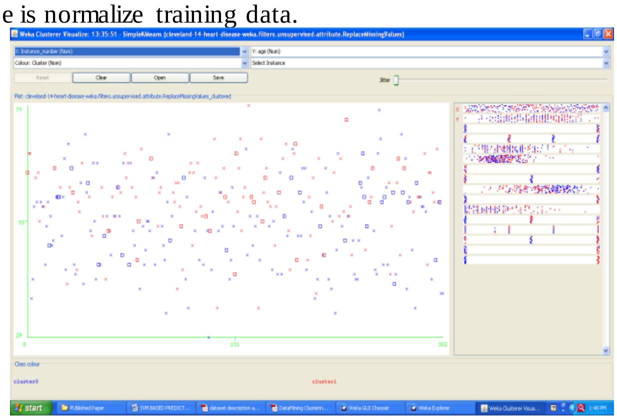

Fig1. Clusterer visualizer of simple k-means clustering

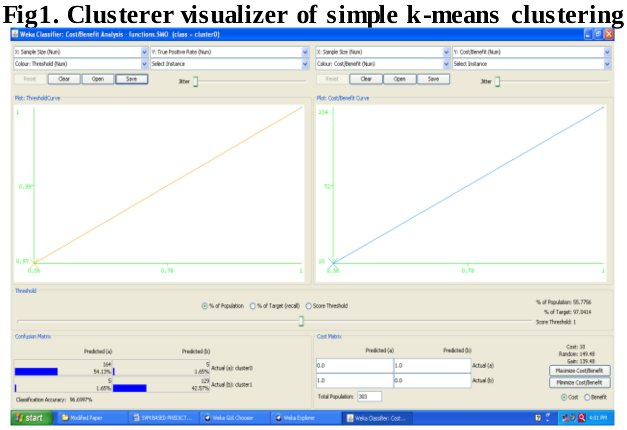

Fig 2. Cost/Benefit Analysis of Function Smo For Class-Cluster 0

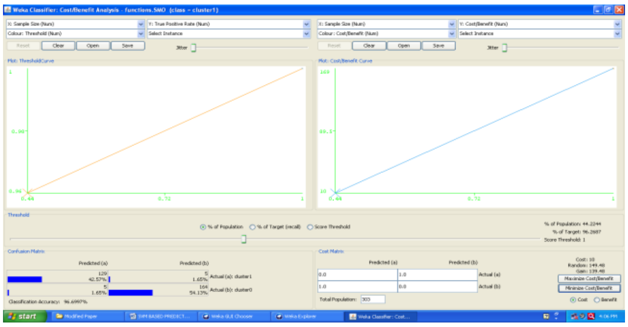

Fig 3. Cost/Benefit Analysis of Function Smo for Class-Cluster 1

Conclusion and Future Works

In this paper we have developed a Prediction Model for the diagnosis of Heart disease by means of Support Vector Machine via Simple K-Means Clustering. Therefore the diagnosis of Heart disease is carried out utilizing different data Mining Techniques results have denoted that SVM via Simple K-Means Clustering is equivalently good as the Simple K-Means Clustering and SVM in the diagnosis of Heart disease. The classification accuracy of Classification via Clustering has been found to be high thus making it a good option for the diagnosis.

This paper investigates integrating k-means clustering with SVM in the diagnosis of heart disease patients. The results show that integrating Simple K-Means Clustering and Support Vector Machine can enhance Support Vector Machine accuracy in the diagnosis of heart disease patients. The results also show that the PolyKernel function and Normalizes Training Data Filter type could achieve higher accuracy than other kernel function in the diagnosis of heart disease patients. The best accuracy achieved is by two clusters regarding Classification via Clustering method showing accuracy of 96.69%. Finally, some limitations on this work are noted as pointers for future research.

References

J. Galindo, P. Tamayo, Credit risk assessment using statistical and machine learning: basic methodology and risk modeling applications, Computational Economics 15 (1 – 2) (2000) 107–143.

D. Michie, D.J. Spiegelhalter, C.C. Taylor, Machine Learning, Neural and Statistical Classification, Elis Horwood, London, 1994.

S.M. Weiss, C.A. Kulikowski, Computer Systems That Learn: Classification and Prediction Methods from Statistics, Neural Networks, Machine Learning and Expert Systems, Morgan Kaufmann, San Mateo, 1991.

V. Vapnik, The Nature of Statistical Learning Theory, Springer- Verlag, New York, 1995.

M.A. Hearst, S.T. Dumais, E. Osman, J. Platt, B. Scho¨lkopf, Support vector machines, IEEE Intelligent Systems 13 (4) (1998) 18–28.

M.P. Brown, W.N. Grudy, D. Lin, N. Cristianini, C.W. Sugnet, T.S. Furey, M. Ares, D. Haussler, Knowledge-based analysis of microarray gene expression data by using support vector machines. Proceedings of National Academy of Sciences 97 (1) (2000) 262–267.

T.S. Jaakkola, D. Haussler, Exploiting generative models in discriminative classifiers, in: M.S. Kearns, S.A. Solla, D.A. Cohn (Eds.), Advances in Neural Information Processing Systems, MIT Press, Cambridge, 1998.

A. Zien, G. Ra¨tsch, S. Mika, B. Scho¨lkopf, T. Lengauer, K.-R. Mu¨ller, Engineering support vector machine kernels that recognize translation initiation sites, Bioinformatics 16 (9) (2000) 799– 807.

⅄T. Joachims, Text categorization with support vector machines, Proceedings of the European Conference on Machine Learning (ECML), 1998.
⅄E. Osuna, R. Freund, F. Girosi, Training support vector machines: an application to face detection, Proceedings of Computer Vision and Pattern Recognition, 1997, pp. 130–136.
⅄F.E.H. Tay, L.J. Cao, Modified support vector machines in financial time series forecasting, Neurocomputing 48 (2002) 847– 861.
⅄T. Van Gestel, J.A.K. Suykens, D.-E. Baestaens, A. Lambrechts, G. Lanckriet, B. Vandaele, B. De Moor, J. Vandewalle, Financial time series prediction using least squares support vector machines within the evidence framework, IEEE Transactions on Neural Networks 12 (4) (2001) 809– 821.
⅄N. Cristianini, J. Shawe-Taylor, An Introduction to Support Vector Machines, Cambridge Univ. Press, Cambridge, New York, 2000.
⅄J.C.Platt, 1998, Sequential minimal optimization: A fast algorithm for training support vector machines, Technical report MSR-TR-98-14, Microsoft Research, 1998.
⅄C.J.C. Burges, 1998, A tutorial on support vector machines for pattern recognition, Data mining and knowledge discovery, 2(2) (1998)121-167.
⅄R. O. Bonow, D. L. Mann, D. P. Zipes, and P. Libby, Braunwald’s Heart Disease: A Textbook of Cardiovascular Medicine, *9th Edition*: New York, Saunders, 2012.
⅄Wu, X., et al., Top 10 algorithms in data mining analysis. Knowl. Inf. Syst., 2007.
⅄Heller, R.F., et al., How well can we predict coronary heart disease? Findings in the United Kingdom Heart Disease Prevention Project. BRITISH MEDICAL JOURNAL, 1984.
⅄Salahuddin and F. Rabbi, Statistical Analysis of Risk Factors for Cardiovascular disease in Malakand Division. Pak. j. stat. oper. res., 2006. Vol.II: p. pp49-56.
⅄Shahwan-Akl, L., Cardiovascular Disease Risk Factors among Adult Australian-Lebanese in Melbourne, International Journal of Research in Nursing, 2010. 6 (1).
⅄Bramer, M., Principles of data mining. 2007: Springer.