International Journal of Soft Computing 7 (5): 242-248, 2012
ISSN: 1816-9503

A Hybrid Classification Model for Multivariate Heart Disease Dataset Using Enhanced Support Vector Machine Technique

[1]G. NaliniPriya, [1]A. Kannan and [2]P. AnandhaKumar
[1]Department of Information Science and Technology,
Anna University, Chennai, Tamil Nadu, India
[2]Department of Information Technology, Anna University,
M.I.T. Campus, Chennai, Tamil Nadu, India

Abstract: In Medical Information Systems, the data available for the learning and prediction are multivariate in nature. Some of the classification models which were generally used in the design of medical decision support systems could not provide a good performance. In this study, researchers address the ways to improve the performance of a supervised learning based classification algorithm. For achieving this, researchers propose the use of statistical technique for performing effective decision making in medical application, screening and manipulating the training samples with little bit of Gaussian Distribution Random Values (GDRV) before using the data for training the neural network. This study present, a way to improve the performance of a neural network based classification model through the proposed biased training algorithm which has been evaluated with the Coronary Artery Disease (CAD) data sets taken from University California Irvine (UCI). The performance has been evaluated with standard metrics.

Key words: CAD, heart disease, classification, multivariate data, BPN, SVM, ANN, random values

INTRODUCTION

The rapidly growing aging population, the increased burden of chronic diseases and the increasing healthcare costs, there is an urgent need for the development, implementation and deployment in everyday medical practice of new models of healthcare services. Classification and other data mining (Chen *et al*., 1996; Fayadd *et al*., 1996; Khan and Kant, 2007; Tsang *et al*., 2009) algorithms play major role in designing computing environments in smart hospitals. For instance, classification algorithms are often useful in patient activity classification and the diagnosis of a disease using a multivariate clinical data which were acquired from the hospital environment using different technologies. This data may be the combination of different types. Development of computer methods for the diagnosis of heart disease attracts many researchers. At the earlier time, the use of computer is to build knowledge based decision support system which uses knowledge from medical experts and transfers this knowledge into computer algorithms manually. Clinical decisions are often made based on doctors' intuition and experience rather than on the knowledge rich data hidden in the database. This practice leads to unwanted biases, errors and excessive medical costs which affects the quality of service provided to patients. There are many ways that a medical misdiagnosis can present itself. Whether a doctor is at fault or hospital staff, a misdiagnosis of a serious illness can have very extreme and harmful effects. This process is time consuming and really depends on medical expert's opinion which may be subjective. To handle this problem, machine learning techniques have been developed in this research to gain knowledge automatically from examples or raw data.

MATERIALS AND METHODS

Coronary Artery Disease (CAD): Heart disease which is usually called Coronary Artery Disease (CAD) is a broad term that can refer to any condition that affects the heart (Das *et al*., 2009). CAD is a chronic disease in which the coronary arteries gradually hardens and narrow (Fig. 1). It is the most common form of cardiovascular disease and the major cause of heart attacks in all countries. Moreover, cardiovascular disease is the leading killer compared to other diseases. Many people with heart disease have symptoms such as chest pain and fatigue as many as 50% have no symptoms until a heart attack occurs. The data generally used for diagnosing the CAD

Corresponding Author: P. AnandhaKumar, Department of Information Technology, Anna University, M.I.T. Campus, Chennai, Tamil Nadu, India

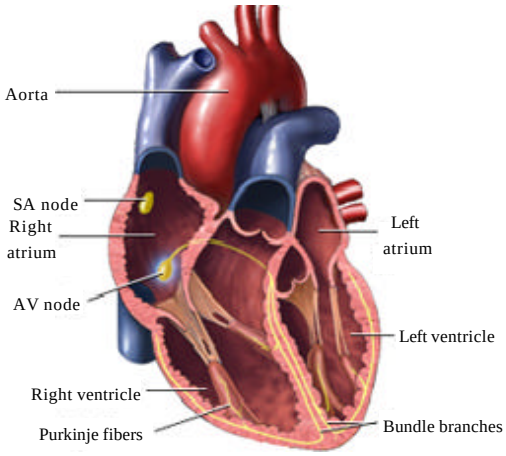

Fig. 1: Heart image

(Bergman and Bertero, 2001; Buetow and Coster, 2001; Bhatti *et al*., 2007) will be multivariate in nature. Having so many factors to analyze and diagnose the heart diseases, physicians generally make decisions by evaluating the current test results of the patients.

The earlier decisions made on other patients with the same condition are also examined by the physicians. These complex procedures are not easy when considering the number of factors that the physician has to evaluate. So, diagnosing the heart disease of a patient involves experience and highly skilled physicians. Recent advances in the field of artificial intelligence and data mining have led to the emergence of expert systems for medical applications. Moreover, in the last few decades computational tools have been designed to improve the experiences and abilities of physicians for making decisions about their patients.

Problem definition: The clinical data which will be used to diagnose a disease will be a mixed type of data which contains different types of attributes. classification of a data can be solved by using a lot of methods from simple methods such as nearest neighbor method to complex methods such as decision trees, neural networks and genetic algorithms. However, it is known that classification of multivariate data are a difficult problem because of several reasons. Generally, this kind of medical data or multivariate data will contain an error and missing values and will not always be pure. Further, the mutual dependence of attributes or variables causes distortion of the space. Due to an effect called boundary effect the nearest points seem to be rather far and farther points near and this causes considerable error in distance calculation during the clustering or classification process. So most of the algorithms which are used for classification cannot be applied on multivariate data.

Motivated by the need of such an expert system, this study, researchers propose a Hybrid Model for improving the performance of machine learning based classification (Sahbi, 2011; Joachims, 1999; Wei *et al*., 2010) system. Researchers have selected to improve the performance of a machine learning based classification system because in the previous evaluation on different classification algorithms, researchers observed that the SVM based classification algorithm produced a maximum accuracy of 85.56 which was almost good compared to that of several earlier researchs. As far as researchers evaluated, the Supervised Learning Methods are providing more promising results than other methods. But the only realized problem in getting more improved accuracy is the selection of training samples because, the performance of these supervised learning methods were very much depend on the training and testing samples and the achieved accuracy also very random with respect to the training and testing samples.

So in this study, researchers address a simple methods for improving the training performance and make the algorithm to provide a constant performance in terms of accuracy and little a bit improvement in accuracy itself. For achieving this, researchers propose the use of statistical technique for screening and manipulating the training samples along with the neural network.

Classification methods under evaluation: Classification is one of the most useful tasks in the data mining process for discovering groups and identifying interesting distributions and patterns in the underlying data. Classification is the process of grouping a set of physical or abstract objects into classes of similar objects. A cluster is a collection of data objects that are similar to one another with the same cluster and are dissimilar to the objects in other clusters. A cluster of data objects can be treated collectively as one group in many applications. Classification is a form of learning by observation rather than learning by examples. Cluster analysis is an important human activity in which researchers indulge since childhood when researchers learn to distinguish between animals and plants, etc. by continuously improving subconscious classification schemes. This has been widely used in numerous applications including pattern recognition, data analysis, image processing, market research, etc.

Classification is a very important application area but widely interdisciplinary in nature that makes it very difficult to define its scope. It is used in several research

communities to describe methods for grouping of unlabeled data. Now these communities have different terminologies and assumptions for the components of the classification process and the contexts in which classification is used.

Supervised Learning Methods: In the training, a Neural Network Model essentially means selecting one model from the set of allowed models that minimizes the cost criterion. There are numerous algorithms available for training neural network models; most of them can be viewed (Setiawan *et al*., 2009) as a straightforward application of optimization theory and statistical estimation. Most of the algorithms used in training artificial neural networks (Masud *et al*., 2011; Rifkin and Klautau, 2004) are employing some form of gradient descent. This is done by simply taking the derivative of the cost function with respect to the network parameters and then changing those parameters in a gradient-related direction. Evolutionary methods simulated annealing and expectation-maximization and non-parametric methods are among other commonly used methods for training (Wang, 2011; Lim *et al*., 2005) neural networks.

Support Vector Machines (SVM): Support vector machines are a set of related supervised learning methods used for classification and regression. Viewing input data as two sets of vectors in an n-dimensional space, an SVM will construct a separating hyper plane in that space, one which maximizes the margin between the two data sets. To calculate the margin, two parallel hyper planes are constructed, one on each side of the separating hyper plane which is pushed up against the two data sets. Intuitively, a good separation is achieved by the hyper plane that has the largest distance to the neighboring data points of both classes, since in general the larger the margin the better the generalization error of the classifier.

SVM is selected as the classifying function. One distinctive advantage this type of classifier has over traditional neural networks (Cheng *et al*., 2010; Chen and Yao, 2010; Yang and Ong, 2011) is that SVMs can achieve better generalization performance. Support vector machine is a pattern classification algorithm developed by Vapnik. SVM originally designed to solve problem where data can be separated by a linear decision boundary. By using kernel functions, SVMs can be used effectively to deal with problems that are not linearly separable in the original space. Some of the commonly used kernels include Gaussian Radial Basis Functions (RBFs), polynomial functions and sigmoid polynomials whose decision surfaces are known to have good approximation properties. Relying on the fact that the training data set is not linearly separable, a Gaussian Radial Basis Function (RBF) kernel is selected in this study. The RBF kernel performs usually better for the reason that it has better boundary response as it allows for extrapolation.

The proposed biased training algorithm: Let D_{train} be the set of healthy records and sick records to be normally trained with the support vector machine. Where m is the total number of healthy records, n is the total number of sick records, d_{ij} is the Euclidean distance between two records. d_{min} is a minimum expected distance between healthy and sick records. N is a set of statistically similar pairs from the healthy and sick records. $D_1 = D_{train} \cap N$ is the records which can be used directly to train the SVM.

But, set N will contain records which are pair of records which will be statistically very very similar to one another. Researchers believe that this insignificant difference between the records of two different categories will lead to poor training in any machine learning based model. To improve the training performance and recognition rate, researchers add bit of Gaussian distribution set of random values (Dutta *et al*., 2003; Bansal *et al*., 2008; Qui *et al*., 1989) GDRV (η) on the statistically weak set of data N:

$$D_2 = N + \eta$$

where, η is a set of random values (Bansal *et al*., 2008; Ince *et al*., 2011; Das *et al*., 2009) of the same size of the set individual attributes of the set N and will be added to the corresponding individual attributes set of N records. This set of values can be chosen randomly or based on the earlier experience of training performance with the SVM:

$$D_{trainNew} \leftarrow D_1 \cup D_2$$

Now, if researchers train the SVM with the newly constructed $D_{trainNew}$, researchers can expect better training of the SVM due to the statistically significant training records. Figure 2 shows the comparative analysis of SVM and the proposed FSR based SVM algorithm of the experiments conducted on CAD (Setiawan *et al*., 2009) data. Figure 2 explains the proposed system steps and also it compares with existing system.

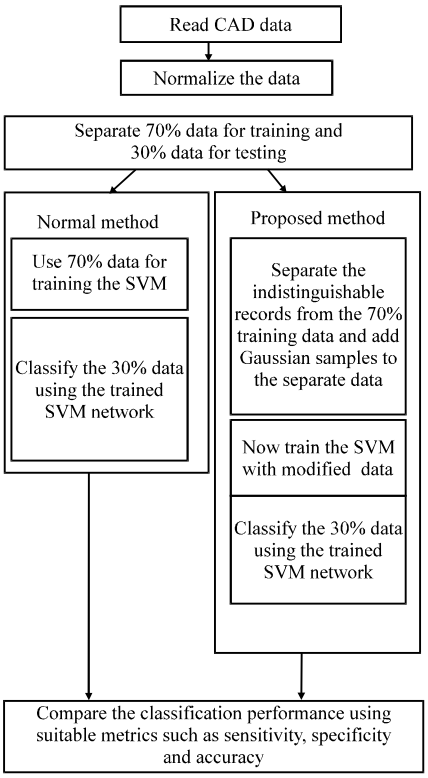

Fig. 2: The comparative diagram of the proposed model

IMPLEMENTATION AND EVALUATION

To evaluate the proposed algorithm, a suitable and standard multivariate data set is needed. A suitable UCI data set called Cleveland data, concerning heart disease diagnosis is used for the evaluation of the algorithms under consideration. This data was originally provided by Clevel and Clinic Foundation. This database contains 303 records with 13 attributes which have been originally extracted from a larger set of 75 attributes and a class attribute among the 303 records, 164 belongs to healthy and remaining are from diseased. Researchers have successfully designed this algorithm with object oriented concepts of rational rose software. The Fig. 3 explains the class diagram of the classification model.

Data preprocessing: As far as the proposed Clevend data set is concerned, only few of the records contained missing values. Researchers just removed them. So that the total records becomes 297. Almost all the classification algorithms will researchers good if the input data is in normalized form. In the proposed evaluation system after loading the CAD data, the data will be normalized by dividing each value of the attribute with the maximum value of that particular attribute (or column).

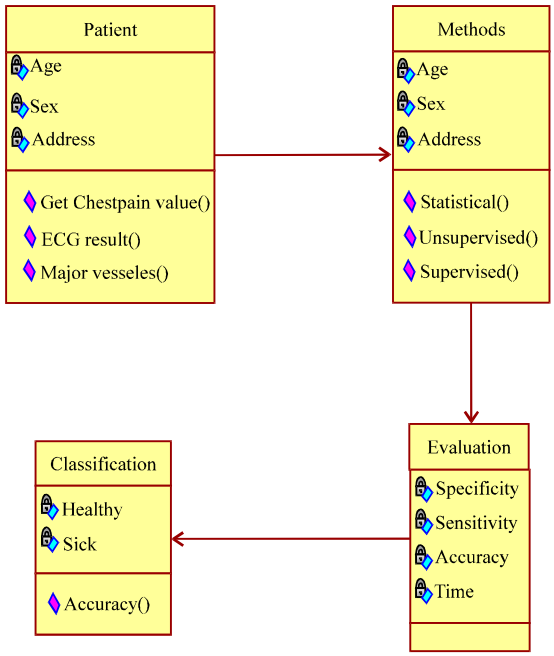

Fig. 3: The class diagram of the classification system

This will lead to values between 0 and 1 which will be suitable for almost all the classification algorithm.

Metrics considered for evaluation

Sensitivity: Sensitivity measures the proportion of actual positives which are correctly identified as such (the percentage of sick people who are correctly identified as having the condition):

$$\text{Sensitivity} = \frac{\text{Number of true positives}}{\text{Number of true positives} + \text{Number of false negatives}}$$

Specificity: Specificity measures the proportion of negatives which are correctly identified (the percentage of healthy people who are correctly identified as not having the condition):

$$\text{Specificity} = \frac{\text{Number of true negatives}}{\text{Number of true negatives} + \text{Number of false positives}}$$

Accuracy: Accuracy of a measurement system is the degree of closeness of measurements of a quantity to its actual (true) value:

$$\text{Accuracy} = \frac{\text{Number of true positives} + \text{Number of true negatives}}{\text{Number of True positives} + \text{False positives} + \text{False negatives} + \text{True negatives}}$$

RESULTS AND DISCUSSION

In this study, two diagnosis classes are considered such as healthy and sick. From the related research survey researchers came to know that various methods have been proposed for diagnosis of the heart disease. The accuracy is tabulated. In this (Das *et al*., 2009) study only statically similar data set is considered and evaluated. If there is no dissimilarity in data then the classification is very easy and researchers can get the accuracy in the higher side. Since, researchers are handling multivariate dataset classification is very challenging.

The result of the classification is measured in terms of metrics such as sensitivity, specificity and accuracy. The results are tabulated in the Table 1. This table showing the comparative result of classification. The following table is showing the performance of the normal SVM based algorithm along with the proposed change in the algorithms. The experiments were repeated for different set of randomly shuffled samples and the significant results were tabulated.

Results with Cleveland dataset: The accuracy is the important collective measure which is directly showing the overall classification performance of the algorithms. In terms of accuracy, the SVM are the only two supervised learning based classification algorithms which produced acceptably good results. The results of normal method and the proposed method more significant and comparable. These performance were almost equal to that of some of the previous methods mentioned in (Bergman and Bertero, 2001; Wang, 2011; Buetow and Coster, 2001).

Performance in terms of metrics: In the implementation, the proposed SVM based method provided almost good results compared with earlier methods. The Fig. 4 shows the performance of the proposed and normal methods of SVM. Even in some cases the implementation of SVM produced better results. But researchers observed that it was purely dependent on the randomly selected traing sets and the testing sets. The Fig. 5 represents the sequence diagram of the classification system.

In terms of time, SVM provided best performance than all other methods. After several, repeated analysis researchers made on the classification algorithms for classifying the CAD data, researchers came to the following conclusion. The supervised learning algorithms lead to better results. In fact the results with highest accuracy were only achieved through the proposed method. But researchers observed a very important fact during experimenting with these supervised learning

Table 1: The results with cleveland dataset

	Normal method			Proposed method		
Trials	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy
1	62.79	93.62	78.89	69.77	93.48	82.02
2	89.47	78.85	83.33	89.47	82.35	85.39
3	78.57	87.23	83.15	76.19	91.49	84.27
4	85.00	90.00	87.78	85.00	93.88	89.89
5	68.29	95.83	83.15	73.17	93.75	84.27
Avg.	76.82	89.11	83.26	78.72	90.99	85.17

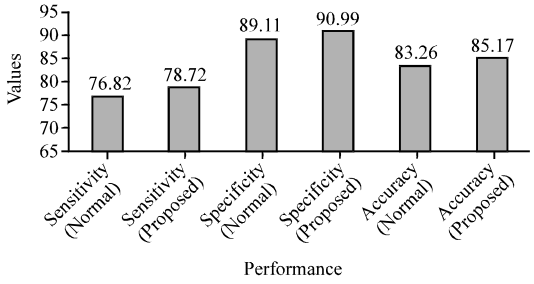

Fig. 4: The performance in terms of accuracy

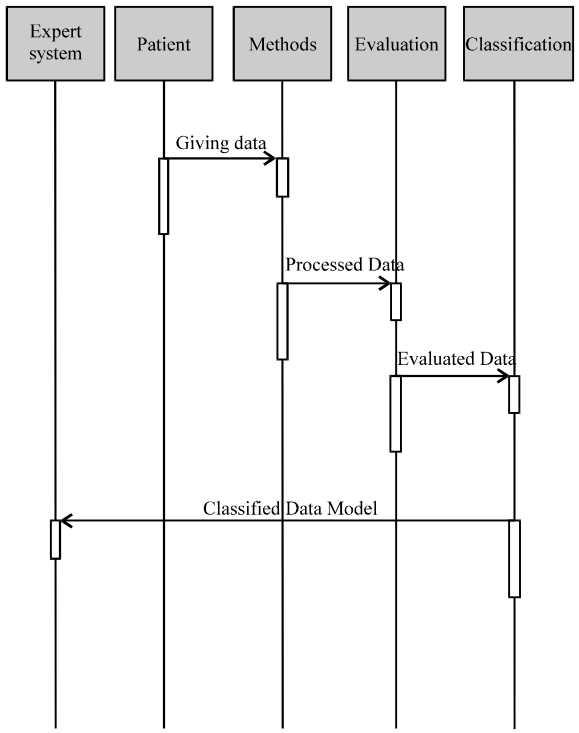

Fig. 5: Sequence diagram of the classification system

algorithm particularly with the CAD dataset, the accuracy of classification was very much depending upon not only the training set but also with the testing set. So that while randomly selecting 70% data for training and remaining 30% data for testing, the achieve accuracy was better that other unsupervised methods but the highest accuracy was achieved only for some particular random sets.

CONCLUSION

In this study, researchers propose a method for enhancing the accuracy of classification algorithms for the diagnosis of coronary artery disease. The result obviously shows the complex nature of data set restricts these algorithms from achieving better accuracy if researchers directly train the SVM with the data. It was realized that the reason for this poor classification is due to the insignificant difference between some of the records of the two classes under classification. The statistically indistinguishable records were separated from the training set and Gaussian distribution random values were added to them and make them somewhat distinguishable from one another. This lead to better performance in terms of accuracy of classification. The average improvement was around 2%. But as far as the complexity of this classification problem is concerned, this little improvement is more significant and considerable.

REFERENCES

Bansal, R., L.H. Staib, D. Xu, AF. Laine, J. Royal and B.S. Peterson, 2008. Using perturbation theory to compute the morphological similarity of diffusion tensors. IEEE Trans. Med. Imag., 27: 589-607.

Bergman, E. and C. Bertero, 2001. You can do it if you set your mind to it: A qualitative study of patients with coronary artery disease. J. Adv. Nurs., 36: 733-741.

Bhatti, R., A. Samuel, M.Y. Eltabakh, H. Amjad and A. Ghafoor, 2007. Engineering a policy-based system for federated healthcare databases. IEEE Trans. Knowledge Data Eng., 19: 1288-1304.

Buetow, S.A. and G.D. Coster, 2001. Do general practice patients with heart failure understand its nature and seriousness and want improved information? Patient Edu. Counseling, 45: 181-185.

Chen, H. and X. Yao, 2010. Multiobjective neural network ensembles based on regularized negative correlation learning. IEEE Trans. Knowl. Data Eng., 22: 1738-1743.

Chen, M.S., J. Han and P.S. Yu, 1996. Data mining: An overview from a database perspective. IEEE Trans. Knowledge Data Eng., 8: 866-883.

Cheng, H., P.N. Tan and R. Jin, 2010. Efficient algorithm for localized support vector machine. IEEE Trans. Knowl. Data Eng., 22: 381-389.

Das, R., I. Turkoglu and A. Sengur, 2009. Diagnosis of valvular heart disease through neural networks ensembles. Comput. Methods Programs Biomed., 93: 185-191.

Dutta, H., H. Kargupta, S. Datta and K. Sivakumar, 2003. Analysis of privacy preserving random perturbation techniques: Further explorations. Proceedings of the Workshop on Privacy in the Electronic Society, October 27-30, 2003, New York, USA., pp: 31-38.

Fayadd, U., G. Piatesky-Shapiro and P. Smyth, 1996. From Data Mining to Knowledge Discovery in Databases. AAAI/ MIT Press, Massachusetts, USA.

Ince, T., S. Kiranyaz, J. Pulkkinen and M. Gabbouj, 2011. Evaluation of global and local training techniques over feed-forward neural network architecture spaces for computer-aided medical diagnosis. Expert Syst. Appli. : Int. J., 37: 8450-8461.

Joachims, T., 1999. Transductive inference for text classification using support vector machines. Proceedings of 16th International Conference on Machine Learning, Jun. 27-30, Morgan Kaufmann Publishers Inc. San Francisco, CA, USA., pp: 200-209.

Khan, S.S. and S. Kant, 2007. Computation of initial modes for k-modes clustering algorithm using evidence accumulation. Proceedings of the 20th International Joint Conference on Artifical intelligence, January 6-12, 2007, Hyderabad, India, pp: 2784-2789.

Lim, C.P., J.H. Leong and M.M. Kuan, 2005. A hybrid neural network system for pattern classification tasks with missing features. Trans. Pattern Anal. Mach. Intell., 27: 648-653.

Masud, M.M., J. Gao, L. Khan, J. Han and B.M. Thuraisingham, 2011. Classification and novel class detection in Concept-drifting data streams under time constraints. IEEE Trans. Knowledge Data Eng., 23: 859-874.

Qui, J., S.M. Shahidehpour and Z. Schuss, 1989. Effect of small random perturbations on power dynamics and its reliability evaluation. IEEE Trans Power Syst., 4: 197-204.

Rifkin, R. and A. Klautau, 2004. In defence of One-vs-all classification. J. Machine Learn. Res., 5: 101-141.

Sahbi, H., 2011. Context-dependent kernels for object classification. IEEE Trans. Pattern Anal. Mach. Intell., 33: 699-708.

Setiawan, N.A., P.A. Venkatachalam and M.H. Ahmad Fadzil, 2009. Rule selection for coronary artery disease diagnosis based on rough set. Int. J. Recent Trends Eng., 2: 198-202.

Tsang, S., B. Kao, K.Y. Yip, W. Ho and S.D. Lee, 2009. Decision trees for uncertain data. Proceedings of the 25th IEEE International Conference of Data Engineering, March 29-April 2, 2009, Shanghai, China, pp: 441-444.

Wang, B., 2011. ELITE: Ensemble of optimal input-pruned neural networks using trust-tech. IEEE Trans. Neural Networks, 22: 96-107.

Wei, J.M., S.Q. Wang and X.J. Yuan, 2010. Ensemble rough hypercuboid approach for classifying cancers. IEEE Trans. Knowl. Data Eng., 23: 381-391.

Yang, J.B. and C.J. Ong, 2011. Determination of global minima of some common validation functions in support vector machine. IEEE Trans. Neural Networks, 22: 654-659.