SVM Based Decision Support System for Heart Disease Classification with Integer-Coded Genetic Algorithm to Select Critical Features

Sumit Bhatia, Praveen Prakash, and G.N. Pillai

***Abstract*—This paper presents a decision support system for heart disease classification based on support vector machine (SVM) and integer-coded genetic algorithm (GA). Simple Support Vector Machine (SSVM) algorithm has been used to determine the support vectors in a fast, iterative manner. For selecting the important and relevant features and discarding the irrelevant and redundant ones, integer-coded genetic algorithm is used which also maximizes SVM's classification accuracy. The Cleveland heart disease database is used in this study and it consists of 303 cases divided in 5 classes, each with 13 diagnostic features. The results of the 5-class classification problem indicate an increase in the overall accuracy when using the optimal feature subset, accuracy achieved being 72.55% indicating the potential of the system to be used as a practical decision support system. As a two class problem, disease or no disease, the proposed method gives an accuracy of 90.57% which shows an improvement over the existing methods.**

***Index Terms*—Medical diagnostic decision support systems, SVM, Feature subset selection, Integer coded genetic algorithm, Multi class classification.**

I. Introduction

In modern times, the number of people suffering from heart disease is on a rise. A large number of people die every year due to heart disease all over the world and it is the leading cause of death in United States [1]. However, accurate diagnosis at an early stage followed by proper subsequent treatment can result in significant life saving [2]. Unfortunately, correct diagnosis of heart disease at an early stage is quite a demanding task due to complex interdependence on various factors [2]. Hence, there is a pressing need to develop medical diagnostic decision support systems which can aid medical practitioners in the diagnostic process.

A system for automatic diagnosis of heart diseases using neural network is described in [3]. The system uses features extracted from the ECG data of the patients. Another decision support system using multi layer perceptrons (MLPs) and back propagation algorithm is described in [2]. The system is used for classifying 5 major heart diseases using 38 input variables with an appreciable accuracy level (63.6% - 82.9%). Support Vector machines have also been utilized in decision support systems such as [4]. A genetic algorithm to select optimal feature subset for use with back propagation artificial neural networks has been described in [5]. The experiments however, have been performed taking the Cleveland database as a 2 class dataset. A genetic algorithm for feature selection as well as for optimization of SVM parameter has been proposed in [6]. Very recently, a real coded Genetic algorithm for critical feature analysis for heart disease diagnosis has been described in [7].

In this paper we present a decision support system for heart disease classification using support vector machine. The dataset used is the Cleveland Heart Database taken from UCI learning data set repository which was donated by Detrano [8], [9]. The dataset is being divided into five classes, 0 corresponding to absence of any disease and 1,2,3,4 corresponding to four different types of diseases. To the best of our knowledge, all the published works have used the dataset to differentiate between the absence (0) and presence (1, 2, 3 or 4) of a disease. In the present work, we classify the data into 5 classes, thus classifying the type of disease as well. We use a fast iterative algorithm known as Simple Support Vector Machine algorithm (SSVM) [10] to find the support vectors for building the SVM. Further, since a practical decision support system must be highly time efficient, therefore in order to avoid the curse of dimensionality, we use a Genetic Algorithm to select an optimal feature subset which maximizes the SVM classification accuracy with a reduced number of features.

The rest of the paper is organized as follows. Section 2 describes the theory of Support Vector Machine and introduces SSVM algorithm. Section 3 describes the genetic algorithm employed and section 4 discusses the computational experiments. Section 5 concludes the paper with some general remarks.

II. Support Vector Machines

Support Vector Machine (SVM) is a class of supervised learning algorithms first introduced by Vapnik [11]. SVM is a learning technique which trades off accuracy for generalization error. SVMs build a hyperplane which divides examples such that examples of one class are all on one side of the hyperplane, and examples of the other class are all on the other side.

Consider input data of the form (x_i, y_i) where the vectors

Sumit Bhatia and Praveen Prakash have completed his B. Tech. in Electrical Engineering from Indian Institute of Technology, Roorkee, India in June2008. (e-mail: sumit.summit@gmail.com, pravnuee@gmail.com)

G. N. Pillai was with the Electrical Engineering Department, N. I.T. Kurukshera, India. He is now with the Electrical Engineering Department, Indian Institute of Technology Roorkee, Roorkee (phone: +91-1332-285198, +91-1332-273560, e-mail: gnathfee@iitr.ernet.in)

x_i are in a dot product space H, and y_i are the class labels. Formally, any hyperplane in H is defined as

$$\{x \in H \mid \langle w, x \rangle + b = 0\} \quad w \in H,\ b \in R$$

where w is a vector orthogonal to the hyperplane and $\langle\ \rangle$ represents the dot product. In an SVM, the idea is to find the hyperplane that maximizes the minimum distance from any training data point (Fig 1). The following constraint problem describes the optimal hyperplane:

$$\min_{w \in H, b \in R} \tau(w) = \frac{1}{2}\|w\|^2$$

subject to $y_i\left(\langle x_i, w \rangle + b\right) \geq 1$

for i = 1, 2, …, m where m is the number of training examples.

The above problem can be solved by introducing the Lagrange multipliers ($\alpha_i \geq 0 (i = 1, ..., m)$) and maximizing the following dual problem

$$\max_{\alpha \in R} W(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle$$

subject to $\alpha_i \geq 0, i = 1, ..., m,$

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

The patterns x_i which correspond to non-zero Lagrange coefficients are called support vectors. The resultant decision function has the following form

$$y(x) = \operatorname{sgn}\left(\sum_{i=1}^{m} \alpha_i y_i \langle x_i, x \rangle + b\right)$$

Thus the optimal margin hyperplane is represented as a linear combination of training points. Consequently, the decision function for classifying points with respect to the hyperplane only involves dot products between points. The algorithm that finds a separating hyperplane in the feature space can be stated entirely in terms of vectors in the input space and dot products in the feature space.

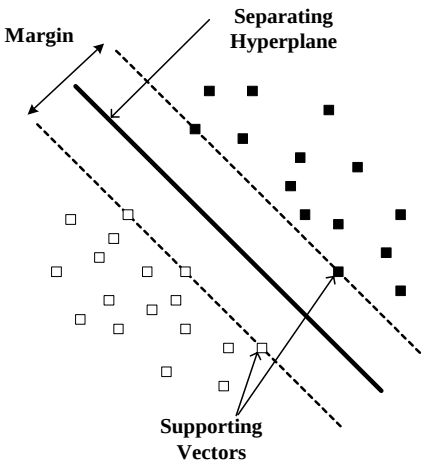

Fig. 1 Maximum margin and optimal hyperplane

When the samples are not linearly separable, a kernel function is used to transforms the data to a higher dimensional space where it is linearly separable. The kernel function gives the dot product of the two examples in the higher dimensional space without actually transforming them into that space. This notion, dubbed the kernel trick, allows us to perform the transformation for purposes of classification to large dimensional spaces. In the nonlinear case, resultant decision function has the following form

$$y(x) = \operatorname{sgn}\left(\sum_{i=1}^{m} \alpha_i y_i K(x_i, x) + b\right)$$

where the kernel function $K(x_i, x) = \langle \varphi(x_i), \varphi(x) \rangle$ and $\varphi(x)$ is the nonlinear map from original space to the high dimensional space. Two of the most commonly used kernel functions are polynomial functions and Gaussian radial basis functions and are given by

Polynomial kernel: $K(x, x') = \left[1 + x^T X'\right]^k \quad k = 2, 3$

Radial basis kernel: $K(x, x') = \exp\left[-\frac{1}{2}\|x - x'\|^2\right] \Big/ \sigma$ where σ s the spread of the Gaussian function.

The maximum margin allows the SVM to select among multiple candidate hyperplanes; however, for many data sets, the SVM may not be able to find any separating hyperplane at all, either because the kernel function is inappropriate for the training data or because the data contains mislabelled examplese The latter problem can be addressed by using a 'soft margin' that accepts some misclassifications of the training examples. In this case introducing slack variables ξ_i and error penalty C, the optimal hyperplane can be found by solving the following new quadratic optimization problem [12]:

$$\text{minimize } \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m} \xi_i$$

$$\text{subject to } y_i\left(\langle x_i, w \rangle + b\right) \geq 1 - \xi_i$$

Though the concept of SVM was originally proposed for binary classification, various methods have been proposed to use SVM for multi-class problems also. "One against One" and "One against All" methods are among the most popular methods for multi-class classification problems [13]. The former involves constructing ${}^{p}C_2$ binary classifiers, one for each pair of a total of p classes. The final class of the test point is determined by a pre-defined voting mechanism. In the "One against All" method, there is a binary classifier for each class to separate the members of that class from all other classes. 'One against one method' is a better method in many applications [14].

The SSVM algorithm was introduced by Vishwanathan and Murty to compute support vectors for a given set of points efficiently in an iterative manner [10]. The algorithm works by maintaining a candidate Support Vector set and updates this candidate set with every iteration. It uses a greedy approach to pick points for inclusion in the candidate set and backtracking approach is used to prune away points which are already present in the candidate set. The candidate set is initialized with nearest pair of points from opposite classes so as to speed up the whole process. The algorithm makes repeated passes over the data to satisfy the KKT constraints [10].

III. The Integer Coded Genetic Algorithm

Genetic Algorithms (GA) are a family of computational models inspired by evolution. These algorithms encode a potential solution to a specific problem on a simple chromosome like data structure and apply recombination operators to these structures so as to preserve critical information [15]. In this paper, we use an integer coded genetic algorithm to select top *N* best features for classification out of total *M* features. Since we are interested in selecting an optimal feature subset for classification, we use classification accuracy achieved by a feature subset as its fitness value. The details about SVM used for classification follow later. The algorithm proceeds with the generation of initial population which consists of 50 chromosomes. Each chromosome is an array of size equal to *N* and represents a feature subset. The elements of the array are assigned randomly generated values from 1 to *N*, each element corresponding to a feature.

The initial population thus generated is subjected to *Tournament Selection* where the fitness values of consecutive pairs of chromosomes are compared and the winner is selected for Crossover operation. All the selected chromosomes are then subjected to *One-point Crossover* where the portions of two consecutive chromosomes after a randomly generated crossover site are interchanged.

Crossover operation is followed by the mutation operation which maintains the diversity from one generation of population to next by randomly changing a gene sequence of a chromosome with certain probability. Accordingly, a mutation site is generated for each chromosome with a probability and the value at that site is replaced by a randomly generated integer from 1 to *N*.

Generation of chromosome having similar genes is undesirable as this means that a particular feature is repeated in the sub set under consideration. Hence this repetition is removed through a step which we call as 'Remove Repetition'. This operation is performed on each chromosome where each repetition, if any, is replaced by a randomly selected number from the set ($Z_n - G_i$), where Z_n is the set of integers from 1 to *N* and G_i is the set of numbers present in the i^{th} chromosome. Fig. 2 represents this step termed as 'Remove Repetition'.

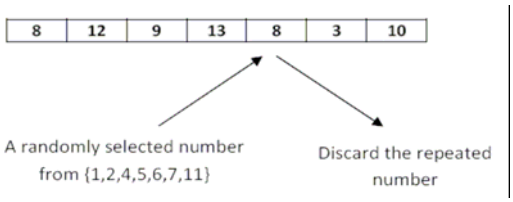

Fig. 2: Remove Repetition

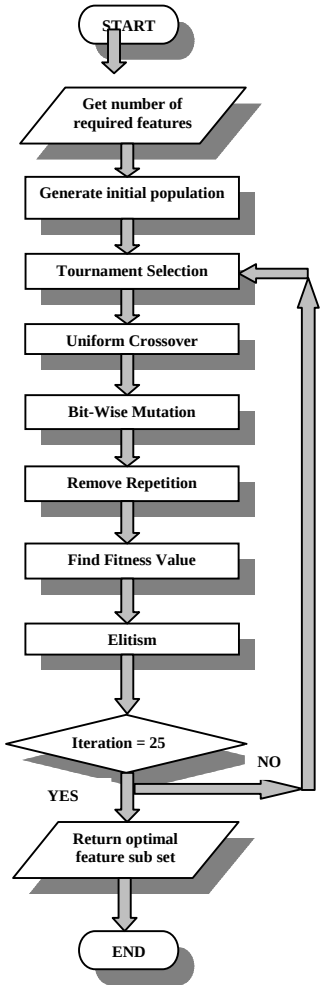

Fig. 3: Flowchart describing the Genetic Algorithm used.

The populations before crossover and after mutation are then combined together to form a population of size double the size of initial population with all the individuals arranged in ascending order of their fitness values. The top half individuals are then selected for the next generation and the cycle continues with the population of the original size. The step-by-step procedure followed is depicted in Fig. 3.

IV. Computational Experiments

A. Dataset used

In order to test our approach we use the Cleveland Heart Database taken from UCI learning data set repository which was donated by Detrano [8-9]. The data set consists of 13 numeric attributes which include age, sex, chest pain type, resting blood pressure, cholesterol, fasting blood sugar, resting ECG, maximum heart rate, exercise induced angina, oldpeak, slope, number of vessels coloured and thal respectively. The classes include integers valued from 0 (no presence) to 4 (types of heart diseases). Total number of patients instances is 303 and 250 of them are used for training and rest are used for testing the SVM.

B. Experimental Setup

The classification task was performed using Support Vector Machines (SVM) implemented using simpleSVM software kit [16]. The software provides various options to

tune the SVM's performance and following experiments were conducted to fine tune the performance of SVM:

a) Type of Multiclass SVM

The simpleSVM toolkit provides option of using either the *One against One* or the *One against All* model for multi class classification problems. The penalty term can be varied to obtain best classification results. Different kernels can be used and the parameters of the defining kernel functions can be changed. We tested the performance of SVM on all different combinations of these parameters keeping all other parameters constant and the results obtained are summarized below:

b) The Kernel Function

The software kit provides the option to use polynomial or radial basis function (RBF) kernel functions which are defined as follows:

I. The Polynomial Kernel

$$K_p(X,X_i) = (1 + X^T X_i)^p$$

The exponent p is specified by the user.

II. The Radial Basis Function Kernel

$$K_r(X,X_i) = \exp\left(\frac{-\|X - X_i\|^2}{2\sigma^2}\right)$$

The width σ^2 is width specified by the user.

We tested the performance of SVM with different parameter values for the polynomial and RBF kernels and it was concluded that the RBF kernel with parameter value 0.025 outperforms in terms of classification accuracy.

The performance of SVM was evaluated for a series of values of the penalty parameter C and it was concluded that optimal performance is obtained around C = 150. The algorithm gives robust result around this value of C. Variation of SVM's classification accuracy is being summarized in Fig. 4. After fine tuning the SVM's parameters, the Genetic Algorithm described in section 3 was run using following parameters:

- Number of generations = 25
- Population Size = 50
- Crossover Probability = 0.8
- Mutation Probability = 0.2

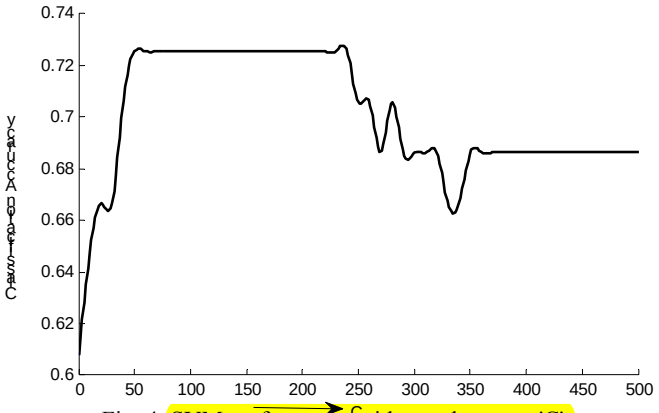

Fig. 4: SVM performance with penalty term 'C'.

The values of crossover probability and mutation probability were taken after extensive experimentation.

C. Experimental Results

The algorithm was used to select top *N* features out of a total of 13 features of the Cleveland Heart database. The results for different values of *N* are summarized in Table 1. It is to be noted that as the value of *N* is increased, the new optimal feature subset is obtained by addition of new features in the previous subset. This shows that addition of a feature affects the capability of that particular subset for classification either in a positive or negative manner and thus, there exists an optimal feature subset at which the accuracy is maximized.

Table 1: SVM performance with different feature subsets.

N	Feature Subset	Classification Accuracy
5	{3,7,8,12,13}	69.37%
6	**{4,3,7,8,12,13}**	**72.55%**
7	{1,4,3,7,8,12,13}	70.36%
13	{1,2,3...,12,13	61.93%

As a two class classification problem, 0 corresponding to absence of any disease and 1,2,3,4 corresponding to presence of heart disease, the proposed method gives a classification accuracy of 90.57% which is better than the previous results [4, 17]. Classification accuracy of 83.8% is obtained in [17] using 80 percent of data for training for the two class problem.

V. Conclusion

In this paper, a decision support system for heart disease classification is described. Integer coded Genetic Algorithm is used to find the optimal feature subset for maximizing SVM's classification accuracy with a reduced number of features. SSVM algorithm is used to find the support vectors in a fast, iterative manner. The algorithm is implemented using simpleSVM toolkit and the performance of SVM is fine tuned after carrying out detailed experimentation with different parameters provided with the toolkit. The maximum accuracy is obtained using 'one against one' multi-class SVM with RBF kernel with width 0.025 and penalty factor 150.The application of Genetic Algorithm for feature selection enhanced the performance of the SVM to a great extent and a high accuracy of 72.55% was obtained using only 6 out of 13 features, as against an accuracy of only 61.93% using all the features. As a two class problem, the proposed method gives an accuracy of 90.57% which is better than the existing methods. Future research involves more intensive testing using a larger heart disease database to get more accurate results.

References

[1] *Medline Plus: Heart Diseases.* http://www.nlm.nih.gov/medlineplus/heartdiseases.html

[2] Yan, H., Zheng, J., Jiang, Y., Peng, C. and Li, Q., 'Development of a decision support system for heart disease diagnosis using multilayer perceptron'; *IEEE Symposium on Circuits and Systems* (5) 2003,pp V709-V712

[3] Kumaravel, N., Sridhar,K.S., and Nithiyanandam, N., 'Automatic diagnoses of heart diseases using neural network', *Proccedings of the*

fifteenth Biomedical Engineering Conference, 29-31 March,1996, pp 319-322.

[4] Comak, E., Arslan, A. and Ibrahim, T. (2007).'A decision support system based on support vector machines for diagnosis of the heart valve diseases'. *Computers in biology and Medicine* (37), pp 21-27.

[5] Yang J. and Honavar, V., 'Feature Subset Selection using a Genetic Algorithm', *IEEE Intelligent Systems*, March-April 1998, pp 44-49.

[6] Huang, C.-L. and Wang, C.-J.,'A GA-based feature selection and parameters optimization for support vector machines', *Expert Systems with applications*, Vol.-31,(2006), pp -231-240.

[7] Yan, H., Zheng, J., Jiang, Y., Peng, C. and Xiao, C., ' Selecting critical clinical features for heart diseases diagnosis with a real-coded genetic algorithm',Applied Soft Computing, Vol.-8,(2008), pp 1105-1111.

[8] UCI Machine Learning Repository: Heart Disease Data Set.. http://archive.ics.uci.edu/ml/datasets/Heart+Disease

[9] Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J., Sandhu, S., Guppy, K., Lee, S., & Froelicher,V. International application of a new probability algorithm for the diagnosis of coronary artery disease. *American Journal of Cardiology*, 64,304—310, 1989.

[10] Vishwanathan, S. V., & Murthy, M. N. SSVM : A Simple SVM Algorithm. *International Joint Conference on Neural Networks*, *IJCNN'02*, Honolulu, HI, USA. pp. 2393-2398, 2002.

[11] Vapnik, V. N. The nature of statistical learning theory. New York: Springer, 1995

[12] Haykin, S., Neural Networks – A Comprehensive Foundation, Pearson Education (Singapore) Pte. Ltd., 2002.

[13] Liu, Y., & Zheng, F. Y., 'One-Against-All Multiclass SVM Classification Using Reliability Measures ', Proccedings of *IEEE International Joint Conference on Neural Networks, IJCNN '05*, 31 July-4 Aug. 2005, pp. 849-854.

[14] Hsu, C.W. and. Lin, C.J.,(2002), 'A comparison of methods for multi-class support vector machines', *IEEE Transactions on Neural Networks*, 2002, Vol. 13, No. 2., pp 415-425.

[15] Goldberg, D. E. (1989). Genetic algorithms in search optimization and machine learning. Reading, MA: Addison-Wesley.

[16] Gaelle Loosli, Simple SVM Toolbox, http://asi.insa-rouen.fr/enseignants/~gloosli/simpleSVM.html

[17] E. Boros, P. L. Hammer, T.Ibaraki, A. Kogan, 'An Implementation of Logical Analysis of Data' IEEE transactions on knowledge and data engineering, vol. 12, no. 2, March April 2000, pp. 292-306