GENETIC BASED SUPPORT VECTOR MACHINE CLASSIFIER FOR HEART DISEASE CLASSIFICATION

Dr. S. Nithya
Assistant Professor
Department of Information Technology
PSGR Krishnammal College for Women.

Abstract: Classification is one among the hot research topic in the field of data mining. Classically classification task represents the data to be categorized based on its features or characteristics. This proposed research work aims in developing genetic based support vector machine classifier. Support vector machine is a type of supervised machine learning technique and once when the dataset is given as input it performs the classification task by itself. The proposed classifier aims in improving the classification accuracy of the support vector machine by making use genetic algorithm. Genetic algorithm is used in order to perform fuzzy association rule extraction, candidate rule pre-screening, rule selection and lateral tuning. The proposed classifier has been tested onnamely PIMA Indian diabetes to classify the occurrence of heart disease among the patients. Performance metrics classification accuracy are taken for comparison of the proposed genetic based support vector machine classifier (GSVM) with SVM classification algorithm. Results showed that the proposed GSVM classifier gives better classification accuracy than that of support vector machine.

Keywords: Genetic, Classification, PIMA, Classification Accuracy.

I. INTRODUCTION

Data mining is the methodology used for discovering hidden information from the existing data. Knowledge discovery in data (KDD) is the trivial process involved in data mining for extracting the hidden knowledge from the data. Several conventional data mining algorithms are there which performs several tasks such as neighbourhood selection, classification, clustering, pattern matching, and information retrieval and so on. Broadly data mining can be classified as supervised data mining and unsupervised data mining. Supervised mechanisms are the one that has the ability to perform classification and prediction tasks. In recent years, machine learning algorithms are used as a supplement to perform data mining. Healthcare industry has abundant research problems that can be solved using data mining. Data mining using machine learning algorithms are used in order to undertake the research problem of classification in medical dataset which has several inputs. Fuzzy rule-based classification systems (FRBCSs) [Ishibuchi et al.,2004], [Kuncheva.,2000] are useful and well-known tools in the machine learning framework, since they can provide an interpretable model for the end user [Jin et al.,1999], [Ho et al.,2004], [Wang et al.,2005], [Zhang et al.,2011]. Support vector machine is one such machine learning algorithm and can be used for classification in data mining. Support vector machine can be trained from the historical / past data with the anticipation that it will determine hidden dependencies and that it will be able to use them for classification. The healthcare industry has volumes data and that need to be mined to discover hidden information for effective decision making. This research work aims to improve the performance of the support vector machine classifier by making the classical simplex method (SM) to modified simplex method (MSM). Genetic algorithm is used in order to perform fuzzy association rule extraction, candidate rule pre-screening, rule selection and lateral tuning. This paper is organized as the follows. This section introduces the scope of the research. Section 2 discusses on the related works pertaining to the chosen research problem. Section 3 discusses on the proposed work. Section 4 details on the chosen dataset with the results and discussions. Section 5 describes the concluding remarks.

II. LITERATURE REVIEW

Akhil jabbar et al.,2012 proposed a efficient associative classification algorithm using genetic approach for heart disease prediction. The main motivation of authors for using genetic algorithm in the discovery of high level prediction rules is that the discovered rules are highly comprehensible, having high predictive accuracy and of high interestingness values. Anbarasi et al.,2010 predicted more accurately the presence of heart disease with reduced number of attributes. Authors used Genetic algorithm to determine the attributes which contribute more towards the prediction. Amma.,2012 presented a diagnosis system for predicting the risk of cardiovascular disease. This system was built by combining the genetic algorithm and neural network. Multilayered feed forward neural networks are particularly suited to complex classification problems. Author used Genetic based neural network for training the system. Dalakleidi et al.,2013 presented a hybrid approach based on the combined use of a genetic algorithm (GA) and a nearest neighbours classifier for the selection of the critical clinical features which are strongly related with the incidence of fatal and non fatal Cardiovascular Disease (CVD) in patients with Type 2 Diabetes Mellitus (T2DM). In Niranjana Devi and Anto.,2014, an evolutionary fuzzy expert system was proposed for the diagnosis of the Coronary Artery Disease (CAD) based on Cleveland clinic foundation datasets for heart diseases. Decision tree was used to select the most significant attributes and the output was converted into crisp if-then rules. The crisp sets of rules are transformed into the fuzzy rules and these rules constitute the fuzzy rule base. Genetic Algorithm (GA) was used to tune the fuzzy membership functions and the optimized of membership functions using GA helps to achieve better accuracy. Murthy and Meenakshi.,2014 presented Neuro-genetic model for the prediction of coronary heart diseases. The novelty of this work is feature subset selection using multi-objective genetic algorithm without sacrificing the accuracy of ANN based heart disease predictor. Subsequently, the selected feature subset was used to predict the level of angiographic coronary heart disease using neural networks.

Amin et al.,2013 presented a technique which involves two most successful data mining tools, neural networks and genetic algorithms. The hybrid system uses the global optimization advantage of genetic algorithm for initialization of neural network weights. Dewan et al.,2015 developed a prototype which can determine and extract unknown knowledge (patterns and relations) related with heart disease from a past heart disease database record. It can solve complicated queries for detecting heart disease. Hanguang Xiao.,2012 proposed a diagnosis method using genetic algorithm (GA) and support vector machine (SVM) based on the acoustic characteristics of Parkinson's patients for improving the diagnosis accuracy. GA was applied into feature selection for improving the performance of SVM. Anirudha et al.,2014 proposed a Genetic Algorithm based Wrapper feature selection Hybrid Prediction Model (GWHPM). This model initially uses k-means clustering technique to remove the outliers from the dataset. An optimal set of features are obtained by using Genetic Algorithm based Wrapper feature selection. It was used to build the classifier models such as Decision Tree, Naive Bayes, k nearest neighbor and Support Vector Machine. Roohallah Alizadehsania et al.,2013 studied and applied several algorithms Zalizadeh Sani dataset(which utilizes several effective features.).

III. PROPOSED WORK

3.1.1 Genetic based SVM

The learning process in SVM involves the solution, which offers the architecture and parameters of a decision function representing the largest possible margin. Such parameters are represented by the vectors in the class boundary and their associated Lagrange multipliers. In order to take into account nonlinearities, a higher dimension space is obtained; this is done by transforming data vectors $x_i \in \Re^n$ through a function ϕ(x). In this transformation, explicit calculation of ϕ (x) is not necessary; instead of that, just the inner product between mapped vectors is required. For this inner product, kernel functions fulfilling the Mercer condition are usually used.To perform kernel functions analysis, an improved performance based approach is developed. This approach includes the common stages of the SVM training process:

01. Reading and Modelling data vectors using a kernel function,
02. Learning (Training), solving QPP obtained with data, and
03. Classification (where a classifier is built with Support Vectors found in the training stage).

This classifier is used to classify new data.The transformation into an Equivalent Linear Model allows using a variation of the classical Simplex Method (SM), and it is named Modified Simplex Method (MSM). Due to, MSM is based on SM, the former inherits a very important feature from later, which is that guarantees the global optimum solution (if it exists), and has been used in many practical problems. MSM is different to SM, mainly in the pivoting rule, particularly the way that the incoming variable is selected. In any iteration, a variable is a candidate to be the next incoming variable if it is a non-basic variable and it can potentially improve the objective function in the next iteration. In the λ-Y case, the selection process is as follows: If λ_i is an incoming variable candidate, it can be selected as the incoming one only if Y_i (i.e. the Y variable with the same index) is not in the basis. MSM indicates those columns (or variables) that are active (i.e. the basic variables) and inactive in the problem solution. On the other hand, if Y_i is an incoming variable candidate, it can be selected as the incoming one only if λ_i is not in the basis. Similar conditions must be fulfilled for variables α_j and u_j The next section makes use of the above said fuzzy logic technique and Improved SVM.

3.1.2 Codification and Initial Gene Pool:

To combine the rule selection with the global lateral tuning, a double coding scheme for both rule selection C_S and lateral tuning C_T is used.

1) For the C_s part, each chromosome is a binary vector that determines when a rule is selected or not (alleles “1” and “0,” respectively). Considering the M rules that are contained in the candidate rule set, the corresponding part, i.e., $C_s = \{c_1, \ldots, c_M\}$, represents a subset of rules composing the final RB so that IF $c_i = 1$ THEN ($R_i \in$ RB) else ($R_i \in$ RB), with Ri being the corresponding ith rule in the candidate rule set and RB being the final RB.
2) For the C_T part, a real coding is considered. This part is the joint of the α parameters of each fuzzy partition. Let us consider the following number of labels per variable: (m_1,m_2, . . . , m_n) with n being the number of system variables. Then, this part has the following form, where each gene is associated with the tuning value of the corresponding label: $C_T = (c_{11}, \ldots, C_{1m^1}, c_{21}, \ldots, C_{2m^2}, \ldots, c_{n1}, \ldots, C_{nm^n}.$

Finally, a chromosome C is coded in the following way : $C = C_s C_T$. To make use of the available information, all the candidate rules are included in the population as an initial solution. To do this, the initial pool is obtained with the first individual having all genes with value “1” in the C_s part and all genes with value “0.0” in the C_T part. The remaining individuals are generated at random.

3.1.3 Chromosome Evaluation:

To evaluate a determined chromosome penalizing a large number of rules, we compute the classification rate and the following function is maximized:

$$Fitness(C) = \frac{\#Hits}{N} - \delta . \frac{NR_{initial}}{NR_{initial} - NR + 1.0} \quad (20)$$

where #Hits is the number of patterns that are correctly classified, $NR_{initial}$ is the number of candidate rules, NR is the number of selected rules, and δ is a weighting percentage given by the system expert that determines the tradeoff between accuracy and complexity. If there is at least one class without selected rules or if there are no covered patterns, the fitness value of a chromosome will be penalized with the number of classes without selected rules and the number of uncovered patterns.

3.1.4 Crossover Operator:

The crossover operator will depend on the chromosome part where it is applied.

01. For the CT part, we consider the Parent Centric BLX (PCBLX) operator [Lozano et al.,2004] (an operator that is based on BLX-α). This operator is based on the concept of neighborhood, which allows the offspring genes to be around the genes of one parent or around a wide zone that is determined by both parent genes. Let us assume that $X = (x_1, \dots, x_n), and\ Y = (y_1, \dots, y_n)$, where $x_i, y_i \in [a_i, b_i] \subset R, i = 1, \dots, n$, are two real-coded chromosomes that are going to be crossed. We generate the following two offspring.
    a) $O_1 = (o_{11}, \dots, o_{1n})$, where o_{1i} is a randomly (uniformly) chosen number from the interval $[l_i^1, u_i^l]$, with $l_i^1 = \max\{a_i, x_i - I_i.\propto\}, u_i^l = \min\{b_i, x_i - I_i.\propto\}\ and\ I_i = |x_i - y_i|$.
    b) $O_2 = (o_{21}, \dots, o_{2n})$, where o_{1i} is a randomly (uniformly) chosen number from the interval $[l_i^2, u_i^2]$, with $l_i^2 = \max\{a_i, x_i - I_i.\propto\}, u_i^2 = \min\{b_i, x_i - I_i.\propto\}\ and\ I_i = |x_i - y_i|$.
02. In the CS part, the half-uniform crossover scheme (HUX) is employed [Eshelman and Schaffer.,1993]. The HUX crossover exactly interchanges the mid of the alleles that are different in the parents (the genes to be crossed are randomly selected from among those that are different in the parents). This operator ensures the maximum distance of the offspring to their parents (exploration).

In this case, four offspring are generated by the combination of the two from the part CT with the two from the part CS. The two best offspring obtained in this way are considered as the two corresponding descendents. Notice that since we consider a real coding scheme for the CT part, the incest prevention mechanism has to transform each gene considering a Gray code (binary code) with a fixed number of bits per gene (BITSGENE) that is determined by the expert to calculate the hamming distance between two individuals in order to apply the crossover operators.

IV DATASET

A dataset is a collection of data. This research work uses PIMA dataset and Z-Alizadeh Sani dataset for evaluating and comparing with existing algorithms.

4.1 PIMA dataset

This multivariate data set is used for diabetes detection, and is the result of a research survey carried out in the National Institute of Diabetes and Digestive and Kidney Diseases, United States on the female patients of Pima Indian heritage having age greater than 21. This dataset is commonly used among researchers who used machine learning method for diabetes disease classification, so it provides us to compare the performance of our method with that of others. The class distribution is: Class 1: normal (500), Class 2: Pima Indian diabetes (268) [Santi Wulan Purnami, et al.,2010]. The dataset contains 768 samples and two classes. All patients in this database are Pima-Indian women at least 21 years old and living near Phoenix, Arizona, USA. It has got 768 tuples and contains 9 numeric-valued attributes including the class. The class attribute has got two values, namely ‘‘tested negative for diabetes’’ and ‘‘tested positive for diabetes’’ and denoted by values ‘0’ and ‘1’ respectively. Many constraints were added for selecting the tuples from a large database [Soumadip Ghosh et al.,2014].

V Results and Discussions

The classification accuracy are the performance metrics used to evaluate the performance of this research work namely GSVM. The classification is performed on dataset namely PIMA Indian diabetes dataset. The classification accuracy, sensitivity and specificity can be calculated using the following metrics.

- ✓ True Positive
- ✓ True Negative
- ✓ False Positive
- ✓ False Negative

Table 1. Classification Accuracy Analysis

Algorithms	PIMA dataset
SVM	94.4
GF-ISVM	97.76

Table 1. depicts Classification Accuracy Analysis SVM and GSVM on PIMA Indian diabetes dataset. It can be clearly understood that the proposed work GSVM provides better results.

VI CONCLUSIONS AND FUTURE RESEARCH DIMENSIONS

The proposed research work presents genetic based support vector machine (GSVM) classifier. GSVM classifier considers the conventional simplex method as the modified simplex method. candidate rule pre-screening, rule selection and lateral tuning. The GSVM is evaluated using the performance metrics sensitivity, specificity and classification accuracy. Dataset was chosen for evaluating the performance of the proposed GSVM classifier with support vector machine algorithms. From the results it is evident that the proposed GSVM achieves better classification accuracy than rest of the algorithms.

REFERENCES

[1] Akhil jabbar, Priti Chandrab, Deekshatuluc, "Heart Disease Prediction System using Associative Classification and Genetic Algorithm", International Conference on Emerging Trends in Electrical, Electronics and Communication Technologies-ICECIT, 2012

[2] Amin, Agarwal, Beg, "Genetic neural network based data mining in prediction of heart disease using risk factors," Information & Communication Technologies (ICT), 2013 IEEE Conference on , vol., no., pp.1227,1231, 11-12 April 2013

[3] Amma, "Cardiovascular disease prediction system using genetic algorithm and neural network," Computing, Communication and Applications (ICCCA), 2012 International Conference on , vol., no., pp.1,5, 22-24 Feb. 2012

[4] Anbarasi, Anupriya, Iyengar, "Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm", International Journal of Engineering Science and Technology, Vol. 2(10), 2010, 5370-5376

[5] Anirudha, Kannan, Patil,"Genetic algorithm based wrapper feature selection on hybrid prediction model for analysis of high dimensional data," Industrial and Information Systems (ICIIS), 2014 9th International Conference on , vol., no., pp.1,6, 15-17 Dec. 2014

[6] Dalakleidi, Zarkogianni, Karamanos, Thanopoulou, Nikita, "A hybrid genetic algorithm for the selection of the critical features for risk prediction of cardiovascular complications in Type 2 Diabetes patients," Bioinformatics and Bioengineering (BIBE), 2013 IEEE 13th International Conference on , vol., no., pp.1,4, 10-13 Nov. 2013

[7] Dewan, Ankita, Sharma, Meghna, "Prediction of heart disease using a hybrid technique in data mining classification," Computing for Sustainable Global Development (INDIACom), 2015 2nd International Conference on , vol., no., pp.704,706, 11-13 March 2015

[8] Eshelman, Schaffer, “Real-coded genetic algorithms and interval schemata,” in Foundations of Genetic Algorithms, vol. 2, D. Whitley, Ed.San Mateo, CA: Morgan Kaufmann, 1993 pp. 187–202.

[9] Hanguang Xiao, "Diagnosis of Parkinson's disease using genetic algorithm and support vector machine with acoustic characteristics," Biomedical Engineering and Informatics (BMEI), 2012 5th International Conference on , vol., no., pp.1072,1076, 16-18 Oct. 2012

[10] Krissna Priya.R., “*A Improved Classification of Network Traffic using Adaptive Nearest cluster Based Classifier*”, International Journal of Computer Trends and Technology ISSN:2231-2803,Vol.18, Issue No.1 January 2015.