A Fuzzy Support Vector Machine model for Improving Prediction accuracy of Heart Disease

Purusothaman G[1], Dr. A. Nithya[2]
[1, 2] Department of Research & CS RVSCAS
[1, 2] Sulur, Coimbatore, India

Abstract- *Data mining is an important research area for medical domain. The Risk Prediction of Heart Disease is challenging task in the medical field. Genetic optimized neural network classification technique gives risk prediction of heart disease when data set is small. Since medical data is associated with imprecision, vagueness and uncertainty. In neural network, back propagation algorithm is very slow in convergence and there is possibility that network never converges. And the existing technique does not provide any improvement based on the size of data set. So there is a need for new intelligent risk prediction model for heart disease. In classification methods, Support Vector Machine (SVM), is a another technique to overcome over fit problem in training level. The technical problem is how computationally to treat such high-dimensional spaces[1]. Adding Fuzzy membership function with SVM may result in new hybrid model for predicting heart disease using risk factors. The hyper plane of FSVM can be skewed towards the minority class, and this skewness can degrade the performance of FSVM with respect to the minority class [2]. One remarkable property of SVMs is that their ability to learn can be independent of the dimensionality of the feature space [3] outperforms compared to any other data mining in dimensionality reduction on various data sets. Also it can be used as best feature selection method. The reduction of attributes in a model may be able to increase the accuracy in prediction. In Enhanced FSVM, fuzzy membership function is used as kernel function. Fuzzy Inference system consists of fuzzy rule and membership function. The Fuzzy based Enhanced SVM is a new hybrid model for heart disease risk prediction problem. We apply a fuzzy membership to each input point and reformulate the SVMs such that different input points can make different contributions to the learning of decision surface [4]. This model provides its best result compared with other Data mining algorithm in terms of accuracy and efficient. The new model is implemented using WEKA and MATLAB. The results are reported as in terms of histogram, line and confusion matrix.*

Keywords- data mining, SVM, Fuzzy model, FSVM, hybrid model

I. EXISTING AND PROPOSEDMETHODOLOGIES

The Existing system in data mining provides less accuracy in various levels like classification of data, data mapping and data extraction takes more time to execute i.e. algorithms speed and stores unwanted information in memory to provide solution like noisy data, sharp boundary data. In human health care domain, there are huge amount of data available to classify to get hidden information. In particular human's heart related data can perform vital role in the medical diagnosis and earlier detection of disease.

The Proposed method overcomes all the existing problems with the help of Support Vector Machine and fuzzy approach. In particular, multi objective approach handles multi objective or multi disciplinary problems without extra time. The existing problems can be solved in object oriented approach.

II. PRE-PROCESSING WORKS

A. Baseline correction

We can remove baselines (or "backgrounds") from data by either by including a baseline function when fitting a sum of functions to the data, or by actually subtracting a baseline estimate from the data.

B. Smoothing

The presence of this noise influences both data mining algorithms and human observers in finding meaningful patterns in mass spectra. The heuristic high frequency noise reduction approaches employed most commonly in studies to date are smoothing filters ,the wavelet transform (WT), or the deconvolution filter . Here we employ a locally weighted linear regression method with a span of 10 M/Z to smooth the spectra.

C. Normalization

We normalized a group of mass spectra by standardizing the area under the curve (AUC) to the group median.

III. FEATURE EXTRACTION AND SELECTION

To avoid regional correlation between the selected features, we used regional information to outweigh the value of potential features using following factor:

W= (1-Exp(-(Dist/2)2)) (1)

Where Dist is the distance between the candidate feature and previously selected features. A small Dist (close to 0) outweighs the significance statistics of only close features. This means that features that are close to already picked features are less likely to be included in the output list. Combining this weighting factor with the ―T test‖ feature selection algorithm over the training set.

Consequently, we utilized fuzzy logic as the powerful tool of soft computation using linguistic variables and rules. To design two fuzzy sets, high and low, as the linguistic values of the features .We utilized Gaussian membership functions and adjusted their parameters with respect to histogram analysis of the feature value distribution over the training samples.

IV. PROCESS AND EVALUATION

Data mining is the search for relationships and global patterns that exist in large databases but are ‗hidden‘ among the vast amount of data, such as a relationship between patient data and their medical diagnosis. These relationships represent valuable knowledge about the database and the objects in the database and, if the database is a faithful mirror, of the real world registered by the database. Machine learning and classification techniques have been widely used to assist the interpretation and analysis of biomedical data. Since medical data is associated with imprecision, vagueness and sometimes uncertainty the Fuzzy Logical Data Classification could be more realistic.

If our data is time series then it is not important what classifier we should use. It is more important how we understand the data and extract good and rich feature set from the data.

In this proposed method (figure-1) there are some challenges to deal with in the medical field with regards to discrimination of symptoms diseases, etc. These are:

1) Is the data imbalanced? For example, how many people in your data set [5] have heart disease vs. those who do not? One possible statement that could be made is that diseases are outliers of the general population.
2) Has there been a causal relationship established or is it just correlation between variables on which your judgment relies about the performance of your baseline.
3) The characteristics of the data that have been gathered, are they enough to discriminate the disease?
4) Is the available data tagged or not. This will dictate whether supervised or unsupervised data is used. SVMs work by implicitly (using the kernel trick) mappingall training data from input space into a (usually) higher dimensional feature space [6].

These will determine the type of technique utilized as well as the validation level required to establish confidence in the procedure carried out. In the medical field, class imbalance is often an issue, so it is important not only to look at the classification methods themselves, but also sampling methods (under-sampling, over-sampling, etc.). Additionally, one needs to be careful when evaluating and interpreting the results, as accuracy is no longer suitable, F1 score and the Matthews correlation coefficient need to be used instead.

In the previous research, the heart disease prediction was implemented with some drawbacks using some hybrid techniques like neural network and genetic algorithm. Such as

1. ANNs often converge on local minima rather than global minima, meaning that they are essentially "missing the big picture" sometimes (or missing the forest for the trees)

2. ANNs often over fit if training goes on too long, meaning that for any given pattern, an ANN might start to consider the noise as part of the pattern.

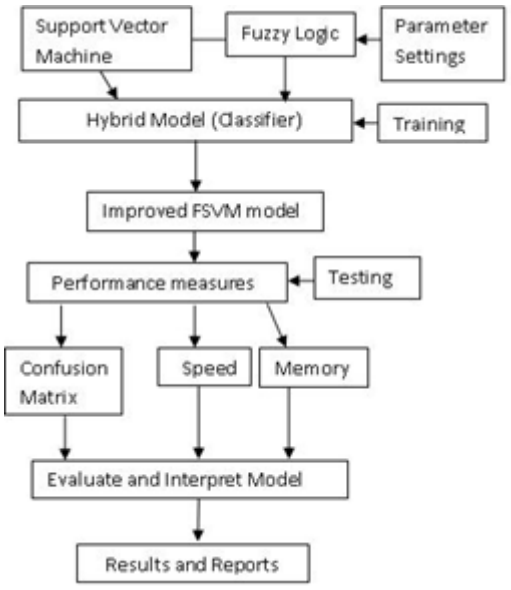

Figure 1. Proposed model

Sometimes, improved data quality is itself the goal of the analysis, usually to improve processes in the designing of medical decision support. Real life medical data often not is good quality. So we would suggest looking in to the data pre processing prior to select good classification model. In medical research, Researchers thus use information from a sample of individuals to make some inference about the wider population of like individuals.

Medical data by and large involves the linguistic components besides numeric. Information given in linguistic form is always associated with imprecision, vagueness and uncertainty. All these inherent properties of medical data can be easily handled by Fuzzy Sets and decision mechanism can be implemented by Fuzzy Reasoning. So in our opinion Fuzzy Logic Technique with Support Vector Machine method is going to be good techniques what we are looking for. Because Generally, the SVM classifier with optimizing feature yielded slightly higher prediction accuracy than ANN. Using the decision functions obtained by training the SVM, for each class, we define a truncated polyhedral pyramidal membership function [7]. Fuzzy support vector machine as a new classification model with high generalization power, robustness, and good interpretability seems to be a promising tool for gene expression microarray classification [8].

V. CONCLUSION AND FEATURE WORK

A new fuzzy membership function is employed in the linear and nonlinear fuzzy support vector machine respectively [9] Although SVM outperformed the ANN classifiers with regard to overall prediction accuracy, both methods were shown to complement each other, as the sets of true positives, false positives (over prediction), true negatives, and false negatives (under prediction) produced by the two classifiers were not identical.

Since boundaries between the classes of medical data are no sharp the partition of data over different classes using Fuzzy Sets is going to be more realistic and close to the human nature of elasticity in classifying the data on intuition based decisions. Optimize the parameters of the SVM(AMPSVM) model in order to improve the learning performance and generalization ability of the SVM model [10].FSVMs (Fuzzy SVM) work well when the average training error is high, which means it, can improve performance of SVMs for noisy data.

REFERENCES

[1] C. Cortes and V. Vapnik, ―Support vector networks,‖ Machine Learning, vol. 20, no. 3, pp. 273–297, 1995.

[2] Xiaoqing Gu, Tongguang Ni, and Hongyuan Wang, ―New Fuzzy Support Vector Machine for the Class Imbalance Problem in Medical DatasetsClassification,‖ Hindawi Publishing Corporation The Scientific World Journal Volume 2014, Article ID 536434, 12 pages.

[3] T. Joachims, ―Text categorization with supportvector machines:learning with many relevantfeatures,‖ in Proceedings of ECML-98, 10th EuropeanConference on Machine Learning, C. Nedellec and C. Rouveirol, Eds., no. 1398. Chemnitz, DE: Springer Verlag, Heidelberg, DE, 1998, pp. 137–142. [Online].Available: citeseer.ist.psu.edu/article/joachims98text.html

[4] Lin CF1, Wang SD, ―Fuzzy support vector machines,‖ IEEE Trans Neural Netw. 2002;13(2):464-71 doi: 10.1109/72.991432.

[5] C. Blake and C. Merz, ―UCI repository of machine learning databases (http://www.ics.uci.edu/mlearn/MLRepository.html),‖UniversityofCalifornia, Department of Information and Computer Science., 1998.

[6] Alistair Shilton and Daniel T. H. Lai, ―IterativeFuzzy Support Vector Machine Classification", IEEE.

[7] T. Inoue Graduate Sch. of Sci. & Technol., Kobe Univ., Japan S. Abe, ―Fuzzy support vector machines for pattern classification,‖ Neural Networks,2001. Proceedings. IJCNN '01. International Joint Conference on Date of Conference: 15-19 July 2001

[8] Mohsen Hajiloo, Hamid R Rabiee and Mahdi Anooshahpour,"Fuzzy support vector machine: an efficient rule-based classification technique for microarrays", BMC Bioinformatics. 2013; 14(Suppl 13): S4.Published online 2013 Oct 1. doi: 10.1186/1471-2105-14-S13-S4

[9] Wan Mei Tang, "Fuzzy SVM with a New Fuzzy Membership Function to Solve the Two-Class Problems", Journal Neural Processing Letters archive Volume 34 Issue 3, December 2011Pages 209-219.

[10] Senhua Wang and and Rui Li, "A Novel Classification Method based on Improved SVM and its Application", International Journal of Database Theory and Application Vol.8, No.4 (2015), pp.281-290.