IOSR Journal of Electronics and Communication Engineering (IOSR-JECE)
e-ISSN: 2278-2834,p- ISSN: 2278-8735.Volume 12, Issue 1, Ver. I (Jan.-Feb. 2017), PP 01-07
www.iosrjournals.org

An Approach for Disease Data Classification Using Fuzzy Support Vector Machine

[1]Er. Jitender, [2]Er. Neeraj Julka

[1]Research Scholar Deptt. Of Electronics and comm.ASRA College of Engineering & TechnologyBhawanigarh (India)

[2]Assistant prof. Deptt. Of Electronics and comm.ASRA College of Engineering & TechnologyBhawanigarh (India)

Abstract: *Data Mining has great scope in the field of medicine. In this article we introduced one new fuzzy approach for prediction of hepatitis disease. Many researchers have proposed the use of K-nearest neighbor (KNN) for diabetes disease prediction. Some have proposed a different approach by using K-means clustering for reprocessing and then using KNN for classification. In our approach Naive Bayes classifier is used to clean the data. Finally, the classification is done using Fuzzy SVM algorithm. Hepatitis diseases data set is used to test our method. We are able to obtain model more precise than any others available in the literature. The Fuzzy SVM approach produced better result than KNN with Fuzzy c-meansand Fuzzy KNN with Fuzzy c-means. Theintroduction of Fuzzy Support Vector Machine algorithm certainly has a positive effect on the outcome of hepatitis disease. This fuzzy SVM model led to remarkable increase in classification accuracy.*

Keywords: *Hepatitis disease, Diabetes disease, Liver disorder, KNN algorithm,fuzzy c-means algorithm, fuzzy KNN algorithm, fuzzy SVM algorithm.*

I. Introduction

Data mining has been used for a longer time for the study of diseases. For example, in case of heart disease, classification techniques such as KNN, decision tree has been widely used in diagnosis. Blood pressure, cholesterol, pulse rate are the major reason for the heart disease. Some non-modifiable factors are also there such as smoking, drinking also reason for heart disease. The heart is an operating system of our human body. If the function of heart is not done properly means, it will affect other human body part also. Some risk factors of heart disease are Family history, High blood pressure, Cholesterol, Age, Poor diet, Smoking. When blood vessels are overstretched, the risk level of the blood vessels is increased. This leads to the blood pressure. Blood pressure is typically measured in terms of systolic and diastolic. Systolic indicates the pressure in the arteries when the heart muscle contracts and diastolic indicates the pressure in the arteries when the heart muscle is in resting state. Various data mining techniques such as Naïve Bayes, KNN algorithm, Decision tree, Neural Network are used to predict the risk of heart disease. The KNN algorithm uses the K user defined value to find the values of the factors of heart disease. Decision tree algorithm is used to provide the classified report for the heart disease. The Naïve Bayes method is used to predict the heart disease through probability. The Neural Network provides the minimized error of the prediction of heart disease. In all this above mentioned techniques the patient records are classified and predicted continuously. The patient activity is monitored continuously, if there is any changes occur, and then the risk level of disease is informed to the patient and doctor. The doctors are able to predict heart diseases at an earlier stage because of machine learning algorithms and with the help of computer technology.

Many nearest neighbor search algorithms have beenproposed over the years; these generally seek to reduce the number of distance evaluation actually performed. Using anappropriate nearest neighbor search algorithm makes KNNcomputationally tractable even for large data sets. Acombination of k-means and KNN algorithms provide animproved accuracy. In our approach Naive Bayes classifier is used to clean the data. Finally, the classification is done using Fuzzy SVM algorithm. It was discussed when Naive Bayes classificationalgorithm is combined withFuzzy SVM algorithms it may provide a betteraccuracy.

Liver and thyroid are among the most important organs of human body which have a high influence on the performance of other body parts. Liver diseases which include various kinds of hepatitis disease are seriously dangerous and fatal. On the other hand, thyroid gland is one of the most important glands in the body. Since thyroid hormones are responsible to control the body metabolism, the performance of the thyroid gland directly influences each of the main body organs. Consequently curing these two diseases is very important. In order to diagnose the thyroid disease 15 experiments are done for each patient, by which we could determine 5diagnosable cases. The dataset gathered consist of 221 cases all examined carefully by specialists.

Hepatitis disease is a fatal and deadly disease and is thought to be the fifth deadly diseaseworldwide. Hepatitisdisease is the inflammation and damage to hepatocytes in the liver and canbe caused by infections with viruses, bacteria, fungi, exposure to toxins, alcohol consumptionand autoimmunity. Clinical symptoms of hepatitis are nausea, fever, general weakness, and jaundice.Five viruses have been identified and named hepatitis A through E.

The data set we considered is the hepatitis data set. This disease has serious consequences and it can lead to death.Many countries have thousands of people who suffer fromliver diseases. This disease effects developed as well asdeveloping countries.

The rest of the paper is organized as follows. Section 2 explains the proposed method in detail. Section 3 explains the Methodology. In Section 4, theresults obtained with the above mentioned dataset is analyzedand compared with the existing methods. Section 5 concludesthe paper with future scope.

II. Proposed Method

Our proposed implementation involves optimization of the predictor function via parallel processingusing the ARM/GPU hardware. The prediction algorithm involves multiplication of akernel function, obtained during the training algorithm, with the sub-band signals and computingthe class they belong to. Kernel multiplication with the sub-band signals is computationally verydemanding as the length of features combined is equal to 4096. The same challenge holds truefor regression analysis (SVR). GPU hardware has been chosen to do this task. There is lot ofopportunities to parallelize both the sub-band decomposition and SVM classification operations,as well as the training algorithm on the GPU since TegraSoC consists of 192 CUDA (a parallelcomputing platform and programming model) cores. Our proposed method has been designed based onthe algorithms Naive Bayesclassifier, Fuzzy Support Vector Machine. These algorithms are briefedhere.

2.1 Support Vector Machine (SVM)

In certain applications, an expert or system designer has knowledge of conditions under whichconfusion between certain classes is inconsequential. An example is illustrated in Figure 1 for amaterial processing application. In this example, a classifier is desired to distinguish betweenmaterial A (circles) and material B (triangles) given a data point that is a tuple consisting of atemperature and another measurement. However, for some part of our 2-dimensional space (thespace where the temperature T is less than some critical temperature Tcrit) the distinction is irrelevant.

However, if we use the data below Tcrit the algorithm will still evaluate the errors thereand try to minimize them, even though they do not matter to the end result. This could lessenperformance in regions where errors do matter. One strategy for dealing with this issue could beto eliminate all the data where T <Tcrit and certainly this would work for this particular example.However, this does not provide a general methodology to tackle problems of this type, wherecertain subsets of the nf dimensional space could be arbitrarily defined in which the error matters to varying degrees. In addition, this example and its trivial solution would not address the addedcomplexity of a multiclass arrangement.

As the dimensionality of the system increases, the complexity increases as well, and datasetsneed not be well behaved as that in the example. This is evident in a wide variety of mechatronicproducts and processes. Large scale roll to roll printers are one application area where classificationof web media type is important in order to select the correct operating parameters. However,the operating parameters that are selected are also dependent on environmental variables such astemperature and humidity, though in some ranges the difference between paper types changes.Because much of this knowledge is not coded into simple functions and rather expert-defined

look-up tables, it cannot be directly added to training cost functions for classification algorithms.The possibility for modification of the weighting of errors during training is common in the literature.Weighting by training example is also used to reduce sensitivity to outliers or artifactsusing metrics that measure how likely a data point is to be an outlier. Other uses of weightinginclude compensating for unbalanced (in terms of number of examples) data sets and attempting to reduce false alarm rates. However, the work presented here deals with a largely unrelated problem from that of the prior art.

In binary classification, Support Vector Machine (SVM)is a family of related algorithms that attempt to produce a hyperplane that best separates two sets of data. The implementation and usage of Support Vector Machines for classification begins in a similar manner to most supervised learning machine learning problems. A training data set, comprised of individual training examples, is used to create the classifier. A training data is denoted by x ∈ Rnf , where nf is the number of features. The features are the quantities used in order to make a decision. In anembedded system, for example, the features might be sensor data, or quantities derived from sensor data. The vector xk represents the kth training data, and yk ∈ {1,−1} is the correct classification in thebinary scheme corresponding to the training data xk. N is the total number of training data examples. Support vector machine (Support Vector Machine, SVM) is based on statistical learning theory machine learning method, which solves the nonlinear, high dimension and local minima problems. And it is have incomparable advantages for the characteristics of relevance and sparsity, and high dimensionality problem as other machine learning methods.

SVM is proposed for binary classification problem, and how to effectively promote the multiple classification is developing.

SVM multiple classification used so much variables in the process of solving the optimization problem in solving method, so the computational complexity is too high and not practical. The 1-a-r (One-against-rest) and 1-a-1 (One-against-one) method is currently the most commonly used two SVM multi-class classification method. For Nclass classification, the 1-a-r and 1-a-1 need to construct the N and N*(N-1)/2 classifiers, when a large number of categories, the speed of these two methods are low, but it has a unrecognized domain. When the number of classes is more, the training rate remains of the DAG (Directed Acyclic Graph) method is low, although with a higher speed to be classification. Classification tree as a multi-class classification method is widely used in pattern classification, especially for large classification problems, it is possible to improve the classification efficiency greater extent. For Nclass classification, The binary tree SVM multi-class classification only need to construct the N-1 SVM classifiers, and the method does not exist unrecognized domains. It does not need to traverse all The SVM classifier, and the better level of hierarchy SVM multi-class classification can significantly increase speed for the training and classification. However, how to design effective the hierarchy SVM multi-class classification is still a problem to be solved.

Support Vector Machine (SVM) paradigm in pattern recognition presents a lot of advantages over other approaches some of which are: 1) the assurance that once a solution has been reached, it is the unique (global) solution, 2) good generalization properties of the solution, 3) reduced number of tuning parameters and, last but not least, 4) clear geometric intuition on the classification procedure.

The contribution of this work consists of the following: 1) It exploits the intrinsic geometric intuition to the full extend, i.e., not only theoretically but also practically (leading to a novel algorithmic solution), in the context of classification through the SVM approach, 2) it provides, for the first time, the theoretical background for a geometric solution of the non-separable (both linear and non-linear) classification problems with linear (1st degree) penalty factors, by means of the reduction of the size of the convex hulls of the training patterns, 3) it provides an easy way to relate each class with a different penalty factor, i.e., to relate each class with a different risk (weight), 4) it develops, for the first time, an efficient algorithm for the computation of the minimum distance between the RCHs and finally 5) it opens the road for applying other geometric algorithms, finding the closest pair of points between convex sets in Hilbert spaces, for the non-separable SVM problem.

2.1.2 Types of Support Vector Machines (SVM)

Simply stated, a SVM finds the best separating (maximal margin) hyperplane between the two classes of training samples in the feature space, which leads to maximal generalization. The patterns in the original, low dimensional space X , are mapped (Φ:X $\rightarrow$H) in a high-dimensional feature space H , which is a Reproducing Kernel Hilbert Space (RKHS). It is not necessary to know the map itself analytically, but only its kernel, i.e., the value of the inner products of the mappings of all the samples

(k(x1,x2)= Φ(x1),Φ(x2) for all pairs of samples x1 , x2 $\in$ X). Through this “kernel trick”, it is possible to transform a nonlinear classification problem to a linear one, but in a higher (maybe infinite) dimensional space.

Although some authors have presented the theoretical background of the geometric properties of SVMs, exposed thoroughly, the main stream of solving methods comes from the algebraic field (mainly decomposition). One of the best representative algebraic algorithms with respect to speed and ease of implementation, also presenting very good scalability properties, is the Sequential Minimal Optimization (SMO). The geometric properties of learning and specifically of SVMs in the feature space, have been pointed out early enough, through the dual representation (i.e., the convexity of each class and finding the respective support hyperplanes that exhibit the maximal margin) for the separable case and also for the non-separable case through the notion of the Reduced Convex Hull (RCH). Actually, the geometric algorithms presented until now are suitable only for solving directly the separable case and indirectly the non-separable case through the trick proposed in. However, the latter (artificially extending the dimension of the input space by the number of training patterns) is equivalent to a quadratic penalty factor and, besides the increase of complexity, due to the artificial expansion of the dimension of the input space, it has been reported that the generalization properties of the resulting SVMs can be poor. Authors of numerous papers use different ANN structures to recognize the hepatitis disease. Reference analyses a large database with hepatitis C virus infected patients. There are made a lot of statistical analyses on the records of this database in order to determine the evolution of biological parameters during the treatment. The results of the statistical analyses and the expert system predictions indicate the use of such a system to facilitate the physician work. Some authors use a feature selection (FS) and artificial immune recognition system (AIRS) with fuzzy resource allocation mechanism diagnosis of hepatitis disease with total accuracy of 92.59% on data set of UCI. ANN has been used to diagnosis chronic hepatitis disease. They report total accuracy of 93%.

In fact, In addition to recognizing the hepatitis cases, it is important to identify the phase and the type of the hepatitis the person caused by, which is the main proposes of this paper. There are fifteen parameters measured for each patient in order to diagnose hepatitis, as follow:

1-sex, 2-age, 3-ALK, 4-AST, SGOT, 5-ALT, SGPT, 6Bi, T, 7-Bi, D, 8-G.G.T, 9-HBSAg, 10-Alb, 11-LHD, 12-PT, 13-FBS, 14-CHO, and 15-HCVAb.

These experimental data show six cases: 1-non hepatitis person, 2-person who carries hepatitis B (no symptoms), 3-person affected by hepatitis B, 4-person who carries hepatitis C (no symptoms), 5-person affected by hepatitis C and 6-non viral hepatitis.

Traditional Chinese Medicine (TCM) has been widely used to treat various diseases, such as hepatitis, cancer, diabetes mellitus and even some intractable diseases. Its effectiveness has been validated in clinical practice. In the past decades, there are tremendous amount of clinical data in TCM field collected from clinical practice for medical studies. Generally these clinical data are about symptoms, diagnoses and treatments as the main information elements and they are in large scale and multidimensional. Many clinical researchers intend to explore the complicated relationships among these information elements, e.g. the relationships between herbs and the symptoms of hepatitis, but some frequently-used tools like Microsoft Excel are not suited for multidimensional analysis.

Data warehouse which can be considered as a database for reporting and analyzing in computing is the basic platform for data mining, On-line Analytical Processing (OLAP) and decision support. Clinical data warehousing is a difficult task due to some complicated issues, such as complicated semantics, many-to-many relationships and bi-temporal data. Our work is based on TCM clinical data warehouse which takes the structured electronic medical record data as the core data source. Currently, the TCM clinical data warehouse has incrementally loaded vast amounts of formatted clinical data related to hepatitis diseases.

As an essential technology for multidimensional analysis, OLAP is one of the main technologies for decision support. OLAP can quickly answer multidimensional analysis queries in computing and it has been widely used in different fields. For example, OLAP is one component of business intelligence techniques which has some typical applications including business reporting for sales, financial reporting and so on. In 2011, the market leading vendors for OLAP systems include SAP Business Objects, Oracle, IBM Cognos, MicroStrategy, Microsoft, SAS, Pentaho and Jaspersoft. Business Objects (BO) is one of the leading business intelligence software companies which help the enterprises tracking and having a better viewing of the business, improving the decision-making level and optimizing enterprise benefit. The products of BO are the main tools to develop OLAP reports. Besides mining the potential information of clinical data, BO can intuitively display the mining results. We believe that BO product packet can also be well used to make OLAP on hepatitis diseases.

One of the most critical features in medicine is the diagnosis of a disease. Diagnosis is needed as the analysis of the physiological or biochemical cause of a disease. It is a complicated task and involves certain level of expertise on the part of a doctor. A sophisticated system is needed to assist doctors for diagnosing a disease accurately and efficiently. The use of technology, especially Artificial Intelligence (AI), can minimize cost, time, human expertise and incorrect diagnosis. Artificial Neural Networks (ANN) is a type of AI which has extensively been applied to solve medical problems. They have been used in multiple applications like diagnosis, forecasting, image analysis etc.

The Ns are systems made up of neurons which work in a similar way as the brain. Hepatitis is considered as one of the most deadly diseases. Early detection increases the chances of recovery manifold. Hepatitis causes inflammation and destruction of hepatocytes (liver cells). Hepatitis can be caused due to viruses, bacteria, drugs, etc. This disease can be categorized as Acute or C onic. Acute hepatitis is the rapid, sharp, and painful onset of the disease. Acute symptoms are more painful for patients but it has a limited course and rarely lasts beyond 1 or 2 months. Usually, there is only minimal liver cell damage and little evidence of immune system activity. Chronic hepatitis is inflammation of the liver that persists more than six months. Chronic inflammatory cell in latex comprising lymphocytes, plasma cells and sometimes lymphoid follicles are usually present in the portal tracts. There are five different types of hepatitis viruses A, B, C, D and E. Hepatitis A and E are of acute type whereas Hepatitis B, C and D are of chronic type. The chronic hepatitis leads to cirrhosis which causes distinction of liver parenchymal cells.

III. Methodology

Our work is divided into two parts:

Part 1:

1. Import data from the database.
2. Find out the class label.
3. Find out the tuples with presence of 0, if yes, and then remove the tuple.
4. Remaining data need to be separated base on the class label in the data.
5. Set up Naive Bayes classifier and find out the misclassified data, mark them.
6. Calculate the classification accuracy.
7. The misclassified tuples were removed.

Part 2:

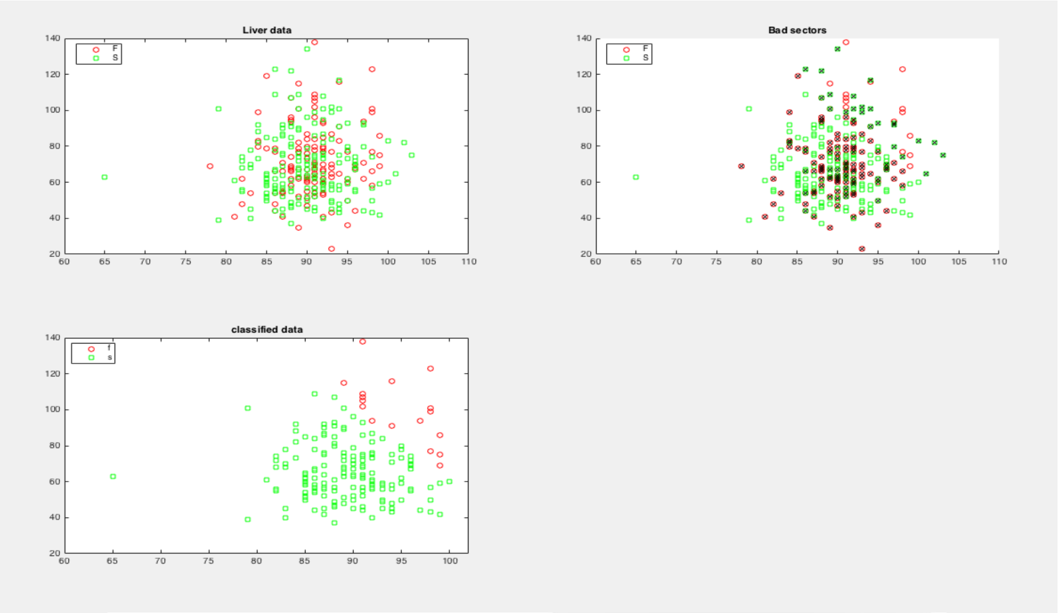
1. Remaining data is fed to our new Fuzzy SVMapproach.
2. Classification accuracy is calculated and compared with the previous method.

Datasets: The performance of the approach mentioned in thispaper has been tested with medical dataset downloaded fromUCI machine learning data repository for hepatitis dataset.
https://archive.ics.uci.edu/ml/machine-learning-databases/hepatitis/hepatitis.data

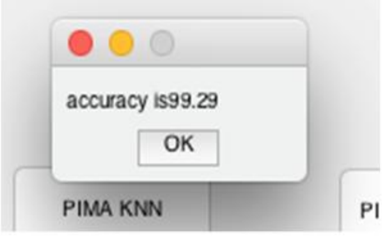
A. About software

We need MATLAB 2015b or latest version of MATLAB with commercial license. MATLAB is a numerical computing environment which widely used in the field of data science and other research fields by researchers and engineers.

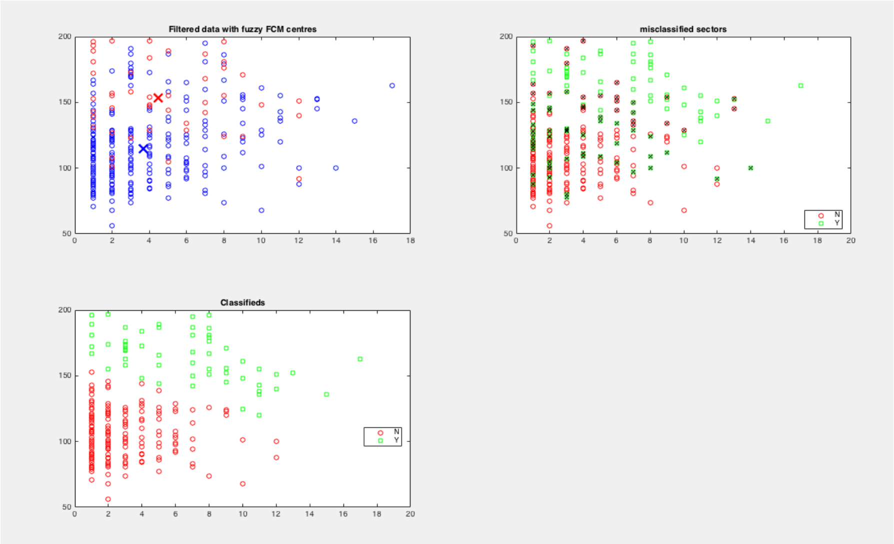
Implementation of the PIMA dataset



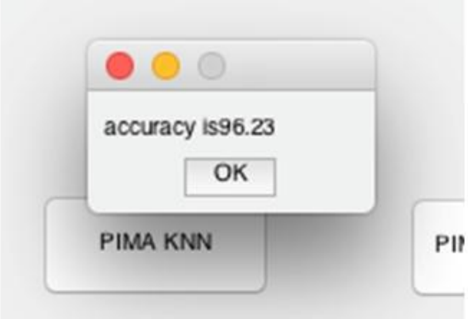
Accuracy of the Classification


accuracy is99.29

Implementation of FCM dataset

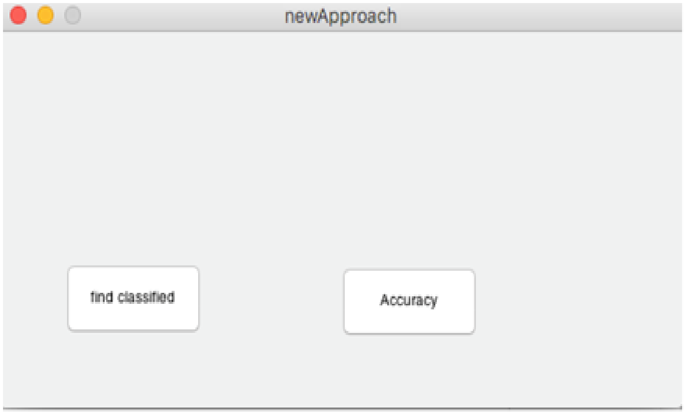


Accuracy of the Classification

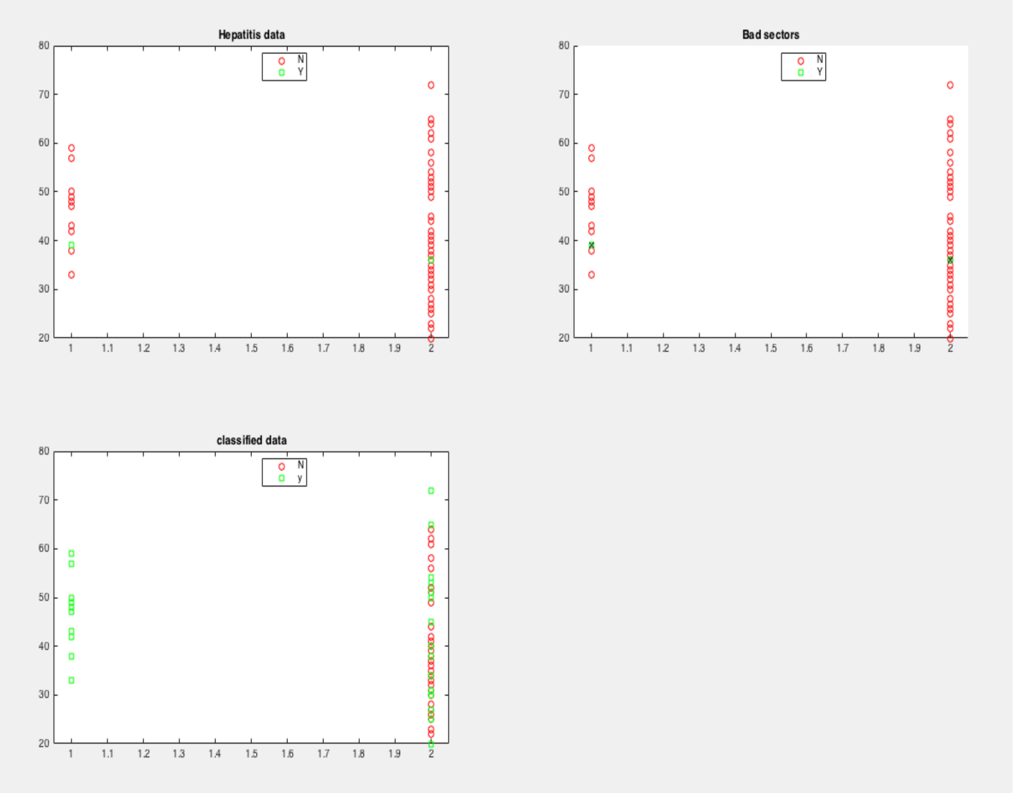


accuracy is96.23

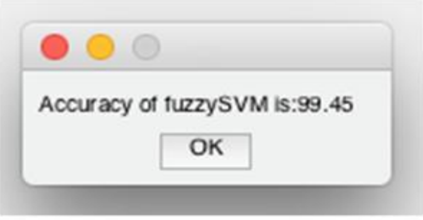
GUI of proposed method



Classified Hepatitis disease data



Accuracy of the Fuzzy SVM Classification



Accuracy of fuzzySVM is:99.45

IV. Analysis And Results

In this research paper we will discuss about accuracy. Table shows the result generated by the Fuzzy SVM approach and compares it with previously used KNN with Fuzzy c-means and Fuzzy KNN with fuzzy c-means methods.

Comparison table

	Tuples before pre-processing	**Tuples after pre-processing**	**Model**	**Classification accuracy**
PIMA	336	262	KNN with Fuzzy c-means	97.04
			Fuzzy KNN with fuzzy c-means	99.29
Liver-disorder	336	189	KNN with Fuzzy c-means	96.23
			Fuzzy KNN with Fuzzy c-means	98.8
Hepatitis	77	75	Fuzzy-SVM	99.45

V. Conclusion

In this research, we studied classification of disease datasets of diabetes and liver disorder. These two datasets were classified using fuzzy c means and fuzzy KNN methods. We proposed a new method named fuzzy SVM. We are successful in developing accurate models which showed better results in classifying a dataset based on Hepatitis dataset. Our results clearly indicate that the proposed methods work better and are more accurate than other existing methods with equal effort. In future, same dataset can be used for classification using different classification algorithm.

References

[1]. A Critical Study of Classification Algorithms Using Diabetes Diagnosis by PanigrahiSrikanth,2016 IEEE 6th International Conference on Advanced Computing (IACC).

[2]. An Efficient Rule Saba Bashir, UsmanQamar, Farhan Hassan Khan, M.YounusJaved, F2014 12th International Conference on rontiers of Information Technology (FIT),-based Classification of Diabetes Using ID3, C4.5 & CART Ensembles.

[3]. A Particle Swarm Optimization Based Classifier for Liver Disorders Classification Jyun Jie Lin and Pei-Chann Chang, 2010 International Conference on Computational Problem-Solving (ICCP).

[4]. A study on feature extraction and disease stage classification for Glioma pathology images by Kiichi Fukuma, V. B. Surya Prasath,HiroharuKawanaka, Bruce J. Aronow, Haruhiko Takase, 2015 17th International Conference on E-health Networking, Application & Services (HealthCom).

[5]. Citrus Gummosis disease severity classification using participatory sensing, remote sensing and weather data by Jayantrao Mohite, Bhushan Jagyasi, Sonali Kulkarni, Srinivasu Pappula, 2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS).

[6]. Classification of the Liver Disorders Data Using Multi-Layer Adaptive Neuro Fuzzy Infeence System by ParisaTavakkoliDavood M. Souran, SaeedTavakkoli, Majid, 2015 6th International Conference on Computing, Communication and Networking Technologies (ICCCNT).

[7]. Classification of renal diseases using first order and higher order statistics Komal Sharma, JitendraVirmani, 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom).

[8]. Comparison of a new ad-hoc classification method with Support Vector Machine and ensemble classifiers for the diagnosis of Meniere's disease using EVestG signals by Z. A. Dastgheib, O. RanjbarPouya, B. Lithgow, Z. Moussavi, 2016 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE).

[9]. Classification and prediction of heart disease risk using data mining techniques of Support Vector Machine and Artificial Neural Network by S. Radhimeenakshi, 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom).

[10]. Chronic Kidney Disease analysis using data mining classification techniques by VeenitaKunwar, KhushbooChandel, A. SaiSabitha, AbhayBansal 2016.

[11]. Disease recognition and classification from movement patterns by Garima Bhatia, Sangeeta Rani, 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom).

[12]. ECG signal analysis using wavelet coherence and s-transform for classification of cardiovascular diseases SakshamAgarwal, VigneshramKrishnamoorthy, SawonPratiher, 2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI).

[13]. Hatamian, Armin MehrabianValentina E. Balas 6th ICCCNT 2015 July 13 15.

[14]. High-accuracy voice-based classification between patients with Parkinson's disease and other neurological diseases may be an easy task with inappropriate experimental design b Jan Rusz, Michal Novotny, Jan Hlavnicka, TerezaTykalova, EvzenRuzicka, IEEE Transactions on Neural Systems and Rehabilitation Engineering year: 2016, Volume: PP, Issue: 99.

[15]. Human Heart Disease Prediction System using Data and mining techniques by Theresa Princy. R and J. Thomas, 2016 International Conference on Circuit, Power and Computing Technologies [ICCPCT].

[16]. IoT based classification of vital signs data for chronic disease monitoring by A. Raji, P. Golda Jeyasheeli, T. Jenitha, 2016 10th International Conference on Intelligent Systems and Control (ISCO).

[17]. Liver Disorder Detection Based on Artificial Immune Systems by Shane Dixon, 2015 11th International Conference on Natural Computation (ICNC).

[18]. NonInvasive Diabetes Detection and Classication Using Breath Analysis Lekha .S and Suchetha. M, 2015 International Conference on Communications and Signal Processing (ICCSP).

[19]. Ontology-based Fuzzy Inference Agent for Diabetes Classification by Mei-Hui Wang and Chang-Shing Lee, 2007 Annual Meeting of the North American Fuzzy Information Processing Society.