Effectiveness of Support Vector Machines in Medical Data mining

Padmavathi Janardhanan, Heena L., and Fathima Sabika

***Abstract*— The idea of medical data mining is to extract hidden knowledge in medical field using data mining techniques. One of the positive aspects is to discover the important patterns. It is possible to identify patterns even if we do not have fully understood the casual mechanisms behind those patterns. In this case, data mining prepares the ability of research and discovery that may not have been evident. This paper analyzes the effectiveness of SVM, the most popular classification techniques in classifying medical datasets. This paper analyses the performance of the Naïve Bayes classifier, RBF network and SVM Classifier. The performance of predictive model is analysed with different medical datasets in predicting diseases is recorded and compared. The datasets were of binary class and each dataset had different number of attributes. The datasets include heart datasets, cancer and diabetes datasets. It is observed that SVM classifier produces better percentage of accuracy in classification. The work has been implemented in WEKA environment and obtained results show that SVM is the most robust and effective classifier for medical data sets.**

***Index Terms*— Medical data mining, Navie Bayes, RBF, Support Vector Machines**

I. INTRODUCTION

WE are living in data-rich times and each day, more data are collected and stored in databases. The increased use of data toward answering and understating important questions has driven research towards the development of data mining techniques. The purpose of these techniques is to find information within the large collection of data. Although data mining is a new field of study of medical informatics, the application of analytical techniques to discover patterns has a rich history. Perhaps it was one of the most successful uses of data analysis for discovering and understanding of the medical science, especially infectious diseases. (E. Donald, 2009) [1]. Medical diagnosis is known to be subjective and depends not only on the available data but also on the experience of the physician and even on the psycho-physiological condition of the physician. A number of studies have shown that the diagnosis of one patient may differ significantly if the patient is examined by different physicians or even by the same physician at various times (R. Zhang & Y. Katta,2002) [2].

Manuscript received February 6, 2015; revised April 12, 2015.

Authors are with the M.O.P. Vaishnav College for women, Chennai, India (e-mail: {padmalaya90, slashe, ridafathima17}@gmail.com)

The idea of medical data mining is to extract hidden knowledge in medical field using data mining techniques. One of the positive aspects is to discover the important patterns. It is possible to identify patterns even if we have not fully understood the casual mechanisms behind those patterns. In this case, data mining prepares the ability of research and discovery that may not have been evident. R. Zhang & Y. Katta [2], have highlighted that irrelevant pattern can also be discovered.

Clinical repositories containing large amounts of biological, clinical, and administrative data that are increasingly becoming available as health care systems integrate patient information for research and utilization objectives (Harleen Kaur and Siri Krishan Wasan, 2006) [3]. Data mining techniques applied on these databases discover relationships and patterns which are helpful in studying the progression and the management of disease (J.C.Prather et al., 1997) [4]. A typical clinic data mining research includes the following cycle: Observations, structured data, narrative text, hypotheses, tabulated data statistics, analysis interpretation, new knowledge more questions, outcomes. Prediction or early diagnosis of a disease can be a kind of evaluation. About diseases like skin cancer, breast cancer or lung cancer early detection is vital because it can help in saving a patient‘s life (Y. Mahajani & G. Aslandogan , 2004) [5] . In order to predict unknown or future values of interest, prediction can be used, whereas description focuses on finding patterns describing the data and the subsequent presentation. For instance, one may classify diseases as per the symptoms provided, which describe each class or subclass.

A. Overview

This work is to study the working of Support Vector Machines in data classification. The main idea is to experiment the predictive model’s working and analyze its performance in terms of accuracy, specificity and sensitivity. The datasets for test and train were grouped randomly such that they formed uniformly distributed datasets and skewed datasets with both positive and negative class attributes. These datasets were tried on different classifiers and outcomes recorded for comparison. It was observed that the performance of SVM better than the other classifiers.

II. LITERATURE SURVEY

Many researchers have analyzed different data classification algorithms in various domains. Jesmin Nahar et

al (2013) [15], have analysed the performance of classifier with predictive Apriori for classifying heart disease in men and women. Ms. Ishtake et al (2013) [13] have analyzed the working of Decision tree, Neural network and Naïve Bayes network in predicting heart attack. Rashedur M. Rahman, and Farhana Afroz (2013) [16], have studied the performance of NN, Fuzzy logic and Decision Tree in diagnoising diabetes. Monali Dey and Siddharth Swarup Rautaray (2014) [17], have studied the working of MLP and Naïve Bayes Network in health care domain.

Rich Caruana and Alexandru Niculescu-Mizil (2006) [8] stated that, ―bagging works well with most decision tree types and requires little tuning, but neural nets and SVMs require careful parameter selection. Their research on STATLOG data set concludes that learning methods such as boosting, random forests, bagging, and SVMs achieve excellent performance that would have been difficult to obtain just 15 years ago. Support Vector Machine (SVM) and integer-coded genetic algorithm (GA) were used for heart disease classification by Sumit Bhatia, Praveen Prakash, and G.N. Pillai (2008) [9]. S. Ghumbre, C. Patil, and A. Ghatol (2011) [10],have proved that RBF and SVM have provided high classification accuracy.

III. RESEARCH WORK

Many researchers have focused on analyzing the behavior of Predictive models in data mining. The recent technology advances help obtain large volumes of medical data. These data contain valuable information. Therefore data mining techniques can be used to extract useful patterns. The application of data mining on diseases such as predicting breast cancer, heart attacks, oral diseases, diabetes are widely concentrated. A categorization has been provided based on the different data mining techniques. Generally association mining is suitable for extracting rules. It has been used especially in cancer diagnosis. Classification is a robust method in medical mining. The different methods like Multilayer Perceptron Neural Networks (MLPNN), Decision trees, Radial Basis function Networks (RBF), Naïve-Bayes Claasifier (NB) and Logistic Regression (LR) have been areas of interest for many researchers. The previous research was in studying the classification of performance of RBFNN and MLPNN [14].

A. *Naïve Bayes Network*

This method is based on the work of Thomas Bayes who proposed Bayesian theorem which is used for classification. Bayesian classification is used in application where the relationship between the attribute set and the class variable is non-deterministic. The class label of a test record cannot be predicted with certainty, even though the attribute set is identical to some of the training examples. This may be due to noise or due to the presence of some confounding factors that affect classification but are not included in the analysis. A naïve-bayes classifier estimates the class-conditional probability by assuming that the attributes are conditionally independent, given the class label y. Instead of computing the class-conditional probability of every combination of X, we have to estimate the conditional probability of each Xi given Y. This method is more practical because it does not require a very large training set to obtain a good estimate of the probability. To classify a test record, the naïve bayes classifier computes the posterior probability for each class Y as, $P(Y|X) = \frac{P(Y)\prod_{i=1}^{d} P(X_i|Y)}{P(X)}$ - eqn(1). For a categorical attribute Xi, the conditional probability is estimated according to the fraction of training instances in the class y that takes on a particular attribute value Xi. The class-conditional probability of continuous attributes is calculated using Gaussian distribution and estimate parameters μ *and* σ^2 are used.

Naïve bayes classifier is robust to isolate noise points and irrelevant attributes because such points are averaged out when estimating the conditional probabilities of the data. Correlated attributes can degrade their performance.

Domingos, has verified that the Bayesian classifier performs quite well in practice even when strong attribute dependencies are present. It has also been proved Bayesian classifier does not require attribute independence to be optimal under zero-one loss. Some necessary and some sufficient conditions have been derived for the Bayesian classifier‘s optimality.

B. *Radial Basis Function Network*

The RBF network has a feed forward structure consisting of a single hidden layer of *J* locally tuned units, which are fully interconnected to an output layer of *L* linear units. All hidden units simultaneously receive the *n*-dimensional real valued input vector *X*. The hidden-unit outputs are not calculated using the weighted-sum mechanism/sigmoid activation; rather each hidden-unit output *Zj* is obtained by closeness of the input *X* to an *n*-dimensional parameter vector μ_j associated with the j^{th} hidden unit [10,11]. The response characteristics of the j^{th} hidden unit (j = 1, 2, .., J) is assumed as, $Z_j = K.\| X - \mu_j \|/\sigma_j$,- eqn(2) where *K* is a strictly positive radially symmetric function (kernel) with a unique maximum at its ‘centre’ μ_j and which drops off rapidly to zero away from the centre. The parameter σ_j is the width of the receptive field in the input space from unit *j*. This implies that *Zj* has an appreciable value only when the distance $\| X - \mu_j \|$ is smaller than the width. Given an input vector *X*, the output of the RBF network is the *L*-dimensional activity vector *Y*, whose l^{th} component (l = 1, 2 ..L) is given by, $Y_j = \sum_{j=1}^{l} W_{ji}\, Z_j\,(X)$ - eqn(3).

For l = 1, mapping of eqn. (2) is similar to a polynomial threshold gate. However, in the RBF network, a choice is made to use radially symmetric kernels as ‘hidden units’. RBF networks are best suited for approximating continuous or

piecewise continuous real-valued mapping $f : R^n \rightarrow R^L$, where n is sufficiently small. These approximation problems include classification problems as a special case. From eqns (2) and (3), the RBF network can be viewed as approximating a desired function $f(X)$ by superposition of non-orthogonal, bell-shaped basis functions. The degree of accuracy of these RBF networks can be controlled by three parameters: the number of basis functions used, their location and their width [10–13]. Gaussian basis function for the hidden units given as Zj, for $j = 1, 2,..J$, where $Z_J = \exp\left(\frac{\|X-\mu_j\|^2}{2\sigma_j^2}\right)$ - eqn(4) and μ_j and σ_j are mean and the standard deviation respectively, of the j^{th} unit receptive field and the norm is the Euclidean.

Training RBF Networks

In RBF networks the hidden and output layers play very different roles, and the corresponding ―weights‖ have very different meanings and properties. It is therefore appropriate to use different learning algorithms for them. The input to hidden ―weights‖ (i.e. basis function parameters $\{\mu_{ij}, \sigma_j\}$ can be trained (or set) using any of a number of unsupervised learning techniques. Then, after the input to hidden ―weights‖ is found, they are kept fixed while the hidden to output weights are learned. Since this second stage of training involves just a single layer of weights $\{w_{jk}\}$ and linear output activation functions, the weights can easily be found analytically by solving a set of linear equations. This can be done very quickly, without the need for a set of iterative weight updates as in gradient descent learning.

C. SUPPORT VECTOR MACHINES

The SVM model is a supervised machine learning technique which is based on the statistical theory. It was first proposed by Cortes and Vapnik(1995) [6] from their original work on Structural risk minimization and later modified by Vapnik(1998) [7].

SVM basically works as the linear separator between two data points to identify two different classes in the multidimensional environment. SVM uses a very big set of non-linear features that is task-independent. They have a clever way to prevent over-fitting. They have a very clever way to use a huge number of features without requiring nearly as much computation as seems to be necessary. The prime objective of this approach is to maximize the margin between the classes and to minimize the distance between the hyper plane points.

SVM basically defines the dealing of interaction respective to the features and the repetitive features. SVM splits the dataset in two vector sets under ‘n’ dimensional space vector. The SVM algorithm basically constructs a hyper plane environment so that each element is been compared respective to the separated linear line. Hyper-plane concept is presented to perform the data separation based on largest distance analysis to identify the classes. To reduce the error ratio, the largest margin classifier is defined.

SVM classifiers are based on the class of hyperplanes, $(w.x) + b = 0$ $w \in RN$, $b \in R$, corresponding to decision functions $f(x) = \text{sign}((w.x) + b)$. We can show that the *optimal hyperplane*, defined as the one with the maximal margin of separation between the two classes.
In practical use, the user specifies the kernel function; the transformation φ(.) is not explicitly stated. Given a kernel function K(xi, xj), the transformation φ(.) is given by its eigen functions (a concept in functional analysis). Eigen functions can be difficult to construct explicitly. This is why we specify the kernel function without worrying about the exact transformation. There exists another view that, kernel function being an inner product, is really a similarity measure between the objects.

Some of the kernel functions are:-

a) Polynomial kernel with degree d

$K(X,Y) = (X^T Y + 1)^d$

b) Radial basis function kernel with width σ

$K(X,Y) = \exp(- \|X - Y\|^2 / (2\sigma^2))$

o Closely related to radial basis function neural networks

o The feature space is infinite-dimensional.

c) Sigmoid with parameter θ and .

$K(X,Y) = \tanh(kx^T y + \theta)$

The ‘d’, ‘σ’ and ‘θ’ are parameters chosen by the user.

The feature space is often very high dimensional and does not suffer from curse of dimensionality. A classifier in a high-dimensional space has many parameters and is hard to estimate. Vapnik argues that the fundamental problem is not the number of parameters to be estimated, rather the problem is about the flexibility of a classifier. Typically, a classifier with many parameters is flexible, but there are also exceptions.

“Structural risk minimization” is to get a low error rate on unseen data. It would be really helpful if we could get a guarantee of the following form:
Test error rate = < train error rate + f(n, h, p) ,where, n is size of training set, h is measure of the model complexity, p is the probability that this bound fails.

Algorithm

a. Prepare the pattern matrix

b. Select the kernel function to use

c. Select the parameter of the kernel function and the value of C. [use the values suggested by the SVM software, or set apart a validation set to determine the values of the parameter].

d. Execute the training algorithm and obtain the α_i.

e. Unseen data can be classified using the α_i and the support vectors.

COMPLEXITY OF THE MODEL:

If we choose 'n' data points and assign labels of + or – to them at random and if our model class (e.g. a neural net with a certain number of hidden units) is powerful enough to learn any association of labels with the data, then it is too powerful. The power of a model class is characterized by the number assignment labels that are perfectly learned. This number of data-points is called the Vapnik-Chervonenkis dimension. The model does not need to shatter all sets of data-points of size 'h'.

SVM Training

The goal is to find the separating plane with the largest margin (i.e., find the support vectors). Training a SVM is equivalent to solving a quadratic programming problem with linear constraints (the number of variables is equal to the number of training data).

Strengths of SVM are:

1. Training is relatively easy.
2. No local optimal, unlike in neural networks.
3. It scales relatively well to high dimensional data.
4. Tradeoff between classifier complexity and error can be controlled explicitly.
5. Non-traditional data like strings and trees can be used as input to SVM, instead of feature vectors.

The weakness of SVM is that, the need to choose a "good" kernel function for good performance.

To design learning algorithms, we thus must come up with a class of functions whose capacity can be computed.

IV. EXPERIMENTAL EVALUATION

The working of SVM with different kernels namely, the polynomial kernel, the RBF kernel and the Sigmoidal kernel on different datasets was observed and recorded. The study was applied on the medical datasets available in the UCI repository. The SPECT and STATLOG heart datasets, the WISCONSIN breast cancer dataset and the PIMA Diabetes dataset was considered for the study.

The heart dataset with 14 attributes (age, sex, chest pain type (4 values), resting blood pressure ,serum cholestoral in mg/dl, fasting blood sugar > 120 mg/dl, resting electrocardiographic results (values 0,1,2), maximum heart rate achieved , exercise induced angina, oldpeak = ST depression induced by exercise relative to rest , the slope of the peak exercise ST segment, number of major vessels (0-3) colored by flourosopy , thal: 3 = normal; 6 = fixed defect; 7 = reversable defect, Class variable(absence or presence)), Wisconsin dataset with 10 attributes (age, menopause, tumor-size, inv-nodes, node-caps, deg-malig, breast, breast-quad, irradiat, class(recuurent/ non-recurrent) and PIMA with 9 attributes (Number of times pregnant , Plasma glucose concentration a 2 hours in an oral glucose tolerance test, Diastolic blood pressure (mm Hg) ,Triceps skin fold thickness (mm), 2-Hour serum insulin (mu U/ml) Body mass index (weight in kg/(height in m)^2), Diabetes pedigree function, Age (years), Class variable (absence or presence)) were taken for the study.

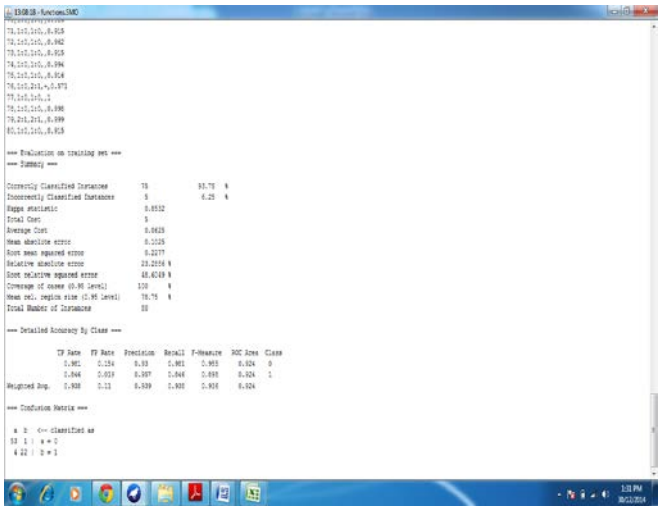

Fig.1. SVM classification in Weka 3.7

10-fold cross validation is performed. This involves breaking the data into 10 sets of size 'N/10'. Then the 9 datasets are trained and 1 set is tested. The process is repeated 10 times and the mean accuracy is taken.

PERFORMANCE MEASURES

a. Accuracy

Accuracy is the overall correctness of the model and is calculated as the sum of correct classifications divided by the total number of classifications. Accuracy is calculated as (TP+TN) / (TP+FN+TN+FN), where TP is true positive, TN is true negative, FN is false positive and FN is false negative.

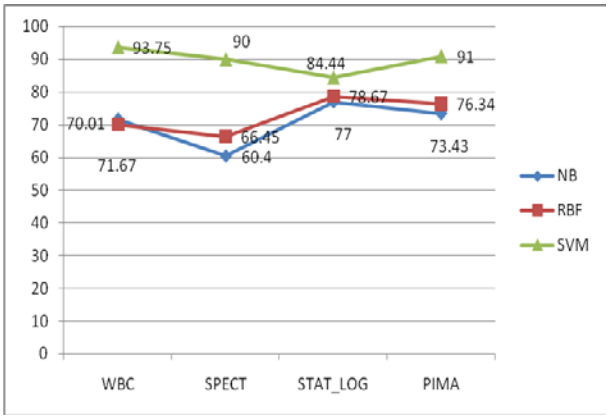

Fig: 2. Accuracy of the Classifiers

b. Precision

Precision is a measure of the accuracy provided that a specific class has been predicted. It is defined by: Precision =

TP/(TP + FN) where TP and FN are the numbers of true positive and false positive predictions for the considered class.

c. Recall

Recall is a measure of the ability of a prediction model to select instances of a certain class from a data set. It is commonly called as sensitivity, and corresponds to the true positive rate. It is defined by the formula: Recall = Sensitivity = TP/(TP+FN) where TP and FN are the numbers of true positive and false negative predictions for the considered class. TP + FN are the total number of test examples of the considered class.

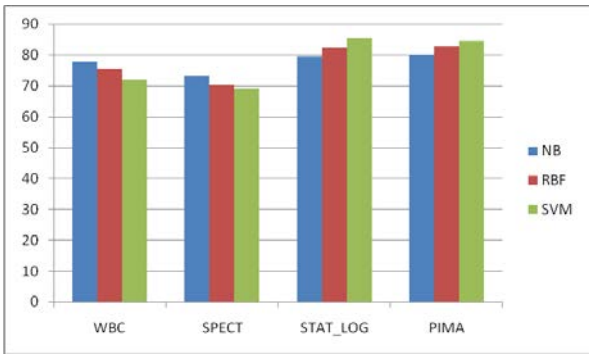

Fig:3. Sensitivity of the classifiers

d. Specificity

Recall/sensitivity is related to specificity, which is a measure that is commonly used in two class problems where one is more interested in a particular class. Specificity corresponds to the true-negative rate. Specificity=TN/(TN+FN).

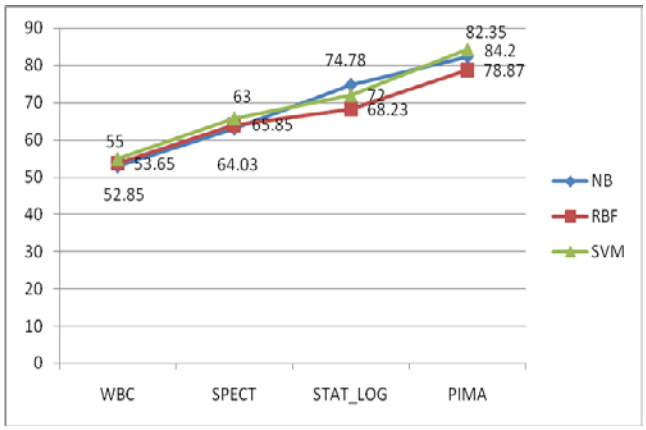

Fig:4. Specificity of the Classifiers

The performance measure in terms of accuracy, specificty and sensitivity of the dataset is shown in Table-1 .

TABLE 1
PERFORMANCE OF SVM ON DATASETS

Performance Measure	Model	WBC	SPECT	STAT_LOG	PIMA
Accuracy	NB	71.67	60.4	77	73.43
	RBF	70.01	66.45	78.67	76.34
	SVM	93.75	90	84.44	91
Specificity	NB	52.85	63	74.78	82.35
	RBF	53.65	64.03	68.23	78.87
	SVM	55	65.85	72	84.2
Sensitivity	NB	77.77	73.23	79.47	80
	RBF	75.34	70.21	82.3	82.65
	SVM	91.3	68.88	93.3	84.6

The graphs below (fig:5 & fig:6) show the ROC curves of RBF and SVM plotted on the same dataset. The accuracy of the test depends on how well the test separates the group being tested into those with and without the disease in question. Accuracy is measured by the area under the ROC curve. An area of 1 represents a perfect test; an area of 0 .5 represents a worthless test. The graph shows an excellent accuraccy with 0.924.

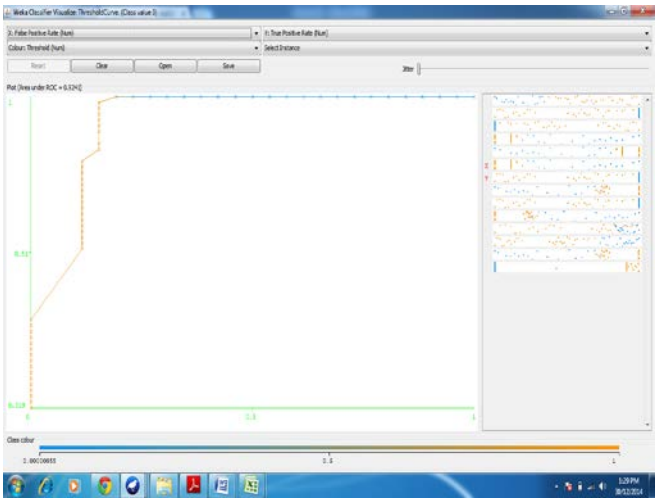

Fig:5. Area under ROC= 0.924 using SVM Classifier

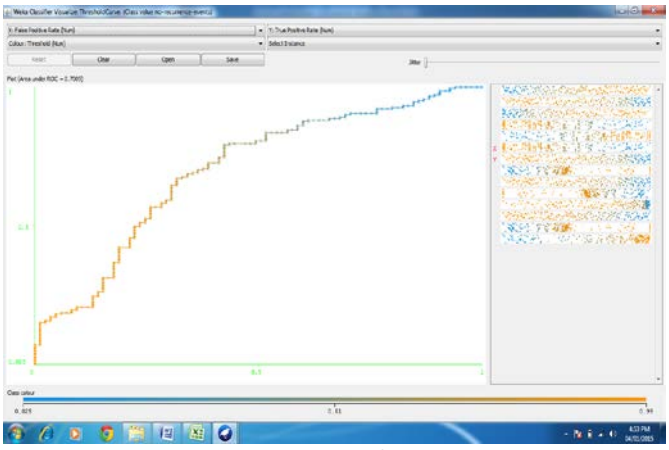

Fig.6. Area under ROC in RBF Classifier

V. CONCLUSION

The study is a binary classification of medical data using the Naïve bayes, RBF and SVM classifier. We have used the heart data set, breast cancer data set and the diabetes dataset provided by the machine learning laboratory at University of California, Irvine. The absence or presence of the disease is classified. It is clear that the accuracy of SVM classifier is more than RBF and NB, while sensitivity is almost equal and

is dependent on the number of positive and negative class samples.

When all the attributes are used to build up classifiers, they operate in high dimensions, and the learning process becomes computationally and analytically complicated, resulting often in the drastic rise of classification error. Hence, there is a need to reduce the dimensionality of the feature space before classification. Future is work is to analyze the performance of the classifier after applying feature extraction techniques.

REFERENCES

[1] Donald, E. (2008), *Introduction to Data Mining for Medical Informatics* , ClinLab Med, pp. 9-35.

[2] Zhang, R., Katta, Y,(2002),*Medical Data Mining*, Data Mining and Knowledge Discovery,pp. 305-308

[3] Harleen Kaur & Siri Krishan Wasan,(2006), *Empirical Study on Applications of Data Mining Techniques in Healthcare*, Journal of Computer Science 2 (2): 194-200, ISSN 1549-3636 Publications.

[4] Prather, J.C., Lobach, D.F., Goodwin, L. K. Hales, J.W, Hage M.L, Edward Hammond W, (1997), *Medical Data Mining: Knowledge Discovery in a clinical Data Warehouse*, Proceedings: a conference of the American Medical Informatics Association / ... AMIA Annual Fall Symposium.

[5] Aslandogan Y, Mahajani G (2004), *Evidence combination in medical data mining*, Proceedings of international conference on information technology.

[6] V. N. Vapnik, The Natural of Statistical Learning Theory, Springer, New York, NY, USA, 1995.

[7] V. N. Vapnik, Statistical Learning Theory, Wiley, New York, NY, USA, 1998.

[8] Rich Caruana and Alexandru Niculescu-Mizil, (2006), *An Empirical Comparison of Supervised Learning Algorithm*, 23rd International Conference on Machine Learning, Pittsburgh.

[9] S. Bhatia, P. Prakash and G.N. Pillai, *SVM based Decision Support System for Heart Disease Classification with Integer-coded Genetic Algorithm to select critical features*, Proceedings of the World Congress on Engineering and Computer Science, San Francisco, USA, pp.34-38, 2008.

[10] S. Ghumbre, C. Patil, and A. Ghatol, *Heart disease diagnosis using support vector machine*,Proceedings of the International Conference on Computer Science and Information Technology (ICCSIT '11), Pattaya, Thailand, 2011.

[11] Xiaoqing Gu, Tongguang Ni, and Hongyuan Wang, *New Fuzzy Support Vector Machine for the Class Imbalance Problem in Medical Datasets Classification* , The Scientific World Journal, Volume 2014 .

[12] V. Anuja Kumari, R.Chitra, *Classification Of Diabetes Disease Using Support Vector Machine,* International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 , Vol. 3, Issue 2, March -April 2013, pp.1797-1801.

[13] Ms. Ishtake S.H ,Prof. Sanap S.A., "Intelligent Heart Disease Prediction System Using Data Mining Techniques", International J. of Healthcare & Biomedical Research,2013

[14] J.Padmavathi, *A Comparative study on Breast Cancer Prediction Using RBF and MLP*, International Journal of Scientific & Engineering Research, Volume 2, Issue 1, January-2011.

[15] Jesmin Nahar, Tasadduq Imama, Kevin S. Tickle, Yi-Ping Phoebe Chen, Association rule mining to detect factors which contribute to heart disease in males and females, Elsevier, 2013.

[16] Rashedur M. Rahman, Farhana Afroz, Comparison of Various Classification Techniques Using Different Data Mining Tools for Diabetes Diagnosis, Journal of Software Engineering and Applications, 2013.

[17] Monali Dey, Siddharth Swarup Rautaray, Study and Analysis of Data mining Algorithms for Healthcare Decision Support System, International Journal of Computer Science and Information Technologies(2014).

Padmavathi Janardhanan has done here research in Data mining in Bharathiar University. She has proposed a hybrid algorithm for data classification in medical domain. She worked in department of computer Science in SRM University. Currently she is working in M,O.P. Vaishnav College

Heena L. and **Fathima Sabika** are under graduate students in the department of Computer Science in M.O.P. Vaishnav College. Both have been working in data mining tool, Weka analyzing its efficient and effective use data analysis.