Recognizing Influences of Liver Enzyme Variation in Physical Disorder – Loom of Data Mining

A . S . Aneeshkumar,
Research Scholar
PG and Research Department
ofComputer Science
Presidency College
Chennai, India

Dr . C . Jothi Venkateswaran,
Dean
Department of Computer Science & Applications,
Presidency College,
Chennai, India

Dr . A . Clementking
Associate Professor,
Department of Computer Science
Computer Science College,
King Khalid University,
Kingdom of Saudi Arabia

***Abstract*— Data mining is an established technology for identifying relevant factors and predicting the occurrence of associated aspects for the estimation and future planning in government and private sector. It is about finding insights which are statistically reliable, significant and previously unknown data. Health care management is one among the major user of the Data mining techniques for diagnosing the attributes for the various medical issues and treatment planning. Disparity in liver enzymes is a liable physical problem which affects other proportional activities of the body.**

Keywords— Data Mining, Classification, Liver Disorder.

I INTRODUCTION

HIV has been an important, familiar health and social crisis for last two decades. HCV infection is less familiar, but also important. These two viruses are similar in a number of ways, and infection with both is a serious problem. Lot of HIV transmissions in India occur within group or network of individuals who is having higher level of risk due to more number of sexual partners or the sharing of drug injection equipments. Drug use and needle sharing is a well-established risk factor for the spread of all sexually transmitted infections (STIs). HIV infected persons, especially injection among drug users, may also be infected with HCV. An estimated 50% to 90% of people who acquired HIV through injected drug usage are co-infected with HCV[1]. According to World Health Organization's (WHO's) estimation, there are 10-24 million HCV infected persons living in India. It isone-fourth of all chronic liver disease in India, because the prevalence and risk factor is not well characterized.

There are possibilities to have HIV in HCV infected persons and vice-versa. But we cannot judge, that all the HCV patients have HIV and all the HIV patients have HCV. Its presentation is almost like a Confusion Matrix, which contains information about the combination of true and false classifications done by a classification system. Clinical database are elements of the domain where application of data mining has a vital role of intelligent diagnosis in healthcare and clinical research. It is possible to acquire knowledge and information concerning a disease from the patient relatedinformation as far as medical data is concerned and therefore, data mining act as a perfect domain in healthcare[2]. Data mining can deliver an accurate assessment from effective comparison and evaluation of causes, symptoms, and courses of treatments [3]. The detection of a disease from several factors or symptoms is a multi-layered problem. Here we are presentingclassification with the correlation of multiple symptoms.

II DATA SET

The total of 128 medical dataset having many attributes with Boolean value in addition to age and sex. Researcherssuggest that compared to people infected with HCV only, co-infected persons had twice the risk for cirrhosis and approximately six times the risk for liver failure [4].Basically these infections doesn't show any symptoms in majority of the cases, when they first acquire. But over time, people with chronic infection may begin to experience the effects of the disease presence. There are some common symptoms of fatigue, Mild fever, Muscle or joint aches, skin disorders, headache, swollen lymph nodes, and sore throat for all Sexually Transmitted Infections (STIs). But some special symptoms for HIV are like, Excess sweating or night sweating, Diarrhoea, Mental illness, lack-of well-being, general feeling of discomfort, Herpes zoster infection, Mouth disorder like gingivitis, oral hairy leukoplakia of the tongue, oral thrush, loss of sensation and inability to control muscles, when compare to HCV [5]. Some of theseare shown in cases of some other STIs. For HCV, Dark urine, dizziness, excessive bleeding, red palms, slow healing and recovery, gray, yellow, white coloured stools.

III APPLIED METHODOLOGY

The proposed hybrid model used for this study is presented in figure 1. Data consist of many missing and inconsistent fields, so it is necessary to reprocess the data, when after collecting it from the repository.Then we are applying various classifications tools.

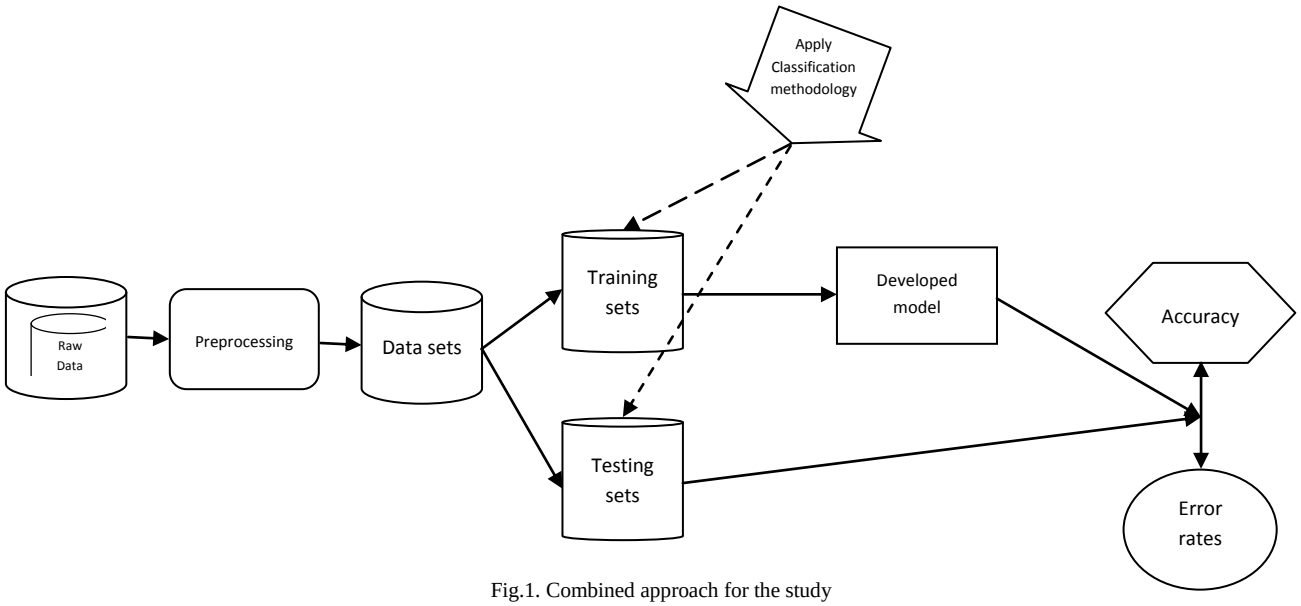

Fig.1. Combined approach for the study

IV CLASSIFICATION

Classification follows two steps, which are learn a function to separates the data classes and calculate the accuracy of the model used for the classification. The accuracy of a classifier on a given test set is the percentage of test set tuples that are correctly classified by the classifier. Here we are comparing some of famous classification algorithm with their accuracy and other statistical measurements.

A. Rough Set

Rough set theory is a mathematical tool to discover hidden patterns in data. The rough set theory is used for several types of data with various sizes [6] for approximate or rough classification. In a given class C, the approximation follows two sets, which are lower approximation of C and upper approximation of C. The lower approximation of C consists of all of the data tuples that, based on the knowledge of the attributes, are certain belong to C without ambiguity. The upper approximation of C consists of all of the tuples that, based on the knowledge of the attributes, cannot be described as not belonging to C [7].That is, let $S = (U, R)$ be an approximation space and X be a subset of$U$. The lower approximation of X by R in S is defined as

$\underline{R}X = \{e \square\square U \mid [e] \square\square X\}$ and

The upper approximation of X by R in S is defined as$\bar{R}X = \{e \square\square U \mid [e] \square\square X \square\square\square\}$ where [e] denotes the equivalence class containing e.

A subset X of U is said to be R-definable in S if and only if $\underline{R}X = \bar{R}X$.The boundary set $BN_R(X)$ is defined as $\bar{R}X - \underline{R}X$.Therefore The pair ($\underline{R}X$, $\bar{R}X$) defines a rough set in S, which is a family of subsets of U with the same lower and upper approximations as $\underline{R}X$ and $\bar{R}X$ [8].

B. Decision Tables

The rows of a decision table induces a decision rule for specify a decision with the participation of satisfied conditions. The columns are labelled by attributes and these attributes are divided into two disjoint groups called condition and decision attributes respectively [9]. If the decision rule uniquely determines decision in terms of conditions, the rule is known as certain, else is uncertain. The decision table is denoted as $S = (U, C, D)$ where C and D are disjoint sets of condition and decision attribute respectively. In each$x \in U$ determines a sequence $C_1(x)$, ……., $c_n(x), d_1(x), \ldots\ldots\ldots, d_m(x)$ where $C = \{c_1, \ldots\ldots\ldots, c_n\} and\ D = \{d_1, \ldots\ldots\ldots\ldots, d_m\}$.

C. Support Vector Machine

Support vector machine (SVM) transforms data set into higher dimension with the help of nonlinear mapping. A linear optimal hyperplane is used to separate these data into two classes and which is useful for linear and nonlinear data [10]. SVM always searches for the largest marginal hyperplane. This hyperplane defines as,

$$W.X + b = 0$$

Where $W = \{w_1, w_2, \ldots\ldots\ldots, w_n\}$, is the weighted vector and b denotes scalar or bias.

For an example, in case of two dimensional training attributes, $X = \{x_1 \& x_2\}$ it defines as,

$$w_0 + w_1 x_1 + w_2 x_2 = 0$$

Where the attributes lies above the hyperplane is denoted as $w_0 + w_1 x_1 + w_2 x_2 > 0$ and below hyperplane as$w_0 + w_1 x_1 + w_2 x_2 < 0$

D. K-Nearest Neighbor

K-Nearest Neighbor (KNN) for non- numeric data is based on learning by analogy, which means it compare the given test tuples with training set tuples, which are similar to that [11]. That is, KNN searches the pattern space for the k-training attributes in the n dimensional space, which are close to these tuples. This closeness is known as *Eucliden distance*and which is calculated between two attributes as,

$$dis(X_1, X_2) = \sqrt{\sum_{i=1}^{n} (x_{1i} - x_{2i})^2}$$

where

$$X_1 = (x_{11}, x_{12}, \dots\dots\dots, x_{1n}) and$$
$$X_2 = (x_{21}, x_{22}, \dots\dots\dots, x_{2n})$$

So it is known as instance based or lazy learning algorithm [12]. Here a tuple is classified according to the majority of votes from its neighbours and all computation should get approximate classification.

E. Multilayer Perceptrons

It is a feed forward neural network and it consists at least one hidden layer [13]. Multilayer perceptron work with non-linear classification problems and therefore, deal more complex decision regions. Each node in the first layer will create hyperplane and second layered nodes combine these hyperplanes to create convex decision regions. Finally the nodes in the last layer will combine convex regions to form concave regions [14]. The activation level for perceptron O_j is defined as,

$$O_j = F\left(\sum_{i=1}^{n} w_i x_i\right)$$

Hidden and output units are calculated according to the activation function, where w_i is the weight on the connection unit i and x_i is the input value. The error rate of hidden layer is calculated as,

$$Err_j = O_j(1 - O_j)\sum Err_j w_{jk}$$

Where w_{jk} is the weight of the connection from unit j to a unit k in the next higher layer and Err_j is the error of unit k.

F. Adaboost

This boosting algorithm uses a weight with each training samples to determine the probability of being selected for a training set. Then it classifies based on weighted vote of weak classifiers[7]. For an example, D is a data set of d's and having class labelled tuples $(X_1, y_1), (X_2, y_2), \dots\dots\dots, (X_d, y_d)$where y_i is the class label of tuple X_i. Initially this algorithm assigns each training tuples an equal weight of $1/d$. Generating k classifiers for the ensemble requires k rounds through the rest of the algorithm. In round i the tuples from D are sampled to form a training set D_i of size d [15]. In this method sampling with replacement is used and so the same tuple may select more than once. But the chance selecting each tuple is according to its weight. The error rate of D_i for the misclassification of the model M_i is defined as,

$$Err = \sum_{j}^{d} w_j \times err(X_j)$$

Where $err(X_j)$ is the misclassification error of tuple X_j and it gives the value of 1 for misclassification and 0 for correct classification.

G. Bagging

Bootstrap aggregation algorithm generates multiple training sets by sampling with replacement from the available training data and assigns vote for each classification. The majority vote made by the large group of classifiers may be more reliable than a majority vote made by a small group of classifiers. Suppose we are having training set D_i of d tuple for all iterations with replacement sampling from the original data D. The classification of unknown tuple, X each classification gives its class prediction and which counts as one vote. The bagging classifiers count the votes and assign the class with the most votes. It often has greater accuracy than a single classifier derived from the original data. Bagging can be applied for continuous values by taking the average of each prediction for a given test data [7].

H. Random Forest

A random forest is an ensemble or collection of unpruned decision tree and it often used for very large training datasets and a very large number of input variables [16].Unlike the construction of single classification trees, Random forest grows many classification trees. To classify a new object from an input vector, put the input vector down each of the trees in the forest. Each tree gives a classification [17]. The trees in random forest gives vote for each class and finally choose the classification having maximum votes.It uses out-of-bag (OOB) samples to measure prediction accuracy [18]. That is each tree in it is constructing with a different bootstrap sample from the original data and one-third of the cases are left out of the bootstrap sample. So at the end of the run, consider class$j$, which got most of the votes every time, then case n is OOB. Errors in class$j$'s proximities to all other classes are small. So the average proximity from case n in class j to the rest of the training data is defined as,

$$\bar{P}(n) = \sum prox^2(n, k)$$

The outlier measure for case n is defined as,

$$n\_samples/\bar{P}(n)$$

I. CART

CART tree uses greedy approach of top-down recursive divide and conquer manner of splitting variable to determine the split [7, 19]. The classification and Regression Tree partitioning will repeated until the node reached for no split, which improve the homogeneity [20]. Prediction of accuracy will be done with the help of termination nodes and it can be considered as a class level in classification and in regression, the average of the response variable in least squares.CARTcan easily handle both numerical and categorical variables, but it is much more complicated for hundreds of levels and variables[21].The pseudo procedure for CART is (Soman *et al*., 2006):

Step 1: Start with root node (t = 1)

Step 2: Search for a split s* among the set if allpossible candidates

Step 3: Split node 1 (t = 1) into two nodes (t = 2, t = 3)using the split s*.

Step 4: Repeat the split search process (t = 2, t = 3) as in steps 1-3 until the tree meets growing rules.

V ESTIMATION METHODOLOGIES

The following tables show the performance of nine major algorithms used for classification. Here we are using different estimation methods for predicting the performance of classification.

A. Cross-validation

In this evaluation, the data partitioned randomly into k subsets with approximate equal size and is known as folds. Then the training and testing will be performed up to k times. For the development of this model, we consider k as 10 and so the folds are $D_1, D_2, D_3, \dots\dots\dots, D_{10}$. Iteration of each fold, that mutually exclusive subset reserve for testing and remaining folds are considered as training sets. That is in first iteration, $D_2, D_3, \dots\dots\dots, D_{10}$ collectively used to develop the model and where D_1 is for test the model.It will continue up to the last fold. So each sample has equal chance for training and one chance for testing. Finally the accuracy will be calculated from the total number of correct classifications from these 10 iterations and error can be computed from the overall loss of iterations.

B. Other Data Splitting

Here we are using 10 splitting methods with different ratios of the total data set for the estimation of classification resultsin above mentioned data set. In first case we are using full data set for the model development. Then as in table 2, third iteration uses 90 percentages of data for testing and 80 in next iteration and so on upto 10 percentage of total number of initial tuples.

VI ANALYSIS AND RESULT

In this study, table 1, shows the time take toforthe model, where K-Nearest Neighbor (K-NN) and CART showsmore fastness in K-fold crossvalidation

TABLE I. Performance evaluation of k- folds cross validation in full datasets

S. No.	classification Methods	10- fold					Full data set				
		A	T	MAE	RMSE	ROC	A	T	MAE	RMSE	ROC
1	Rough set theory	88.46	0.08	0.115	0.340	0.878	93.59	0.08	0.064	0.253	0.917
2	Decision table	87.18	0.06	0.190	0.337	0.822	92.95	0.06	0.130	0.240	0.939
3	Support Vector Machine	85.26	0.02	0.148	0.384	0.836	93.59	0.03	0.064	0.253	0.917
4	K-Nearest Neighbor	**89.10**	0	0.113	0.328	0.901	**100**	0	0.005	0.005	1
5	Multilayer Perceptron	88.46	1.03	0.121	0.315	0.921	98.72	1.05	0.030	0.104	0.999
6	AdaBoost	86.54	0.02	0.139	0.324	0.922	93.59	0	0.088	0.221	0.982
7	Bagging	85.90	0.02	0.160	0.296	0.946	92.31	0.02	0.124	0.225	0.978
8	RandomForest	86.54	0.02	0.126	0.305	0.94	**100**	0.02	0.028	0.081	1
9	CART	88.46	0	0.161	0.319	0.843	91.03	0.02	0.162	0.284	0.881

TABLE II. Performance analysis of testing dataset in 90, 80 and 70 percentage split

S.N.	90%				80%				70%			
	A	MAE	RMSE	ROC	A	MAE	RMSE	ROC	A	MAE	RMSE	ROC
1	87.14	0.129	0.359	0.844	88	0.12	0.346	0.858	89.91	0.101	0.318	0.885
2	85.71	0.225	0.354	0.833	88	0.189	0.327	0.848	89.91	0.180	0.300	0.876
3	87.14	0.129	0.359	0.844	88	0.12	0.346	0.858	90.83	0.092	0.303	0.893
4	**89.29**	0.150	0.314	0.873	88.8	0.135	0.326	0.865	**92.66**	0.089	0.266	0.93
5	88.57	0.125	0.33	0.958	87.2	0.161	0.343	0.864	86.24	0.138	0.350	0.921
6	85.71	0.143	0.378	0.833	88	0.123	0.345	0.82	88.99	0.108	0.282	0.931
7	87.86	0.276	0.366	0.9	87.2	0.203	0.299	0.934	89.91	0.171	0.275	0.943
8	88.57	0.117	0.302	0.923	**89.6**	0.116	0.271	0.945	91.74	0.101	0.271	0.939
9	85.71	0.143	0.378	0.833	88	0.144	0.336	0.848	88.07	0.169	0.329	0.861

TABLE III. Performance analysis of testing datasetin 60, 50 and 40 percentage split

S.N.	60%				50%				40%			
	A	MAE	RMSE	ROC	A	MAE	RMSE	ROC	A	MAE	RMSE	ROC
1	90.43	0.096	0.309	0.889	88.46	0.115	0.340	0.859	87.10	0.129	0.359	0.856
2	89.36	0.178	0.308	0.869	87.18	0.185	0.338	0.832	83.87	0.200	0.358	0.868
3	90.43	0.096	0.309	0.877	88.46	0.115	0.340	0.842	83.87	0.161	0.402	0.815
4	**93.62**	0.077	0.249	0.936	**93.59**	0.074	0.250	0.937	**93.55**	0.073	0.252	0.946
5	91.49	0.100	0282	0.946	88.46	0.118	0.310	0.93	91.94	0.100	0.264	0.963
6	89.36	0.121	0.284	0.95	87.18	0.138	0.315	0.934	87.09	0.132	0.307	0.957
7	89.36	0.147	0.274	0.958	87.18	0.164	0.314	0.934	87.09	0.172	0.311	0.917
8	90.43	0.089	0.249	0.962	91.02	0.092	0.243	0.968	90.32	0.098	0.239	0.982
9	89.36	0.152	0.311	0.869	87.18	0.168	0.342	0.832	87.09	0.171	0.343	0.849

TABLE IV. Performance analysis of testing datasetin 30, 20 and 10 percentage split

S.N.	30%				20%				10%			
	A	MAE	RMSE	ROC	A	MAE	RMSE	ROC	A	MAE	RMSE	ROC
1	89.36	0.106	0.326	0.892	93.55	0.065	0.254	0.932	**93.75**	0.063	0.25	0.938
2	87.23	0.183	0.337	0.869	90.32	0.175	0.297	0.89	**93.75**	0.125	0.243	0.906
3	89.36	0.106	0.326	0.889	90.32	0.097	0.311	0.89	**93.75**	0.063	0.25	0.938
4	**93.62**	0.071	0.251	0.949	**96.77**	0.038	0.178	0.971	**93.75**	0.068	0.248	0.953
5	89.36	0.106	0.294	0.96	93.55	0.071	0.244	0.991	**93.75**	0.069	0.237	1
6	87.23	0.120	0.310	0.983	93.55	0.103	0.249	0.921	**93.75**	0.042	0.128	1
7	87.23	0.152	0.279	0.979	93.55	0.122	0.203	0.991	**93.75**	0.092	0.181	1
8	87.23	0.1	0.256	0.965	93.55	0.055	0.194	0.998	**93.75**	0.05	0.2	1
9	87.23	0.166	0.342	0.869	**96.77**	0.085	0.178	0.958	**93.75**	0.121	0.236	0.938

when compare to others. But in other estimation methodologies, Bagging is equally competitive with K-NN. In case of accuracy K-NN gives maximum result (89.10) for cross evaluation and for full dataset model K-NN and Random forest shares best accuracy as 100 percentages. Where the largest value for the mean absolute error(0.190) produced by thedecision tree and root mean square error rate 0.384 by SVM in cross validation (cv) and CART shares 0.162 and 0.284 respectively in complete training set (ct). The ROC area varies from 0.94 to 0.946 in cv and in ct 0.881 to 1.

All the classification models having accuracy fluctuation ineach split, and shows lesser accuracy in 50 to 70 ratios. But according to table 4, we can see that all the classifications having equal and good accuracy in 10 percentage test set of whole data set. Multilayer perceptron shows more fluctuation and the overall performance of K-Nearest Neighbor is good for this particular data set except in 80 percentage split. This accuracy and error differences are shown in figure 2 and 3 respectively.

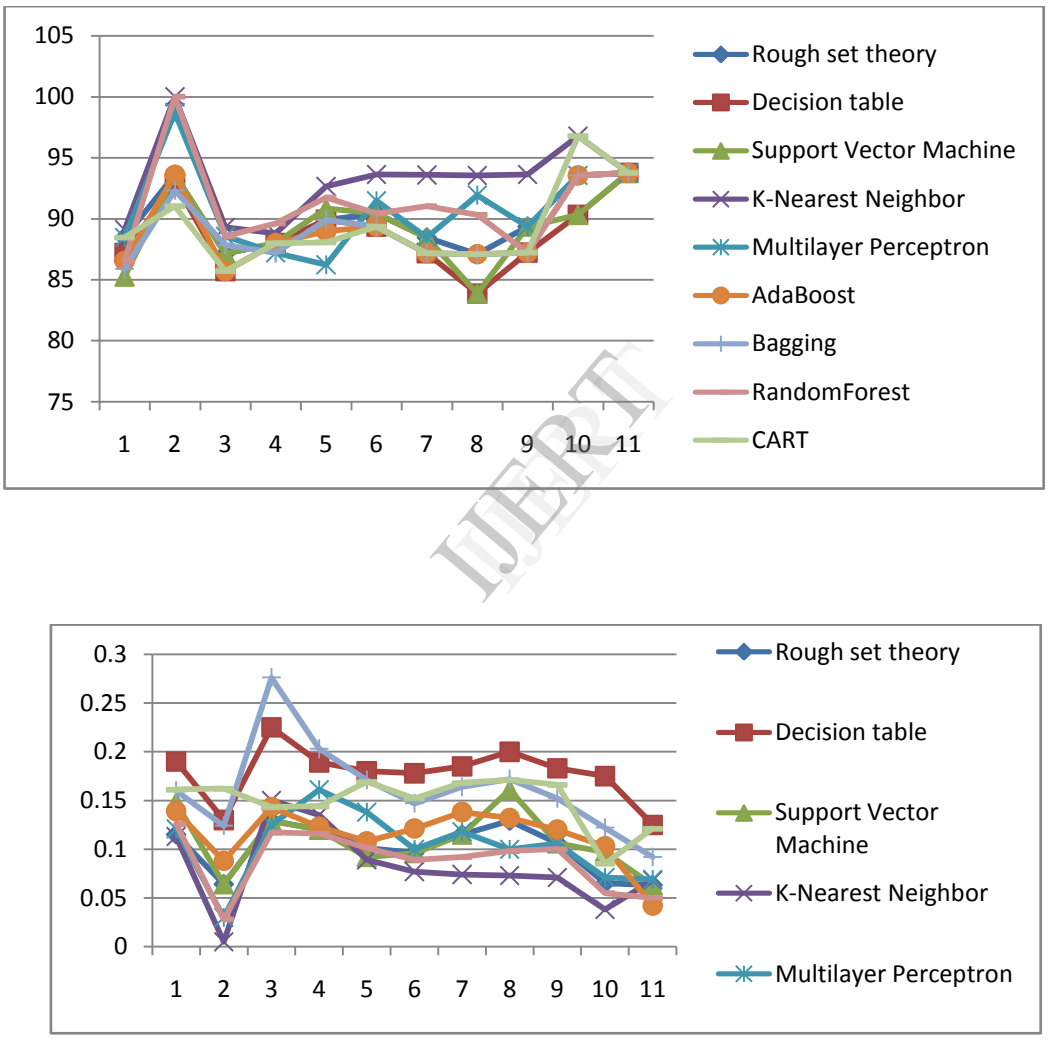

Fig.3. Error evaluation

Decision table produced 7 rules from the total of 191 subsets, where the feature sets are 12, 14, and 16. It gives the merit for the best subset is 92.949. The number of kernel evaluation in SVM is 5907 that is 81.705 percentage of the set. In case of K-NN, it used one nearest neighbour for instance based classification. Adaboost performed 10 number of iteration and calculate 0.35 as the weight.Random forest generated 10 trees, when each one constructed while considering 5 random features and the out of bag error is 0.1603.

VII CONCLUSION

In this paper we develop a clinical decision support model for HIV and HCVclassification from symptoms.According to this evaluation K-nearest neighbour algorithm is more suitable for the prediction of diseases with more complicated symptoms. So this model may helpful for medical experts to guess the problem of the patients.. In future, we will extend this work to predict outliers and co-occurrences of other similar diseases.

ACKNOWLEDGEMENT

We express our sincere thanks to the Director Dr.(Capt.) K. J. Jayakumar M.S., M.N.A.M.S., F.A.I.S. and Chief Manager Dr. R. Rajamahendran, B.Sc., M.B.B.S., D.M.C.H., D.H.H.M., P.G.D.H.S.C.(Diab.), F.C.D., Sir Ivan Stedeford Hospital, Chennai for providing permission to collect data. We are grateful to the chief Manager for his guidance and also would like to thank other hospital staffs for their valuable suggestions and help throughout this study.

RFERENCES

[1] Melissa A.Marx, K.G.Murugavel, Patrick M.Tarwater, A.K.SriKrishnan, David L.Thomas and Suniti Solomon, "Association of Hepatitis C Virus Infection with Sexual Exposure in Southern India", 514, CID 2003:37 (15 August), Marx et at.

[2] Richard Adderley and Peter B.Musgrove, "Data Mining Case Study: Modeling the Behavior of Offenders Who Commit Serious Sexual Assauits", Violent Crime Linkage Analysis System, ACM 2001, KDD 01, San Francisco CA, USA.

[3] Y.Alp Aslandogan and Gauri A.Mahajani, "Evidence Combination in Medical Data Mining", Proceedings of the Intrenational Conference on Information Technology: Coding and Computing (ITCC'04), IEEE.

[4] Brain Leke-Betechuoh, Tshilidzi Marwala, Taryn Tim and Monica Lagazio, "Prediction of HIV Status fron Demographic Data Using Neural Networks", IEEE, December 2005.

[5] Ahmed Mohamed samir ali gama eldin, "A Data mining Approach for the Prediction of Hepatitis C Virus protease Cleavage Sites", International Journal of Advanced Computer Science and Applications, Vol.2, No.12, December 2011.

[6] Peiman Mamani Barnaghi, Vahid Alizadeh Sahzabi and Azuraliza Abu Bakar, "A Comparitive study for Various methods of Classification", 2012 International Conference on Information and Computer Networks (ICICN 2012) IPCSIT Vol.27(2012), IACSIT Press, Singapore.

[7] Jiawei Han and Micheline Kamber, Data Mining Concepts and Techniques, Published by Elsevier, second edition – 2006.

[8] Chien-Chung Chan, "Rough Sets Theory and Its Applications", www.cs.uakron.edu/ chan/DataMining/References/RoughsetanditsApplications.pdf

[9] Zdzislaw Pawlak, "Rough set theory and its applications", Journal of Telecommunications and Information Technology, March 2002.

[10] Chih-Wei Hsu, Chih-Chung Chang, and Chih-Jen Lin, "Practical Guide to Support Vector Classification", http://www.csie.ntu.edu.tw/~ cjlin

[11] Thomas M.Cover and Peter E.Hart, "Nearest neighbor pattern classification, "IEEE Transaction on Information Theory, Vol. 13, issue 1, 1967, pp.21-27.

[12] Thomas B.Fomby, "K-Nearest Neighbor Algorithm: Prediction and Classification", www.faculty.smu.edu/tfomby/eco5385/lecture/K-NN Methods.pdf.

[13] Kohonen.T, "Self-organization and Association Memory", Springer-Verlag, New York, 1988.

[14] LiMin Fu, "Neural Networks in Computer Intelligence, Tata McGraw-Hill Edition 2003, New York.

[15] Asha.T, S.Natarajan and K.N.B.Murthy, "A Data Mining Approach to the Diagnosis of Tuberculosis by Cascading Clustering and Classification", proc. International Conference on Information Processing, Aug 2010.

[16] Graham Williams, "DATA MINING Desktop Survival Guide", http//:www.Data Mining Survivor Contents - Random Forests.htm

[17] Leo Breiman and Adele Cutler, "Random Forests", http//: www.Random forests - classification description.html

[18] Andrew, "Random Forest- Variable Importance", http//:www.Random Forest Variable Importance (news & tutorials).htm

[19] T. Santhanam and Shyam Sundaram, "Application of CART Algorithm in Blood Donors Classification", Journal of ComputerScience 6 (5): 548-552, 2010 , ISSN 1549-3636, 2010 Science Publications

[20] David G.T.Denison, Bani K.Mallick and Adrian F.M.Smith," A bayesian CART algorithm", Biometrica (1998), 85,2 pp. 363-377, Great Britain.

[21] Roman Timofeev and Dr. Wolfgang Hfardle, "Classification and Regression Trees (CART) Theory and Applications", Master of Art Theses, Center of Applied Statistics and Economics, Humboldt University, Berlin, December 20, 2004.