PERFORMANCE ANALYSIS OF VARIOUS DATA MINING CLASSIFICATION TECHNIQUES ON HEALTHCARE DATA

Shelly Gupta[1], Dharminder Kumar[2] and Anand Sharma[3]

[1]AIM & ACT, Banasthali University, Banasthali, India
shelly.gupta24@gmail.com
[2]Department of CSE, GJUS&T, Hisar, India
dr_dk_kumar_02@yahoo.com
[3]Department of CSE, GJUS&T, Hisar, India
andz24@gmail.com

ABSTRACT

Health care data includes patient centric data, their treatment data and resource management data. It is very massive and information rich. Valuable knowledge i.e. hidden relationships and trends in data can be discovered from the application of data mining techniques on healthcare data. Data mining techniques have been used in healthcare research and known to be effective. The present study aimed to do the performance analysis of several data mining classification techniques using three different machine learning tools over the healthcare datasets. In this study, different data mining classification techniques have been tested on four different healthcare datasets. The standards used are percentage of accuracy and error rate of every applied classification technique. The experiments are done using the 10 fold cross validation method. A suitable technique for a particular dataset is chosen based on highest classification accuracy and least error rate.

KEYWORDS

KDD, Data Mining, Classification, Healthcare Datasets and Machine Learning Tools

1. INTRODUCTION

In today's information age, there is a need for a powerful analytical solution for the extraction of the useful information from the large amount of data collected and stored in an organization's databases or repositories. This has led to the emergence of Knowledge Discovery in Databases (KDD) which is responsible for transforming low-level data into high-level knowledge for decision making. Knowledge discovery in databases consists of the list of iterative sequence steps of processes and data mining is one of the KDD processes. Data mining is the application of algorithms for extracting patterns from large volume of data. There is a wealth of data available within the healthcare systems[1]. The healthcare environment is information rich yet knowledge poor. Hence, for healthcare research, data driven statistical research has become a complement. As with the use of computers powered with automated tools the large volumes of healthcare data are being collected and made available to the medical research groups. As a result, Knowledge Discovery in Databases (KDD), which includes data mining techniques, has become a popular research tool for healthcare researchers to identify and exploit patterns and relationships among large number of variables, and also made them able to predict the outcome of a disease using the historical cases stored within datasets.

In this paper, we carried out the performance analysis of various participating data mining classification techniques on healthcare data. This work has helped in determining the best

DOI : 10.5121/ijcsit.2011.3413

classification techniques in terms of its accuracy and error rate in vis-a-vis to the participating techniques on a specific dataset. For this, we have practiced four different healthcare datasets taken from UCI Machine Learning Repository. The examined classification techniques are k-Nearest Neighbours (kNN), Naive Bayes(NB), Support Vector Machine(SVM),CART, Decision tree, Multiple Layer Perceptron (MLP), ID3 etc. The training performance of these techniques measured according to their accuracy. Machine learning tools like WEKA, TANAGRA and CLEMENTINE are used to handle classification problems. This study will help the researchers to determine the better results from the available data within the datasets after knowing the most suitable classification technique for a particular dataset. Hence, the use of suitable classification technique over the healthcare datasets will make the researchers to do the data mining analysis more effectively.

2. LITERATURE REVIEW

This section summarises various review and technical articles on KDD process and data mining classification techniques applied on healthcare datasets. It gives an overview of the current research being carried out on various healthcare datasets using the data mining techniques.

2.1. KDD and Data Mining

The motivation for handling data and performing computation is the discovery of knowledge. For this, we store data about a certain process and retrieve later that information in order to use it in a meaningful way. The KDD process employs data mining methods to identify patterns at some measure of interestingness. The KDD is the process of turning the low-level data into high-level knowledge. Hence, Frawley et al.[2] referred the knowledge discovery in databases as the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data.

The terms Knowledge Discovery in Databases (KDD) and Data Mining are often used interchangeably. This dilemma is because of the three different perspectives to look at the data mining but in real data mining is an important step in the KDD process. Han et al.[3] defined the data mining as the process of discovering interesting knowledge from large amounts of data stored either in databases, data warehouses, or other information repositories. Witten et al.[4]defined data mining as the process of extracting implicit, previously unknown and potentially useful information from data. Hand et al.[5]defined data mining as the analysis of observational data set to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner. On studying the various different definitions on data mining Zhou[6] finally suggested that the database, machine learning and statistics perspectives of data mining put particular emphases on efficiency, effectiveness and validity respectively.

Data mining refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases. It uses machine learning, statistical and visualization techniques to discover and present knowledge in a form which is easily comprehensible to us. The richness and fast evolution of the data mining discipline comes from its large variety of research areas of interest. Data mining applications can use different kind of parameters to examine the data. They include association (patterns where one event is connected to another event), sequence or path analysis (patterns where one event leads to another event), classification (identification of new patterns with predefined targets) and clustering (grouping of identical or similar objects).

2.2. Classification Techniques on Healthcare Data

Data mining has been applied to a variety of healthcare domains to improve decision making. But one of the major challenges in healthcare domain is the extraction of comprehensible

knowledge from diagnosis data. Here, the application of data mining classification techniques is reviewed which are applied on the different diagnostic datasets.

Orlando Anunciacao et al.[7] explored the applicability of decision trees for detection of high-risk breast cancer groups over the dataset produced by Department of Genetics of faculty of Medical Sciences of Universidade Nova de Lisboa with 164 controls and 94 cases in WEKA machine learning tool. To statistically validate the association found, permutation tests were used. They found a high-risk breast cancer group composed of 13 cases and only 1 control, with a Fisher Exact Test(for validation) value of $9.7*10^{-6}$ and a p-value of 0.017. These results showed that it is possible to find statistically significant associations with breast cancer by deriving a decision tree and selecting the best leaf.

A. Soltani Sarvestani et al.[8] provided a comparison among the capabilities of various neural networks such as Multilayer Perceptron (MLP), Self Organizing Map(SOM), Radial Basis Function (RBF) and Probabilistic Neural Network(PNN) which are used to classify WBC and NHBCD data. The performance of these neural network structures was investigated for breast cancer diagnosis problem. RBF and PNN were proved as the best classifiers in the training set. But the PNN gave the best classification accuracy when the test set is considered. This work showed that statistical neural networks can be effectively used for breast cancer diagnosis as by applying several neural network structures a diagnostic system was constructed that performed quite well.

Dr. Medhat Mohamed Ahmed Abdelaal et al.[9] investigated the capability of the classification SVM with Tree Boost and Tree Forest in analyzing the DDSM dataset for the extraction of the mammographic mass features along with age that discriminates true and false cases. Here, SVM techniques show promising results for increasing diagnostic accuracy of classifying the cases witnessed by the largest area under the ROC curve comparable to values for tree boost and tree forest.

K. Rajiv Gandhi et al.[10] constructed classification rules using the Particle Swarm Optimization Algorithm for breast cancer datasets. In this study to cope with heavy computational efforts, the problem of feature subset selection as a pre-processing step was used which learns fuzzy rules bases using GA implementing the Pittsburgh approach. It was used to produce a smaller fuzzy rule bases system with higher accuracy . The resulted datasets after feature selection were used for classification using particle swarm optimization algorithm. The rules developed were with rate of accuracy defining the underlying attributes effectively.

Manaswini Pradhan et al.[11] suggested an Artificial Neural Network (ANN) based classification model as one of the powerful method in intelligent field for classifying diabetic patients. The neural network, used in back propagation algorithm, is m-n-1 type network. The GA is used for optimally finding out the number of neurons in the single hidden layered model. For training and testing 10-fold cross validation method was adopted for Pima Indian Diabetes. For Pima dataset the ANN gives the best accuracy with 5 neurons in the hidden layer. Best accuracy being 72% with average accuracy of 72.2%. The designed model was compared with the Functional Link ANN (FLANN) and several classification systems like NN (nearest neighbor), kNN(k-nearest neighbor), BSS(nearest neighbor with backward sequential selection of feature, MFS1(multiple feature subset) , MFS2(multiple feature subset) for Data classification accuracies. It was revealed from the simulation that their suggested model performed better than compared to all of the participating techniques for comparison.

J. Padmavati[12] performed a comparative study on WBC dataset for breast cancer prediction using RBF and MLP along with logistic regression. Logistic regression was performed using logistic regression in SPSS package and MLP and RBF were constructed using MATLAB. It was observed that neural networks took slightly higher time than logistic regression but the sensitivity and specificity of both neural network models had a better predictive power over

logistic regression. When comparing RBF and MLP neural network models , it was found that RBF had good predictive capabilities and also time taken by RBF was less than MLP.

Humar Kahramanli et al.[13] presented an algorithm for extracting comprehensible classification rules for diagnosis of liver disorders. This algorithm considered all input attributes and extracts rules from the trained neural network with adaptive activation function efficiently. They observed that the neural network trained with adaptive activation function achieved high classification accuracy. Therefore, Neural network was trained by adaptive activation function in hidden layer and fixed sigmoid activation function in output layer. OptaiNET that is an Artificial Immune Algorithm (AIS) used in extracting rules from the trained neural networks. This approach was applied to BUPA Liver Disorders classification problems. The results of comparison experiments showed that the developed approach generated more accurate rules.

N. Suneetha et al.[14] proposed a tree - base approach to analyze multiple response using classification algorithms compared with a modified decision tree method for classification to overcome the known problems for the Gini-based decision tree method and normalizing the Gini indexes by taking into account information about the splitting status of all attributes. Instead of using the Gini index for attribute selection as usual, they used ratios of Gini indexes and their splitting values in order to reduce the biases. They experimented this approach on medical Heart Diseases dataset and results showed that the modified decision tree method reacted differently with the heart dataset when compared to other known decision tree methods. The modified Gini index method performed well for some data bases. Compared to previous multivariate decision tree methods that have limitations on the type of response and size of data the proposed method can analyze any type of multiple response by using this split formulae.

Alaa M. Elsayad[15] investigated three different data mining methods; multilayer perceptron neural network, C5.0 decision tree and linear discriminate analysis in order to build an ensemble model to the problem of differential diagnosis of these erythemato-squamous diseases. The dermatology dataset investigated in this study was taken from the University of California at Irvine (UCI) machine learning repository. The proposed ensemble combined the models using a confidence-weighted voting scheme. The classification performance of the proposed system wass presented using statistical accuracy, specificity and sensitivity. The performance of MLPNN was enhanced using the scored predictions of C5.0 DT and LDA models in the proposed ensemble arrangement. Classification accuracies of the ensemble were very close to those achieved by the work of E. Ubeyli[16] using multiclass support vector machine model.

Chul-Heui Lee et al.[17] proposed a new classification method based on the hierarchical granulation structure using the rough set theory. The hierarchical granulation structure was adopted to find the classification rules effectively. The classification rules had minimal attributes and the knowledge reduction was accomplished by using the upper and lower approximations of rough sets. A simulation was performed on WBC dataset to show the effectiveness of the proposed method. The simulation result showed that the proposed classification method generated minimal classification rules and made the analysis of information system easy.

Sepehr M. H. Jamarani et al.[18] presented an approach for early breast cancer diagnosis by applying combination of ANN and multiwavelet based sub band image decomposition. The proposed approach was tested using the MIAS mammographic databases and images collected from local hospitals. The best performance was achieved by BiGHM2 multiwavelet with areas ranging around 0.96 under ROC curve. The proposed approach could assist the radiologists in mammogram analysis and diagnostic decision making.

M. Lundin et al.[19] have applied ANN on 951 instances dataset of Turku University Central Hospital and City Hospital of Turku to evaluate the accuracy of neural networks in predicting 5, 10 and 15 years breast cancer specific survival. The values of ROC curve for 5 years was evaluated as 0.909,for 10 years 0.086 and for 15 years 0.883, these values were used as a

measure of accuracy of the prediction model. They compared 82/300 false prediction of logistic regression with 49/300 of ANN for survival estimation and found ANN predicted survival with higher accuracy.

W. Nick Street[20] applied ANN classification to Wisconsin Prognostic Breast Cancer and SEER datasets for the analysis of survival. He developed a novel encoding as good and poor prognosis of censored data in an ANN architecture to provide a framework for prognostic prediction. Chih-Lin Chi et al.[21] used the Street's ANN model for Breast Cancer Prognosis on WPBC data and Love data. In their research they used recurrence at five years as a cut point to define the level of risk. The applied models successfully predicted recurrence probability and separated patients with good(>5 yrs) and bad(<5 yrs) prognoses.

Jong Pill Choi et al.[22] compared the performance of an Artificial Neural Network, a Bayesian Network and a Hybrid Network used to predict breast cancer prognosis. The hybrid Network combined both ANN and Bayesian Network. The Nine variables of SEER data which were clinically accepted were used as inputs for the networks. The accuracy of ANN(88.8%) and Hybrid Network(87.2%) were very similar and they both outperformed the Bayesian Network. They found the proposed Hybrid model can also be useful to take decisions.

Delen et al.[23] compared ANN, decision tree and logistic regression techniques for breast cancer survival analysis . They used the SEER(Surveillance Epidemiology and End Results) data's twenty variables in the prediction models. The decision tree with 93.6% accuracy and ANN with 91.2% were found more superior to logistic regression with 89.2% accuracy. C4.5 is a well known decision tree induction learning technique which has been used by Abdelghani Bellaachia et al.[24] along with two other techniques i.e. Naïve Bayes and Back-Propagated Neural Network. They also presented an analysis of the prediction of survivability rate of breast cancer patients using above data mining techniques and used the new version of the SEER Breast Cancer Data. The pre-processed data set consists of 151,886 records, which have all the available 16 fields from the SEER database. They have adopted a different approach in the pre-classification process by including three fields: STR(Survival Time Recode), VSR(Vital Status Recode), and COD(Cause Of Death) and used the Weka toolkit to experiment with these three data mining algorithms. Several experiments were conducted using these algorithms. The achieved prediction performances are comparable to existing techniques. However, they found out that model generated by C4.5 algorithm for the given data has a much better performance than the other two techniques. The results obtained in their work differed from the study of Delen et al because they used a newer version of same dataset, a different pre-classification and different toolkit. Their experimental results showed that their approach outperformed the approach used by Delen et al. They also proposed that after including the missing data in EOD attribute of used dataset can also increase the performance more.

2.3. Conclusion from literature Review

The application of more techniques, new approaches and different tools over the newer version of same dataset can improve the decision making. This can help the healthcare researchers to do better decision making. From above discussion it is clear that more efficient work can be done over the healthcare problems by using new approaches in data mining.

3. RESEARCH METHODOLOGY

We have followed the Knowledge Discovery in Database (KDD) approach as the research methodology. Knowledge Discovery in Databases is responsible to transform low-level data into high-level knowledge for decision making[25]. KDD is the nontrivial process of identifying valid, new, potentially useful, and ultimately understandable patterns in data. Knowledge discovery process consists of the list of iterative sequence steps of processes and data mining is one of the KDD processes.

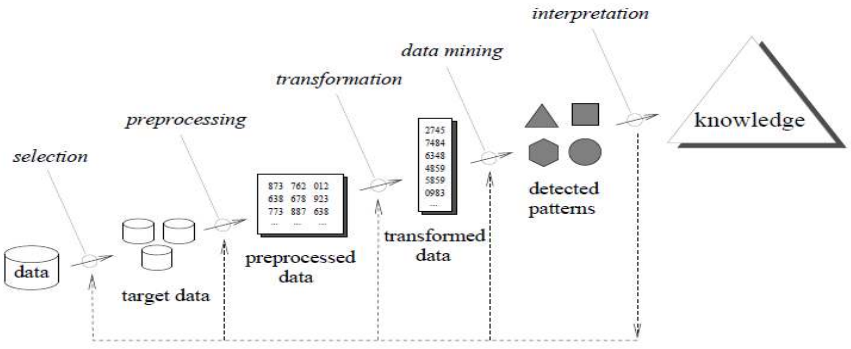

Figure 1. Steps in KDD Process

A brief description of various steps in KDD iterative process is given below:

3.1. Selection Step

In this step the data relevant to the analysis task are retrieved from the database. In selection the target dataset is created which will undergo analysis. In this study, four different healthcare datasets are selected from the UCI repository[26] for the performance analysis of several data mining classification techniques.

3.2. Preprocessing step

The available databases are highly susceptible to noisy, missing, and inconsistent data due to their typically huge size, complexity and their likely origin from multiple heterogeneous sources. So, in this step the target dataset which is selected during the selection step is pre-processed to handle the above problem. In this study, the chosen healthcare datasets are handled for missing values, noise and transformed into a form that is presentable to the classification techniques mostly with the help of Clementine11.1 data mining tool.

3.3. Transformation step

In this step data are transformed or consolidated into forms appropriate for mining by performing smoothing, summary or aggregation, generalization, normalization, discretization and feature construction operations. In this study, Clementine and Tanagra data mining tools are used for the above purpose. In Clementine there are modules like derive, feature selection, type etc. which are linked to the dataset for data transformation as required for the analysis. Where as in Tanagra there are options in the dataset descriptor for the same.

3.4. Data Mining step

In the KDD process, the data mining methods are for extracting patterns from data. In this step of KDD process intelligent methods are applied in order to extract data patterns. The dataset is analyzed based on applied data mining task. In this work, data mining classification techniques like J48, kNN, FT, NB, LMT, SVM, C-RT, QUEST, MLP, ID3, Bayes Net, C4.5, CHAID, LDA, NN-RBFN, Prototype-NN, SPegasos etc. are used to extract the data patterns on healthcare datasets using three different machine learning tools.

3.5. Interpretation step

This step involves pattern evaluation and knowledge representation. This essential step uses visualization techniques to help users understand and interpret the data mining results. In this

analysis, the classification technique with highest percentage of classification accuracy is considered as the most suitable classification technique for a particular dataset. The comparison is shown with help of charts. Based upon the classifier generated by that classification technique, healthcare researchers perform the decision making.

4. EXPERIMENTAL RESULTS AND ANALYSIS

This section presents the experimental results and analysis done for this study. The research methodology for experiments has been explained in section 3. For the experiments, various different classification techniques have been applied on the four different healthcare datasets taken from UCI repository. Table 1 shows the description of datasets selected for this work.

Table 1. Dataset Description

Dataset	Attributes	Instances	Classes
PIMA Indian Diabetes	9	768	2
Wisconsin Breast Cancer	11	699	2
StatLog Heart Disease	14	270	2
BUPA Liver Disorder	7	345	2

In this study WEKA, Tanagra and Clementine machine learning tools for data mining are used to achieve the proposed objectives. The percentage of accuracy rate and error rate for classification techniques are used as the measurement parameters for analysis. These parameters suggests that a high value of accuracy rate and low value of error rate for a classification technique applied on a dataset show that the dataset is highly correctly classified by the obtained classifier. And in reverse a low value of accuracy rate and high value of error rate for a classification technique applied on a dataset show that the dataset is less correctly classified by the obtained classifier.

For experiments, the data is firstly divided into training data and testing data. The training set is used to build the classifier and test set used to validate it. In this study the percentages used for training and testing data are 66% and 34% respectively. Then, the participating classification techniques are applied using the 10 fold cross validation method to generate the classifiers via above mentioned machine learning tools. At last the results are recorded in terms of percentage of accuracy and error rates. The results are shown as below:

4.1. Results for classification techniques applied on PIMA Indian Diabetes dataset

Table 2, 3 and 4 show the results for classification techniques applied on PIMA Indian Diabetes dataset in WEKA, Tanagra and Clementine respectively. On the basis of comparison done over accuracy and error rates; the classification techniques with highest accuracy are obtained for PIMA Indian Diabetes dataset in given different machine learning tools. Figure 2 shows comparison between the best classification techniques applied on this dataset. From Figure 2 it is observed that SVM classification technique applied in Tanagra for this dataset is best among all other participating techniques applied in other tools.

Table 2. Results obtained in WEKA

Technique Applied	Accuracy Rate	Error Rate
Bayes Net	74.34	25.65
Naïve Bayes	76.30	23.70
J48	73.82	26.17
MLP	75.39	24.61
SMO	77.34	22.66
Logistic	77.21	22.79
LMT	77.47	22.53
S Pegasos	77.73	22.27
FT	77.34	22.66

Table 3. Results obtained in Tanagra

Technique Applied	Accuracy Rate	Error Rate
C4.5	84.5	15.5
ID3	77.2	22.78
SVM	96.74	3.18
kNN	80.3	19.66
Prototype NN	63.28	36.71
CRT	78.51	21.48
LDA	78.3	21.7

Table 4. Results obtained in Clementine

Technique Applied	Accuracy Rate	Error Rate
NN-RBFN	78.26	21.74
C5.0	82.29	17.71
C&RT	81.25	18.75
QUEST	76.17	23.83
CHAID	77.6	22.4
LDA	76.82	23.18
Logistic	78.26	21.74

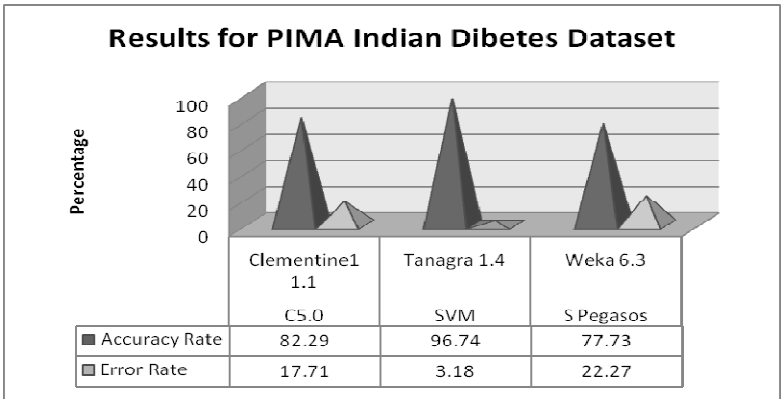

Figure 2. Comparison between best classification techniques applied for PIMA Indian Diabetes Dataset

4.2. Results for classification techniques applied on Wisconsin Breast Cancer dataset

Table 5, 6 and 7 show the results for classification techniques applied on Wisconsin Breast Cancer dataset in WEKA, Tanagra and Clementine respectively. On the basis of comparison done over accuracy and error rates; the classification techniques with highest accuracy are obtained for this dataset in given different machine learning tools. Figure 3 shows comparison between the best classification techniques applied on this dataset. From Figure 3 it is observed that all the three best classification technique applied in given tools performed equally for this dataset.

Table 5. Results obtained in WEKA

Technique Applied	Accuracy Rate	Error Rate
Bayes Net	97.13	2.87
Naïve Bayes	85.40	14.59
J48	94.56	5.44
MLP	95.27	4.73
SMO	96.99	3.01
Logistic	96.56	3.54
LMT	95.99	4.01
S Pegasos	96.85	3.15
FT	96.99	3.01

Table 6. Results obtained in Tanagra

Technique Applied	Accuracy Rate	Error Rate
C4.5	95.42	4.57
ID3	92.41	7.29
SVM	97.13	2.80
kNN	97.28	2.71
Prototype NN	96.28	3.71
CRT	92.41	7.29
LDA	95.99	4.01

Table 7. Results obtained in Clementine

Technique Applied	Accuracy Rate	Error Rate
NN-RBFN	97.28	2.72
C5.0	94.28	5.72
C&RT	95.99	4.01
QUEST	96.14	3.86
CHAID	95.99	4.01
LDA	94.71	5.29
Logistic	97	3

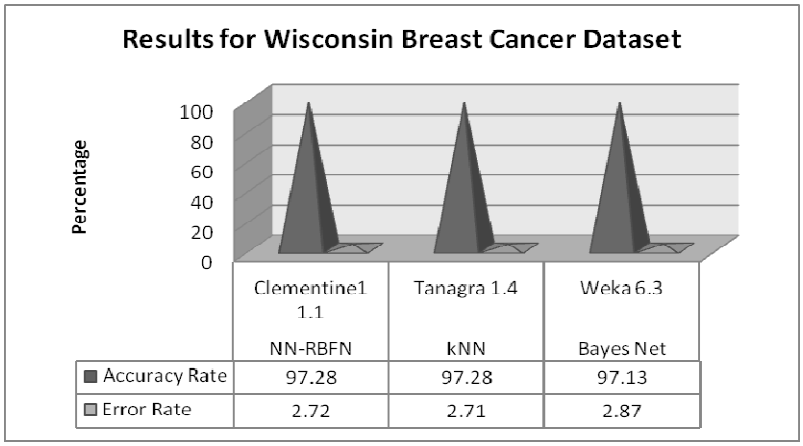

Figure 3. Comparison between best classification techniques applied for Wisconsin Breast Cancer Dataset

4.3. Results for classification techniques applied on BUPA Liver Disorder dataset

Table 8, 9 and 10 show the results for classification techniques applied on BUPA Liver Disorder dataset in WEKA, Tanagra and Clementine respectively. On the basis of comparison done over accuracy and error rates; the classification techniques with highest accuracy are obtained for this dataset in given different machine learning tools. Figure 4 shows comparison between the best classification techniques applied on this dataset. From Figure 4 it is observed that C4.5 classification technique applied in Tanagra for this dataset is best among all other participating techniques applied in other tools.

Table 8. Results obtained in WEKA

Technique Applied	**Accuracy Rate**	**Error Rate**
Bayes Net	56.23	43.77
Naïve Bayes	53.04	46.96
J48	63.47	36.53
MLP	68.12	31.88
SMO	58.26	41.43
Logistic	68.69	31.31
LMT	66.37	33.63
S Pegasos	65.21	34.78
FT	70.72	29.27

Table 9. Results obtained in Tanagra

Technique Applied	**Accuracy Rate**	**Error Rate**
C4.5	79.71	20.29
ID3	60.57	39.43
SVM	61.16	38.84
kNN	67.54	32.46
Prototype NN	61.45	38.55
CRT	64.93	35.07
LDA	61.74	38.26

Table 10. Results obtained in Clementine

Technique Applied	Accuracy Rate	Error Rate
NN-RBFN	61.45	13.3
C5.0	75.36	24.64
C&RT	74.78	25.22
QUEST	62.61	37.39
CHAID	66.09	33.91
LDA	64.06	35.94
Logistic	63.19	36.81

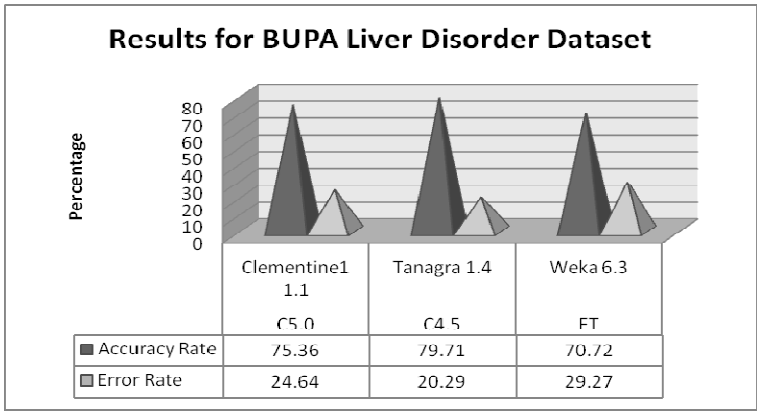

Figure 4. Comparison between best classification techniques applied for BUPA Liver Disorder Dataset

4.4. Results for classification techniques applied on StatLog Heart Disease dataset

Table 11, 12 and 13 show the results for classification techniques applied on StatLog Heart Disease dataset in WEKA, Tanagra and Clementine respectively. On the basis of comparison done over accuracy and error rates; the classification techniques with highest accuracy are obtained for this dataset in given different machine learning tools. Figure 5 shows comparison between the best classification techniques applied on this dataset. From Figure 5 it is observed that SVM classification technique applied in Tanagra for this dataset is best among all other participating techniques applied in other tools..

Table 11. Results obtained in WEKA

Technique Applied	Accuracy Rate	Error Rate
Bayes Net	81.11	18.89
Naïve Bayes	83.71	16.29
J48	76.67	23.33
MLP	78.14	21.86
SMO	82.96	17.04
Logistic	83.70	16.30
LMT	83.33	16.67
S Pegasos	82.96	17.04
FT	82.96	17.04

Table 12. Results obtained in Tanagra

Technique Applied	Accuracy Rate	Error Rate
C4.5	88.88	11.12
ID3	76.29	23.71
SVM	99.25	0.75
kNN	86.66	13.34
Prototype NN	85.18	14.82
CRT	80.37	19.63
LDA	84.81	15.19

Table 13. Results obtained in Clementine

Technique Applied	Accuracy Rate	Error Rate
NN-RBFN	83.33	16.67
C5.0	90.37	9.63
C&RT	90.74	9.26
QUEST	82.96	17.04
CHAID	88.15	11.85
LDA	85.19	14.81
Logistic	85.56	14.44

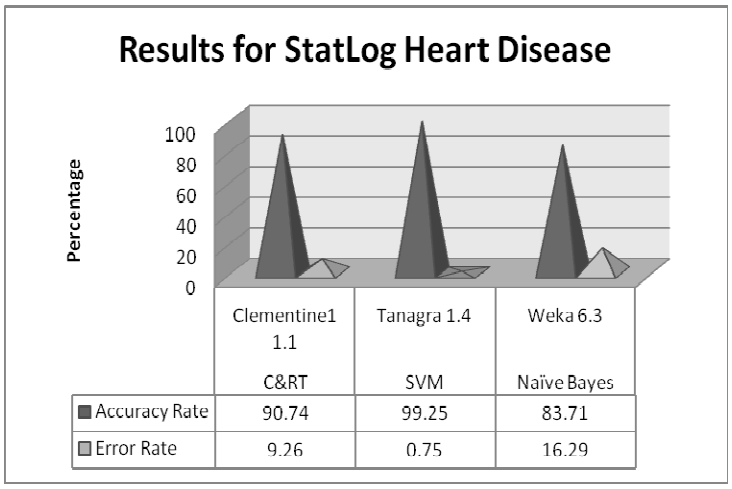

Figure 5. Comparison between best classification techniques applied for StatLog Heart Disease Dataset

5. CONCLUSIONS

The experimental results have shown that different classification techniques behave differently on different datasets depending on the nature of their attributes and size. The classification technique which has shown the highest accuracy rate and lowest error rate over a dataset has been selected as the best classification technique for that dataset. Table 14 shows a summary of results in terms of the best classification techniques' accuracy and error rate on the given datasets.

Table 14. Results showing the best classification techniques over given datasets

Database Used	Technique Applied	Accuracy Rate	Error Rate	Tool Used
Indian Diabetes	SVM	96.74	3.18	Tanagra
Wisconsin Breast Cancer	NN-RBFN	97.28	2.72	Clementine
BUPA Liver Disorder	C4.5	79.71	20.28	Tanagra
StatLog heart disease	SVM	99.25	0.75	Tanagra

From the results obtained after applying different classification techniques on given datasets SVM showed the most promising results for PIMA Indian Diabetes dataset and StatLog Heart Disease dataset with 96.74% and 99.25% accuracy rate respectively and C4.5 decision tree for BUPA Liver-disorders dataset with an accuracy rate of 79.71% whereas for Wisconsin Breast Cancer dataset Bayes Net, SVM, kNN and RBF-NN all shown the almost similar results with high accuracy rate and the highest accuracy rate achieved is 97.28% .

By knowing the best classification technique over a dataset a set of rules can be generated for that particular dataset and these rules will complement the healthcare researchers' study for intelligent decision making. At last for future work it is suggested that more experiments can also be done on healthcare datasets using different parameters and techniques.

ACKNOWLEDGEMENTS

This research work is supported by the UGC Grant under Major Research Project Scheme (F. No. : 33-60/2007(SR) dated Feb. 28, 2008).

REFERENCES

[1] Banu Rahaman S. and ShashiM., "Sequential mining equips e-Health with knowledge for managing diabetes," *4th International Conference on New Trends in Information Science and Service Science (NISS),* 2010, pp.65-71.

[2] Frawley, W., Batheus, C., 1991. *Knowledge Discovery in Databases: An Overview*. In Piatetsky-Shapiro, G. and Frawley, W. (Eds.), Knowledge Discovery in Databases, MIT Press, Cambridge, MA, pp1-27.

[3] Han, J., Kamber, M., 2001. Data Mining: Concepts and Techniques, Morgan Kaufmann, San Fco., CA., USA.

[4] Witten, I. H., Frank, E. 2000. Data mining: Practical machine learning tools and techniques with Java implementations. Morgan Kaufmann, San Francisco, CA. USA, 371 pp.

[5] Hand D.J., Mannila H., Smyth P. 2001. *Principles of data mining*, MIT Press, Boston, MA.,USA.

[6] Zhou, Z. H., 2003. Three perspectives of data mining. *Artificial Intelligence,* Nº 143(1), pp. 139-146.

[7] Anunciacao Orlando, Gomes C. Bruno, Vinga Susana, Gaspar Jorge, Oliveira L. Arlindo and Rueff Jose, "A Data Mining approach for detection of high-risk Breast Cancer groups," *Advances in Soft Computing*, vol. 74, pp. 43-51, 2010.

[8] Sarvestan Soltani A. , Safavi A. A., Parandeh M. N. and Salehi M., "Predicting Breast Cancer Survivability using data mining techniques," *Software Technology and Engineering (ICSTE), 2nd International Conference*, 2010, vol.2, pp.227-231.

[9] Abdelaal Ahmed Mohamed Medhat and Farouq Wael Muhamed, "Using data mining for assessing diagnosis of breast cnacer," *in Proc. International multiconfrence on computer science and information Technology*, 2010, pp. 11-17.

[10] Gandhi Rajiv K., Karnan Marcus and Kannan S., "Classification rule construction using particle swarm optimization algorithm for breast cancer datasets," *Signal Acquisition and Processing. ICSAP, International Conference*, 2010, pp. 233 – 237.

[11] Manaswini Pradhan and Dr. Ranjit Kumar Sahu, "Predict the onset of diabetes disease using Artificial Neural Network (ANN)", *International Journal of Computer Science & Emerging Technologies (E-ISSN: 2044-6004),* pp.303 -311, vol. 2, iss. 2, 2011.

[12] Padmavati J., "A Comparative study on Breast Cancer Prediction Using RBF and MLP," *International Journal of Scientific & Engineering Research*, vol. 2, Jan. 2011.

[13] Kahramanli Humar and Allahverdi Novruz, "Mining Classification Rules for Liver Disorders", *International Journal of Mathematics and Computers in Simulation*, vol. 3, 2009.

[14] Suneetha N., Hari V. M. K. and Kumar V.S., "Modified Gini Index Classification: A Case Study of Heart Disease Dataset", *International Journal on Computer Science and Engineering,* issue 6, vol. 2, pp. 1959-1965, 2010.

[15] Elsayad A. M., "Diagnosis of Erythemato-Squamous Diseases using Ensemble of Data Mining Methods", *ICGST-BIME Journal*, issue 1, vol. 10, 2010.

[16] Ubeyli E., "Multiclass support vector machines for diagnosis of erythemato-squamous diseases". Expert Systems with Applications, 35(4):1733–1740, 2008.

[17] Lee Heui Chul, Seo Hak Seon and Choi Chul Sang, "Rule discovery using hierarchial classification structure with rough sets," *IFSA World Congress and 20th NAFIPS International Conference*, 2001, vol.1 , pp. 447-452.

[18] Jamarani S. M. h., Behnam H. and Rezairad G. A., "Multiwavelet Based Neural Network for Breast Cancer Diagnosis", *GVIP 05 Conference*, 2005, pp. 19-21.

[19] Lundin M., Lundin J., BurkeB.H.,Toikkanen S., Pylkkänen L. and Joensuu H. , "Artificial Neural Networks Applied to Survival Prediction in Breast Cancer", *Oncology International Journal for Cancer Resaerch and Treatment,* vol. 57, 1999.

[20] Street W.N., "A Neural Network Model for Prognostic Prediction", *Fifteenth International Conference on Machine Learning, Madison, Wisconsin, Morgan Kaufmann*, 1998.

[21] Chi C.L., Street W.H. and Wolberg W.H., "Application of Artifical Neural Network- based Survival Analysis on Two Breast Cancer Datasets", *Annual Symposium Proceedings / AMIA Symposium*, 2007.

[22] Choi J.P., Han T.H. and Park R.W., " A Hybrid Bayesian Network Model for Predicting Breast Cancer Prognosis", *J Korean Soc Med Inform*, 2009, pp. 49-57.

[23] Delen Dursun , Walker Glenn and Kadam Amit , "Predicting breast cancer survivability: a comparison of three data mining methods," *Artificial Intelligence in Medicine* ,vol. 34, pp. 113-127 , June 2005.

[24] Bellaachia Abdelghani and Erhan Guven, "Predicting Breast Cancer Survivability using Data Mining Techniques," *Ninth Workshop on Mining Scientific and Engineering Datasets in conjunction with the Sixth SIAM International Conference on Data Mining,*" 2006.

[25] Osmar R. Zaïane, *Principles of Knowledge Discovery in Databases*. [Online]. Available: *webdocs.cs.ualberta.ca/~zaiane/courses/cmput690/notes/Chapter1/ch1.pdf.*

[26] UCI Machine Learning Repository.[Online]. Available: http://archive.ics.uci.edu/ml/datasets.html

[27] Zupan Blaz and Demsar Janez,"Open-Source Tools for Data Mining," *Clin Lab Med*, pp. 37–54, 2008.

[28] Tanagra - A free data mining software for teaching and research. [Online]. Available: http://eric.univ-lyon2.fr/~ricco/tanagra/en/tanagra.html.

[29] SPSS Clementine. [Online]. Available: http://www.the-data mine.com/bin/view/Software/ClementineSoftware.

Authors

Shelly Gupta received her B.Tech. degree in Information Technology from Guru Jambheshwar University of Science & Technology, Hisar (Haryana), India in 2009 and M.Tech. degree in Computer Science from Banasthali University, Banasthali (Rajasthan), India in 2011. Her interests are in Data Mining and Software Engineering.

Dharminder Kumar received his Ph.D in the area of Computer Science in Computer Networks. He is recipient of Gold Medal at his Master's degree. Presently he is heading the Faculty of Engineering and Technology as Dean since 2008, Dean of Colleges since April, 2011 and also the Chairman of two departments i.e. Printing Technology & Bio-Medical Engineering. He has been the Chairman of the Department of Computer Science & Engineering for more than six years. He has guided eight students in Ph.D, about 40 at M.Tech. level, more than 85 at MCA level and 6 students are currently pursuing their Ph.D under his supervision in Computer Science and Engineering. He has published more than 75 research papers in Journals/Conferences/Seminars at National /International levels. He is having more than 22 years of teaching and research experience. His area of interest includes ICT, Data Mining and Computer & Communication Networks. He is member of various professional/regulatory bodies (at State/National/Universities levels) such as UGC, AICTE, CSI, NMEICT, HSCS etc. He is now working as Professor with department of Computer Science & Engineering, Guru Jambheshwar University of Science & Technology, Hisar, Haryana, INDIA.

Anand Sharma received his Master's degree in Computer and Business Administration with specialization in IT and Systems. He is currently a Project Fellow, Department of Computer Science and Engineering at Guru Jambheshwar University of Science & Technology, Hisar (Haryana), India. He is having more than 10 years of professional and research experience in the field of ERP Implementation, Wireless Networks and Healthcare Domain. His area of interest includes Data Mining, Network Security, Software Engineering and Open Source Technologies.