INTERNATIONAL RESEARCH JOURNAL IN ADVANCED ENGINEERING AND TECHNOLOGY (IRJAET) E - ISSN: 2454-4752 P - ISSN : 2454-4744
VOL 2 ISSUE 5 (2016) PAGES 1169 - 1177
RECEIVED : 30/08/2016. PUBLISHED : 04/09/2016

September 4, 2016

MULTI DISEASE PREDICTION AND TREATMENT ANALYSIS USING DATA MINING TECHNIQUES

[1]Vaishnavi G, [2] P.Hariharan
[1]M.Phil Scholar, [2]Assistant Professor
Department of Computer Science & Applications, Adhiparasakthi College of Arts & Science
G.B.Nagar, Kalavai 632506, Vellore District, Tamilnadu.

ABSTRACT

Disease prediction and diagnosis is one of the complex applications where data mining tools and techniques are used to providing successful results because of significant improvements in technology. This research identifies gaps in the research on disease prediction, diagnosis and treatment and it also proposes a model to systematically close those gaps. Data mining have great potential for healthcare industry to enable health systems to systematically use data and identify the efficiency and improve care with reduce cost. The data mining techniques to Multi disease treatment it can provide reliable performance. So the system can be effective in reducing the death toll. The healthcare industry collects huge amounts of healthcare data which, unfortunately are not “mined” to discover hidden information for effective decision making. Advanced data mining techniques can be helpful and can provide an efficient remedy to these kinds of problems. This proposed work has developed a prototype for the Multi Sickness Prediction System (MSPS) using data mining techniques by to compute the chance of prevalence of explicit unwellness from medical knowledge by using k-means, Large Memory Storage And Retrieval (LAMSTAR) and Medical diagnosis methodology. The system uses service oriented architecture (SOA) whereby the system elements of diagnosis, data portal and alternative miscellaneous services are provided. This reduces the multiple diseases showing the similar symptoms problem and it will increase the accuracy of such diagnosis. This proposed system will provide some reliable decision finding the disease for healthcare support. Results show that each technique has its unique strength in realizing the objectives of the defined mining goals from the analysis.

Keywords: Fuzzy Neural Networks(FNN), Service Oriented Architecture (SOA),k-Means.

1. INTRODUCTION

Discovery of new information in terms of patterns or rules from large amounts of data is based on the machine learning technique. Disease prediction plays an important role in data mining. Diagnosis of a disease requires the performance of a number of tests on the patient. However, use of data mining techniques, can reduce the number of tests. This reduced test set plays an important role in time and performance. Disease data mining is important because it allows doctors to see which features or attributes are more important for diagnosis such as symptoms, lab test, etc. This will help the doctors diagnose disease more efficiently.

The research presented in this thesis is intended to address the challenge of improving the prediction model to predict the multi disease for patients and providing timely response in predicting the diseases.

The important research functions are,

- Describe about various data mining techniques can be used in health care industry and to identify their performance in prediction.

- Describe about classification techniques help in developing the prediction model so as to predict accurately the risk of heart disease among diabetic patients.

Application of data mining in analyzing the medical data is a good method for investigating the existing relationships between variables. Nowadays, data stored in medical databases are growing in an increasingly rapid rate. It has been widely recognized that medical data analysis can lead to an enhancement of health care. The primary objective of the research work is the effective development of prediction model using various classification techniques to predict the multi disease and performance in prediction. It also shows that data mining can be applied to the medical databases to predict or classify the data with reasonable accuracy.

2. BACK GROUND AND LITERATURE SURVEY

The medical data, for the purpose of extracting knowledge can be obtained from multiple sources. The data sources include hospital's in-house patient data, out-patient data, clinical laboratories and other researchers in the clinical domain. As the data sources are different the data that are combined to form the base data for analysis could be heterogeneous in nature. The data that is used for mining purpose could be obtained from heterogeneous sources and hence it is essential to pre-process the data before using it to extract the knowledge. The patient data collected at different hospitals and clinical laboratories might be heterogeneous in nature. The data need to be pre-processed due to the following reasons: name of the attributes, domain of attributes and the number of attributes used to store the patient records might be different in the databases at different hospitals and clinical laboratories; all the features (attributes) recorded in the database might not be required for extracting knowledge that is needed to predict the presence or severity of a disease. Some of the patient records might have entries for some of the attributes to be missing.

Shomona Gracia Jacob et al [1]. Compare the error rates and related performance measures produced by the various classification algorithms on the Wisconsin breast tissue dataset and the effect of feature selection algorithms on improving the accuracy of classification of carcinoma in the breast tissue dataset. The existence of other tissue features like Fibro- adenoma, Mastopathy that indicate a higher risk of developing cancer in future is also classified. This research highlights the significance of classification in data mining and knowledge discovery. In research investigate the performance of various data mining classification algorithms viz. Rnd Tree, Quinlan decision tree algorithm (C4.5), K-Nearest Neighbor algorithm etc., on a large dataset from the Wisconsin Breast tissue dataset that comprises of 11 attributes and 106 instances [3]. The results of this study indicate the level of accuracy and other performance measures of the algorithms in detecting the presence of breast cancer and the associated breast tissue conditions that increase the risk of developing cancer in future. Their work classifies medical images and is not applicable to textual medical data. The breast tissue dataset has not provided feature selection as a pre-classification condition.

Shweta Kharya [2] discussed some of effective techniques that can be used for breast cancer classification. Among the various data mining classifiers and soft computing approaches, Decision tree is found to be best predictor with 93.62% Accuracy on benchmark dataset (UCI machine learning dataset) and also on SEER dataset. The predictor can be used to design a web based application to accept the predictor variables and automated system Decision Tree based prediction can be implemented in remote areas like rural regions or country sides, to imitate like human diagnostic expertise for prediction of ailment. The Bayesian network is also found to be a popular technique in medical prediction Particular it has been successfully utilized for Brest cancer prognosis and diagnosis.

M.Akhil jabbar et al [3] developed as generalization of mathematical models of human cognition or neural biology. Two models of neural network have been developed. These models are feed forward neural networks trained with back propagation and radial basis function neural networks. The two models are three layer networks which are made up of an input that is connected with the hidden layer with aid of connection weight. The hidden layer is also connected with the output layer also with the aid of connection weight. The different between the two networks is the kind of an activation function that is present in the neurons of the hidden layer and the neurons in the output layer of the two networks.

3. NEURO FUZZY APPROACHES FOR PREDICTING MULTIPLE DISEASES

In this work a system that predicts the risk of disease was implemented by tailoring the proposed model. The techniques PCA, Fuzzy C-Means Clustering (FCM) and Neuro Fuzzy Inference System were used in this work. The pre-mining subsystem was implemented using PCA and FCM for reducing the dimensionality of input data from eighteen to two. The mining and validation subsystems were integrated and implemented as Neuro-Fuzzy architecture that was tuned manually to extract classification knowledge for predicting the survival of hepatitis. Manual tuning was performed for identifying the optimal premise and consequent parameter values.

The disease symptoms data used for validating the system developed as part of this work is the one provided at the UCI Machine learning repository. In this work gathered some disease symptoms and lab test dataset for evaluation. The system uses service oriented architecture (SOA) whereby the system elements of diagnosis, data portal and alternative miscellaneous services are provided. This reduces the multiple diseases showing the similar symptoms problem and it will increase the accuracy of such diagnosis. Medical decision support system refers to both the process of attempting to determine or identify possible diseases or disorder and the opinion reached by this process.

3.1 PRINCIPAL COMPONENT ANALYSIS TECHNIQUE

The PCA technique is used to reduce the input data features of the symptoms from dataset. PCA transforms a number of (possibly) correlated variables into a (smaller) number of uncorrelated variables called principal components. Since symptoms data has patterns that are hard to find, and as graphical representation is not possible, PCA is a powerful tool for analyzing and performing dimensionality reduction on the disease data.

Input: Disease Symptoms

Process Logic:

Step 1: The input data is normalized so that the domain of the data does not affect the characteristics of the underlying information.

$$\mathbf{V'} = \frac{V-A'}{\sigma A} \ldots(3.1)$$

Where, A represent the attribute, A is the mean value for the attribute, A' represent the standard deviation of the attribute, v and ' v represent the original and normalized values of the attribute respectively.

Step 2: Split the data into two matrices one containing the attributes.

Step3: Calculate the covariance matrix (A) which captures the correlations between all possible pairs of measurement.

$$\mathbf{Cov\ (X) = E\ [\ X - (\ E\ [X]\ )\ (X - E[X])^{T}\ ]\ ....}(3.2)$$

Step 4: The Eigenvectors (V) and Eigen values (D) are computed for the covariance matrix (A), such that equation is satisfied.

$$\mathbf{A*V = V*D\ ...(3.3)}$$

Step 5: The Eigen values are ordered from highest to lowest. The highest Eigenvalue is considered as the principle component of the data set. The feature vector is a matrix containing the values corresponding to the principle component.

Step 6: The feature vector is multiplied with the disease symptoms dataset.

Output: Matrix (B) that represents the Data with features

4. DISEASES PREDICTION BY USING K- MEANS ALGORITHM

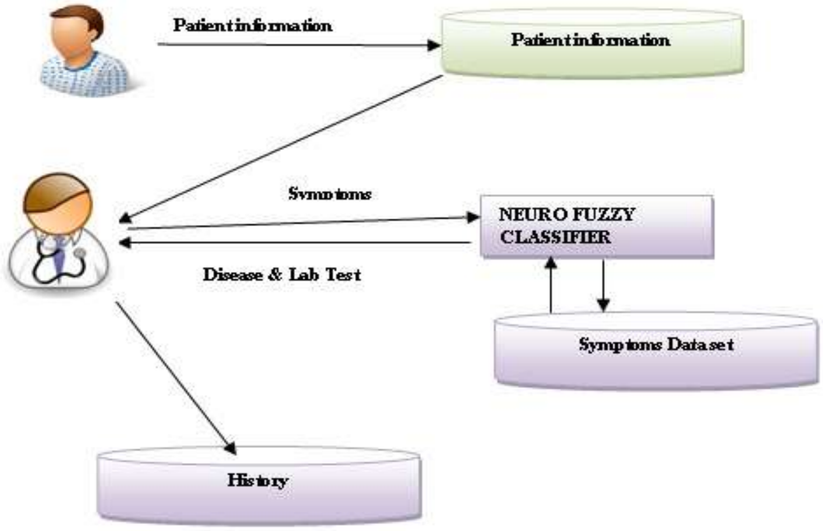


Fig.1. Architectural Model.

First the doctor retrieves the symptoms from the patient record database. After retrieving the symptoms, the doctor identify whether any symptom related diseases contains in the Diseases/Symptoms database. Here the pattern matching service is activated. If any diseases match with Symptoms means list out all the possible matched symptoms and presents the result to the doctors. If the doctors not satisfied with results, compare to recent history and recent trend service must be activated. This service makes use of the Diseases/symptoms database and Patient Record database and the result obtained from pattern matching service to get results. After comparing the diseases to the recent history, cluster the shortlisted diseases. This list is used to compute the probability of each occurrence of particular diseases from the medical data. The probability may be computed based on the distance vector. The highest priority cluster produces the accurate result. Finally, to avoid the vagueness in decisions, the doctor use differential diagnosis and recent diagnosis features use Diseases/symptoms database and Patient record database and result acquired from recent trend services to gain the results. Since the large medical data, using simple client server architecture would not produce the effective aforesaid services and would increase the response time of the system.

Finally conclude that SOA was well suited to apply this system because it improve the delivery of important information and sharing of data across the community of healthcare professionals more practical in cost, security and risk deployment. In various existing EHRs, SOA is more essential for data providers to this system, are already using this very successful and efficient architecture. The system enforced as various services in the existing SOA, result in easy implementation, integration and scalability with existing EHRs.SOA also handles the related issues to data security and patient confidentiality.

4.1 SYMPTOMS COMPARING USING ITERATIVE SEARCH

In this phase symptom matching using iterative search utilize data that is stored. The first step of the algorithm involves selecting the symptoms shown by the patient. The algorithm gives the list of all possible diseases ranked according to the number of symptoms matched in the database. The list is generated after input of every symptom. After the first iteration for the second iteration the next list of symptoms will be shortlisted according to the disease list that was obtained in the previous iteration .The new symptom list will contain symptoms of only those diseases that were obtained in the previous list, if *headache*, *fever* and *pain* in the *sinuses* are entered, then the weights W15, W16 and W19 will be considered. Next all the weights will be added and compared to all subclasses C1, C2, C3 and C4 is most likely the answer depending on its weight. Finally all the diseases in class C4 are considered and if sinusitis (D4) weight is closer to the sum of all the input symptoms weights, then it is possible diagnosis.

Disease	**Symptoms and weight**	**Class weight**
AIDS	Fatigue, swollen lymph nodes, ulcers in the mouth or on the genitals, muscle aches and joint pain, nausea and vomiting, night sweats, body rash, fever	A1 Cd4 Cells
Dengue Fever	high fever , Severe headaches Pain behind the eyes, Severe joint and muscle pain, Fatigue ,Nausea ,Vomiting, Skin rash	D1

Table.1. Sample Data Set: Iterative Pattern Search.

4.2 DISEASES SHORTLISTED USING LAMSTAR NETWORK

In this phase the correct disease shortlisted by the doctor is obtained who confirms it by taking the necessary tests. The final report is then mined to obtain the correct symptoms. The correct symptom thus obtained is then compared with the original symptoms entered. This information is now fed to the LAMSTAR Network for assigning weights. If any stored pattern matches the input subs word within a present tolerance.

The system updates weights according to the following procedure:

$$W_{i,m}(t+1) = W_{i,m}(t) + \acute{\alpha}_i (X_i(t) - W_{i,m}(t)), \text{for } m:\text{€}min$$

Where,

Wi, m (t+1) = Modified weights in module I for neuron m;
Ά i = Learning coefficient for module I;
€min = Minimum error of all weights vectors Wi in module.

The query input which contains patients' examination information was pre-processed. The conditions of the attributes were evaluated and verified whether all of the conditions in a rule was satisfied. If all the conditions in a rule had been satisfied for the given data then corresponding rule would be fired. The rules that were fired had been processed to derive the output. The predicted output which indicates the presence or absence of the disease is displayed as output to the user of the system.

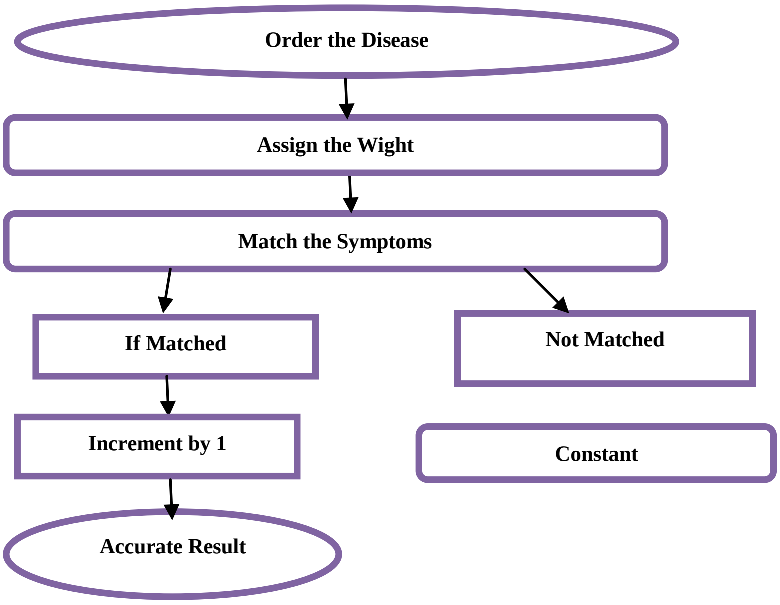


Fig.2. Diseases Shortlisted.

5. K – MEANS CLUSTERINGALGORITHM AND LAMSTAR NETWORKS

The main motivation in this research is based on the assumption that the instance with similar attribute values is more likely to have similar class label. Similarity is measured based on Euclidean distance. Therefore, the misclassified instances after clustering are deleted and correctly classified instances are considered for further classification using decision-tree classifier. Many researchers have used clustering method on unlabelled data to assign class labels. This approach tried clustering by k-mediod algorithm but the misclassification rate was 50%. This approach considered the result of k-means because the misclassification rate was less. The k-means algorithm takes the input parameter, k, and partitions a set of N points into k clusters, so that the resulting intracluster similarity is high but the intercluster similarity is low. Clustering is the process of grouping same elements. This technique may be used as a preprocessing step before feeding the data to the classifying model. The attribute values need to be normalized before clustering to avoid high value attributes dominating the low value attributes.

Final Prediction Algorithm

```
Step 1: Configure the dataset 1-m-n
        Where 1=no of input , m – no of hidden inputs, n - no of output values.
Step 2: The no of weights are calculated based on storing table W.
Step3: Assume no of digits in weight from Dataset D.
Step4: Choose symptoms Si form population Dataset pi,
Step 5: for each weighted symptoms
            {
            Extract weight W-I;
            Keep the weight for each input for train Dataset.
            Calculate the fitness
            Value Fi for each of symptoms from population dataset;
            }
Step 6 : apply LAMSTAR
Step 7: output Fi for each Si;
```

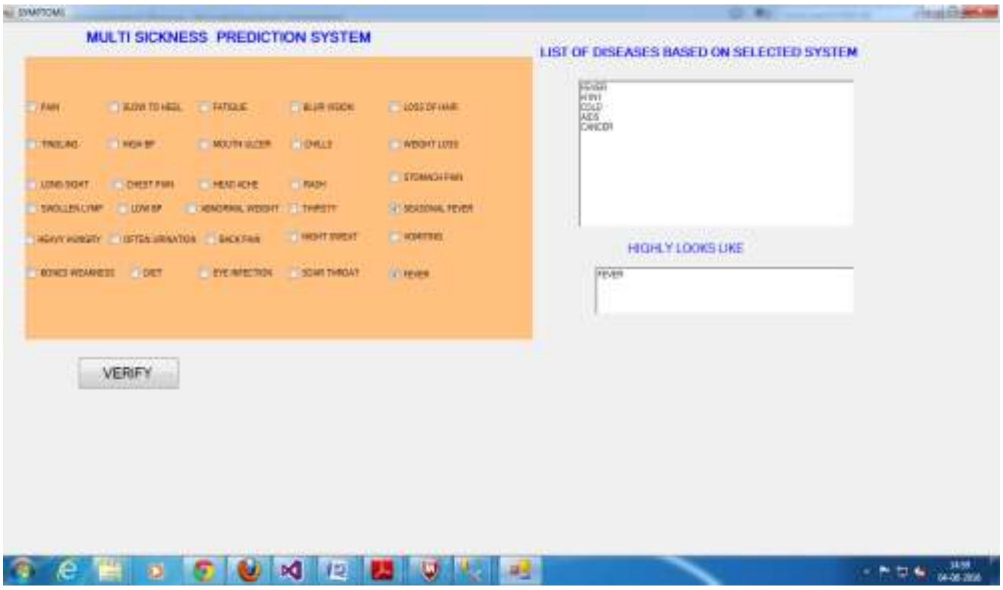
5.1 ADVANTAGES OF PROPOSED ALGORITHM

- The main advantage of this algorithm is simplicity and its speed which allows running large datasets.
- K-Means may be faster than hierarchical clustering (if K is small).
- K-Means may produce tighter clusters than hierarchical clustering, especially if the clusters are globular.

6 EXPERIMENTAL RESULTS AND DISCUSSIONS

In the first experiment, in the current work Multi Sickness Prediction System (MSPS) has been used as a tool and the results of According to Neuro Fuzzy Classifier (NCN) and K-means clustering. The proposed According to Neuro Fuzzy Classifier (NCN) and k – means algorithm was able to classify 94% of the input instances correctly. It exhibited a precision of 91% on an average, recall of 86% on an average, and F-measure of 91.2% on an average. The results show clearly that the proposed method performs well compared to other similar methods in the literature, considering the fact that the attributes taken for analysis are not direct indicators of disease symptoms. The rules extracted by the mining subsystem were validated with the help of an expert. The common approach of “train and test” was used for rule validation. The model to predict the class was built with training samples and validated with an independent test sample. Experts were consulted to validate the rules generated by the system and the validated rules were stored in the knowledge base.

Inference system used the pattern discovered using statistical analysis. This inference system prepared a confidence measure which gave the probability of correct suggestions by examining the values with the inference calculation. This was done by calculating the residual value. Neuro Fuzzy Classifier was used to calculate the membership probability of a given patient record. Conventional approaches of pattern classification involve clustering training samples and associating clusters to given categories. The complexity and limitations of previous mechanisms are largely due to the lacking of an effective way of defining the boundaries among clusters.



SYMPTOMS

MULTI SICKNESS PREDICTION SYSTEM

□ PAIN □ SLOW TO HEAL □ FATIGUE □ BLUR VISION □ LOSS OF HAIR
□ TINGLING □ HIGH BP □ MOUTH ULCER □ CHILLS □ WEIGHT LOSS
□ LONG SIGHT □ CHEST PAIN □ HEAD ACHE □ RASH □ STOMACH PAIN
□ SWOLLEN LYMP □ LOW BP □ ABNORMAL WEIGHT □ THIRSTY ☑ SEASONAL FEVER
□ HEAVY HUNGRY □ OFTEN URINATION □ BACK PAIN □ NIGHT SWEAT □ VOMITING
□ BONES WEAKNESS □ DIET □ EYE INFECTION □ SORE THROAT ☑ FEVER

VERIFY

LIST OF DISEASES BASED ON SELECTED SYSTEM

[illegible]
H1N1
COLD
AIDS
CANCER

HIGHLY LOOKS LIKE

FEVER

Fig.3. Symptoms based the disease Fever.

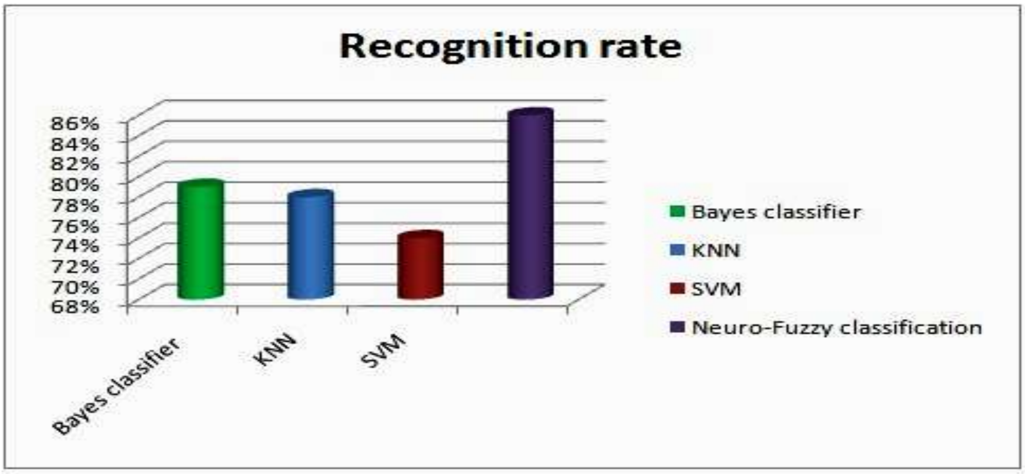


Fig.4. Recognition Rates of Classification Algorithms.

In general the Bayes classifier and the KNN classifier could not handle the prostate data as good as the Neuro-Fuzzy classification systems. This supports earlier findings using artificial neural networks. This effect was not observed with the Iris data which can be contributed to the different statistical properties of the two data sets. With the Iris data every system had problems to identify the same four outliers which limited the achievable recognition rate to 97.33 %.

7. CONCLUSION AND FUTURE WORK

Health care relevant data are enormous in nature and they arrive from various birthplaces all of them not wholly relevant in structure or quality. These days, the performance of knowledge, observation of various specialists and medicinal screening data of patients grouped in a database during the analysis process, has been widely accepted. In this Research thesis is presented an efficient approach for Multi disease prediction based on the patient symptoms warehouses for the efficient prediction of diseases. This research uses, LAMSTAR Network, K-Means algorithm and Neuro Fuzzy Classifier to assist the doctors to perform differential diagnosis along with the possible implementation using SOA technique. By using these techniques, it improves the overall speed and increase the accuracy of algorithm. Especially in large datasets, LAMSTAR network gave faster and better result. It reduces the effects of misdiagnosis, especially practioners and students can also easily identify the diseases. The results obtained for the prediction of the multiple disease show that the system can classify the positive samples with better accuracy as compared to classification of negative samples classification. It can be observed that the classification accuracy of the neuro-fuzzy approach for hepatitis data is relatively better when compared to the other approaches that use neural networks with back propagation training. It will also help the medical fraternity in the long run by helping them in getting accurate diagnosis and sharing of medical practices which will facilitate faster research and save many lives. In future work, this research will extend to conduct experiments on large real time health datasets to predict the diseases and compare the performance of this algorithm with other related data mining algorithms.

REFERENCES

[1] Shomona Gracia Jacob and R.Geetha Ramani,” Discovery of Knowledge Patterns in Clinical Data through Data Mining Algorithms: Multi-class Categorization of Breast Tissue Data”, International Journal of Computer Applications (0975 – 8887) Volume 32– No.7, October 2011.

[2]. Shweta Kharya,” Using Data Mining Techniques For Diagnosis And Prognosis Of Cancer Disease”, International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.2, No.2, April 2012.

[3]. M.Akhil jabbar,B.L Deekshatulu and Priti Chandra,” Classification of Heart Disease Using K-Nearest Neighbor and Genetic Algorithm”, International Conference on Computational Intelligence: Modeling Techniques and Applications (CIMTA) 2013.