International Journal of Applied Engineering Research ISSN 0973-4562 Volume 10, Number 14 (2015) pp 34174-34179

A Comparative Analysis Of Data Mining Techniques Used In Health Care Systems

[1]Mr. S. Mohandoss, [2]Dr. V. SaiShanmuga Raja, [3]Dr. SP. Rajagopalan

[1]Research Scholar, Department of Computer Applications, smohandossmca@gmail.com
[2]Associate Professor, Department of Computer Science and Engineering, vssraja2001@gmail.com
[3]Professor Emeritus, Department of Computer Applications,
Dr. MGR Educational & Research Institute University, Chennai, TamilNadu, India.

ABSTRACT
Health care systems proved health services to target people with health needs with the help of organization of people, resources and institutions. Health care systems have an enormous amount of health related data from which valuable data can be discovered using data mining techniques. In this paper we present a comparative study of data mining techniques such as: Decision tree, Naïve Bayes and Neural network to be used in health care systems for knowledge discovery. Data mining techniques may benefit health industry in different ways. For example data mining techniques can help health insurers to detect fraud, physicians to give effective treatment and best practices, health care organizations to make the customer relationship stronger. The health care transactions produce a large amount data which are complex and voluminous which cannot be processed and analyzed by traditional methods. Identifying appropriate Data Mining technique will transform these voluminous data into useful information for decision making. By this study we aim to analyze the uniqueness of health care data mining, Identifying and selecting the most efficient data mining technique to be used in the modern Health care system.

Key words: Health care systems, Decision tree, Naïve Bayes and Neural network.

1. INTRODUCTION

In medical and health industry, many applications using artificial intelligence have helped in the development of knowledge based systems and expert systems. Health care systems are greatly benefited to many different people involved in Health Insurance, Physical treatment, Health organizations etc. The discovery of new knowledge by mining large health record databases is very important to make effective use of the databases and enhancing health management tasks. They are proved to be very essential for patients as well as for health experts in taking right decisions. A Health care decision support system is any computer application that helps experts in making health care decisions. The main objective of this paper is to analyze the recent trends and Methodologies of Health care systems, to analyze and use the electronic data in health care systems.

Data mining involves six common classes of tasks:

- ***Anomaly detection*** – The identification of unusual data records, that might be interesting or data errors that require further investigation. (Outlier/ change/ deviation detection)
- ***Association rule learning*** - Finding frequent patterns, associations, correlations, or causal structures among sets of items or objects in transactional databases, relational databases, and other information repositories.
- ***Clustering*** – is the task of discovering groups and structures in the data that are in some way or another "similar", without using known structures in the data.
- ***Classification*** - Classification is a data mining technique used to predict group membership for data instances. For example, we may wish to use classification to predict whether the weather on a particular day will be sunny, rainy or cloudy. Popular classification techniques include decision trees and neural networks.

Data mining technology provides a user-oriented approach to the novel and hidden patterns in the data. The discovered knowledge can be used by the healthcare administrators to improve the quality of service. The discovered knowledge can also be used by the medical practitioners to reduce the number of adverse drug effect, to suggest less expensive therapeutically equivalent alternatives. Following are some of the important areas of interests where data mining techniques can be of tremendous use in health care management [1].

1. Fraud detection in Health Insurance
2. Exclusive Health Information system
3. Forecasting Health expenses
4. Prediction of patient's future health using history
5. Public Health Informatics
6. Health care e-governance

2. DATA MINING TECHNIQUES IN HEALTH CARE SYSTEM

Data mining is an essential step of knowledge discovery. In recent years it has attracted great deal of interest in Information industry [2, 3]. Knowledge discovery process consists of an iterative sequence of data cleaning, data integration, data selection, data mining pattern recognition and knowledge presentation. In particulars, data mining may

accomplish class description, association, classification, clustering, prediction and time series analysis. Data mining in contrast to traditional data analysis is discovery driven. Data mining is a young interdisciplinary field closely connected to data warehousing, statistics, machine learning, neural networks and inductive logic programming.
Data mining technique is used in Knowledge Data Discovery (KDD). KDD involves in data cleaning, Data Integration, Data Selection, Data Transformation, Data mining, Pattern Evaluation and Data Presentation.

2.1 Decision Tree
It is a knowledge representation structure consisting of nodes and branches organized in the form of a tree such that, every internal non-leaf node is labeled with values of the attributes. The branches coming out from an internal node are labeled with values of the attributes in that node. Every node is labeled with a class (a value of the goal attribute).

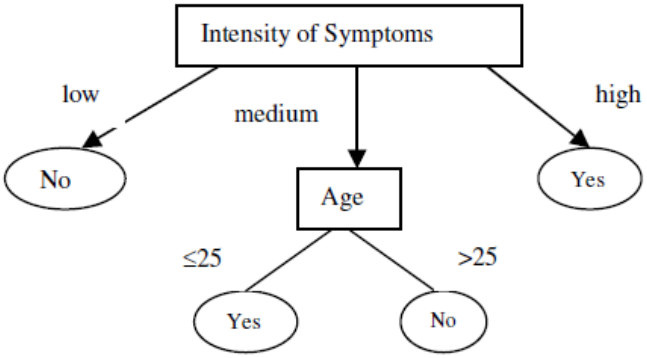

Fig. 1 A decision for the data in Table 1.

Treebased models which include classification and regression trees, are the common implementation of induction modeling [5]. Decision tree models are best suited for data mining. There are many Decision tree algorithms such as CART, ID3, C4.5 (a later version ID3 algorithm), SLIQ, and SPRINT [4]. A Decision tree is built of nodes which specify conditional attributes – symptoms X={x1,x2,.....xk},branches which show the values of B (ih) i.e. the *h*-th range for *i*-th symptom and leaves which present decisions Y={y1,y2.....yk} and their binary values Zdk={0,1} . A sample decision tree is presented in the Fig.1.

Table 1. Data Set used build a Decision Tree

Age	Gender	Intensity of symptoms	Disease (goal)
25	Male	medium	yes
32	Male	high	yes
24	Female	medium	yes
44	Female	high	yes
30	Female	low	no
21	Male	low	no
18	Female	low	no
34	Male	medium	no
55	Male	medium	no

2.2 Multilayer Perceptron neural network (MLP-NN)
MLP-NN is a collection of neuron –like processing units with weight connections between the units. These models mimic the human brain and learn the patterns of a data set in order to make predictions. Artificial Neural Networks (ANN) are analytical techniques modeled after the (hypothesized) processes of learning in the cognitive system and the neurological functions of the brain and capable of predicting new observations from previous observations after executing a process called learning from existing data[5]. Neural networks or artificial neural networks are also called connectionist system, parallel distributed systems or adaptive systems because they are composed by a series of interconnected processing elements that operate in parallel as shown in Fig. 2.
The ANN is a mathematical model which inspired from the structure and functions of the neurons in the human brain [6]. A Neural Network consists of number of neurons connected through weights which process information as a response to external stimuli. Stimuli are transmitted from one processing element to another via synapses or interconnection, which can be excitatory or inhibitory. If the input to neuron is excitatory, it is more likely that this neuron connected to it. Neural networks are good for clustering, sequencing and predicting patterns but their drawback is that they do not explain how they have reached to a particular conclusion.

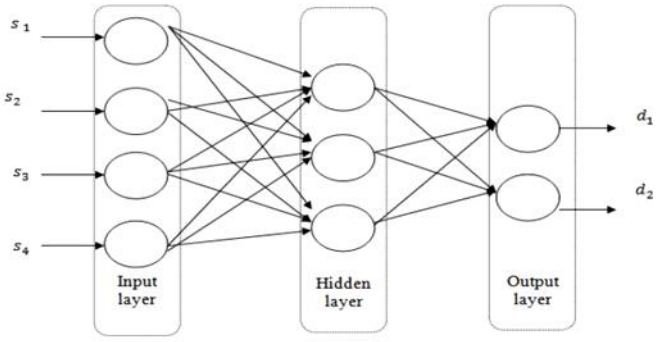

Fig. 2 MLP Neural Network for medical diagnosis

Advantages & disadvantages
The advantages of ANN include the elimination of needing to program the systems and providing input from experts. The ANN can process incomplete data by making educated guesses about missing data and improves with every use due to its adaptive system learning. Additionally, NN systems do not require large databases to store outcome data with its associated probabilities. A neural network can perform tasks that a linear program cannot. When an element of the neural network fails, it can continue without any problem by their parallel nature some of the disadvantages are that the training process may be time consuming leading users to not make use of the systems effectively. The NN systems derive their own formulas for weighting and combining data based on the statistical recognition patterns over time which may be difficult to interpret and doubt the system's reliability [7]. The neural network needs training to operate. The architecture of a neural network is different from the architecture of

microprocessors therefore needs to be emulated. Examples include the diagnosis of appendicitis, back pain, myocardial infarction, psychiatric emergencies and skin disorders. The NN's diagnostic predictions of pulmonary embolisms were in some cases even better than physician's predictions.

2.3 Naive Bayes

Naive Bayes or Bayes' Rule is the basis for many machine-learning and data mining methods [8]. The rule (algorithm) is used to create models with predictive capabilities. It provides new ways of exploring and understanding data. It learns from the "evidence" by calculating the correlation between the target (i.e., dependent) and other (i.e., independent) variables.

3. EVALUATION OF THE DATA MINING TECHNIQUES IN WEKA

It is very important to evaluate the Data Mining Techniques used in Health Care Systems. It is important to know what part of cases was classified correctly. The above said data Mining techniques are evaluated in WEKA.

Performance Evaluation Measures

Evaluation of data mining algorithms can be done in many ways. The following measures were used in this paper to evaluate the performance of the data mining techniques.

Mean absolute error - averages the magnitude of individual errors neglecting their sign (takes the absolute values to diminish the negative effect of outliers).

$$\frac{|p_1 - a_1| + \ldots + |p_n - a_n|}{n}$$

Root mean squared error - easy to calculate error commonly used in various mathematical computations.

$$\sqrt{\frac{(p_1 - a_1)^2 + \ldots + (p_n - a_n)^2}{n}}$$

False Positive Rate - $FPR = \frac{FP}{TN + FP}$

False Negative Rate - $FNR = \frac{FN}{TP + FN}$

True positive rate - $TPR = \frac{TP}{TP + TN}$

True negative rate - $TNR = \frac{TN}{TN + FP}$

Accuracy - $\frac{(TP + TN)}{(TP + FP + FN + TN)}$

ROC – Curves - The ROC analysis which stands for Receiver Operating Characteristics are used especially in medical classification where the distinction between the classification of healthy patient as ill and reverse needs to be done [8]. Using WEKA we can graphically evaluate data mining algorithms performance [8] by selecting WEKA Classifier Visualize option. These graphs can be used to show a ROC (Receiver Operating Characteristics) curve. The ROC curve is the ratio of False Positive Rate (axis *x*) and True Positive Rate (axis *y*). The area under the ROC curve (AUC) is a metric of algorithms performance.

We have made use of the Heart disease database to test the algorithms. A total of 303 records with 14 medical attributes (factors) were obtained from the Cleveland Heart Disease database [9]. The attributes are listed in Table 2. The records in the set belong to one of five angiographic disease statuses. The status of the disease takes values 0, 1, 2, 3, 4, which stand for how advanced a disease is. The higher number the more advanced heart failure. The value 0 indicates absence of the disease.

3.1 C4.5 Algorithm

The C4.5 algorithm in WEKA which is the new version of decision tree algorithm tested with the heart disease database. It generated a pruned tree consisting of 57 leaves that corresponded to 57 rules. The tree size was 92. The distribution of values is shown in Fig 3, the attribute *oldpeak*, decreases with the increase of the values. The distribution of values of the attribute *chestpaintype*, increases with the increase of the values. The distribution of values of the attributes: *age*, *trestbps*, *cholesteral* and *maxheartrate* is of a bell-shape. The distribution of the decisional attribute *diagnosis* also decreases with the increase of the values. The attributes *sex*, *fasting bloodsugar lessthan120*, *exercisedinductedangina* are binominal. *Restingecq*, *slope* and *thal* have three values. The tree generated by C4.5 algorithm is shown in the Fig 4.

3.2 Naïve Bayes

The next algorithm used is Naïve Bayes. The configurations and the evaluation of the algorithm are shown in the Table 4. Compared to C4.5 the rate of correctly classified instances is very higher. However, in this case the errors are much lower than in case of the C4.5. This especially concerns the Mean A*bsolute Error* whose values do not exceed 10%. Better results are gained in terms of the *True Positive* rate. The ratio between the *TP rate* and the *FP rate* is more beneficial than in case of the C4.5. Poor results were gained for the *Precision*, *F-measure* and *Recall*. All of the testing configurations reached the level of 80% in this regard. The AUC had high values for all of the testing configurations. One of the main advantages of the Naïve Bayes is that a small amount of data is enough for estimation of a mean and a variance. The reason for this is the independence of the variables which is assumed. The Naïve Bayes for each class value estimates whether a given instance belongs to it.

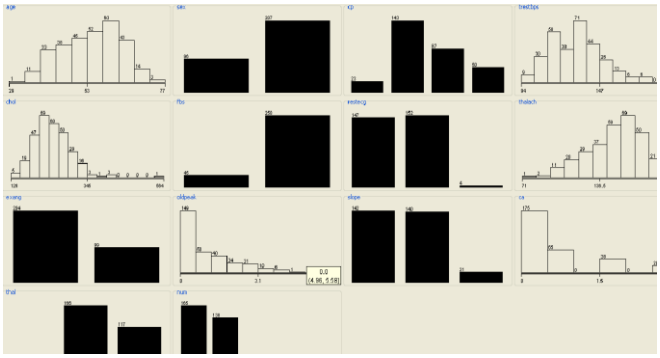

Fig 3: Distribution of the attributes of the heart diseases data

Predictable attribute

1. Diagnosis (value 0: < 50% diameter narrowing (no heart disease); value 1: > 50% diameter narrowing (has heart disease))

Key attribute

1. PatientID – Patient's identification number

Input attributes

1. Sex (value 1: Male; value 0 : Female)
2. Chest Pain Type (value 1: typical type 1 angina, value 2: typical type angina, value 3: non-angina pain; value 4: asymptomatic)
3. Fasting Blood Sugar (value 1: > 120 mg/dl; value 0: < 120 mg/dl)
4. Restecg – resting electrographic results (value 0: normal; value 1: 1 having ST-T wave abnormality; value 2: showing probable or definite left ventricular hypertrophy)
5. Exang – exercise induced angina (value 1: yes; value 0: no)
6. Slope – the slope of the peak exercise ST segment (value 1: unsloping; value 2: flat; value 3: downsloping)
7. CA – number of major vessels colored by floursopy (value 0 – 3)
8. Thal (value 3: normal; value 6: fixed defect; value 7: reversible defect)
9. Trest Blood Pressure (mm Hg on admission to the hospital)
10. Serum Cholesterol (mg/dl)
11. Thalach – maximum heart rate achieved
12. Oldpeak – ST depression induced by exercise relative to rest
13. Age in Year

Table 2: Heart-disease database from Cleveland Clinic Foundation

Table 3 Performance of the C4.5 with respect to a testing configuration for the heart disease database

Testing Methods	Training Set	5- Fold Cross validation	10- Fold Cross validation	15- Fold Cross validation	30 % Split	50% Split	60 % Split
Correctly Classified Instances	92.07 %	75 %	77.55 %	77.55 %	80.66 %	70.19 %	73.5 %
Mean Absolute Error	5.32 %	12.02 %	10.44 %	10.76 %	9.14 %	12.22 %	11.6 %
Root Mean Squared Error	16.24 %	29.13 %	27.25 %	27.71 %	25.59 %	31.67 %	30.23 %
Average FP Rate	8.5 %	26.2 %	23.5 %	22.7 %	19.1 %	29.4 %	26.1 %
Average TP Rate	92.1 %	74.9 %	77.6 %	77.6 %	80.7 %	70.2 %	73.6 %
Average Precision	92.1 %	74.9 %	77.6 %	77.6 %	80.9 %	70.6 %	74.3 %
Average Recall	92.2 %	74.9 %	77.6 %	77.6 %	80.7 %	70.2 %	73.6 %
Average F-Measure	92.1 %	74.8 %	77.4 %	77.6 %	80.7 %	70.2 %	73.4 %
Average AUC	95.2 %	75.2 %	80.9 %	79 %	82 %	72.5 %	70.9 %

Table 4 Performance of the Naïve Bayes with respect to a testing configuration for the heart disease database

Testing Methods	Training Set	5- Fold Cross validation	10- Fold Cross validation	15- Fold Cross validation	30 % Split	50% Split	60 % Split
Correctly Classified Instances	84.5 %	83.16 %	83.49 %	83.49 %	83.01 %	83.44 %	85.12 %
Mean Absolute Error	6.8 %	7.4 %	7.3 %	7.3 %	7.65 %	7.21 %	6.59 %
Root Mean Squared Error	21.28 %	23.1 %	22.99 %	22.78 %	22.91 %	21.79 %	21.23 %
Average FP Rate	16.5 %	17.5 %	17.1 %	17.1 %	16.9 %	17.2 %	15.1 %
Average TP Rate	84.2 %	83.2 %	83.5 %	83.5 %	83 %	83.4 %	85.3 %
Average Precision	84.2 %	83.2 %	83.5 %	83.5 %	83.1 %	83.5 %	85.1 %
Average Recall	84.2 %	83.2 %	83.5 %	83.5 %	83 %	83.4 %	85.1 %
Average F-Measure	84.1 %	83.1 %	83.5 %	83.5 %	83 %	83.4 %	85.1 %
Average AUC	91.9 %	90.1 %	90.4 %	90.5 %	89.7 %	91.1 %	91.7 %

Table 5 Performance of the Multi Layer Perceptron with respect to a testing configuration for the heart disease database

Testing Methods	Training Set	5- Fold Cross validation	10- Fold Cross validation	15- Fold Cross validation	30 % Split	50% Split	60 % Split
Correctly Classified Instances	98.01 %	79.8 %	80.85 %	80.85 %	83.49 %	84.10 %	80.99 %
Mean Absolute Error	1.26 %	8.45 %	7.72 %	7.74 %	6.91 %	7.02 %	7.4 %
Root Mean Squared Error	8.96 %	26.72 %	25.44 %	25.66 %	24.08 %	24.31 %	25.2 %
Average FP Rate	1.9 %	20.6 %	19.4 %	19.7 %	17.1 %	16.6 %	19.5 %
Average TP Rate	98 %	79.9 %	80.9 %	80.9 %	83.5 %	84.1 %	81 %
Average Precision	98 %	79.8 %	80.9 %	80.8 %	83.5 %	84.2 %	82 %
Average Recall	98 %	79.9 %	80.9 %	80.9 %	83.5 %	84.1 %	81 %
Average F-Measure	98 %	79.8 %	80.9 %	80.9 %	83.5 %	84 %	80.8 %
Average AUC	97.6 %	86.8 %	89.1 %	89.2 %	90.1 %	89.3 %	90.7 %

3.3 Multi Layer Perceptron Neural Network

The next algorithm used in evaluation of effectiveness and accuracy of data mining methods is the Multilayer Perceptron Neural Network. The results of the experiments are shown in the Table 5. It shows that all of the models had moderate predictions. In terms of errors, Multilayer perceptron Neural network has gained. The True Positive rates are comparatively higher than the other two algorithms. The algorithm also outperformed the other two models in case of correctly classified instances. The results of the AUC for all the configurations are also significantly good compared to other two algorithms.

The overall comparison of the performance and efficiency of the three algorithms are shown in the Table 6. Since the 10-Fold Cross validation has shown better results for all the configurations for all the three algorithms as shown in Table1, 2 and 3, alone these two configurations has been considered for the comparison.

Table 6 Performance Comparison of the algorithms with respect to the measures with the use of 10-fold cross-validation

Testing Methods	C4.5 Algorithm	Naïve Bayes Algorithm	Multilayer Perceptron Neural Network
Correctly Classified Instances	77.55 %	83.49 %	80.85 %
Mean Absolute Error	10.44 %	7.3 %	7.72 %
Root Mean Squared Error	27.25 %	22.99 %	25.44 %
Average FP Rate	23.5 %	17.1 %	19.4 %
Average TP Rate	77.6 %	83.5 %	80.9 %
Average Precision	77.6 %	83.5 %	80.9 %
Average Recall	77.6 %	83.5 %	80.9 %
Average F-Measure	77.4 %	83.5 %	80.9 %
Average AUC	80.9 %	90.4 %	89.1 %

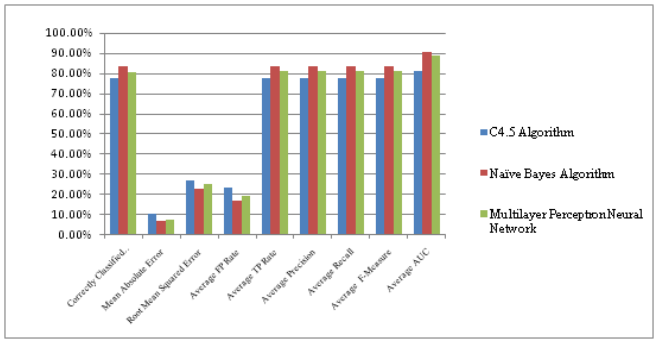

Fig 4: Performance Comparison of the algorithms

4. CONCLUSION

A huge amount of data is collected each day in medical domain. That includes data about patient records, disease, diagnosis, treatment, medicine, doctors, symptoms etc. These data are collected in large databases called as medical databases. These databases provide useful information of symptoms and diagnosis. With the use of Health Care Systems, it is easy to discover relationships in the historical data. This knowledge can be used for diagnosis in future cases. The main goal of this study is to analyze the uniqueness of health care data mining, Identifying and selecting the most efficient data mining technique to be used in the modern Health care system.

Three data mining techniques C4.5 for decision tree classification, Naïve Bayes algorithm and Multilayer Perceptron Neural Network were analyzed using the Cleveland Heart Disease database. WEKA tool was made use to train and test the algorithms. From the results obtained, it is evident that the Naïve Bayses algorithm has outperformed the other two algorithms in all the configurations.

As we can see the highest score of the performance for the medical database and data mining algorithm is achieved for the Naïve Bayes classifier. The second classification is achieved with the use of Multilayer Perceptron algorithm. Finally the C4.5 algorithm's performance is evaluated. Even thought the Multilayer Perceptron Neural Network has closely performed well, the time taken to build the model is higher than the other two algorithms. From Table 6, it is clear that the C4.5 algorithm has shown poor performance for all the configurations.

The results obtained during the studies proved the applicability of the data mining algorithms for the medical datasets. Such results are very useful in creating new Health Care Systems. Our future work may include the evaluation of chosen algorithms on the basis of other medical datasets.

5. REFERENCES

[1] Venkatadri, M., & Lokanatha, C. R., 2011, "A review on data mining from past to the future". *International Journal of Computer Applications*, *15*(7), 19-22.

[2] Glymour, C., D. Madigan, D. Pregidon and P. Smyth, 1996. " Statistical inference and data mining", Communication of the ACM, pp: 35-41.

[3] Shams, K., & Farishta, M., 2001, "Data warehousing: toward knowledge management", *Topics in health information management*, *21*(3), 24-32.

[4] Han, J., Kamber, M., & Pei, J. (2006). *Data mining, southeast asia edition: Concepts and techniques*. Morgan kaufmann.

[5] Lu, H., Setiono, R., & Liu, H., 1996, "Effective data mining using neural networks. Knowledge and Data Engineering", IEEE Transactions on, 8(6), 957-961.

[6] Saishanmuga Raja V, and S. P. Rajagopalan., 2013, "A Comparative Analysis Of Optimization Techniques For Artificial Neural Network In Bio Medical Applications", Journal of Computer Science 10.1, 106.

[7] Saishanmuga raja V. and Rajagopalan SP, 2015, "Selecting a Best Architecture of an Artificial Neural Network Using Genetic Algorithm for Lung Segmentation", Aust. J. Basic & Appl. Sci., 9(11): 39-46.

[8] DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, 837-845.

[9] Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J. J., Sandhu, S., ... & Froelicher, V. (1989). International application of a new probability algorithm for the diagnosis of coronary artery disease. *The American journal of cardiology*, *64*(5), 304-310.