# Comparative Analysis of Data Mining Techniques in Sphere of Medical Science

Divya Arora
Assistant Professor
Maharaja Agrasen Institute of Technology
GGSIPU

Karuna Middha
Assistant Professor
Maharaja Agrasen Institute of Technology
GGSIPU

## ABSTRACT
Data Mining is the extraction of useful patterns from the large data set. With the increase in population, healthcare organizations are collecting a large volume of data which is not optimally used. To use it optimally, Data mining field is gaining popularity in this field of research due to its various approaches or techniques to mine the effective data from the huge set in an efficient way. The goal of the paper is to study techniques, algorithms and accuracy of results since 2011 from selected set of papers for the diagnosis of two diseases i.e. Heart attacks and Liver Disorder.

## Keywords
Data Mining, Neural networks, Regression

## 1. INTRODUCTION
Due to evolution in the various fields of engineering, sociology, medical science, there is an increase in data. Earlier, due to cost constraints it was quite difficult to manage large databases with varied data but knowledge discovery in databases(KDD) have ease the task. Data mining is the core step in KDD which is the non- trivial extraction of implicit unknown useful information about data [1]. Patterns or knowledge is discovered from varied data which undergo various steps in KDD such as Selection, Pre-processing, Transformation, Data Mining and Evaluation. Data mining works on predictive model and descriptive model. The Predictive data mining model works on the theory of prediction about values of data, utilizing results found from different datasets. The Predictive model includes classification, prediction, regression and time series data analysis.

## 1.1 Classification
It is a data mining technique that utilizes set of classified samples to develop a model that can classify the population of records at large. This approach frequently leverages decision tree or neural network-based classification algorithms. Classification technique is used in customer division, modeling enterprises, credit risk analysis, health care organizations and many other applications. The goal of classification is to accurately predict the target class for each item in the data set.

### *1.1.1 Decision tree*
Decision Tree is a commonly used algorithm in data mining. In first stage an attribute must be selected as a root node. The most efficient way to create the root node is to choose one that splits or segregates the data effectively. Each split tries to tone down the actual data until each has the same classification. The most information gain is the best outcome of the splitting process. The model predicts the value of a target variable based on several input variables. Interior node corresponds to one of the input variables while each leaf represents a value of the target variable given the values of the input attributes represented by the path from the root to the leaf. Disadvantages: The performance of decision trees is low if more complex interactions among attributes exist. If the target attributes do not have discrete values, the divide and conquer method cause a lot of clutter. Existence of more highly relevant attributes, leads to better performance. Types of Decision trees includes - a) Iterative Dichotomiser 3 (ID3) is an algorithm that creates the decision tree based on below steps. In the first step all unused attributes are taken to count their entropy concerning test samples. In the second step, minimum impurity is measured by choosing attribute for which entropy is least (or, equivalently, information gain is most). In the final step, make node containing that attribute. b) C4.5 algorithm is used to generate a decision tree for classification, and for this reason, is often referred to as a statistical classifier. The general algorithm for building decision trees is to check for base cases in the first step. For each attribute find the normalized information gain from splitting. Create a decision node that splits on best attribute i.e. the one with the highest normalized information gain. Recursively add the nodes as child node on the sub lists obtained by splitting on best attribute. c) Random Forests grows a decision tree to classify a new factor from an input set by putting the input factor down each of the trees in the forest. Each tree gives a classification, for which each of the tree belonging to forest votes. The classification with most votes is chosen by the forest.

### *1.1.2 Rule-based technique*
It categorizes data by using a collection of conditional rules. The rule is described as positive or negative classification. The rule engine is extremely expressive as they operate with the variables without any transformation or manipulation. Interpretation and generation of rule-based classifiers is easy and they can classify new instances proficiently.

#### *1.1.2.1 Memory based learning*
It is an algorithm that compares new problem instances with instances stored in the memory, instead of performing explicit simplification.

It constructs hypotheses directly from the instances which means that the complexity can grow with the data. Ability to adapt is the advantage memory-based learning has over other methods of data mining. K-nearest neighbor algorithm is an example of an instance based learning algorithm.

#### *1.1.2.2 K-nearest neighbor*
It is a classification technique that decides in which class to place a new case by examining some number ‘k’ out of the similar cases or neighbors. It counts the number of cases for each class, and assigns the new case to the same class to which most of its neighbors belong. Predictor variables of

non-standard data types, such as text can also be used for building models. Disadvantage - A main disadvantage is the large memory requirement for the storage of complete data set as a sample. Response time using sequential search will be very large for a big sample.

### *1.1.3 Neural networks*

These are non-linear statistical data modeling tools. Data mining using neural networks involve manipulation and fertilization of data thereby helping in making informed decisions. Complex relationships between inputs and outputs can be interpreted using neural network modeling technique. Disadvantage -Its training or learning process is very slow, hence very expensive. Secondly, they follow the black box approach i.e. the internal details regarding the occurrence is not visible.

#### *1.1.3.1 Feed forward neural networks(FFNN)*

It has three layers: an input layer, hidden layer and output layer. Each layer has one or more Processing Elements (PEs). PE's are referred to as neurons or nodes. Input is received by a PE from the previous layer. Connections between the PEs at each layer have a weight associated with them which is adjusted during preparation. Information only travels through the network in the one direction i.e. forward direction.

### *1.1.4 Bayesian network*

These are directed acyclic graphs (DAG). In this graph each node represents a random variable. The variables can take any form like observable quantities, latent variables, unknown parameters or hypotheses. Conditional dependencies are represented by Edges while nodes which are dis-connected represent conditionally independent variables. A probability function is associated with each node having input as particular set of values for the node's parent variables and gives the probability of the node represented variable. Disadvantage - The assumption that all the attributes are independent and not correlated is a major drawback of Bayesian network. In various domains the degree of correlation between the attributes is very high.

### *1.1.5 Support Vector Machines*

These are learning models with algorithms that help to analyze data and recognize patterns. It is a non-probability based binary linear classifier that takes a set of input data and predicts, for each given input, the possible output. The basis of SVM is well defined theoretical approach to regularization which separates it from similar models like neural networks and radial basis functions. The ease of use of SVM and it quality is far beyond the capacities of more traditional methods. Disadvantage - If the number of features or attributes is much greater than number of available samples then SVM is more likely to give poor performances. Additionally, it does not directly provide probability estimates. These are calculated using other expensive methods.

## 2. REGRESSION

It is a predictive data mining model which analyzes the dependency or degree of dependency between the attribute values of the same dataset. For example, regression may be used to predict the cost estimate of an artifact or service, provided other variables. The target values are always known in the regression techniques. Linear regression is used to estimate a relationship between two variables. This technique uses the straight-line mathematical formula. This simply means that, given a graph with a Y and an X-axis, the relationship between X and Y is a straight line with few outliers.

## 3. DATA MINING TOOLS

Data mining tools are used to predict the accuracy in various fields such as of diseases in health care organizations.

### *3.1 Weka*

Weka primarily used for data mining tasks, is a collection of machine learning algorithms. Some popular algorithms that WEKA has are logistic regression, decision tree, neural network and support vector machine. The algorithms can be applied either directly to a dataset or called from Java code. Weka is an open source mining tool. Some of tools contained in a Weka can used for data pre-processing, classification, regression, clustering, association rules, and visualization. It is well-suited for developing new machine learning schemes.

### *3.2 Tanagra*

Tanagra is mainly used to perform hypothesis test or calculate confidence intervals. Tanagra data mining software has many components like Data Visualization PLS, Statistics clustering, Non- parametric statistics, Instance Selection, Feature Construction, Regression Association and Factorial Analysis. The main functions like Normal distribution, chi-squared distribution etc. is widely used in many fields like health care.

### *3.3 MATLAB (matrix laboratory)*

It is used for predictive analysis using various popular suit of tools like NN and SVM. MATLAB software has tools that works on the principles of deep learning, which is a machine learning technique that learn features and tasks directly from the data. Data can be images text and sound. The untold is GUI in MATLAB. To use it we don't need any programming knowledge. This tool is very useful in the field of medical science.

### *3.4 DOT NET FRAMEWORK*

Data Mining with Microsoft SQL Server is used for analyzing data to find hidden patterns using automatic methodologies. SQL Server Analysis Services, by moving the activity to a server, makes data-intensive processes run in an environment designed for processing efficiency and connectivity to enterprise systems. Data mining works best in a server environment and thereby Microsoft dot net framework and SQL Server has an edge over data mining applications for either desktop or server application. SME's who understand the data are able to translate the results of data mining into actionable business information in various fields like medical science.

### *3.5 Orange*

Orange, a data mining suite has a GUI based workflow built for data analysis. What it meant is that no coding prerequisites to able to work using Orange and mine data, crunch numbers and derive insights. Tasks like basic visuals to data manipulations, transformations, and data mining can be performed using Orange. One of the selling point of Orange is its wonderful visuals. Silhouettes, heat-maps, geo-maps and all sorts of wonderful visualizations are available. Using this tool, one can write simple and clear scripts in Python, which build upon implementations of computationally-intensive tasks. It provides great flexibility in terms of reusability of components

### *3.6 Rapid Miner*

Rapid Miner uses a wide range of machine learning algorithms for a variety of use cases to analyze unstructured

data. Unstructured data can be in the form of social media discussions, online blogs, product reviews or data from hospitals etc. Marketing and consumer research to discover trends and opinions is one of the leading industry leveraging from this tool. Rapid Miner provides a suite of software tools like Studio, Server and Radoop (Hadoop and Spark tool). Using Radoop, the analytic tasks can be visually represented using data process flows that are easy to develop & maintain, which is finally pushed and executed in Hadoop environment.

## 4. DISCOVERING KNOWLEDGE IN MEDICAL SPHERE VIA DATA MINING ALGORITHM

With the escalation in population, data generated by health care organization is highly un-dimensional, uncertain and massive which should be managed to forecast the precise disease by doctors in one go [2]. Through the application of traditional statistical methods anticipation is not optimized. In contrast, data mining overcomes the problem associated. It is an essential step in knowledge discovery process of medical sphere. Many data mining algorithms that are available which can be applied in the field of medical science for pattern discovery (Figure 1).

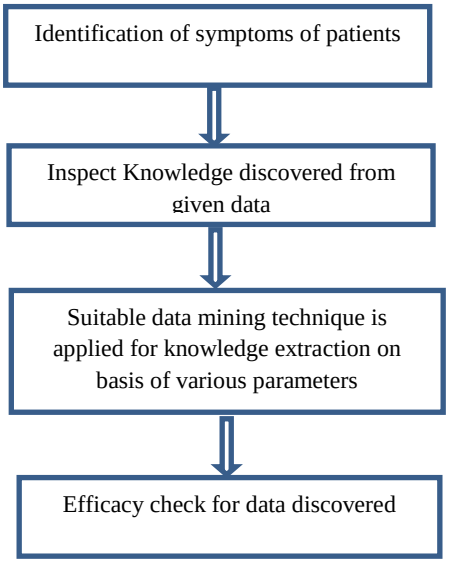


**Fig 1: Process of extraction of useful data in health care**

Here two diseases have been studied and evaluated i.e. Heart Attacks and Liver Disorder.

## 5. DISCUSSION

Our study is presenting analysis of two diseases which shows accuracy of techniques used by the researchers using any of the data mining tool. Findings of the study has been represented in a tabular form.

**Table 1: Heart Disease**

| Year | Author | Technique | Accuracy | Tool (if any) |
|---|---|---|---|---|
| 2011 | Jyoti Soni et.al [3] | Naïve Bayes<br>Decision Tree<br>K-NN | 52.33%<br>52%<br>45.67% | Tanagra |
| 2012 | Nidhi Bhatla, Kiran Jyoti [4] | Naïve Bayes<br>Decision Trees<br>Classification via clustering | 96.5%<br>99.2%<br>88.3% | Weka |
| 2013 | A.V Senthil Kumar [5] | Fuzzy Mechanisms | 94.11% | MATLAB |
| 2014 | M. A. Nishara Banu and B. Gomathy [6] | K-mean based MAFIA<br>K-mean based MAFIA with ID3<br>K-mean based MAFIA with ID3 and C4.5 | 74%<br>84%<br>89% | Tanagra |
| 2015 | Shivnarayan P, Ram BP, U.Rajendra A.[7] | Least Squares SVM | 96.8% | Tunable-Q Wavelet Transform (TQWT) |

**Table 2: Liver Disorder**

| Year | Author | Technique | Accuracy | Tool(if any) |
|---|---|---|---|---|
| 2011 | Shelly Gupta et.al [8] | Naïve Bayes<br>J48<br>C4.5<br>ID3<br>KNN | 53.04%<br>63.47%<br>79.91%<br>60.57%<br>67.54% | Weka<br>Weka<br>Tanagra<br>Tanagra<br>Tanagra |
| 2012 | A.S. AneeshKumar , C. Jothi Venkateswaran [9] | Naïve Bayesian<br>C4.5 | 89.60%<br>99.20% | - |

| | | | | |
|---|---|---|---|---|
| 2014 | S.Dhamodharan [10] | Naïve Bayes | 75.54% | Weka |
| 2014 | E.Rezaei-Darzi et.al[11] | Decision Tree<br>ANN | 83.3%<br>80.33% | Weka |
| 2015 | Dr. S. Vijayarani , Mr.S.Dhayanand[12] | SVM<br>Naïve Bayes | 79.66%<br>61.28% | Matlab Tool 2013 |
| 2016 | Tapas Ranjan Baitharua, Subhendu Kumar Pan [13] | Multilayer Perceptron | 71.59% | Weka |

## 6. CONCLUSION

The study surveyed various data mining techniques for two major diseases Heart Attacks and Liver Disorders. It is seen that classification has been majorly used in health care sector for medical diagnosis. It is seen that Decision Trees gives best results in the predictions since 2011. Each classification technique shows different behavior on different data sets depending upon various attributes for these diseases. Any one of the single model is insufficient to give optimal results. As accuracy depends upon the dataset used for learning in classifiers. To obtain optimized results with higher accuracy, we need to design a model which would be the integration of data various mining techniques. Our future perspective is to design such model for enhanced prediction. This model would be the hybrid of neural networks , random forest and Bayesian networks after considering the drawbacks of each so as to predict optimally.

## 7. REFERENCES

[1] R. Naveen Kumar, M. Anand Kumar, "Medial Data mining techniques for health care systems", IJESC, 2016, volume 6, Issue No.4.

[2] Babajide O. Afeni, Thomas I. Aruleba and Iyanuoluwa A. Oloyede, "Hypertension Prediction System Using Naive Bayes Classifier", Journal of Advances in Mathematics and Computer Science, 2017; Article no. JAMCS.35610.

[3] Jyoti Soni, Uma Ansari and Dipesh Sharma, Predictive Datamining for medical Diagnosis: An overview of Heart Disease Prediction", International Journal of Computer Applications (0975 –8887), Volume 17–No.8, March 2011

[4] Nidhi Bhatla and Kiran Jyoti, "An Analysis of Heart Disease Prediction using Different Data Mining Techniques", International Journal of Engineering Research & Technology (IJERT), ISSN: 2278-0181,Vol. 1 Issue 8, October –2012

[5] A.V Senthil Kumar. Generating Rules for Advanced Fuzzy Resolution Mechanism to Diagnosis Heart Disease. International Journal of Computer Applications, 2013; 77(11): 6-12

[6] M.A.Nishara Banu and B.Gomathy," Disease Forecasting System Using Data Mining Methods", 2014

[7] Shivnarayan P, Ram BP, U. Rajendra A. Automated diagnosis of coronary artery disease using tunable-Q wavelet transform applied on heart rate signals. Knowledge-Based Systems, 2015; 82: 1-10

[8] Shelly Gupta, Dharminder Kumar and Anand Sharma, "Performance analysis of various data mining classification techniques on health care", International Journal of Computer Science & Information Technology (IJCSIT) Vol 3, No 4, August 2011.

[9] A.S. AneeshKumar, C. Jothi Venkateswaran, "Estimating the Surveillance of Liver Disorder using Classification Algorithms", international Journal of Computer Applications (0975 –8887) Volume 57–No.6, November 2012.

[10] S.Dhamodharan,"Liver Disease Prediction Using Bayesian Classification", 4thNational Conference on Advanced Computing, Applications & Technologies, Special Issue, May 2014.

[11] E.Rezaei-Darzi ,F . Farzadfar, A. Hashemi-Meshkini et.al, "Comparison of two data mining techniques in labelling diagnosis to Iranian pharmacy claim dataset: Artificial Neural Network Vs Decision Tree Model", archives of Iranian Medicine, Volume 17, Number 12, December 2014.

[12] Dr. S. Vijayarani, Mr.S.Dhayanand, "Liver Disease Prediction using SVM and Naïve Bayes Algorithms", International Journal of Science, Engineering and Technology Research (IJSETR) Volume 4, Issue 4, April 2015.

[13] Tapas Ranjan Baitharua, Subhendu Kumar Pan, "Analysis of Data Mining Techniques For Healthcare Decision Support System Using Liver Disorder Dataset", International Conference on Computational Modeling and Security (CMS 2016), proedria Computer Science 85 (2016) 862 – 870.