A Review on Prediction of Multiple Diseases and Performance Analysis using Data Mining and Visualization Techniques

Ajinkya Kunjir
BE(Computerscience)
Modern Education Society's
College Of Engineering
Pune-411001,India

Harshal Sawant
BE(Computer Science)
Modern Education Society's
College Of Engineering
Pune-411001, India

Nuzhat F. Shaikh
Head Of Department
Modern Education Society's
College Of Engineering
Pune-411001,India

ABSTRACT
In the field of medical science a tremendous amount of data is generated, doctors need to test the patient physically to find out the injuries and disease of the patient. This paper outlines the idea of predicting a particular disease by performing operations on the digital data generated in the medical diagnosis. In this project an efficient genetic algorithm hybrid with the techniques like back propagation and Naive Bayes approach for disease prediction is proposed. Bad clinical decisions would cause death of a patient which cannot be afforded by any hospital. To achieve a correct and cost effective treatment, computer technology Systems can be developed to make good decision. There is a lot of medical information unexplored, which gives rise to an important query of how to make useful information out of the data. The purpose of this project is to construct a basic prototype model which can determine and extract unknown knowledge (patterns, concepts and relations) related with multiple disease from a past database records of specified multiple diseases. It can solve complicated queries for detecting a particular disease and thus assist medical practitioners to make intelligent clinical decisions which traditional decision support systems were not able to. By providing efficient treatments, it can help to reduce costs of treatment. The medical organizations are ”rich in data” but their ”knowledge utilization is poor ”. There is a lack of sufficiency of improved analysis techniques to find relations, concepts and patterns in the medical data. Data mining is science and engineering study of extracting previously undiscovered patterns from a huge set of data. In this paper, data mining methods namely, Decision tree, Naïve Bayes, and Back-Propagation(ANN) algorithms are implemented on medical data sets .The medical datasets will be represented graphically(graphs , charts , shapes)using different visualization techniques. The algorithms are compared and evaluated on basis of their accuracy and time consumption factors. The algorithm which gives out high accuracy and less duration to give the output is analysed and implemented**.**

General Terms
Data Mining and techniques

Keywords
Prediction, Classification, BP Neural networks, Genetic algorithms, Decision Tree, Regression, Naive Bayes.

1. INTRODUCTION
Data Mining is an important domain of computer science field for many reasons. Knowledge Discovery from Data(KDD) is a process in which selected target data is mined or discovered from the big data. Data mining is used to mine data patterns which are previously undiscovered, novel, valid and potential. Predictions and descriptions are important and focused goals of this field of interest. Prediction uses the previous related knowledge of the subject and uses that knowledge to foresight or predict the unseen subject of the same field. Ankita Dewan, Meghna Sharma in 2015 stated that Data mining has its applications in fields like healthcare, home automation , Finance , Banking. Practitioners in field of medical science generate data with a limited hidden information present, this information is not used effectively for prediction purpose[1]. To tackle this problem, the unused data is converted into a dataset for modeling of data using various data mining technique. People die because of insufficient results and problems that were not taken into considerations. There is a need for medical practitioners and doctors to predict the patients disease and internal injuries before it occurs. The predictions and classifications in data mining help discover relations and patterns in patient medical data in order to improve their health. The problems of data mining represent great challenges for networking and distributed computing. Many researchers had created several methods to deal with the problems of data mining . A Framework for Analyzing Categorical Data (2009)was created by Varun Chandola, Shyam Boriah, and Vipin Kumar. The challenge of grouping and positioning was solved by Hui Xiong et.al in their proposed paper HICAP: Hierarchical Clustering with Pattern reservation (2004)[2]. Techniques like pattern recognition and classifier models were used by the researchers for improved prediction and classification. Jyoti Soni, Ujma Ansari, Dipesh Sharma, Sunita Soni, in their paper proved that The operation on the datasets were carried out using classification algorithms like Decision trees, Naive Bayes , and K-Nearest Neighbours and results proves that Nave Bayes technique outperformed the other techniques used[3].The prediction can be improved by using association rules to find out the frequently used elements . The algorithms such as Apriori,, FP-Growth , MAFIA algorithms are used to mine frequent itemsets from a database .In the implementation and the accuracy testing of Naive Bayes , Decision tree , and Back propagation Neural Network algorithms, the Back Propagation prediction model of Neural Networks achieved higher accuracy than the other two[1].

2. RELATED WORKS
2.1 Background and Research
Data Mining is the variation of Knowledge Discovery from Data(KDD). Tasks of the Domain such as classification, prediction, clustering and regression play an important role in mining patterns and discovering previously undiscovered patterns. The association rules are used to denote an interesting relationship pattern between two items of a

respective database. Abhishek Taneja in his journal mentioned that Algorithms with search restrictions and constraints are also introduced to decrease the number of association rules and also for validation purpose[4].Popular data mining algorithms (Support Vector Machine Analysis, Artificial Neural Network, Naive Bayes) are frequently used by the practitioners to develop a prediction model using attributes and attribute values. SVM became the best prediction model and analysis method after artificial neural networks.

2.1.1 Classification Models : A survey

Classification is the process of predicting discrete or nominal values. Classification is the use of prediction to predict class labels of the new unseen test tuple .Typical applications of classification are target marketing of product, medical diagnosis based on symptoms of patient, and classifying credit approval based on customer data. The classifier model is constructed by performing operations on the training dataset and then validating the test data so that the previously unknown and unseen data-tuple could be classified with respect to its class by referring the class label. Various classification techniques used to construct the classifier models are regression, decision trees, Rules, Neural networks .T.Revathi, S. Jeevitha, in their IJSR's vol- 4[2015] elaborated about the classification models and their accuracy testing performance[5].The popular classification models and algorithms which were compared on the basis of the accuracy factor and validation rate are : • Decision Tree • Naive Bayes • Artificial Neural Networks - BP • Support Vector Machine Analysis

- Decision tree Induction :

Decision tree is a tree structure consisting of set of vertices and links in which , the internal nodes represents the tests on attributes of the dataset , leaf nodes indicates the class labels , the edges or the links denote the outcome of the test . A decision tree represents rule and it is a very popular tool for classification and prediction. To recognize and approve the discovered knowledge got from decision model is very crucial task. A decision node is the node of tree which has leaf node or subtree. Some test to be carried on the each value of decision node to get the decision of class label or to get next sub -tree. Monika Gandhi, Dr.Shailendra Narayan Singh in their paper of INDIACom conference(2015) gave a general description about decision tree, this classifier can handle both categorical data and numerical data[6].The most popular algorithm of all the algorithms of decision tree is ID3(Interactive dichotomized 3) . ID3 uses attribute selection measures such as information gain and entropy to classify data in tree structure. Attribute selection measures are :

a) Information gain –This statistical method can be applied on continuous-valued attributes. The attribute which has the highest information gain is selected for split. Assume that there are two classes, P and N. let 'S' be the number of samples , out of these 'p' samples belong to class P and 'n' samples belong to class N. I(p,n) can be defined as the total information needed to decide if a data example of 'S' belongs to P or N.

$$E(A) = \sum_{i=1}^{v} \frac{p_i + n_i}{p + n} I(p_i, n_i) \qquad (1)$$

As shown in (1) above, Entropy can be defined as the expected amount of information which is needed to assign a class to a randomly derived example in 'S' under the optimal shortest-length code. The information gain is formulated as, Gain(A) = I(p,n) - E(A).

b) Gain Ratio : As shown in (2) , gain ratio is the ratio of gain split and split info , it is an alteration of the information gain that reduces its favouritism on high - branch attributes. Gain ratio should be big when data is evenly spread and small when all data belong to one branch. So it considers number of branches and size of branches when it selects attribute to split.

$$GainRATIO_{split} = \frac{GAIN_{Split}}{SplitINFO}$$

- Naïve Bayes Classifier :

Naive Bayes is a type of classifier in which multiple hypothesis can be predicted by their probability weight. The output of several classifiers can be combined by multiplying the probabilities that all classifiers predict for a given class, this is called as meta-classification. Naive Bayes classification can be formalized with the arrival of conditional probability, which is also called as "Posterior Probability". Naive Bayes classifier uses concepts of statistics and calculations to predict the class label for an unseen set of data examples. Bayesian Classifier performs well in complex world situations. Jyoti Soni et.al in 2015 derived the comparision between the algorithm on the basis of the performance study. Naive Bayes outclasses decision tree and K-NN methods with an accuracy of 52.33 percent and the time taken to give the output was 609 ms[3].Bayes theorem is used to find conditional probabilities . The conditional probability of an event is a likelihood obtained with the additional information that some other event has previously occurred. P(X|Y) is the conditional probability of event X occurring for event Y which has already occurred. The probability which is present at current without any additional information is called as priori probability. The probability which is obtained after getting the additional information is called as posteriori probability

- Artificial Neural Network :

Artificial neural networks are the most powerful learning models. They can have wide variety of complex functions which represents multi-dimensional input-output maps. ANN is an information processing paradigm which is motivated by biological processing system i.e Brain. ANN is represented as a system of interconnected "neurons" which communicate by passing messages to each other. Amrender Kumar in his library paper stated that the neural networks consisting of thousands of neurons communicate by passing messages and sending signals to each other[7].ANN can be represented graphically as a structure of several layers such that the directed links have a direct connection with them. The nodes denote the neurons while link denotes synapses. The nodes are considered as information processing units and links act as communication media. Nuzhat F. Shaikh, Dharmpal D. Doye, in their paper described the working, operation and the implementation of Feed Forward Back Propagation Neural Networks(FFBNN)[8]. The real time applications and the use of the back propagation method in recognition systems is highlighted in the paper. The three layers of the neural networks are the input layer, hidden layers , and an output layer. A neural network can be defined as a set of interconnected neurons which can interact with each other. Artificial Neural Networks is a powerful supervised learning method. Information processing is a paradigm of Neural Network which also covers many of its application and other fields. The types of Artificial Neural Networks are:

a) Perceptron – A perceptron can be defined as neuron with multiple inputs and one output.

b) Multi-layered Perceptron - Includes all the three layers and also called as feed forward networks.

c) Recurrent Neural Networks - This type of network includes backwards links from output to input and also hidden layers.

d) Self Organizing Maps - These types of networks mainly have grid topology with unequal and unbalanced grid weight.

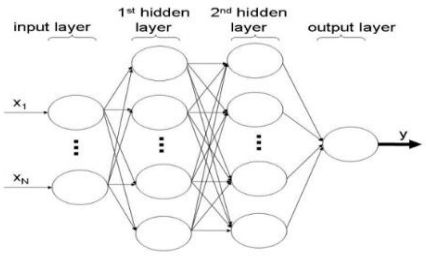

Fig 1:Overview of Artificial Neural Networks

2.2 Comparative Study Of Algorithms

The experiment performed on the dataset consisting of 1000 records and 14 attributes resulted that classification accuracy of Naive Bayes algorithm is better compared to other algorithms i.e Decision tree and K-NN method. The data mining algorithm "Naive Bayes" possessed the highest accuracy of all as compared to Decision tree and K-NN(Nearest Neighbour) method[3].The table (1) below depicts the comparison of algorithms with their accuracy and latency period.

Table 1. Performance study with 1000 records and 14 attributes

ALGORITHM USED	ACCURACY	TIME TAKEN
NAÏVE BAYES	52.33 %	609 ms
DECISION TREE	52 %	719 ms
K-NN METHOD	47.67 %	1000 ms

.

The clinical data related to heart disease which was used in the experiment was derived from Cleveland database which is publicly available dataset on internet i.e UCI Repository. This dataset contained 76 attributes and more than 1000 data samples[5].The algorithms tested on the data were Naive Bayes , Decision tree and Back propagation predictive model of Artificial Neural Network. From table (2), as shown in the table below the back propagation model achieved the greatest accuracy of '100 percent' out of all the algorithms.

Table 2. Performance study with 1000 records and 76 attributes

ALGORITHM USED	ACCURACY
BACK-PROPAGATION NETWORK	100 %
DECISION TREE	99.62 %
NAÏVE BAYES	90.74 %

3. DATA MINING TECHNIQUES: A SURVEY

A Cluster can be defined as a group of similar entities positioned or placed in a specific area .The grouping of entities can be divided into categories , such that similar entities in one group and dissimilar entities into other. Overall, Clustering is the process of categorizing all the similar elements into a set of classes.

Characteristics of Clustering :

- A cluster of data can be partitioned into small sub-clusters, the analysis of the clusters involves partitioning and then classifying the class label.
- Clustering has a characteristic of adaptability.
- The clusters are scalable, interpretable, deal with different types of attributes , high dimensionality , ability to avoid noise interference

A. Regression

Regression is the method of prediction in data mining, the value of the dependent variable is predicted using the independent variable and the relationship between the two variables. The relationship is represented mathematically using statistics and visualization. There are two types of regressions , one which uses single independent variable for prediction called as simple linear regression an the other which use multiple variables is called as multiple regression. The formula for calculating regression is as follows

i)Y = a + bX + u where , Y = dependent variable, X= independent variable, a= intercept, b=slope, u = regression residual.

B. Association Rules

An association between two objects depict the interesting relationship shared between the objects. The two associated objects share a quality bonded pattern between themselves. Extracting or mining frequent itemsets from a transaction database is called as association rule mining. Let I be the total no. of items in the set, in which I1 and I2 share an interesting relationship pattern with each other. These two items I1 and I2 are said to be associated with each other. Multilevel association rules, Support of an item is the count of the occurrences of an item in a transaction .An itemset is called a frequent itemset when its support count is greater than or equal to the minimum support threshold. The algorithms used to generate frequent itemsets and association rules are :

a) Apriori Algorithm

b) FP - growth Algorithm

4. DATA VISUALIZATION

The data is stored as attribute-value pair in the tabular form in the dataset. This data can be viewed graphically in various shapes, graphs etc by the technique of graphical data visualization. This technique elaborates the knowledge representation and utilization of the dataset used for visualization. The applications of visualization supports decision making, data exploration, knowledge representation and data transformation. The data can be visualized in the form of shapes , trees , graphs , maps and pixels. The main advantage of data visualization is the ease of understanding the data selected for the user. The amount of data used and the parts covered by several attributes are displayed graphically by this technique of the field. This technique of data mining has a great promise of advancement in the near future and in diff erent subject domains.

4.1 Data Visualization Techniques

The large data in the data set may have incomplete, noisy, inconsistent values in it. Visualization techniques support such inconsistencies and represents the data in graphical diagrams and plots .These techniques provide user satisfaction and conviniency in understanding the dataset. R.Spence, L. Tweedie, H. Dawkes, and H. Su, in InfoVis 95 stated that techniques used for representation are processed by evaluating the relationships and correlations between the data values[9]. The calculations are done by statistical formulae , geometrical techniques , pixel and image based , and various purposed methods.

- Geometric Technique : This is technique of data visualization in which the dataset selected will be displayed in various geometric formats , shapes , and patterns. Visualization of geometric transformations and projections of the data Types of Geometric Techniques are :

  a)Scatterplot matrices *b*)Hyper-slice *c*)Parallel coordinates

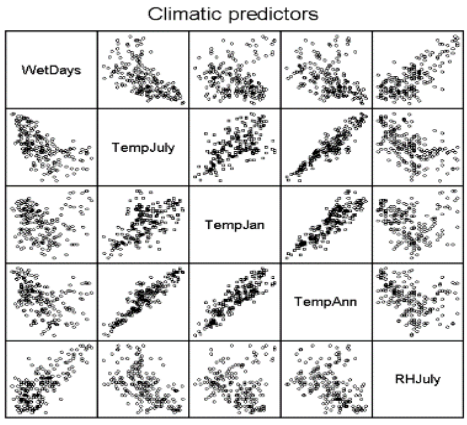

Fig 2 : **Scatterplot Matrix for climate prediction**

- Graph Based Techniques : Data visualization is presenting the data in pictorial format. This type of visualization method deals with graphs consisting of nodes and links. Graph visualization is the easiest type of visualization and which provide a better way of data exploration. The types of Graph visualization techniques are :

  *a)*2D/3D Graphs *b)*line graphs *c)*Bar graphs

- Hierarchical Bases Technique : This Visualization technique displays multiple dimensions simultaneously. This technique is not feasible for large datasets. Hierarchical visualization partitions all dimensions in to subspaces and these subspaces are then visualised in hierarchical manner.

The types of Hierarchical techniques are :

a. Tree maps : Tree maps are well suited for displaying large amounts of structured data. The visualization space is divided into multiple rectangles that are sized and ordered according to a quantitative variable. Each individual rectangle on a level in the hierarchy represents a category in the column.

b. Mosaic plot : These plots are the graphical illustration of successive decompositions. Rectangles in this structure represents the count of the categorical data and at every stage rectangle are split parallel.

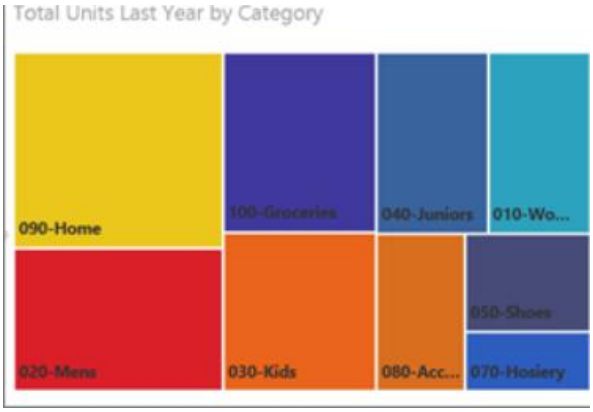

Fig 3: Data categories visualized as tree map

Advantages of Data Visualization :

Few advantages of applications of data visualization are mentioned below :

- Easy data exploration and ease of understanding.
- Visualization does not require any complex mathematical calculations and geometry.
- Visualization is the better way of knowledge utilization and knowledge representation.

5. MOTIVATION

In the medical industry, machine learning algorithms can be used to diagnose some serious diseases. Among all diseases, cardiac arrhythmia, Dengue, Ischemic Heart disease, Stroke, Arthritis are the top one cause of death in the world, claiming more lives than cancer and HIV combined. Thus, how to predict these diseases in real life is of great significance, both to research and application. The medical data of the patients infected by diseases is encapsulated in a dataset tabular form. The various algorithms are compared on the basis of prediction accuracy and the one that gives out highest accuracy is selected to classify the previously unknown and unseen data tuple with examples according to its class.

5.1 Proposed work

We propose an efficient prediction of diff erent Diseases based on the historical and training data. This idea is to analyze and test various data-mining models and algorithms and to implement the algorithm which gives out highest degree of accuracy. The Dataset implemented ideally contains more that 30 medical related attributes for each disease and 5000 data examples. The diseases included in the dataset are skin disease, heart disease, dengue disease, Arthritis and many more. The algorithms and classifier models used for

implementation based on the accuracy and performance on the dataset are Decision tree, Naive Bayes, Back-propagation method(ANN). The dataset is visualized in graphical diagrammatic representation using diff erent visualization techniques for user convenience and better understanding.

5.2 Issues and challenges

The difficult task for doctors and medical practitioners in medical and health care diagnosis is to make a decision based on doctors predictions about the patients suff erings and disease without considering the medical data of the patient located in the patient database. This method of practice leads to unwanted errors and sometimes in crucial stages, might even result in death of the person. The domain and areas of data mining tasks and techniques has provided a solution to this problem of medical diagnosis and prediction. The data mining techniques and data visualization operate on real time data to give out better knowledge representation and prediction.

5.3 Future Scope

In this paper the problem of medical diagnosis without considering the medical digital data is described which in return leads to unwanted errors and issues. The database proposed consists of 3000 training data samples, around 2000 test samples and more than 40 attributes. According to previous research and comparative tests on various algorithms, Back propagation and Naïve Bayes are selected for implementation on the medical dataset. The medical dataset included multiple diseases and the related attributes. Bayesian classifier model improves the quality of prediction after performing with the indulgence of genetic algorithm. The proposed work can further be enhanced for prediction of real time multiple diseases. Real time data from health care and medical organizations is to be collected and all sufficient algorithms, tasks and techniques can be compared to test the best accuracy and measures. Minute Readings and co-ordinates of the digital data like MRI , X-Ray, CT-Scan can be detected and added to the dataset with separate attributes to enhance and detail the outlines of the prediction.

6. CONCLUSION

The recent development and improvement in the data mining algorithms and classifier models have assured an ease in prediction and insights. The paper gives details about all the problems, issues and errors caused by medical diagnosis and predictions without considering the medical data of the patient. The comparative study about the algorithms proposed by the researches in previous editions is mentioned in this paper. The algorithms such as Back-propagation method and Naive Bayes classifier give out the best accuracy results when implemented on the dataset consisting of large number of data examples and labels. The proposed work can be improved and enhanced by gathering all the sensitive medical data from the health care organizations and applying algorithms and techniques on them for the best performance measures. Algorithms like complex neural network and decision tree can be evaluated on huge set of data containing more than 5000 data examples and 50 attributes for accurate prediction and performance. The paper also outlines the applications, advantages, techniques of data mining and data visualization.

7. ACKNOWLEDGMENTS

This research was extensively supported by the institution Modern Education Society's College of Engineering,Pune,India.We are thankful to Dr. Nuzhat F. Shaikh who provided expertise that greatly assisted the research and improved the manuscript significantly.

8. REFERENCES

[1] Ankita Dewan, Meghna Sharma , "Prediction of Heart Disease Using a Hybrid Technique in Data Mining Classification" , 2015 2ndInternationalConference on Computing for Sustainable Global Development (INDIACom), IEEE 2015.

[2] HICAP:Hierarchial Clustering with Pattern Preservation (2004). Hui Xiong, Michael Steinbach, Pang-Ning Tan, and Vipin Kumar, In Proc. of the Fourth SIAM International Conf. on Data Mining (SDM'04), Florida, USA, 2004.

[3] Jyoti Soni, Ujma Ansari, Dipesh Sharma, Sunita Soni , "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction" , International Journal of Computer Applications (0975 8887) Volume 17 No.8, March 2011.

[4] Abhishek Taneja. "Prediction of heart diseases using data mining techniques". Oriental Journal of computer science and technology. December 2013. Vol. 6, No. (4): Pgs. 457-466.

[5] T. Revathi S. Jeevitha, "Comparative Study on Heart Disease Prediction System Using Data Mining Techniques ",International Journal of Science and Research (IJSR) ISSN (Online): 2319-7064 Index Copernicus Value (2013)

[6] Monika Gandhi, Dr.Shailendra Narayan Singh ,"Predictions in Heart Disease Using Techniques of Data Mining" , 2015 1st International Conference on Futuristic trend in Computational Analysis and Knowledge Management (ABLAZE-2015), IEEE 2015.

[7] Amrender Kumar, "ARTIFICIAL NEURAL NETWORKS FOR DATA MINING" ,I.A.S.R.I., Library Avenue, Pusa, New Delhi-110 012.

[8] Nuzhat F. Shaikh, Dharmpal D. Doye,"An Adaptive Central Force Optimization (ACFO) and Feed Forward Back Propagation Neural Network (FFBNN) based iris recognition system", Journal of Intelligent and Fuzzy Systems 30 (2016) 20832094 DOI:10.3233, IOS Press,2083.

[9] R. Spence, L. Tweedie, H. Dawkes, and H. Su, "Visualization for functional design", in Proc. Int. Symp. on Information Vi- sualization (InfoVis 95), 1995, pp. 410

[10] Nuzhat F. Shaikh, Dharmpal D. Doye, "Improving the Accuracy of Iris Recognition System using Neural Network and Particle Swarm Optimization"International Journal of Computer Applications (0975 – 8887)Volume 79 – No3, October 2013