International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 05 Issue: 04 | Apr-2018 www.irjet.net p-ISSN: 2395-0072

DATA MINING TECHNIQUES IN MEDICAL SECTOR

S. Arockia Panimalar[1]

[1]Assistant Professor, Department of BCA & M.Sc SS, Sri Krishna Arts and Science College, Tamilnadu

---

Abstract: *Cutting edge open-source data mining suites of today have made some amazing progress from where they were just 10 years prior. New techniques are required attributable to the size and unpredictability of data accumulations in various dimensions like: organization, business and science. Today is an era of diseases where they are the major causes of death. The rising growth of medical problems has given popularity to use data mining techniques. Data mining has potentially improved the clinical decisions and survival time of patients[2]. Choosing the appropriate data mining technique is the main task because accuracy is the main issue. Earlier diagnosis done was based on doctor's experience or expertise but still wrong cases were reported. The objective is to give an exposure on variety of data mining techniques, so that the researchers can have direction to research on incurable diseases which are costliest diseases so as to save money and lives of the patient.*

Key Words: *Data Mining Techniques, Decision Trees (DTrees), Artificial Neural Network (ANN), K-Nearest Neighbour (KNN), Naive Bayes and Support Vector Machine.*

1. INTRODUCTION

Data mining is the Knowledge Discovery in Databases (KDD) by analyzing the large volumes of data from various perspectives and summarizing it into useful information. Data mining should be possible on data which are in quantitative and multimedia forms[2]. Data mining is an extraction of knowledge from dataset. To make the data reasonable in human structures is the principle target of data mining. Variety of data mining techniques along with a detailed description about their respective objective, need, history, advantages, disadvantages, application areas and role in medical or healthcare field.

2. DATA MINING TECHNIQUES

Varieties of data mining techniques are available to work on different dimensions of research. Medical healthcare is the one boon research area where data mining is proved beneficial. Motive is to give a detailed view on popular data mining techniques to the researchers so that they can work more exploratory.

2.1 Decision Trees (DTrees)

Decision tree works on divide and conquer algorithm which is easy to implement. It suffers from the problem of repetition and replication[4].

A. Need

i. Easy understanding of discovered knowledge is required.

ii. Guiding data may hold errors.

iii. Guiding data may include missing attribute values.

B. Characteristics

i. Able to handle both numerical and categorical data.

ii. Able to handle multi-output problems.

iii. Able to generate understandable rules.

iv. Can be combined with other decision techniques.

v. Runs fast even with lots of observations and variables.

vi. Use a white box model.

C. Limitations of Dtree

i. Computationally expensive to train.

ii. Complex calculations if many values are uncertain.

iii. Learning an optimal decision tree is known to be NP complete.

iv. Less appropriate for estimation tasks where the goal is to predict the value of a continuous attribute.

D. Application Areas

DTree locates its applications in the areas like:

(i) Astronomy (ii)Biomedical engineering (iii) Manufacturing and production (iv) Logistics planning (v) Computational biology (vi) Data mining (vii) Control Systems (viii) Pharmacology (ix) Medicine (x) Plant diseases

E) Role of Decision Trees in Medical

There is a need to cover an uncertainty in medical where,

i. Laboratories report comes about just with some level of mistake[1].

ii. Physiologists do not see accurately how the human body functions.

iii. Medical scientists can't absolutely portray how diseases adjust the ordinary working of the body.

2.2 Artificial Neural Network (ANN)

Artificial neural networks are basically inspired from biology where interconnection of artificial neurons processes information. It consists of three layers as Input layer, Hidden layer and Output layer.

A. Need

In order to save more time and money in any effort related to computers and robots, parallel processing is one of the vital role playing processing technique in today's scenario. In literature, many successful applications solved by conventional approaches can be found in convincing well constrained settings, none is sufficiently flexible to achieve good results outside its domain. Frequent attempts to build up intelligent programs based on von Neumann's centralized architecture have not resulted in general-purpose intelligent programs.

B. Characteristics

It possesses the following characteristics:

i. Can be utilized to separate patterns and distinguish patterns that are excessively mind boggling.

ii. Can be utilized to give projections given new circumstances of intrigue and reply "imagine a scenario where" questions.

iii. Can make its own particular association or portrayal of the information it gets amid learning time.

iv. Can be prepared to tackle certain issues utilizing a teaching method and test data.

v. Can learn from experience rather than being explicitly programmed with rules like in conventional artificial intelligence.

vi. Utilized as an arbitrary function approximation tool to assess the most practical and perfect strategies for touching base at solutions while characterizing computing functions or distributions.

vii. It is an iterative learning process in which data tests are exhibited to the system each one in turn, and the weights are balanced keeping in mind the end goal to foresee the right class name.

C. Limitations

i. No single standardized paradigm available

ii. Experimental nature of model development

iii. Not a daily life general purpose problem solver

iv. Requires high processing time for large neural networks

D. Application Areas

ANN finds its applications in the areas like:

(i) Data validation (ii) Forecasting (iii) Pattern Classification (iv) Prediction/Forecasting (v) Industrial Process Control (vi) Optimization (vii) Prediction of stock price index (viii) Regression analysis (ix) Risk management (x) Target marketing

E. Role of ANN in Data Mining

Over the last few years, applications of neural networks to financial forecasting have been admired because of their ability to extract precious information from an accumulation of history information. ANNs are also used to find patterns in the data and to infer rules from them.

F. Role of ANN in Medical

Data mining techniques have been widely used in diagnostic and health care applications because of their predictive power. Data mining algorithms can learn from past examples in clinical data and model the oftentimes non-linear relationships between the independent and dependent variables. Artificial neural networks have also been used to detect and diagnose a number of cancers like: lung cancer, prostate cancer and colorectal cancer [5]. with more accuracy than the current clinical methods. Neural networks are used experimentally to model the human cardiovascular system through which harmful medical conditions can be detected at an early stage and thus make the process of fighting the disease much easier. With time the size of stored data increases, ANNs play an important role in finding the patterns and extracting knowledge to provide better patient care and effective diagnostic capabilities.

2.3 K-Nearest Neighbour (KNN)

KNN should be the foremost choices for a classification study when there is modest or no previous knowledge about the distribution of the data. KNN is also known with different names are Instance-Based Learning, K-Nearest Neighbours,

Lazy Learning, Memory-Based Reasoning.

A. Need

Earlier the rote classifier was exercised to memorize the entire training data and perform the classification only if the attributes of a test instance matches exactly with one of the training examples.

B. Characteristics

i. Easy to implement and debug.

ii. k is usually chosen as an odd number if the number of classes is 2.

iii. Robust with regard to the search space.

iv. Some noise reduction techniques work only for k-NN.

v. Sensitive to the local structure of the data.

C. Limitations

i. Sensitive to irrelevant or redundant features.

ii. Time to find the nearest neighbours in a large training set can be too expensive.

D. Application Areas

KNN finds its applications in the areas like:

(i) Content Retrieval (ii) Databases (iii) Data compression (iv) Data mining (v) Extrapolation (vi) Intrusion detection (vii) Prediction and Forecasting (viii) Pattern recognition (ix) Online marketing

E. Role of KNN in Data Mining

Web mining is a significant concern in data mining as well as other information process techniques to discover useful patterns. Web mining can be divided into three categories such as content mining, usage mining, and structure mining. KNN is used to categorize and classified the documents present on the web. NN forecasts the class of a new document using the class label of the closest document from the training set. KNN has been applied successfully in mining the vital information regarding stock market which includes revealing the market trends, planning investment policies, discovering the best time to purchase the stocks, etc.

F. Role of KNN in Medical

KNN has been widely used in the medical field to:

i. Predict whether a patient (hospitalized due to a heart attack), will have a second heart attack.

ii. Predict the solvent accessibility in protein molecules.

iii. Reduce the wrong diagnosis and treatment in clinical diagnosis.

iv. Train the ECG database for better accuracy.

2.4 Naive Bayes

Naive bayes is an exceptional form of bayesian network that is extensively used for classification and clustering, but its prospective for all-purpose probabilistic modelling continues to remain as unexploited. The naive bayes classification uses the following different distributions for different features as Normal (Gaussian), Kernel, Multinomial and Multivariate multinomial.

A. Need

In spite of the truth that naive bayes generally over estimates the probability of the selected class, the decision making is accurate and thus the model is precise. Multinomial naive bayes is used when the manifold events of the words matter a lot in the classification problem. Naïve bayes is one of the simplest density estimation methods from which customary classification methods in machine learning can be made.

B. Characteristics

i. Empirically successful.

ii. Entire covariance matrix need not to be calculated.

iii. Easy to program and intuitive.

iv. Easy to deal with missing attributes.

v. Fast to train and to use as a classier.

vi. Inference is cheap.

vii. Supports drill through.

viii. Supports the use of OLAP mining models.

C. Limitations

i. Strong feature independence assumptions

ii. Low in accuracy

D. Application Areas

ANN finds its applications in the areas like:

(i) Probabilistic Reasoning (ii) Document Classification (iii) Medical Diagnosis (iv) Healthcare (v) Marketing (vi) Object Tracking and Recognition (vii) System Performance Management (viii) Clinical Applications (ix) Spam Filtering

E. Role of Naïve Bayes in Data Mining

Naive Bayes is one of the most efficient and effective inductive learning algorithms for machine learning and data mining. Improving the predictive accuracy and achieving dimensionality reduction for statistical classifiers has been an active research area in data mining. A straight-forward method to deal with the indecisive information is to effectively convert the uncertain data objects to deterministic point-valued data.

F. Role of Naïve Bayes in Medical

Automated medical diagnosis helps the doctors to calculate the correct disease with less time. With the help of the dataset, the patterns considerable to the disease prediction are extracted. A tree-like Bayesian network classifier algorithm can be developed for medical decision making problems. Naïve bayes could be used to reduce medical errors, enhance patient safety, decrease unwanted practice variation, and improve patient outcome. In view of client answers, naive bayes can find and extract hidden knowledge (patterns and relationships) related with a specific disease from a historical disease database.

2.5 Support Vector Machine (SVM)

SVMs are based on the notion of decision planes that characterize the decision boundaries to analyse the data and recognize the patterns used for classification and regression analysis[3]. A SVM algorithm has turned out to be acclaimed in light of the fact that it gives accuracy practically identical

to classy neural networks with explained includes in a handwriting acknowledgement task. SVMs are among the best off-the-rack regulated learning algorithm where the core is a Quadratic Programming issue (QP), so as to isolate the help vectors from whatever is left of the preparation data.

A. Need

SVM is quickly rising as promising pattern recognition methodology because of its speculation capacity and its capacity to deal with high-dimensional input. SVMs are a genuine contender to artificial neural networks if the considered parameter prescient exactness.

B. Characteristics

i. Handle multiple continuous and categorical variables.

ii. Supports both regression and classification tasks ranking problems.

C. Limitations

i. Parameters of a settled model are hard to decipher.

ii. Do not specifically give probability estimates.

iii. Long training time.

iv. Not easy to incorporate domain knowledge.

v. Lack of transparency of results.

D. Application Areas

SVM finds its applications in the areas like:

(i) Classification of Images (ii)Computer Vision (iii) Facial Expression Classification (iv)Hand-written Characters Recognition (v) Machine Learning (vi)Medical Science to Classify Proteins (vii) Protein Classification (viii) Text and Hypertext Categorization.

E. Role of SVM in Medical

An approach in light of Support Vector Machines (SVMs) for location of Micro-Calcification (MC) groups in automated mammograms, admonishes a succeeding progression learning structure for improved execution. SVM classifier was gifted through directed figuring out how to test at each site in a mammogram whether a MC is accessible or not. Molecule swarm streamlining SVM based hybrid approach for breaking down arrhythmia construct a prescient model for cancer diagnosis[5].

3. CONCLUSION AND FUTURE SCOPE

Different data mining techniques used in healthcare industry are proposed by various researchers. Data mining is proved efficacious as accuracy is the major concern. There is a boon to data mining techniques because it helps in early diagnosis of medical diseases with high accuracy and precision. Decision trees, Naive Bayes and Artificial Neural Network are used for predicting all the diseases. Future work is to find more serious diseases which have threat to lives using more data mining techniques.

4. REFERENCES

[1] http://www.cancer.gov/cancertopics/what-is-cancer.

[2]Principles of Data Mining. Max Bramer, BSc, PhD, Digital Professor of Information Technology, University of Portsmouth, UK. ISBN-10: 1-84628-765-0.

[3]D. L. Pham, "Unsupervised tissue classification in medical images using edge-adaptive clustering," 2003.

[4] K. Kalaivani "Childhood Cancer-a Hospital based study using Decision Tree Techniques" Journal of Computer Science 7(12), ISSN: 1549-3636

[5] Shweta Kharya," Using data mining techniques for diagnosis and prognosis of cancer disease", International Journal of Computer Science, April 2012