ISSN:0975-9646

Survey on Data Mining Techniques for the Diagnosis of Diseases in Medical Domain

Parvathi I, Siddharth Rautaray

Computer Science, KIIT University
Bhubaneswar,Odisha,India

Abstract **-Data mining is the process of extracting hidden information from a large set of database and it can help researchers gain both novel and deep insights of unprecedented understanding of large biomedical datasets. Data mining can uncover new biomedical and healthcare knowledge for clinical decision making. This review first introduces data mining in general (e.g. Definition, tasks of data mining, application of data mining)and gives a brief summarization of various data mining algorithms used for classification, clustering, and association .**
Discussion is made to enable the disease diagnosis and prognosis, and the discovery of hidden biomedical and healthcare patterns from related databases is offered along with a discussion of the use of data mining to discover such relationships as those between health conditions and a disease, relationships among diseases. It further discusses about the tool that can be used for the processing and classification of data and the advantages of WEKA.

***Keywords*—KDD Process, Hybrid Technology, WEKA**

INTRODUCTION

Data mining is one of the most important steps of KDD process and it is the process of extracting hidden information from large database and transform it into understandable format by considering different perspectives.[4]

- It is the computer-assisted process of digging through and analysing enormous sets of data and then extracting the meaning of the data.
- Data mining explores data and analyses large observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner.
- It uses information from past data to analyse the outcome of a particular problem or situation that may occur in future.

We are concentrating on Healthcare databases, which have a huge amount of data but however, there is a lack of effective analysis tools to discover the hidden knowledge. In this survey we present an overview of the current research being carried out using the DM techniques for the diagnosis and prognosis of various diseases, highlighting critical issues and summarizing the approaches in a set of learned lessons. Data Mining used in the field of medical application can exploit the hidden patterns present in voluminous medical data which otherwise is left undiscovered. Data mining techniques which are applied to medical data include association rule mining for finding frequent patterns, prediction and classification. Traditionally data mining techniques were used in various domains. However, it is introduced relatively late into the Healthcare domain.[7]

I. KDD PROCESS

The KDD process consists of the process of selecting the relevant data, processing it and transforming the data into relevant information and extracting the hidden information from the data pre-processed . The KDD process is specified below.

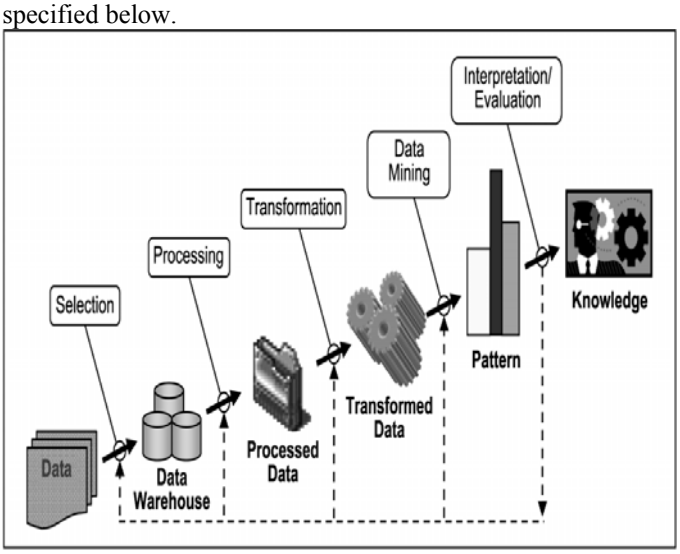

Fig 1: KDD and Data mining process

Fig 1shows the KDD process which consists of the following steps :

1. ***Selection*****:** It is the process of selecting data relevant for the task of analysis from the database.
2. ***Pre-processing***: It Removes noise and inconsistent data and combines multiple data sources.
3. ***Transformation*****:** It transforms data into appropriate forms to perform data mining.
4. ***Data mining*****:** It chooses a data mining algorithm which is appropriate in extracting patterns.
5. ***Interpretation/Evaluation*** : It interprets the patterns into knowledge by removing redundant or irrelevant data and translating the useful patterns into terms that is understandable by human.[4][3].

II. DATA MINING TASKS AND APPLICATION OF DATA MINING

Data Mining Tasks can be classified into :

- Supervised learning
- Unsupervised learning

The distinction is based on how learner classifies the data,

if the classification takes under supervision then it is supervised learning else if the classes are not known then they are unsupervised learning.

In supervised learning the classes are predetermined, the inputs are either assumed or known at the beginning.

In unsupervised learning the classes are not predetermined, the inputs are not known at the beginning and it is not carried out under supervision. The model is not provided with input data ,so it can be used to form clusters based on the statistical properties or similarities among the values

Supervised models:

Classification, etc.

Unsupervised Models:

Clustering, etc.

There are Different Data Mining tools available, which are as follows :

SPSS Clementine, SAS, E-Miner, MATLAB, Oracle DM, SQL server and Open source software such as WEKA, R, Orange.

In this section, we have focused some of the applications of data mining in respective domains:

- Data Mining Applications in Healthcare
- Data mining is used for market basket analysis
- The data mining is used as an emerging trends in the education system
- Usage of Data mining in different areas of manufacturing Engineering.
- Application of Data Mining techniques in CRM
- In language research and language engineering
- Data Mining methods are used in the Web Education
- Credit Scoring
- The Intrusion Detection in the Network
- Sports data Mining
- The Intelligence Agencies
- The Data Mining system implemented at the Internal Revenue Service.
- The Digital Library Retrieves

A. ***Data Mining Applications in Healthcare***

Data mining applications in health have tremendous potential and usefulness. However, the success of healthcare data mining hinges on the availability of clean healthcare data. In this respect, it is critical for the healthcare industry to look into how data can be better captured, stored ,prepared and mined. In health care, data Mining is used for the diagnosis and prognosis of diseases and to identify the relationship that occurs among several diseases. As healthcare data are not limited to just quantitative data ,it is also necessary to explore the use of data mining to expand the scope of what health care data mining can do.

B. ***Data mining is used for market basket analysis***

Data mining technique is can be used in MBA(Market Basket Analysis).When the customer wants to buy some products, then this technique can help us to find out the associations between different items which the customer put in their shopping cart or baskets. Here the discovery of such associations can be identified which promotes the business techniques. The retailers uses the data mining techniques to identify the customers buying pattern .In this way this technique is used for profits of the business and also helps to identify the behaviour of customers .

C. ***The data mining is used an emerging trends in the education system***

In the field of education data mining is tremendously used and is an emerging field . As every year millions of students are enrolled across the country with huge number of higher education aspirants, we believe that data mining technology can help bridging knowledge gap in higher educational systems. Data Mining helps to identify hidden patterns, associations, and anomalies from educational data and can improve decision making processes in higher educational systems. This improvement can bring advantages such as maximizing educational system efficiency, decreasing student's drop-out rate, and increasing student's promotion rate, increasing student's retention rate in, increasing student's transition rate, increasing educational improvement ratio, increasing student's success ratio, increasing student's learning outcome, and reducing the cost of system processes. In recent era we are using the KDD and the data mining tools for extracting the knowledge. The decision tree classification is frequently used in this type of applications.

D. ***Usage of Data Mining in different areas of manufacturing engineering***

When data is retrieved from manufacturing system it is used for different purposes like to find the errors in the data or product, to enhance the design methodology, to make the good quality product . The new methodology was proposed as CRISP-DM which will provides the high level detail steps of instructions for using the data mining in the engineering.

E. ***Application of Data Mining techniques in CRM***

Data mining technique is used in CRM .Now a days it is one of the hot topic to research in the industry because CRM have attracted both the practitioners and academics. It aims to give a research summary on the application of data mining in the CRM domain and techniques which are most often used. Research on the application of data mining in CRM will increase significantly in the future based on past publication rates and the increasing interest in the area.

F. ***In language research and language engineering:***

Sometimes a linguistic information is needed about a text. A linguistic profile that contains large number of linguistic features can be generated from text file automatically using data mining . This technique found quite effective for authorship verification and recognition. The linguistic profiling of text effectively used to control the quality of language and for the automatic language verification. This method verifies automatically the text is of native quality.

G. ***Data Mining methods are used in the Web Education***

Data mining methods are used in the web Education which is used to improve courseware. The relationships are discovered among the usage data picked up during student's sessions. This knowledge is very useful for the teacher or the author of the course, who could decide what modifications will be the most appropriate to improve the effectiveness of the course. Data mining techniques are one of the best learning methods. Web Education which will rapidly grow in by the application of data mining methods

to educational systems can be both feasible and enhanced in the learning process.

H. ***Credit Scoring***

Credit scoring has become very important issue due to the recent growth of the credit industry, so the credit department of the bank faces the huge numbers of consumers' credit data to process, but it is impossible to analyse huge amount of data both in economic and manpower terms. The support vector machine has been widely applied in recent years and which is one of the best technique. Since to improve the performance of this model, it is necessary a method for reduction the feature subset, many hybrid SVM based model are proposed. Many of these proposed models can only classify customers into two classes "good" or "bad" ones. The most used applied methods for doing credit scoring task are derived from classification technique. Generally classification is used when we predict something which is possible by using the previous available information. It is one type of methods which can be defined as classification where the members of a given set of instances into some groups where the different types of characteristics are to be made. Classification task is very suited to data mining methods and techniques

I. ***The Intrusion Detection in the Network***

The intrusion detection in the Network is very difficult and needs a very close watch on the data traffic. The intrusion detection plays an essential role in computer security. The classification method of data mining is used to classify the network traffic either normal traffic or abnormal traffic. If any TCP header does not belong to any of the existing TCP header clusters, then it can be considered as anomaly.

J. ***Sports data Mining :***

The data mining and its technique is used for an application of Sports centre. Data mining is not only used in the business purposes but also it used in the sports .A huge number of games are available where each and every day the national and international games are to be scheduled, where a huge number of data are to be maintained .The data mining tools are applied to give the information as and when its required. The open source data mining tools like WEKA and RAPID MINER are frequently used for sports. This means that users can run their data through one of the built-in algorithms, see what results come out, and then run it through a different algorithm to see if anything different stands out. In the sports world the vast amounts of statistics are collected for each player, team, game, and season. Data mining can be used for prediction of performance, selection of players, coaching and training and for the strategy planning . The data mining techniques are used to determine the best or the most optimal squad to represent a team in a team sport.

K. ***The Intelligence Agencies***

The Intelligence Agencies collect and analyse information to investigate terrorist activities. One of the challenges to law enforcement and intelligent agencies is the difficulty of analysing large volume of data involved in criminal and terrorist activities. Now a days the intelligence agencies are using the sophisticated data mining algorithms which makes it easy, to handle the very large databases for organizations. The different data mining techniques are used in crime data mining .Data Mining helps to generate different types of information in the organization like personal details of the persons along with the vehicle details which can help to identify terrorist activities .The Clustering techniques are used (Association rule mining) for the different objects(like persons, organizations, vehicles etc.) in crime records. The classification technique is used to detect email spamming and String comparator is used to detect deceptive information in criminal record.

L The data mining system implemented at the Internal Revenue Service

The data mining system implemented at the Internal Revenue Service to identify high-income individuals engaged in abusive tax shelters show significantly good results. Data mining can be used to identify and rank possibly abusive tax avoidance transactions. To enhance the quality of product data mining techniques can be effectively used. The data mining technology SAS/EM is used to discover the rules those are unknown before and it can improve the quality of products and decrease the cost. A regression model and the neural network model can also be used for this purpose.

M. ***The Digital Library Retrieves***

The data mining application can be used in the field of the Digital Library where the user finds or collects, stores and preserves the data which are in the form of digital mode. The data and information are available in different formats. These formats include Text, Images, Video, Audio, Picture, Maps, etc.

III. CATEGORIZATION OF DATA MINING TECHNIQUES:

Data mining techniques can be categorized according to various criteria of classification which are as follows:

- Classification according to the type of data source mined: This classification categorizes data mining systems according to the type of data handled such as spatial data, multimedia data, time-series data, text data, World Wide Web, etc.
- Classification according to the data model drawn on: This classification categorizes data mining systems based on the data model involved such as relational database, object-oriented database, data warehouse, transactional, etc.
- Classification according to the king of knowledge discovered: This classification categorizes data mining systems based on the kind of knowledge discovered or data mining functionalities, such as characterization, discrimination, association, classification, clustering, etc.
- Classification according to mining techniques used: Data mining systems employ and provide different techniques. This classification categorizes data mining systems according to the data analysis approach used such as machine learning, association, classification, neural networks, genetic algorithms, statistics, database-oriented or data warehouse-oriented, etc. The classification can also take into account the degree of user interaction involved in the data mining process such as query-driven systems, interactive exploratory systems, or autonomous systems.

Data mining is not specific to one type of media or data. Data mining should be applicable to any kind of information repository. However, algorithms and approaches may differ when applied to different types of data. Indeed, the challenges presented by different types of data vary significantly. Data mining should be carried out for relational databases, object-relational databases and object-oriented databases, data warehouses, transactional databases, unstructured and semi-structured repositories such as the World Wide Web, advanced databases such as spatial databases, multimedia databases, time-series databases and textual databases, and even flat files. Some of them are specified below.

Flat files: Flat files are actually the most common data source for data mining algorithms, especially at the research level. Flat files are simple data files in text or binary format with a structure known by the data mining algorithm to be applied. The data in these files can be transactions, time-series data, scientific measurements, etc.

Relational Databases: A relational database consists of a set of tables containing either values of entity attributes, or values of attributes from entity relationships. Tables have columns and rows, where columns represent attributes and rows represent tuples. A tuple in a relational table corresponds to either an object or a relationship between objects and is identified by a set of attribute values representing a unique key. The most commonly used query language for relational database is SQL, which allows retrieval and manipulation of the data stored in the tables, as well as the calculation of aggregate functions such as average, sum, min, max and count. For example, an SQL query to select the videos grouped by category would be: SELECT count(*) FROM Items WHERE type=video GROUP BY category. Data mining algorithms using relational databases can be more versatile than data mining algorithms specifically written for flat files, since they can take advantage of the structure inherent to relational databases.

Data Warehouses: A data warehouse as a storehouse, is a repository of data collected from multiple data sources often heterogeneous and is intended to be used as a whole under the same unified schema. A data warehouse gives the option to analyze data from different sources under the same location. There are several operations such as drill-down, roll-up, slice, dice etc in OLAP(Online Transactional Processing).

Transaction Databases: A transaction database is a set of records representing transactions, each with a time stamp, an identifier and a set of items. Associated with the transaction files could also be descriptive data for the items. Transactions are usually stored in flat files or stored in two normalized transaction tables, one for the transactions and one for the transaction items. One typical data mining analysis on such data is the so-called market basket analysis or association rules in which associations between items occurring together or in sequence are studied.

Multimedia Databases: Multimedia databases include video, images, audio and text media. They can be stored on extended object-relational or object-oriented databases, or simply on a file system. Multimedia is characterized by its high dimensionality, which makes data mining even more challenging. Data mining from multimedia repositories may require computer vision, computer graphics, image interpretation, and natural language processing methodologies.

Spatial Databases: Spatial databases are databases that, in addition to usual data, store geographical information like maps, and global or regional positioning. Such spatial databases present new challenges to data mining algorithms.

Time-Series Databases: Time-series databases contain time related data such stock market data or logged activities. These databases usually have a continuous flow of new data coming in, which sometimes causes the need for a challenging real time analysis. Data mining in such databases commonly includes the study of trends and correlations between evolutions of different variables, as well as the prediction of trends .

World Wide Web: The World Wide Web is the most heterogeneous and dynamic repository available. Data in the World Wide Web is organized in inter-connected documents. These documents can be text, audio, video, raw data, and even applications. Conceptually, the World Wide Web is comprised of three major components: The content of the Web, which encompasses documents available, the structure of the Web which covers the hyperlinks and the relationships between documents, and the usage of the web describing how and when the resources are accessed. A fourth dimension can be added relating the dynamic nature or evolution of the documents. Data mining in the World Wide Web, or web mining, tries to address all these issues and is often divided into web content mining, web structure mining and web usage mining.

IV. DATA MINING TECHNIQUES:

The most frequently used Data Mining techniques are specified below

Classification learning:- The learning algorithm takes a set of classified examples (training set) and use it for training the algorithms. With the trained algorithms, classification of the test data takes place based on the patterns and rules extracted from the training set.

Numeric predication:- This is a variant of classification learning with the exception that instead of predicting the discrete class the outcome is a numeric value.

Association rule mining:- The association and patterns between the various attributes are extracted and from these attributes rules are created. The rules and patterns are used predicting the categories or classification of the test data.

Clustering: - The grouping of similar instances in to clusters takes place. The challenges or drawbacks considering this type of machine learning is that we have to first identify clusters and assign a new instance to these clusters[8].

Out of these four types of learning methods we need to identify the algorithm which performs better.

The application of data mining techniques depends on the types of data which is fitted to be used in the techniques , and solving data mining problems depend on the types of data to be used and the selection of data mining technique which is most suitable for the data used .

- Mining association rules : The purpose of mining association rules is to find out the associations between items from a huge transactions. It is usually applied to transactional data forms. But it is also suitable for data stored in relational data forms.
- Classification: The basic purpose is to find out the classification principle from a pre-classified data set (training data). This principle can be used to classify the newly coming data. ID3, CART,C4.5, and neural network are all popular classification methods.

There are numerous data mining tools and methods available today. In this survey, we will define the new approach of using hybrid Data Mining system for medical database.

Application of hybrid Data Mining Techniques can enhance the performance and accuracy. It involves the comparisons to be made between the algorithm applied individually and the proposed hybrid techniques. Although there are a number of data mining tools available, we will be using WEKA tool to evaluate and draw to a conclusion on which is the best tool that can be used for diagnosis . Disease diagnosis is one of the applications where data mining tools are proving successful results. By analyzing few algorithms we made WEKA as a tool for implementation as it has many advantages, it implements various machine learning techniques like classification, association and regression. Data mining applications in healthcare include analysis of health care centers for better health policy-making and prevention of hospital errors, early detection, prevention of diseases and prevention of hospital deaths.

V. HYBRID TECHNOLOGY

This survey defines the proposal of building a hybrid methodology, combining data mining techniques such as association rules and classification trees. The methodology is applied to data collected from a hospital and is evaluated by comparing with other techniques. The methodology is expected to help physicians to make a faster and more accurate decisions.

Fig 2 shows the proposed approach of using hybrid technique and compares the single technique and hybrid technique.[18]

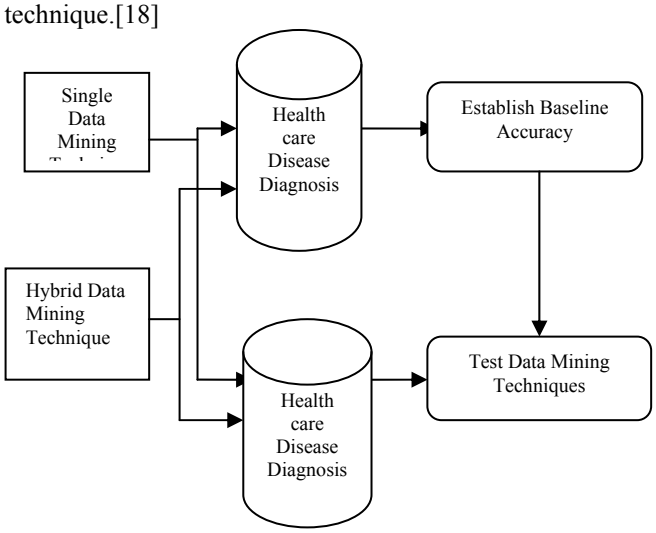

Fig 2: Proposed Approach Of Hybrid Data Mining Techniques

Data mining applications can have tremendous potential and usefulness .However, the success of data mining hinges on the availability of clean data. In this respect, it is critical that the industry look into how data can be better captured, stored ,prepared and mined.[9][10]

In this section, we have focused some of the applications of data mining in respective domains:

VI DATA MINING TASK PRIMITIVES

A data mining task can be specified in the form of a data mining query, which is input to the data mining system. A data mining query is defined in terms of data mining task primitives. These primitives allow the user to interactively communicate with the data mining system during discovery in order to direct the mining process, or examine the findings from different angles or depths.

The set of *task-relevant data* to be mined: This specifies the portions of the database or the set of data in which the user is interested. This includes the database attributes or data warehouse dimensions of interest.

The *kind of knowledge* to be mined: This specifies the *data mining functions* to be performed, such as characterization, discrimination, association or correlation analysis, classification, prediction, clustering, outlier analysis, or evolution analysis.

The *background knowledge* to be used in the discovery process: This knowledge about the domain to be mined is useful for guiding the knowledge discovery process and for evaluating the patterns found.

Concept hierarchies are a popular form of background knowledge, which allow data to be mined at multiple levels of abstraction.

The *interestingness measures and thresholds* for pattern evaluation: They may be used to guide the mining process or, after discovery, to evaluate the discovered patterns. Different kinds of knowledge may have different interestingness measures. For example, interestingness measures for association rules include *support* and *confidence*. Rules whose support and confidence values are below user-specified thresholds are considered uninteresting.

The expected representation for visualizing may include rules, tables, charts, graphs, decision trees, and cubes.

VII. DISCUSSION ON PAPERS

This survey presents a systematic review of the application of Data Mining methods in healthcare domain, with a focus on the application and the techniques used which will optimize the results. These methods are new approaches to solve the problems in healthcare domain. In this literature Survey we present an overview of the current research being carried out using the data mining techniques for the diagnosis and prognosis of various diseases. The following algorithms have been identified: Decision Trees, Support Vector Machine, Artificial neural networks and Naïve bayes. Analysis show that it is very difficult to name a single data mining algorithm as the most suitable for the diagnosis and/or prognosis of diseases. At times some algorithms perform better than others, but there are cases when a combination of the best properties of some of the aforementioned algorithms together results more effective.

A hybrid technology can be developed with the combination of properties of most well performing algorithms. The Table 2.1 below shows the survey made in various techniques used in data mining for the medical diagnosis. And from the below specified techniques, most of the domain are concentrating on classification and association rule mining.

From Table 1 we have noticed that classification is the most frequently used Data Mining Techniques for the diagnosis of Diseases, Classification is used to correctly classify the diseases and the symptoms which can be useful to make decisions for the unknown diseases with the symptoms already classified.

TABLE 1
SURVEY ON DIFFERENT DATA MINING TECHNIQUES

Author	Year	Knowledge Type	Knowledge Resource	DM techniques /applications
Smt Girija D.K [1]	2013	Fibroid diagnosis and prognosis	Tangra tool is used to predict the occurrence of fibroid	Classification: C4.5,ID3, Naive Bayes
Mai shouman	2012	Heart disease diagnosis	Hybridizing more than one techniques to show the enhanced results and accuracy	Classification, Naive Bayes, Decision tree, neural network
Rahul Isola	2011	Automated differential diagnosis in medical system	computes the probability of occurrence of disease to enhance the accuracy and it concentrates on root diseases	Classification, KNN ,Hopfield neural network, SOM
B. M. Patil	2010	Health care Association rule on numeric data	Discretize the continuous valued function into categorical values	
Anjana Gosain et.al	2008	Analysis of health	Analysis of health which predict the occurrence of route transmission based on treatment history of HIV patients. HIVassesses the utilization of Healthcare resources and demonstrate the socioeconomic, and medical histories of patient care using different	Classification: Decision tree, Association
Lavrac et.al	2007	Health care	Public Health Data The health-care provides database Classification – C4.5.The out-patient health-care statistics Database The medical status database **MediMap**:Visualization& database Detection of Outliers .	Classification: C4.5,clustering, agglomerative classification
Subbulakshmi C.V	2012	Comparative analysis of tools	Comparative analysis of XLMiner and WEKA for pattern classification is carried out for heart disease data using open source software packages specified ,WEKA produced good results however the XLMiner execution time is less	Classification:J4.8
Al Jarullah, A.A	2011	Decision tree diagnosis of type 2 diabetes	Weka's J48 decision tree classifier was applied to the modified dataset to construct the decision tree model. The accuracy of the resulting model was 78.1768%.	Classification: J4.8

In paper[1] it describes about the fibroid diagnosis and prognosis and describes about 3 types of Fibroid which can primarily classified as:

- Subserosal
- Intramoral
- Submucosal

Tanagra Tool is used as a Data Mining Tool for classification and diagnosis of Fibroid, the suitable Data format for Tanagra tool is text format and ARFF format.

The Table 1 also specifies the hybrid technique used to enhance the accuracy and performance , also discuses about the Data mining technique WEKA which is used frequently for the diagnosis of diseases and a comparative analysis between XLMiner and WEKA is carried out and the result has been obtained which specifies that the accuracy of WEKA is better in comparison to XLMiner but the time required to obtain the result is better in XLMiner with some fraction of seconds. WEKA generates the results which specifies the number of correctly and incorrectly classified set of attributes and generates a decision tree for classified sets of attributes

Categories Of Medical Domain Data Sets

There are several Data sets available in the field of Medical Data Mining ,few categories are specified below.

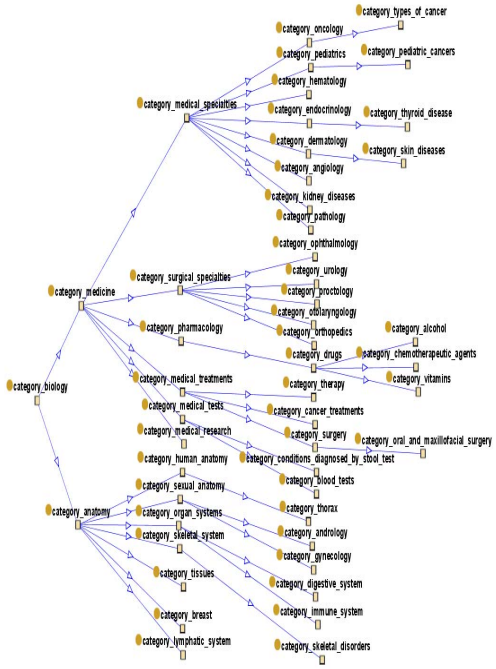

Fig 3: Classification of data sets present in medical domain

VIII. ADVANTAGES AND DISADVANTAGES OF DATA MINING IN MEDICAL DOMAIN:

Advantages:

Data mining can be advantageous as it can help

- Healthcare insurers detect fraud and abuse.
- Physicians identify effective treatments and best practices.
- Patients receive better and more affordable healthcare services.
- One of the key advantages of using data mining is their speed in working with large data sets. Generation of quicker report and faster analysis that can increase operational efficiency and reduce operating cost.
- Data Mining can extract predictive information from large databases which is a very important feature of Data Mining.

Disadvantages:

- Heterogeneity of medical data Volume and complexity

Physician's interpretation Poor mathematical categorization

- Ethical, Legal and Social Issues
- Data Ownership Lawsuits
- Privacy and Security of Human Data Administrative Issues
- Privacy issues
- Security issues
- Misuse of Information/inaccurate information

IX. ADVANTAGES OF USING WEKA :

WEKA is a workbench for machine learning which is intended to aid in the application of machine learning techniques to a variety of real-world problems. Unlike other machine learning projects, this emphasizes on providing a working environment for the domain specialist rather than the machine learning expert.[22]

- WEKA is a computer program that was developed at the University of Waikato[3] .
- WEKA supports data mining tasks such as data pre-processing, classification, clustering, regression, visualization and feature selection.
- It is an open source that can be modified by the users according to their requirements.

X. DISCUSSIONS:

Data Mining extracts hidden information from large set of database and has been of great importance, its importance is increasing day-by-day as the data is gathered systematically in almost every field and we need a technique to intelligently extract hidden information. The application of Data Mining techniques is not straight forward .In order to apply Data Mining techniques one has to understand the nature of data. The most important step involves pre-processing of data, if the pre-processing is not carried out properly then the entire decision making process may go wrong. The pre-processing involves data cleaning, data Integration and transformation of data into understandable format and size reduction, this improves the quality of pattern.

In the above section, we have discussed about the different Data Mining techniques, tasks that are used for the diagnosis of diseases and among those classification is the most frequently used algorithm as it has many advantages in comparison to other algorithms and helps in better classification. The advantages and disadvantages of Data Mining techniques have been discussed above and WEKA tool has been proposed for classification of data sets, initially a hybrid approach is being specified which can enhance the accuracy of the algorithm. WEKA is used as a tool in this work, as it is of great importance while

comparing with other algorithms, it gives a better accuracy and is an open source tool which can be modified by users according their requirements. The survey proposes to build a hybrid data mining model to extract classification knowledge for aid of various disease in clinical decision system. Future research should imply a detailed study of adopted hybrid data mining approach, combining an association rule mining and a classification tree mining. However, another combination of data mining techniques is still possible to accomplish the same task of medical health Informatics. This survey presents a framework of the tool to be used for analysis, a combination of algorithms or enhanced version of an algorithm can be used to overcome the disadvantage of improper pre-processing of data. The section of survey specifies the tools and the techniques that are used for different data mining datasets, in medical domain many algorithms have been used and classification is the most frequently used algorithm as the medical datasets need to be classified and the survey defines the type of knowledge for which the data mining algorithm is applied. To enhance the accuracy of the algorithm we need to calculate the number of correctly classified and incorrectly classified sets of data with the accuracy matrix for performance evaluation.

XI. CONCLUSIONS:

Data mining can be beneficial in the field of medical domain .However privacy, security and misuse of information are the big problems if they are not addressed and resolved properly. This survey describes about the proposal of hybrid data mining model to extract classification knowledge for aid of various disease in clinical decision system and presents a framework of the tool various tools used for analysis .It describes about the advantages and various issues faced by data mining technique and a description of various algorithms that has been applied in the field of medical diagnosis. The accuracy of the algorithms can be enhanced by hybridizing or combining algorithms or their most prevalent features, as a single algorithm may not be accurate for weakly classified sets of data.

FUTURE WORK

Our future work will involve the combination of the above two specified algorithms to enhance the accuracy so that the diagnosis can become more accurate in case of weakly identified data sets. The data sets which cannot strongly identified the classes of diseases.

REFERENCES

[1] Girija, D.K.S.; Shashidhara, M.S., "Data mining techniques used for uterus fibroid diagnosis and prognosis," *Automation, Computing, Communication, Control and Compressed Sensing (iMac4s), 2013 International Multi-Conference on* , vol., no., pp.372,376, 22-23 March

[2] Mon-Fong Jiang; Shian-Shyong Tseng; Shan-Yi Liao, "Data types generalization for data mining algorithms," *Systems, Man, and Cybernetics, 1999. IEEE SMC '99 Conference Proceedings. 1999 IEEE International Conference on* , vol.3, no., pp.928,933 vol.3, 1999

[3] Venkatadri.M and Dr. Lokanatha C Reddy. Article: A Review on Data mining from Past to the Future. *International Journal of Computer Applications* 15(7):19–22, February 2011

[4]http://www.ijser.org/researchpaper%5CA-survey-on-Data-Mining-Tools-Techniques-Applications-Trends-and-Issues.pdf

[5] Huang, Yin-Fu, and Chiech-Ming Wu. "Mining generalized association rules using pruning techniques." *Data Mining, 2002. ICDM 2003. Proceedings. 2002 IEEE International Conference on*. IEEE, 2002

[6] Amin, S.U.; Agarwal, K.; Beg, R., "Genetic neural network based data mining in prediction of heart disease using risk factors," *Information & Communication Technologies (ICT), 2013 IEEE Conference on* , vol., no., pp.1227,1231, 11-12 April

[7] Khaleel et al.," A Survey of Data Mining Techniques on Medical Data for Finding Locally Frequent Diseases"*International Journal of Advanced Research in Computer Science and Software Engineering* 3(8), August - 2013, pp. 149-153

[8] Gosain, A.; Kumar, A., "Analysis of health care data using different data mining techniques," *Intelligent Agent & Multi-Agent Systems, 2009. IAMA 2009. International Conference on* , vol., no., pp.1,6, 22-24 July 2009

[9] P Mishra; N Padhy; R Panigrahi " The Survey of Data Mining Applications And Feature Scope", *International Journal of Scientific & Engineering Research* Volume 4, Issue3, March-2013 8,ISSN 2229-5518

[10]http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/datamining.htm

[11] Silwattananusarn, Tipawan, and Kulthida Tuamsuk. "Data Mining and Its Applications for Knowledge Management: A Literature Review from 2007 to 2012." *arXiv preprint arXiv:1210.2872* (2012)

[12] Shouman, M.; Turner, T.; Stocker, R., "Using data mining techniques in heart disease diagnosis and treatment," *Electronics, Communications and Computers (JEC-ECC), 2012 Japan-Egypt Conference on* , vol., no., pp.173,177, 6-9 March 2012

[13] Robu, R.; Hora, C., "Medical data mining with extended WEKA," *Intelligent Engineering Systems (INES), 2012 IEEE 16th International Conference on* , vol., no., pp.347,350,13-15June2012

[14] Stewart, A.; Herder, E.; Smith, M.; Nejdl, W., "A user study on public health events detected within the medical ecosystem," *Digital Ecosystems and Technologies Conference (DEST), 2011 Proceedings of the 5th IEEE International Conference on* , vol., no., pp.127, 132,May 31 2011- June 3 2011

[15] http://docs.oracle.com/cd/B28359_01/datamine.111/b28129/process .htm

[16] Larose, D. T., "Discovering Knowledge in Data: An Introduction to Data Mining", ISBN 0-471-66657-2,John Wiley & Sons, Inc, 2005

[17]http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.67.4912&rep=rep1&type=pdf

[18] Elshazly, Hanaa; Azar, Ahmed Taher; El-korany, Abeer; Hassanien, Aboul ella, "Hybrid system for lymphatic diseases diagnosis," *Advances in Computing, Communications and Informatics (ICACCI), 2013 International Conference on* , vol., no.,pp.343,347,22-25Aug.2013

[19] Subbulakshmi, C.V.; Deepa, S. N.; Malathi, N., "Comparative analysis of XLMiner and WEKA for pattern classification," *Advanced Communication Control and Computing Technologies (ICACCCT), 2012 IEEE International Conference on* , vol., no., pp.453,457, 23-25 Aug. 2012

[20] Al Jarullah, A.A., "Decision tree discovery for the diagnosis of type II diabetes," *Innovations in Information Technology (IIT), 2011 International Conference on* , vol., no., pp.303,307, 25-27 April 2011

[21] Mallios, N.; Papageorgiou, E.; Samarinas, M., "Comparison of Machine Learning Techniques using the WEKA Environment for Prostate Cancer Therapy Plan," *Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE), 2011 20th IEEE International Workshops on* , vol., no., pp.151,155, 27-29 June 2011

[22] WEKA, by university of Waikato, http://www.cs.waikato.ac.nz/ml/weka/

[23] Salama, G.I.; Abdelhalim, M.B.; Zeid, M.A., "Experimental comparison of classifiers for breast cancer diagnosis," *Computer Engineering & Systems (ICCES), 2012 Seventh International Conference on* , vol., no., pp.180,185, 27-29 Nov. 2012