International Journal of Engineering and Techniques - Volume 4 Issue 2, Mar-Apr 2018

RESEARCH ARTICLE OPEN ACCESS

Literature Review of various Mining Techniques Used in Medical Sector

Ruchika[1], Maneet Kaur[2]

1(M.tech scholar, Lovely Professional University Jalandhar)

2(Assistant Professor, Lovely Professional University Jalandhar)

Abstract:

Currently generations are so abundantly busy in their everyday life routines. They suffer from lots of physical as well mental diseases. Today generation agonize more from mental diseases. Various novel approaches, Data mining and classification techniques are used to diagnosis these diseases and ration their accuracy level. Numerous tools are used for measuring the accuracy, performance and for refining the qualities of facilities in the medical field. In this paper we analysis some of medical field related research papers, their tools and techniques.

Keywords — **Data Mining, Classification, SVM, K-Mean, KNN, clustering**

I. OVERVIEW

Now days, the early generation suffer more from the mental disease. Due to their stressed full busy schedule, heavy workload, social pressure and family issues. Due to this they suffer more mentally pressure and that cause numerous type of mental disorders like they agonize from hypertension, personality disorder, behavioural disorder and autistic spectrum etc.

II. INTRODUCTION

A. *Data Mining*

Data mining is a process of mining the hidden, unknown patterns from the large data base. It helps to discover the patterns that help to select the other new fact and figures that will practice for increasing the throughput of the any organization or firms. It is also called KDD Process. KDD stand for Knowledge discovery in the database. By using data mining techniques we get relevant and useful data which help in making better decision for any organization

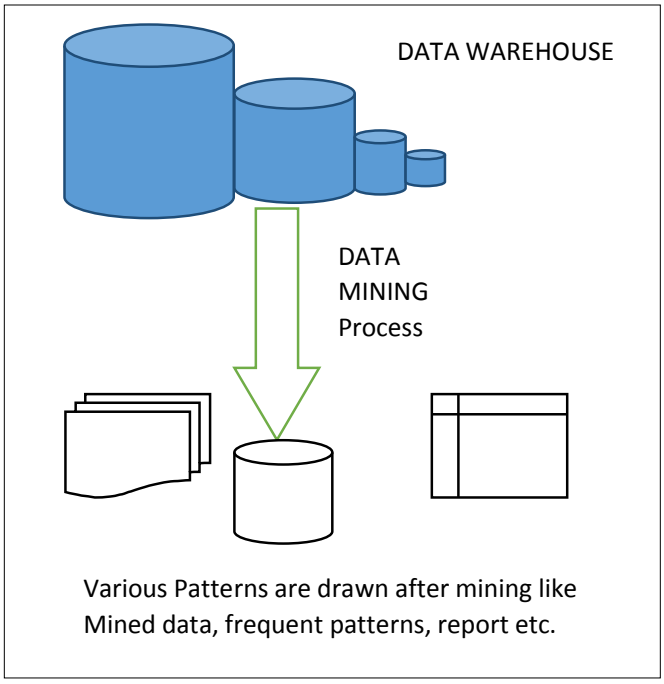

Fig 1 DATA MINING BASIC DIAGRAM

B. *KDD Process*

The procedure of mining the new patterns from the large data base are done in numerous steps, that are explain as below as are:

1) ***Data Cleaning:*** In this step remove the inconsistency, miss

value, Nullvalue, noisy data and irrelevant data from the choose data mart or data warehouse.

2) ***Data Integration***: Integrate the data from the various databases, data marts, data files and from the data warehouse and from other numerous resources.
3) ***Data Transforming***: Transform the data into required data format according to the organization needed format. This step is used to consolidate the data for the mining process by applying the summarization or aggregation operations on it.
4) ***Data Mining:*** In order to extract the new patterns numerous data mining techniques are applied for mining the hidden pattern.

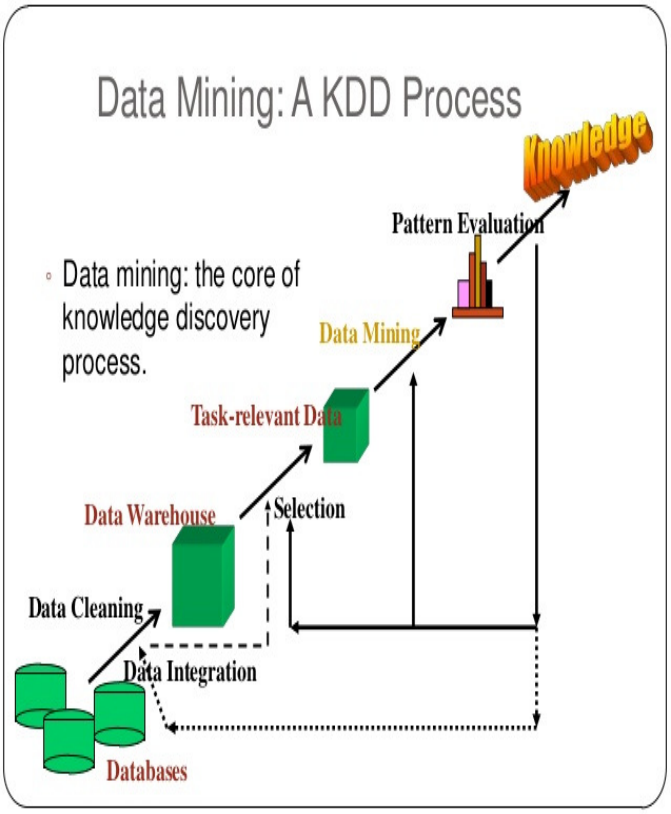

Fig 2. DATA MINING P/KDD PROCESS

5) ***Pattern Evaluation:*** Mined pattern are evaluate after the mining, whether it is valid or invalid.

6) ***Knowledge presentation***: After all these process discovered knowledge is represented in order to increase the productivity of any organization or firm.

III. Common Classes of Data Mining

Some of common classes used in data mining for doing the mining task are:

1 ***Anomaly detection:*** It detects the various anomalies and removes it by applying normalization on data.
2 ***Classification:*** It classifies the data into various predefine classes and structured the new data into these classes.
3 ***Clustering:*** It clusters the similar type of new data pattern in the structured class or cluster. It cluster similar and dissimilar data sets
4 ***Regression***: It calculates the relationship between the data elements and models the data to reduce the errors in the mined data.
5 ***Association learning rules:*** It finds the relationship between the data like finding the new combination etc. Sometime it's also called business bucket analysis.
6 ***Summarization:*** It provides the more compact representation to the mined data. Generate visualization, report generation, diagram and charts etc.

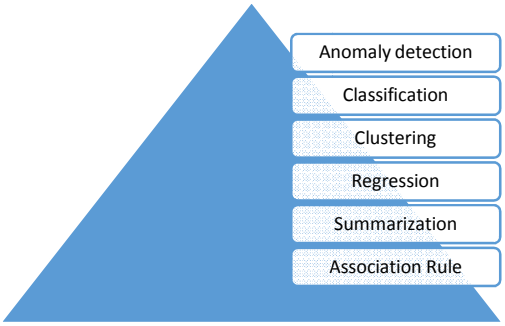

Fig 3. COMMAN CLASSES OF DATA MINING

IV. VariousChallenges In Data Mining

There are many challenges in data mining field are discussed here:

1. Clustering of large data sets in accurate cluster and put similar one in and cluster and dissimilar one in other cluster is some time becomes a big challenge.
2. Classification of data sets to its appropriate class is also difficult because some of data set neither nor belong to any class.
3. Predicting the relation between data is difficult to associate even correlate.
4. After mining validating the needed data become more difficult and sometime some important fact are missed due to mining.

V. LITERATURE SURVEY

Priyanka Dhaka et al. (2016) [1] do survey that Mental health fundamentally degree with high rate of depression, disorder and type of disorder that the humanoid is affected. They use genetic algorithm and big data tool MongoDB for analysis purpose. Genetic algorithm yield optimized result by applying random operations on data and MongoDB is used for together processing, searching and storing the data. Genetic algorithm give optimum solution .Consequence of these applied technique recommend the superior cure of Mental disorder and support doctors to give healthier treatment to their patient with more polished information, in a lesser amount of time and cost.

Lijun Hao,Shumin et al. (2016) [2] they work on open stack and cloud computing, Hadoop technology used for better improvement in the medical fields. Some of most popular data mining tools are used like R language, SQL, Excel, SAS etc. They also normalized the data and remove incomplete, redundant data and make various classes of it.

By using these tools and various mining algorithm they integrate the various most famous analytical mining tools and compare their results. They also configure the personal working platform and reduce cost for device. They configure the virtual machine concept for processing big data related to medical field.

Miroslav Bursa et al. (2011) [3], they use Novel nature inspired techniques for mining the loosely structured medical data. They deploy Ant colony optimization (ACO) approach for loosely structured data records and for clustering they use DTree Method. This approach discovered the shortest path between the large dataset. Output that structure show the further processing.

By using this approach it automatically grouped the relevant literals and makes their clusters. It will increase the speed of loosely structured texted attribute and construct a lexical analysis grammar for comparing the classical methods. Speedup can allow performing more and more iteration loosely structured data for making it better and efficient for processing.

GuiduoDuan et al. (2016) [4] they work on FP-Growth algorithm (Extended Prefix-Tree Structure) for correlating the medical insurance data cost and relevant factors. They use pre –pruning and post- pruning process and classification of data based on three criteria like accuracy, stability and complexity. FT-Growth algorithm work on the concept of Divide- and –Conquer method which compress the frequent data items into smaller data structure called FT-Tree.

MDMP (Medical Decision Model with Pruning) use pruning concept in decision tree construction. They use FT-Growth algorithm for extracting the dataset and then calculate their information gain value. According to that value they decide whether to the discard dataset or not. For future work, they will focus on how to improve the accuracy of classification.

B.V. Kiranmayee et al. (2016) [5], they focus on brain tumour detection. They purposed a methodology which detects the brain tumour on training and testing phases. They purposed an algorithm that classifies the images as tumour prone or not. They work on ID3 (Decision Tree Builder) algorithm for detection purpose and also classification prototype application datasets for health care. They implement it in Java for making decision tree classification prior to diagnosis of brain tumour deceases.

The purposed algorithm can also check the quality of image that is used as for detecting the brain tumour. ID3 algorithm can integrated with real world health care software products.

Hengyi Hu et al. (2017) [6], they work on modular ontologies, authoritative medical ontologies (AMOs), association rules and apriori algorithm. Ontologies represent the highly detailed related to knowledge domain. Its main goal is to describe a method for capture the existing symptom for depression and their related drugs associated with their successful recovery. This approach has two benefits: A) Make assumption regarding to existing patient data. B) Reuse the domain knowledge for discovering the new knowledge.

They work on similarity functions and Semantic Web Rule Language (SWRL) rules in Protégé. Apriori algorithm is used to mine the frequent data set and their association rules. For further, electrical medical record can be mined by using multi-agent system. Those automatically update the ontologies and automate the data entry.

RaziehAsgarnezhad et al. (2017) [7], they work on efficient preprocessing techniques for replacing the missing and selecting the well-known data set for diabetes mellitus. They firstly remove the missing values with Mean, Median, and KNN. Then selection method can include forward selection and backward eliminate brute force and evolutionary techniques. They work on SVM (Support Vector Machines) for predicting the classification and for optimization they use Genetic algorithm (GA).

This will increase the accuracy and precision result for predictive model. GA also improves the performance.

ParisaNaraei et al. (2016) [8], they compare two different algorithms i.e. multilayer perception neural networks and SVM for heart disease detection. SVM is a statistical learning theory which is used for classification. They use WEKA tool for classification and replacing the missing value by using filter in WAKE tool. Back propagation algorithm used in neural network.

It boosts the accuracy and pre-processing techniques process. The results are

comparable with other studied algorithms that are: Neural Networks, Decision Tree, and Naïve Bayes for same data sets. Measure their accuracy performance for those data sets.

HoussemTurki et al. (2014)[9], they work on Knowledge Discovery Data (KDD). They deploy the Bayesian Network (BN) and Dynamic Bayesian Network (DBN) for temporal Knowledge Discovery Data (KDD). BN represent the prior distribution of random variables.

A previous purposed method has limitation of initialization problem. But use of BN can reduce the number of calculations which leave only useful variables for prediction.

They work to achieve the main objective to develop the incremental algorithm for DBN (Dynamic Bayesian Network) structure for:

a) Large number of pragmatic data.
b) Heterogeneous expert of knowledge.
c) State art for contributing in the algorithm developed for learning structure.

Purposed algorithm main goal is to provide the incremental structure learning algorithm for a system to predict the mental retardation in DS children. It also increases the accuracy of diagnosis.

Sneha Chandra et al. (2015) [10] [11], they deploy the enhancement in classification accuracy by using adaptive classifier using various image processing and existing classification algorithms. They work on Bayesian Classification, Decision Tree classification, Ensemble classification, Laplacian correction, Euclidean distance, K-Mean clustering, Mean Grey level algorithm and Rule-Based classification.

Their main objective is to generate the more certain, precise and accurate result. Classifier used for prediction purpose and for increasing the classification accuracy. They use Bagging as the collaborative method for enlightening the classification accuracy.

Higher accuracy can be achieved by AC (Adaptive Classifier but it is still insufficient to achieve 100% accuracy.

Shubpreet Kaur et al. (2015) [12], they work on KDD and WEKA tool is used for measuring the Drug addiction.

Their main objectives are:

a) Generate the efficient way to extract the meaningful data
b) Predict the diseases with higher accuracy and lower cost
c) Simultaneously retrieve the information and minimize the effort.

They compare the various data mining techniques[13] like ANN, Decision Tree, Logistic Regression, KNN, NB, SVM and apriori algorithm and various data mining tools are compared like WEKA, TANAGRA, ORANGE, R, RAPID MINER, KNIME etc.

But most used data mining tool is WEKA and technique is D Tree, NB and ANN. Decision Tree give maximum accuracy and less human efforts are needed in this algorithm.

VI. Comparison Of Literature Reviews

Author	Year of Publication	Tools and Techniques	Objective
Priyanka Dhaka et al [1]	2016	1.MongoDB 2. Genetic algorithm	1.Provide more accurate information, 2. Less cost and time
Lijun Hao et al [2]	2016	1.OpenStac-k 2. Hadoop	1. integrate the mining tools 2. develop the personal working platform
Miroslav Bursa et al [3]	2011	Ant Colony optimization (ACO)	1.find shortest path 2.increase the speed of loosely structure 3. develop lexical grammar for comparison to classical methods
GuiduoDuan et al [4]	2016	1.Purninig method 2.FT-Growth algorithm 3.Decision tree algorithm	1.Optimize the performance 2. increase efficiency 3. avoid errors 4. increase flexibility
B.V. kiranmayee et al [5]	2016	ID3 algorithm	1.increase quality of service 2. increase efficiency 3. integrate the real world health soft-ware
Hengyi Hu et al [6]	2017	1.Modular ontology 2.Apriori Algorithm	1.improve accuracy 2.reuse, access and evolution of domain knowledge
RaziehAsgarnezhad et al [7]	2017	1.SVM algorithm 2. KNN 3. Genetic algorithm	1.calculate the best value for missing one 2. improve performance 3. increase accuracy and precision result
ParisaNaraei et al [8]	2016	1.SVM 2. Neural Network 3. Decision tree 4. WEKA	1.compare multilayered Neural Network and SVM 2.Find hidden Pattern
HoussemTurki et al [9]	2014	1.Bayesian Network 2. Dynamic Bayesian Network	1. Process the temporal KDD 2.Achive incremental learning 3.Make link with expert knowledge
Sneha Chandra et al [10][11]	2016	1.Decision Tee classifier 2.Laplacian Correction 3.Mean Gray level algorithm 4.K-Mean clustering 5.Bayesian classification	1.Increase quality of service 2.Generate more certain and precise result 3. increase accuracy
Shubpreet Kaur et al [12]	2015	1.Data mining tools 2.Data mining Models 3.Data mining technique	1. compare the various tools and techniques 2.compare their performance 3. compare their accuracy and efficiency 4. select the efficient algorithm for extracting the data.

Table 1. Comparison table of various papers related to medical sector

VII. Conclusion

In this research paper, it concludes that various data mining tools and techniques are used for increasing the accuracy and performance of the data sets. By using Novel approaches increase the efficiency and accuracy also. These data mining techniques also increase the quality of services provided in the sector of Medical. Data mining process is included and then it categorized the data in six common classes. Various challenges are also discussed.

References

1. *P. Dhaka and R. Johari, "Big Data Application : Study and Archival of Mental Health Data ,* using MongoDB," pp. 3228–3232, 2016.

2. *L. Hao, S. Jiang, B. Si, and B. Bai, "Design of the Research Platform for Medical Information Analysis and Data Mining," pp. 1985–1989, 2016.*

3. *M. Bursa and L. Lhotska, "Novel Nature Inspired Techniques in Medical Data Mining," vol. 7, pp. 286–288, 2011*
.

4. *G. Duan, D. Ding, and A. F. P. G. Algorithm, "An Improved Medical Decision Model Based on Decision Tree Algorithms," no. 2014, pp. 151–156, 2016.*

5. *B. V Kiranmayee, "A Novel Data Mining Approach for Brain Tumour Detection."*

6. *H. Hu, L. Kerschberg, A. A. Medical, and O. Amos, "Standardizing the Crowdsourcing of Healthcare Data Using Modular Ontologies," pp. 107–112, 2017.*

7. *R. Asgarnezhad, M. Shekofteh, and F. Z. Boroujeni, "IMPROVING DIAGNOSIS OF DIABETES MELLITUS USING COMBINATION OF PREPROCESSING TECHNIQUES," vol. 95, no. 13, pp. 2889–2895, 2017.*

8. *P. Ferguson et. al., "Network Ingress Filtering: Defeating Denial of Service Attacks which employ IP Source Address Spoofing", Technical report, The Internet Society, 1998.*

9. *H. Turki and M. Ben Ayed, "Using Dynamic Bayesian Networks for the Down Syndrome," pp. 163–167, 2014.*

10. *S. Chandra, "Creation of an Adaptive Classifier to Enhance the Classification Accuracy of Existing Classification Algorithms in the Field of Medical Data Mining," pp. 188–193, 2015.*

11. *S. Chandra and M. Kaur, "Enhancement of Classification Accuracy of our Adaptive Classifier using Image Processing Techniques in the Field of Medical Data Mining," pp. 948–954, 2015.*

12. *S. Kaur and R. K. Bawa, "Future Trends of Data Mining in Predicting the Various Diseases in Medical Healthcare System," vol. 6, no. 4, pp. 17–34, 2015.*

13. *Ruchika and M. Kaur, "A Relative Scrutiny on Big Data and Hadoop Paraphernalia and Techniques," pp. 327–330, 2017.*