Medical Data Mining

Kanika Chuchra[1], Richa Vasuja[2], Ayesha Bhandralia[3]

[1,2,3]*Assistant Professor, Department of Computer Science & Engineering, Chandigarh University, Mohali, Punjab.*

Abstract - *Data mining is a method to determine exciting information, such as associations, patterns, anomalies, deviations and important structures from huge amount of data. Various tools can be used in business questions that conventionally were too much time consuming to resolve. Databases for unseen patterns, discovery of analytical information that professionals may slip because it lies beyond the prospects. Healthcare supervisors use the revealed information to develop the quality. This information can also be used by the medical specialists to decrease the amount of adverse drug consequence, to advise less exclusive medicinally corresponding alternatives. In the WEKA data mining tool, J48 is an open source Java implementation of the C4.5 algorithm. The supplementary characteristics of J48 are accounting for mislaid values; decision trees pruning, constant attribute value varieties, derivation of rules, etc. The WEKA tool offers a number of choices related with tree pruning. J48 works like a flow-chart-like tree structure where every internal node signifies a test on an attribute, edge signifies conclusion of the test and leaf nodes signify output or class distribution. WEKA is an innovator system in the past of the data mining. It is open source and easily accessible platform-independent software. It suggests flexible facilities for scripting experiments and is smoothly functioning.*

Key Words: Data Mining, Medical Data Mining, classification, identification etc.

1. INTRODUCTION

Data mining

In this process we find information from huge amount of data. The extracted information is quite useful and new. And this information can be used later we can store this information for further use.

Since long time it is catching the eyes of the various intellectuals as it been used in wide fields. Data mining has been useful in a lot of research fields like while analysis of data, recognition of various patterns, artificial intelligence, machine learning, and database along with management information system. Data mining is all about extracting data, patterns, variation and significant model and structure from huge amount of data warehousing database. In healthcare industry Data mining helps in maintaining health systems to systematically use data and analyse it to find out all the inefficiencies and best practices that improve care and reduce costs. In information technology data mining helps in developing and then branching off into sub-processes of the collection of data, creating, managing, analyzing and interpreting database. Which involve six steps implementation. 1. Data selection, 2. Filtration of undesirable data, 3.adding value to the processed data, 4. Programming 5.mining of data 6.report generation from mined data. First the data is composed and selected, and irrelevant data is removed. Then data normalization happens to remove unnecessary data. Now the third step is integration of complete data into the existing data. The fourth step is programming where we makeover data into desired form that are appropriate for mining. In the fifth step we apply intelligent methods to identify and discover patterns in data. Last phase is producing suitable reports [2, 4].

Medical Data mining

Methods of data mining can be used to make rules, or identify patterns, from medical data to help clinical diagnostic. But we have limited dataset available in case of initial stage of clinical study due to lesser number of patients involved in that.

We cannot trust the rules discovered from small medical database unless a good range successful result is found.

Medicinal awareness is the most valuable asset or information of a medical association. Mining unknown info from patients', with the given raw data we process this raw data which help physicians to use diagnostic treatments effectively. In other words treatment of patient can be enhanced by using their data as information. And this is why data mining interests the healthcare organisation. In this age most of the healthcare organizations are generating and collecting a huge amount of medical data such as demographics, vital signs, laboratory data, and radiology reports to improve medical treatment.

There are various field in which we are mining data and in biomedical context many procedures have been in data mining. Example decision tree induction. The dataset is recursively classified into subsets. Which is based on the value of an attribute in the dataset. These attributes are selected on the basis of its predictability to a certain subcategory. Hence, the final result is a set of series of categories based on values of the attributes. This generates a classification value. A lot of algorithms have been developed.

In field of genetics we can use data mining in order to determine some associations among the changes in DNA sequences of dissimilar individuals. The intention in research in this arena is to improve diagnosis of disease so that it can be treated more easily

There is huge number of applications in which we are using this technique. While using Classification and association, generally problems come during extraction of information. Regression and classification are also most significant tools for assessment and in predicting results. As human has very limited lookout of instinctive and visual understanding capability on problems with bulky dimension or huge dimensions of databases, the visualization of data mining is newly highlighted in practices. Some unusual determinations of data mining is to mining textual data, for a new search technique in WWW multimedia or texture mining for image processing, and spatial mining for the time-series analysis mainly the text mining is one of worthy methods for natural language processing[1,3,8].

As we have very partial lookout of instinctive and pictorial understanding capability on complications with bulky dimension or huge dimensions of databases.

2. Classification

Classification is used while making analysis. Numerous classification algorithms have been considered to deal with the difficulty by assistants in diverse cases, a heave of data mining exploration in the database. We can re-examined the problem in the framework of huge databases. Different researchers in other fields, database scientists give additional consideration to the issues related to the bulk of data. They are also worried with the effectiveness in use of database practices, such as well-organized retrieval of data by using some mechanism [13]. Classification is a two-Step ProcessModel construction*:* labelling a set of pre-setclasses. Every entity will belong to a predefined classi.e. In terms of output specified, this is identified by the class label i.e. Output. Few tuples are used for model construction i.e. training set. The model is characterised as classification method, decision trees, or mathematical formula.

Model usage: for categorising expectations or unidentified objects. Estimation of correctness of the model. The known outcome of test sample is linked with the classified result. Data for testing should be different from training data otherwise over-fitting may happen [14].

We can use classification technique on each item of dataset and further results can be computed. Classification algorithm plays a major role [15].

3. Tree based algorithms

This algorithm is just like flow chart ie. Step by step, for every step there will be some output. There are various node one of them is internal node which represents a check on an attribute, and edge denotes a result or output of the test which is made on predecessor, and leaf nodes will give output. It is based on "if-then" rules, where "if" part represents the condition on the specified set of attributes and then part is for outcome or output of the conditions [9]. This process is iterative and is stopped when it meets the stopping criteria [6]. When the tree is constructed by using training data, it is then heuristically pruned to evade over-fitting problem, which inclines to present classification fault on the test data [10]. Classification by these tree based classifiers is to done in two phases: tree-growing then tree-pruning. The tree-growing is based on methodology i.e. top down. In this phase, we recursively dive the problem in sub problem and these sub problems will given outcome which are specified by the class label. It is finalised when the subclass at a node has all the identical value of the objective variable, or when splitting no longer adds value to the expectations. In the tree-pruning, when tree is grown fully, it is cut back to evade over fitting problem on the dataset and by this we can improve the exactness of the outcome in bottom up methodology.

```
INPUT:

Data //input dataset to the tool

OUTPUT

Tree //Visualize tree

BUILDDT (*DT')

{

T'=empty;

T'= Build a root node and mark it with attribute;

T'= make a branch to root that is marked with some splitting attribute in the previous step;

For each branch do

D'= Create database and store branches;

If stopping criteria is met, then

T''= mark leaf node with class label;

else

T''=BUILDDT(DT');

}
```

WEKA tool provide us with j48 algorithm which is enactment of the C4.5. The Greedy technique is to induce decision trees in which every decision is made for ordering of data. [11].J48 classifier is a simple C4.5 decision tree for classifying the problem into different class labels. It makes a binary tree. The decision tree method is most valuable in classification problem. By using this, a tree is constructed to model the classification process. To make the verdict the attribute which is having highest information gain is to be used. Then the algorithm recurs on smaller subsets [12].When the tree is made, it is useful to all tuple in the database and outcomes in classification for that tuple. While building a tree, missing values are ignored i.e. values which are missing we can remove whole tuple of it or we can replace the value with most frequent appearing values. The basic hint is to split the data into variety based on the attribute values for that item that are found in the training sample. It lets classification concluded either via decision trees or directions produced from them [5].

The central idea is to divide the data into range based on the values for that item that is found in the training sample. To get the best split information gain should be evaluated properly [5]. The node which is designated to have large information gain of the attribute of dataset. D be the dataset, the assessed information mandatory to properly categorize an instance Xi belongs to D, is given in Equation, where pi symbol is eschance that Xi belongs to D, and is estimated [7]:

$$Info(D) = -\sum_{i=1}^{m} \quad Pi(Pi)$$

4. Conclusion

Data-mining approaches can be used to create patterns, or recognize the data after mining, from therapeutic data to assist scientific analysis and in better decision making. Though, in the early phases of a medical study on a new analytical approach, there could be a partial medical dataset existing; or themedicinaluniquenessmean that the amount of patients concerned in the study will never be huge. Analyses made by the directions revealed from such insignificant medical databases should be careful suspect unless a assurance range for a particular identification can be well-known

5. References

[1]. Zandi, Faramak. "A bi-level interactive decision support framework to identify data mining-oriented electronic health record architectures." *Applied Soft Computing* 18 (2014): 136-145.

[2] Wang, L., and T. Z. Sui. "Application of data mining technology based on neural network in the engineering." In *Wireless Communications, Networking and Mobile Computing, 2007. WiCom 2007. International Conference on*, pp. 5544-5547. IEEE, 2007.

[3] Rahman, Rashedur M., and Fazle Rabbi MdHasan. "Using and comparing different decision tree classification techniques for mining ICDDR, B Hospital Surveillance data." *Expert Systems with Applications* 38, no. 9 (2011): 11421-11436

[4] Paramasivam, Vijayajothi, Tan Sing Yee, Sarinder K. Dhillon, and Amandeep S. Sidhu. "A methodological review of data mining techniques in predictive medicine: An application in hemodynamic prediction for abdominal aortic aneurysm disease." *Biocybernetics and Biomedical Engineering* 34, no. 3 (2014): 139-145.

[5] Patil, Tina R., and M. S. Sherekar. "Performance analysis of Naive Bayes and J48 classification algorithm for data classification." *Int J ComputSci Appl* 6 (2013): 256-261.

[6] Mastrogiannis, Nikolaos, BasilisBoutsinas, and Ioannis Giannikos. "A method for improving the accuracy of data mining classification algorithms." *Computers & Operations Research* 36, no. 10 (2009): 2829-2839.

[7] Farid, Dewan Md, Li Zhang, Chowdhury Mofizur Rahman, M. A. Hossain, and Rebecca Strachan. "Hybrid decision tree and naïve Bayes classifiers for multi-class classification tasks." *Expert Systems with Applications* 41, no. 4 (2014): 1937-1946.

[8] Smith, M. R., X. Wang, and R. M. Rangayyan. "Evaluation of the sensitivity of a medical data-mining application to the number of elements in small databases." *Biomedical signal processing and control* 4, no. 3 (2009): 262-268.

[9] Shah, Chirag, and Anjali G. Jivani. "Comparison of data mining classification algorithms for breast cancer prediction." In *Computing, Communications and Networking Technologies (ICCCNT), 2013 Fourth International Conference on*, pp. 1-4. IEEE, 2013.

[10] Yeh, Jinn-Yi, Tai-Hsi Wu, and Chuan-Wei Tsao. "Using data mining techniques to predict hospitalization of hemodialysis patients." *Decision Support Systems* 50, no. 2 (2011): 439-448

[11] Nookala, Gopala Krishna Murthy, Bharath Kumar Pottumuthu, Nagaraju Orsu, and Suresh B. Mudunuri. "Performance analysis and evaluation of different data mining algorithms used for cancer classification." *International Journal of Advanced Research in Artificial Intelligence (IJARAI)* 2, no. 5 (2013).

[12] Sharma, Trilok Chand, and Manoj Jain. "WEKA approach for comparative study of classification algorithm." *International Journal of Advanced Research in Computer and Communication Engineering* 2, no. 4 (2013): 1995-1931.

[13] Lu, Hongjun, Rudy Setiono, and Huan Liu. "Effective data mining using neural networks." *Knowledge and Data Engineering, IEEE Transactions on* 8, no. 6 (1996): 957-961.

[14] Abe, Hidenao, Hideto Yokoi, Miho Ohsaki, and Takahira Yamaguchi. "Developing an integrated time-series data mining environment for medical data mining." In *Data Mining Workshops, 2007. ICDM Workshops 2007. Seventh IEEE International Conference on*, pp. 127-132. IEEE, 2007.

[15] Vijayarani, S., and M. Muthulakshmi. "Comparative Analysis of Bayes and Lazy Classification Algorithms." *International Journal of Advanced Research in Computer and Communication Engineering* 2, no. 8 (2013).