Analysis on Data Mining Application

MANDAKINI PRIYADRSHINI BEHERA, *Gandhi Institute of Excellent Technocrats, Bhubaneswar,India*

SASWATI JENA, *Kalam Institute of Technology,Berhampur, Ganjam, Odisha, India*

Abstract— Data mining is a powerful and relatively new field that employs a variety of methodologies. It transforms raw data into information that can be used in a variety of study domains. It aids in the discovery of patterns that can be used to predict future medical trends.

Data mining, information prediction, and raw data are all keywords.

INTRODUCTION

In numerous study domains, the advancement of information technology has resulted in a great volume of data-base and data. Knowledge mining research has led to the ability to store and change previously stored data in order to aid in the decision-making process.

I. DATA MINING PROCESS

Data mining is a technique for extracting implicit and previously unknown data. Data mining is a method of attracting consumers' attention due to the abundance of large amounts of data and the necessity to convert that data into usable information.

As a result, many people refer to data mining as "knowledge discovery device" or KDD. In data mining, seven successive steps are performed to extract or uncover knowledge:

1) 1) Data cleaning: At this stage, we remove noise and irrelevant data from the collected raw data.
2) 2) Data integration: At this stage, we consolidate different data sources into single target data storage.
3) 3) Data Selection: Data relevant to the analytical task is retrieved as pre-processed data from the data base.
4) 4) Data transformation: By summarizing and aggregating activities, data is consolidated into standard representations suitable for mining.
5) 5) Data Mining: Various intelligent approaches and technologies are used in this step to extract data patterns or rules.
6) 6) Pattern evaluation: Identify tree patterns that represent knowledge in this step.
7) Knowledge representation: This is the final step, in which visualization and knowledge representation techniques are utilized to assist users in comprehending and representing information..

The purpose of the knowledge discovery and data mining processes is to uncover hidden patterns in large amounts of data and interpret them into meaningful knowledge and information..

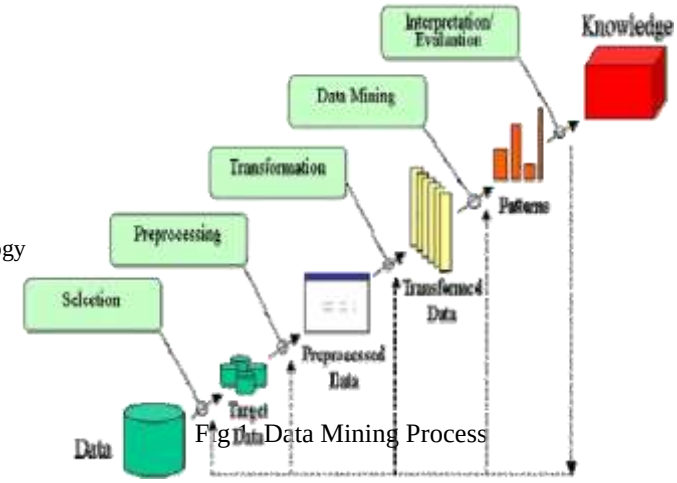

Fig Data Mining Process

The key aspect of the knowledge discovery process in the diagram is data mining.

Application of data mining:

- Customer profiling, retention, identification of potential customers, and market segmentation are all aspects of marketing.
- Intrusion detection and credit card fraud detection
- Scientific data analysis: Identify the data used to make research decisions.
- Text and web mining: used to find text or information on the internet or to analyse raw data.
- Any other applications that require a lot of data.

II. DATA MINING TECHNIQUES [1]

For information discovery from databases, numerous important data mining techniques have recently been created and employed in data mining projects, including association, rule classification, clustering, prediction, and evaluation pattern, among others.

One of the most widely used data mining approaches is association. We use this strategy to find intriguing associations and connections within data by mining common patterns.

In marketing analysis, the association technique is used to find things that are commonly purchased in the same transaction.

buys(X; "computer"))buys(X; "software") [support = 1% ; confidence = 50%] is an example of such a rule extracted from the All Electronics transactional database. X is a variable that represents a customer. A confidence level of 50% suggests that if a consumer buys a computer, there is a 50% chance she will also buy software. A 1% support indicates that computer and software were purchased jointly in 1% of the transactions examined. as laws of single-dimensional association The above rule can be stated simply as "computer)software[1%, 50%]" if the predicate notation is removed.

1. Classification: It is the process of finding a model or function that describes & distinguish data classes or concepts for the purpose of being able to use the model to predict the class of object whose class label is unknown.

In classification, we make software that can learn how to classify the data items into group . Derived model can be presented as classification or rules. So,

Classification techniques:

- Regression
- Distance
- Decision
- Rules
- Neural networks

2. Clustering: Process of grouping a set of physical or abstract object into classes of similar objects is called clustering.

A cluster is a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters.

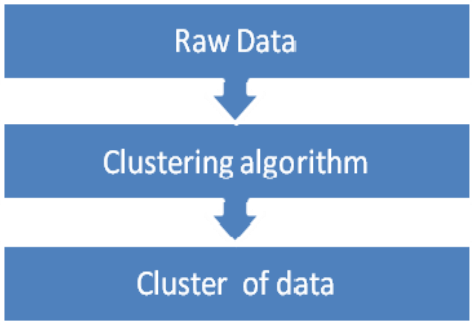

Fig 2. Clustering

A cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters.

By clustering we can identify dense and spare regions in object space and discover distribution patterns and interesting correlations among data attributes. It means data segmentation.

In earth observation, it helps in identification of areas of similar land use and identify group of houses in a city according to house type and geographic location, etc.

Prediction: The classification predicts categorical (discrete, unordered) labels, prediction models continuous-valued functions. That is, prediction is used to predict missing or unavailable numerical data values rather than class labels. But, the term predictionmayrefer to both numeric prediction and class label prediction.

Example: Regression analysis is a statistical methodology that is most often used for numeric prediction, although other methods exist as well. Prediction also encompasses the identification of distribution trends based on the available data.

Applications of prediction:

- Credit approval
- Target marketing
- Medical diagnosis
- Treatment effectiveness analysis

3. Evaluation Pattern:

Data evolution analysis describes and models regularities or trends for objects whose behavior changes over time. Although this may include characterization, discrimination, association and correlation analysis, classification, prediction, or clustering of time related data, distinct features of such an analysis include time-series data analysis, sequence or periodicity pattern matching, and similarity-based data analysis.

Example: Evolution analysis. Suppose that you have the major stock market (time-series) data of the last several years available from the New York Stock Exchange and you would like to invest in shares of high-tech industrial companies. A data mining study of stock exchange data may identify stock evolution regularities for overall stocks and for the stocks of particular companies. Such regularities may help predict future trends in stock market prices, contributing to your decision making regarding stock investments.

Selected data mining techniques in medicine

There are various data mining techniques available with suitable dependent on domain application.

By using data mining we can examine large amount of routine samples collected in disease prediction. Best results are achieved by balancing knowledge of experts for describing the problem and goals with search capabilities.

Hospitals must also want to minimize cost of clinical test. It can be achieved by employing appropriate computer based information and decision sport system. Here, data mining plays an important role to give many results faster and accurate by using various algorithms.

There are two primary goals for data mining **prediction and description.** Prediction involves fields or variables in the data sets to predict unknown or future values of other diseases possibilities. On the other hand description involves finding of pattern describing the data that can be present in knowledge base provided for disease prediction.

We can predict diseases like hepatitis, Lung cancer liver disorder, breast cancer or heart diseases, diabetes etc,.

We can use Naïve algorithm, Robin Karp algorithm, K-NN algorithm and decision tree are most popular classifier which are easy and simple to implement. They can handle huge amount of dimensional data.

Example: we can use naïve algorithm to predict attributes like age, sex, blood pressure and blood sugar, changes of diabetes patient getting heart disease.
Naive algorithm is used to analyze alpha hemoglobin or beta hemoglobin in test of hemoglobin red blood cells. And it can be used for DNA test.
Decision tree can be used to represent results in form of tree. Leaf nodes or internal nodes are labeled with values of attributes. Branches coming out from internal nodes are labeled with values of attributes in the node. This technique is best suited for data mining in medicine or diseases prediction.
Example:The finding of a solution with the help of decision trees starts by preparing a set of solved cases.[5]
The whole set is then divided into 1) a training set, which is used for the induction of a decision tree, and 2) a testing set, which is used to check the accuracy of an obtained solution. Each attribute can represent one internal node in a generated decision tree, also called an attribute node or a test node (Fig-3). Such an attribute node has exactly as many branches as its number of different value classes. The leaves of a decision tree are decisions and represent the value classes of the decision attribute – decision classes (Fig-3).
The decision tree is very easy to interpret. For example, from the tree shown in (Fig-3) wecan deduce the following two rules:

1. if the patient has inter-systolic noise and MCI and heart malformations then she/he has a prolapse, and
2. if the patient has inter-systolic noise and MCI and no heart malformations then she/he does not have a prolapse.

Here, the MCI and Pre-cordial Pain are attribute (test) nodes in a growing decision tree and leaf nodes are the decision nodes.

1

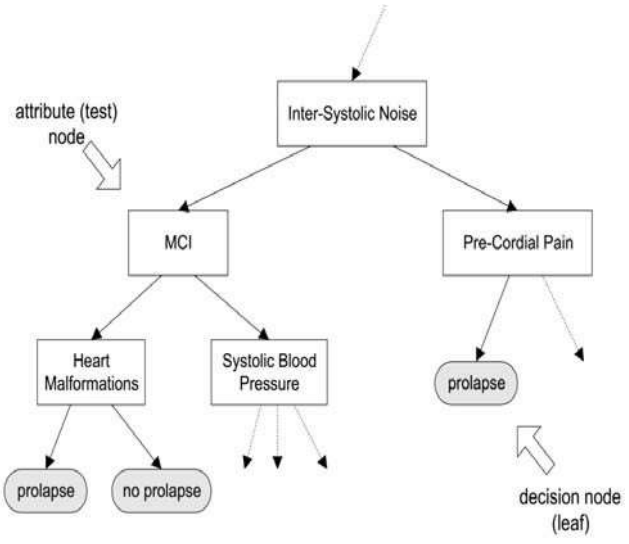

Fig 3. An example of a (part of a) decision tree.[5]

and conquer" approach. If all objects are of same class decision tree consist of single node or leaf node. Otherwise, attribute node have at least two leaf nodes as growing decision tree. For branch from that node the inducing procedure is repeated upon the remaining objects regarding division or output as leaf node comes.
There are many other techniques used to represent data in analyzing the results .
Such as:

- Genetic algorithms.
- Fuzzy sets.
- Neural networks.
- Rough sets.
- Support vector machine(SVM)

We can implement these techniques to classify member sets of objects as either +ve or –ve results of test performed to check fitness or illness of patient, these techniques are used to extent the purpose to analyze the diseases with multi-class decision making algorithms.

III. Conclusion

Data mining is a "decision support" process in which we search for patterns of information in data. Data mining techniques such as classification, clustering, prediction, association and sequential patterns etc.
The commercial, educational and scientific applications are increasingly dependent on these methodologies.
Decision trees are a reliable and effective decision making technique which provide high classification accuracy with a simple representation of collected KDD. It help experts to validate and classify the results and outcomes of tests and analyze various new symptoms of diseases based on data.
Thus , data mining can help to play an important role in the field of medicine or health care and disease prediction.

References

(Journal papers):
[1]. Kalyani et al., International Journal of Advanced Research in Computer Science and Software Engineering, ISSN: 2277 128X ,Volume 2, Issue 10, October 2012 .
[2]. Shalini Sharma, Vishal Shrivastava, International Journal on Recent and Innovation Trends in Computing and Communication , ISSN 2321 – 8169 Volume: 1 Issue: 4, March 2013.
[3].Megha Gupta, Vishal Shrivastava, International Journal on Recent and Innovation Trends in Computing and Communication, ISSN 2321 – 8169 Volume: 1 Issue: 8,August 2013.
[4]. S.Vijiyarani S.Sudha, Disease Prediction in Data Mining Technique – A Survey, International Journal of Computer Applications & Information Technology, ISSN: 2278-7720 Vol. II, Issue I, January 2013 .
[5]. Vili Podgorelec, Peter Kokol, Bruno Stiglic, Ivan Rozman, Decision trees: an overview and their use in medicine, Journal of Medical Systems, Kluwer Academic/Plenum Press,Vol. 26, Num. 5, pp. 445-463, October 2002.
(Books):
[6]. Han and Kamber, "Data Mining and Concepts".