ISSN:0975-9646

Application of Data Mining – A Survey Paper

Aarti Sharma, Rahul Sharma,Vivek Kr. Sharma,Vishal Shrivatava

Department of CS &IT .,
A.C.E.I.T.,Jaipur

Abstract— **Data mining is a powerful and a new field having various techniques. It converts the raw data into useful information in various research fields. It helps in finding the patterns to decide future trends in medical field.**

Keyword: Data mining, information prediction, raw data .

I. INTRODUCTION

Development of information technology has generated large amount of data-base and huge amount of data in various research fields. To research in knowledge mining has give rise to store data and manipulate previously stored data for further decision making process.

II. DATA MINING PROCESS

Data mining is used to extract implicit and previously unknown information from data. Data mining is the process which provides a concept to attract attention of users due to high availability of huge amount of data and need to convert such data into useful information.

So, many people use the term “knowledge discovery device” or KDD for data mining.

Knowledge extraction or discovery is done in seven sequential steps used in data mining:

1) Data cleaning: we remove noise data and irrelevant data from collected raw data, at this step.
2) Data integration: At this step, we combine multiple data sources into single data store called target data.
3) Data Selection: Here, data relevant to analysis task are retrieved from data base as pre-processed data.
4) Data transformation: Here, data is consolidating into standard formats appropriate for mining by summarizing and aggregated operations.
5) Data Mining: At this step, various smart techniques and tools are applied in order to extract data pattern or rules.
6) Pattern evaluation: At this step, strictly identify tree patterns representing knowledge.
7) Knowledge representation: This is the last stage in which, visualization and knowledge representation techniques are used to help users to understand and interpret the data mining knowledge or result.

The goal of knowledge discovery and data mining process is to find the patterns that are hidden among the huge set of data and interpret useful knowledge and information.

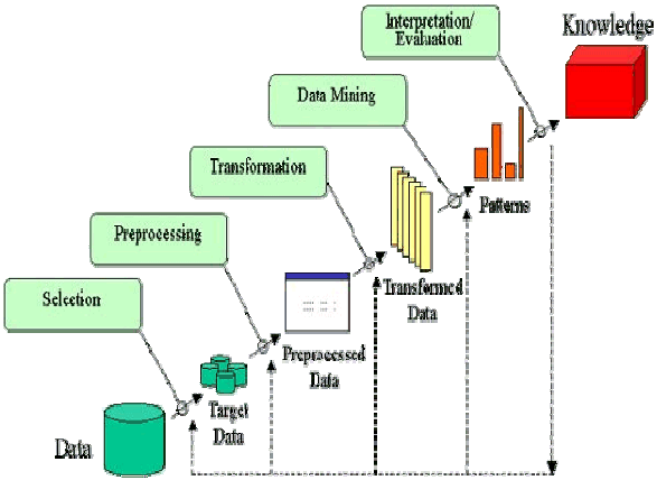

Fig 1. Data Mining Process

In the diagram data mining is the main part of knowledge discovery process.

Data mining applications:

- **Marketing:** Customer profiling, retention, identification of potential customer, market segmentation.
- **Fraud detection:** Identify credit card fraud and intrusion detection.
- **Scientific data analysis:** Identify the research decision making data.
- **Text and web mining:** used to search text or information on web or given raw data.
- Any other applications that involve large amount of data.

III. DATA MINING TECHNIQUES [1]

There are various major data mining techniques that have been developed and used in data mining projects recently including **association, rule classification, clustering, prediction and Evaluation pattern etc.,** are used for knowledge discovery from database.

1. Association: It is one of the most popular data mining techniques. In this technique we mine frequent patterns lead to discovery of interesting association and correlations within data.

Example:

Association technique is used in marketing analysis to identify items which are frequently purchased within the same transactions.

An example of such a rule, mined from the *All Electronics* transactional database, is
***buys*(*X*; "*computer*"))*buys*(*X*; "*software*") [*support* = 1%; *confidence* = 50%]**where *X* is a variable representing a customer. A confidence, or certainty, of 50% means that if a customer buys a computer, there is a 50% chance that she will buy software as well. A 1% support means that 1% of all of the transactions under analysis showed that computer and software were purchased together.as single-dimensional association rules. Dropping the predicate notation, the above rule can be written simply as ***"computer)software*[1%, 50%]".**

2. 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

3. 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.

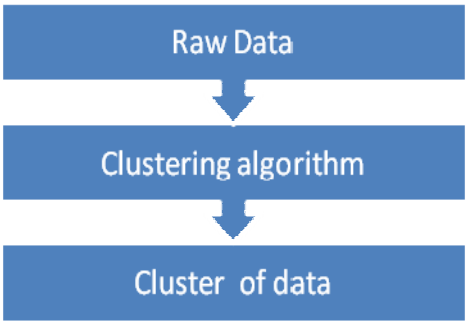

Fig 2. Clustering

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

4. 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.

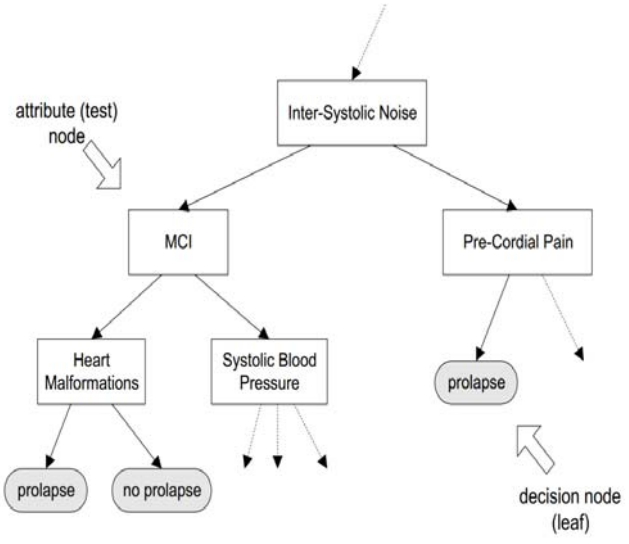

Fig 3. An example of a (part of a) decision tree.[5]

It is built from the set of training objects with "divide 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.

IV. 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".