ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 1, January 2014

A SURVEY ON DATA MINING TECHNIQUES

Ranshul Chaudhary[1], Prabhdeep Singh[2,] Rajiv Mahajan[3]

Student, M.Tech Department, Global College of engineering and technology, Amritsar, India [1]

Assistant professor, M.Tech Department, Global College of engineering and technology, Amritsar, India [2]

Head of the Department, M.Tech Department, Global College of engineering and technology, Amritsar, India [3]

Abstract: Data Mining refers to the analysis of observational datasets to find relationships and to summarize the data in ways that are both understandable and useful. Compared with other DM techniques, Intelligent Systems (ISs) based approaches, which include Artificial Neural Networks (ANNs), fuzzy set theory, approximate reasoning, and derivative-free optimization methods such as Genetic Algorithms (GAs), are tolerant of imprecision, uncertainty, partial truth, and approximation. This paper is concerned with the ideas behind design; implementation, testing and application of a novel ISs based DM technique.

Keywords: Data mining, clustering, classification, MST

I. INTRODUCION

The wide-spread use of distributed information systems leads to the construction of large data collections in business, science and on the Web. These data collections contain a wealth of information, which however needs to be discovered. Businesses can learn from their transaction data more about the behaviour of their customers and therefore can improve their business by exploiting this knowledge. Science can obtain from observational data (e.g. satellite data) new insights on research questions. Web usage information can be analyzed and exploited to optimize information access [3]. Data mining provides methods that allow extracting from large data collections unknown relationships among the data items that are useful for decision making. Thus data mining generates novel, unsuspected interpretations of data [1][2] .

II. SURVEY OF EXISTING RESEARCH

Fayyad et.al[4,5] describes the various data mining techniques that allow extracting unknown relationships among the data items from large data collection that are useful for decision making. The wide-spread use of distributed information systems leads to the construction of large data collections in business, science and on the Web. These data collections contain a wealth of information, which however needs to be discovered. Businesses can learn from their transaction data more about the behaviour of their customers and therefore can improve their business by exploiting this knowledge. Science can obtain from observational data (e.g. satellite data) new insights on research questions. Web usage information can be analyzed and exploited to optimize information access. Thus data mining generates novel, unsuspected interpretations of data.

In practice the two fundamental goals of data mining tend to be: *prediction* and *description.*

1. Prediction makes use of existing variables in the database in order to predict unknown or future values of interest.
2. Description focuses on finding patterns describing the data and the subsequent presentation for user interpretation. The relative emphasis of both prediction and description differ with respect to the underlying application and the technique.

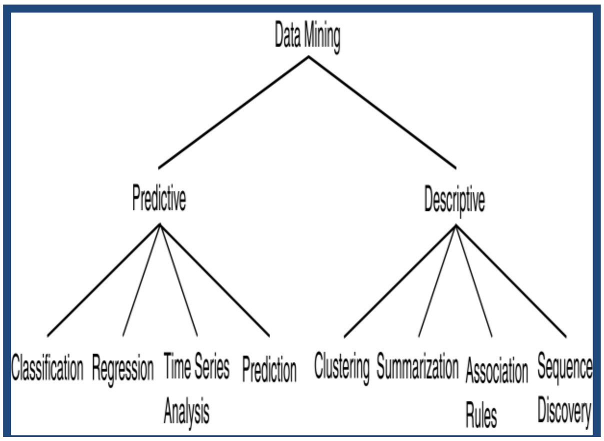

Fig 1. Data mining techniques

A. METHODS OF DATAMINIG (KEY TECHNIQUES):

- Association

Association (or relation) is probably the better known and most familiar and straight forward data mining technique. Here, you make a simple correlation between two or more items, often of the same type to identify patterns.

- Classification

Piatetsky et.al[6] proposes a classification technique by providing training to various data set. You can use classification to build up an idea of the type of customer, item, or object by describing multiple attributes to identify a particular class. For example, you can easily classify cars into different types (sedan, 4x4, convertible) by identifying different attributes (number of seats, car shape, driven wheels). Additionally, you can use classification as a feeder to, or the result of, other techniques. For example, you can use decision trees to determine a classification. Clustering allows you to use common attributes in different classifications to identify clusters.

- Clustering

G.P and MARTY et.al[12] examines in the paper ,how Clustering technique is useful to identify different information by considering various examples and one can see where the similarities and ranges agree. By examining one or more attributes or classes, you can group individual pieces of data together to form a structure opinion. At a simple level, clustering is using one or more attributes as your basis for identifying a cluster of correlating results. Clustering can work both ways. You can assume that there is a cluster at certain point and then use our identification criteria to see if you are correct[16][17].

- Prediction

T.HASTIE et.al [13] proposes prediction method in combination with the other data mining techniques, involves analysing trends, classification, pattern matching, and relation. Prediction is a wide topic and runs from predicting the failure of components or machinery, to identifying fraud and even the prediction of company profits. By analyzing past events or instances, you can make a prediction about an event.

- Sequential patterns

DUDAR and HART P[14] describes the various uses of sequential patterns for identifying trends, or regular occurrences of similar events. For example, with customer data you can identify that customers buy a particular collection of products together at different times of the year. In a shopping basket application, you can use this information to automatically suggest that certain items be added to a basket based on their frequency and past purchasing history[17].

III. OBJECTIVES OF STUDY

- Analytic study of data mining techniques.
- Train and test the data with different techniques.
- How to predict the unknown values i.e. analysis of output of different techniques.
- Comparing different techniques on different factors e.g. input and time taken to train and test.

IV. RESEARCH METHODOLOGY

Under this thesis technique of data mining is implemented and study of data mining tool weka is done. With the help of weka different techniques of data mining implemented. We train a model with an algorithm and dataset and test a particular dataset and do a analysis how a algorithm helps to predict the unknown values. So firs we analyse different techniques of classification - reduction rule, decision tree and bayes net and then compares their outputs analyse which techniques works better on what type of input and at what situation. Similarly different techniques of classifications are implemented and comparison is done.

V. CONCLUSION

In this paper we studied some well known algorithms concerned with data mining. Under the clustering techniques of data mining various algorithms namely: k-means, Hierarchical clustering, COBWEB and DBSCAN algorithms are studied. The results are compared and analyzed in accordance to their efficiencies. For classification, the Decision Tree and Bayesian algorithm were implemented and compared. Under the clustering techniques of data mining various algorithms namely- k-means, k-medoid and DBSCAN algorithms are implemented using WEKA. The results are compared and analyzed in accordance to their efficiencies.

REFERENCES

[1] Meta Group Inc. Data Mining: Trends, Technology, and Implementation Imperatives. Stamford, CT, February 1997.

[2] Goebel, M. and Grunewald, L., A Survey of Knowledge Discovery and Data Mining Tools. Technical Report, University of Oklahoma, School of Computer Science, Norman, OK, February 1998.

[3] Waikato ML Group. User Manual Weka: The Waikato Environment for Knowledge Analysis. Department of Computer Science, University of Waikato (New Zealand), June 1997.

[4] Thearling, K. Data Mining and Database Marketing WWW Pages. http://www.santafe.edu/~kurt/dmvendors.shtml, 1998.

[5] Fayyad, U. Data Mining and Knowledge Discovery: Making Sense Out of Data. IEEE Expert, v. 11, no. 5, pp. 20-25, October 1996.

[6] Survey of classification techniques in data mining in Proceedings of the International Multi Conference of Engineers and Computer Scientists 2009 Vol I IMECS 2009, March 18 - 20, 2009, Hong Kong-classification

[7] S. Guha, R. Rastogi, K. Shim, CURE: an efficient clustering algorithm for large databases, in: Proceedings of the 1998 ACM-SIGMOD International Conference Management of Data (SIGMOD'98), 1998, pp. 73–84.

[8] DEFAYS, D. 1977. An efficient algorithm for a complete link method. The Computer Journal, 20, 364-366.

[9] DHILLON, I., GUAN, Y., and KOGAN, J. 2002. Refining clusters in high dimensional data. 2nd SIAM ICDM, Workshop on clustering high dimensional data, Arlington, VA

[10] BOLEY, D.L. 1998. Principal direction divisive partitioning. Data Mining and Knowledge Discovery, 2, 4, 325-344.

[11] BRADLEY, P. and FAYYAD, U. 1998. Refining initial points for k-means clustering. In Proceedings of the 15th ICML, 91-99, Madison, WI.

[12] BABU, G.P. and MARTY, M.N. 1994. Clustering with evolution strategies Pattern Recognition, 27, 2, 321-329.

[13] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning, Data Mining, Inference and Prediction, Springer, New York, 2001.

[14] DUDA, R. and HART, P. 1973. Pattern Classification and Scene Analysis. John Wiley & Sons, New York, NY.

[15] X. Wu, V. Kumar, J.R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G.J. McLachlan, A.F.M. Ng, B. Liu, P.S. Yu, Z.-H. Zhou, M. Steinbach, D.J. Hand, and D. Steinberg, "Top 10 Algorithms in Data Mining," Knowledge and Information Systems, vol. 14, no. 1, pp. 1-37, 2008.

[16] Q. Yang and X. Wu, "10 Challenging Problems in Data Mining Research," Int'l J. Information Technology and Decision Making, vol. 5, no. 4, pp. 597-604, 2006.

[17] G.P.C. Fung, J.X. Yu, H. Lu, and P.S. Yu, "Text Classification without Negative Examples Revisit," IEEE Trans. Knowledge and Data Eng., vol. 18, no. 1, pp. 6-20, Jan. 2006.

[18] H. Al Mubaid and S.A. Umair, "A New Text Categorization Technique Using Distributional Clustering and Learning Logic," IEEE Trans. Knowledge and Data Eng., vol. 18, no. 9, pp. 1156-1165, Sept. 2006.