International Journal of Electronics Engineering (ISSN: 0973-7383)
Volume 10 • Issue 2 pp. 706-708 June 2018-Dec 2018 www.csjournals.com

Data Mining Techniques

Ranjandeep Kaur Khera
PG Dept. of Computer science, Khalsa College for women,Amritsar

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.

Introduction

The construction of large data collection in business, science and on web is the result of wide spread usage of distributed information systems. 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 [5]. 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 [6][2] .

Survey of Existing Research

Fayyad et.al[4,7] 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 has resulted in the construction of large data collections in business, science and on the Web, Which contain a pool of information, which however needs to be discovered. Business can improve further by exploiting the knowledge provided by their transaction data about the behavioural pattern of various customers. In the field of science, new insight on research questions can be obtained from the observational data. Web usage information can be analyzed and exploited to optimize information access. Thus data mining generates novel, unsuspected interpretations of data.

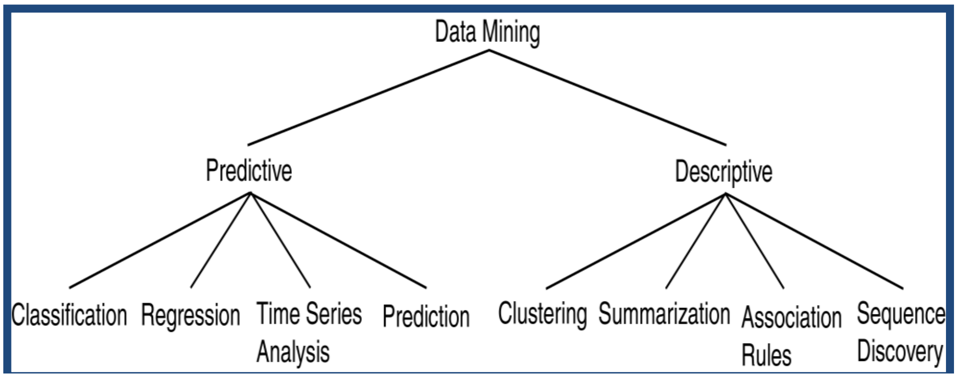

FIG-1: Data Mining Techniques

In practice the two fundamental goals of data mining tend to be: *prediction* and *description.*
1. Prediction utilises the existing variables in the database for predicting unknown or future values of interest.
2. The primary focus of description is to find the pattern which describes the data and further presentations for user interpretations. The relative emphasis of both prediction and description differ with respect to the underlying application and the technique. There are several data mining techniques fulfilling these objectives. Some of these are associations, classifications, and clustering.

METHODS OF DATAMINIG (KEY TECHNIQUES):

Association

Association (or relation) is probably the widely known, most recoganizable and simple data mining technique. Here in order to identify pattern a simple correlation is made between two or more items which are often of the same type.

Classification

Piatetsky et.al[11] proposes a classification technique by providing training to various data set. Classification can be used to construct an idea about the type of customer, item or object by describing numerous attributes for identifying a particular class. For example, you can easily classify cars into different types (sedan, 4x4, convertible) on the bases of distinctive qualities (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. In clustering, common attributes in different classifications can be utilized to identify cluster.

Clustering

G.P and MARTY et.al[15] scrutinizes, in his paper ,the way Clustering technique is helpful to discern different information by considering various examples .It is also useful to find out 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 employs one or more attributes as the basis for identifying a cluster of correlating results. Interestingly, Clustering can work both ways as you can assume that there is a cluster at certain point and then use your identification criteria to see if you are correct[9][14].

Prediction

T.HASTIE et.al [3] proposes prediction method in combination with the other data mining techniques, which involves analysing trends, classification, pattern matching, and relation. Prediction is a broad topic as it runs from predicting the failure of components or machinery, to identifying fraud and even the prediction of company profits. In this the prediction about an event can be made by analysing the past event or instances.

Sequential patterns

DUDAR and HART P[16] describes the various uses of sequential patterns for identifying trends, or regular occurrences of similar events. For Example: Customer data can be used to determine the particular type of product that customer buy at different times of a year. Further in a shopping basket application this information can automatically suggest the type of item that can be added to the basket based on their frequently and the past purchasing history of the customer.

Objectives of the 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.

Research Methodology

The study of data mining tool WEKA is done and technique of data mining is implemented. We train a model with an algorithm and dataset. To test a particular dataset we need to know, how an algorithm helps to predict the unknown values. So first we analyse different techniques of classification - reduction rule, decision tree and bayes net. Then we compare their outputs, to know which techniques works better on what type of input and at what situation. Similarly different techniques of classifications are implemented and comparison is done.

Conclusion and future work

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. 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.
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. T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning, Data Mining, Inference and Prediction, Springer, New York, 2001.
4. Thearling, K. Data Mining and Database Marketing WWW Pages. http://www.santafe.edu/~kurt/dmvendors.shtml, 1998.
5. Waikato ML Group. User Manual Weka: The Waikato Environment for Knowledge Analysis. Department of Computer Science, University of Waikato (New Zealand), June 1997.
6. Meta Group Inc. Data Mining: Trends, Technology, and Implementation Imperatives. Stamford, CT, February 1997.
7. Fayyad, U. Data Mining and Knowledge Discovery: Making Sense Out of Data. IEEE Expert, v. 11, no. 5, pp. 20-25, October 1996.
8. BRADLEY, P. and FAYYAD, U. 1998. Refining initial points for k-means clustering. In Proceedings of the 15th ICML, 91-99, Madison, WI.
9. 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.
10. DEFAYS, D. 1977. An efficient algorithm for a complete link method. The Computer Journal, 20, 364-366.
11. 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.
12. 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.
13. BOLEY, D.L. 1998. Principal direction divisive partitioning. Data Mining and Knowledge Discovery, 2, 4, 325-344.
14. 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.
15. BABU, G.P. and MARTY, M.N. 1994. Clustering with evolution strategies. Pattern Recognition, 27, 2, 321-329.
16. DUDA, R. and HART, P. 1973. Pattern Classification and Scene Analysis. John Wiley & Sons, New York, NY.
17. 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.
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.