

Analysis of Data Mining Techniques and Applications

SANKARA RAO LAMBURU

Assistant Professor,
Department of Computer Science and Engineering,
Viswanadha Institute of Technology and Management, Visakhapatnam
Jawaharlal Nehru Technological University Kakinada, Andhra Pradesh, India

Abstract: - Data Information mining is a moderately youthful and interdisciplinary field of software engineering. It is a procedure that endeavors to find new examples in huge informational collections. Various sorts of mining calculations have been proposed by various specialists as of late. A solitary calculation may not be applied to all applications because of trouble for appropriate information kinds of the calculation. Along these lines, the choice of a right information mining calculation relies upon the objective of an application, yet in addition to the similarity of the informational collection.

The point of this paper is the way to utilize reasonable information mining calculations on the instructive dataset. This paper centers around a similar investigation of different information mining procedures and calculations. and algorithms.

Key words: Data mining, Clustering, Classifications, KDD

1. INTRODUCTION

Data mining is the term used to describe the process of extracting value from a database. Data Mining (DM) is the process of analyzing data from different perspectives and summarizing it into useful information. It is the efficient discovery of valuable, non-obvious information from a large collection of data. A data-warehouse is a location where information is stored [1]. The type of data stored depends largely on the type of industry and the company. Many companies store every piece of data they have collected, while others are more ruthless in what they deem to be “important”.

Consider the following example of a financial institution failing to utilize their data-warehouse. Income is a very important socio-economic indicator. If a bank knows a person's income, they can offer a higher credit card limit or determine if they are likely to want information on a home loan or managed investments. Even though this financial institution had the ability to determine a customer's income in two

ways, from their credit card application, or through regular direct deposits into their bank account, they did not extract and utilize this information.

Another example of where this institution has failed to utilize its data-warehouse is in cross-selling insurance products (e.g. home, life and motor vehicle insurance). By using transaction information they may have the ability to determine if a customer is making payments to another insurance broker. This would enable the institution to select prospects for their insurance products. These are simple examples of what could be achieved using data mining [2].

The field of data mining is an emerging research area with important applications in Engineering, Science, Medicine, Business and Education. The size of data base in educational application is large where the number of records in a data set can vary from some thousand to thousand of millions. The size of data is accumulated from different fields exponentially increasing [3].

Data mining has been used different methods at the intersection of Machine Learning, Artificial Intelligence, Statistics and Database Systems. The overall aim of the data mining process is to extract information from huge datasets and transform it into understandable structure for further use. Data mining techniques which extract information from huge amount of data [4].

2. LITERATURE SURVY

This section summarizes various review and technical articles on data mining techniques applied in education field. Several works have been carried out by many

researchers. This section presents a brief summary on the basis of literature.

M.Sukanya et al. used classification and clustering algorithms of data mining for the performance improvement in education sector. By using these algorithms an educational institute could predict the number of enrolled students.

Manoj Bala et al. applied an application of data mining in educational institute to extract the useful information from the huge dataset and provided analytical tool to view and used this information for decision making process. They also conducted a research on student learning result based on data mining.

Hamidah Jantan used a potential classification technique for academic talent forecasting in higher educational institutes. He proposed a classification model to increase the academic talent in higher educational institutions.

Tai Chang Hsia applied data mining techniques to analyze the course preferences and course completion rates of enrollees in extension education courses at a University in Taiwan. Some of the data mining algorithms like decision tree, link analysis, and decision forest were used for further analysis

3. RELATED WORK

DATA MINING TECHNIQUES

Generally data mining contains several algorithms and techniques for picking out interesting patterns from large data sets. Data mining techniques are classified into two categories: supervised learning and unsupervised learning. In supervised learning, a model is built prior to the analysis. We then apply the algorithm to the data in order to estimate the parameters of the model. Classification, Decision Tree, Bayesian Classification, Neural Networks, Association Rule Mining etc. are common examples of supervised learning. In unsupervised learning, we do not create a model or hypothesis prior to the analysis [4].



Figure .1: Data Mining Process

We just apply the algorithm directly to the dataset and observe the results. Then a model can be created on the basis of the obtained results. Clustering is one of the examples of unsupervised learning [5]. Various data mining techniques such as Classification, Decision Tree, Bayesian Classification, Neural Networks, Clustering, Association Rule Mining, Prediction, Time Series Analysis, Sequential Pattern and Genetic Algorithm and Nearest Neighbors have been used for knowledge discovery from large data sets. Some of the common and useful data mining techniques have been discussed below [2].

- a) Classification of various operations
- b) Decision Tree on existing roots
- c) Bayesian Classification
- d) Neural Networks association with Database
- e) Clustering process
- f) Association Rule Mining in various factors
- g) Prediction Analysis
- h) Time Series Analysis with ROI
- i) Sequential Patterns Analysis

Classification: Classification is a data mining (machine learning) technique used to predict group membership for data instances. For example, you may wish to use classification to predict whether the weather on a particular day will be sunny, rainy or cloudy.

Popular classification techniques include decision trees and neural networks.

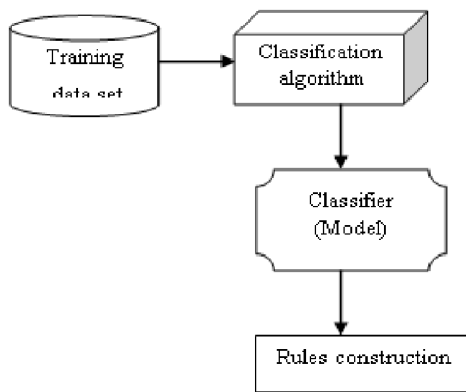


Figure.2: Classification Techniques

Regression: Regression is a data mining function that predicts a number. Regression is a data mining (machine learning) technique used to fit an equation to a dataset. The simplest form of regression, linear regression, uses the formula of a straight line ($y = mx + b$) and determines the appropriate values for m and b to predict the value of y based upon a given value of x . Advanced techniques, such as multiple regression, allow the use of more than one input variable and allow for the fitting of more complex models, such as a quadratic equation [6].

Time Series Analysis: A time series represents a collection of values obtained from sequential measurements over time. Time-series data mining stems from the desire to reify our natural ability to visualize the shape of data. Humans rely on complex schemes in order to perform such tasks. We can actually avoid focusing on small fluctuations in order to derive a notion of shape and identify almost instantly similarities between patterns on various time scales.

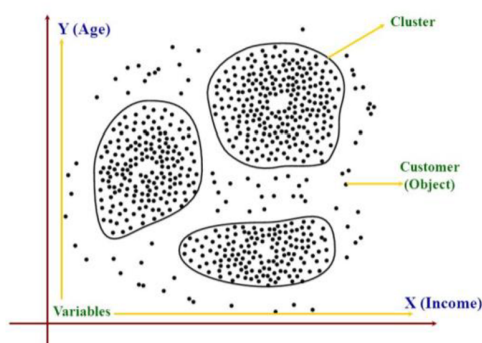


Fig 3: Clustering Technique

Developing theories behind different types of data mining (for example, lower bounds on complexity of mining sequences) would enable us to determine if our data is mineable.

We could use this to develop "probes" of hard-to-mine data (for example, a characterization of data that is likely to contain association rules). This will enable us to build systems that provide the data (or access to the data) in a way that ensures that mining is both unlikely to and interesting results, or will and "planted" results that do not concern us [7].

Data mining to detect inference problems

Given a database containing both "public" information and "sensitive" information, we can use data mining to search for inference paths from the public to the sensitive information. In other words, a rule with antecedents in the public information and consequents in the sensitive information points to a potential problem. Clustering the public information could conceivably bring together the antecedents.

This by itself does not pose a problem, but knowing sensitive information about one of the instances could then be conjectured (correctly) to apply to all of them (greatly expanding the "leak" of sensitive information). Proper use of data mining technology can be used to detect such situations. We can then apply techniques from the previous section to ensure that the data cannot be clustered in meaningful ways [8].

Our introductory example of mining supermarket sales shows how this may work. We may decide to protect brand information; only releasing sales by product type. However, Deters may know that Green Paper dominates a particular product type, and can use data mining combined with this fact to and "brand a Unity" information.

Preemptive use of data mining (to deduce brand from the information we intend to release) would show this product type) brand relationship, allowing us to farthing our information release strategy [9].

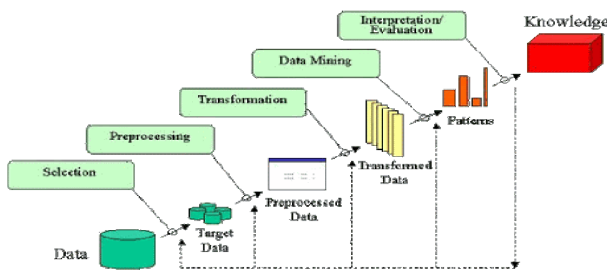


Figure.4: KDD Process in Data Mining

- Data cleaning
- Data Integration
- Data Selection
- Data Transformation
- Data Mining
- Pattern Evaluation
- Knowledge Presentation

Advantages of Neural Networks:

- a) **High Accuracy:** Neural networks are able to approximate complex non-linear mappings
- b) **Noise Tolerance:** Neural networks are very flexible with respect to incomplete, missing and noisy data.
- c) **Independence from prior assumptions:** Neural networks do not make a priori assumptions about the distribution of the data, or the form of interactions between factors.
- d) **Ease of maintenance:** Neural networks can be updated with fresh data, making them useful for dynamic environments.
- e) **Neural networks** can be implemented in parallel hardware
- f) When an element of the neural network fails, it can continue without any problem by their parallel nature.

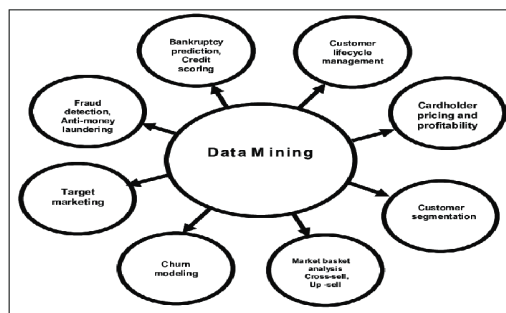


Figure.5: Application of data mining

4. CONCLUSION

In this paper, we have suggested Data mining technology provides a whole new way of exploiting large databases. In the wrong hands, this can be a problem. However, it is not the tools themselves, but the combination of tools and data that pose a problem. The obvious solution is to restrict access to data. Many data mining techniques can be implemented data to predict their future performance.

References

- [1]. Velmurugan T. et al., "Performance Evaluation of KMeans & fuzzy C-means Clustering Algorithm for Statistical Distribution of Input Data Points", European Journal of Scientific Research, Vol. 46, 2010.
- [2]. "Neural networks in data mining", dr. yashpal singh, alok singh chauhan, <http://www.jatit.org/volumes/research-papers/Vol5No1/1Vol5No6.pdf>
- [3]. Comparative Analysis of Data Mining Techniques on Educational Dataset, International Journal of Computer Applications (0975 – 8887) Volume 74– No.5, July 2013
- [4]. Jiawei Han and Micheline Kamber, "Data Mining Concept and Technique", Published by Morgan Kaufman, 2006.
- [5]. Arun K Pujari, "Data mining Technique", Published by Universities Press (I) Pvt. Ltd, Hyderabad, India.
- [6]. Hamidah Jantan, "Classification and Prediction of Academic Talent using Data Mining Technique", 14th International Conference on Knowledge based and Intelligent Information and Engineering Information pages 491-500.
- [7]. R.K. Somani, "Data Mining & Warehousing", College Book Centre, Chaura Rasta, Jaipur, India.
- [8]. Haykin, S., Neural Networks, Prentice Hall International Inc., 1999
- [9]. Berson, "Data Warehousing, Data-Mining & OLAP", TMH