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

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 7, Issue 11, November 2018

Overview of Data Mining

M. Smitha[1], V.J Rajakumar[2], Abdoulelbastoisaidabdou[3]

Student, Department of Computer Technology, Sri Krishna Arts & Science College, Coimbatore, Tamil Nadu, India[1,3]

Assistant Professor, Department of Computer Technology,

Sri Krishna Arts and Science College, Coimbatore, Tamil Nadu, India[2]

Abstract: Data mining is a process of extracting useful knowledge from a large number of data sets, by using any of its methodology. It is a field of intersection between machine language, database and statistics. it is used in many fields for extraction knowledge and information about that particular area. Data mining is used to discover knowledge out of data and presenting it in an understandable way to humans. There are different process and techniques used to carry out data mining successfully.

Keywords: Machine Learning, Statistics, Database

1. INTRODUCTION

Data mining is the process of extracting hidden in-formation in large number of data sets. We can easily retrieve data from database by means of data mining. Data mining is a process that takes data as input and outputs knowledge Some of the popular data mining techniques are classification algorithms, prediction analysis algorithms, clustering techniques. Data mining is also known as Knowledge Discovery in Database.

2. WHAT IS DATAMINING?

Data mining is the process of sorting through large data sets to identify patterns and establish relationships to solve problems through data analysis Data mining techniques are used in many **re-search areas, educational institutions, hospitals and marketing**. Data mining is used to identify the hidden patterns of data according to different methodologies for categorization into useful information, which is collected and arranged in common areas, such as data warehouses.

3. DATA MINING TASKS

- ✓ **Classification**
- ✓ **Prediction**
- ✓ **Clustering**
- ✓ **Association rules**
- ✓ **Sequence discovery**

Classification

Classification maps data into predefined groups or classes. It is often referred to because the classes are determined before the examining the data. Two real time aspects of classification applications are determining whether to make a bank loan and identifying credit risks.

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

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 7, Issue 11, November 2018

Prediction
Many real-world data mining aspects can be seen as predicting future data states based on past and current data. Prediction can come under the type of classification. The difference is that prediction is predicting a future state rather than current state. Prediction applications consist of flooding, speech recognition, machine learning and pattern recognition.

Clustering
Clustering is an unsupervised machine learning method that attempts to uncover the natural groupings and statistical distribution of data. Clustering is defined as group of data. The most similar data are grouped into clusters.

Association Rules
Association rules are created by analyzing data for frequent if/then patterns and using criteria support and confidence to identify the most important relationships an association rules is a model that identifies specific type of data association. This example is mainly implemented in retail sales community to determine the frequently purchased items together.

Sequence Discovery
Sequence discovery is otherwise called as sequence analysis is used to determine the sequential patterns in data. This pattern is based on time sequence of action. These patterns are similar to association in that data are found to be related, but the relationship is based on time. Unlike a market basket analysis which requires the items to be purchased over time in some order.

4. KNOWLEDGE DISCOVERY IN DATA-BASES(KDD) PROCESS

The term KDD process is often used in inter-changeably. This is mainly to discover useful patterns.

- ✓ **Selection**
- ✓ **Transformation**
- ✓ **Geometric**
- ✓ **Graphical**
- ✓ **Pixel-based**
- ✓ **Hybrid**

Selection: The data needed from datamining process is obtained from different and heterogeneous data source. This first step obtains the data from various databases, files and nonelectric sources takes the data from various databases, files.

Transformation: Data from different sources should converted into a common format for processing. Some of the data should be encoded and converted into most usable formats. This technique is used to make the data easier to mine and more useful, and to provide more meaningful results.

Geometric: Geometric techniques include the scattered box plot and geographical diagram techniques.

Pixel-Based: In this technique each data value is displayed by uniquely colored pixel.

Hybrid: The preceding approaches cab be combined into one display. Visualization tools can be used to summarize data as a data as a datamining technique itself.

5. DATAMINING ISSUES

- ✓ **Over fitting**
- ✓ **Outliers**
- ✓ **Visualization of results**
- ✓ **Missing data**
- ✓ **Noisy data**
- ✓ **Changing data**

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

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 7, Issue 11, November 2018

Over-fitting: Over fitting occurs when the model does not fit it to the future states. This may be caused by assumption's that are made about the data or may simply be caused by the small size of the training database. Over-fitting can arise under other problems as well, even though the data are not changing.

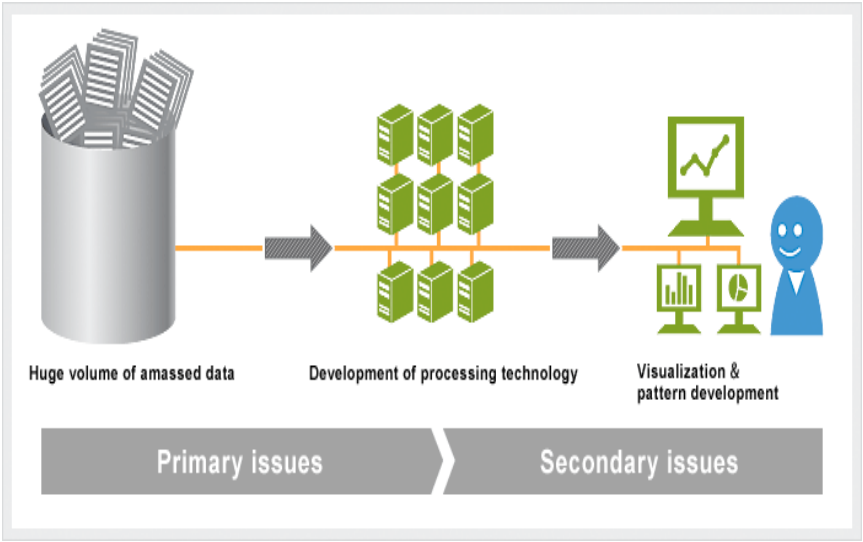
Outliers: There are many data that do not fit into the fit in to the desired model. This becomes more issues with large databases. If a mode is developed that included these outliers, that the model may not be-have well for data that are not outliers.

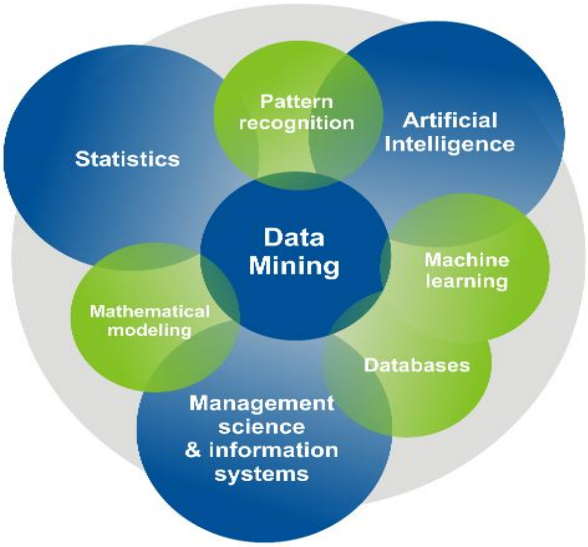
Visualization of Results: To easily view and understand the output of datamining algorithms, visualization of the results is helpful.

Missing Data: During the processing phase of KDD, missing data may be replaced with another estimates. To handle the approaches to missing data in KDD process can lead to invalid results in the data mining step.

Noisy Data: Some attributes values might be invalid or incorrect. These values are often corrected before running data mining.

Changing Data: Database cannot be assumed to be static. However, most data mining algorithms do assume a static database. It helps the algorithm be completely rerun anytime the database changes.





DOI 10.17148/IJARCCE.2018.71123

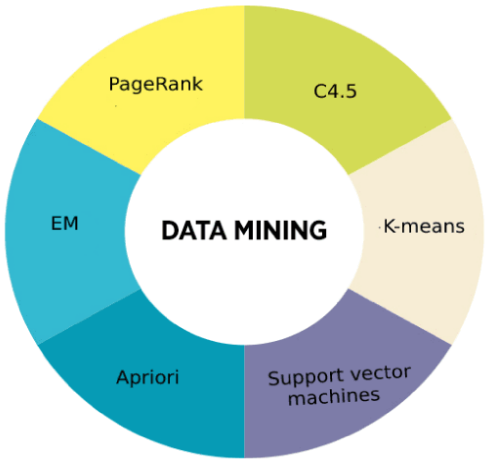
ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 7, Issue 11, November 2018

6. ALGORITHMS USED IN DATAMINING



Top Data mining algorithms

Data mining algorithms offer different techniques for identifying previously unknown characteristics, patterns, association, in the data which may be mined. These techniques are generally identified with data mining operations as database segmentation, predictive modelling, link analysis and deviation detection.

I. C4.5:

C4.5 is an algorithm that is used to implement a classifier in the form of a decision tree and has been developed by Ross Quinlan. When the decision tree is built, missing data. is simply ignored. That is the gain ratio is calculated by looking only a records that have a value for. That attribute. To classify a records that have a value for that attribute value, the value for that item can be predicted based on what is known about the attribute values for another records. Continuous data: the basic idea is to divide the data into ranges based on the attributes values for that item are found in the training sample there are two primary pruning strategies proposed in C4.5

II. K-MEANS:

k-means clustering that is also called as nearest centroid algorithm in which items are moved among sets of clusters in until desired set is reached. As such, it may be viewed as a type of squared error algorithm, although the convergence criteria need not be defined based on squared error. A high degree of similarity among elements in clusters is obtained, while a high degree of dissimilarity among elements in different clusters is obtained.

III. SUPPORTVECTOR MACHINES:

When it comes to machine learning, support vector machines that are also known as support vector networks are basically it will comes under supervised learning models that come with associated algorithms that are known for analyze data that are used for the analysis of regression and classification.

IV. APRIORI

Apriority is an algorithm that is used for frequent item set datamining and association rule learning overall transactional databases. The algorithm used for the identification of the individual items that are frequent used in the database and then extending them to larger item sets as long as frequently those item sets appear often enough in the database table. These frequent item sets that are determined by Apriori algorithm can be used for the determination of association rules which then highlight general trends of apriori.

V. EM(EXPECTATION-MAXIMIZATION):

An expectation–maximization (EM) algorithm, when it comes to statistics is an iterative method of algorithm that is used to find maximum a posteriori(MAP) or maximum similar estimates of parameters in statistical models, that basically depends on unobserved variables.

VI. PAGE RANK(PR)

Page Rank (PR) that was named after Larry Page who is one of the founders of Google is a PageRank algorithm it is used mostly to the purpose of search engines PageRank, it is the first algorithm that was used by the company is not the only algorithm that is being used by Google to order search engine results, but it is the best method of measuring the importance of website pages.

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

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 7, Issue 11, November 2018

VII. ADABOOST

Adaptive Boosting or otherwise called as AdaBoost, that has been discovered by Yoav Freund and Robert Schapiro is a machine learning meta-algorithm. The algorithm can be often used for the composition with many other types of learning algorithms in order to improve performance. AdaBoost is sensitive outliers as well as noisy data.

VIII. KNN

The k-nearest neighbors algorithm (k-NN) is a type of lazy learning algorithm or instance-based learning and it was seen as a non-parametric method that is used for classification and regression of datamining task .In both the mentioned cases, the input consists of the k closest training examples in the feature space and the output depends on whether the algorithm is being used for classification or regression. This K-Nearest Neighbor Algorithm is considered and is also among the simplest of all machine learning algorithms.

IX. NAİVE BAYES

When it comes to machine learning, Naive Bayes classifiers that are considered to be highly scalable are comes under the simple probabilistic classifiers that are based on the of Bayes' theorem has the help of independent assumptions between the features.

X. CART

CART is an algorithm that generally depends upon the classification and regression trees. CART is a decision tree learning technique that either displays outputs classification or regression trees and similarly like C4.5, CART is also a classifier. Many of the reasons that a user would use C4.5 algorithms are also applying to that of CART, since both are coming under decision tree learning techniques and features like ease of interpretation and explanation are also implemented to CART as we

CONCLUSION

This paper gives introduction about all techniques involved in data mining, the process of discovering interesting knowledge from large amounts of data stored in information repositories. One of the big-gest deal for data mining technology is monitoring the uncertain data which may be caused by outdated resources, sampling errors, or imprecise calculation. Future research will involve the development of new techniques for incorporating uncertainty management in datamining.

REFERENCES

[1]. Yao, Hong, Hamilton, H., and Butz, C. J. 2004. A Foundational Approach to Mining Item set Utilities from Databases, Proceedings of the Third SIAM International Conference on Data Mining, Orlando, Florida, pp. 482-486.
[2]. Chu, C., Tseng, V. S., and Liang, T. 2008. An efficient algorithm for mining temporal high utility item sets from data streams. J. Syst. Softw. 81, 7 (Jul. 2008), 1105-1117
[3]. Hu, J., Mojsilovic, A. High-utility Pattern Mining: A Method for Discovery of High-utility Item Sets, Pattern Recognition, Vol. 40, 3317-3324. Jyothi Pillai / International Journal on Computer Science and Engineering (IJCSE) ISSN: 0975-3397 Vol. 3 No. 1 Jan 2011 399
[4]. Ale, J. M. and Rossi, G. H. (2000). An Approach to Discovering Temporal Association Rules. In Proceedings of the 2000 ACM Symposium on Applied Computing, Vol.1, J. Carroll, E. Damiani, H. Haddad, and D. Oppenheim, Eds. SAC '00. ACM Press, New York, NY, pp 294-300.
[5]. Yao, H. and Hamilton, H, J. 2006. Mining Item set Utilities from Transaction Databases, Data and Knowledge Engineering, 59(3): 603-626
[6]. Liu, Y., Liao, W., and Chaudhary, A. 2005. A Fast High Utility Item sets Mining Algorithm. Proceedings of the Utility-Based Data Mining Workshop.
[7]. Teng, W. G., Chen, M. S., and Yu, P. S. 2003. A Regression-Based Temporal Pattern Mining Scheme for Data Streams. Proceedings of the 29th International Conference on Very Large Databases, pp 93-104.
[8]. Ahmed, C. F., Tanbeer, S. K., Jeong, B-S, and Lee, Y-K. 2008. Handling Dynamic Weights in Weighted Frequent Pattern Mining, IEICE Trans. Information and Systems, Vol. E91-D:2578-2588.
[9]. Han, J., Pei, J. and Yiwen, Y. 2000. Mining Frequent Patterns Without Candidate Generation. Proceedings ACM-SIGMOD International