IJCSE International Journal of Computer Sciences and Engineering Open Access
Survey Paper Volume-3, Issue-12 E-ISSN: 2347-2693

A Survey on Data Mining Algorithms

Mrs. Uma[1], Dr. L. Jayasimman[2], V. Upendran[3]

[1,2*,3,4]*Department of Computer Science,*
Srimad Andavan Arts & Science College, India
www.ijcseonline.org

Received: 22/Nov/2015 Revised:05/Dec/2015 Accepted:22/Dec/2015 Published: 31/Dec/21015

***Abstract*— –** This research paper presents a review of data mining algorithms which is useful to predict the accuracy in the educational domain. It discusses and compares various data mining algorithms in the educational domain. Nowadays it is becoming increasingly important to develop powerful tools for analysis of the enormous data that is stored in databases and data warehouses, and mining such data and arriving at values focused on interesting knowledge from it. Modern organizations focus and put on much thrust in developing data mining procedures and derive benefit from it. Data mining is a process of inferring knowledge from such huge data. There are three major components in Data Mining Viz Clustering, Classification and Association Rules. By discussing the various algorithms this research paper has the scope to extend it further, in order to develop it further in the direction of new and innovative algorithms. The main aim of the paper is to study the effectiveness of various algorithms in the educational environment.

***Keywords*—** *Clustering, Classification, Association rule, Data mining, innovative algorithms*

I. INTRODUCTION

Data mining is a knowledge discovery process from the hidden large data. It has many processes such as finding correlations, finding similarities, finding new interpretations, etc. Nowadays many data mining software tools are used to extract the useful hidden and unknown data accurately. Predicting the results with maximum accuracy needs well-constructed computer algorithms and mathematical concepts. Statistical applications play a vital role in algorithm development. Combination of various fields like computer science, mathematics and biological science can be used to find the useful data with maximum accuracy. Genetic algorithm is one of the effective ways of extracting new information. Biological concepts such as gene combination, mutation and pairing are useful in the development of innovative algorithms.

Universal concepts in data mining:

Universal concepts among all the data mining methods are as follows,

(1) Prediction
(2) Clustering
(3) Relationship Mining

(1) **Prediction:**

Prediction is useful to construct a model. Types of Prediction:

- Classification
- Regression
- Density estimation

(2) **Clustering:**

Splitting the large data using similar points called clusters is the main aim of clustering data. **I**n the process of clustering data analyzing a cluster of data and generate a list of grouping rules which can be used to classify future data [1]. For example, one may classify students' results and provide the grades which describe each class or subclass. This has much common ways of traditional work involving statistical calculations and machine learning. However, there are important new issues which arise because of the volume or size of the data. One of the important problems in data mining is the classification-rule learning which progresses through finding rules that partition given data into predefined classes. In the data mining domain where millions of records compromising a large number of attributes are involved, the execution time of existing algorithms can become prohibitive, especially in interactive applications.

(3) **Relationship mining:**

It is a useful method to find the relationships between any two or more variables. Types of Relationship mining:

- Finding Association
- Finding Sequential pattern
- Finding Correlation method

© 2015, IJCSE All Rights Reserved

- Finding Casual data [11].

Association rule is defined as a rule which implies certain association relationships within a set of objects in a database. The process is to discover a set of association rules at various levels of abstraction within the respective set(s) of data in a database. For example, one may discover a set of student grade often occurring together with certain range of marks in various subjects and further study the reasons behind them. In the process, discovering interesting association rules in databases may show some useful patterns for decision support, marketing analysis, financial views, medical diagnosis, and various other applications, it has attracted a lot of importance nowadays in data mining research. Data interpretation and analysis require rules involving iterative scanning of huge transaction or relational databases which is relatively costly in processing. Hence, efficient mining of various association rules involving transaction and/or relational databases are analyzed and studied [10, 12].

II. CLASSIFICATION ALGORITHMS

Classification is a data mining (machine learning) technique, used to predict group membership for data instances.

In Data classification our aim is towards developing a description or model for each class in a database, relevant to the features present in a set of class-labeled training data.[1] There have been many data classification methods analyzed, studied, including decision-tree methods ,such as, statistical methods, rough sets, neural networks, database-oriented methods ...

Data Classification Methods

- **Statistical Algorithms:** Statistical analysis systems such as SAS and SPSS have been used by analysts in the process of detecting unusual patterns and explain patterns using statistical models such as linear models.
- **Neural Networks:** Artificial neural networks replicate the pattern-finding capacity of the human brain and hence suggested by researchers applying Neural Network algorithms similar to pattern mapping. Neural networks have been applied successfully in a few applications where classification is involved [2].
- **Genetic algorithms:** Optimization techniques use processes such as genetic combination, mutation, and natural selection in a select design based on the concepts of natural evolution.
- **Nearest neighbor method:** A technique which classifies each record within a dataset relevant to a combination of the classes of the k record(s) most similar to it in a historical dataset. This is referred to the k-nearest neighbor technique.
- **Rule induction:** The extraction of useful *if-then* rules from data, with reference to the statistical significance.
- **Data visualization:** Data visualization refers to the visual interpretation of complex relationships in multidimensional data.

Clustering Algorithms

According to Mike Chapple, 'clustering is a data mining (machine learning) technique used to place data elements into related groups without advance knowledge of the group definitions'. Popular clustering techniques include k-means clustering and expectation maximization (EM) clustering. Clustering algorithms are categorized based on their cluster model [15,18].

Connectivity based clustering

Connectivity based clustering, also referred to as hierarchical clustering and is based on the core idea of objects being more related to objects nearby more than to objects farther away [16, 17]. As such, these algorithms connect -objects- to form –clusters- based on their distance. A cluster can generally be described largely by the maximum distance needed to connect parts of the cluster [3]. Pointing to various distances, different clusters occur, which can be represented by using a dendrogram that explains where the common name hierarchical clustering" is taken from These algorithms instead of providing a single partitioning of the data set, they provide an extensive hierarchy of clusters which merge with each other at certain distances. In a dendrogram, the y-axis represents the distance at which the clusters merge, while the objects are represented along the x-axis such that the clusters don't mix.

In addition to the usual choice of distance functions, the user also decides upon to the linkage criterion that is a cluster consists of multiple objects and there are multiple candidates to compute the distance use. Usually popular choices are known as single-linkage clustering (minimum of object distances), complete linkage clustering (maximum of object distances) or UPGMA Unweight Pair Group Method with Arithmetic Mean which is also known as average linkage clustering [9, 14]. Moreover, hierarchical clustering may be computed agglomerate (starting with single elements subsequently aggregating them into clusters) or divisive that is starting with the complete data subsequently divide it into partitions. When these methods

© 2015, IJCSE All Rights Reserved

are roughly easy to understand, arrived at results are not always so easy to use, as they will not produce a unique partitioning of the data set, but in a hierarchy the user is still required to choose appropriate clusters from. These methods are not very robust towards outliers, which results in either showing up as additional clusters or even cause other clusters to merge which is also called as "chaining phenomenon", in particular with single-linkage clustering [4, 5]. When studying data mining community these methods are considered as a theoretical foundation of cluster analysis, but very often considered obsolete. They do however provide inspiration for many later methods such as density based clustering.

III. ASSOCIATION RULE ALGORITHMS

Association rule is defined as a rule which implies certain association relationships within a set of objects which occur together or one implies the other in a database. Given a set of transactions, where every transaction is a set of literal which is also called items, an **association rule** is an expression represented by the form X Y where X and Y are sets of items [6]. The meaning of such a rule is that transactions of the database which contain X also tend to contain Y.

Apriori Algorithm

An association rule mining algorithm, **Apriori** is developed for rule mining involving large transaction databases by IBM's Quest project team [3]. An item set is a non-empty set of items.

We can realign the problem of mining association rule into two parts

- List all combinations of items that have transaction support above the minimum support. Name those combinations frequent item sets.
- Apply the frequent itemsets to generate the desired rules. The general idea behind is that if, say, ABCD and AB are frequent itemsets, then we can determine if the rule AB CD holds by computing the ratio r = support(ABCD)/support(AB). The rule holds only if r >= minimum confidence. Here it be noted that the rule will have minimum support because ABCD is frequent. Given below is the Apriori algorithm used in Quest for finding all frequent item sets.

Procedure AprioriAlg()

begin

L1 := {frequent 1-item sets};

for (k := 2; Lk-1 0; k++) **do** {

Ck= apriori-gen(Lk-1) ; // new candidates

for all transactions t in the dataset **do** {

for all candidates c Ck contained in t **do**

c:count++

}

Lk = { c Ck | c:count >= min-support}

}

Answer := k Lk

end

The process makes multiple passes over the database. During the first pass, the algorithm simply counts the item occurrences to determine the frequent 1-itemsets (itemsets with 1 item). In the subsequent passes, say pass k, consists of two phases. Initially, the frequent itemsets Lk-1 (the set of all inclusive frequent (k-1)-itemsets) found in the (k-1)th pass are utilized to generate the candidate itemsets Ck, using the apriori-gen() function. This role of the function is to first join Lk-1 with Lk-1, joining condition being the lexicographical ordered first k-2 items are the same. Next, it deletes all those itemsets from the join result that have some (k-1)-subset that is not in Lk-1 yielding Ck. [7].

Now the algorithm scans the database. For every transaction, it determines which of the candidates in Ck are contained within the transaction using a hash-tree data structure and increments the count of such candidates. Towards the end of the pass, Ck is examined to determine among the data which of the candidates are frequent, yielding Lk [8]. The algorithm finally terminates its process when Lk becomes empty and further process is null.

IV. CONCLUSION

Any algorithm which is proposed for mining data will have to account for out of core data structures. Many data classification methods are analyzed through application of various algorithms. Wherein, statistical algorithms explore the data by analysis and linear models, data visualization techniques apply the model of 'visual interpretation of data. Connectivity based clustering algorithm Clustering algorithm is a non-hierarchical method redefining data into components equal to the required number of clusters. By studying the various classification technique, data miners can predict the innovative algorithm in order to enhance the accuracy.

© 2015, IJCSE All Rights Reserved

REFERENCES

[1]. Hemlata Sahu et al. A Brief Overview on Data Mining Survey, International Journal of Computer Technology and Electronics Engineering (IJCTEE) Volume 1, Issue 3, pp. 114-121. ISSN 2249-6343.

[2]. Chen, Shaoyong, Min Lin, Huanming Zhang. Research of mobile learning system based on cloud computing,e- Education, Entertainment and e-Management. International conference on IEEE,vol.,no.,pp.121-123,27-29 Dec.2011

[3]. Shuai,Qin,zhou Ming-quan. Cloud computing promotes the progress of m-learning, Uncertainty Reasoning and knowledge Engineering. International conference on IEEE, vol.2, no.,pp. 162-164,4-7 Aug.2011.

[4]. Hirsch, Benjamin, Jason WP Ng., "Education beyond the Cloud: Anytime-anywhere learning in a smart campus environment", Internet Technology and Secured Transactions (ICITST), International conference on IEEE, vol., no., pp. 718-723, 11-14 Dec.2011.

[5]. Carsten Ullrich, Ruimin Shen, Ren Tong and Xiaohong Tan. A mobile live video learning system for large scale learning system design and Evaluation. IEEE Transactions on learning Technologies. Vol .3, No.1, January –March 2010.

[6]. N. Mallikharjuna Rao, C. Sasidhar, V. Satyendra Kumar. Cloud Computing Through Mobile – Learning.

[7]. Mohamed Osman M. EI-Hussein and Johannes C.Cronje. Defining Mobile Learning in the Higher Education Landscape. Educational Technology & Society.13 (3), 12-21.

[8]. Mrs. Bhuvana Raghvendra Bajpai. M- Learning & Mobile Knowledge Management: Emerging new stages of E-Learning & Knowledge Management. IEEE ICC 2011

[9]. Kritika Verma, Sonal Dubey, Dr.M.A.Rizvi. Mobile Cloud a New vehicle for Learning: M-Learning its issues and Challenges. International Journal of Science and Applied Information Technology, Volume 1,No.3, July – August 2012.

[10]. Minjuan Wang, Yong Chen and Muhammad Jahanzaib Khan. Mobile Cloud Learning for Higher Education: A Case Study of Moodle in the cloud. The International Review of Research in Open and Distance Learning.

[11]. Hoand T.Dinh, Chonho Lee, Dusit Niyato and Ping Wang. A Survey of Mobile Cloud Computing: Architecture, Applications and Approaches. Wireless Communication and Mobile Computing, 2013; 13: 1587-1611.

[12]. Meilian Chen, Yan Ma, Yikun Liu, Fan Jia, Yanhui Ran and Jie Wang. Mobile learning System Based on Cloud Computing. Journal of Networks, Volume 8, No. 11, November 2013.

[13]. Mohssen M. Alabbadi. Mobile Learning (m-Learning) Based on Cloud Computing: M-Learning as a Service (mLaaS). The Fifth International conference on mobile Ubiquitous Computing, Systems, Services and Technologies, 2011.

[14]. Pragaladan R, Leelavathi. M. A Study of Mobile Cloud Computing and Challenges. International Journal of Advanced Research in Computer and Communication Engineering. Volume 3, Issue 7, July 2014.

[15]. Anwar Hossain Masud, Xiaodi Huang. A Cloud Based M-Learning Architecture for Higher Education. http://www.researchgate.net/publications/235758554

[16]. Hossein Movafegh Ghadirli and Maryam Rastgarpour. A Paradigm for the Application of Cloud Computing in Mobile Intelligent Tutoring Systems. International Journal of Software Engineering & Applications (IJSEA), Volume 4, No.2, March 2013.

[17]. Mike chapple, Data mining definition, downloaded from the web site http://databases.about.com / od / datamining /g/clustering.html.

[18]. Manjunath et al. world wide web information retrieval using clustering RD, International Journal for Scientific Research & Development| Vol. 2, Issue 04, 2014 | ISSN (online): 2321-0613,

© 2015, IJCSE All Rights Reserved