International Journal of Science and Research (IJSR)
ISSN: 2319-7064
Impact Factor (2018): 7.426

An Approach to Data Mining Algorithm for OLAP Databases

Gaurav Agnihotri

Department of Computer Science & Engineering, GNA University, Punjab, India

Abstract: ***As there are number of Data mining algorithms have been proposed in previous years, but it is tough to define comprehensive study to compare these algorithms. k-means algorithm is the first algorithm proposed in this field. With respect to time number of changes proposed in k-means to enhance the performance in term of time. This paper proposes an innovative utility sentient approach for the mining of interesting association patterns from OLAP database. This algorithm uses the results of this analysis to define the parameters of the mining model. While you can use different algorithms to perform the same business task, each algorithm produces a different result, and some algorithms can produce more than one type of result.***

Keywords: k-means Rules; rule mining algorithms; minimum support; computation power; frequent Items

1. Introduction

In Data Mining, **k-means** is a classic algorithm for learning association rules. K-means is designed to operate on databases containing transactions (for example, collections of items bought by customers, or details of a website frequentation). Other algorithms are designed for finding association rules in data having no online analytical processing, or having no timestamps (DNA sequencing).

As is common in association rule mining, given a set of *itemsets* (for instance, sets of retail transactions, each listing individual items purchased), the algorithm attempts to find subsets which are common to at least a minimum number k of the itemsets. K-means uses a "top up" approach, where frequent subsets are extended one item at a time (a step known as *candidate generation*), and groups of candidates are tested against the data. The algorithm terminates when no further successful extensions are found.

A large supermarket tracks sales data by item-keeping unit (IKU) for each item, and thus is able to know what items are typically purchased together. Apriori is a moderately efficient way to build a list of frequent purchased item pairs from this data. Let the database of transactions consist of the sets {5,6,7,8}, {5,6}, {6,7,8}, {6,7}, {5,6,8}, {7,8}, and {6,8}. Each number corresponds to a product such as "butter" or "bread". The first step of Apriori is to count up the frequencies, called the supports, of each member item separately.The table 1 explains the working of apriori algorithm.

Table 1: Support-Item

Item	*Support*
5	3
6	6
7	4
8	5

We can define a minimum support level to qualify as "frequent," which depends on the context. For this case, let min support = 3. Therefore, all are frequent. The next step is to generate a list of all 2-pairs of the frequent items. Had any of the above items not been frequent, they wouldn't have been included as a possible member of possible 2-item pairs. In this way, Apriori *prunes* the tree of all possible sets. In next step we again select only these items (now 2-pairs are items) which are frequent:

Table 2: Relations of Support-Item

Item	*Support*
{1,2}	3
{2,3}	3
{2,4}	4
{3,4}	3

And generate a list of all 3-triples of the frequent items (by connecting frequent pairs with frequent single items). In the example, there are no frequent 3-triples. Most common 3-triples are {5,6,7} and {6,3,7}, but their support is equal to 6 which is smaller than our min support.

- In data mining, association rules are useful for analyzing and predicting customer behavior. They play an important part in shopping basket data analysis, product clustering, catalog design and store layout.

- K-means picks points in multidimensional space to represent each of the k-clusters. These are called centroids. Every patient will close to 1of the k centroids. They hopefully won't all be closest to the same one so they'll form a cluster around their nearest centroid.
- K-means then find the center of each of the k cluster based on its cluster members (yep, using the patient vectors). This center becomes the new centroid for the cluster. Since the centroid is in different place now, patients might be closer to other centroids. In other words they may change cluster membership. Most would classify k-means as unsupervised. Other than specifying the number of clusters, k-means learns the cluster on it own without any information about which cluster an observation belongs to.

2. Problem Definition

- Repeatedly scanning the OLAP database. For each element of each circular candidate set k*, it must be verified whether to join the frequent item set Lk through scanning the database. If there is a large frequent item set contained

10 items, then it is need to scan the transaction database at least 10 times, which will bring a great I/O load.

- Most of the strong rules found can be inferred from domain knowledge and do not lead to new insights. It only tells the presence and absence of an item in transactional Database.
- The strong rules found do not lend themselves to any actions and are hard to interpret**.** These can be removed by using attributes like weight and quantity, weight attribute will give user an estimate of how much quantity of item has been purchased by the customer, profit attribute will calculate the profit ratio and tell total amount of profit an item is giving to the customer.

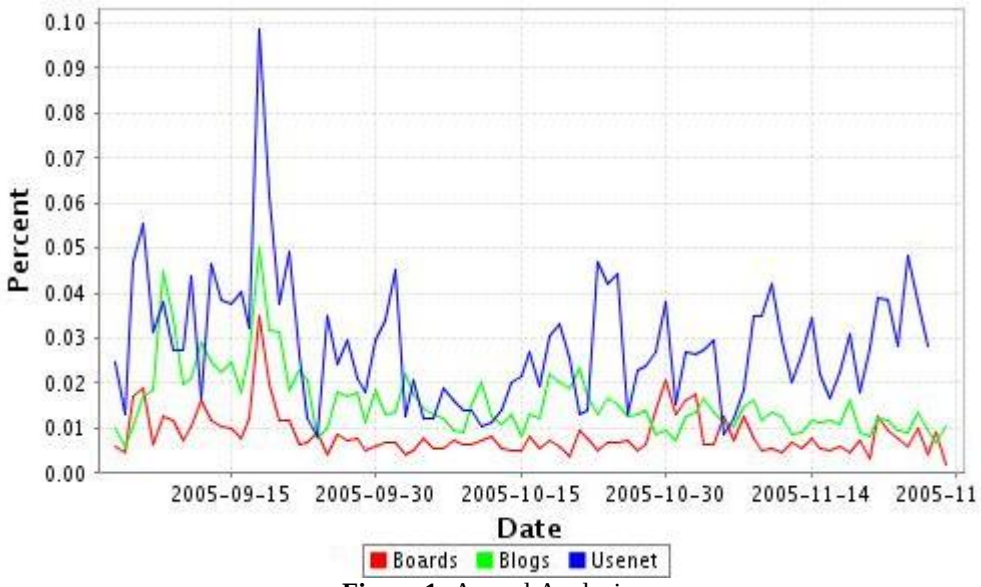

Figure 1: Annual Analysis

3. Methodology

Ck: candidate item set of size k, Lk: frequent item set of size k
L1 = {frequent items};

For (k=1; Lk != null; k++) do begin Ck+1 = candidates generated from Lk; For each transaction t in database do Increment the count of all candidates in Ck+1 that are contained in t

- Specify number of cluster k;
- Initialize centroids by first shuffling the data set and then randomly selecting k data points;
- MAX RULE LENGTH states the maximum number of items (attribute = value) of rules ;
- LIFT FILTERING states the minimum of LIFT.
- LEARNING SET RATIO states the proportion of the dataset used for the learning phase.

Pseudo-code for k-means

$$J = \sum_{i=1}^{m} \sum_{k=1}^{K} w_{ik} \|x^i - \mu_k\|^2 \qquad (1)$$

L1 = {frequent items};
for (k = 1; Lk !=∅ ; k++) do
Ck+1 = candidates generated from Lk; for each transaction t in database do increment the count of all candidates in Ck+1 that are contained in t

Lk+1 = candidates in Ck+1 with min_support return ∪k Lk;

- 15 attribute/subset evaluators + 10 search algorithms for feature selection
- 3 algorithms for finding association rules
- More algorithms being added
- Options to customize using the Java source code is made available
- Software Platform: Java based
- 49 data preprocessing tools
- 76 classification/regression algorithms.
- Custom extensions and plug ins can be developed
- Excellent mailing and discussion lists available.

Tanagra The main purpose of Tanagra Tool is to give researchers and students an easy-to-use data mining software, conforming to the present norms of the software development in this domain (especially in the design of its GUI and the way to use it), and allowing to analyses either real or synthetic data.

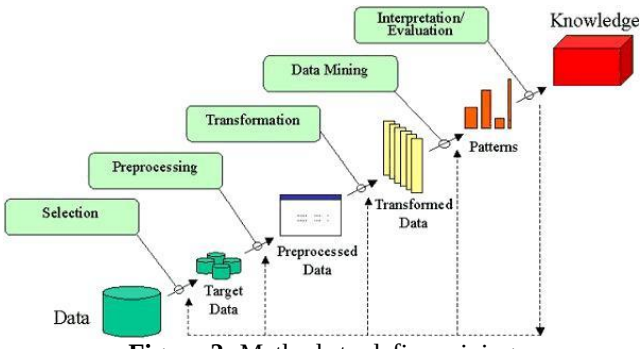

Figure 2: Methods to define mining

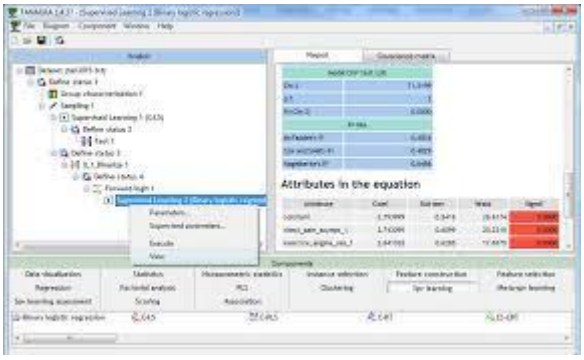

Figure 3: Working of Tanagra tool

The second purpose of TANAGRA is to propose to researchers an architecture allowing them to easily add their own data mining methods, to compare their performances. TANAGRA acts more as an experimental platform in order to let them go to the essential of their work, dispensing them to deal with the unpleasant part in the program of this kind of tools: the data management.

4. Tool of Data Collection & Analysis

- Tanagra: A collection of open source ML algorithms
- Pre-processing
- Classifiers
- Clustering
- Association rule

Figure 4: Data Management

The third and last purpose, in direction of novice developers, consists in diffusing a possible methodology for building this kind of software. They should take advantage of free access to source code, to look how this sort of software is built, the problems to avoid, the main steps of the project, and which tools and code libraries to use for. In this way, Tanagra can be considered as a pedagogical tool for learning programming techniques.

5. Conclusion

The conclusion to this work is that Apriori algorithm is applied on the transactional database. By using measures of apriori algorithm, frequent itemsets can be generated from the database. K-means algorithm is associated with certain limitations of large database scans. This algorithm reduces the number of scans in the database and improves efficiency and computing time by taking the advantage of Advance Mining Techniques. By experiment results, it can obtain higher efficiency.

References

[1] Knowledge and Cache Conscious Algorithm Design and Systems Support for Data Mining Algorithms, 1-4244-0910-1/07/$20.00_c 2007 IEEE.
[2] Data Mining Technique for Expertise Search in a Special Interest Group Knowledge Portal, 2011 3rd Conference on Data Mining and Optimization (DMO), 28-29 June 2011, Selangor, Malaysia.
[3] Anomaly Detection for PTM's Network Traffic Using Association Rule, 2011 3rd Conference on Data Mining and Optimization (DMO) , 28-29 June 2011, Selangor, Malaysia.
[4] Mining Association Rules Based On Cloud Model and Application In Credit Card Marketing, Yan-Li Zhu,Yu-Fen Wang, Shun-Ping Wang, Xiao-juan Guo, 2010 Asia-Pacific Conference on Wearable Computing Systems.
[5] Review of Association Rule Mining Algorithm in Data Mining, xu Chil Network Center,Hunan City University Hunan City University Yiyang, P.R .China, 2011 IEEE.
[6] The Research of Data Mining in AHM Technology based on Association Rule, Jia Baohui,Wang Yuxin Aeronautical engineering college Civil Aviation University of China.
[7] An Improved Association Rule Mining Technique for Xml Data Using Xquery and Apriori Algorithm R.Porkodi, V.Bhuvaneswari, R.Rajesh, T.Amudha , 2009 IEEE International Advance Computing Conference (IACC 2009).
[8] The Design of Algorithm for Data Mining System Used for Web Service, Ren Yanna, Lv Suhong, Wang Qiang College of Information and Management Sciences Henan Agricultural University Zhengzhou 450002, China, 2011 IEEE.
[9] Study on Application of Apriori Algorithm in Data Mining Yanxi Liu School of Science Changchun University,China130022, 2010 Second International Conference on Computer Modeling and Simulation.
[10] Data Warehousing, Data Mining & OLAP, Alex Berson, P No.
[11] **Gaurav Agnihotri** is a Astt. Professor in department of information technology at GNA University, Phagwara, Punjab, India. He has done B.Tech in Computer Science & Engineering, M.Tech in Information Technology and currently pursuing PhD degree in computer engineering from the Punjabi University, Patiala. Mr. Gaurav has more than 9 years teaching and research experience. He has supervised more than 20 M.Tech. students in Data mining , Data Structure and cloud computing