International Conference on Computer Applications 2016 [ICCA 2016]

ISBN	978-81-929866-5-4
Website	icca.co.in
Received	14 – March– 2016
Article ID	ICCA002

VOL	05
eMail	icca@asdf.res.in
Accepted	02 - April – 2016
eAID	ICCA.2016.002

Study on Positive and Negative Rule Based Mining Techniques for E-Commerce Applications

Kavita Yadav[1], Pravin G Kulurkar[2]
[1,2]Vidharbha Institute of Technology, Nagpur

Abstract- *In the recent years the scope of data mining has evolved into an active area of research because of the previously unknown and interesting knowledge from very large database collection. The data mining is applied on a variety of applications in multiple domains like in business, IT and many more sectors. In Data Mining the major problem which receives great attention by the community is the classification of the data. The classification of data should be such that it could be they can be easily verified and should be easily interpreted by the humans. In this paper we would be studying various data mining techniques so that we can find few combinations for enhancing the hybrid technique which would be having multiple techniques involved so enhance the usability of the application. We would be studying CHARM Algorithm, CM-SPAM Algorithm, Apriori Algorithm, MOPNAR Algorithm and the Top K Rules.*

Keywords: *Data Mining, CHARM Algorithm, CM-SPAM Algorithm, Apriori Algorithm, MOPNAR Algorithm and the Top K Rules.*

1. INTRODUCTION

In today's world human beings are using multiple applications to ease their work. Every day a lot of data is generated in every field. The data can be in the form of Documents, graphical representation like picture or video and there can be multiple records also. Since there are multiple types of data there can be multiple types of format proper action should be taken for their better utilization of the available data. Since when the user wants to use the data the data can be retrieved in the proper format and information.

The technique to retrieve knowledge from the data is termed as data mining or knowledge hub or simple Knowledge Discovery process (KDD).The important reason that attracted a great deal of attention in information technology the discovery of useful information from large collections of data industry towards field of "Data mining" is due to the perception of "we are data rich but information poor". This perception is there because we have a very huge amount of data but we cannot convert it to useful information for decision making in different fields. To produce knowledge we require a lot of data and which could be in all possible formats like audio video images documents and much more. In data mining to get the full advantage not only the retrieval but also the tool for extraction of the essence of information stored, summarization of data and discovery of patterns in the data too is required for the knowledge extraction.

Since there is no lack of supply of data we are having a lot of data in different formats therefore it is important to develop a system which can convert this data into knowledge to help in decision making processes. The data mining tools can help in predicting behavior and future trends which can help organizations to make future knowledge-driven decisions. The data mining tools provides various features like automated, prospective analyses can help a better decision making scenario. In this paper we would be studying various data mining techniques and will review which technique can be used in the hybrid of the data mining technique.

Cite this article as: Kavita Yadav, Pravin, G Kulurkar. "Study on Positive and Negative Rule Based Mining Techniques for E-Commerce Applications". *International Conference on Computer Applications 2016*: 06-09. Print.

2. Charm Algorithm

CHARM is an efficient data mining technique which is used for enumerating the set of all frequent closed data item-sets. There are multiple innovative ideas implemented in the development of charm. This technique simultaneously explores both the item-set space and transaction space over item set-tides tree that is the search space of the database.

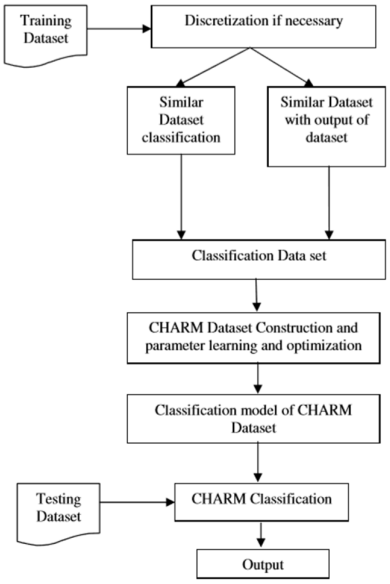

Figure 1: Charm algorithm

This technique uses a highly efficient hybrid search method that skips many levels of the tree to quickly identify the frequent closed set instead of having multiple possible subset to analysis. Fast Hash-based approach is used to eliminate item set during the execution. Charm also able to utilize a novel vertical data representation called diffset for fast frequency computations. Diffsets also keep track for differences in the tids of a candidate pattern from its prefix pattern. Since diffset reduce the size of the memory required to store intermediate result, therefore the entire working in some memory even for huge database.

3. CM-Spam Algorithm

In data mining getting useful patterns is a challenging task. In sequential database many techniques have been proposed for getting the patterns. A subsequence is called sequential pattern or frequent sequence if it frequently appears in a sequence database and its frequency is no less than a user-specified minimum support threshold minsup. This sequential pattern is very important in datamining as it helps in analysis of multiple applications like web medical data, program executions, click-streams, e-learning data and biological data. There are several efficient algorithm present for getting patterns amongst them the most efficient is the CM-SPAM Algorithm.

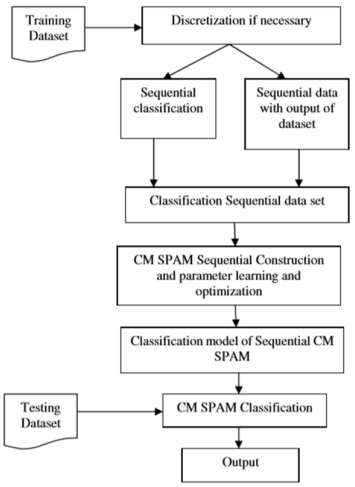

Figure: CM-spam algorithm

Cite this article as: Kavita Yadav, Pravin, G Kulurkar. "Study on Positive and Negative Rule Based Mining Techniques for E-Commerce Applications". *International Conference on Computer Applications 2016*: 06-09. Print.

The figure above shows the CM-SPAM Algorithm here first the trained dataset with known input and output is given as the input to the system. The data is divided equally if it is necessary as it gives the sequential classification from the known input and sequential classification form the known output. The combination of both the above dataset it provides us the classification sequential dataset.Then CM SPAM algorithm is applied to get the sequential construction, parameter learning and optimized dataset from the sequential dataset. All these combinations of parameters give us the Classification model of Sequential CM SPAM which is then used on testing the dataset for the output.

4. Apriori Algorithm

The apriori algorithm is one of the influential algorithm for mining for Boolean association rules. The key concept of the algorithm lies within Frequent Itemsets they are those sets of item which has minimum support and are denoted by Lifor ithItemset in the data. Secondly we have Apriori Property it is nothing butany subset of frequent itemset must be frequent and lastly we have join Operation it is required to find Lk which is a set of candidate k-itemsets is generated by joining Lk-1with itself. Here we will first find the frequent items the sets of items that have minimum support. Here a subset of a frequent itemset must also be a frequent itemset else the algorithm will not work that is if {t1 t2} is a frequent itemset both {t1} and {t2} should be a frequent itemset else the technique will fail to give a proper output. Once this is done iteratively find the frequrntitemset from 1 to k and use them to generate association rules.

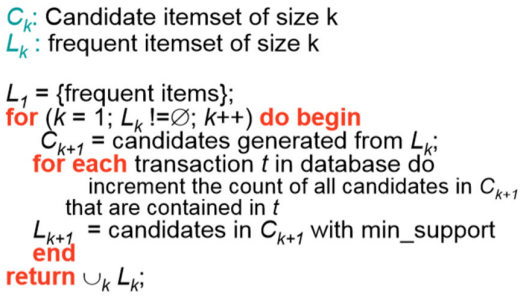

Figure: Apriori algorithm

The above figure shows the Apriori Algorithm. To enhance the efficiency of the apriori algorithm we have multiple techniques like Hash Based itemset counting, here a k itemset which has a hashing bucket count below the threshold cannot be tearmed as frequent. Another method is the Transaction reduction, here those transaction are ignored which doesn't contains the k itemset. The third method is the Partitioning in which if any item is frequent in the database it should also be frequent in partitions. Another method is sampling here the mining is done on the samples or the subsets f the data. Lastly we have Dynamic itemset counting here if all the subset of set is frequent then only it is selected.

5. Mopnar Algorithm

MOPNAR is an extension of multi-objective evolutionary algorithm (MOEA). It helps in mining with a low computational cost a reduced set of positive and negative QARs that are easy to understand and have good tradeoff between the number of rules, support, and coverage of the dataset. The main focus of the algorithm is to obtain a reduced set of PNQARs which are having good tradeoff considering three objectives which are comprehensibility, interestigness and performance. In order to perform a learning of rules it extends the traditional MOEA model. It also introduces two new components namely EP and Restarting process.

To decompose the MOEA it decomposes the muliobjective optimization problem into N scalar optimization. It uses EA to optimize the subproblems gathered. In this system to store all the nondominated rules found, provoke diversity in the population, and improve the coverage of the datasets the EP and the restart is introduced. Here EP will contain all the nondominated rules found and it will also generate the updated offspring for each solution. Since the size of EP is not fixed we can store a large number of rules and can reduce the size of population. Whereas restarting process here deals with the local optima and provoke diversity in the population. This process is applied when number of new individuals of the population in one generation is less than α% of the size of the current population.

The algorithm which MOPNAR uses is as follows:
Input:

1. N population size;
2. nTrials number of evaluations;

Cite this article as: Kavita Yadav, Pravin, G Kulurkar. "Study on Positive and Negative Rule Based Mining Techniques for E-Commerce Applications". *International Conference on Computer Applications 2016*: 06-09. Print.

3. m number of objectives;
4. Pmut probability of mutation;
5. λ1, …λN a set of N weight vectors;
6. T the number of weight vectors in the neighborhood of each weight vector;
7. δ the probability that parent solutions are selected from the neighborhood;
8. ηr the maximal number of solutions replaced by each child solution;
9. γ factor of amplitude for each attribute of the dataset;
10. α difference threshold.

Output: EP

6. Top K Rules Algorithm

As we have studied above all the above algorithms are depending on the threshold which leads to that the current algorithm leads to very slow execution and it generates excess, less or no results depending on the conditions and it also sometime omits valuable information to solve all the above disadvantages Top k rules came into the picture here the k would be the number of association rules to be found and is set by the user. It does not follow the traditional association rules here we can use rules with a single consequent else mining association rules from a stream instead of a transaction database.

```
TOPKRULES(T, k, minconf) R := Ø. L := Ø. minsup := 0.
1.  Scan the database T once to record the tidset of each item.
2.  FOR each pairs of items i, j such that |tids(i)| ×|T| ≥ minsup and |tids(j)| ×|T| ≥ minsup
3.      sup({i}→{j}) := |tids(i) ∩ tids(j)| / |T |.
4.      sup({j}→{i}) := |tids(i) ∩ tids(j)| / |T|.
5.      conf({i}→{j}) := |tids(i) ∩ tids(j)| / |tids(i)|.
6.      conf({j}→{i}) := |tids(i) ∩ tids(j)| / |tids(j)|.
7.      IF sup({i}→{j}) ≥ minsup THEN
8.          IF conf({i}→{j}) ≥ minconf THEN SAVE({i}→{j}, L, k, minsup).
9.          IF conf({j}→{i}) ≥ minconf THEN SAVE({j}→{i}, L, k, minsup).
10.         Set flag expandLR of {i}→{j} to true.
11.         Set flag expandLR of {j}→{i} to true.
12.         R := R∪{{i}→{j}, {j}→{i}}.
13.     END IF
14. END FOR
15. WHILE ∃r ∈ R AND sup(r) ≥ minsup DO
16.     Select the rule rule having the highest support in R
17.     IF rule.expandLR = true THEN
18.         EXPAND-L(rule, L, R, k, minsup, minconf).
19.         EXPAND-R(rule, L, R, k, minsup, minconf).
20.     ELSE EXPAND-R(rule, L, R, k, minsup, minconf).
21.     REMOVE rule from R.
22.     REMOVE from R all rules r ∈ R | sup(r) <minsup.
23. END WHILE
```

Figure: Top K rules algorithm

Mining in this algorithm is a tedious job as the algorithm cannot rely on both threshold that is minsup and minconf here minsup is more efficient and reliable. If the worst case scenario is present a naïve top k algorithm would generate all the rules for the basic algorithm. The figure shows the main algorithm. The algorithm runs as follows it first scans the database once to calculate the tids for each database item termed as c. It then generates all valid rules of size 1x1 with each having at least minsupx|T| tids the procedure save is called next to store the rules generated. The frequent rules are added to R set. The idea is to always expand the rule having the highest support because it is more likely to generate rules having a high support and thus to allow to raise minsup more quickly for pruning the search space

7. Conclusion

By reviewing all the above algorithms we found that if we have to increase the mining feature and create a hybrid approach we have to use Top K rules with the MOPNAR. As the MOPNAR has a high efficient rule mining results whereas with the help of Top K we can increase the speed and can save the memory of the system hence we propose improving the rule accuracy using positive and negative subgraph mining with top k-rules

8. References

1. "CHARM: An Efficient Algorithm for Closed Association Rule Mining" by Mohammed J. Zaki and Ching-Jui Hsiao
2. "Extraction and Classification of Best M Positive Negative Quantitative Association Rules" by Ms. SheetalNaredi, Mrs. Rushali A. Deshmukh
3. "FastAlgorithms for Mining Association Rules" by RakeshAgrawal Ramakrishnan Srikant
4. "Mining Top-K Association Rules" by Philippe Fournier-Viger , Cheng-Wei Wu and Vincent S. Tseng
5. "An Efficient Mining of Sequential Rules Using Vertical Data Format" by SurbhiJigneshkumarSheth, Shailendra K Mishra
6. "Implementation of Different Data mining Algorithms with Neural Network" by Ms. Aruna J. Chamatkar
7. "Performance Analysis of Data Mining Algorithms with Neural Network" By Dr. P K Butey
8. "A New Multiobjective Evolutionary Algorithm for Mining a Reduced Set of Interesting Positive and Negative Quantitative Association Rules" by Diana Mart′ın, Alejandro Rosete, Jesus Alcal ′ a-Fdez

Cite this article as: Kavita Yadav, Pravin, G Kulurkar. "Study on Positive and Negative Rule Based Mining Techniques for E-Commerce Applications". *International Conference on Computer Applications 2016*: 06-09. Print.