IJCTA, 9(12), 2016, pp. 5655-5663

Mining Multilevel Association Rules with Hidden Granules for Recommendations

***V. Ajitha **P. Rajkumar**

Abstract : Granular association rule mining uses granules to represent the knowledge implicitly contained in the databases. It is a two stage process where in the first stage all the frequent granules are discovered and in the second stage association rules are generated from the frequent granules. To generate all kinds of rules we developed fast forward, backward, sandwich methods. This research provides efficient methods to interpret meaningful discovered knowledge such as methods to represent the discovered granules and associations. By this way, meaningless association rules can be adjusted. Experiments are done on a publicly available dataset. The results indicates the viability of the proposed research by setting the appropriate threshold which helps in obtaining high accuracy. Recommender system is used to suggest services or items to the service consumer based on their interest. The most successful recommendation system is collaborative filtering system that predicts a user interest in new items based on the recommendation of the other people with the similar interests. Agglomerative hierarchical clustering is used as a pre-processing step to improve the prediction quality in which similar items are grouped into same clusters hierarchically .Most collaborative filtering algorithm requires an explicit user feedback such as rating inorder to make prediction. Sometimes the users either write only a small portion of all available products (sparsity problem)or present incorrect ratings. This insufficient explicit feedback leads to unsatisfactory recommendations. In this approach, mining users implicit interests from usage records or reviews may be a complement to the explicit interests(ratings).By this means, sparsity problem has been solved to the extend and also recommendations can be generated even if there are only few ratings are available.

Keywords: Granules, Partial match, complete match, recommendations.

1. INTRODUCTION

Association rule mining is one form of data mining that finds association among attributes of transactions. In Boolean association rule, the information is stored in a Boolean database which reveals the connection between two disjoint subsets of the same universe. Quantitative association rules are Multidimensional association rules in which numeric attributes are dynamically discretized.The relational association rule mining looks for patterns that involves multiple tables. Efficient rule mining algorithms are developed to discover knowledge from the databases. But there are some difficulties when we apply these to solve real world problems. The major challenging problem in rule generation is the assessment of derived rules based on the quality. Predictive accuracy of a decision rule is measured by applying separate training dataset and separate test dataset, which contains data instances that were not seen during training. Although this is a widely used measure to assess the quality of a rule, it does not take into account the problem of uncertainty. The larger the results, the greater redundancy exists in the patterns and rules which are not interesting for users. The other problems that are encountered are rule generation takes too much of time. Interpretability will be the issue if there are huge number of patterns. So when the dimension of the input data increases, the accuracy and efficiency of the results decreases rapidly. Thus the worth and the knowledge discovery depreciates. Finally, as the approaches uses only two measures like support and confidence, knowledge coverage becomes incomplete. However if the support and confidence value are low, then it results in large volume of results.

* Asst Prof,CSE Department, Saveetha Engineering College, 602105, India

** Professor, MBA Department, Bannari Amman Institute of Technology,India ajithanice@gmail.com1, profprajkumar@gmail.com2,

In some cases, the entire knowledge is not always necessary to define various processes in the dataset. This motivates the need for efficient ways of representing and interpreting the discovered patterns and the rules. A recommender system is an intermediary program or an agent that intelligently complies a list of requisite information which suits users taste and needs. Many recommender systems have been designed and implemented for various types of items including newspapers, research papers, emails, books, movies, music, restaurants, web pages and other e-commerce products. It proposes a new approach to develop a framework for an efficient recommender system that assist users in decision making process where they want to choose some items among a potentially overwhelming set of alternative products or services. The collaborative filtering approach has been used to achieve the desired framework for our recommender system. Recommender system predicts new items of interest for a user on the basis of predictive relationships discovered between the user concerned and the other users sharing the same tastes and interests. The aim of CF is to recommend items to a target user based on the opinion of other users. Cluster is used as a pre-processing step that will reduce the overall size of the data inorder to increase the prediction accuracy.A fuzzy method proposed by Ma et al (2010) discovered some potentially more interesting association rules. Zailani et al (2010) proposed a trie-based algorithm that generates the significant patterns using support and correlations. Aouad et al (2010) compared the proposed approach with a classical Apriori-like distributed algorithm. Many applications directly or indirectly rely on finding the frequent items. Tremblay et al (2010) proposed a methodology to discover patterns in related attribute values. Yin Kuo-Cheng et al (2010) proposed temporal association rule mining algorithm which automatically generates all the intervals without using any domain specific information. WEI Yong-Qing, et al (2010) proposed an improved apriori algorithm is used minimum supporting degree and degree of confidence, for extracting association rules .But it has suffered from “frequent pattern sets explodes “and “rare item dilemma “.XING Xue et al (2010) reveal knowledge hidden in the massive database and proposed an approach for Evaluation of exam paper. This paper introduces a new direction, applies interesting rules mining to evolution of completive exam and finds out some useful knowledge. But this algorithm need repeated database scan and takes more time to perform I/O operation. Wang et al (2011) presented Apriori association rule algorithm for analysing the performance of college students. Rama subbareddy et al (2011) proposed an approach for mining the positive and the indirect negative associations between itemsets. Abdullah et al (2011) proposed a new measure to discover the significant association rules. Even though several different approaches to association rule mining are presented, starting from traditional approaches, followed by multilevel and cross-level approaches, all those focused on the proposal of different types of algorithms for Association Rule Mining with the measures support and confidence. However, the focus of recent research is on improving the efficiency of these algorithms using measures like source coverage, target coverage, source confidence and target confidence. Therefore, new algorithms have been proposed in this research work to enhance the capabilities of the existing Association Rule Mining algorithms in terms of the number of rules, checking time and basic operations. Mittal et al proposed to achieve the predictions for a user by first minimizing the size of item set the user needed to explore. K-means clustering algorithm was applied to partition movies based on the genre requested by the user. But it requires users to provide some extra information. We et al proposed an improved item based collaborative filtering algorithm based on clustering method. This method constructs the dynamic clustering based on similarity value then it performs collaborative filtering for prediction.it lowers the MAE and computation. But the efficiency of the recommendation will be low because of the items amount of the cluster. Mai et al proposed the neural network based clustering and collaborative filtering. In this approach, clustering is formed based on the back propogation method using users web visiting data. But it is not sure that users preference on web visiting is relevant to preference on purchasing. Lee et al proposed time base recommender system using implicit feedback. In this approach pseudo rating matrix was constructed based on the appearance and purchase time of an item from the implicit feedback of users. This approach does not consider various user activities and use a pseudo rating matrix by adding up the number of item consumption. This may have potential risk that it can misunderstand the users preference.

2. MATERIALS AND METHODS

2.1. Association rule mining algorithms

A rule consists of a pair of Boolean valued propositions, LHS the antecedent and RHS the consequent. The rule states that when the LHS is true then the RHS will be also true. The association rule is of the form A->B where A belongs to itemset, B belongs to itemset and A ? B = ø. The information that customers who purchase computer also tend to buy printer at the same time is represented in association rule mining as

Computer -> printer [Support = 10%, confidence = 80%]

The equation to calculate support and confidence are

$$\text{Support}(A \rightarrow B) = P(AUB)$$

$$\text{Confidence} = P(B/A)$$

$$= \frac{\text{Support (AUB)}}{\text{Support (A)}}$$

The support and confidence are the two measures of rule interestingness. A support of 10% for association rule means that 10% of all the transactions under analysis show that computer and printer are purchased together. A confidence of 80% means that 80% of customers who purchased a computer also bought the printer. Typically association rules are interesting if they satisfy both a minimum support threshold and a minimum confidence threshold. Such thresholds can be set by users or domains experts. Additional analysis can be performed to uncover interesting statistical correlation between associated items.

2.2. Generating Frequent granules using Apriori

The Apriori Algorithm is the best known algorithm for mining frequent itemsets for boolean association rules. It uses breadth first strategy to find the frequent itemsets which is the set of items that have minimum support *i.e.*, if {AB} is a frequent itemset, both {A} and {B} should be a frequent itemset and uses a candidate generation function *i.e.* frequent subsets are extended one at a time. The group of candidates are tested against the data. The algorithm terminates when no successful extensions exists.

The Apriori Algorithm : Pseudo code

Join Step : Candidate itemset C_k is generated by joining $L_k - 1$

Prune Step : Any $(k-1)$– itemset that is not frequent cannot be a subset of a frequent k–itemset

Pseudo-code :

C_k : Candidate itemset of size k

L_k : frequent itemset of size k

L_1 = {frequent items};

for ($k = 1$; L_k ! = ∅; k ++) do begin

C_{k+1} = candidates that are generated from L_k; for each of the transaction t in database do increment the count of all candidates in C_{k+1} that are contained in t

L_{k+1} = candidates in C_{k+1} with minimum support

end

return $\cup_k L_k$;

2.3. Generating Frequent granules using Frequent-Pattern Growth (FP-Growth) Method

It compress a large database into a compact tree structure. It is highly condensed, but complete for frequent pattern mining and avoid costly database scans. FP-tree-based frequent pattern mining method is an efficient divide-and-conquer methodology which decompose mining tasks into smaller ones and avoids the costly process of candidate generation. To construct FP tree, first, create the root of the tree, with "null". Then scan the database D a second time. The items in each transaction are processed in L order (*i.e.* sorted order).A branch gets created

for each transaction having their support count separated by colon. Whenever the same node is encountered in other transaction, we increment the support count of the common node or Prefix. For tree traversal, an item header table is built so that each item points to its occurrences in the tree via a chain of node-links. Thus mining of frequent patterns is transformed to that of mining the FP-Tree. Thus the performance study shows that FP-growth is an order of magnitude faster than Apriori, and is also faster than tree-projection due to no candidate generation, no candidate test, uses compact data structure, eliminates repeated database scans and the basic operation is counting and FP-tree building.

The FP tree construction algorithm :

Input : DB, minsup ms
Output : FP–Tree
Method :
Scan DB once
Collect frequent items F
Sort F in support descending order
Create root of Tree and, for each transaction T of DB
insert frequent items of T in Tree sorted by descending freq
Inserting items <*p*1, *p*2,..>in Tree
if root of Tree has a child N with *p*1, increase counter of node
else create a new child N with *p*1 and count 1
insert <*p*2,…> in subtree with root N

2.4. Proposed System

Transactions possess some common similarities and dissimilarities irrespective of the domain. Concept hierarchies are used in mining Multilevel association rules more effectively which defines mapping of higher level concepts to low level concepts. Based on the ROOT node, the transactions can be categorized. Each item in a transaction will have a parent to some specific transactions and others may be a leaf node or it can be node with no childrens at all. In some cases, the products will be generalized at the top level and aligned to the lower level details. Thus building a taxonomy leads to better mining of association rules by providing proper ordering among the transactions. Ancestors are the headers of the specified group of transactions and descendants are the childrens of those actually occurred before. Ancestors are denoted by a familiar name which represents the combination of several transactions. To differentiate among each other, the attributes are directly mentioned in the descendants.

Let the list of customer be C = {*c*1,*c*2, *c*3...*cn*} and let the list of Products be P = {*p*1, *p*2, *p*3...*pn*}. Thus CP = {*c*1*p*1,*c*1*p*2,…. *cnpn*} is a set of customers who buys products. Each and every transaction is comprised of customer *id* and product *id* which is denoted by (*pid*,*cid*}. All products possess certain characteristics and they belong to a certain ancestors. Separation of all these products with respect to different criteria may or may not be suitable in all situations.

Table 1. Customer *id* with Transactions

Cid	*Name*
*c*1	T-shirt
*c*2	Jeans
*c*3	Bread
*c*4	Butter
*c*5	Jam

Table 2. Product *id* with product name

Pid	*Name*
*p*1	T-shirt, Jeans
*p*2	Jeans, Butter
*p*3	Bread, Butter
*p*4	Butter, Bread
*p*5	Jam, Butter

Table 3. buys

PID/CID	*C1*	*C2*	*C3*
P1	1	1	0
P2	0	0	0
P3	0	0	1
P4	0	1	0
P5	0	0	0

To construct the candidate sets, metrics like support and confidence are needed. When the transactions are analyzed at a very fine level, it requires a very large amount of calculations to find out the frequency of transactions. When mining multidimensional rules, for a single transaction itself there exists a large number of multiple support. As a result the different levels should be analyzed and carefully association rules are framed because an item which may occur less frequently may have a serious impact on being avoided. Thus with Apriori algorithm, by setting appropriately the support value, all the necessary and significant transactions can be mined. Those items which does not come under the candidate set generation are pruned from further estimation.

To evaluate the quality of rules, six measures are used.

Support which reflects the usefulness of the rule

Confidence which reflects the certainty of the rule

Source coverage : It is denoted by the left-hand side of the association rule given the number of objects in the first set.

Target coverage : It is denoted by the right-hand side of the association rule given the number of objects in the second set.

Source confidence : It indicates the strength of a rule in the first set.

Target confidence : It indicates the strength of a rule in the second set.

These types of rules are semantically richer than the existing one because it is more specific than spanning across the database.

Consider the association rule Customer buys Products

This definition is ambiguous and to address these issues, we present four subtypes of rules

Complete match rules : (*eg*) All customer buys all kind of Products.

Left Hand side partial Match rules: (*eg*) 40% of the customer buys all kind of products.

Right Hand side partial Match rules : (*eg*) All customer buys atleast 60% of products.

Partial match rules : (*eg*) 40% of customer buys atleast 30% of products.

2.5. Algorithm

Read the two information system directly

Construct the Boolean information system and read the support compressed Boolean information system.

Read the third information system

Convert it into Boolean one.

if they are not needed for internal representation ,delete the ID of first two sets.

Overall Algorithm : Sandwich

Input : Source coverage threshold, target coverage threshold and the base algorithm .

Output : returns all rules in a string.

Steps involved :

1. Compute source coverage
   Compute the first set frequent granules
   Compute the first set frequent granule extension.
2. Compute target coverage (second set granules)
   Compute the frequent granules for the second set
   Compute the frequent granules for the second set extension.
3. Check all possible rules in the first and second set granules extension length and output the valid rules in terms of checking time used, the number of basic operations performed and the time used.

The time complexity of this method is $O(IRU || RV || U || VI) $ where $(IRUI)$ and $(IRVI)$ are the sizes of the frequent items in the first and second set respectively.

Overall Algorithm : Fast Forward

Input : Source coverage threshold, Target coverage threshold and the base algorithm.

Output : returns all rules in a string.

Steps involved :

1. Compute source coverage
   Compute the first set frequent granules
   Compute the first set frequent granule extension.
2. Compute target coverage
   Compute the frequent granules for the second set
   Compute the frequent granules for the second set extension.
3. For each of the source granules obtained, construct a set of objects that are instances of the granule and store it in one dimensional positive arrays rather than storing the relation in boolean array.

The time complexity of this method is $O(IRU || RV || U |+| VI) $ where $(IRUI)$ and $(IRVI)$ are the frequent items in the first and second set respectively.

Overall Algorithm : Fast backward

Input : Source coverage threshold, Target coverage threshold and the base algorithm.

Output : returns all rules in a string.

Steps involved :

1. Compute source coverage
   Compute the first set frequent granules
   Compute the first set frequent granule extension.

2. Compute target coverage (second set granules)
   Compute the frequent granules for the second set
   Compute the frequent granules for the second set extension.
3. Avoid doing computation of different rules which is having the same Right hand side. For each of the granules obtained, construct a set of objects that are instances of the granule.
4. Check all possible rules in the first and second set granules extension length and output the valid rules in terms of checking time used, the number of basic operations performed and the time used.

The time complexity of this method is $O(IRU || RV || UI)$ where $(IRUI)$ and $(IRVI)$ are the sizes of the frequent items in the first and second set respectively. The space complexity of the above methods are |U|*|V| as it needs to store all the Boolean matrix in the relation.

2.6. Issues to be addressed for Recommendations

Decision is not made within the acceptable time.

Finest recommendation is not generated from so many sevices.

To compute similarity between every pair of users or services may take too much time, even exceed the processing capability of current recommendation system.

Many users assign arbitrary ratings that do not reflect their true opinions.

It is not practical to expect users active participation in ratings.

User rating only fey of all available products(problem)

Collaborative filtering algorithm will suffer from serious scalability problems with an increasing number of users and items.

2.7. The methodologies used for recommendations are

(*a*) System design : Cluster process constructs cluster of similar items based on the similarity value of the item hierarchically. Recommendation engine starts its prediction for a new item or unused items using collaborative filtering when service consumers request an item.The prediction has to be done within a cluster belongs to a target item.

(*b*) Cluster analysis : Clustering is the process of grouping a set of data objects into multiple groups or clusters so that objects within a cluster have high similarity compare to objects in other clusters. Dissimilarities and similarities are evaluated based on the attribute values describing the objects and often involve distance measures. Clustering is the process of partitioning a set of data objects into subsets. Each subset is considered as a cluster, such that objects in a cluster or similar to one another and dissimilar to objects in other clusters. The set of clusters obtained from a cluster analysis can be referred to as clustering. Hence clustering is useful method in that it can lead to the discovery of previously unknown groups within the data.

(*c*) Similarity measurement : Similarity is a measure that reflects the strong point of relationship between the two objects or two features. The similarity values between services are measured in which high similarity clusters are made hierarchically. Jaccard similarity coefficient measures similarity as the intersection divided by the union of the objects. For text data, the jaccard coefficient compares the weight of shared terms to the sum weights of terms that are present in either of the two documents but are not the shared terms.

(*d*) Agglomerative hierarchical clustering : It is used as a preprocessing step that separate big data into manageable parts.It uses a bottom up strategy. It typically starts by letting each service from its own cluster and iteratively merge two most similar clusters until all the services are in a single cluster or certain termination conditions are satisfied.

The Steps in this approach are :

1. Starts with N clusters
2. Search for the pair in the similarity matrix with the maximum similarity and merge them
3. Create a new similarity matrix where similarities between clusters are calculated by their average value.
4. Save the similarities and cluster partitions for later visualization

   Proceed with 1 until the matrix is of size k which means that only k clusters remains.

(*e*) Filtering approach : The basic idea of collaborative filtering based algorithm is to provide item recommendations or predictions based on the attitude of other like minded users. The opinions or rating of users can be obtained explicitly from the users or by using some implicit measures. The goal of this algorithm is to suggest new items or to predict the utility of a certain item for a particular user based on the users previous linkings and the opinions of other like minded users. In a collaborative filtering scenario, there is a list of m users and list of n items each user has a list of items which the user has expressed their opinion about the item. Opinions can be explicitly given by the user as a rating score, generally within a certain numerical level.

Recommendation is a list of N items that the active user will like the most. The recommended list must be on items not already purchased by the active user. This interface of algorithms is also known as Top-N recommendations.

Implicit feedback

Collaborative filtering based recommender systems require a customer profile to identify preferences and make recommendations. To create a customer profile, two profiling techniques are used implicit and explicit ratings. Implicit ratings are techniques that gather information about a customers shopping behaviors and represent preferences as ratings without the customer intervention. Because there is no explicit user feedback, we construct a simulated rating matrix within a cluster from implicit feedback such as purchase information or user behaviour patterns. This rating matrix will be used for collaborative filtering. Although the simulated rating matrix is not constructed directly by users, to some extent it reflects their preferences. Collect the user purchase data or user behaviour pattern. This data is usually available in a typical ecommerce environment and will be used to construct a rating matrix. Describing a rating based on user behaviour patterns and construct a simulated rating matrix using the computed rating values. The rating function will depend on the type of product or service to be recommended.

3. RESULTS AND DISCUSSION

The algorithms were tested in Weka which contains many data mining and machine learning algorithms like preprocessing, classification association rule mining, clustering, prediction etc. The performance of the different algorithms are evaluated based on execution time. The execution time is measured for different datasets with different number of instances. To the Weka, data set is imported in ARFF format. For evaluating the efficiency of the algorithm, graphical user interface was built with java.

1. The source coverage and the target coverage of the rules are already specified. We set the target confidence threshold to obtain different source confidences. The target confidence threshold are set to 0.1,0.2 to 0.9.A very small non zero number guarantees that atleast one object is covered by the rule. Threshold value cannot be specified as zero. If threshold value is mentioned as zero, then it means that it does not provide any meaningful rules. If we increase the target confidence threshold, the source confidence gets decreased. So there exists tradeoff between the two thresholds.
2. The basic operations refers to the comparison, addition etc. But we will focus on runtime instead of number of basic operations since different operations take different time.
3. The number of basic operations are compared with all the methods. It can be naturally observed that the forward and backward methods are more efficient than sandwich method. However, with the decrease

of thresholds, the number of operations increases and the backward algorithm makes the best choice. The rule checking terminates only when certain conditions are met. However the time complexities are for reference only and the runtime depends on the characteristics of data. Inshort the backward algorithm can generate many rules and is scalable whereas for small datasets and large thresholds the sandwich method is more efficient.

4. CONCLUSION

Most of the association rule mining algorithms suffer from the problems of too much execution time and generating too many association rules. Although conventional algorithm can identify meaningful itemsets and construct association rules, it suffers the disadvantage of generating numerous candidate itemsets that must be repeatedly contrasted with the entire database. The processing of the conventional algorithm also utilizes a large amount of memory. Thus, this approach is very significant for effective analysis and it helps the customers in purchasing their items with more comfort, which in turn increases the sales rate of the markets and helps in recommendation of various products.

5. ACKNOWLEDGEMENT

The author likes to thank the management of Saveetha University for their kind support during the preparation of this paper.

6. REFERENCES

1. WEI Yong-Qing, YANG Ren-hua, LIU Pei-yu: "An Improved Apriori Algorithm for Association Rules of Mining" 978-1-4244-3930-0/09/$25.00 © IEEE (2010).
2. XING Xue CHEN Yao WANG Yan-en:"Study on Mining Theories of Association Rules and Its Application" .International Conference on Innovative computing and communication Asia –Pacific Conference on Information Technology and Ocean Engineering 978-0-7695-3942-3/10 $26.00 IEEE (2010).
3. Xiufend Piao, Zhan long Wang, Gang Liu: "Research on mining positive and negative association rules based on dual confiden ce" Fifth International Conference on Internet Computing for Science and Engineering. 978-1-4244-9954-0/11 $31 © IEEE (2011).
4. Pengfei Guo Xuezhi Wang Yingshi Han: "The Enhanced Genetic Algorithms for the Optimization Design" 978-1-4244-6498-2/10 © IEEE (2010).
5. WEI Yong-Qing, YANG Ren-hua, LIU Pei-yu: "An Improved Apriori Algorithm for Association Rules of Mining" 978-1-4244-3930-0/09/$25.00 © IEEE (2010).
6. Sandeep Singh Rawat and Lakshmi Rajamani: "Probability Apriori based Approach to Mine Rare Association Rules".In 3rd Conference on Data Mining and Optimization (DMO), © IEEE (2011).
7. XING Xue CHEN Yao WANG Yan-en:"Study on Mining Theories of Association Rules and Its Application" .International Conference on Innovative computing and communication Asia –Pacific Conference on Information Technology and Ocean Engineering 978-0-7695-3942-3/10 $26.00 IEEE (2010).
8. CH.Sandeep Kumar, K.Shrinivas, Peddi Kishor T.Bhaskar: "An Alternative Approach to Mine Association Rules" 978-1-4244-8679-3/11 $26.00 © IEEE (2011).
9. B. Goethals, W. L. Page, and M. Mampaey, "Mining interesting sets and rules in relational databases," in Proceedings of the 2010 ACM Symposium on AppliedComputing, 2010,pp. 997-1001.
10. F. Min, H. He, Y. Qian, and W. Zhu, "Test-cost-sensitive attribute reduction," Information Sciences, vol. 181, pp.4928-4942, 2011.
11. F. Min, H. He, Y. Qian, and W. Zhu, "Test-cost-sensitive attribute reduction," Information Sciences, vol. 181, pp. 4928-4942, 2011.
12. F. Min, Q. Hu, and W. Zhu, "Granular association rules on two universes with four measures," submitted to Information Sciences, 2012. [Online]. Available: http://arxiv.org/abs/1209.5598
13. F. Min, Q. H. Hu, and W. Zhu, "Granular association rules with four subtypes," in Proceedings of the 2011 IEEE International Conference on Granular Computing, 2012, pp. 432–437.
14. F. Min and W. Zhu, "Granular association rule mining through parametric rough sets," in Proceedings of the 2012 International Conference on Brain Informatics, ser. LNCS, vol. 7670, 2012, pp. 320–331.