ISSN: 2321-7782 (Online)
Special Issue, December 2013
International Journal of Advance Research in Computer Science and Management Studies
Research Paper
Available online at: www.ijarcsms.com

A Survey on Algorithms for Market Basket Analysis

Gajalakshmi.V[1]
PG Student
Department of Computer Science and Engg
Velammal Engineering College
Chennai - India

M. S. Murali Dhar[2]
Assistant Professor
Department of Computer Science and Engg
Velammal Engineering College
Chennai - India

Abstract: ***Market Basket Analysis (MBA) is well known activity of Association Rule Mining (ARM) ultimately used for business intelligent decisions. MBA can be used in decision support systems, credit card fraud detection, telephone calling pattern analysis, fraudulent insurance claim detection, production, market strategy and financial forecast.***

Mining frequent item sets and hence deduce rules to build classifiers with good accuracy is essential for efficient algorithm. Modifications have been done already on existing traditional market basket analysis algorithms Apriori, CBA, CPAR etc., to improve the efficiency. Analysing the data mining algorithms to obtain optimum rules and fine tune the parameters is specific to the problem. This paper analyses various algorithms for market basket analysis.

Keywords: Associative classification, market basket analysis, Association Rule Mining (ARM).

I. INTRODUCTION

Market basket analysis is a data mining method focusing on discovering purchasing patterns of customers by extracting associations or co-occurrences from a store's transactional data. For example, the moment shopper's checkout items in a supermarket, swipe credit card and also offers the loyalty card, a lot of data about the purchase - demographic details, address of the person goes in to the transaction database. Later, this huge data of many customers are analysed, lot of experiments done to arrive at purchasing pattern of customers. Also decisions like which item to stock more ,cross selling, up selling ,store shelf arrangement, catalogue design are determined.

Association rule mining (ARM) identifies the association or relationship between a large set of data items and forms the base for market basket analysis. Association rule mining has been widely used in various industries besides supermarkets, such as mail order, telemarketing, production, fraud detection of credit card and e-commerce. Industries and business organisations are concerned in mining their database to gain useful information which can help them to make better decisions, improve the quality of interaction between the organisation and their customers [1].

Data mining tasks are different and distinct because of the different patterns in a large database [12]. Based on types of patterns, data mining tasks are classified as

- Summarization
- Classification
- Association
- Clustering
- Trend Analysis

Summarization is the generalization or abstraction of data. A set of task relevant data is generalised and presented as aggregate function. Classification is the derivation of a function or model which determines the class of an object based on its attributes.

Association is the togetherness or connection of objects, such connection or togetherness is termed as associative rule. For example retail store may find people tend to buy tomato sauce when buying instant noodles and puts the noodles on sale to promote sale of tomato sauce. Association rule is simply a *If X then Y rule.*

If noodles ***then*** tomato sauce , i.e., noodles→ tomato sauce

Clustering is identification of classes also called clusters or groups, for a set of objects whose classes are unknown.

Associative classification (AC) is the integration of association and classification where associative rules are used in training step of classification. Associative classification (AC) is an emerging trend in mining which has been proved as more efficient than traditional classification and association algorithms in terms of accuracy and efficiency.

Two standard measures are used in associative classification known as

1. Support
2. Confidence

For example in a transaction database of a supermarket, consider only a 100 transactions.

Support: The frequency of occurrence of an item in total number of transactions is called the support of an item.

Suppose in 100 transactions, the item Bread has occurred 80 times,

Support (Bread) = 80/ 100 x 100 = 80%

Confidence: the frequency of co-occurrence of associated item.

If butter is bought along with bread in 60 transactions of 100,

Confidence (Bread, Butter) = Support (Bread U Butter) /Support (Bread)

= 60 / 80 x 100 = 75%

Where Minsupp = 50% and Minconf = 70% , are two thresholds.

The above support and confidence of items leads to a decision in the form of rule as Bread Butter. i.e, if bread is bought then butter is also bought along with it , the store decides to place butter near the bread ,thereby increasing sale of butter . Imagine mining data from large transaction database such as Walmart or Tesco or Reliance retail and associating a vast range of items and making business intelligent decisions. The entire process has to use data mining technique with an accurate and efficient algorithm.

The following section describes the discovery of item sets, rule discovery, rule pruning, rule prediction and rule evaluation which forms the base for any associative classification algorithm. Section III describes the survey of associative classification algorithms.

II. ARCHITECTURE OF ASSOCIATIVE CLASSIFICATION MINING

The associative classification algorithm consists of 5 steps (Fig.1):

STEP 1: The input data from database is preprocessed, converted to vertical format representation. Thresholds Minsupp and Minconf are accepted as input.

STEP 2: Frequent item sets are discovered by calculating the support of each item. Any item that exceeds the Minsupp qualifies as frequent item set. Frequent-1 item sets are found. The process is repeated for frequent-2 item sets Till all frequent items set combinations are found.

STEP3: Confidence of all frequent item sets is calculated, if it exceeds Minconf, the item set qualifies as a candidate rule item. Thus class of associative rules (CAR) are generated.

STEP4: Rules are pruned, redundant rules removed and ranked based on their confidence level. Rule with highest confidence is ranked first. Based on this potential rule set the classifier model is built.

STEP5: The classifier is evaluated using test data to prevent over fitting and biased classification.

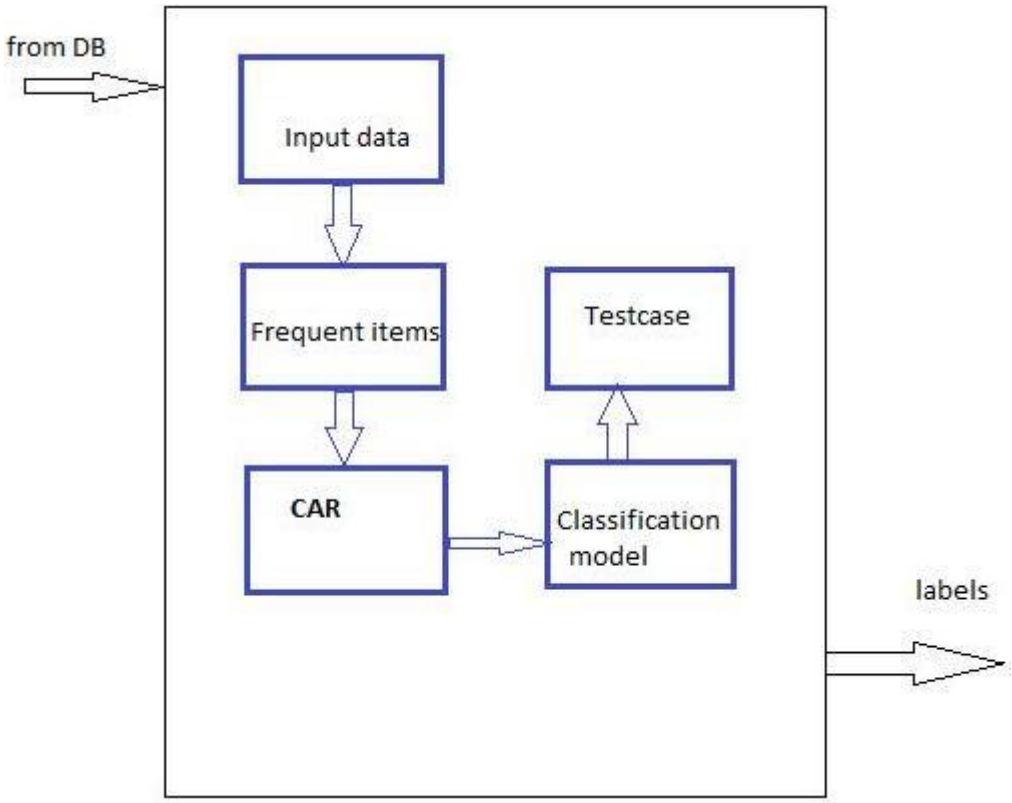

Fig .1 Architecture of associative classification [10] (adopted from Thabtah, 2005)

III. LITERATURE REVIEW

Rakesh Agarwal et al., [3] proposed the Apriori algorithm. Apriori was the first associative algorithm proposed and future developments in association, classification, associative classification algorithms have used apriori as part of the technique.

The discovery of frequent item sets is accomplished in several iterations. Counting new candidate item sets from existing item sets requires scanning the entire training data. In short, the algorithm involves only 2 steps:

- Pruning
- Joining

Apriori uses the ―apriori property to improve the efficiency of the search process by reducing the size of the candidate item sets list for each iteration. Still apriori algorithm has some demerits such as

1. Scanning the database many times results in shortage of memory.
2. Processing time is more and exhibits low efficiency.
3. Huge Time Complexity.

Han et al., [4][5] presented a new association rule mining approach that does not use candidate rule generation called FP-growth that generates a highly condensed frequent pattern tree (FP-tree) representation of the transactional database. Each database transaction is represented in the tree by at most one path. FP-tree is smaller in size than the original database the construction of it requires two database scans, where in the first scan, frequent item sets along with their support in each transaction are produced and in the second scan, FP-tree is constructed.

The mining process is performed by concatenating the pattern with the ones produced from the conditional FP-tree. One constraint of FP-growth method is that memory may not fit FP-tree especially in dimensionally large database.

Liu et al., proposed CBA [6] the first Associative Classification (AC) algorithm. CBA implements the famous Apriori algorithm [3] in order to discover frequent rule items.

The Apriori algorithm consists of three main steps.

Step 1: Continuous attribute in the training data set gets discretised.

Step 2: Frequent rule items discovery

Step 3: Rule generation.

CBA selects high confidence rules to represent the classifier. Finally, to predict a test case, CBA applies the highest confidence rule whose body matches the test case. Experimental results designated that CBA derives higher quality classifiers with regards to accuracy than rule induction and decision tree classification approaches.

Li et al., recommended Classification based on Multiple Association Rules (CMAR). The method is an extension of FP-growth, constructs a class distribution-associated FP-tree, and mines large database efficiently [7].

Moreover, it applies a CR-tree structure to store and retrieve mined association rules efficiently, and prunes rules effectively based on confidence, correlation and database coverage. The classification is performed based on a weighted x2 analysis using multiple strong association rules. Extensive experiments on 26 datasets from UCI machine learning database repository show that CMAR is consistent, highly effective at classification of various kinds of databases and has better average classification accuracy in comparison with CBA and C4.5 methods.

Yin et al., [8] suggested Classification based on Predictive Association Rules(CPAR). CPAR integrates the features of associative classification in predictive rule analysis. It has the following advantages:

1. CPAR generates a much smaller set of high-quality predictive rules directly from the dataset.
2. Avoids redundant rules, CPAR generates each rule by considering the set of " already generated" rules.
3. Predicting the class label of an example, CPAR uses the best k rules the example satisfies.

Moreover, CPAR employs the following features to further improve its accuracy and efficiency:

1. CPAR uses dynamic programming to avoid repeated calculation in rule generation.
2. Rule generation considers all the close-to-the-best literals are selected so that important rules will not be missed.

CPAR generates a smaller set of rules, with higher quality and lower redundancy in comparison with associative classification.

As a result, CPAR is much more time efficient in both rule generation and prediction but achieves as high accuracy as associative classification.

CBA, CMAR approaches have higher accuracy than decision tree classifier due to the fact that decision-tree classifier examines one variable at a time whereas association rules explore highly confident associations among multiple variables at a time [9]. However, these approaches have a severe limitation. All associative classification algorithms use a support threshold to generate association rules. In that way some high quality rules that have higher confidence, but lower support will be missed. Actually the long and specific rules have low support and so they are mostly penalized. But a good classification rule set should contain general as well as specific rules. It should also contain exceptional rules to account for the exceptional instances.

Classification based on Association Rules Generated in a Bidirectional Approach-CARGBA by Gourab et al., [9] is essentially a bidirectional rule generation approach that generates crisp association rules. It not only tries to generalize the dataset but also tries to provide specific and exceptional rules to account for the specific characteristics and anomalies in the dataset. This approach has added a new dimension by considering the exceptional cases too.

The MCAR algorithm introduced by Fadi et al., [10][11] uses an intersection technique for discovering frequent rule items. MCAR takes advantage of vertical format representation and uses an efficient technique for discovering frequent items based on recursively intersecting the frequent items of size n to find potential frequent items of size n+1. MCAR consists of two main phases: rule generation and a classifier builder. In the first phase, the training data set is scanned once to discover frequent one rule items, and then MCAR recursively combines rule items generated to produce potential frequent rule items (candidate rule items) involving more attributes.

The supports and confidences for candidate rule items are calculated simultaneously, where any rule item with support and confidence larger than minsupp and minconf, respectively, is created as a potential rule. In the second phase,the classifier builder ensures that each training instance is covered by at most one rule, which has the highest precedence among all rules applicable to it. Furthermore, there are no useless rules in the MCAR classifier since every rule correctly covers at least one training instance. This approach is similar to the CBA classifier builder as each rule in CBA also covers at least one training instance. However, the way MCAR builds the classifier by locating training instances is more efficient than that of CBA due to abounding going through the training data set multiple times.

Performance studies on 20 data sets from UCI data collection (Table 1) from [11] using WEKA tool indicates that CBA outperforms C4.5 and RIPPER. MCAR outperforms CBA in certain cases. Reduced error rate means the accuracy is better.

Data	Size	Classes	C4.5	RIPPER	CBA	MCAR
Cleve	303	2	23.77	22.45	16.87	17.13
Breast-w	699	2	5.44	4.58	4.16	3.52
Diabetes	768	2	26.18	23.96	24.66	22.18
Glass	214	7	33.18	31.31	30.11	30.33
Iris	150	3	4.00	5.34	6.75	4.68
Pima	768	2	27.22	26.70	24.51	21.46
Wine	178	3	5.62	7.31	1.67	2.89
Austral	690	2	14.79	14.79	14.64	12.57
German	1000	2	29.10	27.80	27.43	27.90
Labor	57	2	26.32	22.81	5.01	12.81
Tic-tac	958	2	16.29	3.03	0.00	0.00
Led7	3200	10	26.44	30.47	28.26	28.76
Heart-s	294	2	18.71	21.77	20.80	18.86
Lymph	148	4	18.92	22.98	23.62	23.98
Vote	435	2	11.73	12.65	13.09	11.30
Zoo	101	7	6.94	14.86	4.04	2.22

Balance-scale	625	3	35.68	25.44	34.34	22.46
Primary-tumor	339	23	58.41	65.20	74.89	58.90
Mushroom	8124	2	0.23	0.10	8.71	2.44
Contact-lenses	24	3	16.67	25.00	20.00	25.00

Table.1 Error rate of MCAR, CBA, C4.5 and RIPPER algorithms using 10-fold cross validation [11].

IV. CONCLUSION

Accuracy and efficiency are crucial factors in classification task in data mining. This paper presents a survey of associative, associative classification algorithms, highlighting the advantages and limitations. Associative classification achieves better accuracy than some traditional rule-based classification in most cases. Every algorithm has different approaches to rule discovery, pruning, ranking, prediction and evaluation. Also memory usage, processing time, database scan time vary depending on the dataset used and the algorithm. MCAR algorithm is highly competitive when compared with traditional classification algorithms such as RIPPER, C4.5 and scales well compared with popular AC like CBA with regards to prediction power, rules features and efficiency. MCAR produces classifiers with slightly more rules than current AC techniques, resulting in reduced error rate.

References

1. Jiawei han , Micheline Kamber, and Morgan Kaufmann," Data Mining: Concepts and Techniques", Publishers, San Francisco, CA, 2000.
2. K.P.Soman et al., "Data mining Theory and Practice",PHI Publishers,NewDelhi,India,2010.
3. R. Agrawal and R. Srikant,"Fast Algorithms for Mining Association Rules in Large Databases", Journal of Computer Science and Technology, vol. 15, pp. 487-499,1994.
4. J. Han, J. Pei, and Y. Yin, "Mining frequent patterns without candidate generation", Proc of the ACM SIGMOD International Conference on, vol. 1, , pp. 1-12, 2000.
5. J. Han, J. Pei, Y. Yin, and R. Mao, "Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach", Data Mining and Knowledge Discovery, vol. 8, pp. 53-87 , 2004.
6. B. Liu, W. Hsu, and Y. Ma, "Integrating classification and association rule mining", Knowledge discovery and data mining, pp. 80–86 ,1998.
7. Li, W., Han, J., & Pei, J. (2001),"CMAR: Accurate and efficient classification based on multiple-class association rule", Proceedings of the ICDM'01, pp.369–376, 2001.
8. X.Yin, J.Han, "CPAR:Classication based on Predictive Association Rules",Proceedings of the Third SIAM International Conference on Data Mining, pp 331-335,2003.
9. Gourab Kundu, Sirajum Munir, Md. Faizul Bari, Md. Monirul Islam, and K. Murase ,"A Novel Algorithm for Associative Classification",14th International Conference, ICONIP 2007, Kitakyushu, Japan, pp 453-459 , November 13-16, 2007.
10. F. Thabtah, P. Cowling, and Y. Peng, "MCAR: multi-class classification based on association rule", Computer Systems and Applications, The 3rd ACS/IEEE International Conference on, pp. 33. , 2005.
11. Fadi Thabtah ,Peter Cowling and Suhel Hammoud,"Improving rule sorting, predictive accuracy and training time in associative classification", ELSEVIER, Expert Systems with Applications xx , pp 1–13, 2005.
12. Fu, Y., "Data Mining: Tasks, Techniques, and Applications", IEEE Potentials, Vol. 16, No. 4, pp 18-20, 1997.
13. Market basket analysis, Tesco, [online] Available: http://loyaltysquare.com/tesco.php accessed on 18/11/2013.

AUTHOR(S) PROFILE

Gajalakshmi.V, Pursuing Master of Engineering in Velammal Engineering College, Affiliated to Anna University, Chennai. Received Bachelor of Engineering Degree in Computer Science and Engineering in 1999. Currently working on a project in Market Basket Analysis.

Murali Dhar M S, received his Master of Engineering degree in Computer Science and Engineering from Velammal Engineering College, Chennai in 2009. He is pursuing his Ph.D. degree in the research area of Cloud Computing and infrastructure from Anna University, Chennai. Presently he is working as Assistant Professor in the Department of Computer Science and Engineering at Velammal Engineering College, Chennai. His research interests are in various applied/systems topics including Cloud Computing, Distributed Systems, Operating Systems, and Network Security/Resilience.

