VOL 3 (2019) NO 3

e-ISSN : 2549-9904
ISSN : 2549-9610

INTERNATIONAL JOURNAL ON INFORMATICS VISUALIZATION

A Data Mining Approach for Prediction Modelling Using Association Rule

Sarkhel H.Taher Karim[#], Rzgar Sirwan Raza[*]

[#] *Department of Computer Science, College of Science, University of Halabja Halabja, Iraq*

[*] *Department of Computer Science, College of Science and Technology, University of Human Development, Iraq*
E-mail sarkhel.kareem@uoh.edu.iq, rzgar.sirwan@uhd.edu.iq

***Abstract*— In the present work, a data mining approach is highlighted, a prediction optimization data mining approach association rule is chosen for performing prediction modeling in a supermarket application, a data mining prediction analysis model is formulated based on association rule is presented in this work. The result of the model formulated is then compared with the result produced on the similar set of input on the traditional optimization problems. While comparing the results it was observed that the result produced by the presented model is much closer to the reality.**

***Keywords*— Data Mining, Prediction.**

I. INTRODUCTION

Data mining (DM) is commonly defined as the extraction of implicit, yet unknown and potentially useful information from data by using methods from statistics, artificial intelligence, machine learning and pattern recognition. Today, DM is seen more and more as one step of a systematic and iterative knowledge discovery process, in which automated pattern recognition methods are combined with the analyst's expert knowledge. This process is known as Knowledge Discovery in Databases (KDD) [1]. And use the result to investigate the reliable patterns and/or efficient associations between dependent and independent variables, and then to confirm the conclusion by applying to detected patterns on the available data in the Data warehouse [5]. The vital goal of data mining is to predict - and predictive mining is the most regular type of data mining and also, prediction modeling is very used fully in all type of business applications. An appeal of market analysis comes from the clarity and utility of its results, which are in the form of association rules. There is an intuitive appeal to a market analysis because it expresses how tangible products and services relate to each other, how they tend to group together. A rule like, "if a customer purchases three-way calling, then that customer will also purchase call waiting" is clear [6]. Even better, it suggests a specific course of action, like bundling three-way calling with call waiting into a single service package. While association rules are easy to understand, they are not always useful [2].

II. HOW DOES ASSOCIATION RULE ANALYSIS WORK?

Association rule analysis starts with transactions containing one or more products or service offerings and some rudimentary information about the transaction. For the purpose of analysis, we call the products and service offerings items. Table 1 illustrates five transactions in a grocery store that carries five products. These transactions are simplified to include only the items purchased. Each of these transactions gives us information about which products are purchased with which other products. Using this data, we can create a co-occurrence table that tells the number of times that any pair of products was purchased together (see Table 2). For instance, by looking at the box where the "Soda" row intersects the "OJ" column, we see that two transactions contain both soda and orange juice.

TABLE 1
GROCERY POINT-OF-SALE TRANSACTIONS

Customer	Items
1	orange juice, soda
2	milk, orange juice, window cleaner
3	orange juice, detergent,
4	orange juice, detergent, soda
5	window cleaner, soda

The values along the diagonal (for instance, the value in the "OJ" column and the "OJ" row) represent the number of transactions containing just that item.

The co-occurrence table contains some simple patterns:

- OJ and soda are likely to be purchased together than any other two items.
- Detergent is never purchased with window cleaner or milk.
- Milk is never purchased with soda or detergent.

These simple observations are examples of associations and may suggest a formal rule like: "If a customer purchases soda, then the customer also purchases milk". For now, we defer discussion of how we find this rule automatically. Instead, we ask the question: How good is this rule? In the data, two of the five transactions include both soda and orange juice. These two transactions support the rule. Another way of expressing this is as a percentage. The support for the rule is two out of five or 40 percent.

TABLE 2
CO-OCCURRENCE OF PRODUCTS

Items	OJ	Cleaner	Milk	Soda	Detergent
OJ	4	1	1	2	1
Window Cleaner	1	2	1	1	0
Milk	1	1	1	0	0
Soda	2	1	0	3	1
Detergent	1	0	0	1	2

Since both the transactions that contain soda also contain orange juice, there is a high degree of confidence in the rule as well. In fact, every transaction that contains soda also contains orange juice, so the rule "if soda, then orange juice" has a confidence of 100 percent. We are less confident about the inverse rule, "if orange juice then soda", because of the four transactions with orange juice, only two also have soda. Its confidence, then, is just 50 percent. More formally, confidence is the ratio of the number of the transactions supporting the rule to the number of transactions where the conditional part of the rule holds. Another way of saying this is that confidence is the ratio of the number of transactions with all the items to the number of transactions with just the "if" items.

III. APRIORI ALGORITHM

Many algorithms for generating association rules were presented over time. Some well-known algorithms are Apriori, Eclat and FP-Growth, but they only do half the job, since they are algorithms for mining frequent itemsets. Another step needs to be done after to generate rules from frequent itemsets found in a database [9].

Apriori is the best-known algorithm to mine association rules. It uses a breadth-first search strategy to counting the support of item sets and uses a candidate generation function which exploits the downward closure property of support [8].

In computer science and data mining, Apriori is a classic algorithm for learning association rules. It 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 transactions (Winepi and Minepi), or having no timestamps (DNA sequencing).

As is common in association rule mining, given a set of item-sets (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 C of the item sets. Apriori uses a "bottom up" approach, where frequent subsets are extended one item [10].

IV. PROBLEM OF LARGE DATASETS

A typical fast-food restaurant offers several dozen items on its menu, says there are a 100. To use probabilities to generate association rules, counts have to be calculated for each combination of items. The number of combinations of a given size tends to grow exponentially. A combination with three items might be a small fries, cheeseburger, and medium diet Coke. On a menu with 100 items, how many combinations are there with three menu items? There are 161,700! (This is based on the binomial formula from mathematics). On the other hand, a typical supermarket has at least 10,000 different items in stock, and more typically 20,000 or 30,000. 14 Calculating the support, confidence, and improvement quickly gets out of hand as the number of items in the combinations grows. There are almost 50 million possible combinations of two items in the grocery store and over 100 billion combinations of three items. Although computers are getting faster and cheaper, it is still very expensive to calculate the counts for this number of combinations. Calculating the counts for five or more items is prohibitively expensive. The use of taxonomies reduces the number of items to a manageable size.

The number of transactions is also very large. In the course of a year, a decent-size chain of supermarkets will generate tens of millions of transactions. Each of these transactions consists of one or more items, often several dozen at a time [7]. So, determining if a particular combination of items is present in a particular transaction may require a bit of effort-multiplied a million-fold for all the transactions.

V. CONCLUSIONS

As the world becomes more of a global village being run by paperless systems, the idea of a cashless society is the hope of the future. Thus more innovations will still evolve which will make cashless transactions easily accessible and affordable.

the supermarket authorities want to find & analyze what similar products should be kept in one single shelf that matches with the common purchasing-tendency or mentality of the customers so that the customers do not waste their invaluable time in walking around a large supermarket and can purchase all the required items in a minimum possible timeframe. This will also eventually increase the sales & goodwill of the supermarket for bringing in ease & efficiency in product purchasing.

We can easily correlate the prototype with a large supermarket with thousands of products in which our algorithm would generate the item-sets that are frequently-purchased by the customers in the pairs of 1, 2, 3 or more.

And using that data, those items can be kept alongside in order to increase the sales & goodwill of the supermarket through saving the customer's time & effort during purchasing.

Working on this project was a good experience. I understand the importance of Planning and design as a part of software development. But it's very difficult to complete the program for a single person. Developing the project has helped us some experience on real-time development Procedures. Well, it's my pleasure to make a project for the title of "Sales Analysis"

REFERENCES

[1] S. S. S. R. H. M. Maike Krause-Traudes1, "Spatial data mining for retail sales forecasting," 11th AGILE International Conference on Geographic Information Science, pp. 1--11, 2008.

[2] B. G. W. a. M. Mateas, "A Data Mining Approach to Strategy Prediction," Computational Intelligence and Games, 2009. CIG 2009. IEEE Symposium on, pp. 140--147, 2009.

[3] W. J. a. P.-S. G. a. M. C. J. Frawley, "Knowledge discovery in databases: An overview," AI magazine, vol. 13, p. 57, 1992.

[4] S. a. P. S. K. a. M. P. Mitra, "Data mining in soft computing framework: a survey," IEEE, vol. 13, pp. 3--14, 2002.

[5] J. a. V. d. P. D. Burez, "Handling class imbalance in customer churn prediction," Elsevier, vol. 36, pp. 4626--4636, 2009.

[6] I. B. (. M. K. Ryszad S. Michalski (Editor), Machine Learning and Data Mining: Methods and Applications, Wiley, 1998.

[7] S. R. Ahmed, "Applications of Data Mining in Retail Business," IEEE, 2004.

[8] J. a. G. M. a. T. M. Garcke, "Data Mining for the category management in the retail market," springer, p. Springer, 2010.

[9] B. K. a. P. S. Bhardwaj, "Data Mining: A prediction for performance improvement using classification," arXiv preprint arXiv:1201.3418, 2012.

[10] K. B.Santhosh Kumar, "Implementation of Web Usage Mining Using APRIORI and FP Growth Algorithms," Int. J. of Advanced Networking and Applications, vol. 01, no. 06, pp. 400-404, 2010.