CONSUMER BUYING PATTERN ANALYSIS USING APRIORI ALGORITHM

AGM. Afrath and AL. Hanees

Department of Mathematical Sciences, Faculty of Applied Sciences
South Eastern University of Sri Lanka.
afrath.agm@gmail.com, hanees.al@gmail.com

Abstract

Business world of this era is very competitive and enterprises are trying to improve the experience of consumer by analyzing their past transactions. Some enterprises have even started to predict what people will buy next. However, so far the analyses were being done by considering all past transactions together. On the other hand some researches show that factors like age, marital status and gender are influencing the buying patterns. This research proposes an efficient approach using which predictions can be done with a reportable accuracy using large data sets. MySQL with Weka are used to generate association rules for an identified minimum support. The confidents of the association rules are very promising and this proves that the scenario based analysis provide high confidence value, in some cases it is 100%, than analyzing the whole transactions together.

Keywords: Apriori, Association rule, buying pattern, Weka

Introduction

Analyzing and understanding the buying patterns of the consumers have become an inevitable task for enterprises to sustain business. Enterprise like supermarkets that have a large collection of items need to plan things like what items to be put at the front, which items need to be kept close and the items to be placed to attract people like teens. Nowadays most of the online shopping portals also do buying analysis. When we pick an item, these portals are able to suggest other items that can be brought along with the item we have taken. The buying patterns analysis have become an inevitable task to improve the business and to complete in the market.

On the other hand, this kind of analysis may help consumers as well. It may help consumers to plan their shopping and sometimes suggest items that they may have forgotten.Most of the supermarkets and other consumer goods enterprises are issuing a loyalty card to track customer transactions. However, in the context of Sri Lanka these loyalties cards are still used to provide discounts and promote businesses, and not yet used to capture and analyze buying patterns.

There are many research works have been carried out to analyze buying patterns. Most of these researches find frequent item set of the whole transactions. On the other hand, there are researches to that the consumer buying patterns influenced by the personal attributes like marital status, age and gender. However, the reviews of online publications showed that the influence of these factors are not separately considered to improve the consumer experience earlier.

This study proposes that helps enterprises to predict consumer buying patterns using the personal attributes like gender, marital status and age. Further, enterprises can also do target marketing and improve the experience of the consumer by providing appropriate suggestions through the predictions.

Frequent item-sets generated by Apriori algorithm completely depend on a minimum support threshold. The researchers observed that the execution time of the proposed approach is inversely proportional to minimum support. Due to candidate item-sets generations, Apriori algorithm consumes a lot of computational power. The value of the confidence may vary between 0 and 1, and the association rules with the confidence of 1 are considered as strong rules.A method for measuring interests of a user are carried out using a concept of tree based on domain ontology, and it proposes a multi-agent based consumer behavior forecasting model in e-commerce.

Mining for Association rules can be extended to detecting fraud when the transaction is being completed. As there are many uses for association rules, then the obviously no single approach will be the most efficient in every situation. The performance of an algorithm depends not only on the execution speed, but also no other factors.

Apriori is suitable for small data sets because it takes a lot of space and time for larger datasets strong rules can be found in only from large datasets, but somehow we should accept that increase of data set take a long time and more spaces.

The problem of rule mining can be decomposed into two sub problems:

1. Generate all combinations of items that have fractional transaction support above a certain threshold, called minsupport. Call those combinations large itemsets, and all other combinations that do not meet the threshold small itemsets.

2. For a given large itemset $Y = I_1,I_2,...I_3,I_k, k \geq 2$, generate all rules(at the most k rules) that uses items from the set $I_1,I_2,...I_3,I_k$. If the itemset Y is large, then every subset of Y will also be large, and must have available support counts as the result of the solution of the first sub problem.

Frequent pattern mining

The problem of frequent pattern mining is that of finding relationship among the items in a database. The problem can be stated as follows.

Given a database D with transactions T_1, $T_2 \ldots T_N$, determine all patterns P that are present in at least a fraction S of the transactions. The fraction S is referred to as the minimum support. The parameter S can be expressed either as an absolute number, or as fraction of the total number of transactions in the database. Each transaction T_i can be considered a sparse binary vector, or as a set of discrete values representing the identifiers of the binary attributes that are instantiated to the value of 1. The frequent mining has implemented on numerous other applications in the context of data mining, the web log mining, sequential pattern mining, and software bug analysis

Association rule mining

Association rule mining is a widely-used approach in data mining. Association rules are capable of revealing all interesting relationship in a potentially large database. The abundance of information captured in the set of association rules can be used not only for describing the relationships in the database, but also for discriminating between different kinds or class of database instances. However, a major problem in association rule mining is its complexity. Even for moderate sized database it is intractable to find all the relationships.

Apriori Frequent Set Mining Algorithm

The Apriori algorithm is one of the most important and widely used algorithm for association rule mining. Most of the other algorithms are based on it or extensions of it. Apriori is a frequent itemset mining algorithm using transaction database. The research initially proposed this algorithm in 1993.

This algorithm executed in two major steps:

1. Frequent itemset generation, whose objective is to find all the itemsets that satisfy the minsup threshold. These itemsets are called frequent itemsets.
2. Association rule generation, whose objective is to extract all the high-confidence rules from the frequent itemsets found in the previous step. These rules are called strong rules.

Strength of association rules

Usually the algorithm will produce a large number of association rules. Therefore, it is important to find the rule set which generated with high strength. There are two ways of measuring the strength of the association rules namelyobjective measure and subjective measure. Subjective measures are not stable, and those are more oriented towards users who measure the strength that is hard to rely on. The unexpectedness and action-ability are two of such subjective measures used to measure the strength. Objective measures involve statistical analysis of the data such as support and confidence.

Support

The count for each item is increased by one every time when the item is encountered in different transaction T in a database D during the scanning process. It means that the support count does not take the quantity of the item into account. For example, in a transaction a customer buys three bottles of beers, but we only increase the support count number of beer by one. In other words, if a transaction contains an item, then the support count of this item is increased by one. Support is calculated by the formula shown below,

$$Support(X \Rightarrow Y) = \frac{\begin{matrix} Total\ number\ of\ transaction \\ that\ contains\ all\ the \\ items\ in\ X\ and\ Y \end{matrix}}{Total\ number\ of\ transaction\ in\ D}$$

Suppose the support of an item is 0.1%, it means only 0.1 percent of the transaction contain purchasing of this item. The retailer will not pay much attention to such kind of items that are not bought so frequently, obviously a high support is desired for more interesting association rules. Before the mining process, users can specify the minimum support as a threshold, which means they are only interested in certain association rules that are generated from those itemsets whose supports exceed that threshold.

However, sometimes even the itemsets are not as frequent as defined by the threshold, the association rules generated from them are still important. For example in the supermarket some items are very expensive, consequently they are not purchased so often as the threshold required, but association rules between those expensive items are as important as other frequently bought items to the retailer.

Confidence

Confidence of an association rule is defined as the percentage/fraction of the number of transactions that contain X ∪Y to the total number of records that contain X, where if the percentage exceeds the threshold of confidence an interesting association rule X ⇒ Y can be generated.

$$confidence\ (X \Rightarrow Y) = \frac{support\ (X \Rightarrow Y)}{support(X)}$$

Confidence is a measure of strength of the association rules, suppose the confidence of the association rule X ⇒ Y is 80%, it means that 80% of the transactions that contain X also contain Y together, similarly to ensure the interestingness of the rules specified minimum confidence is also pre-defined by users.

There are two other measures namely lift and conviction that are also used to measure the strength of an associationrule.

Weka data mining software

Weka is a collection of machine learning algorithms for data mining tasks. The algorithm can either be applied directly to a dataset or called from own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization.Weka is a state-of-the-art tool for developing machine learning techniques and their application to real-world data mining problems.

Mining big data using Weka

Weka is being used to make predictions in real time in very demanding real-world applications. This can be done with almost all Weka models once they have been built.

It is correct that it may be impossible to train models from large datasets using the Weka Explorer graphical user interface, even when the java heap size has already been increased, because the explorer always loads the entire dataset into the computer’s main memory and also incurs significant overhead due to visualization, etc.

When dealing with large datasets, it is the best to either user a command-line interface (CLI) to interact with Weka, use Weka’s knowledge flow graphical user interface, or write code directly in Java or a Java-based scripting language such as Groovy or Jython. Using these methods, it is possible to deal with larger datasets and even datasets that are too big to fit into main memory.

Most Weka classifiers require the entire dataset to be loaded into memory for training, but there are also schemes that can be trained in an incremental fashion, namely all classifiers implementing the interface. These are limited in number; however, for Weka, there is a library that provides access to the MOA data stream software containing state-of-the-art algorithms for large datasets or data streams.

Methodology

From the literature, it was evident that the Apriori algorithm will consume a lot of resources and show poor performance when it comes to a large set of transactions. Therefore, it was decided to divide the data set into many with a small number of transactions. However, this should be done meaningfully.

On the other hand, the literature proves that age, having children and gender are influencing the patterns of consumptions. Therefore, it was decided to filter the transactions based on the age, having children and gender. Therefore, the transaction database was divided into three based on age, number of children and gender, and each of these small set of transactions are called scenario.

Dataset

The dataset has been collected the data from GROCERY SHOPPING DATASETS Repository. The repository has three specialised dataset of shopping details. But the FoodMart (One of the dataset name on the repository) gives the detail information of what I'm expected to build the rule engine. This food market dataset is released from Microsoft and it has the good trade. It contains market baskets from 1560 products and 10281 users. This has been collected in form of structured query language file. Brand name of the goods are difficult to be used and analyzed as the brand name and its popularity may vary among countries. Therefore, it was decided to use categories of items instead of individual items.

Table 1. Data Set Details

Transactions	269720
Products	1560
Categories	110
Customers	10281

Data preparation

Using the help of MySQL Workbench and SQL Query language, I have done the data preparation activities on concerning different factors. So I have filtered the following tables which are needed for my research. The subjectively concentrated attribute could not easily produce association rules. Therefore, I have converted this attributes to the another form. The format conversions are below:

Birthdate attribute is converted into age then it is categorised.

```
if (age $\leq 19$)
    age_cat = 'Teen'
else if ( $20 \leq$ age$\leq 40$ )
    age_cat = 'Young Worker'
else if ( $41 \leq$ age$\leq 64$ )
    age_cat = 'Senior Worker'
else
    age_cat = 'Retired'
```

To reduce the complexity of the analysis, it was decided to track whether or not having children instead of number of children of a person. Therefore, the field where we stored the count of children was set to Boolean to denote whether a customer has a child or not.

Mining rules

For instance, if a person is retired and male, then first it will be checked with the rules that are available on the knowledgebase. Knowledge base is a database where the rules fordifferent conditions are stored. If a rule is available in the knowledge base then it will be used to make prediction.

If rules are not available for the given condition, then the list of transactions that comply with the condition will be extracted from the total transactions. Thereafter using the extracted transactions, rules will be mined using Apriori algorithm. For instance, in the given example, retired and male are two different conditions. Therefore, rules will be built separately for each scenario and the rules will be generated for combined condition as well.

Scenario Builder

Depends on the different behaviour of customers, I have decided to filter transactions from particular customer behaviour and generate rules based on his attribute. From the above equation, I can get the combination of user behaviours.

Rule Mining Engine

It is building rule on every working day and generating the rules in different behaviour of customers. The rules are generated by using the technique of Association Rule Mining.Investigates various amount of tool I have found Weka which is can do the data mining activities in real time what the database as it is.

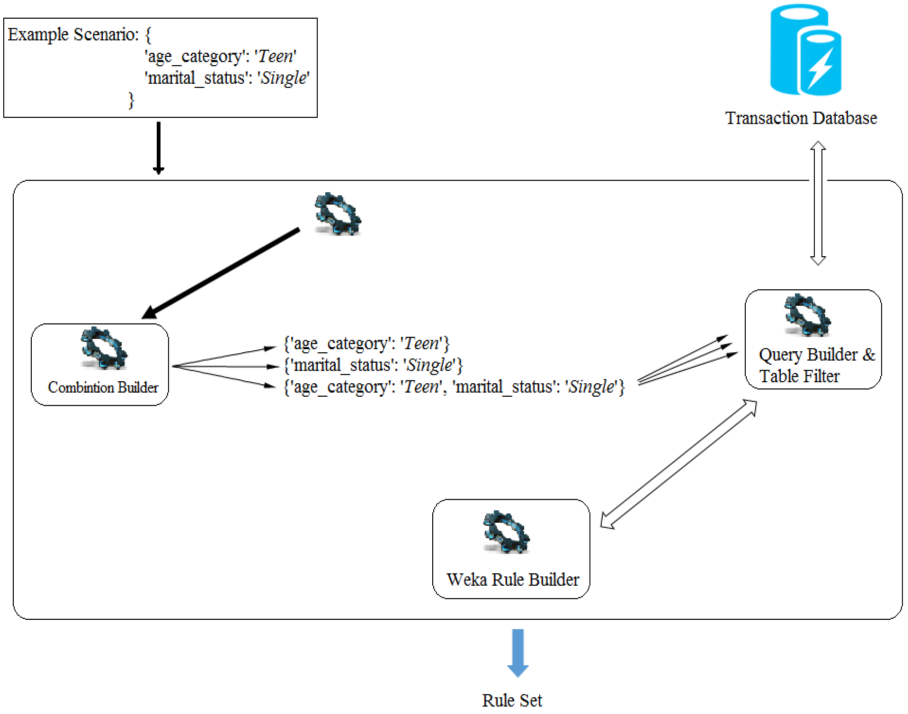
Example Scenario: {
'age_category': 'Teen'
'marital_status': 'Single'
}
Transaction Database
Combintion Builder
{'age_category': 'Teen'}
{'marital_status': 'Single'}
{'age_category': 'Teen', 'marital_status': 'Single'}
Query Builder & Table Filter
Weka Rule Builder
Rule Set

Figure 1: Work flow of rule mining engine

Considering the behaviour makes sense to predict the most relevant items which customers suppose think to buy. There are recommended the system in Market Basket Analysis but they all make the predictions based on only considering the customers transactions. Not considering the customers behaviours. Therefore, it is decided to use to predict customers buying behaviour based on their mental and physical behaviour is make sense and giving more meaningful suggestions to customers.

Implementation

Initially, tools like Weka was used to analyze the buying pattern and to find association rules. However, it was tough to process data and get the required results, as the dataset was in Structured Query Language format. Weka is supported for MySQL Workbench.

Build Scenarios and Extract Tables of corresponding to the Scenarios

The scenario is built by the user behaviour Different Set of users have their set of attributes like:

Person 1: He is a Male
He is in ’Young Worker’ age category
Person 2: She is Female
She is ’Teen.’
She has a car

So like above a different variety of behaviour people. For every set of people I have generated scenarios based on this behaviour subsets.

Example for Person 1:

```
[
        {gender = ’Male’},
        {age_cat = ’Young Worker’},
        {gender = ’Male’ AND age_cat=’Young Worker’}
]
```

Example for Person 2:

```
[
        {gender = ’Female’},
        {age_cat = ’Teen’},
        {cars_cat = ’Yes’},
        {gender=’Female’ AND age_cat = ’Teen’},
        {gender=’Female’ AND cars_cat =’Yes’},
        {age_cat=’Teen’ AND cars_cat =’Yes’},
        {gender=’Female’ AND age_cat=’Teen’ AND cars_cat =’Yes’}
]
```

The above sets are generated by the equation of mathematical combinations technique, if some behaviours are n then the combinations are n! The above scenarios are converted to queries and filter corresponding relations.

Results and Discussion

The whole transactions were first used to generate the association rules using the support of 0.001. It is found that the confidence was very low and not reportable. However, it was noted that almost all the consumers had bought ’fresh vegetables’ and ’fresh fruit’. Therefore, it was removed as it will not be much useful to analyze something that everyone is interested.

After removing the fresh vegetables and fresh fruit, the rules were again generated with the support of 0.001. Then 6124 rules were produced, and the confidence of the best rule was 24.89 percent. It was better than the other existing efforts. It was decided to improvethis by making scenarios as explained in the methodology. The results based on age and gender are shown in Table 2:

Table 2: Results for different categories of people

Category	Total Number of Rules	Best Rule’s Confidence

Teen	7306	83.33%
Young Worker	6798	36%
Senior Worker	6858	51.16%
Retire	6744	34.42%
Male	6108	30.97%
Female	6794	32.58%
Teen-Male	8872	100%
Young Worker – Male	6238	39.29%
Young Worker – Female	7782	60%
Senior Worker – Male	6970	47.61%
Senior Worker – Female	7378	57.89%
Retired Male	7376	60%
Retired Female	6966	55.55%

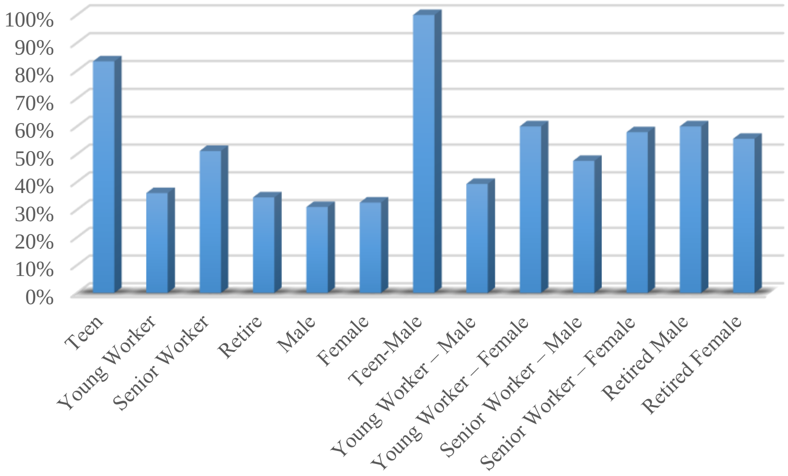

Figure 2. Best rule's confidence of categories

It clearly indicates that there are items that teens buy are not purchased by people with other ages. These rules can be used to do targeted marketing. For example, the adolescents who are females buy the items as shown in Table 3 along with soup that is not purchased by teen males.

Table 3: Teen male buying pattern while boughtsoup

Pre - Buying	Item Predicted
Soup	Milk
Soup	Preserves
Soup	Peanut Butter

Similarly, Table 4 shows the items that are bought by teen males but, not by teen females.

Table 4: Teen female buying pattern while bought soup

Pre - Buying	Item Predicted
Soup	Waffles
Soup	Soda

This kind of prediction helps enterprises to do target marketing and to improve the consumer experience. Interestingly, it is noticed that teen males and teen females buy certain items which are not purchased by people who are teen; The confidence of that rules were 100 percent for the above said cases. For instance, not in their teen Table 5 shows the items to which 100 percent confidence was obtained, which means those items are not bought by other people.

Table 5: Teen male and teen female buying patterns

Pre - Buying	Predicted Item	Confidence
Deli Meats, Juice Soda	Dried Fruit	100%
Popsicles, Tuna	Pasta	100%
Cereal, Cooking Oil, Flavored Drinks	Waffles	100%
Coffee, Dips	Dried Fruits	100%
Popcorn, Sardines	Pizza	100%
Waffles, Ibuprofen	Cooking Oil	100%
Frozen Vegetable, Chocolate Candy, French Fries	Muffins	100%
Light bulbs ,Pancake Mix	Waffles	100%
Cookies, Aspirin	Dried Fruit	100%
Nuts, Home Magazines	Cereal	100%
Dried Fruit, Juice Shampoo	Deli Meats	100%

Conclusion

This research has been done to give the prediction in real-time basis. The results of the research using the proposed scenario based approach where the whole transactions are considered together. For certain association rules as shown in Table 5, 100 percent confident have been obtained. Since the Apriori algorithm is used to mine rules for each scenario. In this research, the categories of an item, like milk powder or soup, are considered instead of individual items or brand names. The research can also be used to carry out targeted marketing.

Future Scope

In future, the research will be extended by adding more factors like income and assets in addition to age, gender and marital status that are considered now. The Apriori algorithm can be slow and the bottleneck is candidate generation. So, In future, we can use another algorithm such as FP GROWTH algorithm to minimize the time and reduce the memory space.

References

[1] R. Agrawal, T. Imieliski and A. Swami (1993), "Mining association rules between sets of items in large databases", *ACM SIGMOD Record*, vol. 22, no. 2, pp. 207-216.

[2] K.S. Adewole, A.G. Akintola A.R. Ajiboye S.O. Abdulsalam (2014). "Frequent Pattern and Association Rule Mining from Inventory Database Using Apriori Algorithm."*Afr J. of CompICTs*. Vol 7, No. 3, Pp.35-42.

[3] Kenneth Lai and Narciso Cerpa, "Support vs Confidence in Association Rule Algorithm,"*In Proceedings of the OPTIMA Conference,* Curic. 2001.

[4] Silberschatz Abraham, Alexander Tuzhilin (1995), "On subjective measures of interestingness in knowledge discovery", *In Proceedings of the First International Conference on Knowledge Discovery and Data Mining*, vol. 95, pp. 275-281.

[5] Rakesh Agrawal and R. Srikant (1994). "Fast Algorithms for Mining Association Rules in Large Databases", *Proceedings of the 20th International Conference on Very Large Data Bases,* pp. 487-499, ISBN: 1-55860-153-8.

[6] Prashasti Kanikar, Parekh, Twinkle Puri, Binita Shah, Ishaan Bajaj, Binita, "Comparative Study of Apriori Algorithm Performance on Different Datasets", *International Journal of Engineering Research and Applications,*ISSN : 2248- 9622,Vol. 4, Issue 4, April 2014, pp.38-43.

[7] Big Data Analytics, Enterprise Analytics, Data Mining Software, Statistical Analysis, Predictive Analtyics", Statsoft.com, 2016. [Online]. Available: http: //www.statsoft.com/. [Accessed: 20- Apr- 2016].

[8] Deepa S. Deshpande, "A Novel Approach for Association Rule Miningusing Pattern Generation," *I.J. Information Technology and Computer Science*, 2014, 11, pp. 59-65.

[9] Grocery Shopping Mall Dataset from Microsoft food market [Online].Available:http://recsyswiki.com/wiki/Grocery_shopping_datasets

[10] A Brief Literature Review on Consumer Buying Behaviour, Available:http://research-methodology.net/a-brief-literature-review-on-consumer-buying-behaviour/

[11] Age Discrimination", Eeoc.gov, 2016. [Online]. Available: http://www.eeoc.gov/laws/types/age.cfm. [Accessed: 20- Apr- 2016]

[12] ‘Weka 3: Data Mining Software in Java’, Available: http://www.cs.waikato.ac.nz/ml/weka/index.html.