# Market Basket Analysis based on Apriori and CART

**Liyuan Wang, Jianqin Sun**

School of Statistics, Shanxi University of Finance and Economics, Wucheng Road, Taiyuan, China

**Keywords:** Market Basket Analysis; Association Rules; Decision Tree; SPSS Modeler

**Abstract:** With the rapid development of economy and information technology, the development of the market retail industry cannot be underestimated. How can it improve the efficiency of the retail industry? The paper which uses Apriori algorithm to find out the data of shopping basket from the massive data of consumers reveals the relationship between the purchased goods, and subsequently applies the association rules and CART decision tree algorithm to reveal the characteristics of the customer group and the target customers classification. In order to dig out more detailed and valuable information, it is convenient for the goods to be better configured and sold, and to improve the operational efficiency of the market.

## 1. Introduction

A shopping basket refers to a basket or a trolley used in a supermarket for customers to purchase goods. When the customer pays, the goods in the shopping baskets are registered and settled by the salesperson through the cash register. The so-called "Market Basket Analysis" is to study the clients' purchase behavior through the information displayed in these shopping baskets. From the analysis and research of consumers' shopping basket information, we can get data such as their purchasing habits, product preferences, brand loyalty and so on. Later we will use the "association rule model" in SPSS Modeler software to analyze commodity transaction flow data to find reasonable merchandise placement rules and help increase sales.

## 2. Introduction of Model Methods

Association rules represent the correlation of different data items in the same event while the Apriori algorithm is a frequent item set algorithm that can mine association rules whose core notion is to repeatedly scan the database. Provided that the association satisfies the minimum support threshold and the minimum confidence threshold, succeeding the association rules is meaningful. The decision tree is a predictive model representing a mapping relationship between object properties and object values. It is a type of supervised learning. The classifier obtained by sample learning can give correct classification to emerging objects. The basic standpoint is to recursively divide the training samples into space of independent variables. Furthermore, use the verification data for pruning. The Bayesian classification is characterized by the use of probabilities to represent all forms of uncertainty, and the core is the conditional probability which is the probability of occurrence of event A under the condition that another event B has previously occurred, expressed as P(A|B).

## 3. Empirical Analysis

### 3.1 Data Processing

After importing the data into SPSS Modeler, add a "Data Audit" node to check whether the data has missing or abnormal values. After inspection, 1000 records are valid, that is data integrity, and there are no missing values and outliers. What's more, set the role of the continuous variable to "None" in the "Type" node and the role of the discrete (category) variable to "Both", afterwards add the table output to check data. Finally connect a "Partition" node and divide the data into training and test sets to assess its performance.

  DOI: 10.25236/etmhs.2019.311

### 3.2 Application of Apriori Algorithm in Market Basket Issue

After establishing the Apriori model in SPSS Modeler, output the network diagram by connecting a "web" node, as shown in Fig 1, from which you can get the relationship between different commodities in the shopping basket. Three thick lines are as mentioned above high-frequency project groups. In accordance with their purchase behavior, we can name the three types of customers. The first type purchases fish, fruit and vegetables called robust diners and named them "healthy"; the second type prefers food and wine giving priority to the quality of lives called "quality". The third category is unhealthy diners eating junk food such as beer and frozen meat. Ultimately output them as nodes and link them in turn.

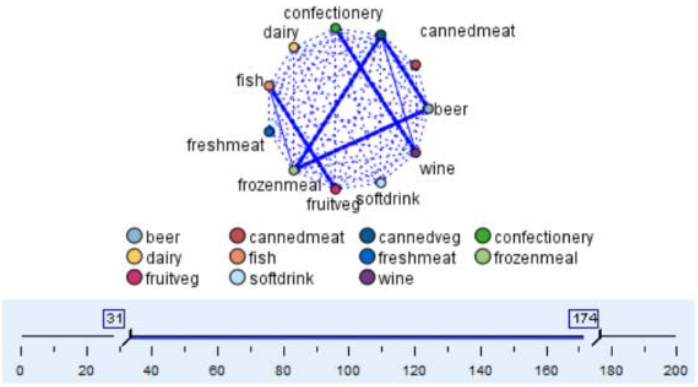


Fig. 1 Network analysis diagram

Although the Apriori algorithm can effectively generate association rules, there are some drawbacks that the algorithm is not efficient: 1. A large number of candidate sets were likely to be generated, and elements that should not participate in the combination are not excluded. 2. It is necessary to repeatedly scan the database. Suppose that the amount of data is large, it takes an army of time to apply the algorithm to generate a candidate set for each iteration, and the cost is relevant to the number of records in the database which represents an increase in the geometric progression.

### 3.3 Construction of Decision Tree

According to the correlation between sales items based on the association rules, the three types of target customers are defined above. In order to get the characteristics of the target customer, variables such as value, pmethod, sex, homeown, income, age are used as input variables, and unhealthy (healthy, quality) is used as the output variable. Since the output variable is not continuous, but classified, thus CART generates a regression decision tree with a depth of 4. The first node is income, which is the most important variable when dividing.

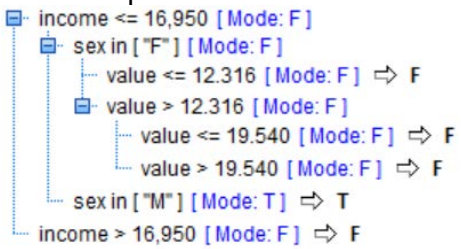


Fig. 2 Rule set of unhealthy

#### 3.3.1 Optimization of CART algorithm - pruning

Since the decision tree is completely dependent on the training samples, the decision tree can bring a perfect fitting effect on the training samples. Nevertheless, such a decision tree is too large and complex for the test sample, even may result in a high error classification rate. This phenomenon is known as overfitting. Therefore, it is necessary to cut the intricate decision tree and remove some nodes to solve the over-fitting issue, that is, pruning. The pruning method is divided

into 2 categories: pre-pruning and post-pruning. Pre-prune is the process of creating a decision tree, prematurely terminating the decision tree to avoid excessive node generation. The pre-pruning method is simple but not practical, since it is difficult to accurately determine when to terminate the length of the tree. After the pruning is completed, the node subtrees whose confidence level is not up to standard are replaced by leaf nodes whose class labels are marked with the highest frequent class in the node subtree. Here we use post-pruning measure of decision tree to prevent overfitting. Since there is a "Prune tree to avoid overfitting" option in the software, we can directly check this option. Figure 3 shows the decision tree after pruning. The tree with a depth of 2, which not only copes with the over-fitting problem, but also simplifies the decision tree.

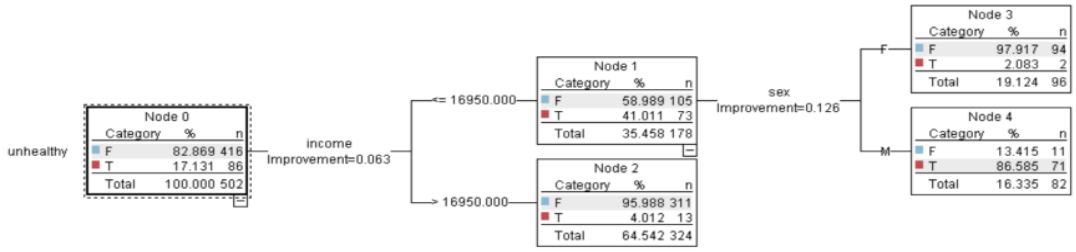


Fig. 3 Decision Tree

The income variable has a Gini value of 0.063 minimum. As the root node, if income is greater than 16950 or less than or equal to 16950 and the gender is female, the customer is considered not to be "unhealthy". If the income is less than or equal to 16950 and the gender is male, it will be considered to be "unhealthy". Of the six attributes entered, only the two attributes of income and gender are used, because these two attributes are the most important in the prediction.

### 3.3.2 Evaluation Models

(1) Confusion Matrix

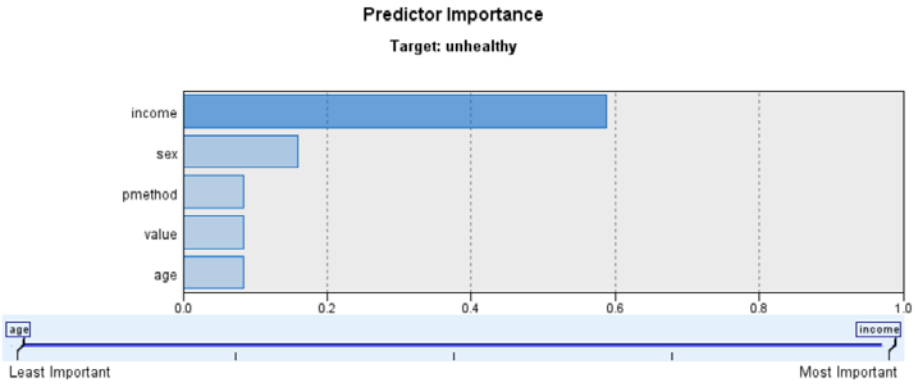


Fig. 4 Predictor Importance

According to the "unhealthy" target customer classification, the order of importance of variables can be obtained. Income is the most important variable when classifying customers, while age is the least significant one. In order to assess the performance of the algorithm, the precision and recall of the algorithm under the test set are tested. Thus the confusion matrix can be obtained from the distribution chart of loss function.

Table 1 Confusion Matrix

| Actuality | Forecast | |
|---|---|---|
| | P | N |
| T | 44 | 8 |
| F | 11 | 237 |

Through the superiority of the confusion matrix analysis algorithm, the recall ratio R is 84.61%,

and the precision P is 80%. Nevertheless, the recall rate and the precision rate are a pair of contradictory measures. Generally, when the recall rate is high, the precision is often low and vice versas. So it requires finding a balance point (BEP), the value of R= P. But BEP is still too simplistic, and F1 metric is widely used more commonly, where N is the total number of samples.

$$F_1 = \frac{2 \times P \times R}{P + R} = \frac{2 \times TP}{N + TP - TN}$$

(2) Receiver Operating Characteristic (ROC) curve - AUC

In different learning tasks, we were likely to pay more attention to the recall rate or the precision rate. At this time, we can use the ROC curve to study the generalization performance of the learning period. ROC full name is the "receiver operating characteristic" curve. When the learner is compared, if the ROC curves of the two learners cross each other, it is difficult to generally assert that which is superior or inferior. However, if it is necessary to make a comparison, a more reasonable basis is to compare the area under the ROC withdrawal, that is, AUC. In order to compare and analyze with other models, we can use an automatic classifier to judge the CART algorithm in the analysis of the shopping basket issue by automatically modeling from quite a few models. Whether the property is excellent, the performance of the model is measured by its AUC value, gain ratio, and overall accuracy.

From the results of the automatic classifier, it can be seen that in many models, based on the AUC, the CART algorithm has the AUC value of 0.916, which is the most excellent, so the CART algorithm is the best; the "Evaluation" node connected after the automatic classifier is used to output scatter plots with six charts, Response, Gains, and ROC, where the Gains reflects the recall ratio and the Response reflects the precision.

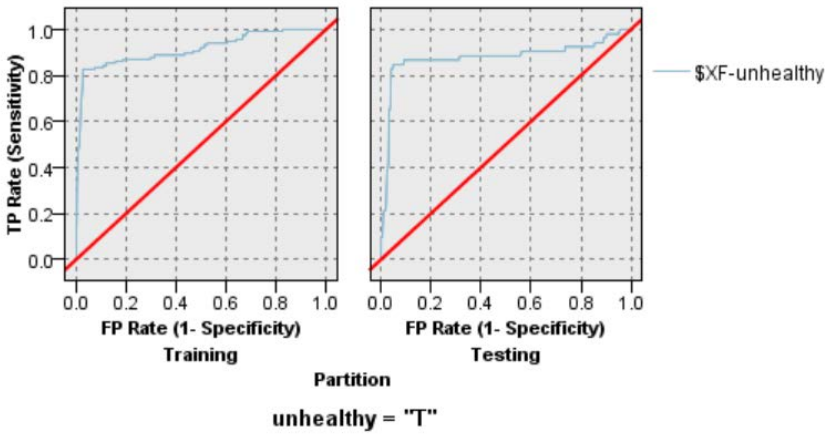


Fig. 5 Receiver Operating Characteristic

(3) Misclassification Matrix

When constructing the decision tree, if the interactive dialogue is applied, the tree growth set and the node information, indicator information and the misclassification matrix of the over-fitting set can be obtained. It can be seen that whether it is a tree growth set or a preventive over-fitting set, the risk estimates are relatively insignificant, 0.052 and 0.045, that is, the CART algorithm has a good performance when analyzing the shopping basket problem. Since the CART algorithm only classifies the target group when the customer property is known, but how to infer more information about the customer when only knowing the customer's minority information is a problem, or Bayesian classification is useful. From the above analysis, we can know that the CART algorithm has an exceedingly good performance when classifying target customers and analyzing the shopping basket issues. Through the CART algorithm, you can analyze the customer's shopping basket information (such as the total price of the basket, mode of payment, etc.) and personal information (such as salary, gender, etc.) to analyze which target group the customer belongs to, and apply it to various fields, especially in the field of retail and e-commerce. Based on the association rules, the correlation between sales and target customers is analyzed for how to best match the sales,

achieve accurate delivery, and accurately analyze the customer's own quality.

### 3.4 Bayesian classification

For Bayesian classification, the value of the target variable unhealthy is known, and the conditional probability is equal to the number of samples for each value of the attribute divided by the total number of samples. Among the 1000 customers, there are 167 unhealthy, so the conditional probability of $unhealthy = T$ is 0.16; similarly, the conditional probability of $unhealthy = F$ is 0.84. The importance of predictor variables is as follows: $\text{income} > \text{sex} > \text{homeown} > \text{pmethod} > \text{age} > \text{value}$ The Bayesian network is a directed acyclic graph. The direction of the arrow in the figure can find the dependencies between variables. As an example, when the customer is known to be unhealthy and the gender is known, the conditional probability that the customer's income belongs to a certain interval can be obtained.

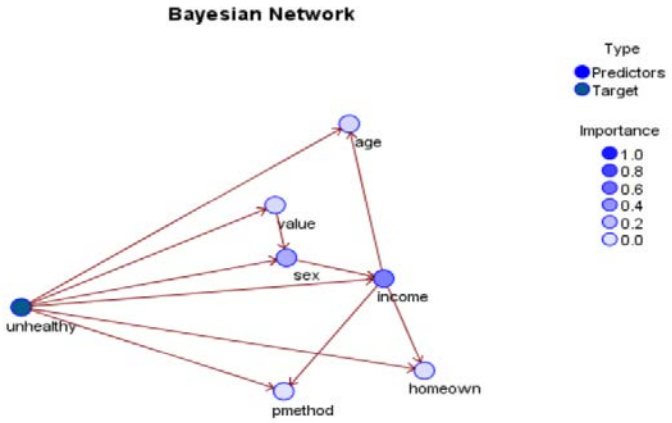


Fig. 6 Bayesian network

From the conditional probability map, the probability that a male who wants to purchase beer, frozen meat and canned vegetables at the same time has a lower income of 20,920 is a lower income group. For women, whether they are unhealthy or not, their income distribution is mainly in two extreme ranges, either high or low. Similarly, the same is true for speculating other variables. Here, merely the income is regarded as an example.

## 4. Conclusion and Suggestion

### 4.1 Result Analysis

The paper first introduces some basic information of data mining and association rules, and carefully analyzes and explains its classical association rule algorithm Apriori algorithm using practical examples to research and analyze. Secondly, for the supermarket shopping basket problem, the association rules are analyzed, and the SPSS Modeler software is used for presentation and interpretation, and the strong correlation between commodities is analyzed.

Through the Apriori algorithm, we find the correlation between the products purchased by some customers, so that they can be applied to various fields. For example, in a supermarket, how to place goods on the shelf, we can think about putting wine and grain together, and canned food and so on together with beer. This can promote the sale of goods, or bundle the relevant products, which will greatly increase the sales volume of goods. Through the analysis of customers with different attributes and their shopping basket contents, the three most relevant combinations are: fish and fruit and vegetables - healthy customers, wine and fruit - high quality customers, beer, frozen meat and canned vegetables - unhealthy customers.

### 4.2 Suggestions of Supermarket Retail

Through the analysis above, we can implement applicable marketing program: 1) Display fish, fruit and vegetable, wine and grain, beer, frozen meat and canned vegetables respectively in one

area, increasing the number of purchases by customers. 2) Carry out bundled sales, combine the related products, pricing, and increase the purchase quantity. 3) Focus on men in certain age groups or income segments, and occasionally hold discount promotions to attract consumers. The phenomenon of "data explosion but lack of knowledge" has become a heart disease for business, but data mining technology has indeed changed the status quo and is enough to solve their anxiety. Technology is changing with each passing day. How to utilize the latest research results and the existing algorithm features of association rules to further promote the development of association rules, especially in artificial intelligence and machine learning, and greatly promote its further research.

## References


[1] SUN Ximing, GONG Chengfang. Application of Association Rules in Shopping Basket Analysis [J].Computer and Digital Engineering, 2008(6):57-60

[2] Huang Wei, Zhang Weiwei. Analysis of Shopping Basket Based on Association Rules——Taking the Shopping Basket of Sunflower Farm as an Example[J].Wireless Interconnect Technology, 2018, 15(08): 105-106.

[3] Li Ming. Application of Association Rules Algorithm Based on R Software in Shopping Basket Analysis [D]. Central China Normal University, 2017.

[4] Zhang Zhibin. Analysis of Data Mining Process Based on SPSS Modeler [J].Digital Technology and Application, 2017(09):72-73.

[5] Luo Mingjian. ROC-based classification algorithm evaluation method [D]. Wuhan University of Science and Technology, 2005.

[6] Li Jie-yi, "Application research in Internet information retrieval of data mining technology," 2013 2nd International Symposium on Instrumentation and Measurement, Sensor Network and Automation (IMSNA), Toronto, ON, 2013, pp. 1071-1074.

[7] Zhang Pengpeng. Application of data warehouse and data mining technology in supermarket CRM [D]. Hebei University of Science and Technology, 2013.

[8] Johnson, R.A. and Wichern, D.W. (2003). AppliedMultivariate Statistical Analysis [M]. 5th ed. Pearson Education, g.P.R.China.

[9] Yi Liu, Jiawen Peng, and Zhihao Yu. 2018. Big Data Platform Architecture under the Background of Financial Technology: In the Insurance Industry As An Example. In Proceedings of the 2018 International Conference on Big Data Engineering and Technology (BDET 2018). ACM, New York, NY, USA, 31-35.