Research Guru
Online Journal of Multidisciplinary Subjects (ISSN : 2349-266X)
Volume-13, Issue-3, December-2019
Impact Factor:4.081
www.researchguru.net

Predictions, Analysis and study of Association Rules using the Non-Linear Regression Model

G.G. Shah

Assistant Professor, Faculty of Business Administration, Dharmsinh Desai University, Nadiad

Dr. H. N. Patel

Assistant Professor, Computer Science Department, Dr. BabaSaheb Ambedkar Open University,Ahmedabad

Abstract :

Because of the rapid growth in worldwide information, efficiency of association rules mining has been concerned for several years. Association rule mining plays vital part in knowledge mining. Apriori algorithm of Association rules based on interestingness measures like support, confidence etc.Traditional association rule mining techniques employ predefined support and confidence values. However, specifying minimum support value of the mined rules in advance often leads to either too many or too few rules, which negatively impacts the performance of the overall system.

Usually, the minsup and minconf are chosen by trial-and-error methods, so the threshold selection is a time consuming task especially when users work on larger scale datasets.

The majority of the existing literature focuses on improving computational efficiency of the mining algorithms, but little work has been done to address the inefficiency in the threshold selection. To solve this problem, proposes a regression based approach to improve the efficiency of the threshold selection of measures like support and confidence in association rule mining. In this research paper proposed non-linear regression analysis to detect potential relationships between support and confidence. This proposed non-linear regression analysis can be used in different types of dataset also.

We use multiple correlation coefficients to test the fitting effects of the proposed equation and use significance test to verify whether the coefficients of parameters are significantly zero or not. We examine the rules from a more rigorous point of view by conducting statistical tests. Specifically, we use F-test, t-test. Different types of advanced Statistical Tests are generated with the help of R-language using and its associated packages.

***Key-word** : Association Rule, Regression Model, Regression Coefficients, Multiple Correlation Coefficient, R -language*

1.Introduction

Data mining is a process of extraction of useful information and patterns from huge data. The process of discovering useful knowledge from a huge data is called as Knowledge Discovery in Database(KDD) and which is often referred to as Data Mining. Data mining is a logical process that is used to search through large amount of data in order to find useful data. Data mining is in fact a broad area which combines research in statistics, database, market basket analysis etc.

Association Rules of Mining introduced by Agrawal[1] is an important research topic among the various data mining problems.Association Rule mining is an important research issue in the area of data mining and knowledge discovery for purposes of data analysis, decision making and pattern discovery[2].Association rules are one of the most important knowledge of data mining's result which can be defined as the relation between the itemsets by given support and confidence in database.

One of the most difficult problems in association rule mining is to handle the vast quantities of items in a dataset .It often requires long execution time and complex data analysis over the entire dataset. Therefore, the research in association rule mining focuses primarily on the efficiency of generating frequent itemsets. To improve the efficiency and its results, a large variety of algorithms and frameworks have been developed for Association Rule mining in last two decades.

Some important algorithms are AIS[1], SETM[3],Apriori[4], Agrawal et at.[4] proposed two algorithms called Apriori and AprioriTid to discover significant association rules between items in large datasets.,DHP[5], CHARM[6],FP Growth[7],RARM[8],Closet+[9]. Shi et al.[10] used association rules for credit evaluation in the domain of farmer's credit history. Z.He., et al.[11] used an approach to finding optimized correlated association rules. Ye. D. Q. et. al[12] used Correlation technique research of association rule based on linear regression. Very few researcher have use the Regression model for the association rule mining using the measures like support, confidence etc.The rest of this paper is organized as follows.

Section 2, we have introduced Association Rule and its basic knowledge with fundamental parameter Support and Confidence. Apriori algorithm is the classic algorithm for Association Rule Mining we find the rules using R-language. Section 3, the core part of the Research paper say Regression Model to be construct and using non-linear regression model, statistical and coefficient test are generate. Section 4,The Case Study for the general dataset is apply to the non-linear regression model and from that Coefficient Evaluation and Coefficient test are used for the prediction of the

regression model. In Section 5 compares our approach with some other related work, Final Conclusion and future work.

2. Association Rule

Association rules[1] provide information in the form of "if-then" statements. These rules are computed from the data and unlike the rules of logic they are probabilistic in nature. Association rule mining searches for interesting relationships among items in a given data set under minimum support and confidence conditions.

The problem of finding association rules X ⟹ Y was first introduced in [1] as a data mining task of finding frequently co-occurring items in a large Boolean transaction database. An association rule X ⟹ Y means if consumer buys the set of items X, then he/she probably also buys items Y.

2.1. Basic knowledge6

Let I $=\{i_1, i_2 \ldots., i_n\}$ be a set of Data items. Each transaction T is a set of items, such that T ⊆ I.For example, this may correspond to a set of items which a consumer may buy in a basket transaction. An Association rule is a condition of the form of X⟹ Y Support :

Formula :

$$\textbf{Support (X}\rightarrow \boldsymbol{Y}) = \frac{\boldsymbol{Frequency\ (X \cup Y)}}{\boldsymbol{|D|}} = \boldsymbol{Prob(X\ \cup Y)}$$

Support means probability of frequents' occurrence, It is very difficult to determine the appropriate threshold of support in practical applications. If the value of the minsup is low, it will generate a large amount of association rules,it also including many rules that we are not interested. In practical life an itemset with a low probability of emergence leads to an itemset with a high probability of emergence. Here,such rules might not be useful for users.If the value of minsup is high, it will lose many low support patterns,particularly those with long patterns and low support. Those patterns may be of great significance for finding combinations of a number of itemsets. Therefore, the setting of minsup will be a repeated adjustment process.

Confidence :

Formula :

$$\boldsymbol{Conf\ (X \rightarrow Y) = \frac{Sup\ (X\ \cup Y)}{Sup\ (X)} = Prob(\frac{Y}{X})}$$

Confidence means the probability that some itemsets emergence will lead to others' occurrence.It signifies the likelihood of items Y being purchased when item X is purchased.It can also be interpreted as the conditional probability P(Y/X) i.e. probability of finding the itemset Y in transactions given that transaction already contains X.

2.3.Apriori Algorithm

The Apriori algorithm is the common algorithm proposed to find frequent itemsets in association rule mining[14].It proceeds by identifying the frequent individual items in the database and extending them to larger and larger item sets as long as those item sets appear sufficiently often in the database.It is very important for effective Market Basket Analysis and it helps the customers in purchasing their items with more ease which increases the sales of the markets.

Apriori Algorithm (Agrawal et al., 1999) is the most classical and important algorithm for mining frequent itemset. Apriori Algorithm is used for to find all frequent itemsets in a given database.Apriori uses breadth first method and a hash tree structure to count candidate item sets efficiently. It assume all data are categorical. Apriori Algorithm works on two concepts (i) Joining (ii) Pruning.

Where (k-1) itemsets are used to generate k itemsets.

Guideline of Apriori Algorithm is

(i). Subset of regular itemsets are frequent itemsets

(ii). Supersets of rare itemsets are infrequent itemsets.

Apriori algorithm utilize level wise search itemsets for items k are used to extend of size L_{k+1}.

2.4. Rules Generation and Analysis using R-language

There are number of software available in the market for Data mining techniques like SPSS, SAS, WEKA, ORANGE, XLMINER, RAPREDUCE and language like R also. Each software are unique and they have own ability and capacity to generate different types of result, which is the requirement of the users. R is an open source programming language and software platform that provides statistical computing and visualization capabilities. In this Research we have used R-language for to generate the Association Rules. R-language mainly related with Statistical Programming and also Data Mining related techniques. Number of features and packages are available in the latest version like 3.05.1.We used R-language due to its versatile different packages, library. We apply Packages like arules and arulesViz are the fundamental packages for Association rule mining.

3.Design of Regression Model

3.1.Regression Analysis Overview

Regression analysis is the process of constructing a mathematical model or function that can be used to predict or determine one variable by another variable or other variable.The most elementary regression model is called simple regression or bivariate regression involving two variables in which one variable is predicted by another variable. In simple regression, the variable to be predicted is called the dependent variable and is designated y. The predictor is called the independent variable, or explanatory variable and is designated as x. In simple regression analysis, only a straight-line relationship between two variable is examined

The first step in determining the equation of the regression line that passes through the sample data is to establish the equation's form. Several different types of equations of line are discussed in algebra, finite math or analytic geometry courses. Recall that among these equation of a line are the two-point form, the point-slop form and the slope-intercept form.

Regression analysis serves three major purposes : (i) description (ii) control (iii) prediction. We frequent use equations to summarize or describe a set of data. Regression analysis is helpful in developing such equation. For example we may collect a considerable amount of health data, in which health dependent upon so many factor like a regression model would probably be a much more convenient and useful summary of those data that a table or even a graph.Besides prediction, regression models may be used for control purposes. A cause and effect relationship may not be necessary if the equation is to be used only for prediction. In this case it is only necessary that the relationships that existed in the original data used to build the regression equation is still valid.

A functional relation between two variables is expressed by a mathematical formula. If X denotes the independent variable and Y the dependent variable, a functional relation is of the form

$$Y = f(X)$$

Given a particular value of X, the function *f* indicates the corresponding value of Y. A statistical relation, unlike a function is not a perfect one. In general, the observations for a statistical relation do not fall directly on the curve relationship.

Depending on the nature of the relationships between X and Y, regression approach may be classified into to broad viz., linear regression model and nonlinear regression models.The response variable is generally related to other causal variables through some parameters. The models that are linear in these parameters are known as linear models, whereas in nonlinear models parameters are appear nonlinearly. Linear models are generally satisfactory approximations for most regression applications. These are occasions, however when an empirically indicated or a theoretically justified nonlinear model is more appropriate.

3.2.Design Regression line Equation.

In this subsequent, a regression model is developed in general domains to explore the relationship between the number of association rules and (support,confidence).For the observed dataset,the number of association rules in a dataset depends on support and confidence. In this research paper Association Rules are the support,confidence frame work.In particular, the smaller the support and confidence, the greater the number of association rules. However, influence of confidence on the number of the association rules is stronger than influence of support on the number of the association rules.For the regression line relationship between support, confidence and the number of association rules are used. Support and Confidence based on probability, we propose a nonlinear regression equation to predict the potential number of association rules on datasets in general domain.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 \frac{1}{x_1} + \beta_4 \frac{1}{x_2} \qquad (1)$$

The reasons we use $1/x_1$ and $1/x_2$ in the models are :

1.Support and Confidence have a non-linear relationship with the number of association rules

2.Support and Confidence have an inverse relationship with the number of association.

3.A smaller threshold of support leads to more number of rules and bigger threshold of support leads to less number of rules.

4.The range of threshold of support is $(\propto,1)$, where α is the smaller value of the threshold of support.

5.Similarly,the effect of smaller threshold of confidence is the same as smaller threshold of support.

6.Confidence (A $\Longrightarrow$ B) =$\frac{P(A\cup B)}{P(A)}$,here we can see that if $P(A \cup B) = \propto$ and $P(A) = 1$,confidence ($\Longrightarrow$ B) will obtain the minimum value $\propto$.Therefore, the range of smaller threshold of confidence is ($\propto$,1).

Here to transform above equation into linear equation can be obtained as.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \in \quad (2)$$

After defining the regression equation, we need to determine the values of five coefficients in order to minimize the residual sum of squares (RSS) say $\sum_{i=1}^{n} {\varepsilon_i}^2$ .And it is calculated by :

$$\sum_{i=1}^{n} {\varepsilon_i}^2 = \sum_{i=1}^{n} (y_i - \widehat{y_i})^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_1 - \beta_2 x_2 - \beta_3 x_3 - \beta_4 x_4)^2 \quad (3)$$

Where y_i is the real number of association rules in a data sample, and $\widehat{y_i}$ is the predicted number of association rules in a data sample? Using the partial derivative for $\sum_{i=1}^{n} {\varepsilon_i}^2$ on $\beta_0, \beta_1, \beta_2, \beta_3 \; and \; \beta_4$ respectively, and let their results equal to zero.

$$\frac{\partial \sum {e_i}^2}{\partial \beta_0} = -2 \sum (y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i} - \beta_3 x_{3i} - \beta_4 x_{4i}) = 0$$

$$\frac{\partial \sum {e_i}^2}{\partial \beta_1} = -2 \sum (y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i} - \beta_3 x_{3i} - \beta_4 x_{4i})\, x_{1i} = 0$$

$$\frac{\partial \sum {e_i}^2}{\partial \beta_2} = -2 \sum (y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i} - \beta_3 x_{3i} - \beta_4 x_{4i})\, x_{2i} = 0 \quad (4)$$

$$\frac{\partial \sum {e_i}^2}{\partial \beta_3} = -2 \sum (y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i} - \beta_3 x_{3i} - \beta_4 x_{4i})\, x_{3i} = 0$$

$$\frac{\partial \sum {e_i}^2}{\partial \beta_4} = -2 \sum (y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i} - \beta_3 x_{3i} - \beta_4 x_{4i})\, x_{4i} = 0$$

After carrying out the partial derivatives, we have

$$\sum y_i = n\beta_0 + \beta_1 \sum x_{1i} + \beta_2 \sum x_{2i} + \beta_3 \sum x_{3i} + \beta_4 \sum x_{4i}$$

$$\sum y_i\, x_{1i} = \beta_0 \sum x_{1i} + \beta_1 \sum {x_{1i}}^2 + \beta_2 \sum x_{1i}\, x_{2i} + \beta_3 \sum x_{1i}\, x_{3i} + \beta_4 \sum x_{1i}\, x_{4i}$$

$$\sum y_i\, x_{2i} = \beta_0 \sum x_{2i} + \beta_1 \sum x_{1i} x_{2i} + \beta_2 \sum {x_{2i}}^2 + \beta_3 \sum x_{2i}\, x_{3i} + \beta_4 \sum x_{2i}\, x_{4i} \quad (5)$$

$$\sum y_i\, x_{3i} = \beta_0 \sum x_{3i} + \beta_1 \sum x_{1i} x_{3i} + \beta_2 \sum x_{2i}\, x_{3i} + \beta_3 \sum {x_{3i}}^2 + \beta_4 \sum x_{3i}\, x_{4i}$$

$$\sum y_i\, x_{4i} = \beta_0 \sum x_{4i} + \beta_1 \sum x_{1i} x_{4i} + \beta_2 \sum x_{2i}\, x_{4i} + \beta_3 \sum x_{3i}\, x_{4i} + \beta_4 \sum {x_{4i}}^2$$

By solving equation we can obtain coefficients $\beta_0, \beta_1, \beta_2, \beta_3 \; and \; \beta_4$ and then regression equation can be solved.

3.3.Multiple Correlation and Regression Coefficient

We presented, in the previous sub-section the Regression Model. In order to apply the proposed above model the potential number of association rules on a particular dataset, we need to evaluate whether the regression model satisfies statistical standards or not. In this section discusses how to evaluate regression model. It consists of (a). To obtain and evaluate Multiple Correlation Coefficient. (b).Hypothetical Test of Regression Coefficients.

(i). To obtain and evaluate Multiple Correlation Coefficient

To obtain the Multiple Correlation Coefficient between dependent variable y and independent variable x_1, x_2 respectively.The MCC is the proportion of variability of the dependent variable y accounted for or explained by the independent variable $x_1$and x_2 respectively.The MCC ranges from 0 to 1.If MCC is 0 that means the particular accounts for none of the variability of the dependent variable and that there is no regression prediction of y on $x_1$and x_2.If MCC is 1 that means there is perfect prediction of y by x_1 and x_2 and that 100% of the variability of y is accounted for by$x_1$and x_2.The researcher must interpret whether a particular R is high or low, depending on the use of the model and the context within which the model was developed.

$$R = \sqrt{1 - \frac{\sum (y_i - \bar{y}_i)^2}{\sum (y_i - \overline{y})^2}} \quad (6)$$

Where y is the average of the real association rule number in a data sample.Again, the closer R to 1, the smaller the residual error is, and the higher the validity of the model is.

(ii).Hypothetical test of Regression Coefficient.

In this sub-section to test whether the regression coefficients $\beta_0, \beta_1, \beta_2, \beta_3 \; and \; \beta_4$ exist in the regression model.Therefore, a pair of hypothesis including null hypothesis H_0 and alternative hypothesis H_1 is created as follows.

a). Testing the Overall Model

It is common in regression analysis to compute an F test to determine the overall significance of the model .In multiple regression, this test determines whether atleast one of the regression coefficients is to test.The hypotheses being tested in simple regression by the F test for overall significance are

$$H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$$

$$H_1: (\beta_1 = 0) \; or \; (\beta_2 = 0) \; or \; (\beta_3 = 0) \; or \; (\beta_4 = 0)$$

The test of H_0 is carried out using the F- value.

$$F = \frac{\frac{R^2}{2}}{\frac{(1-R^2)}{(n-3)}} \qquad (7)$$

Where n is the number of observation and R^2 is the multiple determination coefficient. R^2 can be calculated as follows.

$$R^2 = \frac{\sum(y_i - \bar{y_i})^2 - \sum(y_i - \widehat{y_i})^2}{\sum(y_i - \bar{y_i})^2} \qquad (8)$$

Before carrying out hypothesis test, we need to calculate the critical value $F_{\propto(2,n-3)}$ for this test, where α represents the significance level and n is the number of observations. Mostly the select value of α is 5%.

$(b). Regression \; Coefficient$ Test

In this test coefficient β_1 for support, represent x_1 and β_2 for confidence, represent x_2.We test the hypothesis for each coefficient.

Case 1 : In this test, for β_1 the hypothetical test is

$$H_0: \beta_1 = 0 \; aganist \; H_1: \beta_1 \neq 0$$

The test of H_0 is calculated by using the following statistics

$$t = \frac{\widehat{\beta_1} - \beta_1}{Se(\beta_1)}$$

Where Se(β_1) indicates a standard error of β_1 .Before carrying out hypothesis test, we need to calculate the critical value $-t_{(\propto/2,n-3)} \; and \; t_{(\propto/2,n-3)}$ for this test., where $\propto$ indicates the significance level.According to statistics, the chosen value of $\propto$ is 5%.The critical value of $-t_{(\propto/2,n-3)} \; and \; t_{(\propto/2,n-3)}$ can be found in student distribution table.If $-t_{(\propto/2,n-3)} \leq t \leq t_{(\propto/2,n-3)}$ then Ho is accepted. It states that little evidence exists of interaction between the dependent variable y and the independent variable x_1.In other cases, we accept the alternative hypothesis H_1.It states that there exists interaction between dependent variable y and the independent variable x_1.So, we keep the support variable in the regression variable x_1.

Case 2 :

In this test, for β_2 the hypothetical test is

$$H_0: \beta_2 = 0 \; aganist \; H_1: \beta_2 \neq 0$$

The test of H_0 is calculated by using the following statistics

$$t = \frac{\widehat{\beta_2} - \beta_2}{Se(\beta_2)}$$

Where Se(β_2) indicates a standard error of β_2 .Before carrying out hypothesis test, we need to calculate the critical value $-t_{(\propto/2,n-3)} \; and \; t_{(\propto/2,n-3)}$ for this test., where $\propto$ indicates the significance level. According to statistics, the chosen value of $\propto$ is 5%.The critical value of $-t_{(\propto/2,n-3)} \; and \; t_{(\propto/2,n-3)}$ can be found in student distribution table.If $-t_{(\propto/2,n-3)} \leq t \leq t_{(\propto/2,n-3)}$ then Ho is accepted. It states that little evidence exists of interaction between the dependent variable y and the independent variable x_2.In other cases, we accept the alternative hypothesis H_1.If states that there exists interaction between dependent variable y and the independent variable x_2.So, we keep the support variable in the regression variable x_2

4. Case Study

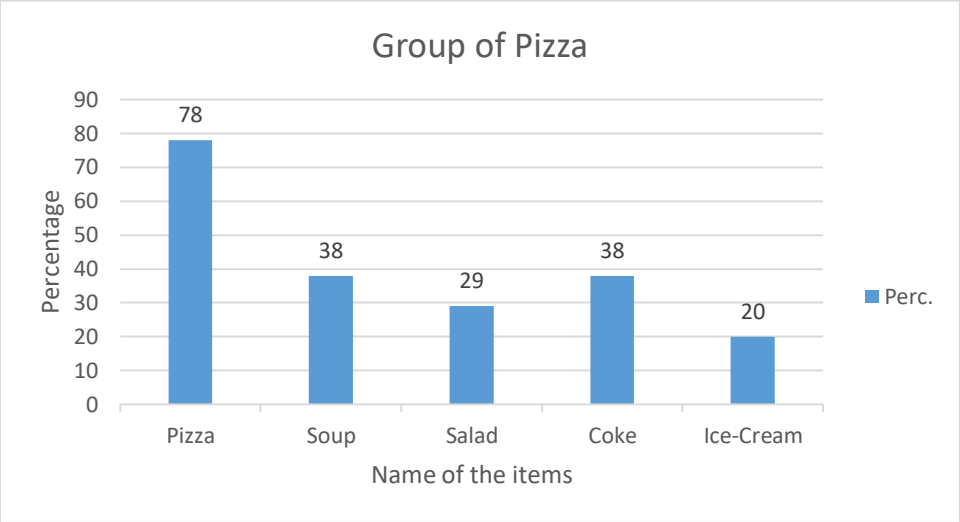
This section based on real world dataset, which is collected from a customer survey in Havmor Restaurant in Nadiad city. This case study not only shows the procedure of using our approach in a particular domain, but also demonstrates the performance of our approach in the real life situations.

The background of the case study is introduced in sub section 4.1.The procedure of generating a concrete model applied in the dataset of the case study is described step by step in subsection 4.2. Regression model selection to predict the number of association rule is given by Subsection 4.3 and using the prediction result improve association rule mining and analysis are given in subsection 4.2

4.1.Background (Overview) of the Case Study

The dataset used in this study is collected from Havmor Restaurant, Nadiad city.The survey's data was collected from different customers arriving to buy pizza with other items.The dataset created by the survey included five different items and 300 data samples.

Figure 1.Data Analysis of the Case Study



Among 300 customers, it has been observed from figure 1, that in the graph of pizza group most customer tend to buy the items like Soup,Salad,Coke and Ice-cream.Using the figure it is clear that 78% buy Pizza,38% buy Soup,29% buy Salad, 38% buy Coke and 20% buy Ice-Cream.

4.2. Concrete Model for the dataset of the Case Study

In this subsection display the procedure of creating a concrete model for this case study.We use the Pizza group to demonstrate our case study.

Step 1 : Regression Model Generation

This step illustrates a concrete regression model of the sample data of Pizza Group from dataset of survey's customers.The sample data is created by using the algorithm of Apriori[4] for association rule mining to get the real number of rules using the R-language in the conditions of different minsup and minconf.The results are shown table 1.

Table 1 : The sample data of Pizza Group

Rules	Supp	Conf	x1	x2
27	0.1	0.5	10	2.000
24	0.1	0.55	10	1.818
20	0.1	0.6	10	1.667
18	0.1	0.65	10	1.538
22	0.15	0.5	6.667	2.000
19	0.15	0.55	6.667	1.818
15	0.15	0.6	6.667	1.667
14	0.15	0.65	6.667	1.538
8	0.2	0.5	5	2.000
6	0.2	0.55	5	1.818
5	0.2	0.6	5	1.667
5	0.2	0.65	5	1.538

Table 2 : Regression Coefficients

```
Call:
lm(formula = Rules ~ Support + Confi + X1 + X2, data = h300)

Residuals:
    Min      1Q  Median      3Q     Max
-19.601  -6.847   2.432   6.508  12.257

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -217.9949   561.3591  -0.388    0.704
Support      996.8851   155.4336   6.414 2.29e-05 ***
Confi         45.0448   403.2585   0.112    0.913
X1             3.4720     0.1563  22.211 1.01e-11 ***
X2            44.8179   192.8942   0.232    0.820
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.48 on 13 degrees of freedom
Multiple R-squared:  0.9871,    Adjusted R-squared:  0.9831
F-statistic: 248.1 on 4 and 13 DF,  p-value: 3.938e-12
```

From the results of Table 1,we can build a regression model by using Equation 1 to present the relationships between the number of association rules, support and confidence .In the case study we use R language to calculate coefficients $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4$ of Equations 1.The results of coefficients $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4$ are shown in table 2.Based on the results in Table 2, a concrete model for the dataset of Pizza Group can be generated as

$$y = -217.99 + 996.88x_1 + 45.04x_2 + 3.47x_3 + 44.82x_4 \quad (9)$$

From the Equation, we can figure out that if the value of support or confidence increase, the number of association rules decrease because the regression model is a none-linear model.The number of association rules can be estimated from Equation 9,If the values of support and confidence for this particular dataset are given.

Interpretation and Summary of Table 2

a). p Value :

The summary statistics above tells us a number of things.One of the is the model p-value and the p-value of individual predictor variables.The p-values are very important because, We consider a non-linear model to be statistically significant only when both these p-values are less that the pre-determined statistical significance level, which is ideally 0.05. This is visually interpreted by the significance starts at the end of the row. The more the stars beside the variable's p-value, the more significance the variable.Here in our model, Support is the more significance variable as a independent variable.

b). t-Value

We can interpret the t-value something like this. A larger t-value indicates that it is less likely that the coefficient is not equal to zero purely by chance. So, higher the t-value, the better.Hre Support , contains higher value of t.

Pr(>|t|) or p-value is the probability that you get a t-value as high or higher than the observed value when the Null Hypothesis is true. So, if the Pr(>|t|) is low, the coefficients are significant. If the Pr(>|t|) is high, the coefficients are not significant. Here, as a independent variable Support is not significant.

In our model, we can say that the p-value of independent variable Support is less than 0.05, therefore we can say that number of association rule is effected by support factor. Over all, at the last line of the table 2, p-value of the model is below the 0.05 threshold, so we can conclude our model is indeed statistically significant.Manually the value of Multiple Correlation Coefficient, Hypothetical Test and Coefficient Test calculated.

Step 2 : Evaluation of the Regression Model.

Before predicting the number of association rules for Pizza Group,we evaluate whether the regression model satisfies the standard evaluation (refer to subsection 2.2)

- Correlation Evaluation

The value of multiple correlation coefficient R shown in Table 2 was calculated by using Equation 6.The detail calculation of R is shown in Equation 6.

$$R = \sqrt{1 - \frac{\sum(y_i - \bar{y}_i)^2}{\sum(y_i - \bar{y})^2}} = \sqrt{1 - \frac{1712.449}{130741}} = 0.9871$$

R=0.9871 means that there is a strong relationship between the number of rules and support and confidence in the dataset of Pizza Group.

- Hypothetical Test (Model Test)

Now we need to carry out for the significance of the regression model. If the model is satisfied by test, we need to carry out the second test. In The second test is to test individual regression coefficients β_1 and β_2.Genearlly,coefficient β_0 needs not to be evaluated because it is a constant.The test data is illustrated as follows

$$H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$$

$$H_1: (\beta_1 = 0)\ or\ (\beta_2 = 0)\ or\ (\beta_3 = 0)\ or\ (\beta_4 = 0)$$

Multiple Correlation Coefficient is 0.99 from Table 2, the statistic to test is

$$F = \frac{\frac{R^2}{2}}{\frac{(1-R^2)}{(n-3)}} = \frac{0.98/2}{(1-0.98)/(18-3)} = 248.1$$

The critical value for this test, corresponding to a significance level of 5% is $F_{\propto(4,n-3)} = F_{0.05,4,15} = 3.06$

Since $F > F_{\propto(4,n-3)}, H_0$ is rejected and it is concluded that at least one coefficient among β_1 and β_2 is significant. In other words, there exists the relationship between the number of association rules and either support factor or confidence factor or both of them.To consider which factor exists in the regression model, we need to carry out regression coefficient test.

- Regression Coefficient test

Coefficient test is carried out to consider which factors (support,confidence) exist in the regression model.

a).The null hypothesis to test β_1 is

$$H_0: \beta_1 = 0$$

$$H_1: \beta_1 \neq 0$$

Based on the Table 2, the statistic to test H_0 is

$$t = \frac{\widehat{\beta_1} - \beta_1}{Se(\beta_1)} = \frac{996.88 - 0}{144.43} = 6.883$$

The critical values of $t_{(\propto/2,n-3)} = t_{(0.025,15)}$ and $-t_{(\propto/2,n-3)} = -t_{(0.025,15)}$ with a significance of 0.05 are 2.13145 and -2.13145 respectively. Since $t = 6.886 > t_{(0.025,15)}$, the null hypothesis H_0 is rejected and it states that the number of association rule is effected by support.

b).The null hypothesis to test β_1 is

$$H_0: \beta_2 = 0$$

$$H_1: \beta_2 \neq 0$$

Based on the Table 2, the statistic to test H_0 is

$$t = \frac{\widehat{\beta_2} - \beta_2}{Se(\beta_2)} = \frac{45 - 0}{403.07} = 0.1116$$

The critical values of $t_{(\propto/2,n-3)} = t_{(0.025,15)}$ and $-t_{(\propto/2,n-3)} = -t_{(0.025,15)}$ with a significance of 0.05 are 2.13145 and -2.13145 respectively. Since $t = 0.1116 < t_{(0.025,15)}$, the null hypothesis H_0 is accepted and it states that the number of association rule is not-effected by confidence.

4.3.Prediction Results and Analysis for Pizza Group

In this sub-section, we use Equation 9 to predict the number of association rules and compare the real number of association rules and the predicted number of association rules. Here, association rule mining is carried out by data mining statistical language R for data analysis.

a.Prediction Results

After the regression model is generated, we can compare the predicted number of rules with the actual number of rules in the association rules in the condition of given support coefficient and confidence coefficient. It means that we replace the pair of value of x_1, x_2, x_3 and x_4 from Table1 into Equation 9 to estimate the number association rules. The result is presented in Table 3.

From the residuals and rate of error in Table 3, we can see that the difference between the actual numbers of rules and the predicted numbers of rule in 18 different cases is acceptable and the average error of 14.71%.

It means the performance of the prediction model is relatively good. Even if the forecast values and actual values are biased, the error is acceptable.

b.Analysis of the Case Study.

The number of estimated rules with the optimal values of support and confidence is estimated from the regression model. After we confirm the values of support x_1 and confidence x_2, the potential number of association rules is obtained for Group of Pizza by using Equation 9. By chance if the minimum support and confidence is 0.015 and 0.75 respectively, the number of predicted association rules for Group of Pizza coming from Equation 9 is

$$y = -217.99 + 996.88x0.15 + 45.00x0.75 + 3.47x\frac{1}{0.015} + 44.80x\frac{1}{0.75}$$

5. Conclusion

This paper proposed a new approach to improve the association rule mining process through the prediction of the potential number of association rules on datasets. The proposed approach is unique and versatile because the design of the regression model to the approach is vast, therefore it can applied into broad domains for predicting the potential number association rules. By using the statistical tests and testing of hypothesis verify the regression model using the significance of the support and confidence in the model.Also, the case study verify the regression model in real life dataset.

The future work will pay attention to develop the new algorithm and compare with candidate generation algorithm specially apriori algorithm.

References

[1] Agrawal, R., Imieliński, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. *ACM SIGMOD Record*,*22*(2), 207-216.

[2] Shankar, S., Purusothaman T.(2009).Utility sentient frequent itemset and association rule mining literature survey and comparative study.International Journal of Soft Computing Applications,vol. 10, no. 4, pp. 81-95.

[3] Houstma, M., Swami A., (1993). Set-oriented mining for association rules in relational databases.Research

Report RJ9567, IBM Almaden Research Center, San Jose, California.

[4] Agrawal, R., Srikant, R.(1994). Fast algorithms for mining association rules. The 20th International Conference on Very Large Databases, pp. 487-499.

[5] Park. J, Chen. M., & Yu. P. An effective has-based algorithm for mining association rules.ACM SIGMOD International Conference on Management of Data, pp. 175-186.

[6] Zaki. M., & Hsiao. C. Charm: an efficient algorithm for closed association rule mining. The 2nd SIAM International Conference on Data Mining. Pp. 457-473.

[7] Han. J., Pei. J., & Yin. Y. Mining frequent patterns without candidate generation. The 2000 ACMM SIGMOD International Conference on Management of Data, pp. 1-12.

[8] Das, A., Ng, W., & Woon, Y.(2001). Rapid association rule mining. The 10th International lnformation and Knowledge Management, 474-481.

[9] Wang, J., Han, J., & Pei, J. Closet+ : Searching for the best strategies for mining frequent closed itemsets. The ninth ACM SIGKOD International Conference on Knowledge Discovery and Data Mining, 236-245.

[10] Shi, L., Jing, X., Xie, X., Yan, J.: Association Rules Applied to Credit Evaluation. In : 2010 International Conference on Computational Intelligence and Software Engineering (CiSE), pp. 1-4 (2010)

[11] He, Z., Huang, H. K., & Tian, S. F. An approach to finding optimized correlated association rules. Chinese Journal of Computers., vol. 29, no. 6, 906-913.

[12] Ye. D. Q., Zhaaou. S. L.(2008). Correlation technique research of association rule based on linear regression. Journal of Computer Research and Development, vol. 45, 291-294.

[13] Aggarwal, C.C., & Yu, P.S.(1998). A new framework for itemset generation. Book A new framework for itemset generation. Series A new framework for itemset generation, ed., Editor ed., ACM.

[14] Han, J., Kamber, M. (2011). Data mining: concepts and techniques. Third edition, Elsevier Inc.