Regression Model Method for Analyze the Association Rules using Major Parameters

G. G. Shah, H. N. Patel

Abstract: ***Using the data mining user can extract the information. Frequent itemsets is one of the popular task in data mining. Association Rule Analysis is the task of discovering association rules that occur frequently in a given large data set.The task is to find certain relationships among a set of itemsets in the database. There are two fundamental parameter(measurement) is Support and Confidence.Traditional association rule mining techniques employ predefined support and confidence values. But, it's observed that specifying minimum support value of the minded rules in advance often leads to either too many or too few rules, which negatively impacts the performance of the overall system.This paper proposes a non-linear regression model using support, confidence and association rules. To predict the number of rules under the given explanatory variables say parameters. Use the R language for the Rules generations and also uses significance test to verify regression coefficients. Using the coefficient test and F-test verify the model.***

Keywords: ***Association Rules, Regression, Regression Coefficients, Multiple Correlation, F-test***

I. INTRODUCTION

Data Mining is a process of extraction of useful information from the large database and pattern from the huge data [1]. Data Mining is a logical process for to search large amount of data in order to find useful data. The aim of this technique is to find patterns that were previously unknown. There are various techniques like Classification, Clustering, Regression, Artificial Intelligence, Neural Networks, Association Rules, Decision Tree, Generic Algorithm, Nearest Neighbor method etc., are used for knowledge discovery from database.

Association Rule is one of the most popular techniques and an important research issue in the area of Data Mining [2].Association Rule is usually to find frequent items from the large data.It is helpful for findings helps businesses to make certain decisions like shopping behavior analysis in the super mall, catalogue design, cross marketing, fraud detections, finance, telecommunication etc. However the number of Association Rules for a given dataset is generally very large.

To improve the efficiency of the Rules various researchers have developed the frameworks and algorithms. These include Agrawal et al.[3] have developed AIS, SETM, Apriori, Apri-oriTid to discover significant association rules between items in large data.

Revised Manuscript Received on May 20, 2020.

G. G. Shah, Faculty of Business Administration, Dharmsinh Desai University, Nadiad, India. E-mail: gopalgshah16@yahoo.com

Dr. H. N. Patel, Department of Computer Science, Dr. Babasaheb Ambedker Open University, Ahmedabad, India. E-mail: himanshu.patel@baou.edu.in

Number of algorithms are developed and compare with above define algorithms and among them areDHP(Park, Chen et al., 1995), CHARM (Zaki and Hsiao, 1999), FP Growth (Han, Pei et al.,, 2000), RARM (Das, Ng et al., 2001), Closet+ (Wang, Han et al., 2003) etc.

In the last two decades, little work has been done on how to choose both appropriate thresholds of support and confidence for the mining algorithms before a real mining process. It is observed the lack of approaches or techniques or mechanism for thresholds selection of support and confidence and it often leads to either too many or too few rules after completing a mining process. Also it can define or lead to either excessive computation time or poor results.

Therefore, current association rule mining technique or approaches are facing following challenges

a). In the Particular Mining algorithm selection of the thresholds of Support and Confidence. b). Extension of the particular approach using the Matrix Application.c). Reduce the time needed for large dataset. In this research paper proposed approach consists of

(i). Generated Regression Model consists of Support, Confidence and Association Rules as an explanatory variable and Response variable.(ii). The Model is applicable for a particular domain.

Therefore, the proposed approach can be applied to different domains with different types of datasets to select thresholds of support and confidence in an association rule mining algorithm. Also, it demonstrates a case study to show the effective performance of the proposed approach on a real-world dataset in support and confidence selection.

The rest of this paper includes as: Section 2 introduces the Regression Model Approach, Section 3 contains the Matrix application for the Regression, Section 4 contains the Regression Statistics, and Section 5 concludes with Conclusions.

II. REGRESSION MODEL APPROACH

In this section, the proposed approach is based on Regression Approach. A regression model is generated in a general level by number of association rules, and support coefficient and confidence coefficient.Naturally in principle the number of association rules in a dataset depends upon the support coefficient and confidence coefficient.

The formula of support and confidence for the form $A \Longrightarrow B$ can be formally define as follow.

$$Support\ (A \Longrightarrow B) = P(A \cup B)$$
$$Confidence\ (A \Longrightarrow B) = P(A \backslash B)$$

Note:

i.The value of the support and confidence are lies in between 0 and 1.

ii. Support describe the significance of the rule and Confidence describe the strengths of the rule.

The Regression Model to be define using Association Rules, Support and Confidence are as follow

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 \frac{1}{x_1} + \beta_4 \frac{1}{x_2} \quad (2.1)$$

Where y is the number of association rules, x_1 and $\frac{1}{x_1}$ represents the support variable and its reciprocal and similarly x_2 and $\frac{1}{x_2}$ represents the confidence variable and its reciprocal respectively. $\beta_0, \beta_1, \beta_2, \beta_3 and \beta_4$ are coefficients of the regression model. The coefficient $\beta_1 and \beta_3$ measures the partial effect of $x_1 and x_3 ony$. Similarly, The coefficient $\beta_2 and \beta_4$ measures the partial effect of $x_2 and x_4 ony$,the coefficient β_0 is rarely explained and evaluated in a regression model.It is known as intercept and it includes all other factors except support and confidence. Using the equation (1) we can see that the number of association rules depends on values of support and confidence, also β_1 and β_2 in a particular domain.After taking the simple variable transformation the equation (1) can be define as

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 \quad (2.2)$$

III. MATRIX APPLICATION FOR THE REGRESSION

Given the actual data values, we may write the model for n number of association rules of the above type, where i^{th} association is

$$Y_i = \sum_{j=1}^{k} \beta_j X_j j = 1,2 \dots, kandi = 1,2, \dots, n$$

Which can be written in matrix notation as

$$\underline{Y} = \underline{\beta} X + \epsilon$$

Where

$Y = (y_1, y_2, \dots . y_n)^T, \epsilon = (\epsilon_1, \epsilon_2, \dots ., \epsilon_n)^T, \beta = (\beta_0, \beta_1, \dots \dots, \beta_n)^T$ And

$$A = \begin{bmatrix} X_{11} & X_{12} & \dots & \dots & X_{1n} \\ X_{21} & X_{22} & \dots & \dots & X_{2n} \\ \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots \\ X_{n1} & X_{n2} & \dots & \dots & X_{nn} \end{bmatrix}$$

We can assume that $E(\epsilon) = 0$ and $V(\epsilon) = \sigma^2 I$since if this were not so, we could simply absorb the nonzero expectation for the error into the mean μ to get a zero expectation.

A. ESTIMATING β

We have to obtain estimates of $\beta_1, \beta_2, \dots \dots, \beta_k$.The regression$Y = X\beta + \epsilon$, partitions, the response into a systematic component $X\beta$ and a random componentϵ. The problem is to find β so that $X\beta$ is as close to the response variable X is as close to the response variable Y as possible. The cure method for estimate$\hat{\beta}$, to be define in the following geometrical figure.

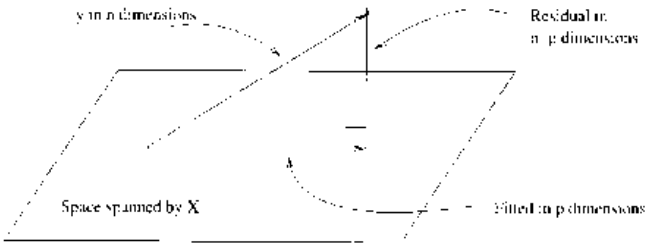

Figure 1 Regression Coefficient

In the model space, $\hat{\beta}$ is the best method to estimateβ. The response variable Y to be predicted by the model $\hat{Y} = X\hat{\beta}$. The difference between the actual response variable and the predicted response is denoted by $\hat{\epsilon}$ and it is known as residuals.

The conceptual purpose of the generalized model is to represent, an accurately as possible. The response variable y is n-dimensional, in terms of very simple way the model, which is k-dimensional.

So, if model is much successful, the structure in the data should be define by k- dimensional and from the above figure it is clear that the random variation in the residuals are lie in (n-k) dimensional space.

B. LEAST SQUARES ESTIMATION

We have to obtain estimates of $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4$ from a non-geometrical point of view. We might define the best estimate β as the one which minimize the sum of the squared errors. The least squares estimate of β, called $\hat{\beta}$.

$$\underline{Y} = X\hat{\beta} + \underline{\epsilon} \quad (3.2.1)$$

$$\underline{\epsilon} = \underline{Y} - X\hat{\beta}$$

Considering sum of squares of errors, that is$\sum \epsilon_i{}^2 = \epsilon^T \epsilon = (Y - X\hat{\beta})^T (Y - X\hat{\beta}) = Y^T Y - 2\hat{\beta}^T X^T Y + \hat{\beta}^T + \hat{\beta}^T (X^T X)\hat{\beta}$ (3.2.2)

Differentiating with respect to $\hat{\beta}$ and setting to zero, we find that $\hat{\beta}$ satisfies:

$$\frac{\partial(\epsilon^T \epsilon)}{\partial \hat{\beta}} = -2X^T Y + 2(X^T X)\hat{\beta} = 0$$

We have,

$$X^T Y = (X^T X)\hat{\beta} \quad (3.2.3)$$

These are called the normal equation. We have to solve them. Let $X^T X$ be positive definite matrix, that is $|X^T X| \neq 0$.In this case pre-multiplying both side of X^T we get

$$\widehat{\beta} = (X^T X)^{-1} X^T Y \quad (3.2.4)$$

And if $|X^T X| = 0$ then the solution of (3.2.2) does not exists, but by taking generalized inverse $(X^T X)^{-1}$of $(X^T X)$ then the generalized solution is

$$\widehat{\beta} = (X^T X)^{-1} X^T Y \quad (3.2.5)$$

IV. REGRESSION STATISTICS

In the given regression model generation Response Variable is defined by $\hat{y}$. The Explanatory Variable is define by $X_i, i = 1,2,3,4$. Using the dataset, the obtain Rules are denoted by Response Variable and Support and Confidence are define by Explanatory Variable. Support and Confidence are in the inverse relation with Association Rules, so generate the non-linear regression model. In the above section define the formula for the Regression Coefficient. By

using matrix calculation obtain Regression Coefficients.

Table 1. Association Rules

Rules	Supp	Conf	x_1	x_2
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
17	0.1	0.7	10	1.429
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
13	0.15	0.7	6.667	1.429
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
5	0.2	0.7	5	1.429

Using the above Table 1, to obtain the Regression Coefficients. Convert the specific columns into Response and Explanatory Variable. Rules are define as the Response and Supp, Conf,x_1 and x_2 are define as Explanatory Variables. They are convert into Matrix form and define in the following steps:

Step 1. Define Rules as Response Variable by y (shown below)

Step 2. Define Support, Confidence,x_1, x_2 as Explanatory Variable by X. (shown below)

$$y = \begin{bmatrix} 27 \\ 24 \\ 20 \\ 18 \\ 17 \\ 22 \\ 19 \\ 15 \\ 14 \\ 13 \\ 8 \\ 6 \\ 5 \\ 5 \\ 5 \end{bmatrix} \quad X = \begin{bmatrix} 0.1 & 0.5 & 10 & 2.000 \\ 0.1 & 0.55 & 10 & 1.818 \\ 0.1 & 0.6 & 10 & 1.667 \\ 0.1 & 0.65 & 10 & 1.538 \\ 0.1 & 0.7 & 10 & 1.429 \\ 0.15 & 0.5 & 6.667 & 2.000 \\ 0.15 & 0.55 & 6.667 & 1.818 \\ 0.15 & 0.6 & 6.667 & 1.667 \\ 0.15 & 0.65 & 6.667 & 1.538 \\ 0.15 & 0.7 & 6.667 & 1.429 \\ 0.2 & 0.5 & 5 & 2.000 \\ 0.2 & 0.55 & 5 & 1.818 \\ 0.2 & 0.6 & 5 & 1.667 \\ 0.2 & 0.65 & 5 & 1.538 \\ 0.2 & 0.7 & 5 & 1.429 \end{bmatrix}$$

Step 3. Obtain Transpose of the given matrix X.

Step 4. Find $(X'X)$

Step 5. Find Inverse of $(X'X)$ say $(X'X)^{-1}$

Step 6. Find $X'Y$

Step 7. Find $(X'X)^{-1}X'Y$ and denoted as $\hat{\beta}$.

$$\hat{\beta} = \begin{bmatrix} -340.76 \\ 58.981 \\ -3.735 \\ 33.859 \end{bmatrix}$$

Interpretation:

In the above matrix the Coefficient define as with the value -340.76, 58.981, -3.735 and 33.859 respectively. The coefficient x_1 is define for Support and x_3 is the inverse of Support. The sign of x_1 and x_3 are negative, that means if the independent variable increases, then the value of the dependent variable tends to decreases.

Similarly x_2 is define for Confidence and x_4 is the inverse of the Confidence. Here, x_2 and x_4 are both positive. The value of x_2 are increase then the value of the independent variable increase.

It means in the above the data, the Response variable say number of rules and Explanatory variable say Support are inverse related with each other. If the Support increase the number of rules are decrease but negative sign indicatesthat the both variable are opposite direction. Here, the obtain Regression Model is

$$y = -340.76\,X_1 + 58.981X_2 - 3.735X_3 + 33.859X_4$$

A. REGRESSION MODEL TEST

Before predicting the number of association rules, the define regression model satisfies the following standard evaluation or not.

i. Multiple Correlation Coefficient evaluation

The detail calculation of $R$is shown as follows.

$$R = \sqrt{1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y}_i)^2}} = 0.9848$$

It means that there is a strong relationship between the number of rules and support and confidence in the given dataset.

B. REGRESSION MODEL TEST

Now we have to test for the significance of the regression model. If the model is satisfied by test, we need to carry out the second test. In other cases, the data sample is considered by increasing the number of observations. The second test is to test individual regression coefficients $\beta_1, \beta_2, \beta_3 and \beta_4$.here, β_0 define as the constant so it is not need to be evaluated. The null hypothesis for the regression model is

$$H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0\ aganist$$
$$H_1: \beta_1 \neq \beta_2 \neq \beta_3 \neq \beta_4 \neq 0$$
$$R^2 = 0.9699, \text{ and}$$
$$F = \frac{R^2/_k}{(1-R^2)/_{(n-k-1)}} = \frac{0.9699/_4}{(1-0.9699)/(15-4-1)} = 80.55$$

The critical value for this test, corresponding to a significance level of 5% is $F_{\alpha(k-1,n-k-1)} = F_{0.05,3,10} = 3.71$.Therefore, $F_{cal} > F_{\alpha(2,n-3)}$. Null Hypothesis H_0 is rejected and it is concluded that at least one coefficient among $\beta_1, \beta_2, \beta_3 and \beta_4$ is significant. In other words, there exists the relationship between the number of association rules and either support factor, confidence factor or both of support and confidence.

C. REGRESSION COEFFICIENT TEST

The aim of the Regression Coefficient Test is to observe which factors (support, confidence) exist in the regression model.

(i) The null hypothesis to test β_1 is

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

$$t = \frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)}$$

Where standard error of the regression coefficients are obtained from the variance-covariance matrix say $(X^T X)^{-1} and \hat{\sigma}$.Similarly $\hat{\sigma}$ is known as standard error of the model and it is obtained as

$$\hat{\sigma} = \sqrt{\frac{\sum {\epsilon_i}^2}{n-k-1}}$$

$$t = \frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)} = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{d_{11}}\hat{\sigma}} = \frac{-340.0744 - 0}{\sqrt{1120.865}\, x\, 1.511307}$$
$$= -6.7211$$

The critical values of $t_{(0.05,15-4-1)} = t_{(0.05,10)}$ with a significance of 0.05 are 2.228 respectively. Since $t = 6.7211 > t_{(0.05,10)}$, the null hypothesis H_0 is rejected and it states that the number of association rule is effected by support.

ii) The null hypothesis to test β_2 is

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

$$t = \frac{\hat{\beta}_2 - \beta_2}{SE(\hat{\beta}_2)} = \frac{\hat{\beta}_2 - \beta_2}{\sqrt{d_{22}}\hat{\sigma}} = \frac{62.00854 - 0}{\sqrt{1342.353}\, x\, 1.511307}$$
$$= 1.1199$$

The critical values of $t_{(0.05,15-4-1)} = t_{(0.05,10)}$ with a significance of 0.05 are 2.228 respectively. Since $t = 1.1199 < t_{(0.05,10)}$, the null hypothesis H_0 is accepted and it states that the number of association rule is not effected by confidence.

V. CONCLUSIONS

Using the generated non-linear Regression model it is clear that Support and Confidence are the fundamental parameter for the Association Rules. In the proposed a new approach the association rule mining process through the prediction of the potential number of association rules on dataset. In our approach is designed and evaluated by Multiple Correlation, Regression Model and Regression Coefficient tests in terms of significance of Support and Confidence in the model.To consider which factor exists in the define Regression Model, it is to be define by the coefficient test. In the future study, our approach to datasets in other general areas like Transportation, Telecommunication, Mutual Funds, and Finance and so on.

REFERENCES

1. B. Ramageri, "DATA MINING TECHNIQUE AND APPLICATIONS," Indian Journal of Computer Science and Engineering, 2010.
2. D. T. Le, F. Ren, and M. Zhang, "A Regression-Based Approach for Improving the Association Rule Mining through Predicting the Number of Rules on General Datasets," Lecture Notes in Computer Science PRICAI 2012: Trends in Artificial Intelligence, pp. 229–240, 2012.
3. Agrawal, R., Imieliński, T. and Swami, A. (1993). Mining association rules between sets of items in large databases. ACM SIGMOD Record, 22(2), pp.207-216.

AUTHORS PROFILE

G. G Shah, Assistant Professor at Dharmsinh Desai University. Having Experience of 21 at U.G Level.. Published 4 research articles at National and International journals. The main interest field is Data Science

Dr.H.N Patel, Assistant Professor at Dr. BabasahebAmbedkar Open University. Having experience of 15 years.4 Paper published in International Journal and 5 Paper published in National Journal with ISSN/ISBN no. and Total 11 Paper presented in Conference/Seminar/ Workshop. He is having Lifetime membership of CSI and was member of ACM.