IJCSI International Journal of Computer Science Issues, Volume 14, Issue 4, July 2017
ISSN (Print): 1694-0814 | ISSN (Online): 1694-0784
www.IJCSI.org https://doi.org/10.20943/01201704.2028

A new Algorithm for Mining Association Rules Based on Hypothesis Test

Medhat H A Awadalla[1,2] and Sara G El-Far[2]
[1] Electrical and Computer Engineering Department, SQU
Muscat, Oman

[2] Communication, Electronics and Computers Department,
Helwan University, Egypt

Abstract

Most association rule mining approaches rely on predetermined support and confidence values to find certain relationships among database itemsets. However, specifying minimum support and confidence values of the mined rules in advance often leads to either too many or too few rules, which negatively affects the performance of the overall system. To address this issue, this paper presents a new algorithm based on null hypothesis to find non-coincidental relations among different itemsets in large databases without a prior defined threshold values. Intensive simulated experiments have been performed on different databases to confirm the validity of the suggested algorithm. The obtained results show that there is a significant improvement in the system performance in terms of the number of frequent items used, the number of generated rules, and the run time.

Keywords: *Association Rule Mining, Null Hypothesis, Chi-Square Test, Non-Coincidental Rules.*

1. Introduction

Data mining focuses on extracting convenient information and discovering hidden relationships that exist in large databases [1]. It is considered as one of the most well-known approaches for data exploration, helping users to get unpredicted rules from large database itemsets.

The association rule mining used to find relationships among the frequent itemsets that are not related in different databases and it mainly relies on the support and confidence values. The frequent itemsets are known as the itemsets that their support value is greater than or equal the minimum threshold support value, and frequent rules are defined as the rules that their confidence values are greater than or equal the minimum threshold confidence value.

Association Rule Mining is all about finding all rules whose support and confidence exceed the threshold, minimum support and minimum confidence values. The procedure of association rule mining contains two steps. Firstly, to find all itemsets with suitable support values and secondly, to produce association rules by combining these frequent (or) large itemsets together [2].

In the traditional association rule mining [3], the minimum support threshold and minimum confidence threshold values are assumed to be available for mining frequent itemsets, which is difficult to be set without specific knowledge. If the value of the support threshold were set too high, a small number only of the rules would be generated or even no rules to conclude. In a such case, a smaller threshold value should be predicted to do the mining again, that may or may not give a significant result. If the threshold was set too small, numerous results would be produced for the users, these huge results would require not only very long time for computation but also for screening these rules. For all these issues addressed above, it is required to develop a new algorithm to produce the minimum support and confidence values based on the itemsets in the addressed databases. Therefore, we developed an algorithm to calculate each itemset support value and the minimum support value. Aggregation functions such as standard deviation, simple mean, mean square error are used to find the minimum support value for each itemset in the addressed database [4]. Similarity or confidence pruning is another constraint that could be produced if association rule mining is addressed without the value support threshold [5].

Originally, the mining of the association rules was suggested for analyzing the data itemsets of the market basket. As stated in [6], if there is a set of transactions D, the role of the association rule mining is to discover the association rules that have support value higher than the minimum well-defined support threshold and confidence value higher than the minimum well-defined confidence threshold. The following is a proper way to describe the association rule mining for database transaction. Let $I = \{i_1, i_2. \ . \ . \ i_m\}$ is the space of itemsets. A set (X, I) of items is entitled an itemset. A transaction $t = (TID, X)$, where *TID* has a distinctive transaction ID and X is an itemset. A transaction database D is a group of transactions. The *count* of an itemset X in D, indicated by count(X), is the number of transactions in D covering X. The support of an itemset X in D, denoted by supp(X), is

the proportion of transactions in D that contains X. The rule X → Y holds in the transaction set D with confidence c, where c = conf (X → Y) and conf (X → Y) = supp (X υ Y) / supp(X). The discovery of positive association rules in frequently occurring itemsets is the dominant theme of the traditional association mining [7].

The association rule algorithms can be categorized into two brands, breadth-first search (BFS) or candidate-generation-and-test approach such as Apriori algorithm, and depth-first search (DFS) or pattern-growth approach [8-13]. Each one of the algorithms has its strategy to traverse the search space and find the support values of the itemsets. Furthermore, each algorithm may apply specific optimization methods for reducing the required run time. Apriori algorithm is considered as one of the most popular algorithms of this type.

The rest of this paper is organized as follows. In section 2, the related background is presented comprising a brief introduction for the Chi-Square Test. In section 3, the developed procedure for testing the hypothesis and calculating Chi-Square in presented. In section 4, the proposed hypothesis testing based algorithm for mining association rules is presented. In section 5, the conducted experiments and discussions are demonstrated. The conclusion of the paper is given in section 6.

2. Background

Since this paper presents new algorithms based Null Hypothesis and Chi-Square testing approaches to find non-coincidental relations among different itemsets in large databases without a prior defined threshold, therefore, in the following sub-sections we give a brief explanation for them.

2.1. Hypothesis Testing

Hypothesis testing [14] is applying the statistics to find the probability that a particular hypothesis is true. There are two kinds of hypothesis testing, Null and Alternative.
The first hypothesis test is the null hypothesis, which is a statistical hypothesis test for possible rejection under the assumption that it is true; generally, those observations come out by the chance. The second type is the alternative hypothesis test. In this kind of test, the observations are mainly due to a real effect in addition to some amount of chance variation superposed. The procedure for hypothesis contains four steps.
1. The first step is to formulate the null hypothesis H0 (observations by accident) and the alternative hypothesis H1 (observations by both real actual and chance variation).
2. The second step is to identify a statistical test to evaluate the truth of the null hypothesis.
3. The third step is to compute the P-value, the probability that the null hypothesis were true. If the P-value is small, then the evidence against the null hypothesis grows strong.
4. The fourth step is the comparison of P-value to a certain acceptable significance value, usually named α, if the value of P is less than or equal the value of α, the observed effect is statistically substantial, and in this case, the null hypothesis is rejected, and hence the alternative hypothesis is true. On the other hand, if the P value is greater than α, then the null hypothesis is not rejected and the difference is "not statistically important". There are some methods to compute 95% confidence interval for any calculated statistic [15]. The value of 95% itself does not have any special meaning however, by convention the intervals of confidence are usually assessed for 95%. Theoretically, the intervals of confidence can be determined for any degree of confidence. If more confidence level is needed, the intervals will be broader. Else, the intervals will be narrower.

2.2. Chi-Square Test

Chi-Square is defined as a statistical test that is normally used for comparing the observed data with the real data that expected to obtain based on a specific hypothesis [16]. Null hypothesis is always implemented by the Chi-Square test, mentioning that, the difference between the expected and observed result is not significant and the Chi-Square (Chi^2) is given in equation 1:

$$Chi^2 = (o-e)^2/e \quad (1)$$

It is the deviation, difference squared between observed data (o) and the expected data (e) data divided by the expected data (e) in all different possible data itemsets.

3. The developed Procedure for testing the hypothesis and calculating Chi-Square

The procedure for testing the hypothesis and calculating Chi-Square is outlined as follows.
1. First of all, the data should be collected through conducting the proper experiments, outline the state the hypothesis being tested and the expected results.
2. The predicted number for each observational class should be determined. Percentages are preferred to use, not numbers. If the expected value in any category is less than 5% [16], the Chi-Square should not be computed.

3. Chi-Square should be computed via equation 1. All calculations should be completed.

4. The distribution table of the Chi-Square is applied to find the significance of the value.
   a. The degree of freedom, df, should be evaluated and its value should be positioned in the proper column.
   b. Determine the nearest value in the degree of freedom row that is the closest to the calculated Chi-Square one.
   c. The value of P should be computed.
5. Conclude the achieved hypothesis.
   a. If the calculated value of *p* exceeds 0.05, your hypothesis is accepted. In this case, the difference between observed and the expected data is small enough that the chance alone accounts for it.
   b. If the calculated value of p is below 0.05, your hypothesis should be rejected. It means, not only the chance but also other factors are contributing for the deviation.

4. Proposed Hypothesis Testing based algorithm for Mining Association Rules

To search for infrequent itemsets, two circumstances should be taken into account. The first is the occurrence of itemsets because of any non-random process that generates them or by coincidence. The itemsets that have a low support value but their confidence values are high are interesting and may be taking by chance and should be considered as noise.

Clearly, it makes sense only to consider the candidate itemsets that appear together more often than coincidence. Null Hypothesis test presents the relation among different items, and distinguishes if the events happen together due to coincidence. Based on the null hypothesis test and Chi-Square test explained before, a new algorithm is proposed and presented in this section.

The studied database is converted into the corresponding bitmap, by reassigning elements into the map to zeros and ones in a bitmap array as we developed in [4]. Every two itemsets are combined together to produce contingency table for every combination.

Applying Chi-Square test on the produced contingency tables will determine for every case both Chi-Square value and P-value. For each combined itemsets with P-Value < significance level, the null hypothesis is assured to be true. For instance, if the significance level (α) is 0.01, it means that these itemsets are 99% related and decline that they occur together as a coincidence.

The significance level should be set at a suitable value to be matched with the needed level of certainty. If the null hypothesis (H0) is assumed on a pair of itemsets, it presents that the two itemsets are uncorrelated, and there is no relation between them. If the alternative hypothesis (H1) is assumed on two items, these items will be related to each other. Then apply Chi-Square statistical test to calculate corresponding P-value.

If the P value is less than the threshold, the null hypothesis is excluded and the difference is "statistically significant" and the itemsets are dependent. However, if the P value is greater than the threshold, the null hypothesis cannot be rejected and the difference is "not statistically significant" and itemsets are independent and may be occurring together due to coincidence. The proposed hypothesis test flowchart is illustrated in figure 1 and the pseudo code is as follows.

Input: Transaction database *D is given*
Output: Generate the non-coincidental frequent rules
Set significance level α
for all data itemsets I ∈ D
{
for (k = 1; k ≤ no of itemsets; k++) do
{
for (count =k+1, count <= no of itemsets; count ++) do{
Combine itemset k with itemset k+count;
Run Chi-Square statistical test;
Calculate P-value for the pair;
if p ≤ α then itemsets are related;
else items are independent;
}
}
}

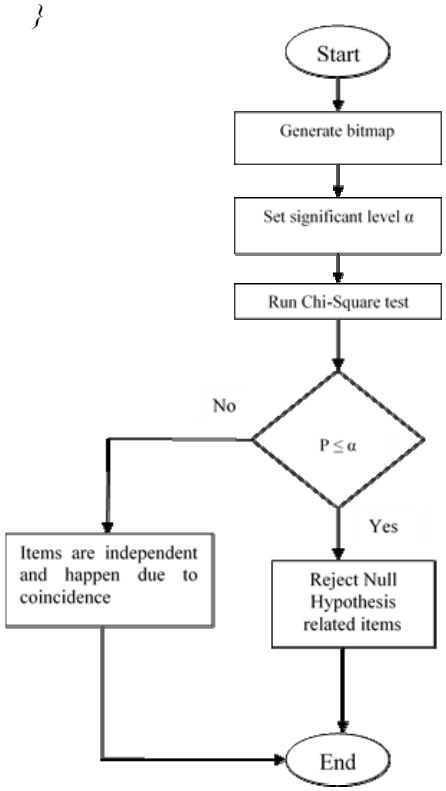

Fig.1. The flowchart of the proposed hypothesis test

5. Experimental and Discussions

Intensive simulated experiments have been conducted using 100 transactions of IBM dataset [17] to validate the proposed algorithm. Data first converted into bitmap, table 1 shows the contingency table to be used during the statistical test. The Chi-Square test was applied with the help of a Statistical Software "Minitab – V.15" shown in figure 2. The program is fed with data extracted from contingency tables to be used to apply the Chi-Square Statistical test, and to calculate the Chi-Square value and their corresponding P-values as well. Snapshot of running the Chi-Square test through Minitab is depicted in figure 3, selecting data to run Chi-Square test is illustrated in figure 4, and then the achieved results for both Chi-Square and P-values are shown in figure 5.

Table 1. Contingency Table for Items 9 and 41

Item 41	Item 9		Sum
	0	1	
0	55	5	60
1	39	1	40
sum	94	6	100

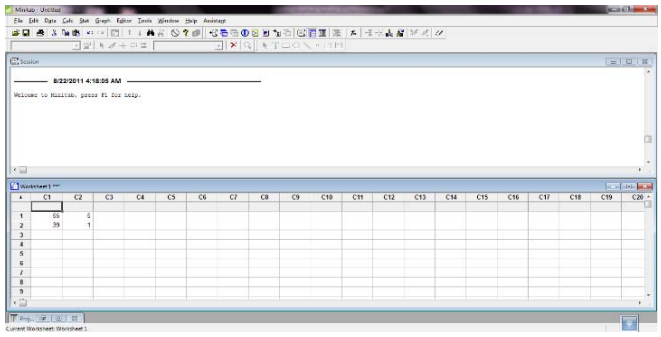

Fig. 2. Statistical Software "Minitab – V.15

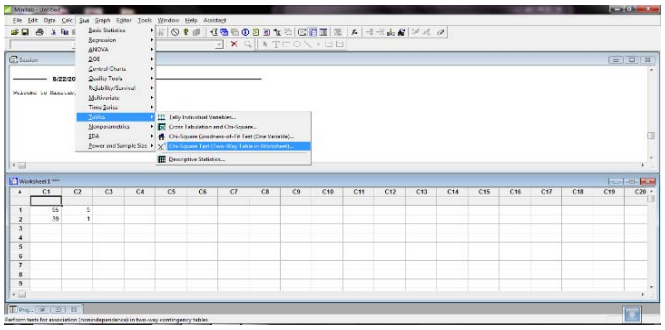

Fig. 3. Running the Chi-Square Test through Minitab

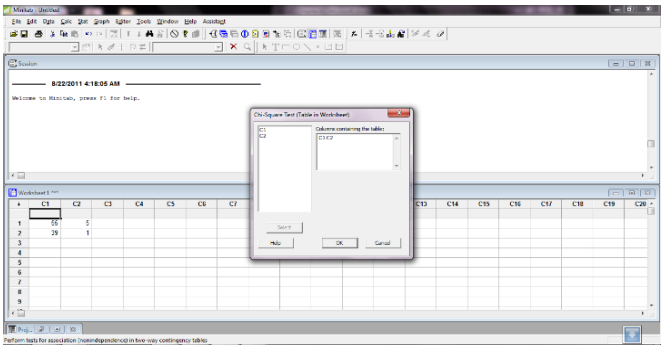

Fig. 4. Data Selection

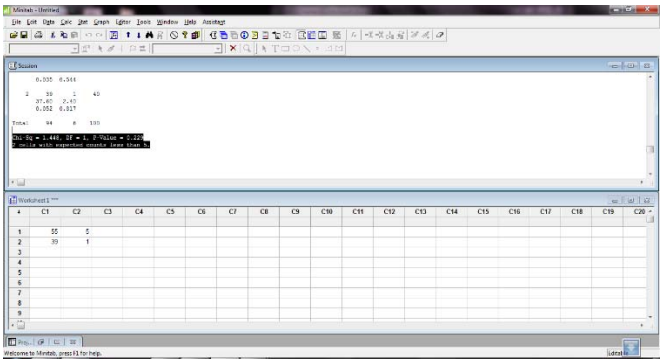

Fig. 5. The achieved results for both Chi-Square and P-values

Table 2 shows the pruning of the unsuccessful itemsets that occur due to coincidence. The significance level α is set to 0.05 (5%) to give certainty of 95% to the produced rules. Only two successfully rules are survived whose P-Value ≤ α.

These successful candidates are:

- Item 39 >> Item 48 with support=67.
- Item 38 >> Item 47 with support = 2.

These items are not appeared together on the transactions due to coincidence; they are related to each other.

Again, the value of the significance level cannot be set to too small or too large values; if significance level is chosen to be too small, a very small number of rules or no rules would be produced, and if the significance level chosen to be too large, a very large number of redundant and weak rules would be produced. If the significance level increases, the number of produced rules increases and redundant and weaker rules are obtained as shown figure 6. Both figure 6 and table 3 show that the number of rules extremely increases as the significance level value increases. For instant, if the significance level is raised from 0.05 to 0.2, some rules as (item 31 >> item 36) with support value of zero, which gives no information as item

31 appears twice and item 36 appears 11 times in one-hundred transaction. If the significance level is 0.4, rules such as (item 32 >> item 39) with support of 10, while item 32 appears 14 times and item 39 appears 85 times in the whole used transactions.

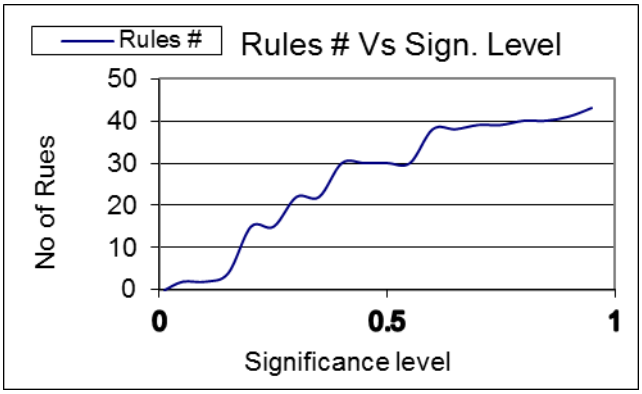

Figure 6. Number of Rules versus the Significance level for sample-1

The proposed algorithm is applied again using another 100 transaction of database IBM dataset and the same procedure is repeated. Table 4 shows the pruning of the unsuccessful itemsets that occur due to coincidence. The significance level α has been set to 0.05 (5%) to give certainty of 95% to the produced rules. Three rules successfully survived whose P-Value ≤ α. These successful candidates are:

- (Item 38 >> Item 41) with support=19,
- (Item 38 >> Item 48) with support=18 and;
- (Item 41 >> Item 48) with support = 59.

Again, as significance level increase, the number of generated rules increases as well and redundant and weaker rules are obtained as shown in table 5 and corresponding graph in figure 7. Both figure 7 and table 5 show that the number of rules increases as the significance level value increased and weak rules to be produced.

Again, if the significance level is raised from 0.2 to 0.05, some rules like (item 9 >> item 48) with support of 3, which produce a weak rule as item 9 appears 6 times and item 48 appears 76 times of the 100 transaction. When the significance level is 0.4, rules as (item 9 >> item 41) with support of 1, while item 9 appears 6 times and item 41 appears 40 times in the transaction.

Table 2. The achieved Results for Successful Candidates (Sample-1)

Item# 1	Item# 2	Sup (item 1)	Sup (item 2)	Occur together	P-Value
11	38	3	42	0	0.135
11	41	3	35	1	0.951
47	48	4	69	3	0.791
18	38	3	42	2	0.379
18	41	3	35	1	0.951
41	47	35	4	1	0.669
41	48	35	69	21	0.153
39	41	80	35	29	0.6
39	48	80	69	59	0.04
38	39	42	80	31	0.188
38	41	42	35	15	0.899
38	47	42	4	4	0.016
38	48	42	69	26	0.192
37	38	5	42	5	0.607
37	39	5	80	3	0.387
37	41	5	35	3	0.204
37	47	5	4	4	0.301
36	38	11	42	11	0.607
36	39	11	80	8	0.387
36	41	11	35	4	0.204
36	48	11	69	6	0.301
35	38	1	42	0	0.607
35	39	1	80	0	0.387
35	41	1	35	0	0.204
35	48	1	69	0	0.301
34	38	1	42	0	0.607
34	39	1	80	0	0.387
34	41	1	35	0	0.204
34	48	1	69	0	0.301
33	38	1	42	0	0.607
33	39	1	80	0	0.387
33	41	1	35	0	0.204
33	48	1	69	0	0.301
32	36	14	11	0	0.156
32	38	14	42	5	0.607
32	39	14	80	10	0.387
32	41	14	35	7	0.204
32	48	14	69	8	0.301
31	36	2	11	0	0.156
31	38	2	42	0	0.607
31	39	2	80	1	0.387
31	41	2	35	1	0.204
31	48	2	69	0	0.301

Applying the proposed algorithm again on two transactions, one hundred itemset each, to get relations among items in the database. Again, table 5 shows the pruning of the unsuccessful itemsets that occur due to coincidence. Eight rules successfully are survived whose P-Value ≤ α. The successful candidates are:

- (Item 39 >> Item 48) with support=67,
- (Item 41 >> Item 48) with support = 45,
- (Item 38 >> Item 47) with support = 2,
- (Item 38 >> Item 48) with support = 44,
- (Item 37 >> Item 38) with support = 7,
- (Item 36 >> Item 38) with support 17,
- (Item 36 >> Item 39) with support 10 and;
- (Item 32 >> Item 48) with support = 11.

These items are not showing together on the transactions due to coincidence, they are related to each other. Again, when significance level increase, the number of produced rules increases and redundant and weaker rules are obtained as shown in the table 6 and figure 7.

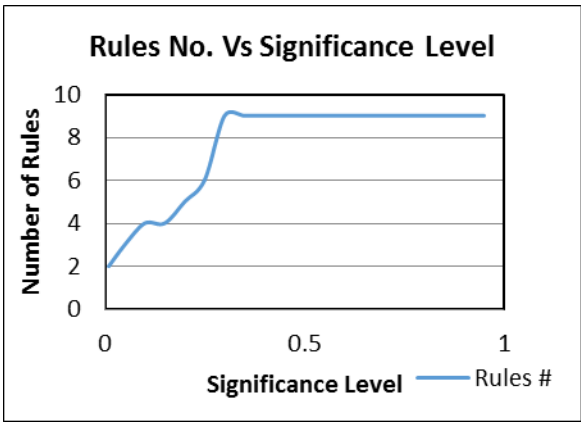

Fig. 7. Number of rules versus the significance level (Sample – 2)

Table 3. Number of rules versus the significance Level (Sample – 1)

significance level:	Number of Rules
0.01	0
0.02	1
0.03	1
0.04	2
0.05	2
0.1	2
0.15	4
0.2	15
0.25	15
0.3	22
0.35	22
0.4	30
0.45	30
0.5	30
0.55	30
0.6	38
0.65	38
0.7	39
0.75	39
0.8	40
0.85	40
0.9	41
0.95	43

Table 4. The achieved results for successful candidates (Sample-2)

Item# 1	Item# 2	Sup (item 1)	Sup (item 2)	Occur together	P-Value
9	38	6	33	2	0.986
9	41	6	40	1	0.229
9	48	6	76	3	0.124
38	39	33	85	25	0.069
38	41	33	40	19	0.012
38	48	33	76	18	0
39	41	35	69	21	0.253
39	48	80	35	29	0.294
41	48	80	69	59	0.002

Table 5. Number of rules versus the significance level (Sample – 2)

significance level:	Number of Rules
0.01	2
0.05	3
0.1	4
0.15	4
0.2	5
0.25	6
0.3	9
0.35	9
0.4	9
0.45	9
0.5	9
0.55	9
0.6	9
0.65	9
0.7	9
0.75	9
0.8	9
0.85	9
0.9	9
0.95	9

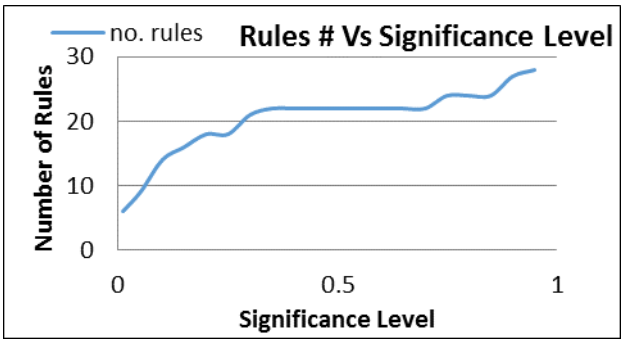

Fig. 8. Number of rules versus the significance level (Combined Data)

Table 6. The achieved results for successful candidates (Combined Data)

Item# 1	Item# 2	Sup (item 1)	Sup (item 2)	Occur together	P-Value
11	38	3	75	0	0.176
11	41	3	75	1	0.881
47	48	6	145	4	0.745
18	38	3	75	2	0.293
18	41	3	75	1	0.881
41	47	75	6	3	0.521
41	48	75	145	45	0.002
39	41	165	75	61	0.054
39	48	165	145	67	0.04
38	39	75	165	56	0.024
38	41	75	75	34	0.076
38	47	75	2	2	0.016
38	48	75	145	44	0.001
37	38	7	75	7	0.001
37	39	7	165	4	0.072
37	41	7	75	4	0.274
37	48	7	145	5	0.948
36	38	17	75	17	0
36	39	17	165	10	0.007
36	41	17	75	7	0.743
36	48	17	145	9	0.059
35	38	1	75	0	-
35	39	1	165	0	-
35	41	1	75	0	-
35	48	1	145	0	-
34	38	1	75	0	-
34	39	1	165	0	-
34	41	1	75	0	-
34	48	1	145	0	-
33	38	1	75	0	-
33	39	1	165	0	-
33	41	1	75	0	-
33	48	1	145	0	-
32	36	23	17	0	0.12
32	38	23	75	9	0.864
32	39	23	165	16	0.083
32	41	23	75	12	0.122
32	48	23	145	11	0.004
31	36	3	17	0	-
31	38	3	75	0	0.176
31	39	3	165	1	-
31	41	3	75	2	0.293
31	48	3	145	0	-

Both figure 8 and table 7 show that the number of rules increases producing weak rules as the significance level value increases. Again, if the significance level is raised from 0.05 to 0.2, some rules like (item 11 >> item 38) with support of zero, which give no information as item 11 appears 3 times and item 38 appears 75 times of the 100 transaction. When significance level is 0.4, rules like (item 31 >> item 41) with support of 2, while item 31 appears 3 times and item 41 appears 75 times of the 100 transaction. Successful produced rules are noticed to be strong rules where the occurrence of items together is not due to coincidence. In addition, the significance level (α value)

provides the level of certainty for the produced item combinations. For such significance level with value equal 0.05, it gives assurance of 95 % for the rule strength. In addition, as experimental results clarify, widen the space of the dataset would produce more efficient, strong and suitable rules and combinations.

Table 7. Number of rules versus the significance level (Combined Data)

Significance Level	no. rules
0.01	6
0.05	9
0.1	14
0.15	16
0.2	18
0.25	18
0.3	21
0.35	22
0.4	22
0.45	22
0.5	22
0.55	22
0.6	22
0.65	22
0.7	22
0.75	24
0.8	24
0.85	24
0.9	27
0.95	28

6. Conclusions

In this paper, a new algorithm is presented to discover all significant association rules among itemsets in large database transactions. The key feature of the proposed algorithm is that it does not require a pre-set threshold value for minimum support value. The proposed algorithm utilizes Minitab to run the Chi-Square test to produce P-value needed for the statistical hypothesis test, which is the core concept, mainly used and strongly applied through the proposed algorithm. It also takes the advantages of bitmapping features for speeding up the database transaction processing, reducing the time and space requirements. Intensive simulated experiments are conducted to validate the proposed algorithm. The achieved results showed that the proposed algorithm efficiently found the minimum number of rules, which are non-coincidental without using arbitrary support thresholds. In addition, it is found that the sensitivity of algorithm is high when tested for different data itemsets with different sizes.

References

[1] W. L. Suqiu and F. Xiaoliang, "Extraction of non-redundant association rules from concept lattices based on IsoFCA system". Fourth International Conference on Computer Science and Network Technology (ICCSNT 2015), IEEE, pp. 479-484, 2015.

[2] T. Slimani and A. Lazzez, "Efficient Analysis of Pattern and Association Rule Mining Approaches", International Journal of Information Technology and Computer Science, pp. 70-81 2014.

[3] D. S. Deshpande "A Novel Approach for Association Rule Mining using Pattern Generation", IJITCS, vol.6, no.11, pp.59-65, 2014.

[4] M. Awadalla and S. El-Far. "Aggregate Function Based Enhanced Apriori Algorithm for Mining Association Rules". IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 3, No 3, pp. 277-287, May 2012.

[5] S. Kannan, and R. Bhaskaran "Association rule pruning based on interestingness measures with clustering". International Journal of Computer Science Issues, IJCSI, 6(1), pp. 35-43, 2009.

[6] S. Pramod, O.P. Vyas, "Survey on Frequent Item set Mining Algorithms". International journal of computer applications, pp. 86-91, 2010.

[7] S. Venkata Krishna Kumar, P. Kiruthika. "A Survey on Association Rule Mining". International Journal of Advanced Research in Computer Science and Software Engineering 5(9), pp. 62-65, September 2015.

[8] A. M. Wanjohi, "Chi-Square Test for Goodness of Fit", 2016.

[9] S. A. Abaya "Association rule mining based on Apriori algorithm in minimizing candidate generation". International Journal of Scientific & Engineering Research Volume 3, Issue 7, July-2012.

[10] M. Shweta1 K. Garg " Mining Efficient Association Rules Through Apriori Algorithm Using Attributes and Comparative Analysis of Various Association Rule Algorithms" International journal of Advanced Research in Computer Science and Software Engineering, Kurukshetra University, India, Volume 3, Issue 6, June 2013 ISSN: 2277 128X

[11] J. Jha and L. Ragha. "Educational Data Mining using Improved Apriori Algorithm". International Journal of Information and Computation Technology. Department of Computer Engineering, Ramarao Adik

Institute of Technology, Navi Mumbai, India. ISSN 0974-2239 Volume 3, Number 5, pp. 411-418, 2013.
[12] S. Ahuja and G. Kaur. "Review of Association Rule Mining Using Apriori Algorithm". International Journal of Advanced Research in Computer Science and Software Engineering. Chandigarh University, India, Volume 4, Issue 3, March 2014.
[13] E. Weisstein, "Fisher's exact test". Math-World, a Wolfram Web Resource, 2005.
[14] M. F. Zibran, "Chi-Squared Test of Independence". Department of Computer Science, University of Calgary, Alberta, Canada, 2007.
[15] R. Lowry, 1999-2000, Chi-Square Procedures for the Analysis of Categorical Frequency Data.
[16] S. Bai, H. Ren, W. Jiang, and Y. Jiang, "Hypothesis Testing Based Knowledge Discovery in Distributed Multiple Data Sources". International Conference on Internet Technology and Applications, pp. 1- 4, 2010.
[17] IBM Research - Almaden.
WWW. Almaden.ibm.com. Retrieved 31-01-2013.

Medhat Awadalla is an Assistant professor and works at Electrical and Computer Engineering Department, Sultan Qaboos University. He obtained his PhD from university of Cardiff, UK, 2005. He obtained his MSc. and BSc. from Helwan University in 1991 and 1996 respectively, Egypt. His research interest includes Energy efficient real time systems, sensor networks, high performance computing.

Sara El-Far is MSc. student at college of Engineering, Helwan University. She obtained her BSc in computer science and engineering from university of Helwan, Egypt. Her research interest includes Data Mining, high performance computing, and embedded systems.