Mining time Series Data without the consideration of support and confidence

Mr. Rohit Garud[1], Prof.Mrs. Rajeswari Kanan[2], Dr.Prof.Mr. V.V.Vaithiyanadhan[3]

Abstract—The data in the real world is in the form of quantitative and time related format. The time related data is also called as time series data. It is a combination of quantitative and dynamic data. The values of time series data change with respect to time. As it is dynamic in nature it is difficult to handle time series data. An example is in IPL 2014 the runs scored in each of the over is shown in a time series manner. For each ball the number of scored runs is different. Apriori algorithm is used to mine binary valued data and also use minimum support and confidence values as threshold to generate rules. Fuzzy logic deals with quantitative data but does not deal with time series data. It also use minimum support and confidence to generate rules. To define an appropriate threshold value is still a research topic. Thus in this paper we propose a new algorithm which takes input as time series data and generate business oriented rules without a particular threshold value.

Index Terms—Binary data, Fuzzy Coherent, Quantitative data and transactions, Minimum Support and Confidence, Time Series.

———————————— ◆ ————————————

1 Introduction

DATA mining is used to help people to find something innovative or interesting from huge collection of raw data, which might help them to generate new information from it. An example is an IPL 2014 time line data then we can get a bowling strategy of a team using this information. The information such as when is it good time to tackle a ball ie which number ball should a batsman hit or takes risk we can also have a same for a particular bowler. The algorithm proposed in this paper is capable to do this which will ultimately help the team to get good score. The algorithm generated the rules which are business oriented which helps in earning profit to a particular business.

2 Literature REview

There are many methodologies to generated rules that use minimum support and confidence to generate rules but the data input to these methods are quantitative valued data. To set or to predict an appropriate threshold is a difficult task. As if threshold is set to a high value number of rules are eliminated but if we keep low threshold less number of rules are eliminated. The threshold can be predicted by experience but it is also difficult to prove that the threshold value predicted is correct. Threshold value changes with respect to application that is why it is application dependent. Fuzzy logic helps to overcome the above issues. It is a very close to human language that is why fuzzy representation is more effective rather than binary representation. We can easily convert the natural language to a coding using fuzzy. The rules generated are represented in the form of itemA → itemB which depicts that itemA in the rule is closely related to itemB in the same rule.

There are two types of rules

1. General rule: The rule is easy to generate as it's a common sense rule for example In cricket whether a ball with free hit gets punished to four or six.
2. Business rule: The rules which are business oriented which can be used to increase profit of the business for example in cricket every six ball in the first over is hit for a maximum score as a new batsmen has to analyze the pitch and the movement of the ball due to this the batsmen will always attack the last ball and play little carefully with the first five balls

Membership function is the representation of the time series value in the into fuzzy format. These fuzzy values are used for further calculation in the algorithm to generation of rules. Triangular membership functions are used for critical and dynamic applications. Membership functions are the most critical part of the fuzzy logic without membership function it is not possible to run the algorithm as it plays an important role of converting the time series data to fuzzy values. The given below figure 1 depicts the same example of membership function.

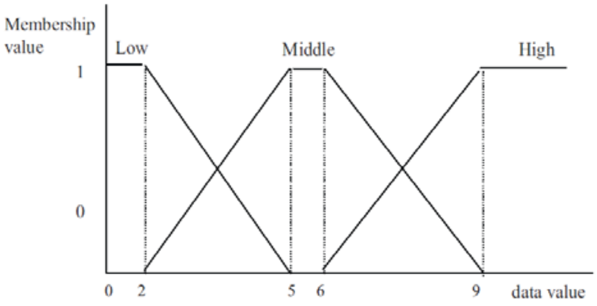

Figure 1 Membership function.

3 PROPOSED ALGORITHM

1. Take input as time series data set.
2. Generate the sub sequence based on the number of attribute of the data set.
3. Then convert each subsequence value from time series to fuzzy values representation with help of appropriate membership function.
4. Then generated the support count for each pair or attribute and check support count(itemA,ItemB) > support count(-itemA,ItemB), support count(itemA,ItemB) > support count(itemA,-ItemB), support count(-itemA,-ItemB) > support count(-itemA,ItemB), support count(-itemA,-ItemB) > support count(itemA,-ItemB)
5. If the following above condition satisfies then we consider it as the rule which is good from business point of view.
6. Otherwise we discard it.
7. Stop

4 EXAMPLE

Data set:

Time series data	1,2,5,3,2

The subsequence :

Subsequence : 125 253 532

Fuzzy values:

Eqvivalent fuzzy values:	1.0 , 0.0 , 0.0 1.0 , 0.0 , 0.0 0.0 , 1.0 , 0.0 1.0 , 0.0 , 0.0 0.0 , 1.0 , 0.0 0.75 , 0.25 , 0.0 0.0 , 1.0 , 0.0 0.75 , 0.25 , 0.0 1.0 , 0.0 , 0.0

All the possible combination with condtion.

All possible pair support	Sum in A=1.0 B=1.0 C =0.75 D =0.25 Rule not generated is A1.low->A2.low Sum in A=1.0 B=1.0 C =0.25 D =0.75 Rule not generated is A1.low->A2.Medium Sum in A=0.0 B=2.0 C =0.0 D =1.0 Rule not generated is A1.low->A2.high Sum in A=0.75 B=1.25 C =1.0 D =0.0 Rule not generated is A1.low->A3.low Sum in A=1.25 B=0.75 C =0.0 D =1.0 The Rule Generated is A1.low->A3.medium Sum in A=0.0 B=2.0 C =0.0 D =1.0 Rule not generated is A1.low->A3.high Sum in A=0.75 B=1.0 C =1.0 D =0.25

5 RESULTS

The previous section shows the example of each step proposed in the algorithm. This algorithm does not take any threshold value as input but generate the rules which are good from business point of view. The previous methodology and the proposed is represented in a graphical format in the graph given in figure 2.

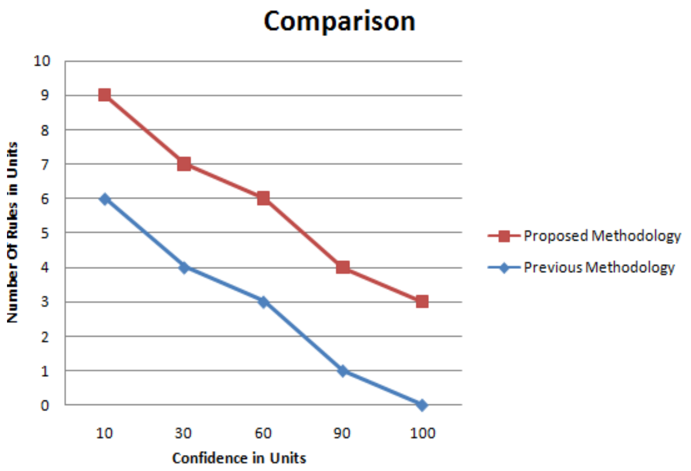

Figure 2: Comparison Graph

The above graph consist of the representation of the number of rules generated based on a particular threshold value given to the algorithm on different algorithms. The x axis depicts the number of rules generated and the y axis represents confidence.

Factors of dependency:

1. Time: The time take for the proposed algorithm is less than the previous as in the previous we need to calculate the support and confidence for each pair and then check which is not in proposed method only once comparison is done so takes less time.
2. Space: The space to store the calculate support and confidence is more than that of proposed algorithm as only support if each pair is stores so as proposed algorithm need to store only support value the no confidence so the space is reduced .
3. Rules : The rules generated will be more business oriented than common since as the propositional logic has been incorporated in the algorithm .
   Propositional logic: This logic help in generation of rule which is business oriented rules.
4. Stability: The rules remain stable even though the data is dynamic for a particular time. In the previous system the data changes with respect to change in minimum support and confidence value. The number rules generated are fixed irrespective of the threshold value.

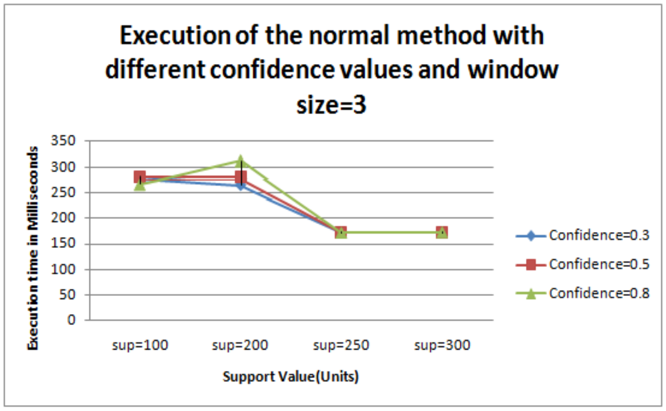

Figure 3: Execution time of algorithm based on different support

The above Figure 3 depicts the execution time of the previous algorithm with different confidence and support values which states that with more confidence the execution time increases for different support. Each color line defines different confidence value with window size as 3.

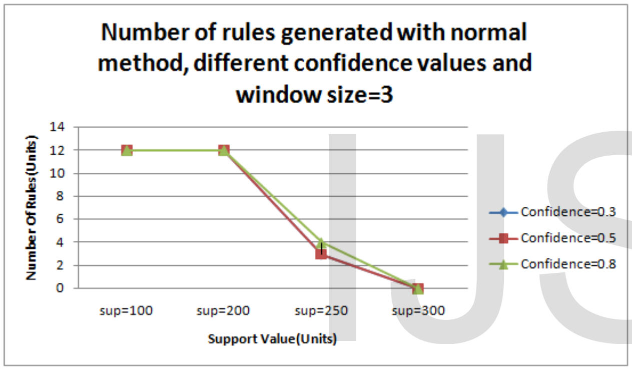

Figure 4: Number of rules generated

The above graph depict 4the number of rules generated with different support and confidence of 0.3, 0.5, 0.8 based on the previous methodology as the support and confidence increases the number of rules keep on decreasing so it is inversely proportional in nature Each line depicts the different confidence value with respect to window size as 3.

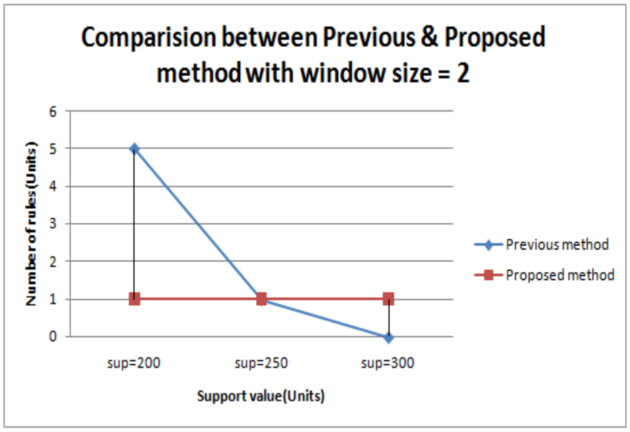

Figure 5 Number of rules generated with window size-2

The above graph depicts 5 states the relation between number of rules generated with different support based on the previous and proposed methodology. With previous method as the support increases the number of rules starts decreasing on the contrary with proposed methodology the number of rules generated remains constant irrespective to the support value.

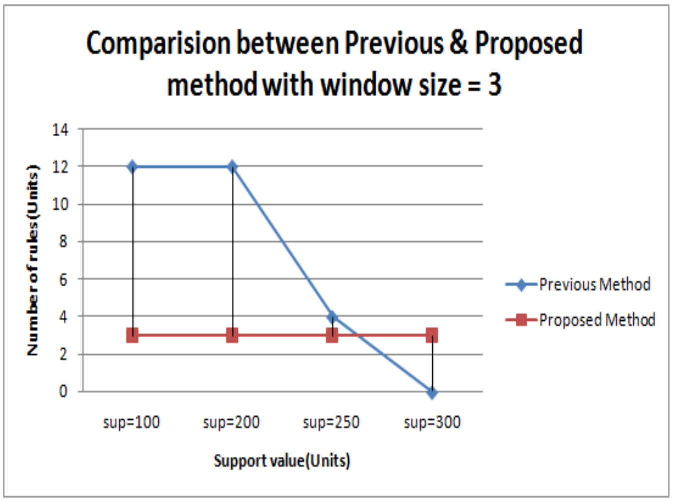

Figure 6 Number of rules generated with window size-3

The Figure 6 above depicts the number of rules generation with respect to the window size. As the window size increased the number of rules generates is increased but the increase in previous methodology is in linear fashion and with the proposed method it very less it does not increase in linear fashion but very little it increases at each different window size.

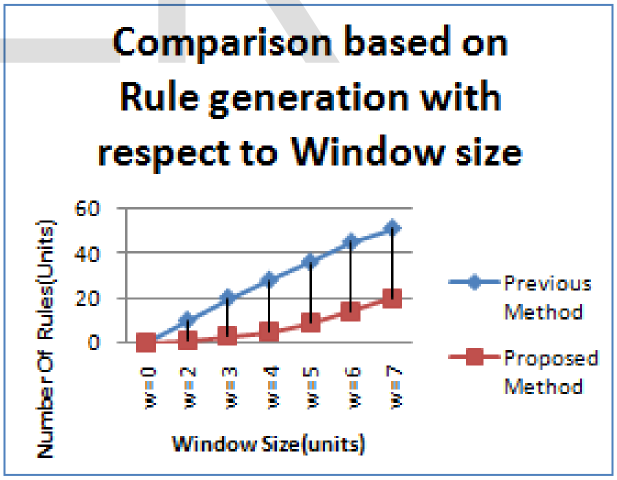

Figure 7 Number of rules generated based on window size

The above Figure 7 shows number of rules generated for the different algorithm the with respect to the window size. The number of rules keeps on increasing as the size of the window increases it. The parameter number of rules generated and window size is directly proportional to one another. The curve representing rules with respect to window size of the proposed algorithm is more like a sine wave on the other hand for the previous algorithm it increases diagonally.

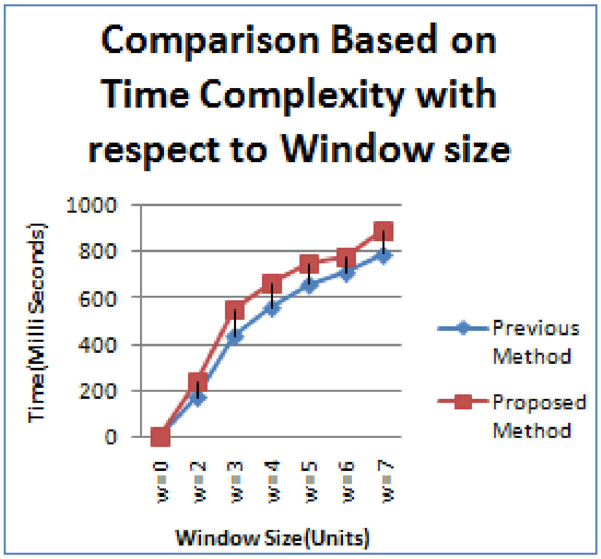

Figure8 Time Complexity with respect to window size

The Figure 8 defines the time taken by both the previous and proposed in completion of execution of the algorithm. Both the techniques as the window size increases the time taken for the execution also increases. The time required for execution is nearly close to one another the difference lies only in the rules generated. With increase in the window size the time taken by the algorithm also increases.

CONCLUSIONS

Thus we propose a new algorithm which accepts time series data as input process the data to generate association rules without applying threshold parameter of association rule mining i.e. minimum support and confidence value. The proposed algorithm uses propositional logic to process the data and generated rules that can help the manager of any organization to analyze their growth and helps in prediction purposes and decision making.
As the window size increases the time taken to process the data also increases due to which harms the performance of the algorithm in order to overcome this technologies like Hadoop, Parallel computing or distributed processing can be incorporated

ACKNOWLEDGMENT

I would like to thank my guide Prof Rajeswari Kanan and Dr. V. Vaithiyanathan for the help and support.

REFERENCES

[1] Chun-Hao Chen, Tzung-Pei Hong , Vincent S. Tseng, ”Fuzzy data mining for time-series data”, Applied Soft Computing 12 (2012) 536–542.

[2] Chun-Hao Chen, Ai-Fang Li , Yeong-Chyi Lee, ” A fuzzy coherent rule mining algorithm”, Applied Soft Computing 13 (2013) 3422–3428

[3] Chun-Hao Chen, Guo-Cheng Lan, Tzung-Pei Hong, Yui-Kai Lin “Mining high coherent association rules with consideration of support Measure” Applications 40 (2013) 6531–6537

[4] Hau lin lin “Generating Weighted Fuzzy Rules for Handling Classification Problems” International Journal of Electronic Business Management, Vol. 3, No. 2, pp. 116-128 (2005) 116

[5] J. Alcala-Fdez, R. Alcala, F. Herrera, A fuzzy association rule-based classification model for high-dimensional problems with genetic rule selection and lateral tuning, IEEE Transactions on Fuzzy Systems 19 (5) (2011) 857–872.

[6] K.C.C. Chan, W.H. Au, An effective algorithm for discovering fuzzy rules in relational databases, The IEEE International Conference on Fuzzy Systems 2 (1998) 1314–1319.

[7] C.H. Cai, W.C. Fu, C.H. Cheng, W.W. Kwong, Mining association rules with weighted items, in: The International Database Engineering and Applications Symposium, 1998, pp. 68–77.

AUTHORS

Mr. Rohit Garud is pursuing his Master Of Engineering Computer Engg from Pimpri Chinchwad College Of Engineering, pune also completed Bachelor Of Engineering from Genba Sopanrao Moze College Of Engineering, Pune paased in 2012 and has done Diploma from Cusrow Wadia Of Technology, Pune.

Prof.Mrs. K. Rajeswari received BE and Mtech in Computer Science .She has published several papers in Data Mining. She is winner of Cambridge International Certification for Teachers with Distinction. She has about 15 years of Teaching experience Currently she is working as an Associate Proffessor Computer Science and Engg in PCCOE,Akurdi,Pune.She is pursuing her Ph.D in SASTRA University,Tanjore,India.

Dr. V. Vaithiyanathan is a Professor Associate Dean Re-search, CTS Chair Professor, SASTRA University. His area of interest are data mining and Image Processing. He is associated with many funded projects. He has published many International papers in reputed journals and conferences