A review on mining Quantitative, Time series and Binary data using fuzzy mining algorithms with and without minimum support and confidence

Mr. Rohit Garud, Prof.Mrs. Rajeswari Kanan, Dr.Prof.Mr. V.V.Vaithiyanadhan,

Abstract—To mine or to extract knowledge from the data there are many ways Association rule mining, Decision trees, Fuzzy logic, Fuzzy logic with weights, etc in which some of them requires minimum support and minimum confidence values as input from the user as a threshold values. This papers thus gives a review on these techniques with its pros and cons.

Index Terms— Association rule mining, Decision trees, Fuzzy logic, Fuzzy logic with weights, etc

1 INTRODUCTION

DATA mining is a way through which we can discover or invent a new thing from existing things it is just a kind of recycling the things which can be re-used from garbage of data. There are many ways in which we can mine the data like Association rule mining, Decision tree, Neural Network and fuzzy logic etc[1,4]. to mine a data we require something called as minimum support and minimum confidence which helps to supply a threshold or cut off from which we can distinguish the important and common rules generated from the mining algorithm here are some of the few techniques given below which deals with mining data using fuzzy logic mining with and without supplying minimum support and minimum confidence[1].

2 LITERATURE REVIEW

The fuzzy logic was introduced by Zadeh in 1965. Fuzzy logic is the logic with talks about partial truth or false where the truth/false is rated in percentages or between 0, 0.1, 0.2, 0.3……0.9, 1. So if the truth value id 0.6/60% then the false value is 0.4/40%. This means that the value is 60% truth but 40% false so fuzzy logic deals with this kinds of things. A fuzzy set A in the universe of discourse U, U = {u1, u2,…,un}, can be characterized by a membership function μA where μA: U [0,1], μA(ui) indicates the grade of membership of an element U in the fuzzy set A, and μA(ui) ∈ [0,1]. A fuzzy set A of the universe of discourse U, U = {u1, u2,…,un}, can be represented by content here[4].

$$A = \mu_A(u_1)/u_1 + \mu_A(u_2)/u_2 + \ldots + \mu_A(u_n)/u_n = \sum_{i=1}^{n} \mu_A(u_i)/u_i$$

Definition 1: Let A and B be fuzzy sets of the universe of discourse U with the membership functions μA and μB respectively. Then, A is a subset of B, denoted as A is subset of B , if and only if μA (u) <= μA (u) for all u belongs to U[4].

Definition 2:Let A and B be fuzzy sets of the universe of discourse U, and let μA and μB the membership functions of the fuzzy sets A and B, respectively. The fuzzy subsethood S(A, B) measures the degree in which A is a subset of B[4].

$$S(A,B) = \frac{M(A \cap B)}{M(A)} = \frac{\sum_{u \in U} \min(\mu_A(u), \mu_B(u))}{\sum_{u \in U} \mu_A(u)}$$

where S(A, B) ∈ [0, 1]

Definition 3:Let A be a fuzzy set of the universe of discourse U, where U = { u1, u2, u3……. un },
A = μA(u1) /u1 + μA (u2)/u2 + ... + μA (un)/un and let α be a real value between zero and one[4]. The α -level-cut of the fuzzy set $Aαlevel cut is defined as follows:

$$A_{\alpha\text{-level}} = \sum_{i=1}^{n} \mu_{A_{\alpha\text{-level}}}(u_i)/u_i$$

2.1 Generating Weighted Fuzzy Rules for Handling Classification Problems[4].

In recent years, many methods have been proposed to generate fuzzy rules from a set of training data. In this Author present a new method to automatically generate weighted fuzzy rules from a set of training data, where the attributes appearing in the antecedent parts of the generated fuzzy rules have different weights, respectively. Author also applies the generated weighted fuzzy rules to deal with the "Saturday Morning Problem". The problem states that user have some plans to do on Saturday ie either Swim, Weight lift or Volleyball and we need to find out what will the user do depends on some weather conditions[4]. So to do that we have a new proposed algorithm which is as follows:

1. Accepts the threshold value R cut to convert the data set into 0's and 1's.
2. Convert the data set to Frequency Distribution Table.
3. Convert Frequency Distribution Table to Probability Distribution Table.
4. Find Fuzzy Subset hood values for each attribute with class label.

Advantages:

1. The proposed method can get a higher classification accuracy.

Disadvantages:

1. Difficult to define a particular support value.
2. Generated Negative rules are not sufficient to define a rule correctly.
3. Generated rules do not give Business logic.

2.2 Mining High Coherent Association Rules with Consideration of Support measure[3]

Data mining has been studied for a long time. Its goal is to help market managers find relationships among items from large databases and thus increase sales volume. Association-rule mining is one of the well known and commonly used techniques for this purpose[3,2]. The Apriori algorithm is an important method for such a task. Based on the Apriori algorithm, lots of mining approaches have been proposed for diverse applications[3]. Many of these data mining approaches focus on positive association rules such as ''if milk is bought, then cookies are bought''. Such rules may, however, be misleading since there may be customers that buy milk and not buy cookies[3]. Author in this paper thus takes the properties of propositional logic into Consideration and proposes an algorithm for mining highly coherent rules. The derived association rules are expected to be more mean-ful and reliable for business[2,3].

Advantages:

1. Highly coherent association rules were developed to help decision-makers make marketing strategies.
2. To reduce the time required to generate candidate coherent itemsets, the corresponding lower and upper bounds of supports of subitemsets in the consequent part of an itemset are defined for removing itemsets that cannot become highly coherent itemsets..

Disadvantages:

1. Lower and Upper bounds are considered as to remove some rules which may in lead to the case of less rule generation.

2.3 A Fuzzy Coherent Rule Mining Algorithm[2]

In real-world applications, transactions usually consist of quantitative values. Many fuzzy data mining approaches have thus been proposed for finding fuzzy association rules with the predefined minimum support from the give quantitative transactions[2]. However, the common problems of those approaches are that an appropriate minimum support is hard to set, and the derived rules usually expose common-sense knowledge which may not be interesting in business point of view[2]. Author in this paper have suggested an algorithm for mining fuzzy coherent rules is proposed for overcoming those problems with the properties of propositional logic. It first transforms quantitative transactions into fuzzy sets[2]. Then, those generated fuzzy sets are collected to generate candidate fuzzy coherent rules[2,3]. Finally, contingency tables are calculated and used for checking those candidate fuzzy coherent rules satisfy the four criteria or not. If yes, it is a fuzzy coherent rule.

Advantages:

1. The proposed approach can derive interesting rules effectively without setting minimum support.
2. More rules which are best for business.
3. No duplicate rules generated.
4. Rule with length 4 is also generated.

Disadvantages:

1. More time is required for execution ie High Time Complexity.

2.4 A Fuzzy Data mining for Time Series Data[1]

Time series analysis has always been an important and interesting research field due to its frequent appearance in different applications[1]. In the past, many approaches based on regression, neural networks and other mathematical models were proposed to analyze the time series[1]. Author in this paper attempt to use the data mining technique to analyze time series. Many previous studies on data mining have focused on handling binary-valued data. Time series data, however, are usually quantitative values, thus extend our previous fuzzy mining approach for handling time-series data to find linguistic association rules[1,2]. The proposed approach first uses a sliding window to generate continues subsequences from a given time series and then analyzes the fuzzy itemsets from these subsequences. Appropriate post-processing is then performed to remove redundant patterns[1].

Advantages:

1. The final results are represented by linguistic rules, they will be friendlier to human than quantitative representation.
2. Through the post-processing procedure in the algorithm, lots of redundant rules can be filtered such that the mined rules can be compact. Users can thus utilize the rules more easily

Disadvantages:

1. Specify the threshold values.
2. Duplicate Rules are generated.

3. COMPARITIVE STUDY

Table 1 depicts the comparative e study of all the reference papers were support is the threshold value given to Apriori length of rule is number of antecedent and consequent values, Type of rule states whether rule is Commen-sence represented as CS and Business as B, Duplicity defines redundancy

.

Table -1: Comparative analysis

Paper Reference no	Support	Length Of Rule	Type Of Rule	Duplicity
1	Y	2	CS	N
2	Y	3	B	N
3	Y	4	B	Y
4	N	4	B	N

4 CONCLUSION

There is different technique to mine data sets. The data can be in different format. The right Mining algorithm with right data will result in better prediction and generation association rules. We thus need to make a comparative analysis before using a particular technique to find its advantages and distadvantages and use it in required manner.

ACKNOWLEDGMENT

I would like to thank my guide Prof Rajeswari Kanan and Dr.Prof.Mr. V.V.Vaithiyanadhan 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