Global Journal of Enterprise Information System

DOI: 10.18311/gjeis/2017/15480

Sensitivity Association Rule Mining using Weight based Fuzzy Logic

Meenakshi Bansal[1*], Dinesh Grover[2] and Dhiraj Sharma[3]

[1]Research Scholar, IK Gujral, PTU, Jalandha, Punjab, India; ermeenu10@gmail.com

[2]Professor, IK Gujral, PTU, Jalandhar, Punjab, India; dineshgrover@yahoo.com

[3]Assistant Professor, Punjabi University, Patiala, Punjab, India; Dhiraj.pbiuniv@gmail.com

Abstract

Mining of sensitive rules is the most important task in data mining. Most of the existing techniques worked on finding sensitive rules based upon the crisp thresh hold value of support and confidence which cause serious side effects to the original database. To avoid these crisp boundaries this paper aims to use WFPPM (Weighted Fuzzy Privacy Preserving Mining) to extract sensitive association rules. WFPPM completely find the sensitive rules by calculating the weights of the rules. At first, we apply FP-Growth to mine association rules from the database. Next, we implement fuzzy to find the sensitive rules among the extracted rules. Experimental results show that the proposed scheme find actual sensitive rules without any modification along with maintaining the quality of the released data as compared to the previous techniques.

Keywords: Fuzzy Logic, Sensitive Rules, WFPPM, Weights

Paper Code: 15480; **Originality Test Ratio:** 19%; **Submission Online:** 17th March 2017; **Manuscript Accepted:** 25-March-2017; **Originality Check:** 28-Mar-2017; **Peer Reviewers Comment:** 16-April-2017; **Double Blind Reviewers Comment:** 02-May-2017; **Author Revert:** 12-May-2017; **Camera-Ready-Copy:** 28-May-2017

1. Introduction

Data is an important part of an organization. Due to the growth of data in databases, there comes the need of data mining. Data mining extract important data items from large databases using different association rule mining techniques. In this paper association rules are mined by using FP- Growth unlike other studies which mostly use apriori data mining technique. This data is shared among different companies which help them in better decision making. Along with the normal data this data also contains some sensitive information. Major issue raised during information sharing is leakage of sensitive information. So it is required that this sensitive information must be extracted and hidden. In this paper main focus is on extracting sensitive information from the database which is further hidden using cryptography technique. Privacy preserving data mining provided a solution to this problem. In the previous research methods thresh hold support and confidence are used as a benchmark for extracting sensitive rules. They increase or decrease the support or confidence of the rules for making it sensitive. These methods only consider the frequency of the items in the data set to calculate its sensitivity. Doing this causes serious side effects to the original databases like generating ghost rules, lost rules and false rules. But in some cases like banking or any other financial organization sensitivity of the rule does not only depend upon the frequency of the item set in the given database. In such cases values of the parameters are more important upon which the sensitivity of the rule depends. To overcome the limitation of existing methods the approach of weighted fuzzy association rules are used.

Fuzzy logic is based on the concept of decision making on the basis of range of values instead of stop ranges. After extracting the rules from the available dataset using FP-growth sensitive rules are mined using WFPPM technique.

The remaining parts of this paper are organized as follows. Related studies regarding Fuzzy correlation algorithm, Fuzzy FP-Growth and Fuzzy Apriori will be in Section 2. Data set used for the implementation of this paper is explained in Section 3. Section 4 gives the graphical representation of the dataset. The proposed algorithm is given in Section 5. The experimental results demonstrate the calculated weight of individual attribute used to find the sensitivity of the rule is given in Section 6. The conclusion is then given in Section 7.

2. Related Studies

Existing studies mostly depended upon two values support and confidence for deciding the sensitivity of the rule. (Karthikeyan et al., 2012; Chueh,2007) proved that only fuzzy support and confidence measures are insufficient for filtering out uninteresting fuzzy correlation rules, so a new method has been proposed for discovering fuzzy association rules using fuzzy correlation rules. In this framework a fuzzy correlation measure for fuzzy num-

*Author for correspondence

bers, is used to augment the fuzzy support- confidence for fuzzy association rules. Author described that according to the aforementioned framework if support and confidence of the two fuzzy items is equal to or greater than minimum threshold values than they are considered as an interesting fuzzy association rule. But the presence of this fuzzy item-set does not necessarily imply the presence of other fuzzy item-sets which are also included in these fuzzy association rules. Hence there is an urgent need for analysing the relationships between fuzzy item-sets. Therefore authors try to find out the linear relationship between two fuzzy item-sets using fuzzy correlation analysis.[3] Applied fuzzy FP- growth to mine fuzzy association rules instead of apriori data mining algorithm and compare the difference. From the experimental results it is illustrated that the proposed method outstand the apriori algorithm in aspect of execution time. It proved that by applying FFP-growth to generate frequent patterns can highly promote the overall efficiency execution. [4] Analyzes time series data using extended Fuzzy Frequent Pattern (FP) growth approach against the existing approach called Fuzzy Apriori (FA). Extended fuzzy FP- Growth approach is 2-step process. Firstly frequent itemsets were found using FP tree algorithm. Secondly extended fuzzy FP-Growth is applied. Experimental results show that this method is efficient and scalable for extracting both small and long frequent patterns. The Fuzzy FP-Growth approach provided solution to the problem of finding long frequent patterns by using least frequent items as a suffix and there by offering good selectivity. The approach reduces the search costs to a great extend. FP growth approach outperforms FA approach in terms of execution time.[5] Proposed new Fuzzy ARM algorithm which is 8-19 times faster for the very large standard real-life dataset as compared to fuzzy apriori. This algorithm contains the novel combination of features like two-phased multiple partition tid list-style processing, byte-vector representation of tid lists, and fast compression of tid lists that contribute a lot to the efficiency in performance. This algorithm also included an effective pre-processing technique for converting a crisp dataset to a fuzzy dataset. Fuzzy Apriori itself was a very inefficient and slow algorithm when it comes to dealing with very huge datasets. Thus, any fuzzy adaptations of Apriori would be inadequate to deal with newer real-life datasets which are becoming larger day-by-day. Very large data sets also can't be handled by in memory algorithms like FP-Growth but only by algorithms which are not totally memory dependent. But the proposed algorithm was based on a two-phased processing technique, and uses a tid list approach for calculating the frequency of item sets. Due to which the proposed algorithm was much faster than the fuzzy apriori algorithm. [6] Extended his research by using Borgelts prefix trees for computing fuzzy association rules. In order to adapt Borgelts algorithm to mine fuzzy association rules, they modified the prefix tree structure to store the fuzzy frequencies of each item set and modified the algorithm to propagate these values through the tree as it is updated. It is proved experimentally that fuzzy rule sets provide better running time and accuracy. Authors included fuzzy logic management in the software provided by Borgelt, and have obtained similar accuracy with respect to the basic Apriori algorithm. However, the running time of Borgelts algorithm is much lower, and its use may enable online detection while[7] integrated Decision Tree and Naive Bayes with fuzzy logic for diagnosis heart disease. In their research, six attributes were reduced to four attributes which automatically reduced the number of tests to be taken by a patient. From their experiment they proved that these techniques outplayed other data mining techniques. Further[8] proposed new fuzzy mining algorithm based on the AprioriTid approach to find fuzzy association rules from given quantitative transactions. The proposed algorithm can also solve conventional transaction-data problems by using degraded membership functions. Author in his work has assumed that membership functions are known in advance. But in future research attempt can be made to dynamically adjust the membership functions in the proposed mining algorithm to avoid the bottleneck of the acquisition of membership functions. Research analyzed that Apriori performs better than AprioriTid in the initial passes but in the later passes AprioriTid has better performance than Apriori. Due to this reason another algorithm can be used called Apriori Hybrid algorithm in which Apriori is used in the initial passes but in the later passes one can switch to AprioriTid. In[9] authors proposed new algorithm named Fuzzy Cluster Base (FCB) which worked along with Partial Fuzzy Cluster Base (PFCB). Unlike PFCB, FCB performs single database scan for calculating the final support of itemsets. It has been proved in[10] that repeated scan of database for mining association rules decrease the overall performance of the algorithm. Therefore a new algorithm named Fuzzy Cluster Test (FCT) has been proposed which reduced incredible amount of scanning data as compare to existing Fuzzy Association rule Mining. Therefore the running time of mining algorithm having lesser database scan is reduced greatly and shows better performance. In[11] authors used fuzzy programming approach instead of fuzzy algorithm to fulfill 2 objectives: a) data reduction b) privacy preserving for data sharing. While achieving these two objectives Euclidian distances must be preserved. Authors proposed the approach for selection of Fourier coefficient through coefficient suppression to achieve the above said objectives. Many algorithms have been proposed two solve these issues, one of them is random projection method but they did not preserve Euclidean distances due to non-orthogonality of the matrix. Hence fuzzy programming approach has been proposed to select a minimal set of high-energy coefficients across the rows of transformed data. Experimental results demonstrate that the proposed approach outperforms in achieving much better mining quality than the existing techniques like random

perturbation and random projection giving the same degree of privacy in both centralized and distributed cases. In the above quoted studies related to fuzzy association rule mining (FARM) emphasise was on the use of fuzzy in finding association rules by the use of either apriori or FP- growth instead of finding association rules. Some studies have considered minimum threshold support or confidence value for finding sensitive rules. But considering only the threshold value of support and confidence is not sufficient for efficiently extracting the sensitive rules from the database. So in this paper our focus is on extracting sensitive rules from the database using WFPPM technique.

Data used

Sample data set that we use in our experiment is the standard dataset. All the datasets are chosen from UCI Machine Learning[12] repository and KEEL datasets[13,14]. This datasets are explained in Table 1.

3. Proposed Methodology

3.1 FP-Growth

FP-growth algorithm is one of the latest and most efficient algorithms in depth-first algorithm[15]. The algorithm does not subscribe to generate and test paradigm of Apriori. FP-growth adopts a divide-and-conquer strategy. It encodes the data set using a compact data structure called FP-tree and extracts frequent itemsets directly from this structure. The problem of mining frequent patterns in database is transformed into that of mining the FP-tree. Figure 1 illustrates the process of finding frequent pattern using FP-Growth.

3.2 Working of FP-Growth

- Convert csv file to .dat file.
- Pass this .dat file to the main program.
- From main program we will call algorithm that will create FP-tree and will implement FP-growth.
- First records from file will be traversed one by one. From these records we will find frequent items. To find frequent items :-
- First a record will be picked, then record will be divided into tokens, these tokens will be added to map, where each entry will contain the occurrence of the item. We will sort these map entries on the basis of occurrence, keys having occurrence more than threshold will be frequent items.

For tree Creation:-

- Create a root node.
- Get each record from the file, divide record into token, and then create a list which contain frequent item in sorted order for a particular record. Send this list for insertion in tree as nodes.
- Get item from the list then if item is a children of the tree increment it's count else add item to the tree as children and add it to the end of the tree node.

4. Finding Frequent Pattern

- Find frequent pattern in the tree with their support and add to map.
- Read the tree in recursive order, child to parent till root node and create conditional patterns. Add conditional pattern to the map as key and last node count as value of map.

Figure 1. Mining by FP-Growth.

- Divide conditional patterns into tokens; add these tokens to a map as key and count of conditional pattern containing this token as value.
- Ignore the items having values less than threshold, create conditional tree using higher value items.

Table 1. Data Set Descriptions[12]

Data Set Characteristics:	Multivariate including (Numeric, Categorical and binary values)	Number of Instances:	45211	Area:	Banking
Attribute Characteristics:	Real	Number of Attributes:	16 + output attribute	Date Donated	2012-02-14
Associated Tasks:	Association Rule Mining	Missing Values?	N/A	Number of Web Hits:	163828

- Process the map of conditional pattern and conditional frequent items to get frequent items in the record. Add these items to the list and use this list to create frequent patterns.

4.1 Fuzzy weight based prediction

Fuzzy logic is the branch of logic in which the truth value of a logical proposition is represented as a real value on unit interval [0, 1]. Fuzzy logic provides means to represent approximate knowledge[16-18]. It is a logic that arrives at a definite conclusion based on vague, ambiguous, or imprecise input information. The algorithm consists of the following steps:

- Find all the available parameters on which rules can be built. Parameters are simply the fields in our dataset like age, education etc.
- Find data type of each parameter along with the range of values that each parameter holds.
- Find lower and upper bound of the parameters, to create member function. Bound can decide the type of membership function, like if our age parameter is spanned between 10 -100 and our higher bound value is set to 50 then it is a triangular membership function. Starting from 10, peak is at 50 and then again slides down toward higher values. Using fuzzy logic we are assigning a weight to the parameter values.
- For integer and float type our function automatically set the weight to the parameter value based on our lower and upper bounds.

$$Range\ value = ((higher\ bound - lower\ bound)/\ 10) \quad (1)$$

$$Weight = 1 - (critical\ value - value\ for\ parameter) / range\ value * 10) \quad (2)$$

Equ. (1) gives the value of range which depends upon the bounds for parameter which is between the lowest and highest value it has. Lower bound is the lowest value and higher bound is the max value, critical value depends on how we want to set our range.

Equ. (2) finds the weight of individual parameter. Suppose for age between 10 – 100 if we want to give highest preference to people around age 50 then we will set critical value to 50 and function will give highest weight to the age around 50. If we set highest bound to 100 functions will give highest preference to 100.

- For string values, we have a dictionary of string weights, which can help us define the range of weights for different values. Like for a profession parameter we can set weight for various profession which can help to categorize the professions over the range of lower, medium and higher profession. Or A, B, C and D class profession. And as per our requirement we can define the member function to target a particular range. e.g. If managers need to target high class and lower middle class professional. Then we can define our member function in such a way e.g. we can set fuzzy logic like,

  If profession is Class A Then Highly suitable
  If profession is Class B Then not so suitable
  If profession is Class C Then suitable
  If profession is Class D Then not suitable.

So we can define string weight based on string weight dictionary and based on that can define fuzzy logic for string.

- After finding the weight of individual parameter using fuzzy based weight calculation, weight of each rule is calculated. For this we need to read mined rules one by one.
- We will read one rule, split it into parameters and then our weight calculation module will find weight for each parameter based on its type and our defined boundaries. After

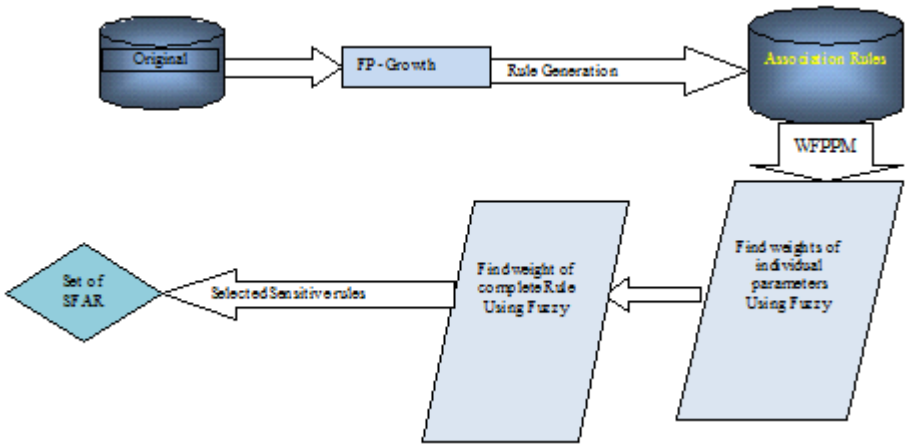
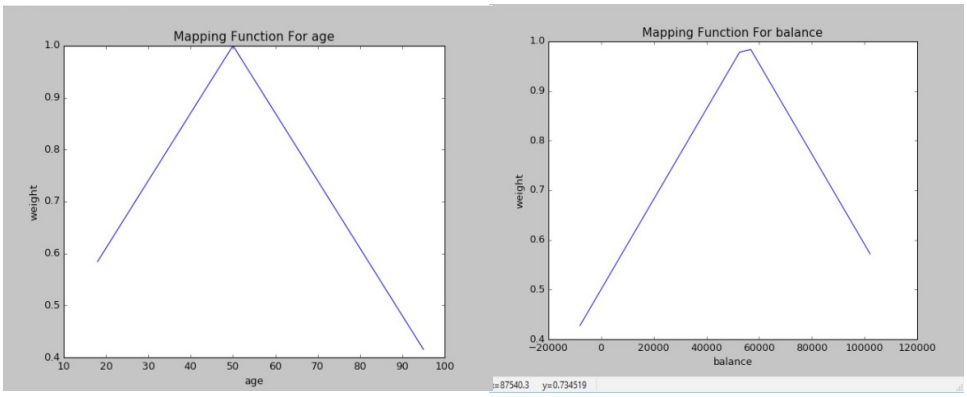


Figure 2. Flow of WFPPM.

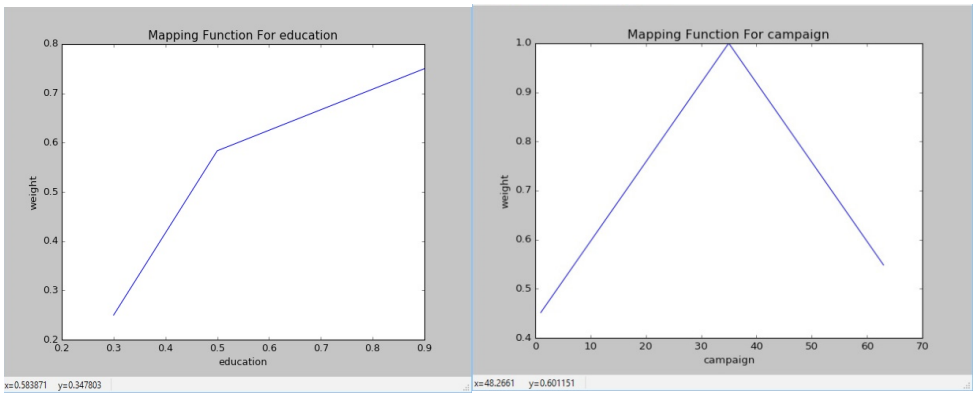
calculating the weight of each parameter, we will find the weight of the rule, by averaging the weight of each parameter.

- After calculating the weight of all rules, next step is to find the sensitive rules; we have split the rules in the range of highly sensitive, sensitive, average and low based on the weights of the rules. So instead of defining a strict threshold to define sensitive rules we have used fuzzy logic to categorize the rules over a range on basis of their weight. Figure 2 summarizes the process for identification of sensitive items.



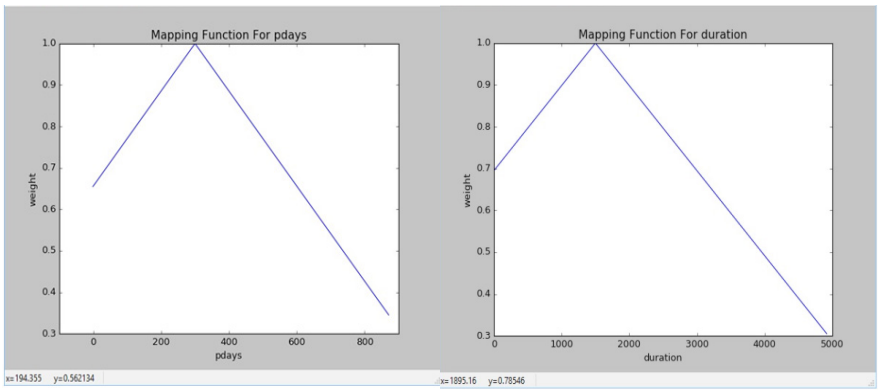
(a) (b)

Figure 3. (a). Mapping function of age. (b). Mapping function of balance.



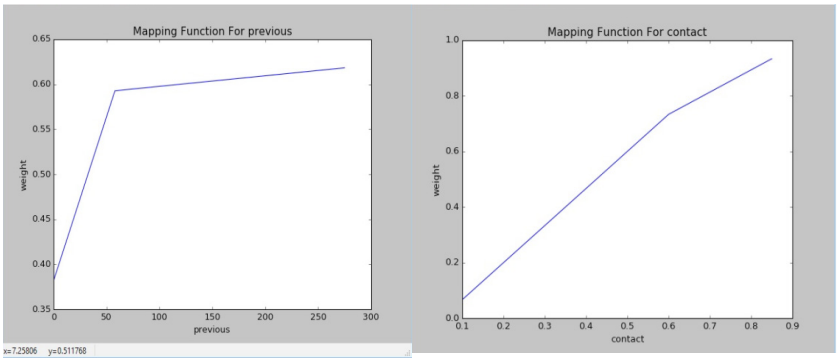
(c) (d)

Figure 3. (c). Mapping function of education. (d). Mapping function of campaign.



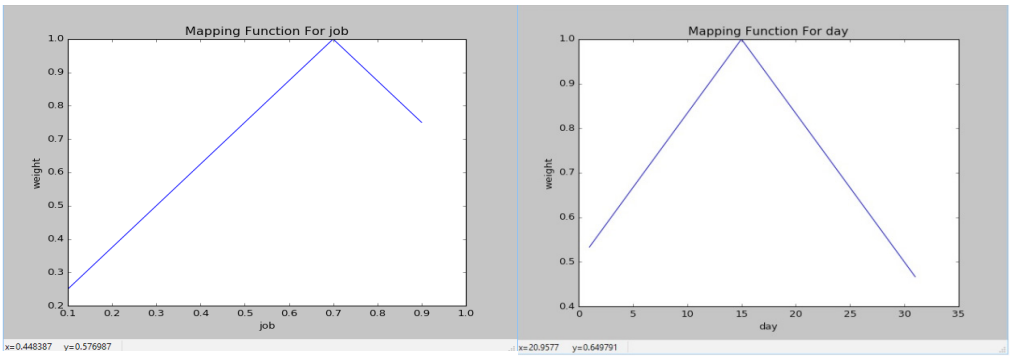
(e) (f)

Figure 3. (e). Mapping function of pdays. (f). Mapping function of duration.



(g) (h)

Figure 3. (g). Mapping function of previous. (h). Mapping function of contact.



(i) (j)

Figure 3. (i). Mapping function of job. (j). Mapping function of day.

4.2 Fuzzy membership function of individual parameter

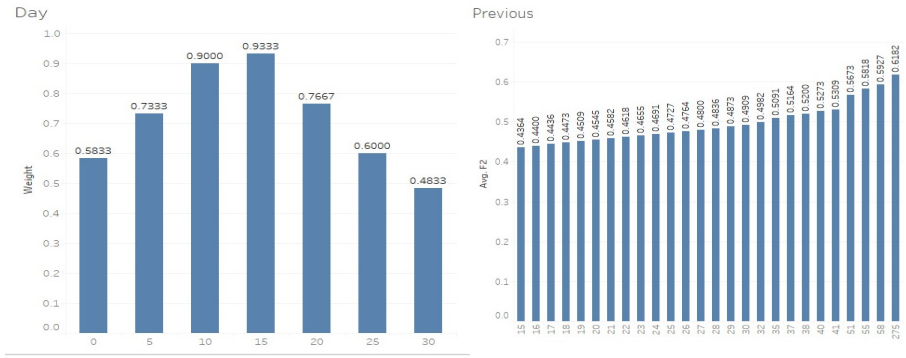
Following are given the membership functions of each parameter in the database which helps in deciding the sensitivity of the attribute. We applied these membership functions to fuzzify the values.

(a) (b)

Figure 4. (a). Graph of weight of Balance attribute for all possible values. (b). Graph of weight of Age attribute for all possible values.

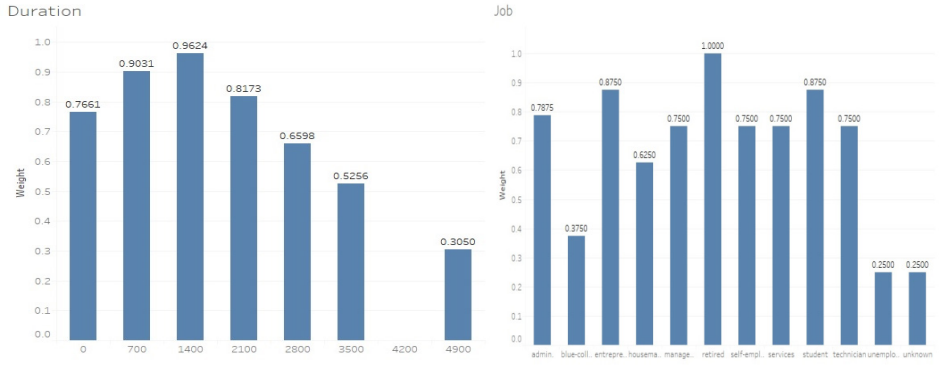
(c) (d)

Figure 4. (c). Graph of weight of campaign attribute for all possible values. (d) Graph of weight of contact attribute for all possible values.



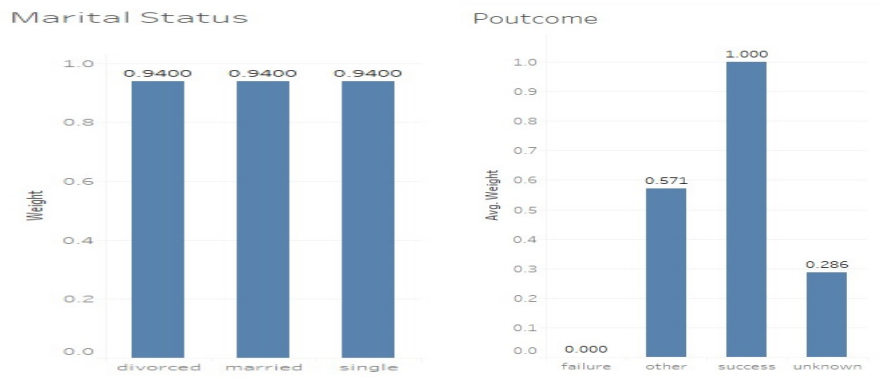
(e) (f)

Figure 4. (e). Graph of weight of Day attribute for all possible values. (f). Graph of weight of Previous attribute for all possible values.



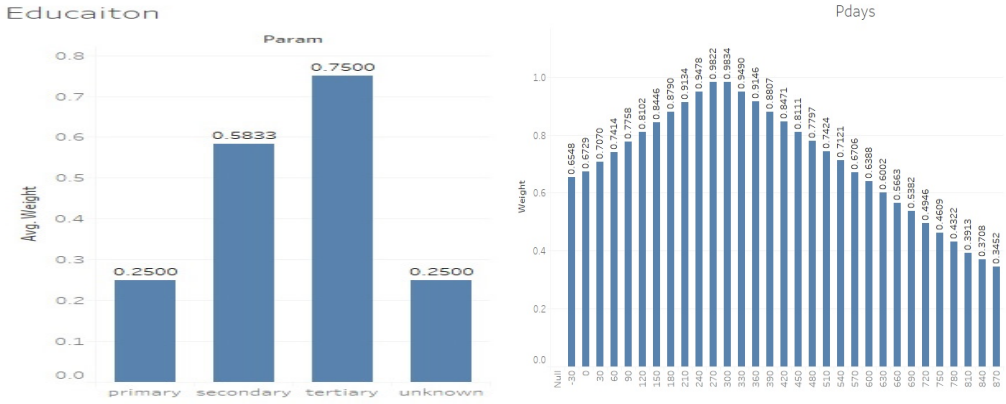
(g) (h)

Figure 4. (g). Graph of weight of Duration attribute for all possible values. (h). Graph of weight of job attribute for all possible values.



(i) (j)

Figure 4. (i). Graph of weight of martial status attribute for all possible values. (j). Graph of weight of Poutcome attribute for all possible values.



(k) (l)

Figure 4. (k). Graph of weight of Education attribute for all possible values. (l). Graph of weight of Pdays attribute for all possible values.

5. Experimental Results

In this section experimental results are given in which weights of the individual parameters are calculated using WFPPM which helped in calculating the weight of a complete rule present in the database. Further these rules have been divided in the range of highly sensitive, sensitive, average and low based on the weights of the rules to find sensitive association rules (SAR). So instead of defining a strict threshold to define sensitive rules we have used fuzzy logic to categorize the rules over a range on the basis of their weight.

6. Conclusion

In this paper, we have presented a novel framework for extracting sensitive fuzzy association rules from "frequent items" with quantitative properties (sub itemsets) using weighted fuzzy sets. The WFPPM algorithm produces a more succinct set of fuzzy association rules using fuzzy measures and weight as the interestingness (certainty) measure and thus presents a new way for extracting sensitive association rules from items with properties. This is different from normal quantitative ARM. Experimental results show that the proposed scheme extracts sensitive rules from the data set by considering the weight of individual parameter rather than depending upon the minimum threshold value of support and confidence. Largely, WFPPM offers potential to apply this framework in varied domains like banking sector or other financial organization where sensitivity of the rules depends upon the value of the parameter rather than its frequency of occurrence in a database. This technique helps the researcher to efficiently extract the sensitive rules which are further to be hidden by using one of the association rules hiding technique for privacy purpose.

7. Acknowledgment

Authors of this paper pay gratitude toward the entire team of RIC of IK Gujral Punjab Technical University (PTU) Jalandhar to provide an opportunity to do research in this area. The product of this research paper would not be possible without all of them.

8. References

1. Karthikeyan T, Chellathurai S, Praburaj B. A study on a novel method of mining fuzzy association using fuzzy correlation analysis. African Journal of Mathematics and Computer Science Research. 2012 Jan 15; 5(2):28–33. https://doi.org/10.5897/AJMCSR11.157
2. Lin NP, Chueh HE. Fuzzy correlation rules mining. In: Proceedings of the 6th WSEAS International Conference on Applied Computer Science. 2007 Apr 15. p. 13–8.
3. Wang CH, Lee WH, Pang CT. Applying fuzzy FP-Growth to mine fuzzy association rules. World Academy of Science, Engineering and Technology. 2010 May 26; 65:956–62.
4. Suresh H, Raimond K. Mining association rules from time series data using hybrid approaches. Proceedings of International Journal of Computational Engineering Research. 2013; 3(3):181–9.
5. Mangalampalli A, Pudi V. Fuzzy association rule mining algorithm for fast and efficient performance on very large datasets. IEEE International Conference on Fuzzy Systems, 2009. FUZZ-IEEE 2009. 2009 Aug 20. p. 1163–8. https://doi.org/10.1109/fuzzy.2009.5277060
6. Florez G, Bridges SA, Vaughn RB. An improved algorithm for fuzzy data mining for intrusion detection. In: Fuzzy Information Processing Society, 2002. Proceedings. NAFIPS. 2002 Annual Meeting of the North American, IEEE. 2002; 457–62. https://doi.org/10.1109/nafips.2002.1018103
7. Bhatla N, Jyoti K. A Novel Approach for heart disease diagnosis using Data Mining and Fuzzy logic. International Journal of Computer Applications. 2012 Jan 1; 54(17). https://doi.org/10.5120/8658-2498
8. Hong TP, Kuo CS, Wang SL. A fuzzy AprioriTid mining algorithm with reduced computational time. Applied Soft Computing. 2004 Dec 31; 5(1):1–10. https://doi.org/10.1016/j.asoc.2004.03.009
9. Ebrahimzadeh A, Sheibani R. Two efficient algorithms for mining fuzzy association rules. International Journal of Machine Learning and Computing. 2011 Dec 1; 1(5):510. https://doi.org/10.7763/IJMLC.2011.V1.76
10. Ebrahimzadeh A. Fuzzy Association Rules: new method and implementation. 2014; 5(1):26–31.
11. Mukherjee S, Chen Z, Gangopadhyay A. A fuzzy programming approach for data reduction and privacy in distance-based mining. International Journal of Information and Computer Security. 2008 Jan 1; 2(1):27–47. https://doi.org/10.1504/IJICS.2008.016820
12. Available from: http://archives .ics.uci.edu/ml/
13. Available from: http://sci2s.ugr.es/keel/datasets.php
14. Moro S, Laureano R, Cortez P. Using data mining for bank direct marketing: An application of the crisp-dm methodology. In: Proceedings of European Simulation and Modelling Conference-ESM'. Eurosis. 2011. p. 117–21.
15. Han J, Pei J, Kamber M. Data mining: concepts and techniques. Elsevier. 2011 Jun 9.
16. Zadeh LA, Fu KS, Tanaka K. Fuzzy sets and their applications to cognitive and decision processes: Proceedings of the US–Japan seminar on fuzzy sets and their applications, held at the University of California, Berkeley, July 1-4, 1974. Academic press. 2014 Jun 28.
17. Zadeh LA. Fuzzy logic. Computer. 1988 Apr; 21(4):83–93. https://doi.org/10.1109/2.53
18. Yager RR, Zadeh LA. An introduction to fuzzy logic applications in intelligent systems. Springer Science and Business Media. 2012 Dec 6; 165

Citation:

Meenakshi Bansal, Dinesh Grover and Dhiraj Sharma
"**Sensitivity Association Rule Mining using Weight based Fuzzy Logic**",
Global Journal of Enterprise Information System. Volume-9, Issue-2, April-June, 2017. (http://informaticsjournals.com/index.php/gjeis)

Conflict of Interest:

Author of a Paper had no conflict neither financially nor academically.