Association Rule Discovery for Customer Relationship Management Using Genetic Algorithm

Maziyar Grami [1] , Reza Gheibi [2,*] , Fakhereh Rahimi [3]
[1,2,3] Department of Computer, Technical and Engineering College, Kermanshah Branch, Islamic Azad University, Kermanshah, Iran.
* gheibi.reza@iauksh.ac.ir

Abstract— **Today, development of internet causes a fast growth of internet shops and retailers and makes them as a main marketing channel. This kind of marketing generates a numerous transaction and data which are potentially valuable. Using data mining is an alternative to discover frequent patterns and association rules from datasets. In this paper, we use datamining techniques for discovering frequent customers' buying patterns from a Customer Relationship Management database. There are lots of algorithms for this purpose, such as Apriori and FP-Growth. However, they may not have efficient performance when the data is big, therefore various meta-heuristic methods can be an alternative. In this paper we first excerpt loyal customers by using RFM criterion to face more reliable answers and create relevant dataset. Then association rules are discovered using proposed genetic algorithm. The results showed that our proposed approach is more efficient and have some distinction in compare with other methods mentioned in this research.**

Keywords: Feature selection; Genetic algorithm; Data mining; Association rules

I. INTRODUCTION

Internet technology has led to many competitive advantages and it makes customers to search various services and products to satisfy their requirements in less time. Customers purchase through the internet from online shops which are a main market channel. After 2002, many online shops appeared and each of them had its own features. The appearance of this shops leads to reduce costs for services and products [1]. These online shops generate plenty of data which known as customer knowledge. Extracting information and pattern recognition from this knowledge is useful and increases profitability for these online shops. In this area, one important challenge is knowledge management and here we are concerned about the term customer relationship management which is a kind of information system that makes organizations able to collect and store customer's data [2]. To utilize this information system, usually data mining techniques, which are one of the main cores of knowledge discovery, are used [3]. An important topic which is widely used in data mining is association rules. Association rule mining is a method for discovering interesting relations between variables in large databases. These rules are used to discover the relations between data in a dataset. Formally, these rules are used to analysis shopping cart.

For this purpose, some algorithms such as Apriori and FP-Growth are presented which are applicable in binary databases. To prepare the database, each initial record should alternate to binary representation which called Item. However usually, those algorithms are efficient enough but in some situations, they may be sluggish to attain the answer. Therefore, numerous heuristic and meta-heuristic algorithms are proposed in previous literatures. Various meta-heuristic algorithms such as Ant-Colony, Particle Swarm and Genetic algorithms are presented for solving this problem [4-8]. These algorithms provide efficient answers in less time and they are usually useful where obtaining the answer is not binomial. In this research, we use genetic algorithm to discover association rules on a customer relationship management dataset. But we concentrate on CRM database and its features.

The rest of this paper is as follows. Section (II), explains related works. Implementation details of the proposed method are presented in Section (III). Finally, results will conclude in Section (IV).

II. RELATED WORK

In this section, we briefly explain previous researches and works done in this area. Data mining was first introduced in 1990 and later appeared as a powerful tool to discover unknown patterns from huge datasets [9-11]. Data mining can be divide into three main categories as clustering, classifications and association rules mining. Recently, association rules and especially discovering frequent patterns has attracted more and more attention among data mining techniques.

There are a lot of literatures that used association rules mining to solve a problem. Wang et al., used association rules to analyze dataset of sequence corrosion in chemical process [12]. They introduced a new fuzzy method to discover association rules and sequence corrosion as well. Parkinson et al., utilize association rules in order to study and monitor authentication in system files [13]. Their approach consisted of two phases, analysis the NTFS to discover access allowance and utilizing association rules to detect illegal accesses. In another research, Ivancevic et al. detected risk factors for early childhood caries [14]. They created a dataset include of %10 of children aged less than 7 years old in Serbia and detected risk factores by using association rules. Kargarfard et al., used a

combination of classification and association rules in a dataset consist of seven thousand records, to detect influenza disease [15]. They beleived that combining classificaion algorithms and association rules can lead to development of decision support and expert systems.

In some other researches, authors have tried to use meta-heuristic algorithms for mining association rules. Martin et al., presented a new meta heuristic method based on genetic algorithm named NICGAR which refers to Niching Genetic Algorithm [16]. Their method achieves more rules in less time. Cheng et al., proposed a method to discover useful knowledge from past history and defects for construction managers [17]. Their approach utilized genetic algorithm and after experiment, results showed rational relationship between discovered factors and existed defects. Kuo and Shih used Ant Colony System for discovering association rules from a healthcare insurance dataset of Taiwan. They put some multi-dimensional constraint on this big dataset. Their results showed that their proposed method have better performance in compare to Apriori algorithm [6]. Bhugra et al. used Biogeography-Based Optimization (BBO) algorithm for association rule mining. They had changed migration part of this algorithm and their result showed a good performance [18]. Kou et al. used Particle Swarm Optimization (PSO) algorithm for association rule mining. The tested their proposed model on dataset from Microsoft and compared it to Genetic algorithm. Their results showed that their proposed algorithm obtain better discovered rules in compare with Genetic algorithm [19]. Djenouri et al., combined Artificial Bee Colony optimization algorithm and Tabu Search for mining association rules. They called their algorithm as HBO-TS in which Artificial Bee Colony was used for creating diversity and Tabu Search to search fast. Their results showed that although this method had some difficulties in parameters settings, but it can create good rules in an acceptable time [20].

III. PROPOSED METHOD

In this section, we explain our approach. This method consists of two main phases. The first phase is, pre-processing data, which dedicates the way to collect and prepare data and the second phase which explains the proposed genetic algorithm and its implementation.

A. *Creating binary dataset*

In this study the main challenge is accessing online shop's data and especially a binary dataset that it's not easily possible and therefore in this paper we used a transactional database of an online shop at (*http://www.community.tableau.com*) and then convert it to a binary database which is explained further (in subsection D). This database consists of 8400 transactions of 795 customers about their buying information in an online shop.

B. *Collect and prepare data*

For the aim of crating the binary dataset, below steps should be done:

- **Seeking dataset:** as mentioned formerly, the data are collected from "*http://www.community.tableau.com*". This dataset is transactional and since association rule mining algorithms require binary dataset, so collected dataset should be changed to binary.
- **Preparing data:** because of high data value in databases, they may have lots of uncorrected and inconsistent data. And this can effect on data mining results. So a preprocessing phase includes filtering and data reduction should be done first. This process is done to clean noisy data, remove outliers, detect important and effective features in order to reduce the volume of data and so on.
- **Creating dataset according to customers:** as mentioned before, the dataset is transactional and it have more than 8400 transactions from 795 customers. But as this dataset is sparse, we created a dataset in which transaction of each customer is merged as a single record. So the final dataset in this step is a dataset with 795 rows (number of customers).

C. *Identifying loyal customers*

To get better results, customers' value is computed as a parameter and then loyal customers are detected based on this parameter. Therefore a process of identifying loyal customers should be done. For this aim let's extract RFM measurement first. RFM is a criterion for identifying customers' value. It stands for Regency, Frequency and Monetary, i.e. how recently did the customer purchase, how often do they purchase and how much do they spend. After that a binary dataset is created. R, F and M are computed as shown in equation 1-4:

$$R_N = \frac{(R - R_{min})}{R_{max} - R_{min}} \quad (1)$$

$$F_N = \frac{(NumberOfTransactions_i - NumberOfTransactions_{min})}{NumberOfTransactions_{max} - NumberOfTransactions_{min}} \quad (2)$$

$$M_N = \frac{(Sales_i - Sales_{min})}{Sales_{max} - Sales_{min}} \quad (3)$$

Then, RFM customer value is calculated as below:

$$RFM = \frac{(R + F + M)}{3} \quad (4)$$

Finally, a column for the criterion RFM is added for each customer in the database.

- **Extracting loyal customers:** for extracting loyal customers, first we order customers according to RFM value. Then we split dataset into two parts using a threshold. The upper part includes loyal customers and lower consist the others. The threshold is obtained by trial and error, i.e. We change this threshold several time and we use Neural Network classifier to test the correctness of this threshold. After several thresholds

and classifying we found the best threshold for identifying loyal customers.

D. Creating binary dataset

As said formerly, we need to convert this transactional database to a binary database. For this aim, we use SQL commands in SQL Server. In this way, since we faced numerous items (products), so we put a row for each product category which is initialized with 'true' (i.e. the customer had bought an item) or 'false' (i.e. the customer did not bought an item).

E. Collecting results

In the final step, the created data in previous steps is collected and the existed binary database is ready to be discovered by the proposed genetic algorithm which is explained in detail in next section.

IV. IMPLEMENTATION THE PROPOSED METHOD USING GENETIC ALGORITHM

We used genetic algorithm (GA) for mining association rules from the prepared database. The genetic algorithm is a heuristic (sometimes called metaheuristic) which utilized to solve optimization problems by using approaches inspired from nature. This algorithm works as below:

1. ***Begin***
2. *Choose initial population*
3. ***Repeat***
4. *Evaluate the individual fitness of a certain proportion of the population*
5. *Select pairs of best-ranking individuals to reproduce*
6. *Apply crossover operator*
7. *Apply mutation operator*
8. *until terminating condition*
9. ***End***

Here every chromosome is defined as a row of this database. So a population or a chromosome consists of a binary array of items which consists of 'One' (i.e. the customer purchased that item) and 'Zero' (i.e. the customer didn't purchase that item).

Therefore, first we need to create an initial population for genetic algorithm which is done by using the Apriori algorithm. We define objective function as 'Support'; 'Support' is defined as the proportion of transactions in the database which contains corresponding item-set. Also termination condition is obtained as several iterations. The proposed genetic algorithm is shown in Fig .1

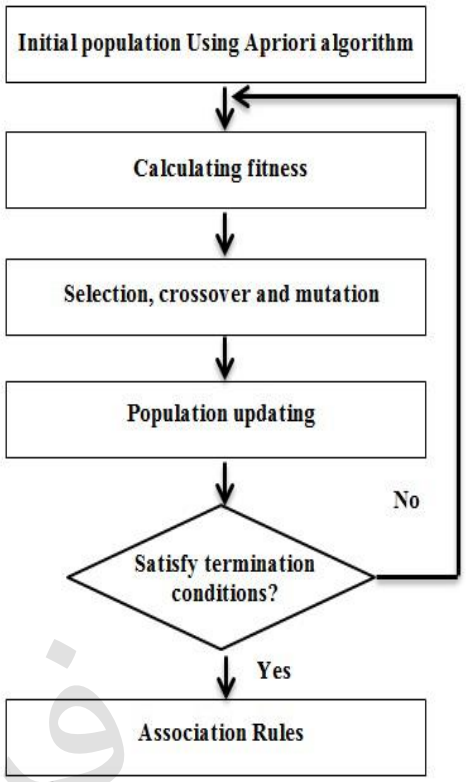
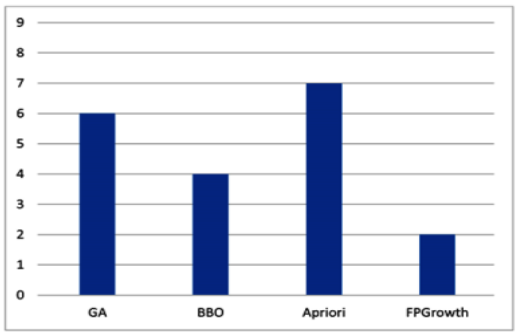


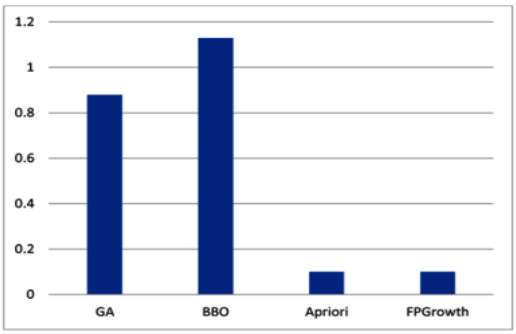
Figure 1. Proposed Genetic Algorithm

Our model is implemented using MATLAB software, version 2015b, by a classical computer with 4x300 Gigahertz CPU and 32 GB ram equipped.

Results shows that genetic algorithm is more useful and faster in discovering association rules in a transactional database and it also reach the answer in less repetition. It obtains better answers in compare with other algorithms like Apriori and FP-Growth. We have compared our model with FP-Growth, Apriori an BBO meta-heuristic algorithm. As we can see in the Fig 2.a, GA created better rules than BBO and FP-Growth, but it couldn't conquer Apriori. Time consumption for creating model using Apriori FP-Growth algorithms is very low and GA time consumption is better than BBO (Fig 2.b). Although both FP-Growth and Apriori algorithms are successful in some situations, but they use a lot of memory. According to this experiment, our method has a good performance in compare with BBO, Apriori and FP-Growth.



a) Number of rules with more than 60% Support



b) Time cnsumtion for creating models

Figure 2. Genetic algorithm in compare with BBO, Apriori and FP-Growth

V. CONCLUSION

In this paper we utilized Genetic Algorithm, for the purpose of mining association rules from the prepared database. We defined every chromosome as a row of this database. Results showed that genetic algorithm is more useful and faster in discovering association rules in a transactional database. In addition genetic algorithm reaches the answer in less repetition and it obtains better answers in compare with other algorithms like Apriori and FP-Growth. We have compared our model with FP-Growth, Apriori an BBO meta-heuristic algorithm. Genetic algorithm created better rules in compare with BBO and FP-Growth, but it couldn't conquer Apriori. Time consumption for creating model using Apriori FP-Growth algorithms is very low and GA time consumption is better than BBO Although FP-Growth and Apriori algorithms are successful in some situations, but they use a lot of memory. Based on this experiment, our method showed better performance in compare with BBO, Apriori and FP-Growth.

ACKNOWLEDGMENT

The authors would like to thank reviewers and the editor for their helpful discussions about the topic. This work is supported by Islamic Azad University, Kermanshah Branch, Kermanshah, Iran.

REFERENCES

[1] Shim, B., K. Choi, and Y. Suh, CRM strategies for a small-sized online shopping mall based on association rules and sequential patterns. Expert Systems with Applications, 2012. 39(9): p. 7736-7742.

[2] Khodakarami, F. and Y.E. Chan, Exploring the role of customer relationship management (CRM) systems in customer knowledge creation. Information & Management, 2014. 51(1): p. 27-42.

[3] Ming-Syan, C., H. Jiawei, and P.S. Yu, Data mining: an overview from a database perspective. IEEE Transactions on Knowledge and Data Engineering, 1996. 8(6): p. 866-883.

[4] Sarath, K.N.V.D. and V. Ravi, Association rule mining using binary particle swarm optimization. Engineering Applications of Artificial Intelligence, 2013. 26(8): p. 1832-1840.

[5] Srinivasan, S. and S. Ramakrishnan, Evolutionary multi objective optimization for rule mining: a review. Artificial Intelligence Review, 2011. 36(3): p. 205-248.

[6] Kuo, R.J. and C.W. Shih, Association rule mining through the ant colony system for National Health Insurance Research Database in Taiwan. Computers & Mathematics with Applications, 2007. 54(11-12): p. 1303-1318.

[7] Minaei-Bidgoli, B., R. Barmaki, and M. Nasiri, Mining numerical association rules via multi-objective genetic algorithms. Information Sciences, 2013. 233: p. 15-24.

[8] Ribeiro, M.H., A. Plastino, and S.L. Martins, Hybridization of GRASP Metaheuristic with Data Mining Techniques. Journal of Mathematical Modelling and Algorithms, 2005. 5(1): p. 23-41.

[9] Tomar, D. and S. Agarwal, A survey on Data Mining approaches for Healthcare. International Journal of Bio-Science and Bio-Technology, 2013. 5(5): p. 241-266.

[10] Muhammed, L.-N. Using data mining technique to diagnosis heart disease. in Statistics in Science, Business, and Engineering (ICSSBE), 2012 International Conference on. 2012. IEEE.

[11] Shouman, M., T. Turner, and R. Stocker. Using data mining techniques in heart disease diagnosis and treatment. in Electronics, Communications and Computers (JEC-ECC), 2012 Japan-Egypt Conference on. 2012. IEEE.

[12] Wang, J., et al., Association rules mining based analysis of consequential alarm sequences in chemical processes. Journal of Loss Prevention in the Process Industries, 2016. 41: p. 178-185.

[13] Parkinson, S., V. Somaraki, and R. Ward, Auditing file system permissions using association rule mining. Expert Systems with Applications, 2016. 55: p. 274-283.

[14] Ivancevic, V., et al., Using association rule mining to identify risk factors for early childhood caries. Comput Methods Programs Biomed, 2015. 122(2): p. 175-81.

[15] Kargarfard, F., A. Sami, and E. Ebrahimie, Knowledge discovery and sequence-based prediction of pandemic influenza using an integrated classification and association rule mining (CBA) algorithm. J Biomed Inform, 2015.

[16] Martín, D., et al., NICGAR: A Niching Genetic Algorithm to mine a diverse set of interesting quantitative association rules. Information Sciences, 2016. 355-356: p. 208-228.

[17] Cheng, Y., W.-d. Yu, and Q. Li, GA‐based multi-level association rule mining approach for defect analysis in the construction industry. Automation in Construction, 2015. 51: p. 78-91.

[18] Bhugra, Divya, Vipul Singhania, and Shivani Goel. "Association rule analysis using biogeography based optimization." Computer Communication and Informatics (ICCCI), 2013 International Conference on. IEEE, 2013.

[19] Kuo, Ren Jie, Chie Min Chao, and Y. T. Chiu. "Application of particle swarm optimization to association rule mining." Applied Soft Computing 11.1 (2011): 326-336.

[20] Djenouri, Youcef, Habiba Drias, and Amine Chemchem. "A hybrid Bees Swarm Optimization and Tabu Search algorithm for Association rule mining."Nature and Biologically Inspired Computing (NaBIC), 2013 World Congress on. IEEE, 2013.