International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Review of Computer Engineering Studies
Vol. 8, No. 3, September, 2021, pp. 65-70
Journal homepage: http://iieta.org/journals/rces

Improving Association Rule Mining Using Clustering-Based Data Mining Model for Traffic Accidents

Mohamad Mohamad Shamie[*], Muhammad Mazen Almustafa

Department of Master Web Science, Syrian Virtual University, Damascus 35329, Syria

Corresponding Author Email: mohamad_70397@svuonline.org

https://doi.org/10.18280/rces.080301

Received: 12 August 2021
Accepted: 15 September 2021

Keywords:
data mining, association rules, clustering, knowledge discovery

ABSTRACT

Data mining is a process of knowledge discovery to extract the interesting, previously unknown, potentially useful, and nontrivial patterns from large data sets. Currently, there is an increasing interest in data mining in traffic accidents, which makes it a growing new research community. A large number of traffic accidents in recent years have generated large amounts of traffic accident data. The mining algorithms had a great role in determining the causes of these accidents, especially the association rule algorithms. One challenging problem in data mining is effective association rules mining with the huge transactional databases, many efforts have been made to propose and improve association rules mining methods. In the paper, we use the RapidMiner application to design a process that can generate association rules based on clustering algorithms.

1. INTRODUCTION

The concept of data mining was first introduced at the ACM conference in the United States in 1995. Data mining is also known as KDD, it refers to mining a large amount of data from the database to uncover implicit, unknown, potentially useful information [1, 2]. Association rule mining is an important branch of data mining research. Association rule mining is used to find a relation hidden in the interest of large datasets, one of the most important applications of data mining is association rules discovery. The goal of association rule mining is to find the relevant elements that satisfy the minimum support and confidence [3].

The purpose of this paper is to address the generation of association rules for big data by proposing a methodology that uses the RapidMiner application that will generate the association rules after using clustering algorithms.

In this paper, we will discuss the algorithm K-Means, FP - Growth and its use to find better rules to help identify common causes of traffic accidents. We will generate association rules based on the K-Means algorithm to cluster data by each cluster and then generate a data table for each cluster using the RapidMiner application.

We used the UK dataset for our study. The Accidents Dataset contains all accidents on public roads between 2005 and 2015 [4]. The dataset was downloaded from the UK Department for Transport. The dataset contains 22,046 records and about 7 important attributes, as shown in Table 1.

Table 1. UK data attributes

Road_Type	Time
Age of Driver	Gender
Age of Vehicle	Light Conditions
Accident Severity	

2. ASSOCIATIOON RULE AND FP-GROWTH ALGORITHM

2.1 Association rule

Association rule mining is an important branch of data mining research. Association rule mining used to find an association hidden in the interest of large data sets.

The topic of mining association rules was first proposed by Agrawal et al. [3] and others in 1993. The association rules mining process mainly consists of two steps [5]. In the first step, all items whose support is greater than the minimum support are found in the transaction database, i.e. all frequent itemsets. In the second step, the expected association rules are generated from these frequent itemsets.

The principle of the search for association rules is based on three basic factors: support, confidence, and lift. Its goal is to reduce the very large number of association rules generated, some of which are not important [6].

Support [7]: it is the number of transactions that contain elements in the {X} and {Y} parts of the rule, as a percentage of the total number of transactions, it is defined by the following rule:

$$Support = \frac{Count(X\ U\ Y)}{N}$$

Confidence [7]: Measures how often each element in Y occurs in transactions that also contain elements in X and is defined by the following rule:

$$Confidence = \frac{Count(X\ U\ Y)}{Count(X)}$$

Lift [7]: It is a measure of the importance of the rule and is defined by the following rule:

$$Lift(X \rightarrow Y) = \frac{Support(X\ U\ Y)}{Support(X) * Support(Y)}$$

Based on the previous rules, association rules are generated that achieve a support value greater than a threshold called minsup, and the user defines a confidence value greater than a threshold called minconf.

2.2 FP-Growth algorithm

It is one of the most important algorithms in the field of association rules, which is a further development of the Apriori algorithm. The resulting tree structure aims to store compressed and important information about recurrent data to develop an effective FP -growth tree for recurrent models [6]. This algorithm differs from Apriori algorithm in that it reduces the size and time required to generate the rules [8].

The FP-growth algorithm scans the database only twice. FP -growth algorithm is a depth-first search algorithm combined with direct counting and uses the recursive strategy of pattern growth. It does not need to generate candidate sets, instead the transaction database is compressed into a tree structure that stores only the frequent elements. All frequent elements are obtained by recursive mining the FP tree [6].

3. RELATED WORKS

Wang and others from Tongji University presented a study entitled: Research on Automated Modeling Algorithm Using Association Rules for Traffic Accidents [9].

They pointed out in the study that when association rules are used in other areas, such as sales, the focus is on maximizing support and confidence. However, in the area of analyzing traffic accident data, things are quite different because, for example, the frequency of serious accidents is very low, but the damage is high. If we want to extract the relationships that lead to serious accidents, we must also consider the cases where support and confidence are low. The research aimed to define or establish specific criteria to determine the minimum support and the minimum confidence. The research followed the method of clustering using the K-Means algorithm to cluster the rules for several categories (weather conditions, speed limits, day or night, road type) and then select a certain number of bases from each class based on the rules with the highest support.

After applying the Apriori algorithm, they found that the most interesting of the rules where the value of confidence is high have low support, but also most of the rules where the value of lift is high have high support. They chose the rules according to the order of lift from highest to lowest, basing some of the rules on expert opinions about traffic accidents. What makes this study stand out, regardless of the type of algorithm used, is the focus on how the value of minsup and minconf is determined for this type of research.

Simon et al. presented a study entitled Discovery of Breast Cancer-Causing Factors Using Association Rules [10]. It aimed to discover the factors that cause breast cancer by comparing the characteristics of patients with breast cancer and those who do not have the disease at all. They used the logit model to select the factors that affect the probability of developing breast cancer. At the beginning, the support was determined at 30% and confidence at 80%, and no rules were determined for patients with breast cancer. After several attempts to change the value of support and confidence, the value of support was changed to 0.001% and the confidence value was set to 90%, obtaining 67 rules in which the characteristics of patients with the disease are.

Our research can benefit from this study by applying narrow values of the above support and confidence to obtain rules whose incidence is very high, but considering that in return we lose some rules that can be useful in detecting some types of rare accidents. The most important thing that stands out in this study is the researcher's focus on extracting rules whose incidence is very large due to the high confidence level.

Soniya Mudgal [11] presented a study: Mining the Correlations for Fatal Road Accident using Graph-based Fuzzified FP - Growth Algorithm. The study aimed to extract the causes of fatal traffic accidents by using fatal accident data from the US government. The study had the result: The FP-Growth algorithm is faster than the Apriori algorithm, FP - Growth generated the association rules within 0.5658 seconds, in contrast Apriori algorithm took 3.2101 seconds to generate the same number of rules. The researchers concluded that the FP Growth algorithm is faster and more accurate than the Apriori algorithm.

4. METHODOLOGY

In this chapter, we will explain the methodology that aims to design a workflow whose task is to use data mining methods to access the factors that cause accidents. The problem with using algorithms for correlation rules for traffic accidents is always that there are rules for a certain classification of data that can be useful and valuable, and therefore cannot be observed in the number of a large number of extracted rules, which is the main obstacle for decision makers.

We used RapidMiner software to design the processes we will apply to the extracted data. RapidMiner is a data science software platform developed by the company of the same name that provides an integrated environment for data preparation, machine learning, deep learning, text mining, and predictive analytics. It is used for business and commercial applications as well as for research, education, training, rapid prototyping and application development and supports all steps of the machine learning process, including data preparation, visualization results, model validation and optimization.

4.1 Methodology design

Table 2. Time category

Category	From	To
Midnight	00:01	03:00
Dawn	03:00	04:00
Morning	05:01	9:00
Mid-morning	9:01	12:00
Noon	12:01	14:00
Afternoon	14:01	17:00
Evening	17:01	21:00
Night	21:01	00:00

Before we start designing the system, we will transform some data types to prepare the data:

1- The age of the driver and the age of the vehicle are converted from numerical form to textual form, so that each set of ages is converted to a textual form that expresses in decades. An example of this is to convert the age between 20 and 30 to the third decade and so on.

2- Convert the accident time into eight sections instead of the time value, as shown in Table 2.

Association rules are generated based on the K-Means algorithm cluster to separate data after each clustering and then generate a dataset for each clustering, and it is designed by RapidMiner application. Our proposed methodology will follow four basic steps:

1- Using the K-Means algorithm to extract clusters.

2- Extracting the data table for each cluster.

3- Traversing each table and extracting the association rules by the FP -Growth algorithm for each cluster.

4- Collecting the association rules in one place.

In this design, the following methodology was followed:

4.1.1 Preparing data as shown in Figure 1

a- Load dataset:

Retrieve traffic accidents data from the database.

b- Select Attributes:

Select the elements to determine the common attributes between the elements. The following elements have been selected which are among the most important elements: (severity of the accident, age of the driver, gender of the driver, time of the accident).

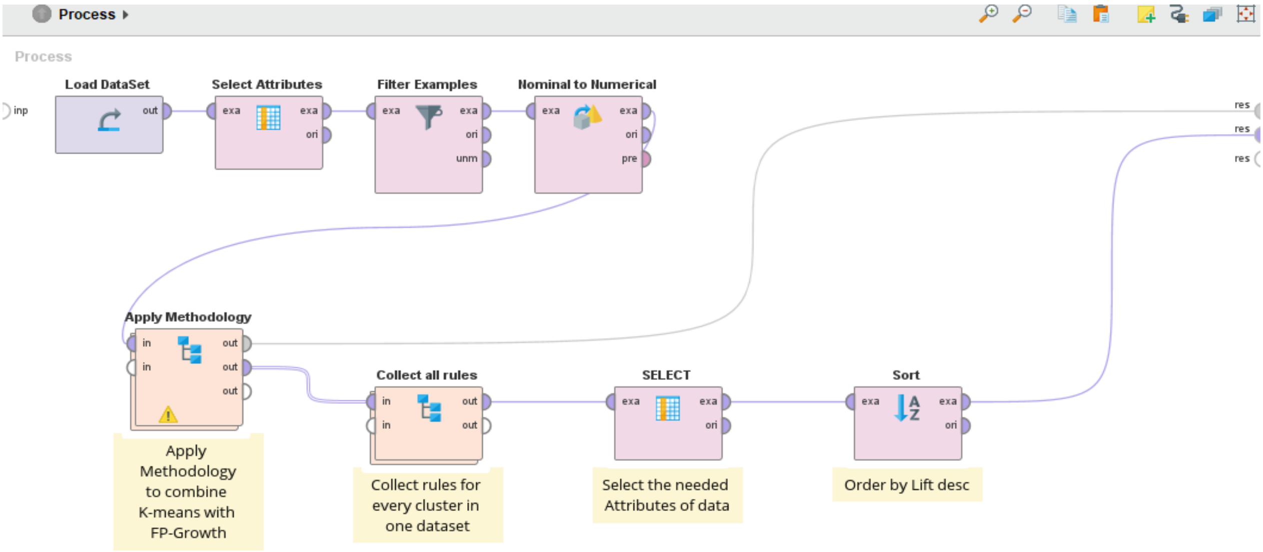


Figure 1. Initializing the data, applying the methodology

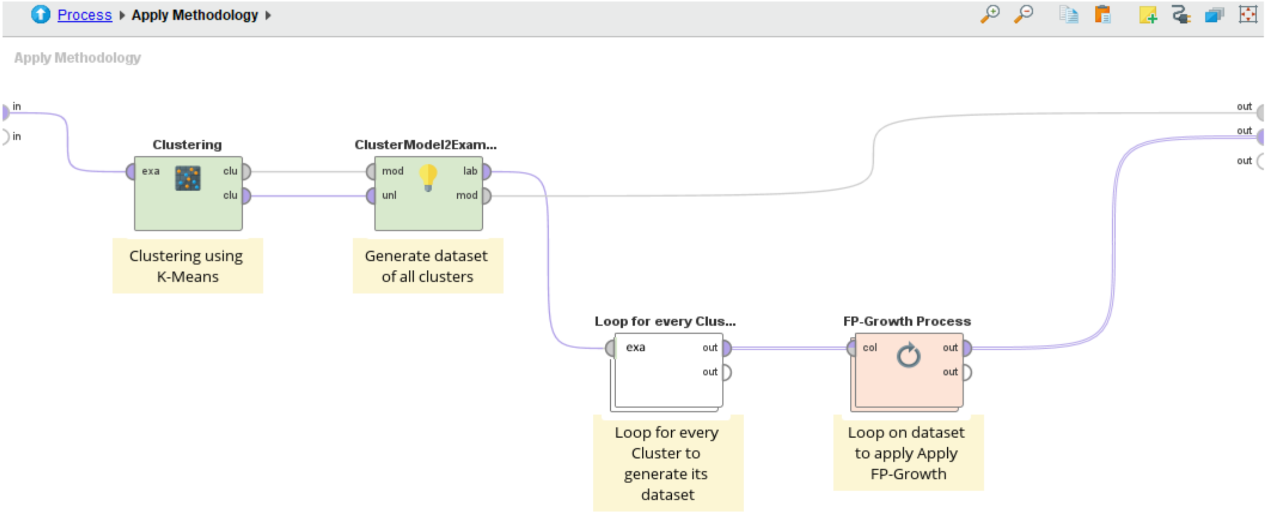


Figure 2. Illustrates the sub-process of applying clustering and generating data for each cluster

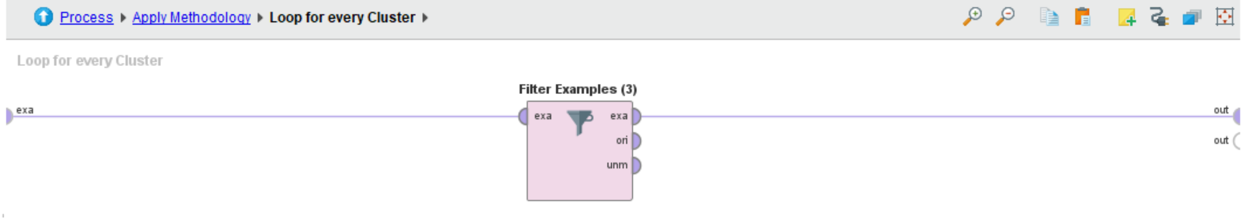


Figure 3. Filter clusters by cluster value

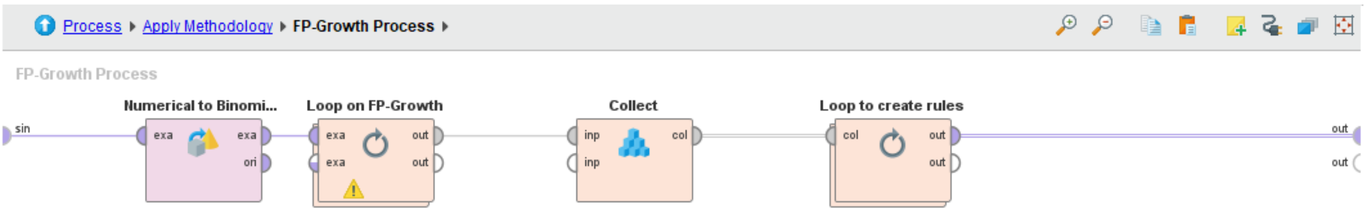
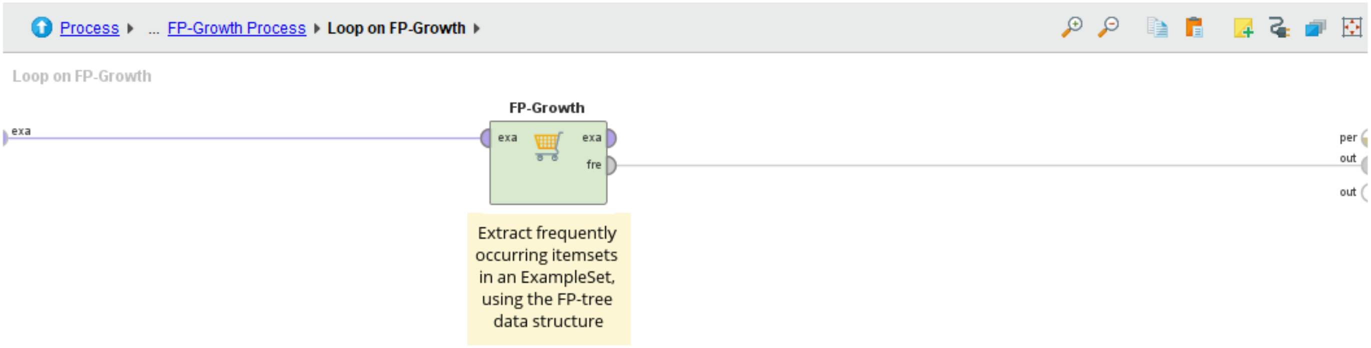
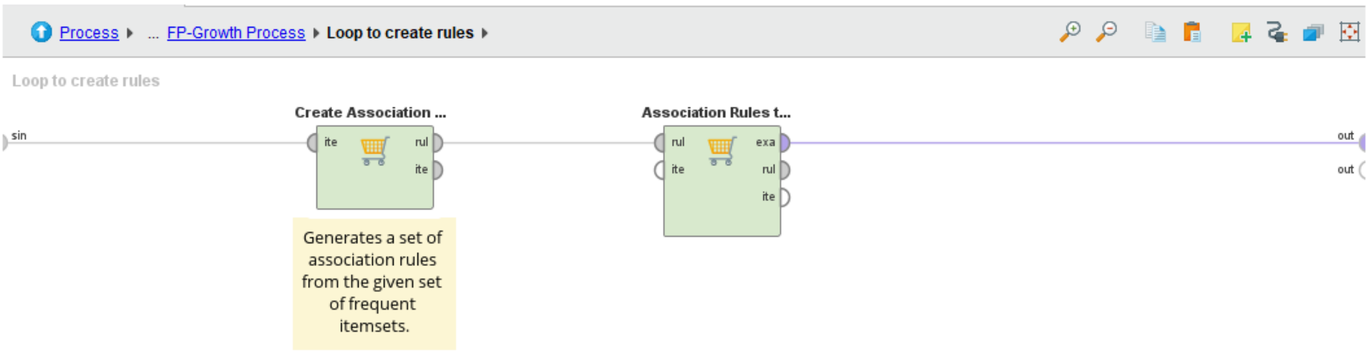


Figure 4. Applying FP-Growth to each cluster dataset



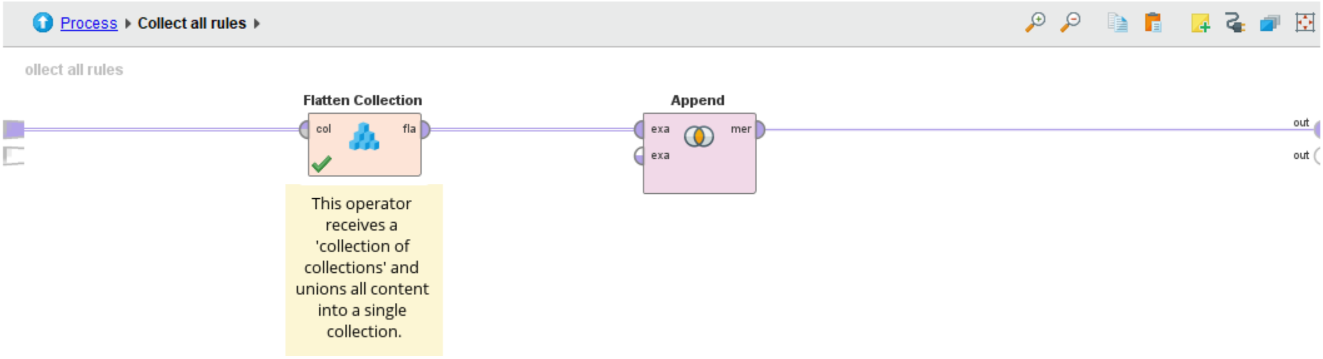
Extract frequently occurring itemsets in an ExampleSet, using the FP-tree data structure

Figure 5. Applying the FP -growth algorithm to each cluster dataset



Generates a set of association rules from the given set of frequent itemsets.

Figure 6. Create association rules



This operator receives a 'collection of collections' and unions all content into a single collection.

Figure 7. Collect rules in one dataset

c- Filter examples:

Filtering data and selecting fatal and serious accidents and their number 2941 rows.

d- Nominal to Numeric:

Converts the specified non-numeric values to a numeric type.

4.1.2 Apply methodology:

The process of applying the methodology:

a- Clustering using the K-Means algorithm:

Selection of the Euclidean distance and generation of data tables for each cluster as shown. Figures 2 and 3.

b- FP-Growth Process:

Passing through each data table of each cluster to perform the process of generating the association rules as shown in Figure 4.

c- FP-Growth:

Extract the repeated models using the FP-Growth algorithm to generate the FP-Tree growth tree with minsup=0.1 as shown in Figure 5.

d- Create Association Rule:

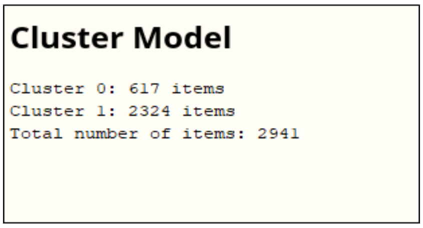
Generate association rules from the resulting tree, using the confidence scale to determine the required rule strength with minconf = 0.7, as shown in Figure 6.

e- Collect rules:

In this process, there is a for loop for each result of the generation association rule to collect all the rules in one dataset as shown in Figure 7.

4.2 Methodology result

By applying the previous design to the UK data by choosing the number of clusters K=2, 271 rules were generated, with two clusters generated, the first cluster containing the male incidents and the second cluster containing the female incidents. The clusters were generated within 1 second and the following results in Figure 8 were generated for the clusters:



Cluster Model

```
Cluster 0: 617 items
Cluster 1: 2324 items
Total number of items: 2941
```

Figure 8. Cluster results for K=2

Table 3 shows the first 10 rules generated by the proposed methodology, sorted by the lift value. From the clustering results, female accidents account for about 20% of the total accidents. However, we have three rules related to females due to the application of the previous methodology.

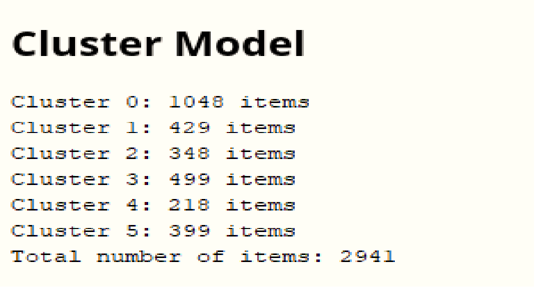
Using this methodology, we can generate more interesting association rules by increasing the number of clusters. For example, the methodology was implemented with a number of clusters K=6, and the clusters were generated within 1 second, as shown in Figure 9.

Table 3. The first ten rules generated by the set K=2

#	Premises	Conclusion	Support	Confidence	Lift
1	Gender = Male, TimeCategory = Afternoon	Accident_Severity = Serious, Road_Type = Single carriageway	0.14	0.76	1.05
2	TimeCategory = Afternoon	Gender = Male, Road_Type = Single carriageway	0.15	0.82	1.04
3	Gender = Male, Accident_Severity = Serious, TimeCategory = Afternoon	Road_Type = Single carriageway	0.14	0.82	1.04
4	Gender = Male, TimeCategory = Evening	Accident_Severity = Serious, Road_Type = Single carriageway	0.19	0.75	1.04
5	Accident_Severity = Serious, TimeCategory = Afternoon	Gender = Female, Road_Type = Single carriageway	0.17	0.84	1.04
6	Gender = Male, Accident_Severity = Serious, TimeCategory = Evening	Road_Type = Single carriageway	0.19	0.81	1.03
7	TimeCategory = Afternoon	Gender = Female, Accident_Severity = Serious, Road_Type = Single carriageway	0.17	0.79	1.03
8	TimeCategory = Noon	Gender = Male, Accident_Severity = Serious	0.10	0.95	1.03
9	Gender = Male, Road_Type = Single carriageway, Age_of_Driver_Category = Fifth decade	Accident_Severity = Serious	0.12	0.94	1.02
10	Gender = Female, TimeCategory = Mid-morning	Accident_Severity = Serious	0.16	0.97	1.02

Table 4. The first ten rules generated by set K=6

#	Premises	Conclusion	Support	Confidence	Lift
1	Gender = Female, Accident_Severity = Serious, Age_of_Driver_Category = Fifth decade	Road_Type = Single carriageway, TimeCategory = Afternoon	0.119	0.897	2.247
2	Road_Type = Single carriageway, Age_of_Driver_Category = Fifth decade	Accident_Severity = Serious, TimeCategory = Afternoon	0.119	1.000	2.158
3	Gender = Female, Accident_Severity = Serious, Road_Type = Single carriageway, Age_of_Driver_Category = Fifth decade	TimeCategory = Afternoon	0.119	1.000	2.076
4	Gender = Female, Accident_Severity = Serious, Age_of_Driver_Category = Fifth decade	Road_Type = Single carriageway, TimeCategory = Afternoon	0.152	0.894	1.600
5	Gender = Female, Road_Type = Single carriageway, Age_of_Driver_Category = Fifth decade	TimeCategory = Afternoon	0.161	0.976	1.573
6	Accident_Severity = Serious, Road_Type = Single carriageway, Age_of_Driver_Category = Fifth decade	TimeCategory = Afternoon	0.161	1.000	1.538
7	Gender = Female, Age_of_Driver_Category = Fifth decade	Road_Type = Single carriageway, TimeCategory = Afternoon	0.185	0.877	1.509
8	Road_Type = Single carriageway, Age_of_Driver_Category = Fifth decade	TimeCategory = Afternoon	0.178	0.976	1.458
9	Gender = Female, Accident_Severity = Serious, Road_Type = Single carriageway, TimeCategory = Mid-morning	Age_of_Driver_Category = Fourth decade	0.110	1.000	1.444
10	Age_of_Driver_Category = Fifth decade	Gender = Male, Road_Type = Single carriageway	0.434	0.891	1.398



Cluster Model

```
Cluster 0: 1048 items
Cluster 1: 429 items
Cluster 2: 348 items
Cluster 3: 499 items
Cluster 4: 218 items
Cluster 5: 399 items
Total number of items: 2941
```

Figure 9. Cluster results for K=6

We were able to generate 1387 rules. Table 4 shows the first ten rules generated by the proposed methodology with the set K=6, sorted by the lift value.

5. CONCLUSION

In this paper, we discuss how to generate rules based on clustering data. As we saw in the association rule results, if we change the number of clusters, we can generate more interesting and rare association rules based on the data of each cluster. Because of this methodology, it was essential for us to combine clustering algorithms with association rule algorithms. We were able to show that the clustering process adds value to the process of obtaining rare association rules by generating interesting rules that might not otherwise be found. Moreover, we can use this design and apply it to any data source.

REFERENCES

[1] Han, J., Pei, J., Kamber, M. (2011). Data Mining: Concepts and Techniques. Elsevier. https://doi.org/10.1016/C2009-0-61819-5

[2] Cai, Q.Q., Cui, H.G., Tang, H. (2017). Big data mining analysis method based on cloud computing. In AIP Conference Proceedings, 1864(1): 020028. https://doi.org/10.1063/1.4992845

[3] Agrawal, R., Imieliński, T., Swami, A. (1993). Mining association rules between sets of items in large databases. In Proceedings of the 1993 ACM SIGMOD international conference on Management of data, pp. 207-216. https://doi.org/10.1145/170035.170072

[4] Anoopkumar, M., Rahman, A.M.Z. (2016). A review on data mining techniques and factors used in educational data mining to predict student amelioration. In 2016 International Conference on Data Mining and Advanced Computing (SAPIENCE), pp. 122-133. https://doi.org/10.1109/SAPIENCE.2016.7684113

[5] Kumbhare, T.A., Chobe, S.V. (2014). An overview of association rule mining algorithms. International Journal of Computer Science and Information Technologies, 5(1): 927-930.

[6] Kotsiantis, S., Kanellopoulos, D. (2006). Association rules mining: A recent overview. GESTS International Transactions on Computer Science and Engineering, 32(1): 71-82. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.103.6295&rep=rep1&type=pdf.

[7] Solanki, S.K., Patel, J.T. (2015). A survey on association rule mining. In 2015 Fifth International Conference on Advanced Computing & Communication Technologies, pp. 212-216. https://doi.org/10.1109/ACCT.2015.69

[8] data.gov.uk. (2019). UK Car Accidents 2005-2015. https://www.kaggle.com/silicon99/dft-accident-data#contextCSVs.zip.

[9] Gao, Z., Pan, R., Yu, R., Wang, X. (2018). Research on automated modeling algorithm using association rules for traffic accidents. In 2018 IEEE International Conference on Big Data and Smart Computing (BigComp), pp. 127-132. https://doi.org/10.1109/BigComp.2018.00027

[10] Kabir, M.F., Ludwig, S.A., Abdullah, A.S. (2018). Rule discovery from breast cancer risk factors using association rule mining. In 2018 IEEE International Conference on Big Data (Big Data), pp. 2433-2441. https://doi.org/10.1109/BigData.2018.8622028

[11] Soniya Mudgal, M.P. (2020). Mining of the correlations for fatal road accident using graph-based fuzzified FP-growth algorithm. International Journal of Engineering and Advanced Technology (IJEAT), 9(5): 279-283. https://doi.org/10.35940/ijeat.E9526.069520