Available Online at http://www.journalajst.com

ASIAN JOURNAL OF
SCIENCE AND TECHNOLOGY

ISSN: 0976-3376

Asian Journal of Science and Technology
Vol. 12, Issue, 02, pp.11527-11534, February, 2021

RESEARCH ARTICLE

DATA PREPROCESSING AND PERCEIVING OUTLIERS USING CLUSTERING ALGORITHM ON REAL – TIME DATASET

*Dr. Rajeswari, J.

Assistant Professor, Department of Computer Science, Karpagam Academy of Higher Education,
Coimbatore -21

ARTICLE INFO

Article History:
Received 07th November, 2020
Received in revised form
19th December, 2020
Accepted 24th January, 2021
Published online 28th February, 2021

Key words:

Data Mining,
Clustering Algorithm,
Outliers, Preprocessing, Kmean,
Density based, EM, WEKA.

ABSTRACT

Data Mining is an interesting, interpretable new knowledge about a large amount of data. Data mining can be considered as part of the knowledge discovery process. Data mining works on patterns which is used to find the patterns in the data set. Data mining is an interdisciplinary subsequence of computer science and statistics with overall extract information from the data set[12]. Data set is nothing but database, which is a collectable information that is organize so that information can be easily accessed, manage and updated. Data set may contain random errors called noisy data which is unnecessary for the data set. These unwanted data are called outliers. The outliers may occur due to incorrect entry, sampling error, mis- reporting, Exceptional but true value. The outlier in the actual world is dirty, inadequate, noisy and unpredictable data which gives no quality mining results. These data has to be clean using the method called preprocessing. Preprocessing involves data cleaning, data integration, data transformation, data reduction and data discretization. Among the preprocessing methods, data cleaning plays a vital role in removing outliers and resolve inconsistencies. These outliers can be removed using an open source software called WEKA (Waikato Environment for Knowledge Analysis) issued under the GNU General Public Licence [11]. The patterns of data mining can be described as rules for clustering, classification, summarization and association. Based on the trends and relationship between the data, clustering produces a group of modules. There are different kinds of clustering algorithm as kmeans, EM, Farthest first, Filtered cluster, hierarchical, density based algorithms. In this paper a real time data is preprocessed and detecting outliers using various clustering algorithm. The various clustering algorithms are applied on the data set using WEKA tool. The analyzing result of various clustering algorithm is used to find out which algorithm is more comfortable and time consuming for the user for performing clustering algorithm.

Citation: Dr. Rajeswari, J., 2021. "Data Preprocessing and Perceiving Outliers Using Clustering Algorithm on Real – Time Dataset.", *Asian Journal of Science and Technology*, 12, (02), 11527-11534.

INTRODUCTION

Clustering is a method of dividing a data set into completely comparable subclasses called clusters. Similarly, the objects of same group is grouped together to form a cluster. In a large range of areas, such as market analysis, WWW, pattern recognition, image processing, data mining, clustering is pragmatic. Scalability, the ability to deal with multiple types of attributes, high dimensionality and interpretability are clustering algorithms. In the WEKA tool[10], different clustering algorithms are present. The WEKA instrument is an assembly of algorithms for machine learning for data mining tasks. In this paper, the sections is organized as introduction of clustering algorithm, introduction of WEKA, Preprocessing the data set, analyzing various clustering algorthims in WEKA tool, Comparing the result, Concludes the paper.

1. Clustering Analysis

Cluster Analysis is a segregated data-driven approach aimed at group-related segregation techniques In object clusters [2]. Data discovery that is directed at pedestrian fashion is the cluster study of flora.

Corresponding author: *Dr. Rajeswari, J.,
Assistant Professor, Department of Computer Science, Karpagam Academy of Higher Education, Coimbatore -21.

Clusterization is not a single grouping, but rather a process of considering object groups that are well interpretable. Each clustering is derived upon based on centroid which is center of the cluster. Clustering belongs to descriptive models used to group according to their similarity.

2. WEKA Tool

An open source tool is the Waikato Environment for Information Research (WEKA).Software produced in New Zealand by the University of Waikato. WEKA is a data-mining workbench. The WEKA tool was first printed in the C language and was rewritten in the future by the Java language that runs on all platforms, such as Linux, Windows, Mac[2]. WEKA covers a hefty number of algorithms for classification, clustering and a lot of algorithms for data preprocessing, feature selection and regression. Implementation of algorithm in WEKA tool starts with raw data. These raw data may contain several null and irrelevant values. These values can be cleaned by using data preprocessing tool provided in WEKA.

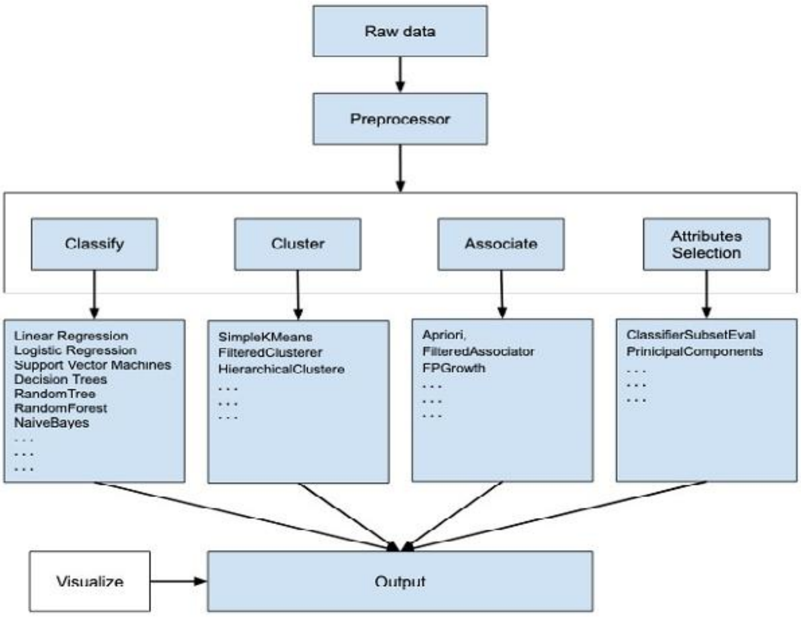
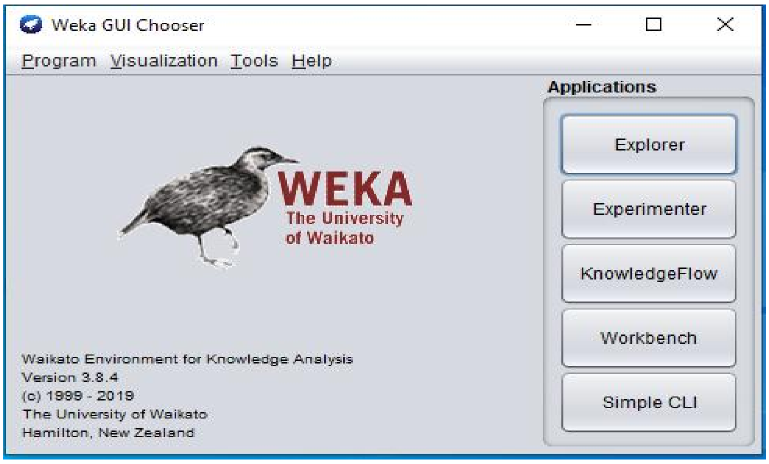


Fig . Processing of Raw Data in WEKA tool

The WEKA GUI chooser application will gave the following five different types of application. The types of applications are:

1. Explorer
2. Experimenter
3. Knowledge Flow
4. Workbench
5. Simple CLI



Waikato Environment for Knowledge Analysis
Version 3.8.4
(c) 1999 - 2019
The University of Waikato
Hamilton, New Zealand

Fig. 2. WEKA GUI Chooser

- **Explorer:**

This environment is used to exploring the data. It contains several tabs like preprocess, classify, cluster, associate, select attributes, visualize and give entree to all of its amenities using menu selection.

- **Experimenter:**

Statistical tests between learning schemes, evaluating machine learning algorithms and performing experiments are carried in this environment.

- **Knowledge Flow:**

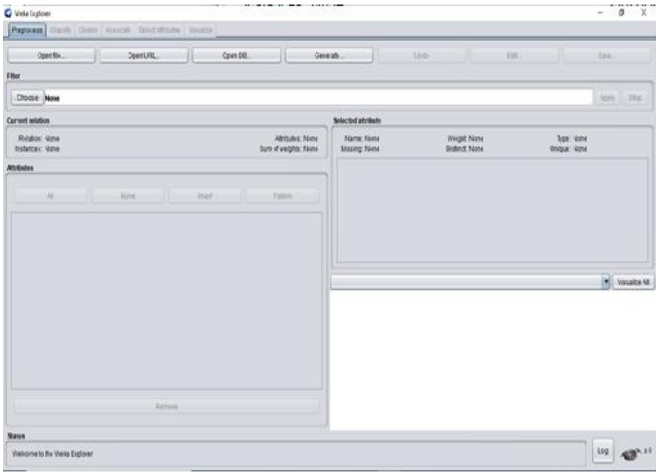
This environment is similar to Explorer, but with an interface for drag-and-drop. One advantage is that it offers gradual learning.

- **Workbench:**

A unified graphical user interface that incorporates interfaces (Explorer, Experimenter, Information Flow) into a single program, enabling the user to determine which applications and additions are to be displayed, in addition to the appropriate settings.

- **Simple CLI:**

Command-line interface is simple interface for writing commands of WEKA which does not provides their own command line interfaces.



Filter: None

Relation: None
Instances: None
Attributes: None
Sum of weights: None

Name: None
Missing: None
Weight: None
Distinct: None
Type: None
Unique: None

Welcome to the Weka Explorer

Fig 3. Explorer Environment in WEKA

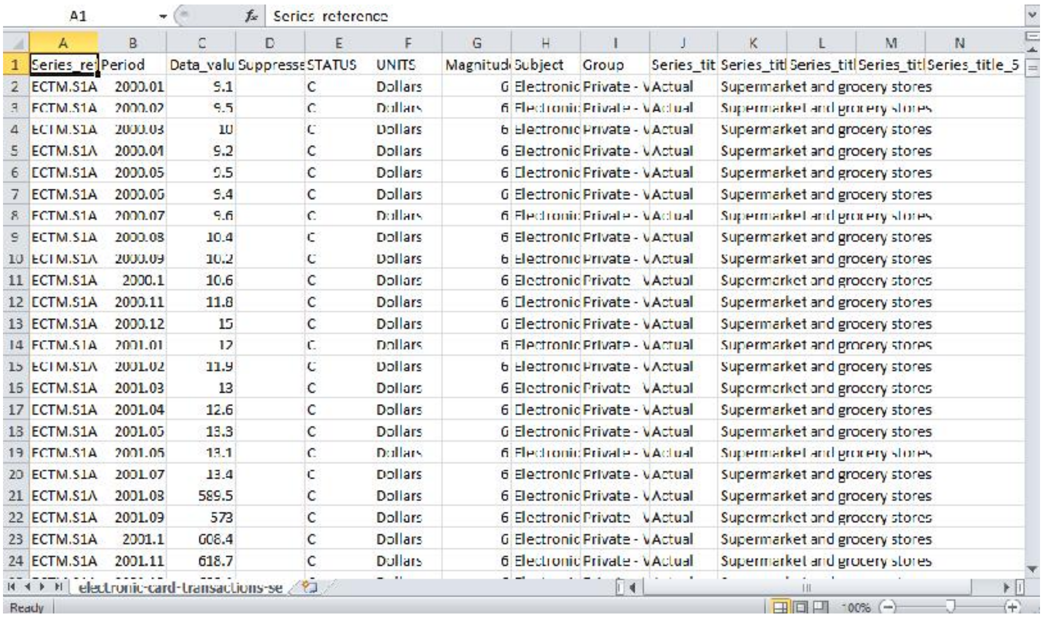
3. Data Preprocessing

The tool used to transform raw data into a clean data set is Data Preprocessing. In other words, once the data is obtained from different sources, it is collected in raw format, which is not feasible for analysis [1]. To convert the data into a small, clean data set, some steps are then taken. This process is performed before Iterative Analysis is executed. The sample data set electronic card transaction is taken which is in CSV format. In this data set it contains 6474 instances, 14 attributes, 48 distinct values, 14 features. In WEKA, on the Preprocess tab, select open file…, then select the dataset.CSV file in .CSV format.

3.1 Filtering Attributes and Discretization

There are few attributes in the dataset that need to be extracted before the data mining steps, which can be achieved by attribute filters in WEKA. There occur a popup window by clicking the filter panel[5]. In this window select “weka. filters. unsupervised. Attributes. Remove”. It removes the unwanted attributes which creates false result during analysis. In this dataset we remove the attributes suppressed, series_title_3, series_title_4, series_title_5. Now, remaining all attributes have to discretization. Discretization is process of converting real-valued attribute into an ordinal attributes called nominal attributes. The following figure shows the discretization on dataset.

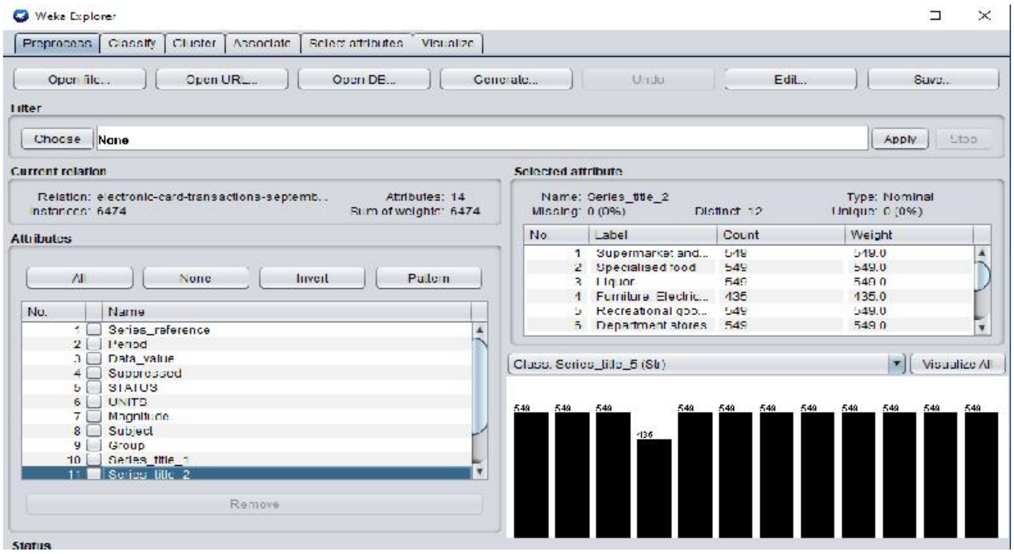
4. **Detecting Outliers using WEKA:** An outlier is an abnormal distance from other values in a random sample. Outliers cause miss significant finding or distort real result in many statistical analysis. These outliers can be handled by filters in WEKA[6]. The filter unsupervised attributes filter – Interquartile Range (IQR) is appled to adds new attributes that indicate the outliers or extreme values. Now, applying filters in the data set and finding out the outliers or extreme values. The following figure shows that 0% outliers in the data set.



A1 Series_reference

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Series_re	Period	Data_valu	Suppresse	STATUS	UNITS	Magnitud	Subject	Group	Series_tit	Series_titl	Series_titl	Series_titl	Series_title_5
2	ECTM.S1A	2000.01	9.1		C	Dollars	6	Electronic	Private - V	Actual	Supermarket and grocery stores			
3	ECTM.S1A	2000.02	9.5		C	Dollars	6	Electronic	Private - V	Actual	Supermarket and grocery stores			
4	ECTM.S1A	2000.03	10		C	Dollars	6	Electronic	Private - V	Actual	Supermarket and grocery stores			
5	ECTM.S1A	2000.04	9.2		C	Dollars	6	Electronic	Private - V	Actual	Supermarket and grocery stores			
6	ECTM.S1A	2000.05	9.5		C	Dollars	6	Electronic	Private - V	Actual	Supermarket and grocery stores			
7	ECTM.S1A	2000.06	9.4		C	Dollars	6	Electronic	Private - V	Actual	Supermarket and grocery stores			
8	ECTM.S1A	2000.07	9.6		C	Dollars	6	Electronic	Private - V	Actual	Supermarket and grocery stores			
9	ECTM.S1A	2000.08	10.4		C	Dollars	6	Electronic	Private - V	Actual	Supermarket and grocery stores			
10	ECTM.S1A	2000.09	10.2		C	Dollars	6	Electronic	Private - V	Actual	Supermarket and grocery stores			
11	ECTM.S1A	2000.1	10.6		C	Dollars	6	Electronic	Private - V	Actual	Supermarket and grocery stores			
12	ECTM.S1A	2000.11	11.8		C	Dollars	6	Electronic	Private - V	Actual	Supermarket and grocery stores			
13	ECTM.S1A	2000.12	15		C	Dollars	6	Electronic	Private - V	Actual	Supermarket and grocery stores			
14	ECTM.S1A	2001.01	12		C	Dollars	6	Electronic	Private - V	Actual	Supermarket and grocery stores			
15	ECTM.S1A	2001.02	11.9		C	Dollars	6	Electronic	Private - V	Actual	Supermarket and grocery stores			
16	ECTM.S1A	2001.03	13		C	Dollars	6	Electronic	Private - V	Actual	Supermarket and grocery stores			
17	ECTM.S1A	2001.04	12.6		C	Dollars	6	Electronic	Private - V	Actual	Supermarket and grocery stores			
18	ECTM.S1A	2001.05	13.3		C	Dollars	6	Electronic	Private - V	Actual	Supermarket and grocery stores			
19	ECTM.S1A	2001.06	13.1		C	Dollars	6	Electronic	Private - V	Actual	Supermarket and grocery stores			
20	ECTM.S1A	2001.07	13.4		C	Dollars	6	Electronic	Private - V	Actual	Supermarket and grocery stores			
21	ECTM.S1A	2001.08	589.5		C	Dollars	6	Electronic	Private - V	Actual	Supermarket and grocery stores			
22	ECTM.S1A	2001.09	573		C	Dollars	6	Electronic	Private - V	Actual	Supermarket and grocery stores			
23	ECTM.S1A	2001.1	608.4		C	Dollars	6	Electronic	Private - V	Actual	Supermarket and grocery stores			
24	ECTM.S1A	2001.11	618.7		C	Dollars	6	Electronic	Private - V	Actual	Supermarket and grocery stores			

Fig 4. Electronic Card Transaction Dataset in .CSV file format



Filter: None

Relation: electronic-card-transactions-septemb... Attributes: 14
Instances: 6474 Sum of weights: 6474

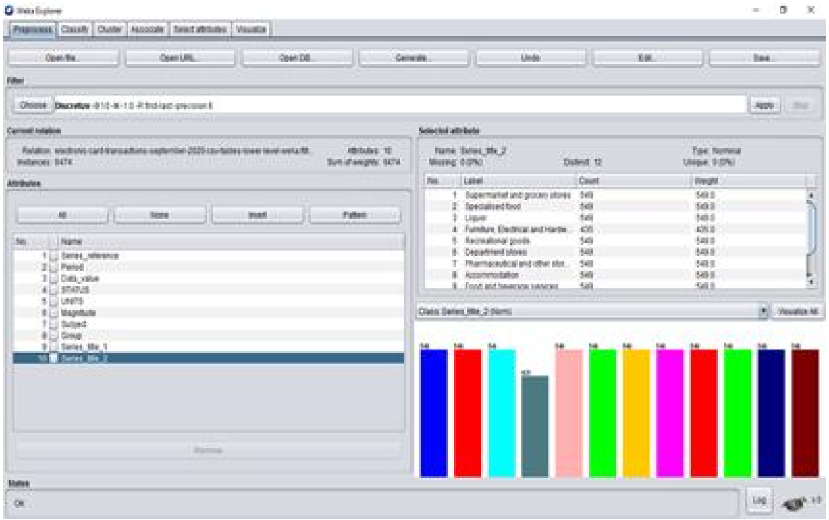
Name: Series_title_2 Type: Nominal
Missing: 0 (0%) Distinct: 12 Unique: 0 (0%)

No.	Label	Count	Weight
1	Supermarket and...	549	549.0
2	Specialised food	549	549.0
3	Liquor	549	549.0
4	Furniture, Electric...	435	435.0
5	Recreational goo...	549	549.0
6	Department stores	549	549.0

No.	Name
1	Series_reference
2	Period
3	Data_value
4	Suppressed
5	STATUS
6	UNITS
7	Magnitude
8	Subject
9	Group
10	Series_title_1
11	Series_title_2

Class: Series_title_5 (Str)

Fig 5. Electronic Card Transaction Dataset in WEKA before Preprocessing



Filter: Discretize -B 10 -M -1.0 -R first-last -precision 6

Relation: electronic-card-transactions-september-2020-csv-tables-lower-level-weka.filt... Attributes: 10
Instances: 6474 Sum of weights: 6474

Name: Series_title_2 Type: Nominal
Missing: 0 (0%) Distinct: 12 Unique: 0 (0%)

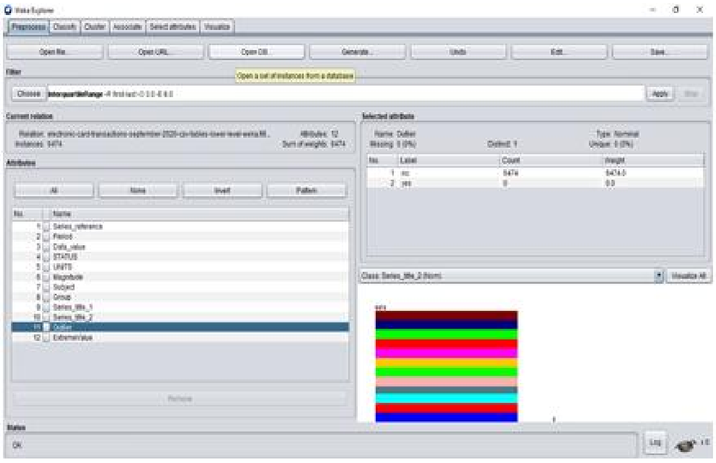
No.	Label	Count	Weight
1	Supermarket and grocery stores	549	549.0
2	Specialised food	549	549.0
3	Liquor	549	549.0
4	Furniture, Electrical and Hardw...	435	435.0
5	Recreational goods	549	549.0
6	Department stores	549	549.0
7	Pharmaceutical and other sto...	549	549.0
8	Accommodation	549	549.0
9	Food and beverage services	549	549.0

No.	Name
1	Series_reference
2	Period
3	Data_value
4	STATUS
5	UNITS
6	Magnitude
7	Subject
8	Group
9	Series_title_1
10	Series_title_2

Class: Series_title_2 (Nom)

Status: OK

Fig 6. Discretization process on dataset



Filter

Choose InterquartileRange -R first-last -O 3.0 -E 6.0

Current relation

Relation: electronic-card-transactions-september-2020-csv-tables-lower-level-weka.filt... Attributes: 12
Instances: 6474 Sum of weights: 6474

Selected attribute

Name: Outlier Type: Nominal
Missing: 0 (0%) Distinct: 1 Unique: 0 (0%)

No.	Label	Count	Weight
1	no	6474	6474.0
2	yes	0	0.0

Attributes

No.	Name
1	Series_reference
2	Period
3	Data_value
4	STATUS
5	UNITS
6	Magnitude
7	Subject
8	Group
9	Series_title_1
10	Series_title_2
11	Outlier
12	ExtremeValue

Class: Series_title_2 (Nom)

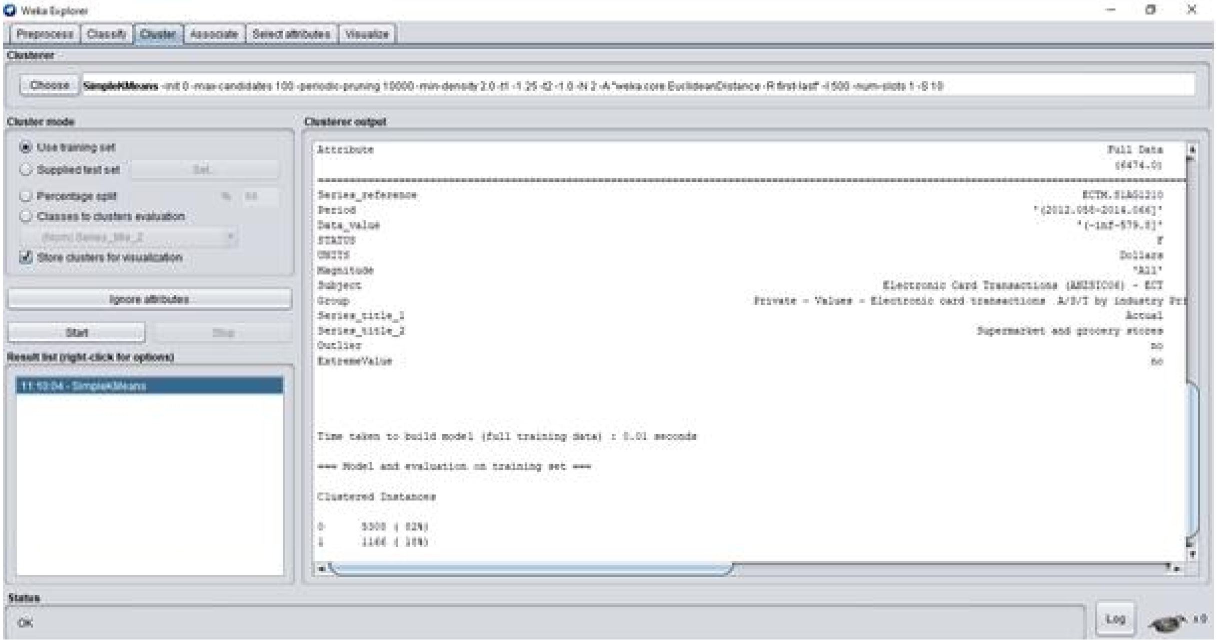
Status

OK

Fig 7. Outliers or Extreme values in Dataset

5. Clustering Algorithms

5.1 **K-Mean Clustering Algorithm:** K-mean is used to find minimum distance in clusters in dataset. K-mean clustering is a part of data warehousing and mining. K-mean clustering is an exploratory data analysis techniques used to analysis a complete data set[4]. It is used to implements nonhierarchical method of grouping objects together. K-mean algorithm uses the centroid method which is used to compute the distance between the objects using Euclidean method. After calculating the Euclidean distance, the objects are grouped based on minimum distance. Now, the data set which has preprocessed and detected the outlier is performed the K-means clustering algorithm using WEKA tool.



Clusterer

Choose SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 2 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10

Cluster mode

● Use training set
○ Supplied test set
○ Percentage split % 66
○ Classes to clusters evaluation
(Nom) Series_title_2
☑ Store clusters for visualization

Result list (right-click for options)

11:53:04 - SimpleKMeans

Clusterer output

```
Attribute                                                                          Full Data
                                                                                    (6474.0)
============================================================================================
Series_reference                                                                ECTM.S1AG1210
Period                                                                '(2012.055-2014.066]'
Data_value                                                                    '(-inf-579.8]'
STATUS                                                                                     F
UNITS                                                                                Dollars
Magnitude                                                                              'All'
Subject                                   Electronic Card Transactions (ANZSIC06) - ECT
Group                  Private - Values - Electronic card transactions  A/S/T by industry Pr
Series_title_1                                                                        Actual
Series_title_2                                                  Supermarket and grocery stores
Outlier                                                                                   no
ExtremeValue                                                                              no



Time taken to build model (full training data) : 0.01 seconds

=== Model and evaluation on training set ===

Clustered Instances

0      5308 ( 82%)
1      1166 ( 18%)
```

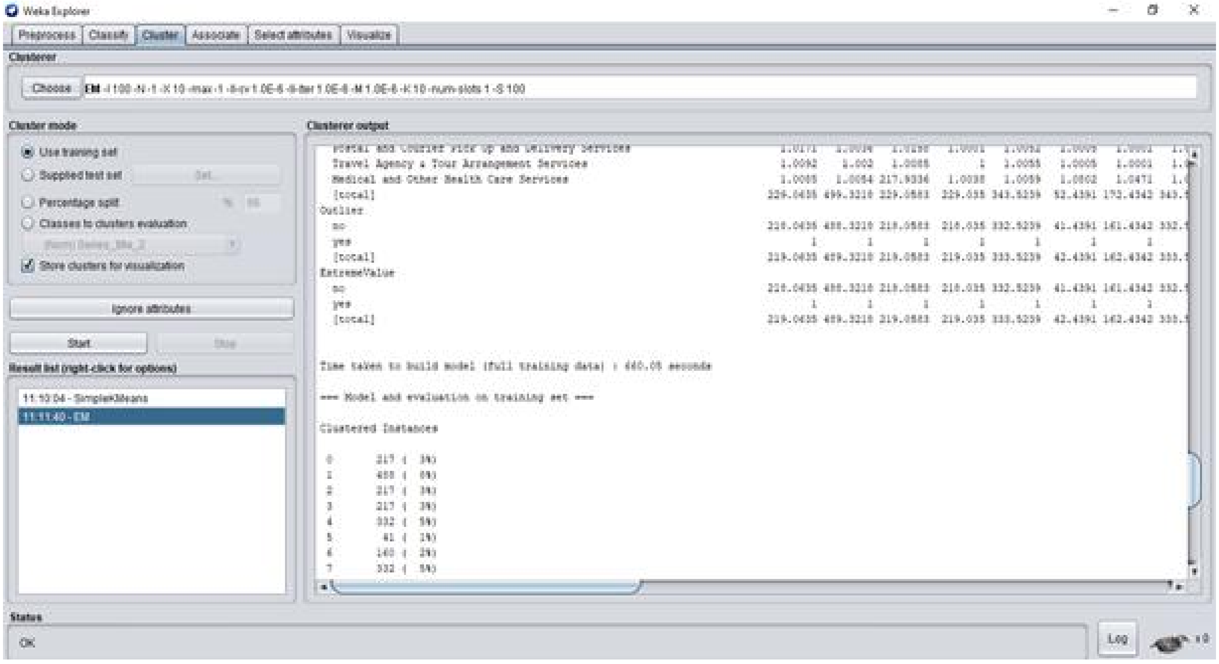
Status

OK

Fig 8. K-Mean Clustering Algorithm using WEKA tool

5.2 **EM Clustering Algorithm:** EM algorithm is called as Expectation – Maximization algorithm. Expectation is our estimation from machines and then classify the data into some classes. Maximization if finding the maximized class from the estimated one[8]. EM process starts on the given data set by repeating the two steps until a clear result is formed. The first step is E-step used to classify the data using the current theory. In this step, generates the expected classification for the given data set. The second step is M-step use to generate best theory using current classification of the data. The missing values exist in this model, which can be formulated simply by this EM steps.

5.3 **Density-Based Clustering Algorithm:** Clustering based on density is built on the function of probability density for the defined data collection. Naturally, density-based algorithms repute clusters as compact regions of data space objects that are separated by low-density regions[2]. Using two parameters, such as Eps and Minpts, they locate and distinguish high density regions from low density regions.



Clusterer: EM -I 100 -N -1 -X 10 -max -1 -ll-cv 1.0E-6 -ll-iter 1.0E-6 -M 1.0E-6 -K 10 -num-slots 1 -S 100

Cluster mode: ● Use training set; ○ Supplied test set; ○ Percentage split % 66; ○ Classes to clusters evaluation (Nom) Series_title_2; ☑ Store clusters for visualization

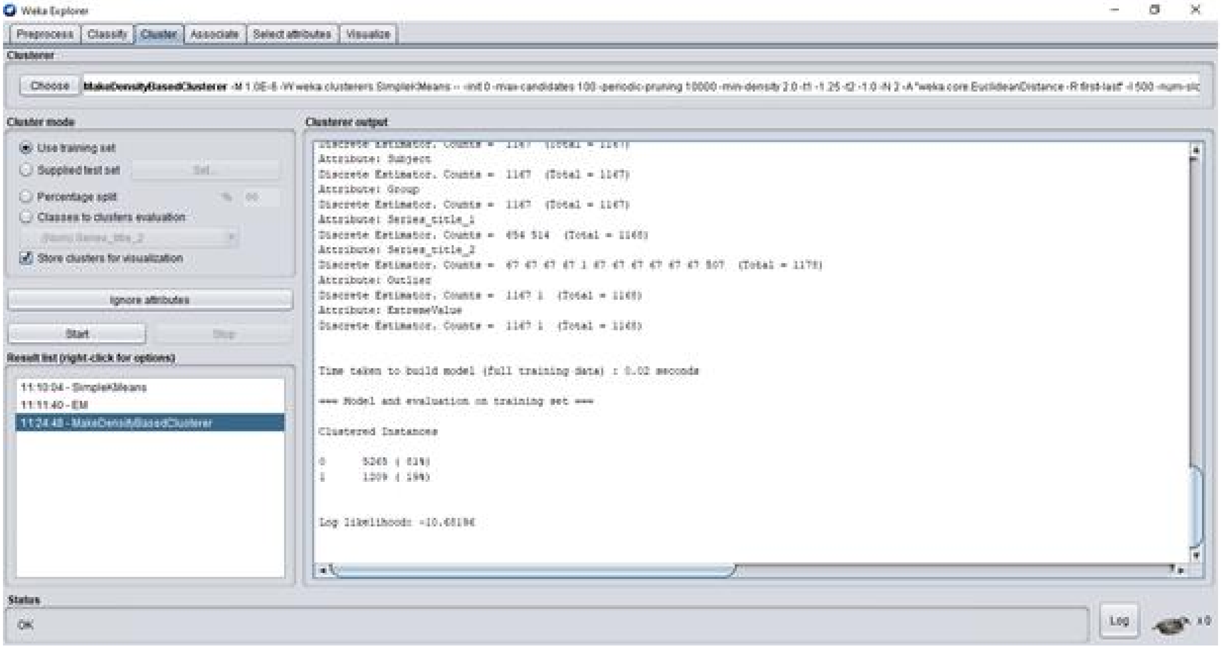
Result list (right-click for options): 11:10:04 - SimpleKMeans; 11:11:40 - EM

Clusterer output:

```
  Postal and Courier Pick up and Delivery Services          1.0171   1.0034   1.0288   1.0001   1.0052   1.0009   1.0001   1.
  Travel Agency & Tour Arrangement Services                 1.0092    1.002   1.0085        1   1.0055   1.0005   1.0001   1.
  Medical and Other Health Care Services                    1.0085   1.0054 217.9336   1.0038   1.0059   1.0802   1.0471   1.
  [total]                                                 229.0435 499.3218 229.0583  229.035 343.5259  52.4391 172.4342 343.5
Outlier
  no                                                      218.0435 488.3218 218.0583  218.035 332.5259  41.4391 161.4342 332.5
  yes                                                            1        1        1        1        1        1        1
  [total]                                                 219.0435 489.3218 219.0583  219.035 333.5259  42.4391 162.4342 333.5
ExtremeValue
  no                                                      218.0435 488.3218 218.0583  218.035 332.5259  41.4391 161.4342 332.5
  yes                                                            1        1        1        1        1        1        1
  [total]                                                 219.0435 489.3218 219.0583  219.035 333.5259  42.4391 162.4342 333.5


Time taken to build model (full training data) : 440.05 seconds

=== Model and evaluation on training set ===

Clustered Instances

 0       217 (  3%)
 1       488 (  8%)
 2       217 (  3%)
 3       217 (  3%)
 4       332 (  5%)
 5        41 (  1%)
 6       160 (  2%)
 7       332 (  5%)
```

Status: OK

Fig 9. EM Clustering Algorithm using WEKA tool



Clusterer: MakeDensityBasedClusterer -M 1.0E-6 -W weka.clusterers.SimpleKMeans -- -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 2 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num-slc

Cluster mode: ● Use training set; ○ Supplied test set; ○ Percentage split % 66; ○ Classes to clusters evaluation (Nom) Series_title_2; ☑ Store clusters for visualization

Result list (right-click for options): 11:10:04 - SimpleKMeans; 11:11:40 - EM; 11:24:48 - MakeDensityBasedClusterer

Clusterer output:

```
Discrete Estimator. Counts =  1167  (Total = 1167)
Attribute: Subject
Discrete Estimator. Counts =  1167  (Total = 1167)
Attribute: Group
Discrete Estimator. Counts =  1167  (Total = 1167)
Attribute: Series_title_1
Discrete Estimator. Counts =  654 514  (Total = 1168)
Attribute: Series_title_2
Discrete Estimator. Counts =  67 67 67 67 1 67 67 67 67 67 67 507  (Total = 1178)
Attribute: Outlier
Discrete Estimator. Counts =  1167 1  (Total = 1168)
Attribute: ExtremeValue
Discrete Estimator. Counts =  1167 1  (Total = 1168)


Time taken to build model (full training data) : 0.02 seconds

=== Model and evaluation on training set ===

Clustered Instances

0      5265 ( 81%)
1      1209 ( 19%)


Log likelihood: -10.68196
```

Status: OK

Fig. 10. Density-Based Clustering Algorithm using WEKA tool

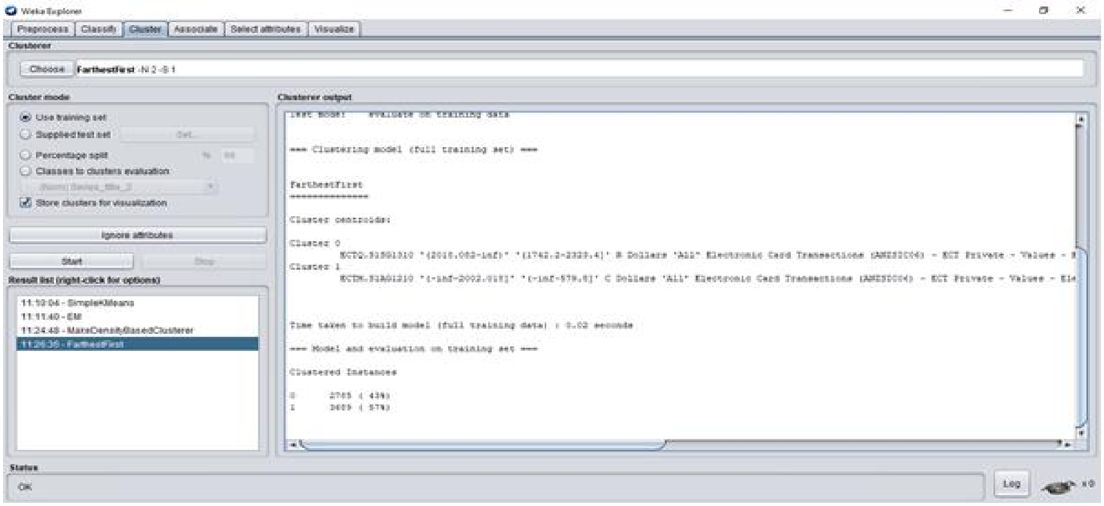
Eps is the maximum neighborhood radius and Minpts is the minimum number of points for a point in the Epsneighborhood. In this algorithm clusters are formed as maximum sets of density connected points and can detect noise and used when outliers are encountered.

5.4 Farthest First Clustering Algorithm

The FarthestFirst Traversal Algorithm was first implemented by Hochbaum and Shmoys, 1985.A version of K-center that is chosen as cluster centers is Farthest First. The Farthest First Traversal (FFT) is a randomly chosen first-center grasping algorithm. Then, greedily, the second center is picked as the point farthest from the first. Similarly, the residual points that are farthest from the points already selected[5] are determined. The residual points are further applied to the nearest center of the cluster, rather than the farthest points. Farthest first algorithm is suitable for the large dataset.

6. Result Analysis of Various Clustering Algorithms using WEKA

The focal goal of this paper is to give introduction of clustering algorithms using



FarthestFirst -N 2 -S 1

● Use training set
○ Supplied test set
○ Percentage split % 66
○ Classes to clusters evaluation
(Nom) Series_title_2
☑ Store clusters for visualization

11:10:04 - SimpleKMeans
11:11:40 - EM
11:24:48 - MakeDensityBasedClusterer
11:26:36 - FarthestFirst

```
Test mode:    evaluate on training data


=== Clustering model (full training set) ===


FarthestFirst
==============

Cluster centroids:

Cluster 0
        ECTQ.S1S01S10 '(2016.002-inf)' '(1742.2-2329.4]' R Dollars 'All' Electronic Card Transactions (ANZSIC06) - ECT Private - Values - E
Cluster 1
        ECTM.S1A01S10 '(-inf-2002.018]' '(-inf-579.8]' C Dollars 'All' Electronic Card Transactions (ANZSIC06) - ECT Private - Values - Ele



Time taken to build model (full training data) : 0.02 seconds

=== Model and evaluation on training set ===

Clustered Instances

0      2785 ( 43%)
1      3689 ( 57%)
```

OK

Fig 11. Farthest First Clustering Algorithm using WEKA tool

Table 1. Comparison result of various Clustering Algorithm

Algorithm	No. of Clusters	Cluster Instance	No. of Iteration	Time Taken to build the model
K-mean Algorithm	2	5308 (82%) 1166 (18%)	3	0.01 seconds
EM Algorithm	33	217 (3%) 288 (8%) . . . 44 (1%)	28	660.05 seconds
Density-Based Algorithm	2	5265 (81%) 1209 (19%)	3	0.02 seconds
Farthest First Algorithm	2	2785 (43%) 3689 (57%)	1	0.02 seconds

WEKA tool. The data set electronic card transaction has 14 attributes, 6474 instances and 48 distinct values. This data set has be preprocessed and outliers has been detected, various clustering algorithms has been applied on this data set. The following table shows the implementation of different data set clustering algorithms that evaluate the best clustering algorithm among the algorithms used for clustering. Based on the time taken to create the model, it can be concluded that the best and fastest clustering algorithm is the K-mean clustering algorithm. In the following table, the product of the clustering algorithm is shown:

Conclusion

Outliers are the interesting theory of data mining. These outliers have been extracted in this paper using the method of data preprocessing in the data collection. We have also predicted different clustering algorithms in this paper using the WEKA tool. When working with WEKA, deep knowledge is not needed. So, WEKA is a very appropriate tool for open source data mining.Based on the time factor, as opposed to other algorithms, we found that the K-mean clustering algorithm is the easiest, highest output and time consuming algorithm.

REFERENCES

1. Dipannita Kar, Mr. Haresh Chande and Mr. Rajendra Gaikwad,, 2017. A study paper on Outlier Detection on Time series data", *International Journal of Creative Research Thoughts*, Vol 5, Issue ISSN: 2320 - 2882.
2. Dr. Deepajothi, S and Dr. Juliana, 2019. "Survey of Clustering Algorithm of WEKA tool on Labor Dataset", International Journal of Applied Engineering Research, Vol 14, No. 5,, ISSN: 0973-4562.
3. Hassan, M. R., Hossain, M. M., Begg, R. K., Ramamohanarao, K., &Morsi, Y. 2010. Breast-cancer identification using HMM-fuzzy approach. Computers in Biology and Medicine, 40, 240–251.
4. In Gao, David B. 2009. Hitchcock ―James Stein Shrinkage to Improve K-means Cluster Analysis‖ University of South Carolina, Department of Statistics November 30,.
5. Jiawei Han and Micheline Kamber, "Data Mining: Concepts and Techniques", Morgan Kaufmann Publishers, second Edition, (2006). 2757 International Journal of Engineering Research & Technology (IJERT) ISSN: 2278-0181 www.ijert.org Vol. 2 Issue 6, June - 2013 IJERT.
6. Sunitha, L.Dr. M. Balraju, Dr. J. Sasikiran, E. Venkat Ramana, "Automatic Outlier Identification in Data Mining using IQR in Real-Time Data", *International Journal of Advanced Research in Computer and Communication Engineering*", Vol 3, No. 6,2014, ISSN: 2278-1021.

7. Mac Queen J. B. 1967. "Some Methods for classification and Analysis of Multivariate Observations", Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. University of California Press. pp. 281–297.
8. Manish Verma, Mauly Srivastava, Neha Chack, Atul Kumar Diswar and Nidhi Gupta, 2012. “A Comparative Study of Various Clustering Algorithms in Data Mining”, International Journal of Engineering Research and Applications (IJERA) Vol. 2, Issue 3, May-Jun, pp.1379-1384.
9. Nguyen, T., Khosravi, A., Creighton, D., &Nahavandi, S. 2015a. Medical data classification using interval type-2 fuzzy logic system and wavelets. Applied Soft Computing, 30, 812–822.
10. Sapna Jain, M Afshar Aalam and M N Doja, 2010. “K-means clustering using weka interface”, Proceedings of the 4th National Conference; INDIACom.
11. Shhweta Srivastava, 2014. “WEKA: A Tool for Data preprocessing, Classification, Ensemble, Clustering and Association Rule Mining” International Journal of Computer Applications(0975-8887), Vol-88, No.10,
12. Zhang Haiyang, 2011. “A Short Introduction to Data Mining and its Applications”, Institute of Electrical and Electronics Engineers (IEEE).

****