Development of a Unifying Theory for Data Mining Using Clustering Techniques

Yaser Issam Hamodi
Ministry of Higher Education & Scientific Research, Baghdad, Iraq.
E-mail: yaserissamtc@gmail.com

Ruaa Riyadh Hussein
College of Education for Girls, Al-Iraqia University, Baghdad, Iraq.

Naeem Th. Yousir
College of Information Engineering, Al-Nahrain University, Baghdad, Iraq.

Received May 04, 2020; Accepted July 10, 2020
ISSN: 1735-188X
DOI: 10.14704/WEB/V17I2/WEB17012

Abstract

A performance evaluation of four different clustering techniques was carried out based on segmenting consumer by product type and by product usage in the research. Cobweb, DBSCAN, EM and k-means algorithms were evaluated based on the computational time, accuracy of the result produced and the purity of the result produced. The experiment was performed using WEKA as a data mining tool. The performance evaluation of the four techniques showed that K-means outperformed others in all considered evaluation measure while the EM technique was the second best in terms of accuracy and purity, outperforming the other two. DBSCAN technique was the 3rd best of the selected algorithms even as its computational time is shorter than that of EM while the fourth best performing calculation has been believed to be the Spider web calculation as respects to immaculateness, exactness and computational time.

Keywords

Cobweb, DBSCAN, EM, K-means Algorithm, WEKA.

Introduction

Cluster is a generally excellent way for ordering homogenous gatherings of articles called groups. Articles or perceptions that are indistinguishable will in general offer numerous qualities however are exceptionally not at all like an item that doesn't have a place with that cluster [1]. Clustering aims to organize collections of data objects or items into different clusters, in ways that items within some specific clusters are more similar than they are to items in the other clusters. Clustering in data mining can be used to discover distribution guide in anyunderlying data. It is additionally utilized in gathering indistinguishable clients and items, as it is fundamentally, an essential action for showcasing. Scientists exclusively structure division of market on down to earth applications, mechanical use with shrewdness subsequently, group examination causes sections to be framed based on information which are less reliant on subjectivity [1]. Mining of data involves two types of learning approaches in supervised and unsupervised learning. Clustering is a type of unsupervised learning approach. It is the method of gathering sets of items in various manners in order to guarantee that the objects of the clusters are more indistinguishable from each other than are similar to the ones in different groups. The techniques for bunching have numerous uses in numerous areas, for example, man-made reasoning, bioinformatics, AI, design acknowledgment and segmentation [2].

Customer segments conclusively analyze data just to understand patterns. Collecting and analysing data doesn’t really make sense except wish to objectively or subjectively study the data, major importance of segmentation analysis is to; note the largely and slightest profitable customers, so as to better focus marketing efforts, Improvement in customer service, maintaining loyal relationships with customers, pricing and for better improvement of the products [3].

Clustering algorithms are used to gather homogenous groups of data and analyze them based on specifically defined criteria. There are numerous clustering algorithms even as there is also much potential for evaluating the algorithms. The determination of the most proper clustering algorithm and viable proportion of assessment relies upon the items for clustering and the kind of clustering undertaking to be executed. Clustering algorithms have been used extensively for a lot of segmentation procedures thereby solving many problems, many algorithms have been used to cluster and even analyze data but little investigation or study has been done on evaluating the performances of the algorithms used to carry out the clustering procedures, in order to determine the most efficient. The aim of this work is to carry out a performance evaluation of some selected clustering algorithms used for segmenting customers based on product usage and type. This research carried out a comparative analysis which established the most efficient of the four different clustering techniques as a unifying theory for data mining.

Theoretical Framework

1) Brief Overview of Clustering and Data Mining

A. Clustering just signifies gathering of gathered items into classes with same articles. Cluster investigation signifies a noteworthy instrument in information examination. Clustering strategy is utilized for characterization of accumulation models in explicit gatherings naturally as for what they share in like manner. Normally, designs in a comparative group will in general be more indistinguishable from each other than designs in an alternate cluster. In this manner, it gets essential to have the option to appreciate the divergence between unsupervised clustering order and directed grouping arrangement. In regular markets, client clustering or division is for the most part noteworthy procedure applied in considering showcasing. The investigation isolated open client clustering or strategies for division into strategically or methodological-based methodology and application or utilitarian based methodology. To recognize improved divided homogenous gathering, most of the methodological arranged examinations utilized arithmetic methodological methodology, for example, Fluffy set, the conventional calculation (GA), neural net and measurements [4]. The calculation is an efficient equation (usual methodology) used to investigate oftentimes including multiple iterations, yet regardless, have an end-point which delivers the outcome as yield. The computerization procedure permits applying complex calculations to significant arrangements of information, subsequently, the outcomes will be immediately created with inconsequential expense. Despite the fact that information mining calculations are constantly helpful to enormous informational collections, a portion of these calculations can likewise be applied to generally little informational indexes. Datasets utilized for use in information mining are normally straightforward in structure.

B. One significant issue identified with data clustering algorithm institutionalization, the created algorithms could give better outcomes in some kind of information in contrast with the outcomes gave in regards to datasets of different sorts. Numerous endeavors have been made to institutionalize the algorithms to perform appropriately in all cases, be that as it may, up to now, no primary accomplishment has been done [5]. A great deal of clustering algorithms were proposed right up 'til the present time, yet, every algorithm has a few characteristics and a few disadvantages and it can not work well in all cases for example institutionalization. To guarantee getting all the benefits of clustering algorithms, a few necessities must be achieved, for example, Versatility: It implies that the information ought to be adaptable, so as to gain right outcomes. A clustering algorithm manages various kinds of characteristics, finds clustered information with discretionary shapes must be obtuse toward clamor and exceptions, the outcomes acquired from a clustering algorithm should likewise be interpretable and useable and it must have the option to manage informational index of high dimensionalities. As expressed by Gartner Gathering [6], information mining can be characterized as the progression in finding significant new patterns, examples, and connections through the way toward examining a generous number of information put away in a distribution center, by using some numerical and measurable methodologies and example acknowledgment innovations. It is the utilization of present techniques for information examination to disentangle once in the past unnoticed relationship among information things. The way toward mining routinely includes looking at information that is loaded up in the information storehouses. Significantly 3 information mining techniques are considered the most significant; clustering, arrangement, and relapse. A few scientists could utilize inapplicable investigation to sets of information which require a totally extraordinary technique, for instance for that is the models might be found to have been based upon totally empty suppositions. Hence, it is a necessity to get a handle on the arithmetical and numerical model structures fundamental the product [7].

C. Information disclosure process as portrayed in Figure 2.1 incorporates an iterative progression of stages like information introduction, design assessment, information mining, information change, information reconciliation and information cleaning. The primary jobs and functionalities with respect to Information mining are group investigation, forecast, order, connection, affiliation, segregation and portrayal, etc. clustering is considered as one of the information mining systems which cluster comparative information to a group and various information to various group [8].

2) Consumer Segmentation and Product Usage

Division by item utilization shows to a gathering of shoppers or a lot of organizations with one of a kind or particular attributes through expending or utilizing items. It is a progression of occasion which incorporates confining and finding the customers of the market to homogenous gatherings as indicated by their inclinations, demeanor and requirements taking into reflection their item use. It assists with understanding what the clients truly need in states of the pace of their buy and what the provider can offer them. It remains as a splendid establishment as to winning and keeping beneficial clients [9]. It is predominantly the refocusing of the market as indicated by how regularly the buyer uses some specific items. Shoppers are grouped into 3 kinds, the first is the people who utilize the item, the people who utilize the item a great deal and the people who have little utilization for that item [10]. A client or market subdivision is utilized to join the clients who share explicit qualities. One of the most significant approaches to gauge the association between the client and the item and friends is to comprehend the clients, their disparities and their connections [11]. It doesn't simply disclose to you how to serve your clients better, yet additionally it enables the revelation of neglected purchaser needs and upgrade the conveyance of improved items and administrations to both old and new gatherings of buyers [12].

3) Consumer Segmentation by Product Type

Another aspect of segmentation, apart from the product usage categorization, in Market segmentation, is the "segmentation by product type". This simply classifies the segments of homogeneous consumers who purchase specific types of goods. For instance, sub-dividing the market to a category or group of individuals with similar demands in the group, but different demands over the groups [13]. The sub-group would then be individuals want cars, but different car types. The diversity of car types could be SUV's, sport, and luxury. This is considered as a sub-division of the market according to the advantages the user hope to get by using that specific product [9] [13]. Some more recent works related to clustering for WSNs studied in [22]-[24].

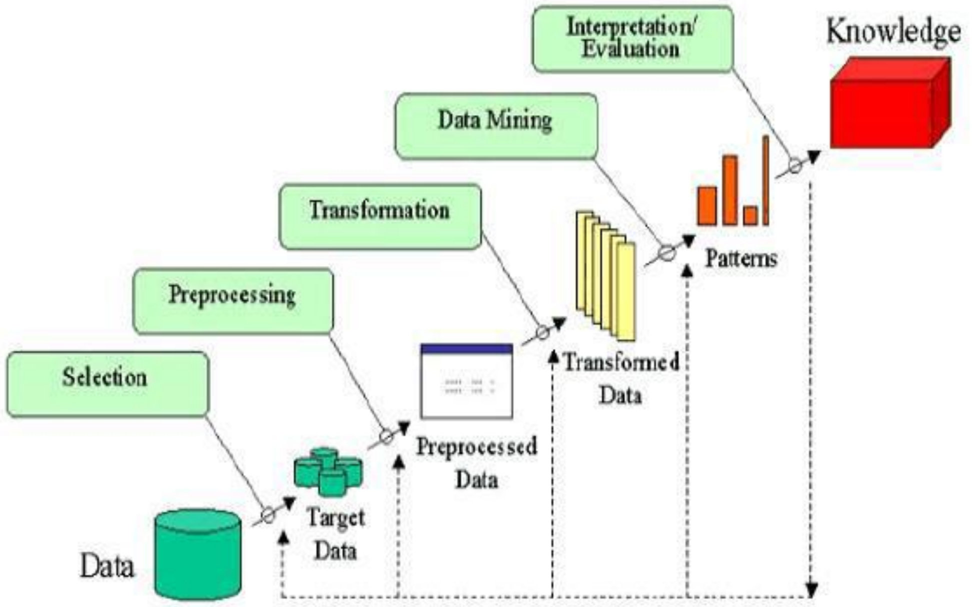

Figure 2.1 Phases of data mining (Raj, Sunil and Juhi, 2014)

4) WEKA Clustering Algorithms

The Weka clustering algorithms include Cobweb, DBSCAN, K-means, and EM.

- **Cobweb Clustering Algorithm**

Spider web can be characterized as a steady technique for progressive applied bunching. It does employ four major operations in constructing the classification tree. Whichever operation is chosen at every single time rely upon the grouping of the classification that is achieved viautilizingit.

- **DBSCAN Clustering Algorithm**

DBSCAN which is an abbreviation for (Density-based spatial clustering of applications with noise) can be defined as a clustering algorithm created in the year 1996 by Xiaowei Xu, Jörg Sander, Kriegel, Hans-Peter and Martin Ester. DBSCAN algorithm is shown in Figure 2.2.

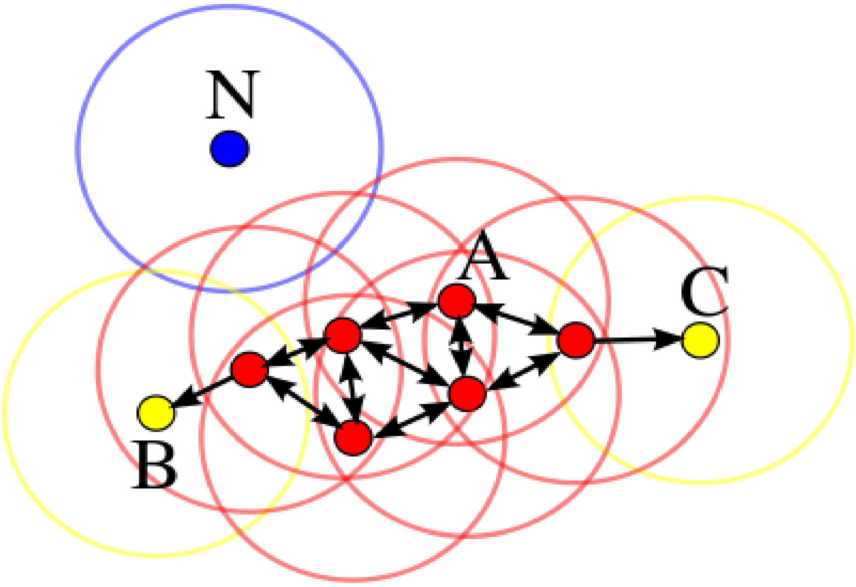

Figure 2.2 Demonstration of DBSCAN Clustering Algorithm (Wikipedia, 2016)

- **K-Means Clustering Algorithm**

In data mining, the K-means clustering algorithm is considered as a process of vector quantization, regularly from signal processing, that is widely used in cluster analysis.

- **Expectation Maximization Algorithm**

Expectation Maximization algorithm which can be abbreviated to (EM) can be defined as a general-purpose maximum likelihood algorithm for missing-data problems. It is effective in resolving predicament of parameter estimation.

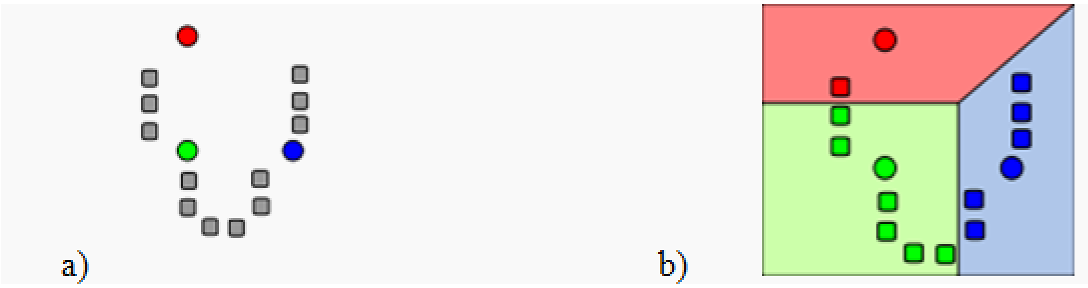

Figure 2.3 (a,b) Demonstration of the Standard K-means Algorithm (Wikipedia, 2016).

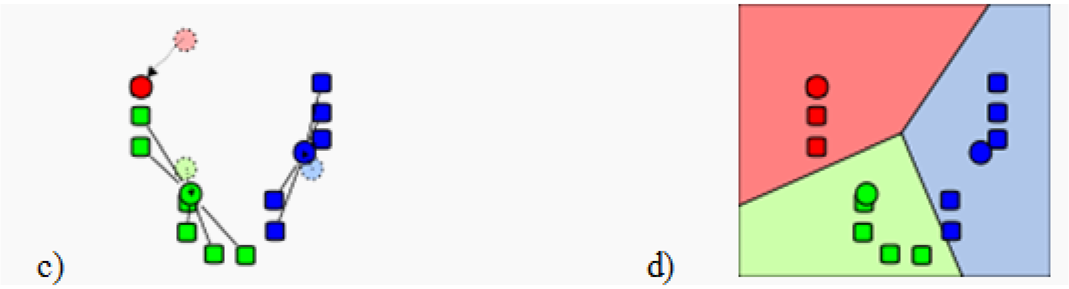

Figure 2.3 (c,d) Demonstration of the Standard K-means Algorithm (Wikipedia, 2016).

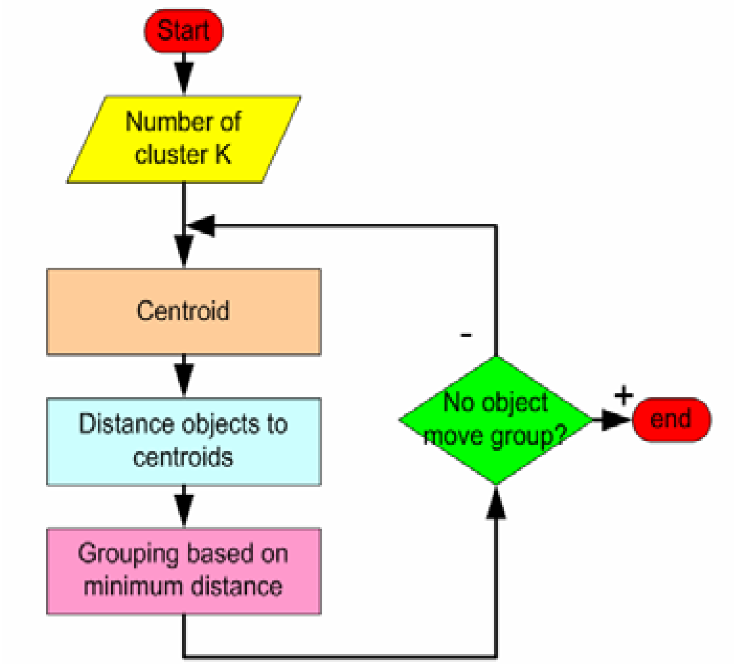
Start
Number of cluster K
Centroid
Distance objects to centroids
Grouping based on minimum distance
No object move group?
-
+
end

Figure 2.4K-means clustering process (Sapna, Aalam, and Doja, 2010).

5) Parameters for Evaluation

When statistical moves towards clustering are used, it is highly justified is by carefully making use of statistical methods and testing hypotheses so as to avoid finding patterns in noise and to compare clustering algorithms, the following factors were considered in the course of this work:

i. Computational time for the result generation
ii. The accuracy of the result generated.
iii. The purity of the result generated.

6) Phases of the Clustering Process

WEKA is the data mining tool in use for this research. In the WEKA platform the four selected clustering algorithms; Cobweb, EM, DBSCAN, and K-means were implemented one after another. After launching the WEKA application, there are four available options: explorer, experimenter, knowledge flow and simple CLI. The option used for this research was the explorer environment as it comprises the; pre-process, classify, cluster, association, select attribute and visualize tabs.

The individual phases of the procedures are as summarised:

i. Open an.arff or CSV file – The preprocess tab in the explorer environment enables data files to be imported into the WEKA application. No operation can be carried out using the WEKA application except a data file has been imported into the application. The file formats acceptable by weka is the.arffand.CSV file format. Hence, this research work employed the.CSV file format.
ii. Cluster – This tab is used for segmentation. The tab was used because segmentation was the primary operation in the research before the performance evaluation was done.
iii. Expand clustering methods- This was done also under the cluster tab, there are available options for the type of clustering algorithms to be implemented at every single time the application is launched.
iv. Choose clustering algorithms – This research employed the EM, DBSCAN Cobweb, and K-means clustering algorithms. The listed algorithms were therefore selected and implemented one after another.
v. Set test options and modes-The test options and modes used for the four clustering algorithms implemented were the training and test set modes. The iteration was carried out twice each with different values of cluster category.
vi. Results - The results generated from the four selected clustering algorithms were analyzed and its performance compared using the computational time, accuracy and purity as performance metrics.

The Selected Clustering Algorithms

Previous researchers evaluated two and also three clustering algorithms. In this research, four clustering algorithms (Cobweb, EM, DBSCAN, and K-means) were selected, examined implemented and the overall result evaluated. The pseudo codes for the selected clustering algorithms are as depicted in Figures: 3.1, 3.2, 3.3 and 3.4.

```
doubleem (int n,double * data, int k, double * prob, double * mean,double *sd,double eps)
{
double 11k=0,prev 11k=0 ;
double ** class_prob = alloc_matrix(n,k);
startem (n,data,k,prob,mean,sd) ;
do{
 prev_11k =ilk ;
update_class_prob (n,data,k,prob,mean,sd,class_prob);
update_parameters (n,data,k,prob,mean,sd,class_prob) ;
llk = mix LLk(n,data,k,prob,mean,sd) ;
} while echecktol(11k,prev_11k,epsb ) ;
returnllk ;
}
```

Figure 3.1 EM pseudo-code of the clustering procedure

```
Input: E= {e1, e2..... en} (set of entitites to be clustered)
        K (number of clusters)
        MaxIters (limit of iterations)
Output: C= {$C_1$, $C_2$.....$C_k$} (set of cluster centroids)
        L= {1(e)| e = 1,2 .... n} (set of cluster labels of E)
Foreach e1 € E do
        C1- ej € E (e.g random selection)
End
Foreach e1 € E do
L(e1)- argminDistance (e1, cj) j€ (1....K)
End
Changed – false;
Iter – o;
Repeat
        For each c1 € C do
                UpdateCluster (C1);
        End
        For each e1 € E do
        minDist – argminDistance (e1, cj) j € {1....K}
        if  minDist = 1(e1) then
        1(e1) – minDist;
        Changed – true;
    End
End
iter + +;
until changed = true and iter ≤ MaxIters
```

Figure 3.2 K-means pseudo-code of the clustering procedure

```
DBSCAN (D, eps, MinPts) {
    C=0
    For each point P in dataset D {
        If P is visited
            Continue next point
Mark P as visited
NeighborPts=regionQuery (P, eps)
If sizeof (NeighborPts) <MinPts
Mark P as Noise
Else{
C=next cluster
Expandcluster(P, Neighbor, C, eps, MinPts)
}
}
}
Expandcluster(P, Neighbor, C, eps, MinPts)
Add P to cluster C
For each point P in NeighborPts {
If P is not visited {
Mark p as visited
Neighbour Pts=region Query (P, eps)
If size of (neighbour Pts)>=MinPts
Neighbour Pts=Neighbour Pts joined with Neighbour Pts
}
If P is not yet member of any cluster
Add P to cluster C
}
}
Region query (P, eps)
Return all points within P eps- neighbourhood (including P)
```

Figure 3.3 Pseudo Code of DBSCAN Algorithm

```
COBWEB (root, record)
Input: A COBWEB node root, an instance to insert record
If root has no children them
Children= {copy (root)}
New category (record)\\ update roots statistics
Else
Insert (record, root)
For child in roofs children do
Calculate category Utility for insert (record, child)
Set best1, best2 children w, best CU
End for
If new category (record)) yields best CU then
New category (record)
Else if merge (best1, best2) yield best CU then
Merge (best1, best2)
COBWEB (root, record)
Else if split (best1) yield best CU then
Split (best1)
COBWEB (best1, record)
Else
COBWEB (best1, record)
End if
end
```

Figure 3.4 Algorithm for the Cobweb procedure

Evaluation Results of the Selected Clustering Algorithms based on Product Type Segmentation

Table 4.1 Comparison of a various clustering algorithm for product type dataset (2cluster) category

Clustering method	Time taken	Accuracy	Purity
Cobweb	23.49 secs	14%	0.99
DBSCAN	1.42 secs	14%	0.14
EM	2.71 secs	69%	1
K-means	0.64 secs	100%	1

Table 4.2 Comparison of a various clustering algorithm for product type dataset (5 clusters) category

Clustering method	Time taken	Accuracy	Purity
Cobweb	38.99 secs	17%	1
DBSCAN	1.28 secs	14%	0.14
EM	6.15 secs	71%	1
K-means	0.13 secs	96%	1

Table 4.3 Comparison of a various clustering algorithm for product usage dataset (2 cluster) category

Clustering method	Time taken	Accuracy	Purity
Cobweb	4.29secs	4%	1
DBSCAN	0.23 secs	All unclustered	0
EM	0.36 secs	84%	1
K-means	0.03 secs	99%	1

Table 4.4 Comparison of various clustering algorithm for product usage dataset (5 cluster) category

Clustering method	Time taken	Accuracy	Purity
Cobweb	4.90 secs	4%	0.99
DBSCAN	0.28 secs	All unclustered	0
EM	0.48 secs	82%	1
K-means	0.06 secs	93%	1

1) Discussion of Evaluation of Results of the Segmentation Techniques

K-means assumed to be the best performed of all the techniques evaluated considering the product type evaluation, it posed to be the best considering all the three performance metrics: accuracy, computational time and purity and also considering all categorical iterations. EM algorithm apart from its computational time which takes longer than that of DBSCAN proofs to be the second best only in terms of accuracy and purity, while the effectiveness of Cobweb and DBSCAN cannot be adequately graded as the DBSCAN algorithm was just slightly better performed than the cobweb algorithm as it can be concluded that even as it had a lower computational time than the cobweb algorithm, Its accuracy rate to that of cobweb are close together with 17% and 14% respectively (both are still less than 20%), DBSCAN also not seen or known to be having so much complexity in processing result had several of the instances un-clustered, this made the purity value less than 1(one)signifying a bad cluster quality. Cobweb can then be said to be the least performed as it took a longer computational time than all the four algorithms compared even as the accuracy rate compared to that of DBSCAN was slight. The same follows for the product usage segmentation which can be concluded to buttress the result of the product type segmentation. The very slight difference was gotten in the experimental result of the product usage segmentation to that of the product type segmentation.

Conclusion

The performance evaluation of the four techniques showed that K-means outperformed others in all considered evaluation measure while the EM technique was the second best in terms of accuracy and purity, outperforming the other two. DBSCAN technique was the 3rd best of the selected algorithms even as its computational time is shorter than that of EM while the 4th best performing algorithm has been seen to be the Cobweb algorithm with respect to both computational accuracy, time with purity. Therefore, performance evaluation carried out on clustering algorithms for segmenting consumers has contributed to knowledge by helping future researchers to decide on the most appropriate clustering algorithm for segmenting consumers or any other classification study.

References

Mooi, E., & Sarstedt, M. (2011). *A Concise Guide to Market Research: The Process, Data and Methods Using IBM SPSS Statistics*. Emerald group publishing.

Raj, B., Sunil, S., & Juhi S. (2014). A Comparative Analysis of Clustering Algorithms. *International Journal of Computer Applications*, *100*(15), 35-39.

Jeff, S. (2014). *Why and How to Segment Your Customers*. Measuring U online Publication.

Sankar, R. (2011). Customer Data Clustering Using Data Mining Technique. *International Journal of Database Management Systems (IJDMS)*, *3*(4), 1-11.

Schonfeld, E. (2016-02-28). "CNN's – The Webtop". con Mindy con mine.com.

Gartner Group. (2015). *IT Glossary-Data Mining. Garther Research Company, Stamford Connecticut United States*. Gartner incorporate.

Narendra, S., Aman, B., & Ratnesh, L. (2012). Comparison the Various Clustering Algorithms Of Weka Tools. *International Journal of Emerging Technology and Advanced Engineering*, *2*(5), 73-79.

Chaudhari, B., & Parikh, M. (2012). A Comparative Study of clustering algorithms Using weka tools. *International Journal of Application or Innovation in Engineering & Management (IJAIEM)*, *1*(2), 154-158.

HubPages Inc. (2016). Principles of Marketing - The Market Segmentation Process.

Miyuri S. (2017). Effects of Price Reforming Tactics on Consumer Perception. *Journal of Retailing and consumer services*, *34*, 82 -87.

Nottingham City Council GIS Team 2016.

Elena, L.M. (2016). *How to Strengthen Customer Loyalty Using Customer Segmentation*. Bulletin for the Transitivity University of Brasor V: Economic Sciences, *9*(2), 51-60.

Sethughes. (2013). *Principles of marketing-the market segmentation process*.Toughnickel inc.

WIKIPEDIA: The Free Encyclopedia.

Sapna, J., Afshar Aalam, M., & Doja, M.N. (2010). K Means Clustering Using Weka Interface. *Proceedings of the International Journal of Computer Applications and Management*.

Witten, I.H., Frank, E., Hall, M.A., & Pal, C.J. (2011). Data Mining: Practical machine learning tools and techniques, 3rd Edition. *Morgan Kaufmann, San Francisco (CA).* Retrieved -01-19, 2011.

Al Hayani, B., & Ilhan, H. (2020). Image transmission over decode and forward based cooperative wireless multimedia sensor networks for Rayleigh fading channels in medical Internet of Things (MIoT) for remote health-care and health communication monitoring. *Journal of Medical Imaging and Health Informatics*, *10*(1), 160-168.

Alhayani, B., & Ilhan, H. (2020). Efficient cooperative imge transmission in one-Way mult-hop sensor network. *International Journal of Electrical Engineering Education*, 1–17.

Alhayani, B., & Ilhan, H. (2017). Hyper spectral image classification using dimensionality reduction techniques. *International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering*, *5*, 71-74.

Alhayani, B., & Milind, R. (2014). Face recognition system by image processing. *International journal of electronics and communication engineering & technology (IJCIET)*, *5*(5), 80–90.

Al-Hayani, B., & Ilhan, H. (2020). Visual Sensor Intelligent Module Based Image Transmission in Industrial Manufacturing for Monitoring and Manipulation problems. *Journal of Intelligent Manufacturing*, *4*, 1-14.

Mahajan, H.B., & Badarla, A. (2018). Application of Internet of Things for Smart Precision Farming: Solutions and Challenges. *International Journal of Advanced Science and Technology*, 37-45.

Mahajan, H.B., & Badarla, A. (2019). Experimental Analysis of Recent Clustering Algorithms for Wireless Sensor Network: Application of IoT based Smart Precision Farming. *Journal of Advanced Research in Dynamical & Control Systems*, *11*(9).

Mahajan, H.B., & Badarla, A. (2020). Detecting HTTP Vulnerabilities in IoT-based Precision Farming Connected with Cloud Environment using Artificial Intelligence.*International Journal of Advanced Science and Technology*, *29*(3), 214 - 226.

Bibliography of Authors

Yaser Issam Hamodi received B.Sc., degrees in Computer Engineering - College of Electrical And Electronic Techniques, Foundation of Technical Education, Baghdad, Iraq, in 2004, and M.Sc. degree in computer engineering from Cankaya University – Ankara –Turkey in 2012. He is currently in Ministry of Higher Education & Scientific Research, Iraq.

Ruaa Riyadh Hussein received B.Sc. degree in Computer Science-College of Education- Ibn al-Haytham, Baghdad University, Baghdad, Iraq, in 2007, and M.Sc. degree in Computer Science- College of Engineering and Computer Science, University of Central Florida, Florida, USA, in 2014. She is currently a teacher in College of Education for Girls, Al-Iraqia University, Iraq.

Naeem Thjeel Yousir received B.Sc., degree in Computer Science and Statistics, Al-Rafidain University College, Baghdad, Iraq in 1998, and M.Sc degree Computer Science in 2003, and Ph.D. degree in Computer Engineering and Communications from Hacettepe University – Ankara –Turkey in 2014.