IJCSE International Journal of Computer Sciences and Engineering Open Access
Survey Paper Vol.-7, Issue-6, June 2019 E-ISSN: 2347-2693

Implementation of Clustering Techniques in Various Fields: A Survey

Syed Zishan Ali[1*], Surbhi Chundawat[2], Shrinjanee Shukla[3], Akshat Choudhary[4]

[1,2,3,4]Dept. of Computer Science & Engineering, Bhilai Institute of Technology Raipur, CSVTU, Chhattisgarh, India

Corresponding Author: zishan786s@gmail.com, Tel.: +919977701133

DOI: https://doi.org/10.26438/ijcse/v7i6.704707 | Available online at: www.ijcseonline.org

Accepted: 13/Jun/2019, Published: 30/Jun/2019

***Abstract*—**In Today's world clustering techniques are used in different fields like Image Classification, AI, E-commerce, etc. The advantage of clustering is that it provides a summarized output for the user, where the user can obtain the exact results and predict the outcome. Once implemented, clustering offers a clarified outcome to the user which strikes out the necessity for further research and development. Clustering plays a major role in areas where there is a probability and necessity of figuring out similar as well as dissimilar objects. One of the major aspects of data mining is to differentiate between objects based on their properties or attributes. This paper focuses on three such fields namely Medical Science, Agriculture and E-Commerce.

***Keywords*—** Clustering, Medical Science, Agriculture, E-commerce.

I. INTRODUCTION

The ability to form meaningful groups of objects is one of the most fundamental modes of intelligence. Humans perform this task with remarkable ease. Clustering analysis is a tool for exploring the structure of data. The core of cluster analysis is clustering; the process of grouping objects into clusters such that the objects from the same cluster are similar and objects from different clusters are dissimilar. Unlike classification, clustering doesn't require assumptions about category labels that tag objects with prior identifiers. Therefore, clustering is an unsupervised learning technique versus classification which belongs to supervised learning.

II. TYPES OF CLUSTERING

A. Hierarchical methods

The methods proceed successively by either merging the smaller clusters into larger ones, or by splitting the larger clusters. The result of algorithm is a tree of clusters, called Dendrogram which creates a hierarchical decomposition of the given data objects. The method can be classified as being either agglomerative or divisive i.e. bottom-up or top-down, based on how the hierarchical decomposition is formed.

B.Partitioning methods

The methods attempt to directly decompose the data into disjoint clusters i.e. it first creates initial *k* partitions, where parameter *k* is the number of partitions to construct; then it uses an iterative relocation technique that attempts to improve the partitioning by moving objects from one group to another. The method works well for finding spherical-shaped clusters in small or medium-sized databases.

C.Density- Based methods

The methods are grouped based on density conditions. It either grows neighbouring objects into clusters based on density conditions. It either grows cluster according to the density of neighbourhood or according to some density function.

D.Grid Based Methods

The grid based clustering approach uses a multi resolution grid data structure. It quantizes the object space into a finite number of Cells that form a grid structure on which all of the operations for clustering are performed. The main advantage of the approach is its fast processing time, which is typically independent of the number of data objects, yet dependent on only the number of cells in each dimension in the quantized space.

E.Model Based Methods

The methods hypothesize a model for each of the clusters and find the best fit of the data to that model. Typical model based methods involves statistical approaches (COBWEB, CLASSIT, autoclass) or neural network approaches (competitive learning and self organizing maps).

II. RELATED WORK

IMPLEMENTATION OF CLUSTERING IN VARIOUS FIELDS:

E-commerce

It has been seen throughout the world that the number of people connected to the internet is rising exponentially. And with rise in popularity of internet, comes rise in use of internet based applications. The life of people is becoming more and more fast and efficient day by day. Because of this there has been a rise in number of people using E-Commerce website and the demand for better E-Commerce websites has also increased. With this surge of customers in E-Commerce, the sellers need to find a way to see which customers are the most profitable and to achieve that we have studied from the paper [1] on how the E-Commerce websites and sellers have tackled this issue.

Problem with E-Commerce [2]

1) The Fake Problem: The Fake Problem has been a huge issue for E-Commerce for a long time. The customers have to suffer a lot because of this issue as it further delays the date on which they get the product even as they will return the product if turns out to be fake and also file complaint against it. The mere existence of fake products has set back the rate growth of E-Commerce by a lot.
2) Problem of Asymmetric Information: In E-Commerce, the buyer does not get to actually see the product that he/she is buying. The consumer has to trust the information provided by the E-Commerce website. This is also an issue as the seller may give a false description of the product which can lead to unhappy customers as it will make the E-Commerce website much less trustworthy and the news can spread rather quickly because of all the development in Social Media which will overall result in the E-Commerce website getting much less traffic.

As per the model described in paper [1], self-organising maps and k-means are used to effectively perform cluster analysis. By applying SOM to a cluster of 730 customers, it was found that the number 9 is the best number of clustering based on the characteristics of LRFM. The result of this method is shown below. After that the number nine of cluster (k) generated by SOM can be used as a parameter for the next step. In this step, K-means method is applies to find the final solution. Now, we have 9 clusters of customers that have similar LRFM behaviour. The figure below gives a summary of the clustering of these nice clusters according to LRFM. If the average L,R,F,M value of the cluster exceeds the average L,R,F,M value then an over bar appears. And if the value is less than the average values then an under bar appears.

Cluster	Size	Average of L	Average of R	Average of F	Average of M	Pattern
Cluster 1	95	697.74	342.14	16.33	3 559.70	$\overline{L}\,\overline{RF}\,\underline{M}$
Cluster 2	83	770.63	368.29	5.81	1 743.49	$\overline{L}\,\overline{R}\,\underline{F}\,\underline{M}$
Cluster 3	88	143.13	84.73	2.41	684.45	$\underline{L}\,\underline{R}\,\underline{F}\,\underline{M}$
Cluster 4	60	220.00	362.67	6.80	6 284.83	$\underline{L}\,\overline{R}\,\underline{F}\,\overline{M}$
Cluster 5	73	211.21	182.42	13.37	4 074.19	$\underline{L}\,\underline{R}\,\overline{FM}$
Cluster 6	72	679.21	381.04	16.03	11 127.82	$\overline{LRFM}$
Cluster 7	104	296.78	367.78	16.76	6 340.59	$\underline{L}\,\overline{R}\,\overline{FM}$
Cluster 8	77	537.95	165.97	7.25	2 076.83	$\overline{L}\,\underline{R}\,\overline{FM}$
Cluster 9	78	355.33	334.62	4.23	924.11	$\underline{L}\,\overline{R}\,\underline{F}\,\underline{M}$
Total	730	438.86	287.90	10.16	3 986.63	

Figure 1.Descriptive Statistics of Nine Clusters Based on K-Means Method

We observe that Cluster 4 has the least important customers, while the number of customers in cluster 1 and 7 are very important. Cluster 3 has the least value of LRFM compared to other clusters. We observe that the customers in Cluster 3 have recently joined the online store and these customers do not spend much money. This cluster can be named as the uncertain new customer's cluster. If we see carefully it can be seen that the Cluster 9 also has new customers that have purchased some product recently.

Agriculture

Agriculture is an important factor in supporting the economy of our country. About two-third of Indian population depends on agriculture for its living. The productivity of agriculture is very low as compared to the food demands as only one-third of the cropped part is actually irrigated. Since the demands are increasing and productivity is decreasing, the researchers and government are putting extra efforts to increase productivity. As a result the agriculture data is increasing. As data increases we need more methods to extract these data whenever required. To ease the handling of bulk data we use various techniques based on data mining. The Data Mining techniques applied on Agricultural data include k-means, bi clustering, and k nearest neighbour, Neural Networks (NN) Support Vector Machine (SVM), Naive Bayes Classifier and Fuzzy c-means.

III. CLUSTERING TECHNIQUES

In clustering we organize the data into groups or clusters such that these clusters are placed close by one another and can be accessed easily. The clustering technique groups similar or dissimilar type of data instances into subsets such that same type of groups are placed together and a separate group is made for other type of similar instances .

k-means

K-means algorithm helps in partitioning the n-dimensional data into *k* sets and minimizes the mean distance within each set. The input parameter, k is taken which partitions set of n objects into k cluster such that the resulting similarity within the clusters is high but the similarities with other the clusters is low. The squared Euclidean distance and the sum

of the squared differences across variables are the most commonly used distance measures [4]

Application of k-means algorithm in agriculture:
The k-means algorithm can be used for soil classifications using technologies based on GPS. K-means can also be employed in Plant, soil, and residue regions of interest which are classified by using color images, grading apples before marketing, Monitoring water quality changes, detecting weeds in precision agriculture. The k-means approach can be done for prediction of problems that can arise during fermentation of wine. Advance knowledge of the fact that the process of wine fermentation could get stuck or be slowed down can help the enologist to correct the problem hence a good fermentation process can be ensured [3].

K-nearest neighbour: Application in agriculture:
The K-NN is a data mining technique that uses the principle in which entries to place an unknown entity in a particular group on the basis of its neighboring groups.
The K-NN is a data mining technique that uses the algorithm which facilitates us in placing an unknown sample in a group just by looking at the other surrounding samples. But for the classification to be accurate it is required that the unknown sample is surrounded with many groups of the known samples. Classification is not possible if there is only one known sample.

In simulating daily precipitations and other weather variables, the k-nearest algorithm is used. It is also used in determining the parameters for soil water ratio and can also be employed in climate forecasting. [5]

Self-Organizing Maps in the field:

A. *Agriculture*

Visualization of Agriculture data can be done by using Self-Organizing Maps. In this, first the data is acquired and pre-processed. Then that data is built, evaluated and optimized. After the collection of data, the neural network predictor uses the collected past data to predict the future yield. RuB et al in their paper "Visualization of Agriculture Data Using Self-Organizing Maps" [6] showed the use of this prediction to optimize the use of fertilizers economically. For each part of the field different amount of fertilizers are used. The authors have visualized the data with the help of labels and u-matrix and established co-relations in the data which is visible in maps.

Medical science
Today's medical field system focuses on building and enhancing the decision making capability which is a requirement in medical and engineering fields. Availability of data regulates the growth of applications. Upliftments have been observed in the precision and sensitivity or reactivity of diagnostic trials. From witnessing outermost symptoms to advanced lab researches and compound imaging techniques are increasing which allows exclusion of internal tests and trials. The betterment in accuracy has assuredly developed broader specifics about the patients. Basically, the procedure of obtaining affirmations/evidences to figure out possible reasons regarding the patient's major symptoms and traits from all the other predictable causes leading to the symptoms is referred as setting up of a medical diagnosis.

Health maintenance related data mining is one of the most demanding fields of the implementations in data mining and knowledge innovations. The cause of challenging and demanding field is due to the data sets of vast, compound, multiple, hierarchical, time series and quality-wise differentiable. The procedure of data amalgamation is extremely challenging job. Medical diagnosis is known to be subjective and depends not only on the available data but also on the experience of the physician and even on the psycho - physiological condition of the physician. A number of studies have demonstrated that the diagnosis of one patient can differ significantly if the patient is tested by different physicians or even by the same physician at various times. As the available healthcare datasets are fragmented and diffused in nature, thereby making the process of data integration is a highly challenging task [7].Considering a small subset of medical datasets, algorithms have been formed and accurate work have been achieved by some of them [8]. Applied to a much larger generalized dataset, for every medical field, obtaining accurate results has yet been very difficult. Lliad[9] is an expert system program that uses Bayesian classification to estimate the posterior probabilities of various diagnoses under consideration, given the symptoms present in a case.

k-means in Medical Science:
The author Dr. Bushra M. Hussan in her paper [10] has performed experiment using k-means algorithm on the Pima Indian diabetes data set, which included 768 complete. Instances (described by 8 features like blood pressure, glucose tolerance test, etc).The values are taken as input in the k-means algorithm in order to create clusters. Initially it takes random value for creating clusters and after some iteration when the mean value becomes constant we get the final clusters.

CONCLUSIONS

This paper concludes a summarized study of clustering techniques of data mining in different areas. Different algorithms have been implemented in E-commerce, Agriculture and Medical Science. We have concluded different characteristics of algorithms and their results have been shown on their respective areas. In E-commerce the algorithms have been employed to find the group of most beneficial customers in a consistent manner, in agriculture the algorithms have been used to detect the quality of soil as well as plants and group them accordingly while in medical science, these algorithms helps in the prediction of causes which may lead to a medical condition. The paper will prove helpful for the researchers to identify the use of different types of clustering techniques and hence be able to fetch the desired information.

REFERENCES

[1] Rachid Ait daoud, Abdellah Amine, Belaid Bouikhalene, Rachid Lbibb," Customer Segmentation Model in E-commerce Using Clustering Techniques and LRFM Model: International Journal of Computer and Information Engineering Vol:**9**, No:**8, 2015 2000International.**

[2] HuaYu, XiZhang, "Research on the application of IoT in E-commerce", 2017 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC).

[3] Patel, Hetal, Patel, Dharmendra (2014). A Brief survey of Data Mining Techniques Applied to Agricultural Data. International Journal of Computer Applications (0975 – 8887) **Volume 95– No. 9, June 2014.**

[4] Yadav, Dileep Kumar (2015). A Comparative Analysis Of Clustering Algorithms For Agricultural Data. International Journal of Current Research Vol. **7**, Issue, **07**, pp.**18361-18364**, July, 2015

[5] Mucherino, A., Papajorgji, P., & Pardalos, P. (2009). Data mining in agriculture (Vol. **34**). Springer.

[6] Ruß, Georg, Kruse, Rudolf, Schneider, Martin, Wagner, Peter (2008). Visualization of Agriculture Data Using Self-Organizing Maps. International Conference on Innovative Techniques and Applications of Artificial Intelligence SGAI 2008: Applications and Innovations in Intelligent Systems XVI pp 47-60. Springer.

[7] K.J. and G.W. Moore Cios, "Uniqueness of medical data mining ," Artificial Intelligence in Medicine, **pp. 1-24, 2002**.

[8] Anamika Gupta, Naveen Kumar, and Vasudha Bhatnagar," Analysis of Medical Data using Data Mining and Formal Concept Analysis, World Academy of Science, Engineering and Technology International Journal of Medical and Health Sciences Vol:**1**, No:**11, 2007** .

[9] H. R. Warner and O. Bouhaddou, "Innovation review: Iliad—A medical diagnostic support program" Top Health Inf. Manage., vol. **14**, no. 4, pp**. 51–58, 1994**.

[10**]** Dr. Bushra M. Hussan," Data Mining based Prediction of Medical data Using K-means algorithm", Computer Science Department - Collegeof Science - Basrah University, Basrah Journal of Science (A) Vol.**30**(1),46-56 2012.

Authors Profile

Mr.Syed Zishan Ali pursued Bachelor of Engineering from CSVTU, Bhilai in 2009 and Masters in Computer Science in 2014 from CSVTU.He is currently working as Assistant Professor in Department of computer Science & Engineering , BIT Raipur since 2012.He has published Research papers in IEEE which is availabe online and also presented paper on SPRINGER conferences.

Ms. Surbhi Chundawat is current a scholar under Mr.Syed Zishan Ali.

Ms.Shrinjanee Shukla is current a scholar under Mr.Syed Zishan Ali

Mr.Akshat Choudhary is current a scholar under Mr.Syed Zishan Ali

