ISSN No.: 2455-6130
Volume 1, Issue 8, 2016, pp. 22-32

website: www.ijoetr.com

International Journal of Emerging Trends in Research

A proposed efficient data algorithm on basis of clustering using data mining

Saloni Chaudhary[1] , Tarun Kumar[2]

[1,2]Department of Computer Science & Engineering, Geeta Institute of Engineering and Technology, Kanipala Haryana, INDIA

Abstract

In this paper, we present a new clustering algorithm for unsupervised clustering process. Recently, a lot of researchers have a significant attention in developing clustering algorithms. The main problem in clustering is that we do not have preceding information knowledge about the particular dataset. In addition, the alternative of input constraint such as the number of clusters, number of nearest neighbors and other issue in these algorithms make the clustering more challengeable area. Thus any defective selection of these constraints gives up bad clustering results. As well these algorithms suffer from insufficient correctness when the dataset hold clusters with dissimilar complex form, densities, dimension, noise and outliers. We present experiments that present the effectiveness of our new algorithm in discovering clusters with different non-convex form dimension, densities, noise and outliers even though the poor initial conditions. These experiments show the superiority of our proposed algorithm when comparing with most competing algorithms.

Keywords: Data clustering; Data mining; Unsupervised clustering; grouping of clustering; WEKA tool.

1. Introduction

Data Clustering is one of the most important issues in data mining. Clustering is a process of discovering homogenous groups of the studied objects. Clustering is an important technique with numerous applications, such as marketing and customer segmentation. Clustering typically group data into sets in such a way that the intra-cluster similarity is maximized and while inter-cluster similarity is minimized. Clustering is an unsupervised learning. Clustering algorithms examines data to find groups of items that are similar.

Grouping can be viewed as the most vital unsupervised learning issue; thus, as every other issue of this kind, it manages discovering a structure in a gathering of unlabeled data. A free meaning of grouping could be "the procedure of sorting out items into gatherings whose individuals are comparative somehow". A cluster is along these lines a gathering of items which are "comparable" in the middle of them and are "divergent" to the articles fitting in with different clusters.

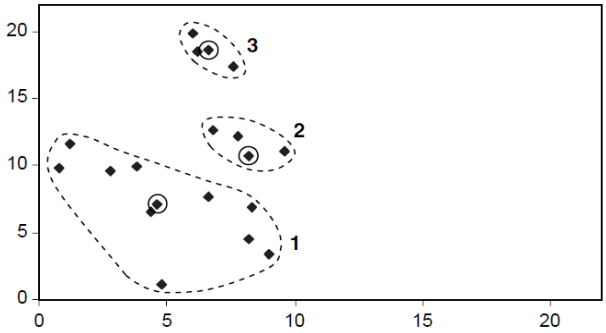

Figure 1: Clustering of Data

2. Related work

Lloyd, S., 1982 [1] describes in the paper that the data mining process is to extract information from a huge data set and convert it into an different utilizable form for further use. Can, F.; Ozkarahan, E. A., 1990 [2] computational many-sided quality is a significant test in transformative calculations because of their requirement for rehashed wellness capacity assessments.In the survey paper, Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, 1996 [3] describes various types of clustering techniques in data mining is done. They states that data mining refers to remove useful information from huge amounts of data.

Andre Baresel, Harmen Sthamer, Michael Schmidt, 2002 [4] proposed find CBLOF Algorithm for identifying anomalies. This calculation processes the estimation of CBLOF for every record which decides the level of record's deviation. He Zengyou, Xu Xiaofei, Deng Shenchun, 2002 [5] highlights the investigation of wellness capacities utilized as a part of the stream and area of mechanical autonomy. M. Davarynejad, M.-R.Akbarzadeh-T, N.Pariz, 2007 [6] describes the role of Data mining Techniques in Education system.

Jerzy Stefanowski, 2009 [7] represents the grouping of data objects such that the objects within a cluster are related to one another and unrelated from the objects in other groups. Many of clustering algorithm is obtainable to analyze data. Aditya Desai, Himanshu Singh, Vikram Pudi, 2011 [8] uses evolutionary algorithms (EA) to look for particular test data that give high basic scope of the product under test. In Hans-Peter Kriegel, Peer Kröger, Jörg Sander, Arthur Zimek, 2011 [9] introduces a near study on number of similitude measures, for example, Goodall, Occurrence Frequency, Overlap, Inverse Occurrence Frequency, Burnbay, Gambaryan, and Smirnov.

3. Need of proposed work

3.1 Problem identification on clustering of data sets

At the first occurrence, there is a Data Vault (D). In the Data Archive, there will be number of exchanges or data sets or value-based data. Every exchange or record or data set is termed as R(Ti). The Data Set paying little respect to the quality and related parameters will be qualified and make headway the wellness capacity displaying. Once the wellness capacity is connected taking into account the percentile based estimation, it will shape the new criteria for the incorporation in the clusters. At last the arrangement of clusters will be created with proficient results and ideal time parameter.

3.2 Objectives

Objectives works in the following steps:

- *To devise and execute a novel and productive procedure for dynamic and in addition powerful group development.*
- *To apply and get the important records in manifestation of the total values or groups for brainpower and expectations.*
- *To investigate the proposed cluster development calculation with the current strategy and to demonstrate the viability of the proposed work.*
- *To devise a novel wellness capacity to the value-based data so that the qualification or significance of the record can be investigated.*

3.3 Classical approach

Classical Calculation over and over peruses tuples one by one from the dataset. At the point when the first tuple arrives, another cluster is framed. In existing approach every tuple has a place with one cluster just.

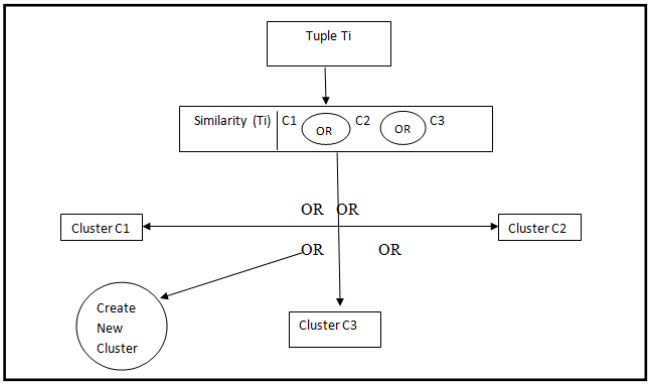

Figure 2: Classical Approach of Clustering

3.4 Proposed approach

In proposed calculation, assume there are n tuples. A wellness quality is appointed to every tuple utilizing the wellness capacity. In the event that the wellness estimation of the tuple is equivalent to or about equivalent to the edge estimation of the produced arrangement of irregular clusters.

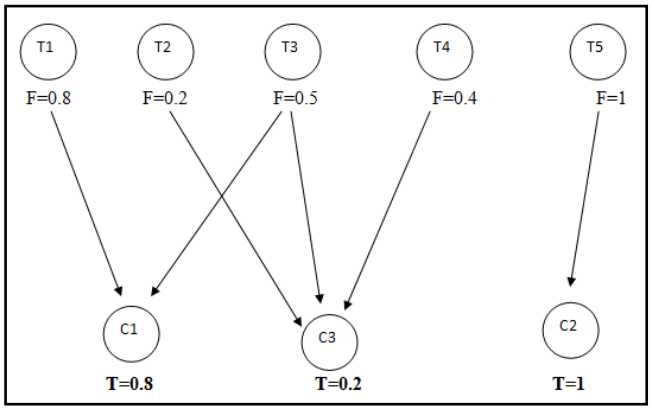

Figure 3: Proposed Approach (T= Threshold, F= Fitness Value)

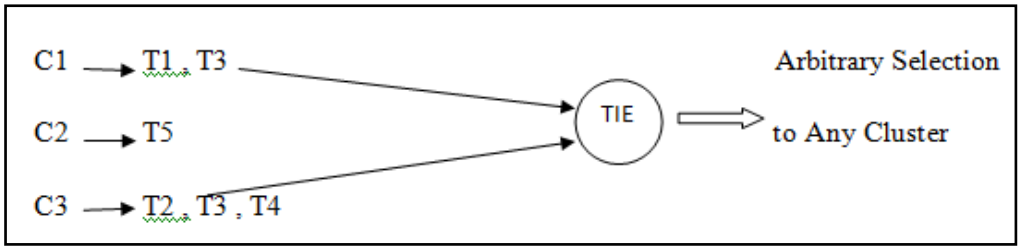

Figure 4: Analysis of Fitness and Selection

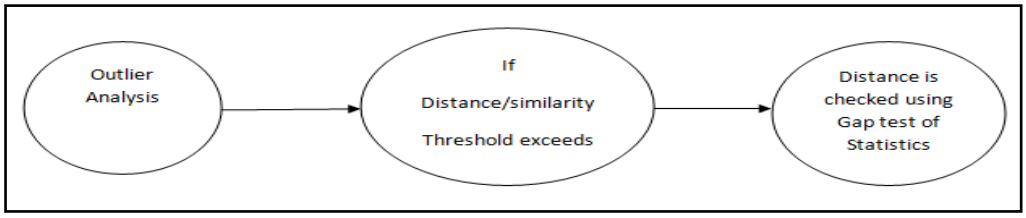

Figure 5: Outlier Detection

The proposed methodology is indicated in figure 6 with the assistance of the stream chart. The preparation dataset is chosen either arbitrarily or consecutively from the information distribution center. At that point compute wellness estimation of every tuple.

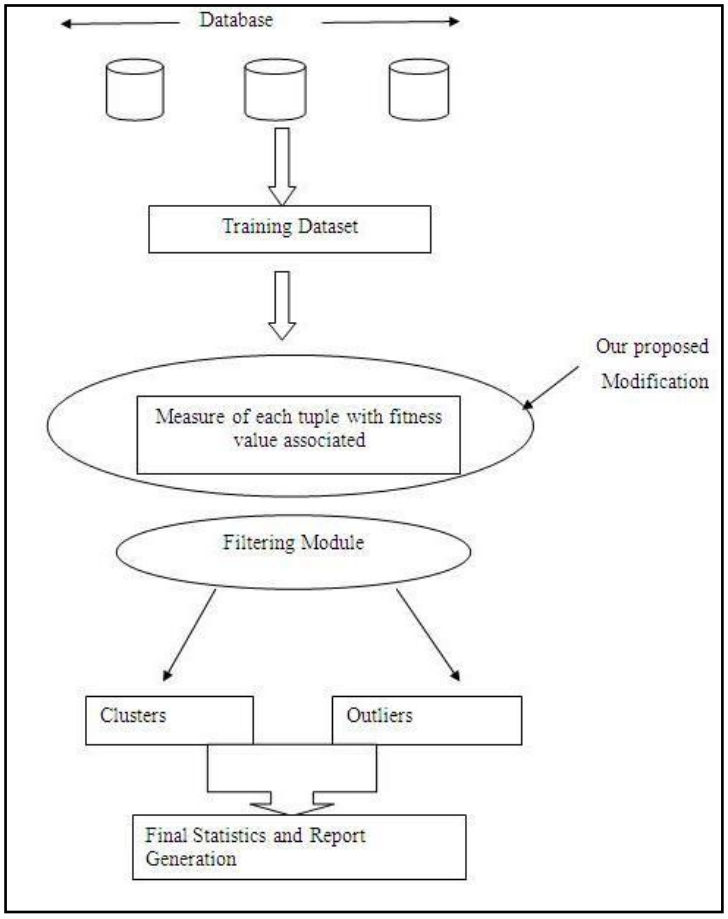

Figure 6: Flow diagram of proposed approach

4. Experimental Work

4.1 Data Set Used for Implementation

Table 1: Implementation Data Set

SHOPPING ID	Price	Weekday	Date	Month	Year	Items
018	600	Friday	17	January	2016	42
019	750	Sunday	1	January	2016	67
019	750	Thursday	26	January	2016	71
020	400	Sunday	5	January	2016	47
020	400	Thursday	30	January	2016	55
020	400	Monday	27	January	2016	72
001	1000	Thursday	15	January	2016	67
006	2000	Saturday	6	January	2016	56
005	1200	Wednesday	23	January	2016	43
010	2500	Monday	15	January	2016	46
019	750	Sunday	11	January	2016	41
001	1000	Monday	14	January	2016	69
007	1400	Monday	17	January	2016	13
007	1400	Wednesday	5	January	2016	90
015	1200	Sunday	9	January	2016	4
001	1000	Sunday	3	January	2016	67
003	2400	Tuesday	4	January	2016	6
001	1000	Wednesday	6	January	2016	325
008	8000	Friday	4	January	2016	7
001	1000	Friday	18	January	2016	78

Cluster 1	**Cluster 2**	**Cluster 3**
Most Favourite Products	Average Products	Very Less Bought Products
PID (012)	SHOPPING ID (017) SHOPPING ID (001)	SHOPPING ID (020) SHOPPING ID (010) SHOPPING ID (009) SHOPPING ID (007) SHOPPING ID (008) SHOPPING ID (014) SHOPPING ID (006)

Cluster 1	Cluster 2	Cluster 3
Most Favourite Products	Average Products	Very Less Bought Products
	No Product meet the Fitness Function Criteria	SHOPPING ID (012) SHOPPING ID (017) SHOPPING ID (001) SHOPPING ID (003) SHOPPING ID (015) SHOPPING ID (018) SHOPPING ID (004) SHOPPING ID (002) SHOPPING ID (016) SHOPPING ID (019) SHOPPING ID (005) SHOPPING ID (011) SHOPPING ID (013) SHOPPING ID (020) SHOPPING ID (010) SHOPPING ID (009) SHOPPING ID (007) SHOPPING ID (008) SHOPPING ID (014) SHOPPING ID (006)

Percentile based execution (proposed) takes less execution time than existing based usage. The qualification and the best fit parameter are no place being measured in the current strategy and in addition giving the turnaround time higher than the proposed method. The established method is creating and grouping the information things in the cluster that may not be valuable in the learning revelation.

In the proposed system, a select estimation is thought seriously about and actualized on the same exchange information for examination of the consequences of the proposed when contrasted with the current procedure.

4.2 Evaluation graph between existing and proposed approach

The graph has been plotted to represent the pattern and behavior of the cluster formation process. Using the implementation of cluster formation, it is shown that the proposed technique is producing better results as compared to the classical technique.

The graph has been plotted for the warehouse records with respect to the execution time. The investigation has been performed for the slabs or layers of the records for the efficiency analysis.

Figure 7: Data items in the best fit cluster based on intelligence

5. Results

5.1 Weka (data mining product) screenshots

The following graphs and associated statistical values are fetched from the machine learning tool WEKA. The graphs individually analyze each parameter of the data set in terms of the maximum, minimum, mean and standard deviation in the overall transactions.

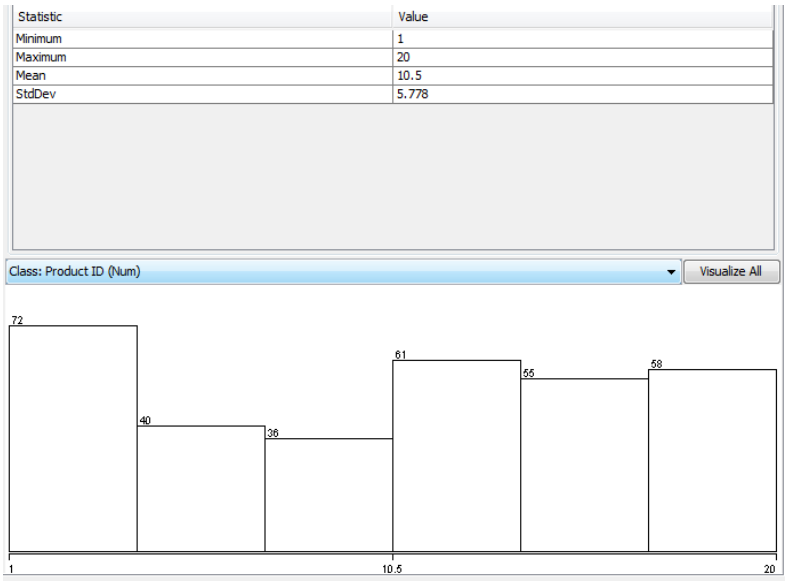

Figure 8: WEKA Analysis 1

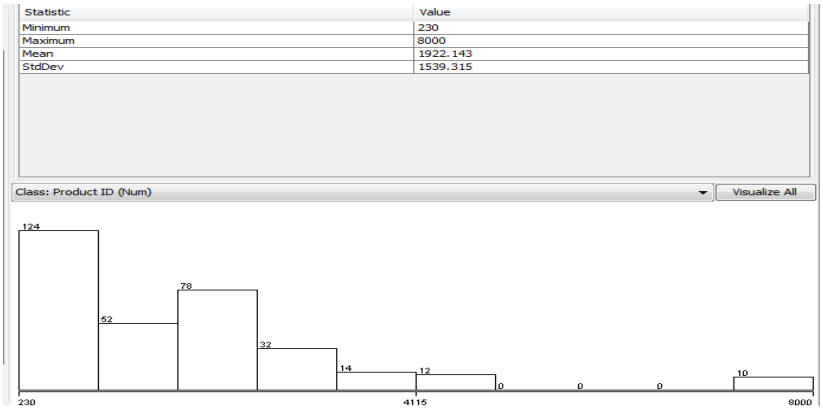

Figure 9: WEKA Analysis 2

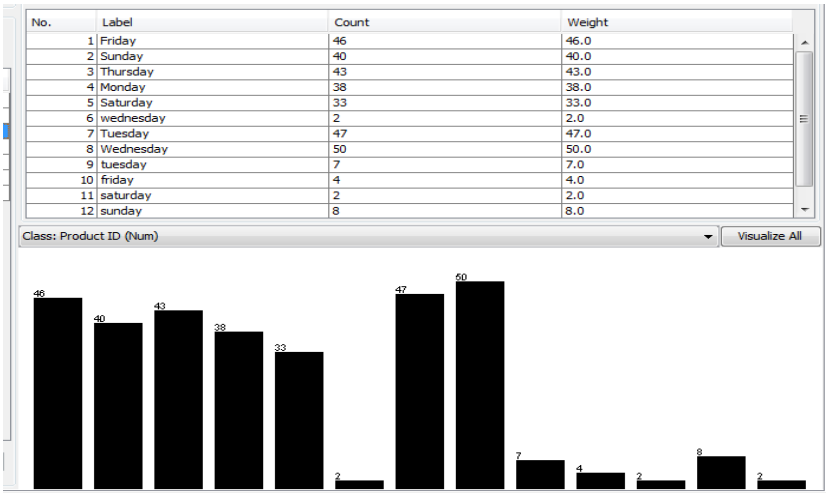

Figure 10: WEKA Analysis 3

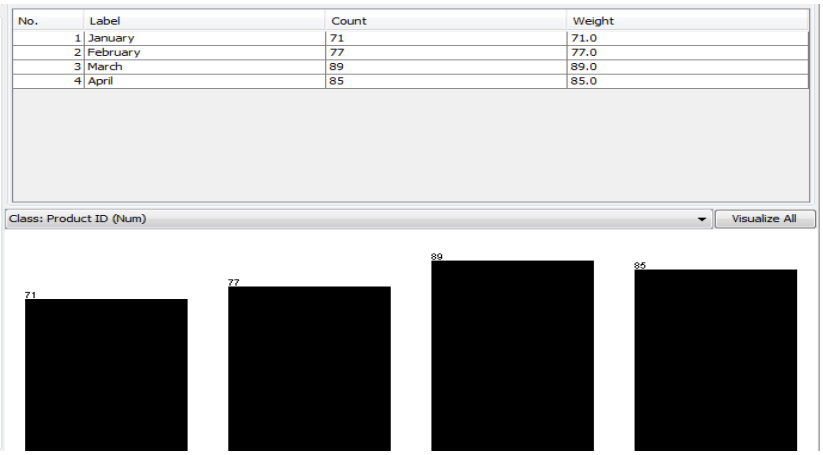

Figure 11: WEKA Analysis 4

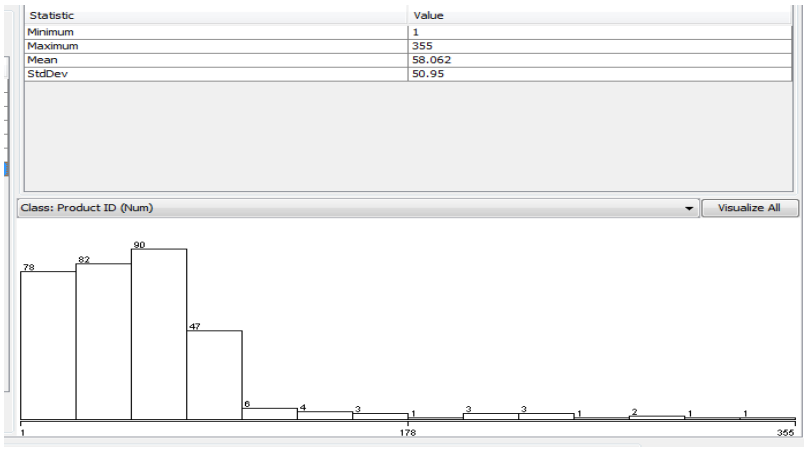

Figure 12: WEKA Analysis 5

Table 2: Execution Time in Classical and Proposed Approach

Attempt ID	Classical Approach	Proposed Approach
1	2.6176271438599	0.0019679069519043
2	3.0634899139404	0.0010800361633301
3	2.1322290897369	0.00083780288696289
4	2.2837190628052	0.00086307525634766
5	2.5840640068054	0.00063014030456543
6	2.9248859882355	0.00073790550231934
7	11.58077788353	0.00071811676025391
8	2.5234549045563	0.00067400932312012
9	2.6024219989777	0.00077295303344727
10	5.7838900089264	0.00071883201599121
11	2.6262028217316	0.0057981014251709
12	2.5887069702148	0.0013151168823242
13	4.1497797966003	0.00069904327392578
14	2.4847829341888	0.0008389949798584
15	2.5324690341949	0.0014100074768066
16	2.3420181274414	0.00042510032653809
17	3.2956490516663	0.00064206123352051
18	4.5505108833313	0.00039505958557129
19	4.3060128688812	0.00038504600524902
20	2.3504519462585	0.00036907196044922
21	11.376751899719	0.00051712989807129
22	2.7007350921631	0.0025358200073242
23	2.5504150390625	0.00075697898864746
24	6.0459179878235	0.00081181526184082
25	3.7791841030121	0.00067901611328125
26	2.3425049781799	0.0012350082397461
27	2.2166659832001	0.00090813636779785
28	2.5292708873749	0.00077009201049805
29	2.4779279232025	0.00055909156799316
30	2.2424299716949	0.00081896781921387
31	2.2484631538391	0.00045394897460938
32	2.127748966217	0.00061202049255371
33	2.3827328681946	0.00057411193847656

6. Conclusion

Cluster investigation itself is not one particular calculation, but rather the general errand to be fathomed. It can be accomplished by different calculations that contrast fundamentally in their thought of what constitutes a cluster and how to effectively discover them. Prevalent ideas of clusters incorporate gatherings with little separations among the cluster individuals, thick zones of the information space, interims or specific factual dispersions. Clustering can accordingly be detailed as a multi-target streamlining issue.

The following points throw light on why clustering is required in data mining −

- **Scalability** − we need highly scalable clustering algorithms to deal with large databases.
- **Ability to deal with different kinds of attributes** − Algorithms should be capable to be applied on any kind of data such as interval-based (numerical) data, categorical, and binary data.
- **Discovery of clusters with attribute shape** − the clustering algorithm should be capable of detecting clusters of arbitrary shape. They should not be bounded to only distance measures that tend to find spherical cluster of small sizes.
- **High dimensionality** − the clustering algorithm should not only be able to handle low-dimensional data but also the high dimensional space.
- **Ability to deal with noisy data** − Databases contain noisy, missing or erroneous data. Some algorithms are sensitive to such data and may lead to poor quality clusters.
- **Interpretability** − the clustering results should be interpretable, comprehensible, and usable.

References

[1] Lloyd, S. (1982). Least squares quantization in PCM. *IEEE transactions on information theory*, *28*(2), 129-137.

[2] Can, F., & Ozkarahan, E. A. (1990). Concepts and effectiveness of the cover-coefficient-based clustering methodology for text databases. *ACM Transactions on Database Systems (TODS)*, *15*(4), 483-517..

[3] Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996, August). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd* (Vol. 96, No. 34, pp. 226-231).

[4] Baresel, A., Sthamer, H., & Schmidt, M. (2002, July). Fitness Function Design To Improve Evolutionary Structural Testing. In *GECCO* (Vol. 2, pp. 1329-1336).

[5] He, Z., Xu, X., & Deng, S. (2006). Improving categorical data clustering algorithm by weighting uncommon attribute value matches. *Comput. Sci. Inf. Syst.*, *3*(1), 23-32..

[6] Davarynejad, M., Akbarzadeh-T, M. R., & Pariz, N. (2007, September). A novel general framework for evolutionary optimization: Adaptive fuzzy fitness granulation. In *2007 IEEE Congress on Evolutionary Computation* (pp. 951-956). IEEE.

[7] Stefanowski, J. (2009). Data Mining-Clustering. *University of Technology, Poland*.

[8] Desai, A., Singh, H., & Pudi, V. (2011, May). DISC: data-intensive similarity measure for categorical data. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining* (pp. 469-481). Springer Berlin

[9] Heidelberg.Kriegel, H. P., Kröger, P., Sander, J., & Zimek, A. (2011). Density-based clustering. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *1*(3), 231-240.

[10] Gupta, S. (2016). A Comparative Analysis of Wired and Wireless Network Architecture. *International Journal of Emerging Trends in Research*, *1*(1), 05-11.

[11] Chaudhary, S., & Kumar, T. (2016) An Observed Study of Clustering in Data Mining. *International Journal of Emerging Trends in Research*, *1*(4), 29-35.