K-Means and K-Medoids Algorithms Comparision on TB Data

Ashwini L P
B.E., (M.Tech), Department of CS&E,
Adhichunchanagiri Institute of Technology,
Chikkamagaluru

Sunitha M R
B.E., M.Tech., (Ph.D)., Department of CS&E,
Adhichunchanagiri Institute of Technology,
Chikkamagaluru

***Abstract*:- In identifying and forming the group's dynamic data clustering is a challenge. Undirected knowledge discovery is the outcome of unsupervised learning usually. By victimization the similarity measures the cluster detection algorithmic rule finds the clusters of information that are nearly same to 1 another. Deciding the appropriate algorithmic rule to bring sensible clump result's a problem. Identifying the dataset provided parameters information and attributes information of the results changes in both K- medoids method and K- means method. The proposed method presents a comparison of each algorithmic rules specializing in the result of dataset on every algorithm**

Keywords: K-medoids algorithm, Clustering method, K-means algorithm.

I. INTRODUCTION

The main objective of cluster objects is to search out objects that demonstrate homogeneity in the attributes and group them into one cluster, dissimilar attributes are grouped into another cluster based on different potentials. Clustering method area unit sometimes unattended learning techniques since they look after info while not noted or pre-classified labels, that's that the foremost distinction between cluster and classification. cluster algorithms unit sometimes classified based on Partitioning, that partitions the record set into initial partitions therefore utilize repetitive relocation technique to reinforce the value of clusters and terminates once it converges; Hierarchical-based, that decomposes the dataset into a hierarchy and will be either agglomerate or factious, the previous uses a bottom-up approach that starts with one object and builds clusters by adding similar objects whereas the latter uses a top-down approach that starts with one cluster and divides the dataset into smaller clusters; a Grid which is based on density builds clusters so that a minimum of particular threshold like neighbors for each different members inside each clusters and conjointly based on Model, which builds a cluster models first and finally teams the points of data which is according to the model. The k-means method and k-medoids method comprise partition based mostly techniques.

Choosing the suitable approach to research knowledge may even be tough while not correct understanding regarding the knowledge, the particular domain and additionally the expected result. during this paper, the comparative analysis among two connected approaches is reviewed. clump technique is one amongst powerful tools in data processing wont to discover information. Grouping knowledge according specific parameters eventually varied supported their knowledge and downside domain. There's no specific account all knowledge varieties. This analysis targeted on two techniques that area unit K-Means algorithmic rule and K-Medoids algorithmic rule. Cluster analysis, to boot referred to as unattended knowledge classification, could be an important subject in data processing [1]. Its aim is to partition a group of patterns into clusters of comparable knowledge points. K-Medoids being planned to urge every centers to be one amongst the aim itself. Even K-Medoids algorithmic rule is similar to K-means; variety of the info could also be appropriate to cluster by exploitation K-Medoids algorithmic rule. Many domains information being utilized to research these two techniques to avoid data specific bias.

II. LITERATURE REVIEW

DATA PREPROCESSING

Once the data is taken or loaded the data must be preprocessed. The available data is unstructured and need to be processed for further use. The data contains many unfilled or missing values, special characters and noise in it. The special characters and noise is removed manually by looking the data. In this paper the data is cleaned by filling the missing values or unfilled attributes values. The missing attributes values are filled by taking the mean of all the attributes values in that corresponding column or row. If more number of attributes values is missing then entire row or column is deleted, so that it reduces the data. There are many preprocessing techniques only two are used in this paper. Clean and reduce data preprocessing techniques are applied in our work.

K MEANS CLUSTERING

K-means method (Mac Queen, 1967) is one in all the most unsupervised learning calculations that make out the quality grouping draw back. The system takes once a transparent and simple approach to characterize a given informational index through associate unequivocal assortment of teams (accept k bunches) mounted from the sooner. K-centroids are characterized for every bunch. These centroids have to be compelled to be set in associate passing crafty strategy as a results of entirely extraordinary space causes entirely distinctive outcome. During this method, the upper determination is to position them the foremost extreme total as potential inaccessible from one another. The subsequent stride is to need every reason satisfaction to a given informational index and partner it to the closest center of mass. At the purpose once no style is incomplete, the initial step is finished associated an early gathering age is finished. Currently we'd like to re-figure k new centroids as centers of the teams ensuing from the past stride. Once the k new centroids are collected, a substitution excluding have to be compelled to be completed between indistinguishable informational index focuses and moreover the nearest new center of mass. A circle has been made. As a consequences of this circle we have a tendency to might see that the k centroids revision their space regular until no additional changes territory unit done. At last, this equation goes for limiting Partner in Nursing target work, amid this case a square blunder work. The target operation is

$$J = \sum_{j=1}^{K} \sum_{j=1}^{n} ||x_i \text{ (j)} - Cx_j||^2$$

Where || xi (j) – Cj || two could be a chosen distance live between a knowledge purpose xi (j) and therefore the cluster center cj, is associate degree indicator of the gap of the n knowledge points from their individual cluster centers.

The formula consists of the subsequent steps:

1. Place K points into the area delineate by the objects that are being clustered. These points represent initial cluster centroids.

2. Assign each object to the cluster that has the closest center of mass.

3. Once all objects are assigned, cipher the positions of the K centroids. Repeat steps two and three till the centroids now not move. This produces a separation of the objects into teams from with the metric to be reduced are often calculated.

K-MEDOIDS CLUSTERING

K-medoids could be a cluster rule associated with the k-means rule and it is called medoid shift rule, each k-means clustering and k-medoids clustering algorithms are partitioned (breaking the dataset up into groups) and every decide to minimize sq. error, the gap between points tagged to be in associate passing cluster and some extent elect as a result of the middle of that cluster. In distinction to the k-means algorithm, k-medoids chooses information points as centers (medoids or exemplars). K-medoids is in addition a partitioning technique of cluster that clusters the information set of n objects into k clusters with k far-famed a priori. It may be plenty of strong to noise and outliers as compared to k-means clustering method as a result of it reduce a complete of general attempt wise different similarities as an alternative of a complete square geometrician distance. Several of the distinction operate is implausibly affluent, but in an application. This tends to use the square Euclidian value. The center of medoid as a finite information that could be for a data purpose from any data set, overall average distinction to all information points is lowest such that it is the main centre set purpose among different data set.

The most general understanding of k-medoid bunch is that the "Partitioning around Medoids" (PAM) algorithmic rule and it is listed below:

1. Initialization step: Initially select a k information point out of n information points as a center point.

2. Allocation step: Group each data to the nearest center value (medoid).

3. Update step: For each center value (medoid) m and each information purpose of related to m swap m and o and calculate the full worth of the configuration. Medoid o with the lowest worth of the configuration is selected.

4. Repeat step 2 and step 3 until there is no alteration among the allocation.

III. METHODOLOGY

The outline overview of the system is called the architecture of the system. It is the model which illustrates the characteristics of the modules, sub modules associated with the main system. It explains the flow of the overall system working, the levels in which the data flow, the modules and the processes which should be executed in what order. It should be able to categorize both visible and non-visible activities of the system. Figure 1, shows the proposed system architecture. The system architecture diagram of the proposed method mainly contains 3 modules. The three different modules are namely Data Collection Approach, Data preprocessing, cluster formation using data mining techniques and result Analysis. Using the health care data is more advantageous it helps the health care professionals and researchers to gain knowledge about the diseases and also the accuracy of the diseases increasing year by year.

Data collection: Data collection is the approach in which the health care data is collected from the specified hospitals or by some websites. The information of websites data is unstructured and fuzzy and hence it is subjected to the pre-processing techniques.

Pre-processing techniques: It includes clean and reducing the data. The data is cleaned by removing any noise present in the data. The unfilled value are filled by taking the mean of all the values and if more values are missing then the entire tuple is deleted, this is how the data is reduced. The xsl sheet data is converted into arff style data. These are the pre-processing steps we undertake for the data set sample.

Clustering: Clustering can be viewed as the most imperative unsupervised learning issue; in this way, as each other issue of this type, it manages to find a structure in an collection of unlabeled information. The other meaning of bunching could be "the procedure of sorting out items into gatherings whose individuals are comparative somehow". A group is in this manner a gathering of articles which are "comparative" amongst them and are "unique" to the items having a place with different bunches. In clustering first we cluster the data by using k-means clustering algorithm, then again we cluster data using k-medoids (PAM) algorithm.

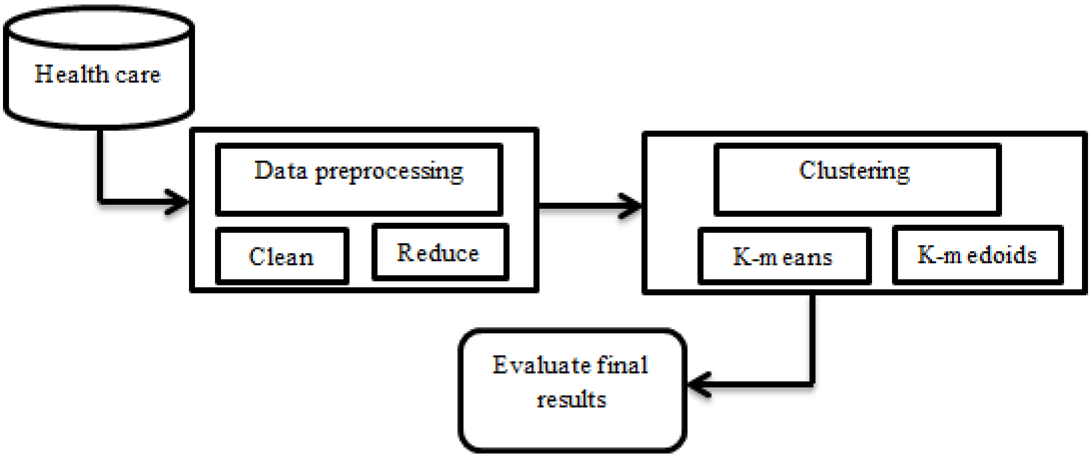


Figure 1: Architecture diagram for clustering TB data sample

IV. PERFORMANCE COMPARISION OF STRATEGIES FOR CHOOSING INITIAL MEDOID

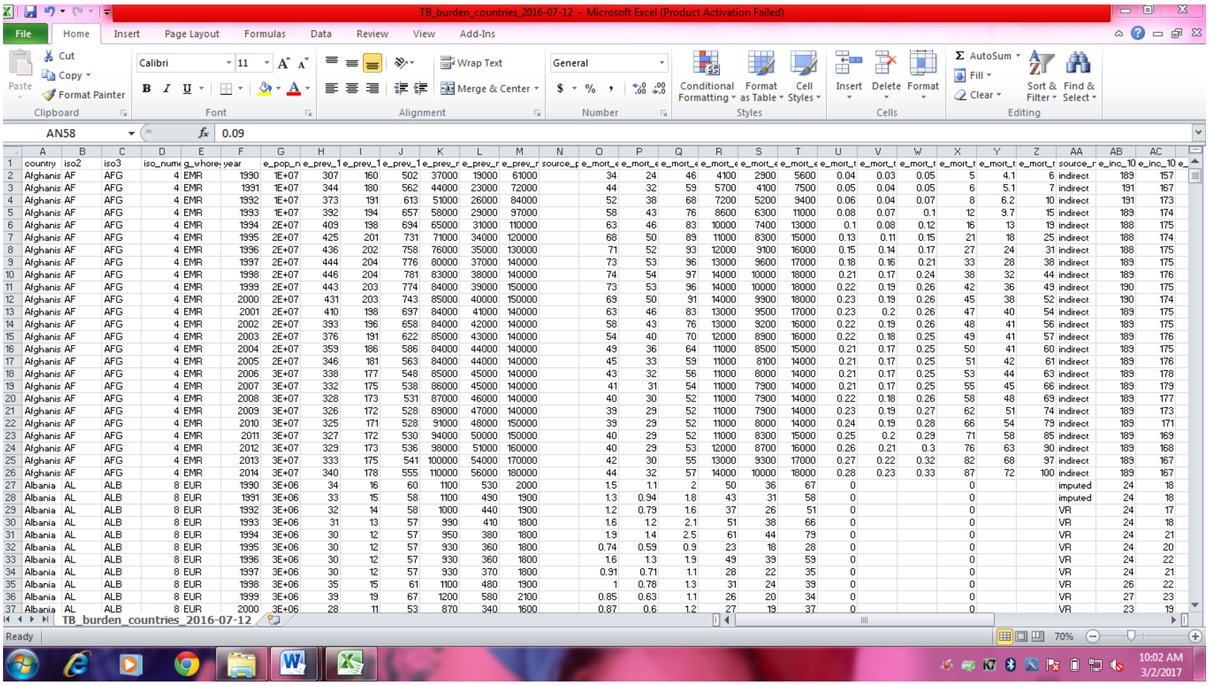
Performance of unvarying cluster algorithmic rule depends up on selection of 'cluster centers' in every step. An experiments square measure conducted on T.B. information set sample. T.B. (TB) dataset is shown in below Figure a pair of. A world wide causes of death due to Tuberculosis (TB) is found one in every of the ten people. T.B. (TB) is caused by microorganisms (Mycobacterium tuberculosis) that almost all typically have an effect on the lungs. T.B. can be cured and prevented. TB is unfolded from one person to another person through air. once individuals with respiratory organ sneeze or spit, TB cough, they spread the germs of TB into an air. If someone has inhale solely many of these germs are to be infected.

Regarding tierce of a world population had latent TB problem, which suggests individuals are infected by TB microorganism however aren't (yet) sick with the sickness and can't transmit the sickness. Individuals infected with TB microorganism have a tenth period of time danger of having sick with TB. On the other hand, persons with high immune power like individuals having an HIV, deficiency disease or polygenic disorder, or peoples using tobacco, have a higher danger for falling sick. Once someone develops an active TB sickness, the symptoms of these disease are (weight loss, cough, night sweats, or fever) could also be delicate for several days. This will cause due of days in seeking a care, and leads to transform of an microorganism to another. Individuals with an active TB problem will be infecting to 10–15 peoples through shut in contacts over a courses of a year. While not correct handling of disease, leads to forty five percent of HIV-negative individuals with TB on this the average and a nearly all the people having HIV-positive with a TB problem will die.

In 2015, 10.4 million people fell ill with TB and 1.8 million died from the disease (including 0.4 million among people with HIV). Over 95% of TB deaths occur in low- and middle-income countries. Six countries account for 60% of the total, with India leading the count, followed by Indonesia, China, Nigeria, Pakistan and South Africa.

Attributes

Country –country name
ISO2- International standard organization (ISO) code alpha 2(country code)
ISO3-country code alpha 3
ISO_numeric-Numeric country code
Year
e_most_exc_tbhiv_100k-Mathematical model for estimating TB incidence
e_inc_tbhiv_100k_hi-HIV positive per 100000 population high bound
e_mort_exc_tbhiv_num- Number of deaths from TB
e_pop_num - Estimated total population number
new_ep_f014- female extra pulmonary aged 0-14 years



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC
1	country	iso2	iso3	iso_nume	g_whore	year	e_pop_n	e_prev_1	e_prev_1	e_prev_1	e_prev_r	e_prev_r	e_prev_r	source_p	e_mort_e	e_mort_e	e_mort_e	e_mort_e	e_mort_e	e_mort_e	e_mort_t	e_mort_t	e_mort_t	e_mort_t	e_mort_t	e_mort_t	source_r	e_inc_10	e_inc_10
2	Afghanis	AF	AFG	4	EMR	1990	1E+07	307	160	502	37000	19000	61000		34	24	46	4100	2900	5600	0.04	0.03	0.05	5	4.1	6	indirect	189	157
3	Afghanis	AF	AFG	4	EMR	1991	1E+07	344	180	562	44000	23000	72000		44	32	59	5700	4100	7500	0.05	0.04	0.05	6	5.1	7	indirect	191	167
4	Afghanis	AF	AFG	4	EMR	1992	1E+07	373	191	613	51000	26000	84000		52	38	68	7200	5200	9400	0.06	0.04	0.07	8	6.2	10	indirect	191	173
5	Afghanis	AF	AFG	4	EMR	1993	1E+07	392	194	657	58000	29000	97000		58	43	76	8600	6300	11000	0.08	0.07	0.1	12	9.7	15	indirect	189	174
6	Afghanis	AF	AFG	4	EMR	1994	2E+07	409	198	694	65000	31000	110000		63	46	83	10000	7400	13000	0.1	0.08	0.12	16	13	19	indirect	188	175
7	Afghanis	AF	AFG	4	EMR	1995	2E+07	425	201	731	71000	34000	120000		68	50	89	11000	8300	15000	0.13	0.11	0.15	21	18	25	indirect	188	174
8	Afghanis	AF	AFG	4	EMR	1996	2E+07	436	202	758	76000	35000	130000		71	52	93	12000	9100	16000	0.15	0.14	0.17	27	24	31	indirect	188	175
9	Afghanis	AF	AFG	4	EMR	1997	2E+07	444	204	776	80000	37000	140000		73	53	96	13000	9600	17000	0.18	0.16	0.21	33	28	38	indirect	189	175
10	Afghanis	AF	AFG	4	EMR	1998	2E+07	446	204	781	83000	38000	140000		74	54	97	14000	10000	18000	0.21	0.17	0.24	38	32	44	indirect	189	176
11	Afghanis	AF	AFG	4	EMR	1999	2E+07	443	203	774	84000	39000	150000		73	53	96	14000	10000	18000	0.22	0.19	0.26	42	36	49	indirect	190	175
12	Afghanis	AF	AFG	4	EMR	2000	2E+07	431	203	743	85000	40000	150000		69	50	91	14000	9900	18000	0.23	0.19	0.26	45	38	52	indirect	190	174
13	Afghanis	AF	AFG	4	EMR	2001	2E+07	410	198	697	84000	41000	140000		63	46	83	13000	9500	17000	0.23	0.2	0.26	47	40	54	indirect	189	175
14	Afghanis	AF	AFG	4	EMR	2002	2E+07	393	196	658	84000	42000	140000		58	43	76	13000	9200	16000	0.22	0.19	0.26	48	41	56	indirect	189	175
15	Afghanis	AF	AFG	4	EMR	2003	2E+07	376	191	622	85000	43000	140000		54	40	70	12000	8900	16000	0.22	0.18	0.25	49	41	57	indirect	189	176
16	Afghanis	AF	AFG	4	EMR	2004	2E+07	359	186	586	84000	44000	140000		49	36	64	11000	8500	15000	0.21	0.17	0.25	50	41	60	indirect	189	175
17	Afghanis	AF	AFG	4	EMR	2005	2E+07	346	181	563	84000	44000	140000		45	33	59	11000	8100	14000	0.21	0.17	0.25	51	42	61	indirect	189	176
18	Afghanis	AF	AFG	4	EMR	2006	3E+07	338	177	548	85000	45000	140000		43	32	56	11000	8000	14000	0.21	0.17	0.25	53	44	63	indirect	189	178
19	Afghanis	AF	AFG	4	EMR	2007	3E+07	332	175	538	86000	45000	140000		41	31	54	11000	7900	14000	0.21	0.17	0.25	55	45	66	indirect	189	179
20	Afghanis	AF	AFG	4	EMR	2008	3E+07	328	173	531	87000	46000	140000		40	30	52	11000	7900	14000	0.22	0.18	0.26	58	48	69	indirect	189	177
21	Afghanis	AF	AFG	4	EMR	2009	3E+07	326	172	528	89000	47000	140000		39	29	52	11000	7900	14000	0.23	0.19	0.27	62	51	74	indirect	189	173
22	Afghanis	AF	AFG	4	EMR	2010	3E+07	325	171	528	91000	48000	150000		39	29	52	11000	8000	14000	0.24	0.19	0.28	66	54	79	indirect	189	171
23	Afghanis	AF	AFG	4	EMR	2011	3E+07	327	172	530	94000	50000	150000		40	29	52	11000	8300	15000	0.25	0.2	0.29	71	58	85	indirect	189	169
24	Afghanis	AF	AFG	4	EMR	2012	3E+07	329	173	536	98000	51000	160000		40	29	53	12000	8700	16000	0.26	0.21	0.3	76	63	90	indirect	189	168
25	Afghanis	AF	AFG	4	EMR	2013	3E+07	333	175	541	100000	54000	170000		42	30	55	13000	9300	17000	0.27	0.22	0.32	82	68	97	indirect	189	167
26	Afghanis	AF	AFG	4	EMR	2014	3E+07	340	178	555	110000	56000	180000		44	32	57	14000	10000	18000	0.28	0.23	0.33	87	72	100	indirect	189	167
27	Albania	AL	ALB	8	EUR	1990	3E+06	34	16	60	1100	530	2000		1.5	1.1	2	50	36	67	0			0			imputed	24	18
28	Albania	AL	ALB	8	EUR	1991	3E+06	33	15	58	1100	490	1900		1.3	0.94	1.8	43	31	58	0			0			imputed	24	18
29	Albania	AL	ALB	8	EUR	1992	3E+06	32	14	58	1000	440	1900		1.2	0.79	1.6	37	26	51	0			0			VR	24	17
30	Albania	AL	ALB	8	EUR	1993	3E+06	31	13	57	990	410	1800		1.6	1.2	2.1	51	38	66	0			0			VR	24	18
31	Albania	AL	ALB	8	EUR	1994	3E+06	30	12	57	950	380	1800		1.9	1.4	2.5	61	44	79	0			0			VR	24	21
32	Albania	AL	ALB	8	EUR	1995	3E+06	30	12	57	930	360	1800		0.74	0.59	0.9	23	18	28	0			0			VR	24	20
33	Albania	AL	ALB	8	EUR	1996	3E+06	30	12	57	930	360	1800		1.6	1.3	1.9	49	39	59	0			0			VR	24	22
34	Albania	AL	ALB	8	EUR	1997	3E+06	30	12	57	930	370	1800		0.91	0.71	1.1	28	22	35	0			0			VR	24	21
35	Albania	AL	ALB	8	EUR	1998	3E+06	35	15	61	1100	480	1900		1	0.78	1.3	31	24	39	0			0			VR	26	22
36	Albania	AL	ALB	8	EUR	1999	3E+06	39	19	67	1200	580	2100		0.85	0.63	1.1	26	20	34	0			0			VR	27	23
37	Albania	AL	ALB	8	EUR	2000	3E+06	28	11	53	870	340	1600		0.87	0.6	1.2	27	19	37	0			0			VR	23	19

Figure 2: Tuberculosis dataset sample

V. COMPARISON BETWEEN K-MEAN CLUSTERS AND K-MEDIOD ALGORITHMIC RULE WITH A VARIETY OF CLUSTER NUMBER AND ITS EXECUTION TIME

Table 1: For different number of clusters and time taken for execution using K-Mean method and K-Medoid Algorithms

Number of clusters	Execution Time (in milliseconds) K-means Algorithm	Execution Time (in milliseconds) K-medoids Algorithm
2	23123	20456
3	42245	41275
4	68632	69452
5	76893	75649

The Figure 3 shows a comparison among K-means and a K-medoids algorithm. Since graph shows the variety of a cluster is a smaller amount of time takes to execute than a K-mean method. If a quantity of a clusters is over its once it is extra accurate that K-medoid takes a lesser time to compute 15345 records than a K-mean cluster method. At a foremost variety of a cluster enhances than time taken for execution by a K-medoid method is a smaller amount than the K- mean.

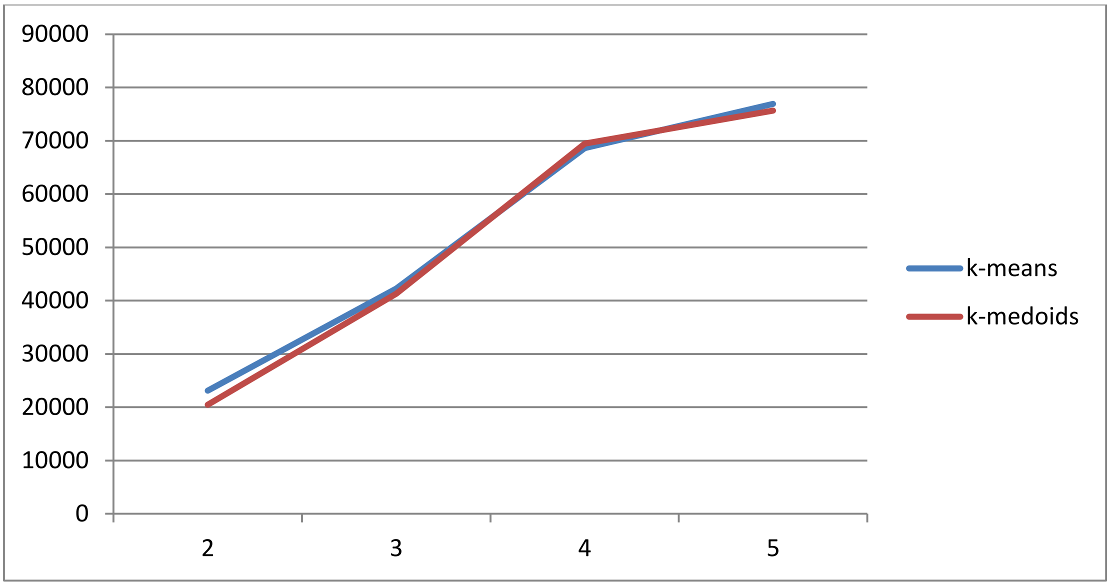


Figure 3: Comparison of a K-means and K-medoid algorithm

VI. CONCLUSION

In this paper comparison of both K-means and K-medoids algorithmic rules are discussed. K-medoid technique has higher performance than K-means agglomeration and this takes a lesser execution time than K-mean method. Conjointly numerous strategies for choosing original clusters are conferred and compared. The K-Medoid agglomeration may be a higher for agglomeration performance and its execution time will be a smaller amount. As a result, even if the strategy of choosing first clusters delineate within the planned method sufficient for the use of once allowing for each performance and also for a computation time.

REFERENCES

[1] S. A. Raut, et.al, “Bioinformatics: Trends of Gene Expression and Analysis,” proceedings of International Conference On Bioinformatics and Biomedical Technology, 16-18 April 2010,Chengdu, China.

[2] S. A. Raut, et.al. “Gene Expression Analysis- A Review of large datasets,” vol.4, Issue 1, Journal of Computer Science and Engineering, November 2010.

[3] Xiong, H., et.al. “K-Means clustering versus validation measures: A Data distribution perspective”, IEEE Trans. Syst., Man,C ybernet. PartB,39:318331.http://www.ncbi.nlm.nih.gov/pubmed/1909553, 2009.

[4] S. Ray, et.al, “Purpose of number of clusters in k-means clustering and application in a color image segmentation,” In Proceedings of the 4th International Meeting on Advances in Pattern Recognition and Digital Techniques, pp.137-143. 2014.

[5] G. Sheikholeslami et.al, “Wave-Cluster: A Multi-Resolution Clustering Approach for Very Large Spatial Databases,” Proc. 24th Int. Conf. on Very Large Data Bases. New York, pp. 428-439,2008.

[6] R. Sibson, “SLINK: An optimally efficient algorithm for the single link cluster method,” The Comp. Journal, 16(1), 2015, pp. 30-34.

[7] T. Zhang, et.al, “BIRCH: An Efficient Data of Clustering Method for Very Large Databases,” Proc. ACM SIGMODInt. Conf. on

[8] Zhang Y, et.al “An efficient Clustering algorithm”, In Proceedings of Second International Conference on Machine Learning and Cyber netics, November 2013.