International Journal of Engineering, Science and Mathematics
Vol. 8 Issue 4, April 2019,
ISSN: 2320-0294 Impact Factor: 6.765
Journal Homepage: http://www.ijesm.co.in, Email: ijesmj@gmail.com
Double-Blind Peer Reviewed Refereed Open Access International Journal - Included in the International Serial Directories Indexed & Listed at: Ulrich's Periodicals Directory ©, U.S.A., Open J-Gage as well as in Cabell's Directories of Publishing Opportunities, U.S.A

NEW APPROACH FOR FINDING NUMBER OF CLUSTERS USING DISTANCE BASED K-MEANS ALGORITHM

Mohamed Cassim Alibuhtto*

Nor Idayu Mahat**

ABSTRACT

Clustering is one of the most universal unsupervised classification methods for partitioning objects into a set of meaningful clusters. The *k*-means clustering algorithm is a commonly used partitioning based clustering method for finding optimal number of clusters. However, number of clusters generated by *k*-means algorithm depends on the choice of centroid value which sometimes could be misled. Therefore, a new approach for identifying the optimal number of clusters based on distance in *k*-means algorithm is proposed. The designed algorithm was tested using twelve sets of simulated data has revealed that the proposed algorithm is able to identify the exact number of clusters.

KEYWORDS:

Clustering;
k-meansalgorithm;
Simulation;
Validation.

Author correspondence:

Mohamed Cassim Alibuhtto,
Department of Mathematical Sciences,
Faculty of Applied Sciences,
South Eastern University of Sri Lanka, Sri Lanka.

Second author affliation:

Department of Mathematics and Statistics,
School of Quantitative Sciences,
Universiti Utara Malaysia, Malaysia.

* Doctorate Program, Linguistics Program Studies, Udayana University Denpasar, Bali-Indonesia (9 pt)
** STIMIK STIKOM-Bali, Renon, Depasar, Bali-Indonesia

1. INTRODUCTION

Problems in clustering objects is associated with finding an explainable structure in a collection of unlabeled data set. Clustering works by splitting *n* objects into *k* clusters such that the objects within the same cluster are similar and show some differences with objects in other clusters. Clustering offers better insight about a huge size of data by giving a compact representation in a form of clusters of objects. Nowadays, the exercise in recognizing clusters of objects has become popular and is widely used in many fields such as in geology, marketing, medical, meteorology, and finance (Gan, Ma & Wu, 2007; Liu et al., 2010; Kodinariya & Makwana, 2013).

A process of clustering objects systematically is termed as cluster analysis. The process could be divided into hierarchical clustering and non-hierarchical clustering techniques. Among the well-known hierarchical clustering are agglomerative and divisive, while famous methods on non-hierarchical method include *k*-means and *k*-medoids. Hierarchical clustering is a tree like structure where it continuously combines similar objects until they are all in the same cluster. Meanwhile, non-hierarchical clustering divides objects into a pre-determined cluster. In dealing with large amount of data, often non-hierarchical clustering is preferable where *k*-means clustering algorithm always turn as the most preferred one. The technique determines a specified number of non-overlapping clusters within data and is widely used in several fields due to its simplicity and efficiency (Jain, 2010; Mihai & Mocanu, 2015).

Many studies were conducted to find number of clusters using *k*-means algorithm (Milligan & Cooper, 1985; Kane & Nagar , 2012; Mehar, Matawie & Maeder, 2013; Muca & Kutrolli, 2015) where the centroids were sometimes based on initial guess. Such choice may end up with non-optimal clusters, hence identifying best possible initial centroids would be a good one. In the past, few studies were conducted to identify number of clusters using distance-based *k*-means algorithm (Napoleon & Lakshmi, 2010; Singh, Rana & Yadav, 2013).

The k-means algorithm was developed by MacQueen in 1967. The steps of *k*-means algorithm are as follows:

1. Randomly select *k* objects from a sample, each of which initially represents a cluster centroid.

2. For each of the remaining objects, assign an object to a cluster to which it shows the most similar, based on the distance between the object and the cluster mean.
3. Compute mean for each cluster.
4. Repeat Step 2 and Step 3 until converges.

The k-means algorithm works in minimizing the sum of the squared error function, which is very simple and can be easily implemented.

$$J = \frac{1}{n}\sum_{i=1}^{k}\sum_{x\in C_i}\|x - m_i\|^2 \quad (1)$$

where J is the sum of the square error for all objects in the data set, *x* is the point in space representing a given object; and mi is the mean of cluster C_i and *k* is the given number of clusters.

However, these common steps face some drawbacks where uncertain number of iterations could be processed to determine the optimal number of clusters especially when inappropriate centroid value (*k*) was used. This paper aims at introducing a new algorithm for determining the number of optimal clusters using Euclidean distance in *k*-means clustering algorithm. The proposed algorithm was tested by using simulated multivariate data.

2. RESEARCH METHODOLOGY

2.1 Euclidean based distance of k-means algorithm

Determining the optimal number of clusters in a dataset is the main issue in the *k*-means cluster algorithm, which requires the user to specify number of clusters to be generated. Thus, this study proposeda new distance based *k*-means algorithm to find the best number of optimal clusters of available data. For this technique, the Euclidean distance was chosen as a measurement of the distance between objects due to its simplicity and easy computation for numerical multivariate data. The proposed study is to find the best *k*, such that adding more clusters will not cause major changes in term of amount of separation between the clusters. The proposed method is describedin Algorithm 2.1.

Algorithm 2.1: Proposed k-means Algorithm using EuclideanDistances

Step 1: Set the counter *k*=2, where *k* represents the number of cluster.

1.1 Identify the centre of each cluster *k*, C_{jk} , where $j=1,2,\ldots,k$.

1.2 Perform *k*-means clustering analysis on the data set using the identified centres from step 1.1.

1.3 Compute distance between centres, label as d_k

Step 2: Add 1 to the previous value of *k*, so $k=k+1$

1.1 Repeat steps 1.1-1.3.

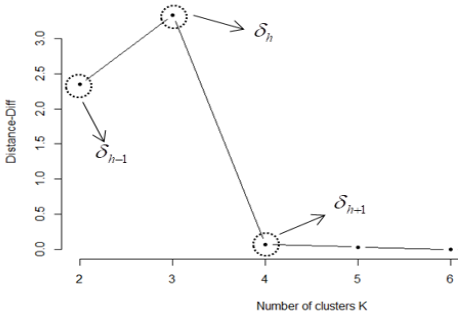
1.2 Compute the different of two distances at *k* and *k-1*.

$\delta_i = |d_k - d_{k-1}|$ where i=1,2,…

2.3 Compare $\delta_i < m$, if true store k-1 as optimal cluster, otherwise repeat step 2.

Detemining the constant value m

The constant value of *m* is derived from the scatter plot of different distance (δ) against *k*-value. The peak value $(say\ \delta_h)$ for the graph is determined from the point where the trend moves in a downward trend without more fluctuation. Then, m can be determined by taking average of the first three points such as $\delta_{h-1}, \delta_h, and\ \delta_{h+1}$. These points are marked in Figure 2.1.



$$m = \frac{(\delta_{h-1} + \delta_h + \delta_{h+1})}{3} \quad ; h = 2,3,...,6 \qquad (2)$$

*Figure 2.1.*Scatter plot

2.2 Cluster Validation Measures

A clustering process was measured to ensure that the obtained clusters are correct to explain patterns of clusters in the data. In this study, Dunn and Calinksi Harbaz indices were used to evaluate the clustering results. Both indices are briefly described below.

2.2.1. Dunn Index

This Dunn index define as the ratio between the minimal intra cluster distance to maximal inter cluster distance. The Dunn index is as follows:

$$D_k = \min_{1 \le i \le k} \left\{ \min_{\substack{1 \le j \le k \\ i \ne j}} \left\{ \frac{d(C_i, C_j)}{\max_{1 \le l \le k} \{diam(C_l)\}} \right\} \right\} \quad (2)$$

Where $d(C_i, C_j)$ is the distance between two clusters C_i and C_j as the minimum distance between a pair of objects in the two different clusters separately and the diameter of cluster Cl, diam(Cl), as the maximum distance between two objects in the cluster. The maximum value of Dunn index indicates that *k* is the optimal number of clusters.

2.2.2 Calinski-Harabasz Index

This index evaluates the validity of the clusters based on two measures that separation between cluster sum of squares (SSB) and cohesion within cluster sum of squares (SSW). The CH index is defined as:

$$CH = \frac{SSB/(k-1)}{SSW/(n-k)} \quad (3)$$

Where n is the number of observations and *k* is the number of clusters. The maximum value of Calinski-Harabasz index indicates that *k* is the optimal number of clusters.

2.3 Data Simulation

In this study, the proposed k-means algorithm was tested by using 2-dimensional and 3-dimensional data sets, defined by 50,100, and 500 objects, generated using a multivariate normal distribution from 2 and 3 groups (clusters) with different means and covariance matrices are tabulated in Table 1.

Table 1: *Features of simulated datasets with 2 and 3-dimensional data*

Data	Number of clusters	Number of variables	Number of Objects	Mean	Covariance Matrix
Data2_1	2	2	50	$\mu1 = (0,0)$ $\mu2 = (3,0)$	$\begin{bmatrix} 0.2500 & 0.0059 \\ 0.0059 & 0.1225 \end{bmatrix}$
Data2_2			100	$\mu1 = (-2,0)$ $\mu2 = (1,0)$	$\begin{bmatrix} 0.2500 & 0.0077 \\ 0.0077 & 0.2025 \end{bmatrix}$
Data2_3			500	$\mu1 = (0,0.5)$ $\mu2 = (2.5,0)$	$\begin{bmatrix} 0.2500 & 0.0005 \\ 0.0005 & 0.1600 \end{bmatrix}$
Data2_4		3	50	$\mu1 = (-2,0,0)$ $\mu2 = (5,0,0)$	$\begin{bmatrix} 0.4900 & 0.0588 & 0.0070 \\ 0.0588 & 0.3600 & 0.0000 \\ 0.0070 & 0.0000 & 0.2500 \end{bmatrix}$
Data2_5			100	$\mu1 = (0,0,0)$ $\mu2 = (2,1,0)$	$\begin{bmatrix} 0.0900 & 0.0168 & 0.0021 \\ 0.0168 & 0.1600 & 0.0000 \\ 0.0021 & 0.0000 & 0.1225 \end{bmatrix}$
Data2_6			500	$\mu1 = (-3,1,0)$ $\mu2 = (4,0,0)$	$\begin{bmatrix} 0.1600 & 0.0168 & 0.0036 \\ 0.0168 & 0.0900 & 0.0095 \\ 0.0360 & 0.0095 & 0.2025 \end{bmatrix}$
Data3_1		2	50	$\mu1 = (1,0)$ $\mu2 = (3,0)$ $\mu3 = (5,0)$	$\begin{bmatrix} 0.0900 & 0.0029 \\ 0.0029 & 0.1600 \end{bmatrix}$
Data3_2			100	$\mu1 = (-1,0)$ $\mu2 = (2,1)$ $\mu3 = (5,0)$	$\begin{bmatrix} 0.1600 & 0.0027 \\ 0.0027 & 0.0400 \end{bmatrix}$

Data3_3	3		500	$\mu 1 = (-3,0)$ $\mu 2 = (1,1)$ $\mu 3 = (5,0)$	$\begin{bmatrix} 0.4900 & 0.0168 \\ 0.0168 & 0.3600 \end{bmatrix}$
Data3_4		3	50	$\mu 1 = (-3,0,0)$ $\mu 2 = (5,1,0)$ $\mu 3 = (15,1,0)$	$\begin{bmatrix} 0.2500 & 0.0028 & 0.0450 \\ 0.0028 & 0.1600 & 0.0306 \\ 0.0450 & 0.0306 & 0.2025 \end{bmatrix}$
Data3_5			100	$\mu 1 = (0,1,0)$ $\mu 2 = (5,0,0)$ $\mu 3 = (9,1,0)$	$\begin{bmatrix} 0.1600 & 0.0022 & 0.0120 \\ 0.0022 & 0.1600 & 0.0480 \\ 0.0120 & 0.0480 & 0.0900 \end{bmatrix}$
Data3_6			500	$\mu 1 = (-3,0,0)$ $\mu 2 = (2,1,0)$ $\mu 3 = (7,0,0)$	$\begin{bmatrix} 0.2500 & 0.0675 & 0.0300 \\ 0.0675 & 0.2025 & 0.0540 \\ 0.0300 & 0.0540 & 0.0900 \end{bmatrix}$

3. RESULTS AND DISCUSSIONS

3.1 Two and three-dimensional scatter plots

Figures 2 and 3 show the two and three-dimensional scatter plots for simulated data (in Table 1) with two different number of clusters (k=2 &k=3).

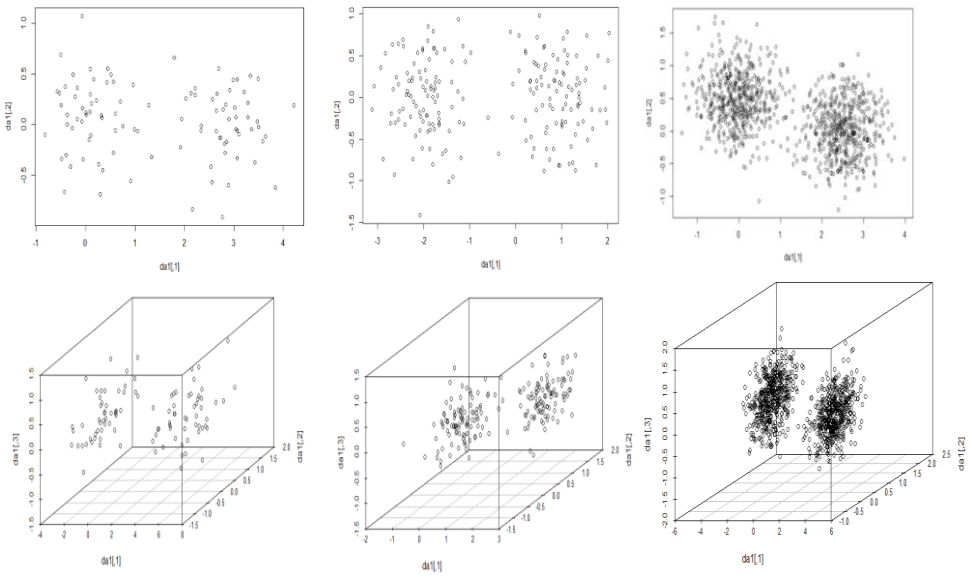


Figure 1. Scatter plot for two group

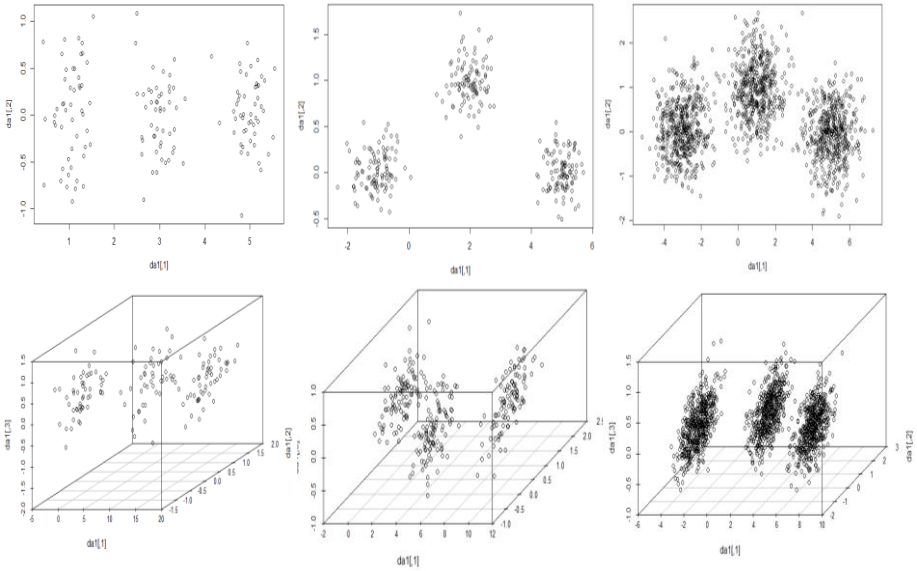


Figure 2. Scatter plot for three groups

3.2 Results of proposed clustering algorithm

The proposed algorithm was tested using two clusters of simulated data sets (such as Data2_1,…,Data2_6) to find the optimal number of clusters. The validity indices (Dunn index and Calinksi-Harabaz) and difference between consecutive clusters (Algorithm 2.1) is computed for each data set. In Table 2, the maximum values of Dunn and Calinksi-Harabaz indices are obtained with *k=2*. These indices confirm that the number of clusters of the datasets is 2. Furthermore, all six datasetsmeet the corresponding condition of $\delta < m$ (in Algorithm 2.1) when *k=2*. Hence, the proposed distance-based *k*-means clustering

algorithm is more appropriate fordetermining the number of clusters.

Table 2: Results for 2 clusters simulated data

Data Set	k	Clusters of sizes	Euc.dis	Dunn	CH	Diff (δ)	*m*
Data2-1	2	51,49	2.724	0.238*	471.994*	-	**0.665**
	3	45,16,39	1.333	0.130	326.302	**1.391**	
	4	20,26,31,23	0.819	0.085	306.397	0.514	
	5	12,24,20,31,13	0.729	0.081	272.245	0.090	
	6	17,24,10,13,12,24	0.666	0.101	257.229	0.063	
Data2-2	2	100,100	2.896	0.174*	907.551*	-	**0.722**
	3	99,54,47	0.880	0.039	596.245	**2.016**	
	4	51,53,47,49	0.776	0.024	508.744	0.104	
	5	47,54,31,36,32	0.821	0.050	471.120	0.045	
	6	32,32,31,36,40,29	0.797	0.044	478.795	0.024	
Data2-3	2	503,497	2.579	0.042*	4161.476*	-	**0615**
	3	494,236,270	0.816	0.009	2720.890	**1.763**	
	4	233,268,266,233	0.760	0.009	2419.885	0.056	
	5	265,159,235,165,176	0.734	0.008	2164.164	0.026	
	6	167,170,159,171,163,170	0.741	0.008	2111.422	0.007	
Data2-4	2	50,50	6.939	1.342*	1146.309*	-	**1.969**
	3	50,13,37	1.471	0.152	723.371	**5.468**	
	4	23,27,37,13	1.246	0.044	633.597	0.225	
	5	18,13,27,23,19	1.033	0.054	562.176	0.213	
	6	18,14,13,21,15,19	1.028	0.076	515.996	0.005	
Data2-5	2	100,100	2.246	0.338*	698.738*	-	**0.542**
	3	61,100,39	0.668	0.023	425.465	**1.578**	
	4	54,39,61,46	0.621	0.024	347.182	0.047	
	5	34,61,33,33,39	0.621	0.026	295.484	0.000	
	6	36,22,34,42,33,33	0.621	0.052	270.131	0.000	
Data2-6	2	500,500	7.084	1.431*	27313.650*	-	**2.129**
	3	500,280,220	0.808	0.021	16654.040	**6.276**	
	4	277,220,280,223	0.741	0.019	13591.900	0.067	
	5	149,280,220,173,178	0.697	0.024	11357.730	0.044	
	6	150,146,149,190,161,204	0.679	0.022	10183.410	0.018	

Similarly, the three clusters ofgenerated data sets (such as Data3_1,…,Data3_6) were tested using the proposed distance based *k*-means clustering algorithm. In Table 3, the maximum values of Dunn and Calinksi-Harabaz indices are obtained with k=3, and this confirms that number of clusters of datasets is equal to 3. Moreover, all six datasets satisfy the corresponding condition of $\delta < m$(in Algorithm 2.1) when k=3. Therefore, the proposed distance-based k-means clustering algorithm is more appropriate to find number of clusters even without using validation indices.

Table 3: Results for 3-clusters simulated data

Data Set	k	Clusters of sizes	Euc.dist	Dunn	CH	Diff (δ)	*m*
Data3-1	2	52,98	2.926	0.067	331.483	-	**0.758**
	3	50,50,50	1.934	0.350*	760.096*	0.992	
	4	27,23,50,50	0.850	0.045	679.316	**1.084**	
	5	27,23,23,27,50	0.651	0.049	630.694	0.199	
	6	27,27,23,23,23,27	0.569	0.064	614.061	0.082	
Data3-2	2	197,103	4.538	0.048	747.163	-	**1.304**
	3	100,100,100	3.166	0.519*	4856.618*	1.372	
	4	100,100,65,35	0.725	0.017	4094.605	**2.441**	
	5	100,35,51,65,49	0.627	0.021	3916.461	0.098	
	6	51,49,40,58,42,60	0.549	0.020	3922.115	0.078	
Data3-3	2	945,555	5.943	0.013	3569.235	-	**1.601**
	3	501,499,500	4.083	0.029*	9532.989*	1.860	
	4	495,497,244,264	1.178	0.012	7391.813	**2.905**	
	5	255,259,495,241,250	1.141	0.009	6496.045	0.037	
	6	268,259,230,249,253,241	1.119	0.007	6280.588	0.022	
Data3-4	2	50,100	14.081	0.686	557.015	-	**4.409**
	3	50,50,50	8.148	1.556*	6312.207*	5.933	
	4	50,25,50,25	1.037	0.109	4866.301	**7.111**	
	5	25,23,27,50,25	0.853	0.081	4072.923	0.184	
	6	23,29,27,21,31,19	0.825	0.087	3632.701	0.028	
Data3-5	2	100,200	6.936	0.446	982.303	-	**2.081**
	3	100,100,100	4.118	0.731*	4821.784*	2.818	
	4	56,100,44,100	0.743	0.049	3601.604	**3.375**	
	5	58,42,55,100,45	0.692	0.027	3065.732	0.051	
	6	54,55,51,46,49,45	0.674	0.052	2754.829	0.018	

Data3-6	2	505,995	7.487	0.030	3891.489	-	**2.195**
	3	500,500,500	5.067	0.600*	23128.390	2.420	
	4	500,248,500,252	0.947	0.015	17942.680	**4.120**	
	5	252,248,248,252,500	0.903	0.015	15791.700	0.044	
	6	252,248,248,252,248,252	0.865	0.015	15004.120	0.038	

4. CONCLUSION

Clustering is an important field in data mining techniques, and many researchers use the cluster techniques to find the optimal number of clusters. The goal of this paper is to determine the appropriate number of clusters using a new approach ofthe distance-based k-means clustering algorithm.This proposed *k*-means algorithm shows that the number of clusters of simulation data is optimal. Furthermore, the optimal number of clusters can be identified without using the validation indices. In addition, this study will be an important platform for highlighting and discussing big data problems that determine the optimal number of clusters. However, this study can be improved when objects are overlapped and more groups in the datasets.

ACKNOWLEDGEMENTS

This research paper is a part of first author's PhD studies under the supervision of the second author. This study was supported by the University Grant Commission of Sri Lanka (UGC/VC/DRIC/PG2017(I)/SEUSL/03).

REFERENCES

[1] Gan, G., Ma, C., &Wu, J. (2007). Data Clustering: Theory, Algorithms, and Applications. *Philadelphia:* ASA-SIAM.

[2] Jain, A. K. (2010). Data clustering : 50 years beyond *k*-means. *Pattern Recognition Letters*, *31*(8), 651–666.

[3] Kane, A., & Nagar, J. (2012). Determining the number of clusters for a *k*-means clustering algorithm. *Indian Journal of Computer Science and Engineering (IJCSE)*, 3(5), 670–672.

[4] Kodinariya, T. M., & Makwana, P. R. (2013). Review on determining number of Cluster in K-Means Clustering. *International Journal of Advance Research in Computer Science and Management Studies*, *1*(6), 90–95.

[5] Liu, Y., Li, Z., Xiong, H., Gao, X., Wu, J., & Wu, S. (2010). Understanding and enhancement of internal clustering validation measures.*IEEE Transactions on Cybernetics*, *43*(3), 982–994.

[6] Mehar, A. M., Matawie, K., & Maeder, A. (2013). Determining an optimal value of K in K-means clustering. *Proceedings - 2013 IEEE International Conference on Bioinformatics and Biomedicine, IEEE BIBM 2013*, 51–55.

[7] Mihai, D., & Mocanu, M. (2015). Statistical considerations on the k -means algorithm. *Annals of the University of Craiova, Mathematics and Computer Science Series*, *42*(2), 365–373.

[8] Milligan, G. W., & Cooper, M. C. (1985). An examination procedure for determining the number of clusters in a data set. *Psychometrika*, *50*(2), 159–179.

[9] Muca, M., & Kutrolli, G. (2015). A Proposed Algorithm for Determining the Optimal Number of Clusters.*European Scientific Journal*,*11*(36), 112–120.

[10] Napoleon, D., & Lakshmi, G. (2010). An Enhanced *k*-means algorithm to improve the Efficiency Using Normal Distribution Data Points. *International Journal on Computer Science and Engineering*, *2*(7), 2409–2413.

[11] Singh, A., Rana, A., & Yadav, A. (2013). *k*-means with Three different Distance Metrics. *International Journal of Computer Applications*,*67*(10), 13–17.