

A Novel Center Point Initialization Technique for K-Means Clustering Algorithm

Dauda Usman¹ & Ismail Bin Mohamad¹

¹ Department of Mathematical Sciences, Faculty of Science, Universiti Teknologi Malaysia, UTM Johor Bahru, Johor Darul Ta'azim, Malaysia

Correspondence: Dauda Usman, Department of Mathematical Sciences, Faculty of Science, Universiti Teknologi Malaysia, UTM Johor Bahru 81310, Johor Darul Ta'azim, Malaysia. E-mail: dauusman@gmail.com

Received: June 5, 2013 Accepted: July 19, 2013 Online Published: August 2, 2013

doi:10.5539/mas.v7n9p10 URL: <http://dx.doi.org/10.5539/mas.v7n9p10>

Abstract

Clustering is a major data analysis tool utilized in numerous domains. The basic K-means method has been widely discussed and applied in many applications. But unfortunately failed to offer good clustering result due to the initial center points are chosen randomly. In this article, we present a new method of centre points initialization and we prove that the distance of the new method follows a Chi-square distribution. The new method overcomes the drawbacks of the basic K-means. Experimental analysis shows that the new method performs well on infectious diseases dataset when compare with the basic K-means clustering method and a histogram measures the quality of the new method.

Keywords: center initialization, chi-square distribution, cluster analysis, data standardization, principal component analysis

1. Introduction

The massive quantity of information gathered and input into databases brings up the necessity of efficient exploration technique which can utilize the information contained unconditionally there. Among the initial data exploration work is clustering, which enables a person to comprehend pattern and natural groupings within the datasets. Hence, enhancing clustering techniques continues to be getting a lot of interest. The aim would be to cluster the items in the databases to some group of significant subclasses (Ankerst et al., 1999).

Information are generally preprocessed by means of data selection, data integration, data transformation as well as data cleaning and ready for the exploration process. The exploration may be carried out in different databases as well as data repositories, though the styles available were laid out in different exploration benefits such as concept/class description, classification, association, prediction, correlation analysis, cluster analysis and so on.

Cluster analysis is a method of grouping certain sets of designs in to different groups. This is accomplished so that designs within the same groups are similar, while designs in different groups are dissimilar. Cluster analysis has become commonly studied problems in various usage areas such as knowledge discovery as well as data mining (Fayyad et al., 1996).

In clustering methods which are determined by reducing a proper objective function, the most commonly utilized and practiced might be k-means method. For some n data items on real d -dimensional space, R^d , with an integer k , the issue is to ascertain some sets of k items on R^d , known as centres, in order to reduce a mean squared distance from every data point to their closest centre.

A basic stage for k-means cluster analysis is straightforward. At first we decide how many groups 'k' so we presume the centres of such groups. It will consider any random items to be the first centre or an initial k items within the series also can function as a first centroid. After that k-means technique performs three of the stages here till converge:

Iterate till stable (Means zero item transfer groups):

- i. Decide a centre coordinate.
- ii. Decide a distance for each item to the centres.
- iii. Cluster the items according to minimal distance.

Although the basic k-means approach features some benefits more than alternate data clustering approaches, also, it seems to have shortcomings; it converges usually to the local optimum (Anderberg, 1973), the end effect is determined by the initial centroids.

The two relatively easy methods for cluster centre initialization are, to randomly decide for the initial values or to select any initial k samples in the data items. Rather than, other sets of initial values are selected (from the data items) also, the set that will be nearest to optimal, can be selected. Then again, examining different initial sets is considered impracticable criteria, especially for a large number of clusters (Rajashree et al., 2010). Again the computational complexity of the basic k-means technique may be very excessive, specifically with regard to huge data sets. Additionally the amount of distance computations rises tremendously having increases on the dimensionality of the dataset. Once dimensionality increase normally, just one particular dimension is significant for specific clusters, however data within insignificant dimension might possibly yield a lot of noise also it will conceal original groups to be found. Furthermore if dimensionality increase, information usually become continuously minimal, which means that data items positioned from various dimensions can be viewed as all equally distanced as well as distance estimate, basically for clustering technique, turns into the incomprehensible. For this reason, feature reduction or perhaps dimension reduction might be a very important data-preprocessing procedure regarding clustering technique with dataset acquiring huge amount of attributes/features (Rajashree et al., 2010).

Principal Component Analysis by Valarmathie et al. (2009) is an unsupervised feature reduction technique concerning predicting higher dimension dataset to a different reduces dimension dataset that represents most of the variance within a dataset with minimal reconstruction error. Dimensionality reduction by (Rajashree et al., 2010) is the transformations of higher dimension dataset to low dimension correspond to the intrinsic dimensionality of the dataset. It is categorized in to two classes, that is feature reduction and feature selection. Feature Selection criteria aims at obtaining the subsets of the extremely representative features based on a few goal functions within discrete space. The methods of these are normally greedy. Therefore, they generally can not actually discover the optimum solutions within a discrete space. Feature Extraction methods aims at extracting features through projecting an initial higher dimension dataset to a lower dimension space by means of algebraic transformation. This reaches an optimum solution of the problems in a continuous space, however the computations intricacy will be more compared to feature selection criteria. Numerous feature reduction techniques were proposed. Principal component analysis is the frequently employed feature reduction technique concerning reducing reconstruction errors.

A number of efforts have been made by research workers to enhance the performance and effectiveness on a basic k-means technique. Yuan (2004) presented an organized way of selecting an initial center point, but his approach fails to propose an enhancement for the time intricacy on the k-means technique. Belal and Daoud (2005) presented a new technique to cluster centers by considering a group of medians obtained from the dimensions having optimum variance. Zoubi (2008) presented a technique to improve k-means cluster analysis when avoiding unnecessarily distance computations using the partial distance logic. Fahim (2009) presented an approach for selecting an excellent initial solution by means of dividing datasets into blocks than also employing k-means for each block, however the intricacy for the time is a bit more. Although the technique above can acquire effective initial centres for some level, they tend to be very complicated while many utilize k-means technique in their techniques, and is also have to utilize a method of randomly selecting center point. Deelers and Auwatanamongkol (2007) presented a method to enhance k-means clustering technique in accordance with data partitioning technique utilized for color quantization. This technique carries out data partitioning on the data axis considering the maximum variances. Nazeer and Sebastian (2009) presented a better k-means technique, includes an organized way of getting initial center points with a new effective approach of assigning data items into their clusters. This approach guarantees the whole procedure for grouping within $O(n^2)$ time while not compromising correctness for the clusters. Furthermore (Xu et al., 2009) stipulate a new initialization structure to choose initial cluster centres using reverse nearest neighbor lookup. Yet the whole techniques above fail to function effectively with huge dimension datasets. Yeung and Ruzzo (2000) presented an empirical exploration with principal component analysis for grouping gene expression datasets, still the initial center points were also selected here at random. Chao and Chen (2005) as well presented an approach regarding dimensions reduction for microarray data exploration employing Locally Linear Embedding.

Karthikeyani and Thangavel (2009) enhanced k-means clustering technique through the use of global normalization prior to carrying out the cluster analysis in distributed dataset, while not always getting each of the information to a one site. The efficiency for the proposed normalization centered distributed k-means clustering technique was evaluated alongside of distributed k-means clustering technique and normalization centered

directed k-means clustering technique. The clustering level has also been evaluated with three normalization methods, the z-score, decimal scaling as well as min-max with the suggested distributed clustering technique. A comparison test revealed that a distributed cluster effects rely upon the kind of normalization method. Alshalabi et al. (2006) designed an experiment to evaluate the impact for various normalization procedures for consistency as well as preciseness. The experiment results suggested choosing the z-score normalization as the method that will give a much better accuracy.

2. Materials and Methods

Standardization of the original dataset: The initial dataset are scaled with mean 0 and variance 1. The position as well as scale information with the initial variables has been missed Jain and Dubes (1988). An essential limitation with the z-score standardization z is that, it is used for global standardization rather than within-cluster standardization (Milligan & Cooper, 1988). The second method applied is the principal component analysis for outliers detection and removal.

Computing principal components of the standardized dataset: The number of principal components obtained will be identical with the initial variables also to clear away the weaker components from the set of principal component, we obtained the corresponding variance, percentage of variance and cumulative variances in percentage shown in Table 2. Then considered principal components with variances below the mean variance and disregarding the others. The reduced principal components are shown in Table 3.

Acquiring the reduced dataset utilizing reduced principal components: The transformation matrix with reduced principal components is formed which can be used for further data analysis. The reduced dataset Y is used for further analysis shown in Table 4.

Initialization of the Center points: The stairways for the k-means clustering center point initialization are highlighted below

Stage 1: Center point initialization

- 1) Set $m = 1$.
- 2) Work out the distance among each data points from the set Y .
- 3) Select any two data point y_i and y_j in a way that distance (y_i, y_j) is at the maximum.
- 4) $Cen[m] = y_i; Cen[m+1] = y_j; m = m + 2$;
- 5) Eliminate y_i, y_j from the set Y .
- 6) If $(m \leq k)$.

For $i = 1$ to $m - 1$, obtain a distance of each object in Y to $Cen[i]$.

Obtain an average of the distances to the centroid for each object in Y .

Pick the data object y_o acquiring highest average distance from earlier centroids.

$Cen[m] = y_o; m = m + 1$;

Eliminate the object y_o out of Y .

Stage 2: K-means clustering considering the initial centroids succumbed $Cen[]$.

7) Find the nearest cluster center for each data point in Y from the list of Cen , which is closest, than assign that data point to the corresponding cluster.

8) Update the cluster centers in each cluster using the mean of the data points, which are assigned to that cluster.

Re-iterate the steps 7 and 8 up to the point there is little or no further variations in the centroids.

3. Results and Discussions

In this section, we show that the new method is normal and follows a Chi-square distribution. We analysed and compare the results of the basic and new methods. We also evaluate the accuracy of the two approaches, whereby accuracy is measured by the error sum of squares for the intra-cluster range, that is a distance between data vectors in a group and the centroid of the cluster, the smaller the sum of the differences is, the better the accuracy of clustering.

3.1 The New Distance Follows a Chi-square Distribution

As original k-means distance follows a Chi-square distribution the new method also follows a Chi-square distribution. Consider Figure 1 below with two groups, cluster 1 and cluster 2 having $\bar{y}_1, \bar{y}_2$, as the random

selected center points by conventional method for cluster 1 and cluster 2 respectively. Also consider y_{c1} and y_{c2} as the new centers by the new method for cluster 1 and cluster 2 respectively. The Sum of Squares Error for a basic k-means method $SSE = \sum (y - \bar{y}_1)^2 + \sum (y - \bar{y}_2)^2$ which is Chi-square.

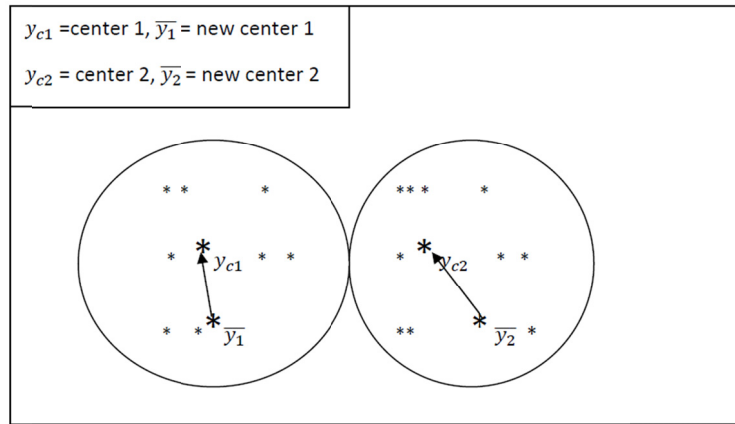


Figure 1. A cluster formation containing two centers for cluster 1 and cluster 2

Proof: The Sum of Squares Error (SSE) for the new method is

$$\sum (y - (\bar{y}_1 + y_{c1}))^2 + \sum (y - (\bar{y}_2 + y_{c2}))^2 \quad (1)$$

Consider the first center point from Equation 1, that is $\sum (y - (\bar{y}_1 + y_{c1}))^2$ which can also be re-writing as

$$\sum (y - \bar{y}_1 - y_{c1})^2$$

Therefore

$$\begin{aligned} \sum (y - \bar{y}_1 - y_{c1})^2 &= \sum (y - \bar{y}_1 - y_{c1}) (y - \bar{y}_1 - y_{c1}) \\ &= \sum (y^2 - y\bar{y}_1 - y y_{c1} - y\bar{y}_1 + \bar{y}_1^2 + y_{c1}\bar{y}_1 - y y_{c1} + y_{c1}\bar{y}_1 + y_{c1}^2) \end{aligned}$$

Collecting the like terms we have:

$$\begin{aligned} &= \sum (y^2 - 2y\bar{y}_1 - 2y y_{c1} + \bar{y}_1^2 + 2y_{c1}\bar{y}_1 + y_{c1}^2) \\ &= \sum (y^2 - 2y\bar{y}_1 - 2y y_{c1}) + \sum (\bar{y}_1^2 + 2y_{c1}\bar{y}_1 + y_{c1}^2) \end{aligned} \quad (2)$$

Than we add $\pm 4y y_{c1}$ into Equation 2, we have:

$$\begin{aligned} &= \sum (y^2 - 2y\bar{y}_1 - 2y y_{c1} + 4y y_{c1}) + \sum (\bar{y}_1^2 + 2y_{c1}\bar{y}_1 - 4y_{c1}\bar{y}_1 + y_{c1}^2) \\ &= \sum (y^2 - 2y\bar{y}_1 + 2y y_{c1}) + \sum (\bar{y}_1^2 - 2y_{c1}\bar{y}_1 + y_{c1}^2) \end{aligned} \quad (3)$$

For independence assumption the covariance between $2y$ and y_{c1} equals to $\bar{y}_1^2$, that is:

$$\begin{aligned} &= \sum (y^2 - 2y\bar{y}_1 + y_{c1}^2) + \sum (\bar{y}_1^2 - 2y_{c1}\bar{y}_1 + y_{c1}^2) \\ &= \sum (y - \bar{y}_1)^2 + \sum (\bar{y}_1 - y_{c1})^2 \end{aligned} \quad (4)$$

Also for the second center point, from Equation 1 $\sum (y - (\bar{y}_2 + y_{c2}))^2$ which can also be re-writing as

$$\sum (y - \bar{y}_2 - y_{c2})^2$$

Therefore

$$\begin{aligned}\sum (y - \bar{y}_2 - y_{c2})^2 &= \sum (y - \bar{y}_2 - y_{c2}) + (y - \bar{y}_2 - y_{c2}) \\ &= \sum (y^2 - y\bar{y}_2 - y y_{c2} - y\bar{y}_2 + \bar{y}_2^2 + y_{c2}\bar{y}_2 - y y_{c2} + y_{c2}\bar{y}_2 + y_{c2}^2)\end{aligned}$$

Collecting the like terms we have:

$$\begin{aligned}&= \sum (y^2 - 2y\bar{y}_2 - 2y y_{c2} + \bar{y}_2^2 + 2y_{c2}\bar{y}_2 + y_{c2}^2) \\ &= \sum (y^2 - 2y\bar{y}_2 - 2y y_{c2}) + \sum (\bar{y}_2^2 + 2y_{c2}\bar{y}_2 + y_{c2}^2)\end{aligned}\quad (5)$$

Than we add $\pm 4y y_{c2}$ into Equation 5, we have:

$$\begin{aligned}&= \sum (y^2 - 2y\bar{y}_2 - 2y y_{c2} + 4y y_{c2}) + \sum (\bar{y}_2^2 + 2y_{c2}\bar{y}_2 - 4y_{c2}\bar{y}_2 + y_{c2}^2) \\ &= \sum (y^2 - 2y\bar{y}_2 + 2y y_{c2}) + \sum (\bar{y}_2^2 - 2y_{c2}\bar{y}_2 + y_{c2}^2)\end{aligned}\quad (6)$$

For independence assumption the covariance between $2y$ and y_{c2} equals to $\bar{y}_2^2$, that is:

$$\begin{aligned}&= \sum (y^2 - 2y\bar{y}_2 + y_{c2}^2) + \sum (\bar{y}_2^2 - 2y_{c2}\bar{y}_2 + y_{c2}^2) \\ &= \sum (y - \bar{y}_2)^2 + \sum (\bar{y}_2 - y_{c2})^2\end{aligned}\quad (7)$$

Adding Equation 4 and 7 we have

$$= \sum (y - \bar{y}_1)^2 + \sum (\bar{y}_1 - y_{c1})^2 + \sum (y - \bar{y}_2)^2 + \sum (\bar{y}_2 - y_{c2})^2 \quad (8)$$

Hence Equation 8 also follows a Chi-square distribution.

3.2 Experimental Analysis

In order to test our algorithm we used an infectious diseases dataset. We compare the analysed results of the k-means algorithm with the two different initialization techniques, which are the random initialization technique and the new technique, respectively. The experimental result of the cluster analysis shows that the new initialization approach outperforms the basic clustering approach.

Table 1. The original datasets with 20 data objects and 8 attributes

	X1	X2	X3	X4	X5	X6	X7	X8
Day 1	9	6	4	3	2	5	2	1
Day 2	7	5	5	5	1	1	1	3
Day 3	7	2	3	2	2	3	2	3
Day 4	6	3	2	1	1	3	2	2
Day 5	10	5	3	3	3	2	5	1
Day 6	12	5	6	1	5	2	4	1
Day 7	8	3	2	3	1	2	3	1
Day 8	9	2	3	1	1	3	3	1
Day 9	11	3	2	1	7	1	3	2
Day 10	7	7	2	2	1	2	1	2
Day 11	10	5	7	9	1	6	1	1
Day 12	13	9	5	4	3	2	5	1
Day 13	11	3	4	3	1	2	3	1
Day 14	8	2	3	5	2	1	2	2
Day 15	7	3	1	2	1	2	3	3
Day 16	15	4	3	4	2	1	3	2
Day 17	9	4	1	7	2	3	1	1
Day 18	14	3	2	2	1	1	2	1
Day 19	15	2	3	8	1	2	2	1
Day 20	9	5	1	1	2	2	1	2

Table 2. The variances cumulative percentages

	Variances	Percentage of Variances	Cumulative Percentage of Variances
PC1	2.2727	28.4088	28.4092
PC2	2.0346	25.4325	53.8421
PC3	1.1806	14.7575	68.6000
PC4	0.7914	9.8925	78.4921
PC5	0.6627	8.2837	86.7752
PC6	0.5350	6.6875	93.4624
PC7	0.3141	3.9265	97.3882
PC8	0.2089	2.6112	100.0000

Table 2 presents the variances, the percentage of the variances and cumulative percentage which corresponds to the principal components of the original dataset.

Table 3. Reduced principal components

PC1	PC2	PC3	PC4	PC5
0.4533	-0.1835	0.5140	-0.1944	0.1108
0.3228	0.0358	-0.5673	-0.2449	0.6863
0.4559	0.2170	-0.2707	-0.3110	-0.3893
0.2302	0.4772	0.3449	-0.4203	-0.0625
0.2133	-0.4609	-0.2340	-0.1613	-0.5043
0.1603	0.5117	-0.3092	0.4568	-0.2823
0.3425	-0.4589	-0.0883	0.2393	-0.0219
-0.4909	-0.0739	-0.2523	-0.5805	-0.1638

Table 3 presents the reduced principal components that have variances greater than mean variance. But the number of principal components found is the same with the number of the original dataset, here we present only the eighty percent (applying pareto law) to be considered for further analysis.

Table 4. The reduced dataset with 20 data objects and 5 attributes

	X1	X2	X3	X4	X5
Day 1	1.0138	1.4300	-1.3632	1.0157	0.0394
Day 2	-1.2300	0.9742	-0.8015	-2.2844	0.0745
Day 3	-1.9341	0.1725	-0.6831	-0.1909	-1.2645
Day 4	-1.7823	0.3215	-0.6716	0.9849	-0.1303
Day 5	1.3276	-1.3779	-0.3584	0.6914	0.2175
Day 6	2.2886	-1.7366	-0.9953	-0.0643	-0.9875
Day 7	-0.4555	-0.0791	0.4903	1.1052	0.3142
Day 8	-0.2591	-0.0398	0.2934	1.6813	-0.4171
Day 9	-0.1070	-2.9462	-0.2386	-0.4974	-1.4578
Day 10	-1.2239	0.5107	-1.2592	-0.3359	1.6031
Day 11	2.1184	3.9972	-0.5116	-0.2817	-1.0310
Day 12	3.1675	-1.0375	-1.2102	-0.6025	1.3251
Day 13	0.5897	-0.0158	0.7215	0.5170	-0.0368
Day 14	-1.0760	0.0168	0.7459	-0.7207	-0.6518
Day 15	-2.2945	-0.5396	-0.3456	0.0001	0.1014
Day 16	0.6047	-0.9815	1.2208	-1.1093	0.3802
Day 17	-0.3022	1.3812	0.8661	0.2468	0.3512
Day 18	0.0336	-0.6991	1.7731	0.3133	0.8160
Day 19	1.0005	0.9373	2.7306	-0.5172	-0.1246
Day 20	-1.4798	-0.2883	-0.4036	0.0486	0.8789

Table 4 presents the transformed dataset having 20 data objects and 5 attributes which are generated using the reduced principal component analysis and the original dataset shown in Table 3 and 1 respectively.

Table 5. Summary of error sum squares and time taken

	Basic k-means algorithm	Proposed technique
SSE	175.00	74.01
Time taken in ms	82	69

Table 5 presents the error sum squares and respective time taken obtained for both basic k-means clustering algorithm and the proposed technique. The result also shows that the new technique provides better error sum squares and the time taken for the execution also reduced.

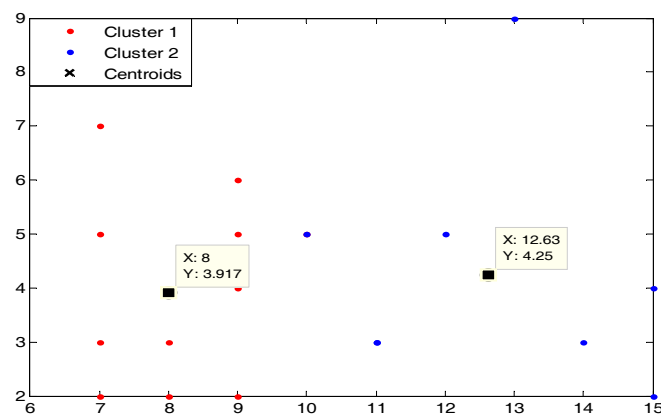


Figure 2. Basic k-means algorithm

Figure 2 presents the result of the basic *k*-means algorithm using the original dataset having 20 data objects and 8 attributes as shown in Table 1. Indicating three points attached to both cluster 1 and 2 are out of the cluster formation, indicating the presence of outliers. The intra-cluster distance is very high while the inter cluster distance is also very small with the error sum of squares equal 175.00.

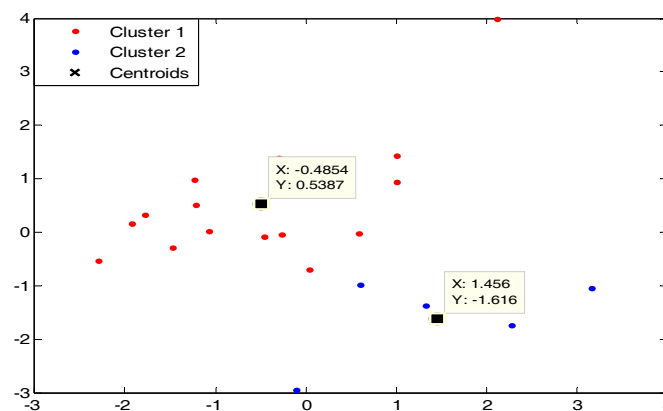


Figure 3. New k-means clustering technique

Figure 3 presents the result of the basic *k*-means algorithm using the reduced dataset, having 20 data objects and 5 attributes as shown in Table 4. The intra-cluster distance is very small and the inter cluster distance is very high with the error sum of squares equal 74.01.

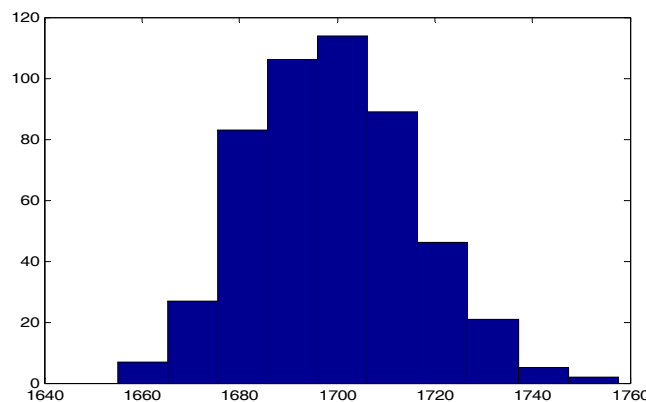


Figure 4. A histogram of ESSs against 100 runs

Figure 4 presents a histogram of error sum of squares when the new method is run one hundred times, the histogram skewed to the right indicates that the new distance method follows a Chi-square distribution.

4. Conclusion

Many applications rely on the clustering techniques. One of the most widely used clustering approaches is k-means clustering. In this article a new method of center point initialization is proposed to produce optimum quality clusters and we prove that the new distance method follows a Chi-square distribution. Comprehensive experiments on infectious diseases datasets have been conducted in a manner that the sum of the total clustering errors was reduced as much as possible whereas inter distances between clusters are preserved to be as large as possible for better performance. The experimental result of the cluster analysis shows that the new initialization approach outperforms the basic clustering approach.

References

- Alshalabi, L., Shaaban, Z., & Kasasbeh, B. (2006). Data Mining: A Preprocessing Engine. *Journal of Computer Science*, 2(9), 735-739. <http://dx.doi.org/10.3844/jcssp.2006.735.739>
- Anderberg, M. R. (1973). *Cluster Analysis for Applications*. Academic Press.
- Ankerst, M., Breunig, M., Kriegel, H. P., & Sander, J. (1999). OPTICS: Ordering Points to Identify the Clustering Structure. *Proc. ACM SIGMOD Int. Conf. Management of Data Mining*, 49-60.
- Belal, M., & Daoud, A. (2005). A new algorithm for cluster initialization. *World Academy of Science, Engineering and Technology*, 4, 74-76.
- Chao, S., & Chen, L. (2005). Feature dimension reduction for microarray data analysis using locally linear embedding. *3rd Asia Pacific Bioinformatics Conference*, 211-217. http://dx.doi.org/10.1142/9781860947322_0021
- Deelers, S., & Auwatanamongkol, S. (2007). Enhancing K-means algorithm with initial cluster centers derived from data partitioning along the data axis with the highest variance. *International Journal of Computer Science*, 2(4), 247- 252.
- Fahim, A. M., Salem, A. M., Torkey, F. A., Saake, G., & Ramadan, M. A., (2009). An Efficient K-means with Good Initial Starting Points. *Georgian Electronic Scientific Journal: Computer Science and Telecommunications*, 2(19), 47-57.
- Fayyad, U. M., Piatetsky, S. G., Smyth, P., & Uthurusamy, R. (1996). *Advances in Knowledge Discovery and Data Mining*, AAAI/MIT Press.
- Jain, A., & Dubes, R. (1988). *Algorithms for Clustering Data*. Englewood Cliffs, NJ: Prentice Hall.
- Karthikeyani, V. N., & Thangavel, K. (2009). Impact of Normalization in Distributed K-means Clustering. *International Journal of Soft Computing*, 4(4), 168-172.
- Milligan, G., & Cooper, M. (1988). A study of standardization of variables in cluster analysis. *Journal of Classification*, 5, 181-204. <http://dx.doi.org/10.1007/BF01897163>

- Nazeer, K. A., & Sebastian, M. P. (2009). Improving the accuracy and efficiency of the k means clustering algorithm. *Proceedings of the World Congress on Engineering*, 1, 308-312.
- Rajashree, D., Debahuti, M., Amiya, K. R., & Milu, A. (2010). A hybridized K-means Clustering Approach for High Dimensional Dataset. *International Journal of Engineering, Science and Technology*, 2(2), 59-66.
- Valarmathie, P., Srinath, M., & Dinakaran, K. (2009). An increased performance of clustering high dimensional data through dimensionality reduction technique. *Journal of Theoretical and Applied Information Technology*, 13, 271-273.
- Xu, J., Baowen, X., Zhang, W., Zhang, W., & Hou, J. (2009). Stable initialization scheme for K-means clustering. *Wuhan University Journal of Natural Sciences*, 14(1), 24-28. <http://dx.doi.org/10.1007/s11859-009-0106-z>
- Yeung, K. Y., & Ruzzo, W. L. (2000). *An empirical study on principal component analysis for clustering gene expression Data*. Tech. Report, University of Washington.
- Yuan, F., Meng, Z. H., Zhang, H. X., & Dong, C. R. (2004). A new algorithm to get the initial centroids. *Proc. of the 3rd International Conference on Machine Learning and Cybernetics*, pp. 1191-1193.
- Zoubi, M. B., Hudaib, A., Huneiti, A., & Hammo, B. (2008). New Efficient strategy to Accelerate K-means Clustering Algorithm. *American Journal of Applied Sciences*, 5(9), 1247-1250. <http://dx.doi.org/10.3844/ajassp.2008.1247.1250>

Copyrights

Copyright for this article is retained by the author(s), with first publication rights granted to the journal.

This is an open-access article distributed under the terms and conditions of the Creative Commons Attribution license (<http://creativecommons.org/licenses/by/3.0/>).