ICIC Express Letters
Part B: Applications
Volume **12**, Number **4**, April **2021**

ICIC International ©2021 ISSN 2185-2766
pp. **317–323**

VISUAL COMPARISON OF CLUSTERING USING LINK-BASED CLUSTERING METHOD (LBCM) WITHOUT PREDETERMINING INITIAL CENTROID INFORMATION

S. M. F. D. Syed Mustapha[1], Biju Theruvil[2] and Mukesh Madanan[2]

[1]College of Technological Innovation
Zayed University
P.O. Box 19282, Dubai, United Arab Emirates
Syed.Duani@zu.ac.ae

[2]Department of Computer Science
Dhofar University
P.O. Box 2509, Salalah 211, Sultanate of Oman
{ b_sayed; mukesh }@du.edu.om

Received September 2020; accepted December 2020

Abstract. *High dimensional data are difficult to view in two-dimensional plot. However, having a mechanism to reduce to a selected number of salient features that can well present the data is essential. We attempted to reduce N dimensional data to two-dimensional data using the combination of Information Gain (IG) and Principal Component Analysis (PCA) and to perform the link-based clustering which is our novel technique presented in this work in determining the linked clusters automatically using visual approach. Link-based Clustering Method (LbCM) is applied on the two-dimensional data to determine the clusters automatically. The significance of the method is that it does not require prior information such as the number of linked clusters. The approach using a combination of IG-PCA for feature selection is also useful to deal with high dimensional data. The LbCM is able to detect the number of linked clusters automatically by analyzing the X-Y coordinate positions of the points and visual information such as gaps between points and of two extreme points for both axes. Since the number of clusters is represented visually in two dimensions, LbCM performance can be compared visually.*
Keywords: Clustering algorithm, Information gain, Density-based clustering, Principal component analysis

1. **Introduction.** The motivation of reducing features to two dimensions is that most clusters are evaluated based on the data that have the tendency to produce clusters, number of clusters and the quality of clusters. One of the means to do so is through visual observation on the data plotted (2 or 3 dimensions) on the graph or to perform Hopkins test [1] to determine the existence of clusters. The determination of cluster has been an ongoing research area as it is subject to the scale of the data plot, distribution of the data and the cluster requirement set by the functional aspects such as business requirements. In our research, we explore the approaches in dealing with datasets taken from one of the banks in Portugal [2] and the datasets were made available in UCI repository [3].

We argue that even though PCA could reduce the dimension, in some cases, when it is used alone without IG, the selected features are purely on eigenvalues and omitting the physical meaning of the features [2,4]. PCA is unable to measure the contribution of each feature towards the decision categories. The integration of IG and PCA is essential and there are attempts by some research works such as in text categorization [5]. The two-stage method suggests the use of IG to reduce the number of words (so-called features) to reduce the computational complexity. It is possible to deploy genetic algorithm to

DOI: 10.24507/icicelb.12.04.317

optimize the search in determining IG values. Another work by Uguz [6] in integrating PCA and IG was in classifying TCD (Thermal Conductivity Detector) signals in which comparison was made on the performances of the individual IG, PCA and the integration of both. It was suggested that the integration of IG-PCA is higher and consistent for the selected feature ranges between 10%-100%. In our view, TCD signals are time series data that have no physical meaning at various points of time except for the peak values exhibited. In our work, each feature has physical meaning and hence, we regard IG will be useful for selecting features while PCA is used for mapping the selected features made by IG to X-Y coordinate for the purpose to determine the clusters. The previous two works are similar to the current proposed one such that IG is used on the dataset to capture the most salient physical meaning based on plain graph line or terms with high occurrences, while the current work regards attributes with physical meaning such as age, and balance account to determine the salient features. In the text categorization, the number of occurrences is used to evaluate the importance of the terms.

The main contributions of this paper are, firstly, to demonstrate the detection of clusters using Link-based Cluster Method (LbCM) which can be done without the need of prior knowledge on the number of clusters and the calculation of the centroids or medoids; secondly, to demonstrate the use of feature selections which are the Information Gain (thereafter, IG) and Principal Component Analysis (thereafter, PCA) in determining the two essential features. It is known that IG reduces number of attributes and subsequently it lightens the computational burden when calculating the PCA. IG does not have the advantage of transforming to 2-dimensional data which can be done using PCA. The remainder of this paper is organized as follows: Section 2 justifies the need of link-based clustering method, followed by the discussion on centroid and medoids in Section 3, approach and technique at Section 4, procedure determining affinity in Section 5, the experimental results at Section 6, and conclusions in Section 7.

2. **Why Link-based Clustering Method (LbCM).** It is common for most clustering methods to pre-determine two pieces of essential information which are the i) possible number of clusters, ii) initial value for medoids or centroids. In addition to these, there are issues such as determining the threshold value for the stopping criteria before the final number of clusters is optimal, estimating an acceptable difference in values between members in the clusters and the centroids/medoids of their cluster and determining the additional weight for some attributes that are significant in determining the similarity or dissimilarity values.

3. **Centroids or Medoids.** As shown in Figure 1(a), the two clusters are clearly separated and if the number of clusters is correctly specified, two clusters can be correctly determined. However, in Figure 1(b), the two clusters are closely joint, and it can be interpreted as two clusters or even one cluster. For a Link-based Clustering Method (LbCM), each member of a data point X is considered as a member of a group of a cluster A if there exists data point within the members of the cluster A that has "close link" with the data point X. In this case, the two clusters in Figure 1(b) will be computed as a single cluster since there are points at the tip of both clusters that are close. Hence, an LbCM shall treat an elongated cluster to be a single cluster as shown in Figure 2 where the centroid is far from the two extreme points at the edge of the cluster but both data points are within the same cluster.

Figure 3 shows how the data can be seen overlapping to each other and this depends on the angle and the dimension chosen for the view. For example, Figure 3(a) shows that the three clusters are disintegrated while Figure 3(b) shows some overlapping on parts of the points at the edges of each cluster. Since there are N possible dimensions that the data points can be associated, N has to be reduced to smaller M dimensions where only

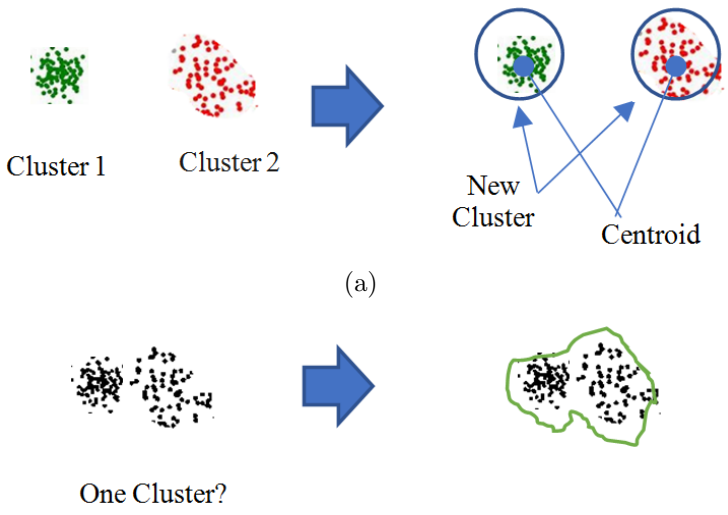

FIGURE 1. Two types of clusters

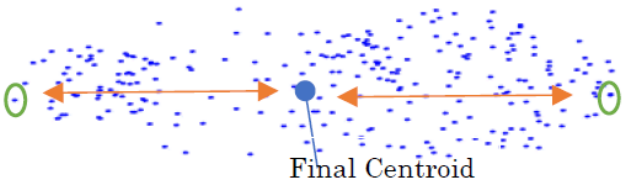

FIGURE 2. Elongated cluster

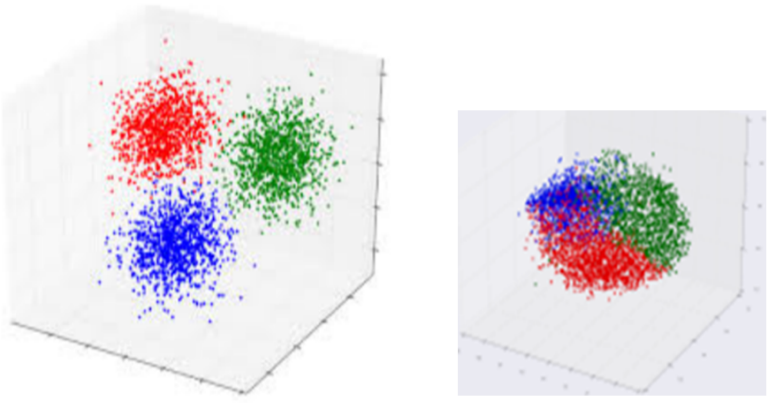

FIGURE 3. (color online) Cluster outlook at different angles and dimensions

the important dimensions will be selected. We adopt IG-PCA as an integrated technique to reduce the dimension based on the following principles.

i) Features are selected based on how much those features could contribute to giving information in order to reduce the uncertainties in selecting final decision. IG is applied to each feature; the information gain for each feature is calculated and the M features are selected based on the score obtained from calculating the IGs.

ii) LbCM analyzes data points based on two-dimensional plot. Without reducing the size of M dimensions which has been determined using IG, PCA is used to transpose the M dimensions plot into X-Y coordinate plot.

In summary, LbCM is used for the following three reasons:

i) The dimensions that are used on LbCM are the salient dimensions that are sufficient to be used for calculating the data points that are considered associated;

ii) LbCM is fast in determining the number of clusters and also it does not use any centroid or mediods in determining the clusters;

iii) LbCM allows a display of the result of clustering through the X-Y dimensional plots.

The following section discussed the LbCM techniques.

4. **Link-based Clustering Method (LbCM) – Approach and Technique.** LbCM applies a simple principle that two data points are associated or close to each other if one or more coordinate points is/are near to each other.

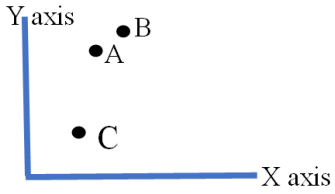

FIGURE 4. An illustration on LbCM data points

If we have three data points, A, B and C as shown in Figure 4, LbCM defines the associations of the data points as follows:

$$A_x \equiv B_x, \quad A_y \equiv B_y, \quad A_x \equiv C_x$$
$$A_y \not\equiv C_y, \quad B_x \not\equiv C_x, \quad B_y \not\equiv C_y$$

$\equiv$ denotes close association on axis.

If P and Q are two points then $P \equiv Q$ iff

$$\begin{cases} X_P - X_Q \leq gap_x \text{ or} \\ Y_P - Y_Q \leq gap_y \end{cases}$$

and X_P is the coordinate at X axis for data point P and gap is determined based on the scale of an axis.

Considering that the scale size of the axis is significant in determining the distance between points, the *gap* is normalized such that $gap_x = \frac{\bar{\bar{X}} - \bar{X}}{X_{\max} - X_{\min}}$ where $\bar{\bar{X}}$ and $\bar{X}$ are the largest and the smallest gaps, respectively. The $X_{\max}$ and $X_{\min}$ are the highest data point and the lowest data point at X axis. Similarly, the gap_y is normalized in a similar fashion.

5. **Procedure Determining Affinity.** The data points are given in a serial form as data input and the following procedure demonstrates steps in determining the affinity between points.

1) $a_i \in A$ where A is a collection of data members;

2) Find $a_j \in A$ where $a_i \equiv a_j$ for all $j = \{1, 2, \ldots, n\}$ where n is $\eta(A)$ which is the number of member data in A;

3) Assume $a_j \in A$ and $j = \{\alpha, \beta, \Omega\}$, find a_j where $a_\alpha \equiv a_j$ where $j = \{1, 2, \ldots, n\}$ and α, β, Ω are any data points;

4) Repeat Step 3) for β, Ω;

5) Assume $a_j \to a_\alpha \to a_{\alpha+k} \to a_{\alpha+n}$ then a_j, a_α, $a_{\alpha+k}$, $a_{\alpha+n}$ are the members of the same cluster;

6) Perform 3) and 4) until there are no more data members in the A list.

The above procedure can be applied in both data members for either one of the axes. The above procedure can be demonstrated in steps as shown in Figure 5 for one cluster.

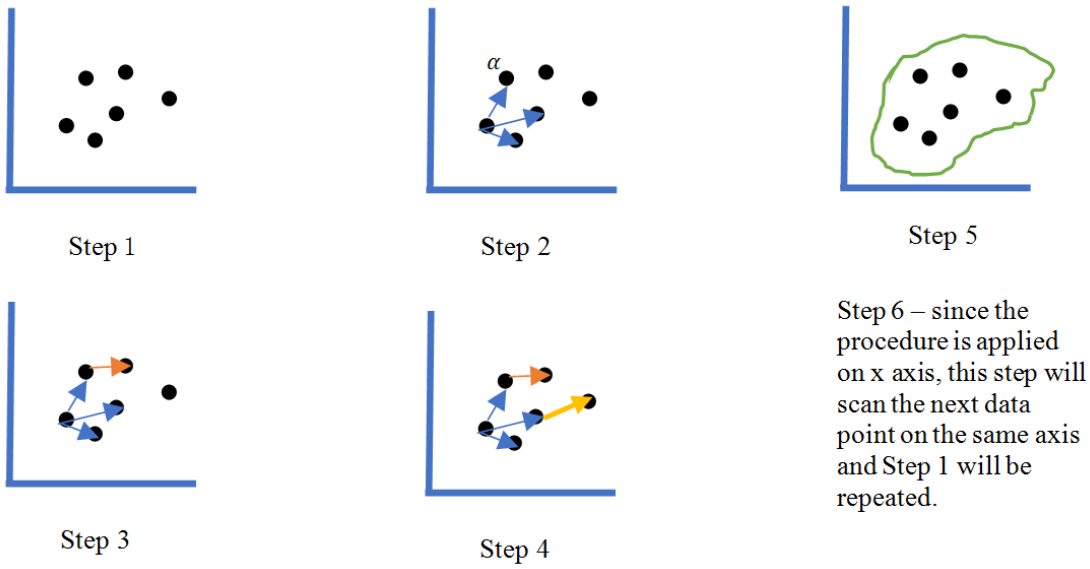

FIGURE 5. Steps illustrating the procedure determining affinity points and cluster building

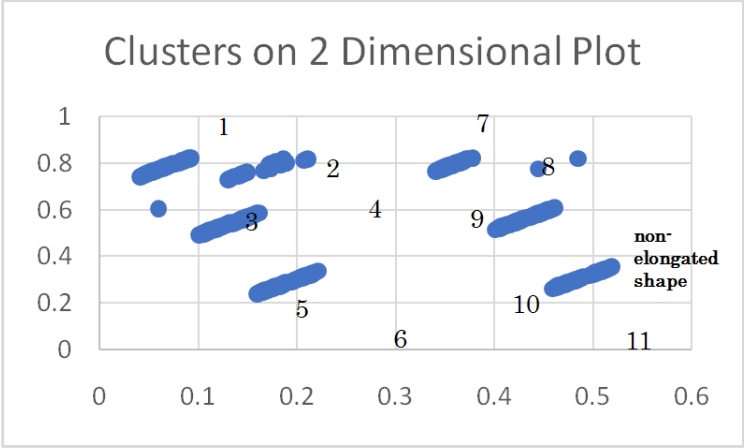

FIGURE 6. Transposition of 8 dimensions to 2 dimensions using IG-PCA

6. **Experiments and Results.** Prior to plotting to two-dimensional data, the X-Y features are determined based on the computation on the IG values and application of PCA to determine the top eigenvalues. Subsequently, the clusters are determined automatically based on the LbCM algorithm. The original data obtained from UCI repository has 16 attributes. The threshold value for the IG value is set to be above 0.01 and there are 8-top IG values that are obtained (refer to [7] for details). Figure 6 shows the data points that are plotted manually on the spreadsheet in order to determine visually the number of link-based clusters. Visually, one can identify that there are 11 main clusters with elongated shapes and 7 island clusters (non-elongated) with round shape. The same data are applied on LbCM procedure to identifying the number of clusters that it can detect automatically, and the result is shown in Figure 7. Based on subjective observation on the cluster plots, there are 12 elongated clusters (small and big) and 7 island clusters with round shape. The single dot does not represent a single datum, but it is a collection of data that have the same values for all attributes while the elongated clusters are representation of the data points of which at least one or more attributes are associated. The result indicates that LbCM is able to determine the clusters without first determining the

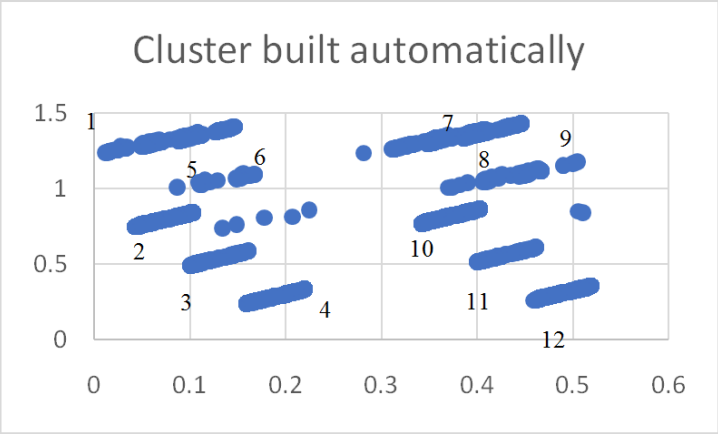

FIGURE 7. Clusters built by LbCM procedure

number of clusters. In comparison to other clustering methods, LbCM works well on the following conditions.

1) The data have some group of serial data points that are elongated in shapes such that the data series in the elongated data points are close to each other with a gap value.

2) There is a gap value such that it represents the maximum value of two coordinate points in the entire collection of data points and the distance of any two points that fall between the gap value is visually seen as being connected or "linked". Hence, a sparse data where every point has equal distance will not work well using LbCM.

3) Data plot is in two dimensions to reduce computational times as the data points require distance calculation to each of its neighboring points.

Our experiment focuses on determining the possibility of detecting the presence of linked clusters by analyzing the coordinate positions of the points and determining a suitable gap value to determine the closeness of two points, hence it is not comparable to the traditional clustering methods that focus on measuring the differences of feature values. Nevertheless, this approach is comparable to density-based clustering [8,9]. Density-based clustering requires the determination of core points and the radius initial estimation to determine the border points (MinPts). To do this, each point is tested whether it has neighboring point such that it can be a core point and the neighboring points will be the border points. This has high computational cost as each point has to be examined and the radius size that is to be used is randomly guessed and adjusted in order to get minimum points. LbCM does not use radius as each point is a potential core point to be tested with its neighboring points. LbCM also does not require to determine the border of each cluster as it continues to determine the neighbor of each point until there is no point that has distance gap less than the gap value from a particular core point. Hence, our investigation on LbCM on this data set has shown that it is possible to determine the cluster on density-based data set. A recent work works on the coordinate distance between points and uses intersection between the means of the points to determine a single cluster [10]. The work also proved our current work in using representative 2-dimension plot to represent the high dimensional attributes of the data as basis to determine cluster. Nevertheless, the authors do not mention any particular method being used in transforming N-dimension to 2-dimension. LbCM can be compared with this recent work as future work.

7. **Conclusions.** Problems in clustering always require having a good estimation on the number of centroids and also in positioning the location of the centroid. This creates a fundamental problem on over-fitting or over-clustering. In our work, the data points are transformed into X-Y plots such that the relations between the data points and associated clusters can be visually observed. The data points in one cluster are associated

by at least one common attribute that forms the elongated shape of the cluster. LbCM is used to detect the presence of such cluster and the finding has shown that these clusters are determined by their positions in the coordinate plots by scanning in the X or Y axis. Without the need to predetermine the cluster and centroids/medoids, the issues on over-fitting or over-clustering do not arise. Nevertheless, LbCM requires some pre-processing tasks which are the feature selection where IG values were used to select essential features and PCA is applied to transforming into two dimensional plots.

The work has some potential research work that can be considered for further investigations. The performance of LbCM can be compared further with density-based clustering method and the variations of its kinds. When plotting the 2-dimensional data, LbCM can be compared with other traditional clustering techniques such as k-means or global clustering in terms of determining the number of clusters and the data points that comprise within the clusters. The comparison can be made in two ways, using the original set data or the transposed data from PCA. The same investigation can be made on various other clustering techniques. Another further work is to further analyze the performance of LbCM based on the sensitivity or recall, accuracy using various assessment technique in machine learning.

REFERENCES

[1] A. Banerjee and R. N. Dave, Validating clusters using the Hopkins statistic, *IEEE International Conference on Fuzzy Systems*, Budapest, Hungary, vol.1, pp.149-153, doi: 10.1109/FUZZY.2004.1375706, 2004.

[2] D. Napolean and S. Pavalakodi, A new method for dimensionality reduction using K-means clustering algorithm for high dimensional data set, *International Journal of Computer Applications*, vol.13, no.7, pp.41-46, 2011.

[3] S. Moro, R. Laureano and P. Cortez, Using data mining for bank direct marketing: An application of the CRISP-DM methodology, *European Simulation and Modelling Conference – ESM'2011*, Guimaraes, Portugal, pp.117-121, 2011.

[4] C. Ding and X. He, K-means clustering via principal component analysis, *Proc. of the 21st International Conference on Machine Learning*, Banff, Canada, ranger.uta.edu/∼chqding/papers/Kmeans PCA1.pdf, 2004, Downloaded on 30 April 2018.

[5] H. Uguz, A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm, *Knowledge-Based Systems*, vol.24, pp.1024-1032, 2011.

[6] H. Uguz, A hybrid system based on information gain and principal component analysis for the classification of trancranial Doppler signals, *Computer Methods and Programs in Biomedicine*, vol.107, pp.598-609, 2012.

[7] S. M. F. D. S. Mustapha and A. Alsufiyani, Application of artificial neural network and information gain in building case-based reasoning for telemarketing prediction, *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol.10, no.3, 2019.

[8] R. Hyde and P. Angelov, Data density based clustering, *The 14th UK Workshop on Computational Intelligence (UKCI)*, doi: 10.1109/UKCI.2014.6930157, 2014.

[9] R. J. G. B. Campello, D. Moulavi and J. Sander, Density-based clustering based on hierarchical density estimates, in *Advances in Knowledge Discovery and Data Mining. PAKDD 2013. Lecture Notes in Computer Science*, J. Pei, V. S. Tseng, L. Cao, H. Motoda and G. Xu (eds.), Berlin, Heidelberg, Springer, 2013.

[10] Z. Nazari, M. Nazari and D. Kang, A bottom-up hierarchical clustering algorithm with intersection points, *International Journal of Innovative Computing, Information and Control*, vol.15, no.1, pp.291-304, doi: 10.24507/ijicic.15.01.291, 2019.