An Effective Analysis of Data Clustering using Distance-based K- Means Algorithm

P.Ramkumar¹, P.Kalamani², C.Valarmathi³, M.Sheela Devi⁴
¹,²,³,⁴ Sri Sairam College of Engineering, Bangalore

e-mail: ¹ramkumarkohila@gmail.com, ²pskalamani@gmail.com, ³vinmathi20@gmail.com, ⁴sheela29183@gmail.com

Abstract. Real-world data sets are regularly provides different and complementary features of information in an unsupervised way. Different types of algorithms have been proposed recently in the genre of cluster analysis. It is arduous to the user to determine well in advance which algorithm would be the most suitable for a given dataset. Techniques with respect to graphs are provides excellent results for this task. However, the existing techniques are easily vulnerable to outliers and noises with limited idea about edges comprised in the tree to divide a dataset. Thus, in some fields, the necessarity for better clustering algorithms it uses robust and dynamic methods to improve and simplify the entire process of data clustering has become an important research field. In this paper, a novel distance-based clustering algorithm called the entropic distance based K-means clustering algorithm (EDBK) is proposed to eradicate the outliers in effective way. This algorithm depends on the entropic distance between attributes of data points and some basic mathematical statistics operations. In this work, experiments are carry out using UCI datasets showed that EDBK method which outperforms the existing methods such as Artificial Bee Colony (ABC), k-means.

Keywords: Artificial Bee Colony, Clustering, Data points, Entropic Distance, K-means, Outliers

1. Introduction

Clustering is one of the unsupervised learning mechanisms which is broadly utilized in various genus of techniques such as recognition of pattern, retrieval of information, image processing and text processing [1-2]. At the time of clustering, data with the major resemblance are keep it in a cluster, as a consequence the data in various clusters have the minimal connection to each other [3]. In recent years, clustering process face the high-dimensional data problems i.e., the objects to be clustered have a huge amount of features [4]. In most of the real-world scenario, only a small segment of the features is considered to be relevant to the cluster structure. A good clustering techniques should be able to recognize the relevant features and discard the negative influences of the noisy features [5-6]. The existing clustering approaches have been partitioned into two varieties such as hierarchical and partitional clustering. In hierarchical clustering, a faction of sequences in clusters will be identified in the form of hierarchical trees in the last part of clustering process [7]. In partitional clustering, dataset
is partitioned into non-overlapping subsets in such a way that each data item will be appropriately in one subset [8]. K-means is a partitioned clustering algorithm that has become one of the most well-known and attractive partitioned clustering methods because of its simplicity, effectiveness and precise computational point [9].

2. Literature Review

Many techniques are addressed by researchers in the data clustering. In this case, brief estimations of some significant contributions to the existing techniques are presented.

Y. Wang, et al., [15] designed a sparse subspace clustering (SSC) algorithm for analyze the time serious data. The problem of measuring the similarity of time series was effectively solved by SSC algorithm. The SSC method was tested on artificial data set and daily box-office data for attaining better results when compared with K-means and spectral clustering algorithms.

Z. D. Bacciu, and Daniele Castellana, [16] designed a two different forms of mixtures of Switching-Parent Hidden Tree Markov Model (SP-BHTMM) to learn structural patterns for clustering applications. The fixed number of components was required by the first form of mixture i.e. finite mixture (MIX-SP-BHTMM), which was used as a hyper-parameters for Expectation Maximization approach (EM).

J. Sui, et al., [17] developed a improved streaming affinity propagation (ISTRAP) depends on an integrated evolution tracking framework. By using ISTRAP, the recurrent clusters were detected in effective way and the outdated clusters were removed rather than being treated as novel clusters. The evolution of data streams such as emergence, disappearance and reoccurrence were detected by ISTRAP which was proven by simulation results on real-world data streams.

3. Proposed Methodology

The existing techniques uses the different distance measures (i.e. Euclidean distance) based on physical characteristics of data to identify the distance which faces many challenges such as time consumption to cluster the data. Mostly, these techniques didn’t concentrate on outliers, hence similarity of data and precision are reduced. The average distance of existing techniques are high which leads to outlier presence and poor accuracy. To overcome the above mentioned issues, the proposed EDBK method focused on entropy based distance for attribute characteristics of data. In this section, this research work first proposes the EDBK to explore the implanted order in sequence for data distance measurement. Then, based on the entropy distance, the outliers are removed. After removing the outliers, the data are clustered by using K-means algorithm. The detail description of proposed method is given below with Figure 1.

The data are collected from UCI dataset which is given as input for normalization process. In this process, the kind of data characteristic are collection of two major classes, i.e., categorical and numerical attributes. The categorical classes are further divided two subclasses, i.e., ordinal attributes and nominal attributes, where ordinal attributes acquired certain tract of nominal attributes. On the one hand, like nominal attributes, the categories (i.e., the possible values) of features in ordinal data, i.e., the data associated with the ordinal attributes only, are all qualitative and not suitable for arithmetic operations such as division, summation and mean.
By using normalization process, these data can be converted into standard forms which are used to cluster the data in an effective way. The data can be clustered by using k-means algorithm, but in these normalized data, there are some outliers will present. To remove the outliers, first the distance between data can be calculated by using entropy based method is introduced.

**Entropy-Based Distance Metric**

In this paper, for a data set \( X = \{x_1, x_2, \ldots, x_N\} \) with \( N \) data objects are signified by \( d \) attributes. From the view of data objects, \( O_r \) is the value space of the \( r \)th value of a data object. The main problem in calculating distance measurement of data is how to estimate the distance contributions of various categories. From the aspects of information theory, a higher entropy value typically indicates a huge amount of information or more ambiguity. A choice with more information or higher uncertainty level usually costs more thinking for a participant. As a result of this, an entropy value of a category is appropriate for indicating its distance contribution.

Thus, the distance between the \( r \)th value of two objects, \( x_i \) and \( x_j \), from a data set \( X \) with \( N \) objects represented by entropy distance which is explained in Eq. (1),

\[
E_{O_r(s)} = - e_{O_r(s)} I_{e_{O_r(s)}} \tag{1}
\]

Where, \( E_{O_r(s)} \) stands for the entropy value of category \( O_r(s) \), item \( e_{O_r(s)} \) stands for the occurrence probability of value \( O_r(s) \) in attribute \( A_r \), which can be written in Eq. (2)

\[
e_{O_r(s)} = \frac{a_{O_r(s)}}{N} \tag{2}
\]

where \( a_{O_r(s)} \) is the number of data objects in the data set \( X \) with their \( r \)th values equal to \( O_r(s) \). Subsequently, the distance between two ordinal data objects \( x_i \) and \( x_j \) can be written in Eq. (3)

\[
D_{i,j}(X_i,x_j) = \sqrt{\sum_{r}(R_{O_r}(I_r) O_t(I_r))} \tag{3}
\]

which has the following properties when \( i, j, C \{1, 2, \ldots, N\} \).

\[
\text{Dist}(x_i, x_j) = 0 \text{ if } x_i = x_j \\
\text{Dist}(x_i, x_j) = \text{Dist}(x_j, x_i) \\
0 \leq \text{Dist}(x_i, x_j) \leq 1.
\]

If, \( ir = ij \), then the above eq. (4) is zero. By using the Eq. (3), the distance between data can be identified. The outliers between two distance data can be removed by using Eq. (1), which makes effective for clustering. The distance between two data can be reduced by using the proposed entropy method.
4. Experimental Analysis

In this section, the performance of EDBK is compared with existing methods such as RMC-MCN [18], MMNMF[19], DED, GAC, KM [20] in terms of F-measure, Precision, Recall, Normalized Mutual Information (NMI), accuracy and error rate. The below sections describes the dataset description, parameters evaluation and results of EDBK with existing methods.

Dataset description

Some experiments have been evaluated to examine the proposed EDBK method about its effectiveness. Twelve target sets of data in the field of clustering are tested in this study, all selected from consistent and real UCI data sets. In this study, it is tried to choose those data sets for analysis which included different attributes of problem space such as sample dimension, shared samples, sample size, borderline samples, feature diversity sample layout, the range of changes in various dimensions of samples, the number of classes and classes’ population. Table 1 describes the summary of datasets used in EDBK method.

Table 1: Summary of UCI dataset

| Data sets | Number of attributes | Number of clusters | Number of data objects |
|-----------|----------------------|--------------------|-----------------------|
| Iris      | 5                    | 3                  | 150                   |
| Wine      | 13                   | 3                  | 178                   |
| Heart     | 13                   | 2                  | 270                   |
| Robot     | 90                   | 5                  | 164                   |
| Breast    | 9                    | 2                  | 277                   |
| German    | 20                   | 2                  | 1000                  |
| Zoo       | 16                   | 7                  | 101                   |
| CMC       | 9                    | 3                  | 1473                  |
| WBC       | 9                    | 2                  | 683                   |
| E-coli    | 7                    | 8                  | 336                   |
| Glass     | 9                    | 6                  | 214                   |

Along with these dataset, the HW data set is consists of two thousand data points for zero to nine digit classes, and each class contains 200 data points. There are six types of characteristics are offered: Fourier coefficients of the character shapes, profile correlations, morphological features, Karhunen–coefficients pixel averages in 2 × 3 windows and Zernike moments. The parameter used to validate the efficiency of EDBK is described in below section.

Performance evaluation of Accuracy

In this section, the clustering accuracy of proposed EDBK is compared with existing methods such as DED, GAC, KM, MMNMF and RMC-MCN for some UCI datasets such as Iris, Wine, Breast, German, Zoo, Heart, Robot. Table 2 describes the clustering accuracy performance and Figure 2 describes the graphical representation of accuracy performance.

| Data sets | DED (%) | GAC (%) | KM (%) | MMNMF (%) | RMC-MCN (%) | EDBK (%) |
|-----------|---------|---------|--------|-----------|-------------|----------|
| Iris      | 80.23   | 89.9    | 79.3   | 59.10     | 88.00       | 90.56    |
| Wine      | 75.23   | 42.30   | 69.10  | 70.20     | 71.10       | 84.94    |
| Breast    | 70.7    | 41.8    | 48.97  | 60.65     | 41.88       | 76.21    |
| German    | 72.6    | 58.69   | 59.7   | 59.70     | 58.64       | 79.45    |
| Zoo       | 86.14   | 76.19   | 70     | 79.21     | 76.19       | 91.75    |
| Heart     | 59.20   | 59.30   | 14.30  | 51.50     | 51.44       | 67.47    |
| Robot     | 29.40   | 43.60   | 33.40  | 39.70     | 54.50       | 76.23    |
From the Table 2, the experimental results clearly showed that the EDBK achieved high clustering accuracy in all datasets which are used for evaluation. For instance, the EDBK achieved nearly 91% accuracy when compared with all other existing systems, but the proposed method achieved very low accuracy in Heart dataset due to complexity of data processing.

Figure 2: Clustering Accuracy performance of EDBK

The existing method DED achieved nearly 30% accuracy in Robot dataset, but the ECBK achieved nearly 77% accuracy because the outliers are removed in this dataset by using entropic distance.

**Performance evaluation of Error Rate**

The proposed EDK is validated by using the parameter error rate with existing techniques such as ABC, K-means and Hd-ABC. Table 3 shows the error rate value for EDBK and figure 3 represents the graphical representation for error rate metrics. The experiments for error rate can be carried out on six UCI datasets such as Iris, Wine, CMC, WBC, E-coli and glass.

**Table 3: Performance of Error Rate for EDBK**

| Data sets | K-Means | ABC | Hd-ABC | EDBK |
|-----------|---------|-----|--------|------|
| Iris      | 11      | 5   | 1      | 0.9  |
| Wine      | 11      | 8   | 1      | 0.78 |
| CMC       | 10      | 7   | 3      | 2.3  |
| WBC       | 9.2     | 5   | 1      | 0.69 |
| E-coli    | 6.48    | 4.6 | 2.5    | 1.43 |
| Glass     | 9       | 6.4 | 2      | 1.56 |
When compared with all other existing techniques, the proposed EDBK method achieved very low error rate for all UCI datasets. The existing method K-means obtained very high error rate for some datasets such as iris, wine and CMC, but obtained low error rate for E-coli dataset. The Hd-ABC method achieved very low error rate when compared with other two existing techniques, but very high error rate while comparing with EDBK. The EDBK achieved nearly 0.6% error rate in WBC and 2.3% error rate in CMC. The data used in CMC having many outliers which needs further improvement for EDBK to reduce the error rate.

4.6. Performance evaluation of NMI and F-measure

In this section, the performance of EDBK is validated by using NMI and F-measure for HW datasets. The existing methods such as MMNMF and RMC-MCN are used in this experiments for validating the EDBK performance in terms of NMI and F-measure.

Table 4: NMI and F-Measure Performance of EDBK

| Methods  | HW Dataset | F-Measure | NMI  |
|----------|------------|-----------|------|
| MMNMF    |            | 70.68     | 74.31|
| RMC-MCN  |            | 71.96     | 78.24|
| EDBK     |            | 75.82     | 84.56|

Table 4 and Figure 4 represents the experimental results of EDBK for HW dataset.
From the above results, it is clearly shows that the EDBK method achieved high NMI and F-measure in HW dataset when compared with existing methods. In this dataset, the proposed and existing methods are achieved nearly high values in both parameters, because the existing method also used the data by removing the outliers. But, the existing methods concentrates on physical similarity measure distance and failed to focus on attribute distance calculation.

5. Conclusion

Earlier research techniques have exhibited that enhanced clustering accuracy can be obtained using integrated information i.e., the data in the hidden patterns can be better exploited by identifying the general latent structure. However, traditional clustering methods are normally sensitive to noises and outliers, which greatly impair the clustering performance in practical problems. To address this problem, design a EDBK from the aspect of information entropy. In the gap with an existing categorical data metrics, the proposed EDBK treats ordinal attributes and nominal attributes in a different way but unifies the concept of the distance, which eliminates data loss during the distance measurement. The K-means algorithm is used to cluster the data in an effective way without outliers and noises. in addition, the proposed EDBK metric is easy to use and non-parametric, which can be easily applied for the clustering analysis of various types of categorical data. Experiments have proven that the proposed EDBK metric outperforms its counterparts on UCI data set in terms of accuracy, f-measure, precision, recall, NMI and error rate. The EDBK achieved 92% accuracy, 84.56% NMI, 80.71% recall with 0.6% error rate on UCI datasets.

References

[1] C. Yin, S. Zhang, Z. Yin, and J. Wang, “Anomaly detection model based on data stream clustering”, Cluster Computing, pp. 1-10, 2017.
[2] Z. Yu, P. Luo, J. You, H. S. Wong, H. Leung, S. Wu, and G. Han, “Incremental semi-supervised clustering ensemble for high dimensional data clustering,” IEEE Transactions on Knowledge and Data Engineering, vol. 28, no. 3, pp. 701-714, 2016.
[3] Rathore, P., Kumar, D., Bezdek, J. C., Rajasegarar, S., & Palaniswami, M. (2019). A rapid hybrid clustering algorithm for large volumes of high dimensional data. IEEE Transactions on Knowledge and Data Engineering, 31(4), 641-654.
[4] Tien, Nguyen Dang. "Tune up fuzzy C-means for big data: some novel hybrid clustering algorithms based on initial selection and incremental clustering." International Journal of Fuzzy Systems 19.5 (2017): 1585-1602.
[5] Lahmar, I., Ayed, A. B., Halima, M. B., & Alimi, A. M. (2017, March). Cluster forest based fuzzy logic for massive data clustering. In Ninth International Conference on Machine Vision (ICMV 2016) International Society for Optics and Photonics, (Vol. 10341, p. 103412J).
[6] Chang, X., Wang, Q., Liu, Y., & Wang, Y. (2016). Sparse Regularization in Fuzzy-c-Means for High-Dimensional Data Clustering. IEEE transactions on cybernetics, 47(9), 2616-2627.

[7] Tsai, Chun-Wei, Shi-Jui Liu, and Yi-Chung Wang. "A parallel metaheuristic data clustering framework for cloud." Journal of Parallel and Distributed Computing 116 (2018): 39-49.

[8] J. Ren, and Youlong Yang. "Multitask possibilistic and fuzzy co-clustering algorithm for clustering data with multisource features." Neural Computing and Applications (2018): 1-20.

[9] Mr. Dharmesh Dhabiya, Mr. Rahul Sharma. (2012). Efficient Cluster Formation Protocol in WSN. International Journal of New Practices in Management and Engineering, 1(03), 08 - 17.

[10] de Gusmão, Renê Pereira, and Francisco de AT de Carvalho. "Clustering of multi-view relational data based on particle swarm optimization." Expert Systems with Applications 123 (2019): 34-53.

[11] Mr. Bhushan Bandre, Ms. Rashmi Khalakar. (2015). Impact of Data Mining Technique in Education Institutions. International Journal of New Practices in Management and Engineering, 4(02), 01 - 07.

[12] Bhatnagar, V., Kaur, S., Saxena, R., & Khanna, D. (2017). DASC: data aware algorithm for scalable clustering. Knowledge and Information Systems, 50(3), 851-881.

[13] Ramadas, Meera, Ajith Abraham, and Sushil Kumar. "FSDE-Forced Strategy Differential Evolution used for data clustering." Journal of King Saud University-Computer and Information Sciences (2016).

[14] Alsawitti, Mohammed, Mohanad Albughdadi, and Nor Ashidi Mat Isa. "Variance-based differential evolution algorithm with an optional crossover for data clustering." Applied Soft Computing (2019).

[15] Natarajan, B., Obaidat, M.S., Sadoun, B., Manoharan, R., Ramachandran, S. and Velusamy, N., 2020. New Clustering-Based Semantic Service Selection and User Preferential Model. IEEE Systems Journal. DOI: 10.1109/JSYST.2020.3025407.

[16] Nataraj, S.K., Al-Turjman, F., Adom, A.H., Sitharthan, R., Rajesh, M. and Kumar, R., 2020. Intelligent Robotic Chair with Thought Control and Communication Aid Using Higher Order Spectra Band Features. IEEE Sensors Journal, DOI: 10.1109/JSEN.2020.3020971.

[17] Babu, R.G., Obaidat, M.S., Amudha, V., Manoharan, R. and Sitharthan, R., 2020. Comparative analysis of distributive linear and non-linear optimised spectrum sensing clustering techniques in cognitive radio network systems. IET Networks, DOI: 10.1049/iet-net.2020.0122.

[18] Sitharthan, R., Yuvaraj, S., Padmanabhan, S., Holm-Nielsen, J.B., Sujith, M., Rajesh, M., Prabaharan, N. and Vengatesan, K., 2021. Piezoelectric energy harvester converting wind aerodynamic energy into electrical energy for microelectronic application. IET Renewable Power Generation, DOI: 10.1049/rpg2.12119.

[19] Sitharthan, R., Sujatha Krishnamoorthy, Padmanaban Sanjeevikumar, Jens Bo Holm-Nielsen, R. Raja Singh, and M. Rajesh. "Torque ripple minimization of PMSM using an adaptive Elman neural network-controlled feedback linearization-based direct torque control strategy." International Transactions on Electrical Energy Systems 31, no. 1 (2021): e12685. DOI: 10.1002/2050-7038.12685.

[20] Jiang, Xingpeng, Xiaohua Hu, and Tingting He. "Identification of the clustering structure in microbiome data by density clustering on the Manhattan distance." Science China Information Sciences 59.7 (2016): 070104.