Distance-based K-Means Clustering Algorithm for Anomaly Detection in Categorical Datasets

Noor Basha
Research Scholar, DBIT, Bangalore

Ashok Kumar P.S.
Professor, DBIT, Bangalore

ABSTRACT
Real-world data sets also provide knowledge in an unsupervised manner with distinct and complementary aspects. In the field of cluster analysis, a number of algorithms have recently arisen. A priori, it is difficult for a user to determine which algorithm will be most suitable for a given dataset. For this job, algorithms based on graphs give good results. Such algorithms are however, vulnerable to outliers and noises with minimal edge information found in the tree to split a dataset. Thus, in several fields, the need for better clustering algorithms increases and for this reason utilizing robust and dynamic algorithms to improve and simplify the whole 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 remove 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 conducted using UCI datasets showed that EDBK method outperforms the existing methods such as Artificial Bee Colony (ABC), k-means etc. The EDBK achieved 80.71% recall, 79.81% precision and 75.82% F-measure. The results show that the EDBK method not only improve the clustering accuracy (i.e. nearly 92%), but also greatly reduce the interference of outliers to clustering results.

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

1. INTRODUCTION
"Clustering is an unsupervised learning method which is widely used in different branches of science such as image processing, pattern recognition, information retrieval and text processing" [1]. "During clustering, data with the most similarity are placed in a cluster; as a result the data in different clusters have the minimum similarity to each other. In recent years, clustering process face the high-dimensional data problems i.e., the objects to be clustered have a large number of features. "In most of the real-world cases, only a small portion of the features is assumed to be relevant to the cluster structure. A good clustering approach should be able to identify the relevant features and avoid the negative influences of the noisy features" [5]. "The existing clustering methods can be divided into two categories such as partitional and hierarchical clustering. In hierarchical clustering, a set of nested clusters will be seen in the form of hierarchical trees at the end of clustering process" [7]. "In partitional clustering, dataset is divided into non-overlapping subsets in a way that each data item will be exactly in one subset" [8]. "K-means is a partitional clustering algorithm that has become one of the most well-known and popular partitional clustering methods due to its simplicity, efficiency and linear computational time" [9].

"Although several clustering methods were reported with good performance, they suffer from the drawbacks such as traditional clustering methods are usually sensitive to noises and outliers, which greatly impair the clustering performance in practical problems" [10]. The existing methods depends on Nonnegative Matrix Factorization (NMF) cannot be utilized to tackle the data matrix with mixed signs. For graph based methods, application of different kernels to build the graph will greatly affect the clustering performance. Furthermore, existing methods focused solely on the similarity distance between data points, resulting in poor clustering accuracy. Hence, this research paper aims to introduce a new algorithm to determine the noises and outliers using the entropic distance based K-means clustering algorithm for categorical dataset.

2. LITERATURE REVIEW
Several techniques are suggested by researchers in the data clustering.

Noor Basha et al., [2] presented dimensionality reduction technique using fast clustering-based feature selection algorithm initially divides the data features into clusters by using graph-theoretic clustering method. For choice calculation in subsection of structures, where a novel grouping approach is initiated to reduction in dimensionality of structures, it eliminating immaterial structures.

Y. Wang, *et al.*, [3] "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. The SSC method was applied in various domains such as movie recommendation, film evaluation and able to guide theater exhibitors and distribution. Some clusters of SSC provides low score with the time series data because of its rapid decay pattern".

D. Bacciu, and Daniele Castellana, [4] "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 were required by the first form of mixture i.e. finite mixture (MIX-SP-BHTMM), which was used as a hyper-parameter for Expectation Maximization approach (EM). The problem of components specifications was addressed to allow an infinite number of components within the model was known as second form as infinite-mixture model (INF-SP-BHTMM). The numbers of clusters were directly learned from the data in INF-SP-BHTMM which was considered as advantage. The SP-BHTMM is unable to use an infinite number of hidden states which leads to incomplete non-parametric model".

Noor Basha et al., [5] made use of KNN, logistic regression

and decision tree algorithm are used to calculate accuracy and performance for heart/chronic disease. Machine learning algorithm provides very valuable knowledge on analysis and prediction of many chronic diseases

S. Huang, *et al.,* [6] "proposed a novel robust multi-view clustering method (RMC) to integrate heterogeneous representations of data. The multi-capped-norm (MCN) was designed to remove the noises and outliers, especially the extreme data outliers. The RMC-MCN was a low complexity algorithm because of using classical K-means algorithm. To solve the multi-view clustering problems, the RMC-MCN designed an effective optimization algorithm. The experiments were conducted on three datasets such as Handwritten numerals (HW) which was selected from standard UCI datasets, Caltech101 and BBCS port to validate the performance of RMC-MCN method when compared with existing method like multi-manifold regularized non-negative matrix factorization framework (MMNMF)" [7]. "The method focused on similarity distance measures which leads poor performance on clustering accuracy because it focused on only physical characteristics of data and suffered from time complexity".

Noor Basha et al., [8] made use of many classification algorithms to predict the severe heart syndromes based on risk rate, where the author specifically used machine learning approach.

J. Ma,*et al.,* [9] "novel two-phase clustering algorithm with a density exploring distance (DED) measure. In this phase, the fast global K-means (KM) clustering algorithm was used to obtain the cluster number and the prototypes. Afterwards, all the prototypes were clustered according to a DED measure which made data points locating in the same structure to possess high similarity with each other. The results demonstrated that the DED algorithm was flexible to different data distributions and has a stronger ability when compared with the genetic algorithm-based clustering (GAC) algorithm. This procedure greatly speeds up the whole algorithm and maintained the accuracy of clustering which saved a lot of time simultaneously. But, the DED algorithm did not remove the outliers and noises from the UCI datasets".

F. Zabihi, and Babak Nasiri, [10] "implemented an History-driven Artificial Bee Colony (Hd-ABC) to improve the ABC's performance by applying a memory mechanism. The Hd-ABC utilized a binary space partitioning (BSP) tree to memorize useful information of the evaluated solutions. Fitness evaluation was a costly and time consuming process in clustering problem, but utilizing the memory mechanism had decreased the number of fitness evaluations significantly. The Hd-ABC algorithm had been applied on nine UCI datasets and two artificial datasets which showed that the Hd-ABC algorithm outperforms the original ABC and its variants. The method cannot able to select the relevant features which leads poor performance on clustering the data".

To overcome the above mentioned issues, this research work focuses on EDBK which remove the outliers and noises before clustering the data by using entropic distance.

3. PROPOSED DESIGN

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 is high which leads to outlier presence and poor accuracy. To overcome the above mentioned issues, the proposed EDBK to exploit embedded order data or the calculation of data distance. Then, based on the entropy distance, the outliers are removed. After removing the outliers, the data are clustered by using K-means algorithm.

The data are collected from UCI dataset which is given as input for normalization process. In this process, two key classes, i.e., categorical and numerical attributes, form the types of data attributes. There are also two subclasses under the categorical class i.e., nominal attributes and ordinal attributes, where ordinal attributes inherit some nominal attribute properties. On the one side, including nominal attributes, both qualitative and unsuitable for arithmetic operations are the types of attributes in ordinal data, i.e. the data correlated with ordinal attributes, for arithmetic operations such as mean, division and summation of the data associated with the ordinal attributes only are all qualitative and unsuitable.

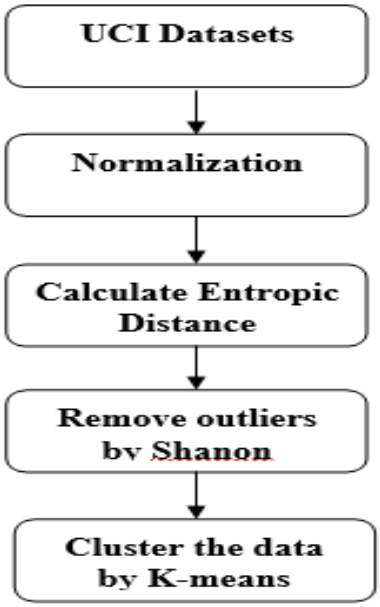

Figure 1: Basic Structure of EDBK method

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.

3.1. Entropy-Based Distance Metric

In this paper, for a data set $X = \{x_1, x_2, \ldots, x_N\}$ with N data objects represented by d attributes. From the viewpoint of data objects, O_r is the value space of the rth value of a data object. The vital problem in data distance measurement is how to measure the distance contributions of different categories. From the perspective of information theory, a higher entropy value usually indicates a larger amount of information or more uncertainty. A choice with more information or higher uncertainty level usually costs more thinking for a participant. Therefore, the entropy value of a category is suitable for indicating its distance contribution.

Therefore, the distance between the rth 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),

$$E_{O_{r(s)}} = -p_{O_{r(s)}} log\, p_{O_{r(s)}} \quad (1)$$

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

$$p_{O_{r(s)}} = \frac{\sigma_{O_{r(s)}}}{N} \quad (2)$$

where $\sigma_{O_{r(s)}}$ is the number of data objects in the data set X with their rth values equal to $O_{r(s)}$. Subsequently, the distance between two ordinal data objects x_i andx_j can be written in Eq. (3)

$$Dist(x_i, x_j) = \sqrt{\sum_{r=1}^{d} \vartheta\left(O_r(i_r), O_r(j_r)\right)^2} \quad (3)$$

which has the following properties when $i, j, \in \{1, 2, \ldots, N\}$.

i. $Dist(x_i, x_j) = 0$ if $x_i = x_j$

ii. $Dist(x_i, x_j) = Dist(x_j, x_i)$

iii. $0 \leq Dist(x_i, x_j) \leq 1$.

Where, $\vartheta(O_r(i_r), O_r(j_r))$ is defined in Eq. (4),

$$\vartheta(O_r(i_r), O_r(j_r)) = \sum_{s=min(i_r, j_r)}^{max(i_r, j_r)} E_{O_{r(s)}},$$

if $i_r \neq j_r$ (4)

If, $i_r = i_j$, 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. Then, these data can be clustered by using K-means algorithm which is described as below.

3.2. K-means Algorithm

The k-means the algorithm determines based on the number and size of data, the optimum number of desired clusters. Suppose that N is the total number of uniformly distributed data nodes in a M*M Square area. The optimum number of clusters k_{opt} can be obtained in Eq. (5)

$$k_{opt} = \frac{\sqrt{N}}{\sqrt{2\pi}} \sqrt{\frac{\varepsilon_{fs} M}{\varepsilon_{mp} d_{x_i, x_j}{}^2}} \quad (5)$$

Where, $d_{x_i.x_j}$ is the distance from data x_i and x_j, ε_{f_s} is the parameter for free space model and ε_{mp} is the parameter for multipath model. The basic idea of K-means clustering algorithm is to classify a given set of data items into k number of disjoint clusters where the value of k is predefined in Eq. (5). Algorithm 1 describes the basic K-means algorithm which mainly consists of two phases as follows.

Algorithm 1: The K-means clustering algorithm

Input:

D = set of n data items

k = number of desired clusters

Output:

A set of k clusters.

Steps:

1. Randomly produce predefined value of k centroids
2. Allocate each object to the closest centroids
3. Recalculate the positions of the k centroids, when all objects have been assigned.
4. Repeat steps 2 and 3 until the sum of distances between the data objects and their corresponding centroid is minimized.

From the above algorithm, the data can be clustered without outliers in an effective way. In next section, the effectiveness of EDBK can be validated by using various experiment analysis on UCI dataset can be described.

4. RESULTS AND DISCUSSIONS

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.

4.1. Dataset Description

In order to analyze the proposed EDBK approach in terms of its efficacy, such experiments and comparisons are performed. This study tested twelve benchmark data sets in the clustering area, all selected from standardized and real UCI data sets. In this report, it was attempted to select certain research data sets that included different characteristics of problem space such as sample dimension, feature diversity, borderline samples, shared samples, sample size, 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

Datasets	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 comprised of 2000 data points for 0 to 9 digit classes, and each class has 200 data points. There are six types of features are available: profile correlations, Fourier coefficients of the character shapes, Karhunen–Love coefficients, morphological features, pixel averages in 2 ×3 windows and Zernike moments. The parameters used to validate the efficiency of EDBK is described in below section.

4.2. Evaluation of Parameters

The set of parameters such as error rate, cluster accuracy, precision, recall and NMI are used to test the effectiveness of EDBK method which is described in this section.

Error Rate: The clustering error rate (ER) is often used to measure the efficacy of the proposed EDBK, which refers to

the percentage of data items that have been misplaced. For selected UCI datasets, the number of clusters and the type of cluster allocated to any given sample are calculated., the EDBK can calculate ER using the following equation (6),

$$ER = \frac{No.of\ misplaced\ data\ objects}{Total\ size\ of\ data\ objects} \quad (6)$$

Accuracy: By matching of data point to the correct cluster, the accuracy (ACC) test measures the largest rate of proper allocation. Denote l_i as the label of an algorithm and T_i as the true label of x_i. ACC is defined in Eq. (7):

$$ACC = \frac{\sum_{i=1}^{n} \delta(T_i, map(l_i))}{n} \quad (7)$$

Where, $\delta(x, y)$ is the indicator function, and n is the total number of data points, $map(l_i)$ is the mapping function that permutes the clustering labels to match the ground truth labels.

Normalized mutual information: The NMI shows the statistics exchanged between the algorithm label and the true label. Given the true labels$\Delta = \{C_1, C_2, \ldots, C_c\}$ and the clustering results $\Delta' = \{C'_1, C'_2, \ldots, C'_k\}$ of $X(|X| = n)$, let n_i and n'_i be the number of data points in cluster C_i and C'_iseparately. Let n_{st} denote the number of data points that are in cluster C_s as well as in cluster C'_t, then the NMI of Δ and Δ' is given in Eq. (8)

$$NMI = \frac{\sum_{s=1}^{c}\sum_{t=1}^{k} \log\left(\frac{n n_{st}}{n_s n'_t}\right)}{\sqrt{\left(\sum_{s=1}^{c} n_s \log\frac{n_s}{n}\right)\left(\sum_{t=1}^{k} n'_t \log\frac{n'_t}{n}\right)}} \quad (8)$$

Precision: Positive predictive value, also known as accuracy, is defined as the measurement of true positive data for both false and positive evaluation quantities. The mathematical equation for this metric is represented in Eq. (9).

$$Precision = \frac{TP}{(TP+FN)} \quad (9)$$

Where, TP is true Positive and FN is False Negative.

Recall: The sensitivity measures the ratio of clusters that samples identify correctly. The mathematical equation of sensitivity is described in Eq. (10).

$$Recall = \frac{TN}{(TN+FP)} \quad (10)$$

Where, TN is True Negative and FP is False Positive.

F-Measure: F-measure is the measure of accuracy test and it considers the both precision PP and recall RR of the test in order to calculate the score. The general formula for F-measure is given in the Equation (11).

$$F - measure = 2.\frac{Precision.Recall}{Precision+Recall} \times 100 \quad (11)$$

The next section will describe the performance analysis of EDBK in terms of various parameters for different datasets with graphical representation.

4.3. Parameters selection

In this section, this research paper will report the results of the experiment under various parameter settings. The EDBK method explores the sensitivity with respect to λ parameters and the ratio of outliers. In fact, since unlike synthetic data, some of the real world datasets are manually collected, these datasets which contain noise and outliers to some degree because of the conditions under which the datasets are captured. That is, there may exist some entries of the data corrupted arbitrarily, and the corruption is usually sparse. Thus, the EDBK can assume that the ratio of outliers is small, and can reduce the influence of noisy data and outliers by utilizing the entropy distance. Here, the ratio of outliers is set by search the grid $\{0, 0.025, 0.05, \ldots, 0.2\}$and also search the logarithm of the parameter λ, i.e., we varied the value of $log_{10}\lambda$ in steps of 0.2 from 0.1 to 2. When the value of one parameter varies, then keep another parameter fixed at the optimal value.

4.4. 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.

Table 2 describes the clustering accuracy performance

Data sets	Clustering Accuracy (%)					
	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

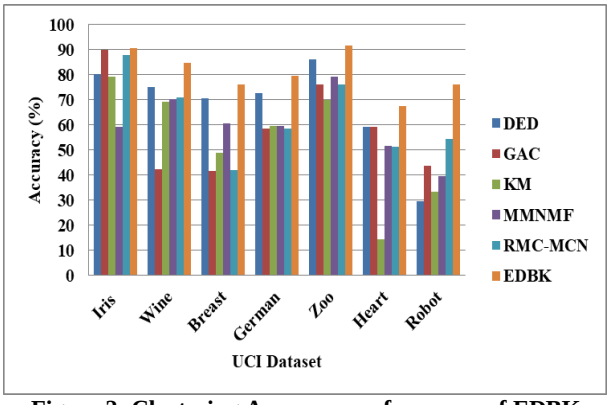

Figure 2: Clustering Accuracy performance of EDBK

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.

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.

4.5. Performance evaluation of Error Rate

The proposed EDK is validated by using the parameter error rate with existing techniques such as ABC [21], K-means [21] and Hd-ABC [21]. 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

Datasets	Error Rate			
	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

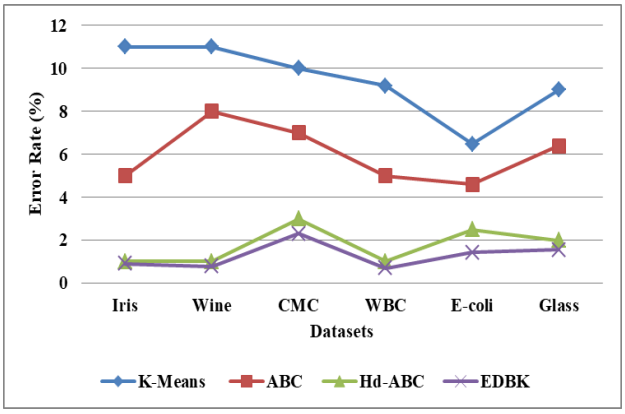

Figure 3: Error Rate of EDBK

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 experiment for validating the EDBK performance in terms of NMI and F-measure. Table 4 and Figure 4 represents the experimental results of EDBK for HW dataset.

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**

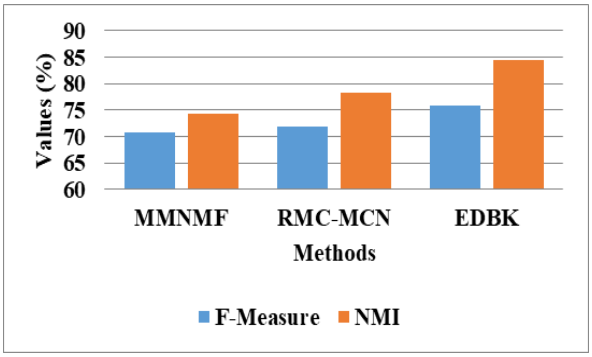

Figure 4: EDBK performance in NMI and F-Measure for HW dataset

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. The EDBK method removed the outliers by entropic distance which is based on attribute characteristics of data. The EDBK method obtained 75.82% F-Measure and 84.56% NMI in HW dataset.

4.7. Performance evaluation of Precision and Recall

The EDBK method undergoes an experiment on HW dataset for validating the precision and recall when compared with existing methods such as MMNMF and RMC-MCN. Table 5 and Figure 5 represent the performance of EDBK which is given below.

Table 5: Performance of EDBK in HW dataset

Methods	HW Dataset	
	Precision	Recall
MMNMF	69.57	71.83
RMC-MCN	70.03	7.57
EDBK	**79.81**	**80.71**

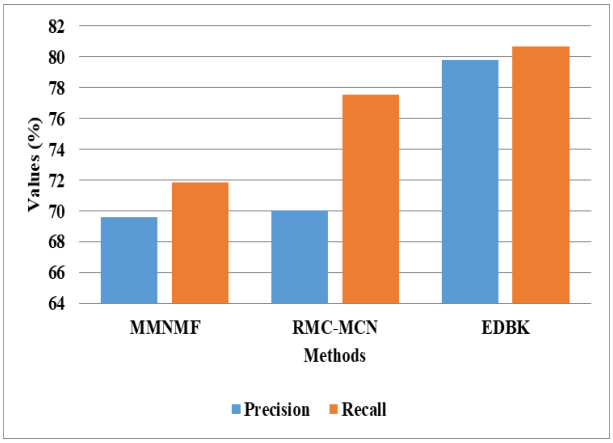

Figure 5: EDBK performance in terms of Precision and Recall

The EDBK method achieved nearly 80% precision and 81% recall in HW dataset, whereas the RMC-MCN obtained 70.03% precision and 77.57% recall. When compared with all other methods, MMNMF method achieved very low precision (i.e. 69.57%) and low recalls (71.83%). The proposed method is a little sensitive with respect to the ratio of outliers. It is worth mentioning that the clustering performance of data set HW seems to be more sensitive with respect to the ratio of outliers than other data sets. This may be because there exists some noises and extreme data outliers within the data set. From the above all experiments on various parameters for different datasets, the proposed EDBK achieved higher clustering accuracy, NMI, precision and recall with low error rate values.

5. CONCLUSION

Previous studies have shown that it is possible to achieve better clustering precision using integrated knowledge, i.e. by finding the common latent structure; the hidden patterns in data can be better explored. Traditional clustering strategies however are typically susceptible to noises and outliers, which significantly impair the efficiency of clustering in realistic problems. In addition, due to the kernel/affinity matrix construction or own decomposition, current clustering methods, e.g. graph-based methods, are of high computational complexity. Designing an EDBK from the perspective of information entropy to resolve this problem. The proposed EDBK handles ordinal attributes and nominal attributes differently, in contrast to the current categorical data metrics, but unifies the definition of distance, preventing information loss during distance measurement. The K-means algorithm is used to cluster the data in an effective way without outliers and noises. Moreover, the proposed EDBK metric is easy to use and non-parametric, which can be easily applied for the clustering analysis of different types of categorical data. Experiments have shown 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. In future work, the categories from different attributes should be improved, because the attributes do not have a unified scale. In addition, the proposed EDBK is further improved for larger number of categories because the larger distance values make poor measured distance of attributes.

6. 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] Noor Basha, PS Ashokkumar, P Venkatesh " Reduction of Dimensionality in Structured Data Sets on Clustering Efficiency in Data Mining " IEEE International Conference on Computational Intelligence and Computing Research (ICCICI), pages 1-4.

[3] Y. Wang, Y. Ru, and J. Chai. "Time series clustering based on sparse subspace clustering algorithm and its application to daily box-office data analysis." *Neural Computing and Applications*,pp. 1-10, 2018.

[4] D. Bacciu, and Daniele Castellana. "Bayesian mixtures of Hidden Tree Markov Models for structured data clustering." *Neurocomputing*, 2019.

[5] Noor Basha, K Manjunath, Mohan Kumar Naik, PS Ashok Kumar "Analysis and Forecast of Heart Syndrome by Intelligent Retrieval Approach" Intelligent Computing and Innovation on Data Science, Springer, Singapore, pages 507-515.

[6] S. Huang, Yazhou Ren, and Zenglin Xu. "Robust multi-view data clustering with multi-view capped-norm k-means." *Neurocomputing* 311 (2018): 197-208.

[7] L. Zong, X. Zhang, L. Zhao, H. Yu, and Q. Zhao, "Multi-view clustering via multi-manifold regularized non-negative matrix factorization,"*Neural Networks*, 88, 74-89, 2017.

[8] Noor Basha, Ashok Kumar P.S, P Venkatesh, "Early Detection of Heart Syndrome Using Machine Learning Technique," 4th International Conference on Electrical, Electronics, Communication, Computer Technologies and Optimization Techniques (ICEECCOT) Pages 387-391.

[9] J. Ma, Xiangming Jiang, and Maoguo Gong. "Two-phase clustering algorithm with density exploring distance measure." *CAAI Transactions on Intelligence Technology* 3.1 (2018): 59-64.

[10] F. Zabihi, and Babak Nasiri. "A Novel History-driven Artificial Bee Colony Algorithm for Data Clustering." *Applied Soft Computing*vol. 71, pp. 226-241, 2018.