Text Clustering using Ensemble Clustering Technique

Muhammad Mateen[1], Junhao Wen[2], Sun Song[4]
School of Big Data and Software Engineering,
Chongqing University,
Chongqing, 401331, China

Mehdi Hassan[3]
Department of Computer Science,
Air University,
Islamabad, 44000, Pakistan

***Abstract*—Clustering is being used in different fields of research, including data mining, taxonomy, document retrieval, image segmentation, pattern classification. Text clustering is a technique through which text/ documents are divided into a particular number of groups, so that text within each group is related in contents. In this paper, the idea of ensemble text clustering of majority voting is defined. For this purpose, different clustering methods such as fuzzy c-means, k-means, agglomerative, Gustafson Kessel and k-medoid are used. After performing the pre-processing of the documents, inverse document frequency (IDF) has been achieved by the provided dataset. The achieved IDF is considered as input to the clustering algorithms. Dunn Index and Davies Bouldin Index have been calculated which are applied to analyze the usefulness of the proposed ensemble clustering. In this work, a dataset "Textclus" which contains four different classes, history, education, politician and art as a text is applied. Additionally, another dataset "20newsgroups" is also applied for analysis. The clustering quality measures have also been calculated from the proposed ensemble clustering results. The attained results show that the proposed ensemble clustering outperforms the other state of the art clustering techniques.**

Keywords—Agglomerative; document clustering; ensemble clustering; gustafson kessel; inverse documents frequency; text clustering

I. Introduction

Clustering is mostly used in the area of pattern recognition and information retrieval. Text clustering is a technique through which text/ documents are divided into a particular number of groups, so that text within each group is related in contents [1]. The goal of text clustering us to make a set containing relative data objects in a particular way like kind of text, a group of text, etc.

In the text clustering unsupervised technique of data is needed for clustering. Usually, text clustering techniques use characteristics like sequences, words, phrases from the documents to apply the clustering [2]. Text clustering is an interesting and advance research area because the availability of a huge amount of information in electronic forms [3]. A lot of several applications have been designed in literature which are applied for document clustering [4]. Several techniques have been used in text/document clustering are given as under:

- Frequent pattern-based clustering
- Constraint based clustering
- Partitioning
- Grid-based
- Model-based
- Hierarchical
- Density-based
- Fuzzy clustering

Currently, text clustering is widely used to solve the problem raised by the area of the database, the main purpose of this research is to scalable clustering of a heterogeneous and multi-dimensional type of data [5]. To cluster out the data, there are various clustering techniques. Hard and soft clustering methods are used for document clustering according to their attributes [6]. There is a close relationship among clustering methods and several fields including heterogeneous data analysis, web application, and DNA analysis in computational biology, white blood cells and red blood cells clustering [7]. Hisham Al-Mubaid et al. [8] proposed a new clustering technique based on a successful feature selection method and logic based learning method Lsquare. Yanjun Li et al. defined a feature selection technique and after that evaluated with K-means clustering algorithm for different standard datasets. Their results showed that TCFS with choice of internal representations (CHIR) has better accuracy of clustering in the form of purity and f-measure [9].

The rest of the paper is managed as; related work of study is discussed in section II, section III is based on proposed technique and in section IV, the experimental results with the help of graphical representation are discussed. In section V, finally the research outcome is concluded.

II. Related Work

A lot of research work has been dedicated to improve and develop the text clustering algorithms. In [10] researchers proposed a novel clustering technique which is based on mode seeking without any parameter. In the mean algorithm, mode seeking technique is especially used for clustering the data; it is also helpful and makes the system efficient when the parameters are not defined in a proper way. In this work, the researcher proposed ensemble clustering contains the repetition of the recent mode seeking technique of kNN. Mode seeking algorithm is more convenient and faster than the regular mean shift, for the high level of dimensional data. Mode seeking algorithm is also famous about the robustness about parameter selection and high level of dimensional data. The proposed method is achieved with the help of the consensus algorithm. Consensus algorithm process is based on two main steps. (1)

Randomly initialize the subsets with different parameters to run the multiple clustering and this session is considered as kNN ensemble mode seeking (2) merge all the results to calculate the consensus of whole repeated clustering.

The complex structure of the dataset includes data dimensionality and distribution managed by clustering algorithms. In [11] authors proposed a new data clustering technique that is based on the k-nearest neighbor chain (KNNC) using heuristic rules. With the inspiration of the PageRank, the algorithm Researcher used random walk design to calculate the value of data points. After that, based on valuable data points, researcher designed a KNNC to arrange the K-nearest neighbors with respect to distance and defined two heuristic approaches to obtain the suitable number of groups having same characteristics of objects and also initial clusters. In the first rule of heuristic approach is the distance of K-nearest neighbor chain which demonstrates the level of separation of groups with convex curves and secondly, the distance of internal compactness of a group to the nearest neighbor of KNNC. The proposed clustering technique achieved better performance than famous clustering algorithms.

Text mining is widely used in the field of social networks, opinion mining considered emotion analysis. It plays an effective role in the evaluation of emotions and opinion in texts. Opinion methodology normally depends on an expression lexicon, which is a group of predefined keywords that define emotions. Opinion mining gather required emotional words defined in advance and also have some complexity to classify the sentences that involve a judgment without the use of any emotional keywords. In [12] the researcher proposes a novel emotion analysis technique, based on hidden Markov design with text data. In this research, there is a text-based representation of emotions via ensemble TextHMMs. It explains hidden variables with the help of semantic group information. For the reflection of diverse models, an ensemble technique is applied that is based on TextHMMs classifier.

Consensus clustering is increasing attention to the heterogeneous data analysis. The co-association technique is used to define the consensus clustering as a graph partitioning problem. In [13] proposed spectral ensemble clustering (SEC) to influence the benefit of co-association for information integration but works in an efficient way. Algorithmic complexity exponentially reduced with the combination of SEC and weighted K-means clustering. The proposed big data grouping technique is introduced to overcome the challenges based on incomplete partitions with the extension of SEC. Multi-view and ensemble clustering techniques demonstrate the advantage of SEC with other big data clustering techniques.

In recent years, [14] ensemble clustering has become an attractive method for robust clustering. In existing ensemble clustering techniques there is a limitation that all the clusterings are regarded as the same as dependability, which caused as vulnerable to the low-level base clusterings. To globally analyze and weight the clusterings, there are some tasks has been performed which are also used to neglect the diversity of groups inside the base clustering. There is a common issue about to analyze the reliability of clusterings and utilize the neighbor diversity in the ensemble technique to improve the consensus quality, particularly, in the scenario, when there is no chance to obtain the data features or precise prediction about data distribution.

To overcome this problem [15] proposes a new ensemble clustering technique on the basis of ensemble-driven clustering estimation and locally weighted strategy. Particularly, the doubtfulness of every cluster is predicted by considering the clustering labels in the internal ensemble through an entropic principle. A new approach named as ensemble-driven clustering is explained and locally weight co-association matrix has defined the conclusion of an ensemble of diverse groups.

In the last few years, fuzzy c-means cluster ensemble and random projection approaches have been designed for data clustering with high dimensionality. Random projection is widely used to reduce the dimensionality because of its efficiency and simplicity. A large amount of space is required to store the huge affinity matrix, and earn large computational time to cluster out the affinity matrix.

In [15], to reduce the dimensionality of huge data, the researcher designed a framework with the combination of fuzzy c-means and random projection based on cluster ensemble. The framework uses the collective agreement to cumulative fuzzy partitions. With the use of internal and external cluster indices, the fuzzy partitions are ranked with random projections. The most excellent partition in the levels of the queue is the core partition achieved by cluster ensemble.

In [16], tree-ensemble clustering technique is introduced for the analysis of static CRAFTER, datasets to handle the high dimensionality. CRAFTER is useful to tackle numerical and categorical features concurrently and analyze the size and dimensionality of datasets.

CRAFTER influences the features of a tree-ensemble to tackle high dimensionality and mixed attributes. The class probability estimation technique is used for the representation of data points of clustering. According to the limitations of the ensemble clustering, one of them is the independence of basic clusterings and ignorance between their relationships. On the other, there is a lack of information corporation between the local and global relationship of clusterings, especially, when renovating from one point to another of the similar matrix between basic clusterings. In [17] a new ensemble clustering technique is introduced as relative density path accumulation (MRDPA). In this technique, density nearest-neighbor and relative k-nearest neighbor are used to create basic clusterings. The clusterings represent multi-scale features for the input dataset of k into the RNKD. To investigate the global information in the creative k-nearest neighbor graph, and finally generated via consensus function.

The basic purpose of ensemble clustering is to combine the various essential parts into the consent. In the related work of different articles, it has been explained that to increase the numerous partitions caused a lower variance and better performance for ensemble clustering. In this scenario, for the given dataset, the best partition among the different partitions is still a challenging problem.

In [18], the researcher proposed a novel approach to solving this problem. The author introduced the infinite ensemble clustering (IEC) to drop out the noising data and represents the infinite partitions. In this technique, de-noising auto encoder is used to generate the prospective representation of infinite partitions. For resultant clustering, concatenation of deep features is applied to k-means.

In [19], researcher introduced a novel technique for ensemble (subspace) clustering with high dimensionality of text data. This technique implements the integration of two stages for feature representation of data including topics and words to make clusters. Ensemble clustering is effective to enhance the strength of clusters. This approach based on topic modeling to lead the two steps attributes the representation of data and make various ensemble components. With the use of both words and topics to cluster out the text data, significant clusters can be achieved with the weight of topics and words in every cluster.

III. Proposed Techniques

In the field of information retrieval, text clustering is an important area of research to categorize and understand the unstructured textual data. In this research, the ensemble clustering technique is investigated. The ensemble clustering is based on k-means, agglomerative, fuzzy c-means, k-medoid, and Gustafson Kessel clustering and has obtained different clustering results separately of a specific data; observed that all results were different from each other's [20]. These processes are used for the quality and performance of clustering algorithms, and these stages are necessary to complete the clustering algorithm [21]. The proposed ensemble clustering technique consists of four various stages and depicted in Figure 1.

1) In data collection phase some processing operations are applied at the given data include crawling, indexing, filtering etc. which are helpful in document clustering. It also indexes the documents to store the data and access in an efficient way and screen out the unnecessary data for example stop words.

2) In pre-processing phase some specific operations have been performed which are used to make the text into a meaningful format like, vector-model, graphical model.

3) Text clustering phase divides a set of text into a specified number of clusters, as each dataset specified number of clusters having different features.

4) In post-processing phase, core applications are included through which documents can be clustered.

To obtain accurate results, the ensemble clustering technique based on majority voting scheme is proposed. To ensure that ensemble clustering results are useful for text clustering. Dunn Index (DI) and Davies Bouldin Index (DBI) are evaluation parameters indicate that ensemble clustering technique offers superior results as compared to individual technique.

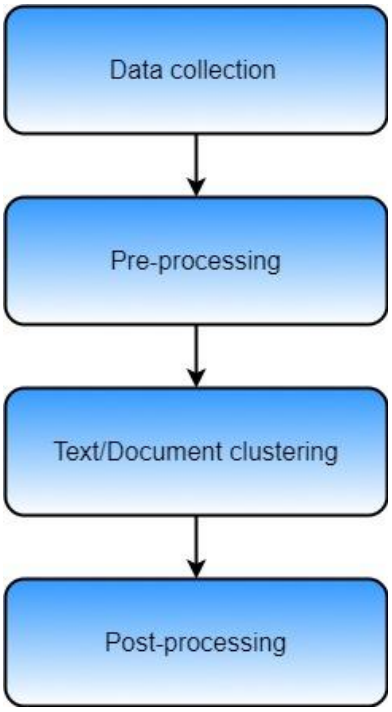

Fig. 1. Process of Clustering

The following clustering techniques have been used for text clustering. A brief description of each algorithm used in ensemble clustering is given as under:

A. K-mean Algorithm

K-means clustering is a very popular and widely used technique which is successfully being used in image segmentation, computer vision and object qualification etc [22].

Algorithm: K-means

1. Cluster centroids $\mu 1, \mu 2 \ldots \mu k \in \mathbb{R}n$ arbitrarily.
2. Recur till union {

   For each i, set

$$c^{(i)} := \arg\ \min ||x^{(i)} - \mu_j\ ||^2.$$

For every j, set

$$\mu_j := \frac{\sum_{i=1}^{m} 1\{c^{(i)} = j\} x^{(i)}}{\sum_{i=1}^{m} 1\{c^{(i)} = j\}}$$

}

B. Fuzzy C-Means Algorithm

Fuzzy c-means (FCM) is a technique of soft clustering in which a portion of data belongs to more than one clusters [23].

Algorithm: Fuzzy c-means

$Input: X, c, m$

$Output: U, V$

$Initialize\ V$

$While: max_1\ £\ \{||v_{k,new} - v_{k,old}||^2\} > \hat{I}\ do$

$$m_{ij} = \left[\sum_{k=1}^{c} \left(\frac{||x_j - v_i||}{||x_j - v_k||} \right)^{\frac{2}{m-1}} \right]^{-1} ,''i, j$$

$$v_i = \frac{\sum_{j=1}^{n} (m_{ij})^m x_j}{\sum_{j=1}^{n} (m_{ij})^m} ,''i$$

where U is the $(c \times n)$ partition matrix, $V = \{v_1, \ldots, v_c\}$ is the set of c cluster center in R^d, $m > 1$ is the fuzzification constant and $\|\cdot\|_A$ is an inner product A-induced norm. μ_{ij}, v_i are iterated until algorithm terminates.

C. Hierarchical Agglomerative Clustering Algorithm

Hierarchical clustering algorithms are either bottom-up or top-down approaches. Bottom-up algorithms are treated as a pair of clusters which are merged until all clusters are merged into a single cluster which contains all documents. Hierarchical agglomerative clustering is more commonly used in information retrieval than top-down clustering [24, 25].

$$K = arg_{K'} min[RSS(K') + lK] \quad (1)$$

Where K′ denotes cut of the hierarchy that results in K′ clusters, RSS represents the residual sum of squares and λ is a consequence for each additional cluster. Another measure of distortion can be used instead of RSS.

D. K-medoids Algorithm

K-medoid comes in an algorithm related to the k-means algorithm, except when fitting the centers *C1,…Ck*, restrict the attention to the points themselves [25].

The initial guess for centers *C1,… Ck* (e.g., randomly select *K* of the points *X1,…Xn*), then repeat:

1) Minimize over C: for each i=1,…n, find the cluster center ck closest to *Xi* and let *C(i)=K*

2) Minimize over *c1,…ck*: for each *k=1,… K* let *Ck=Xk*, the medoid of points in cluster k, i.e., the point Xi in cluster k that minimizes

$$\sum c(j) = k \| X_j - X_i \|_2^2$$

stop when within-cluster variation doesn't change

In words:

1) Cluster (label) each point based on the closest center

Replace each center by the medoid of points in its cluster

E. Gustafson-Kessel Algorithm

The Gustafson-Kessel (GK) algorithm is a dominant clustering method with numerous applications in different domains including, classification, system identification, and image processing. The Gustafson-Kessel is a technique that is used to combine each cluster with both a matrix and a point, correspondingly shows the cluster center and its covariance. But the fuzzy c-means make the inherent assumption that clusters are spherical, the Gustafson- Kessel algorithm is not subject to this restriction and can categorize ellipsoidal clusters [26].

The Gustafson–Kessel algorithm is based on iterative optimization of an objective function of the c-means type:

$$J_m = \sum_{j=1}^{n} \sum_{i=1}^{k} \mu_{ij}{}^m d_{ij}^2$$

Where J_m defines within group sum of squared errors.

F. Ensemble Clustering

Ensemble clustering combines a set of clustering from the same data set and generates a final clustering. The goal of ensemble clustering is to improve the quality of individual data clustering. In this paper, majority voting based ensemble clustering is applied. For this purpose, five clustering techniques are employed which are already defined on specific data sets to obtain individual results. After implementation, the ensemble clustering results are compared with individual techniques. It has been observed from ensemble clustering results that it generates better clustering results.

IV. Experimental Results And Discussion

In this paper, five various clustering algorithms and two datasets are used to obtain the ensemble clustering, but here k-means, agglomerative, fuzzy c-means, k-medoid and Gustafson Kessel clustering techniques. The ensemble clustering is obtained from these five clustering techniques. To access the clustering quality measure, Dunn Index and Davies Bouldin Index are computed for each technique. Results indicate that the ensemble clustering technique offers superior results as compared to all others state of the art technique. All experiments have been performed on Matlab 2015(b) on a PC with RAM 8GB and Windows 7.

A. Datasets

In this research, two datasets are used, one of which is a private dataset named it "Textclus". The “Textclus” dataset has four classes, education, history, art, and politician. A total of sixty files have been collected by various websites. The proposed ensemble clustering technique has been evaluated at the obtained dataset “20newsgroups” freely available [27]. The obtained experimental results on the basis of both datasets are depicted in TABLE I.

TABLE I. Datasets

Clustering Algorithms	Dataset: Textclus		Dataset: 20newsgroups	
	DI	***DBI***	***DI***	***DBI***
K-Means	*0.3738*	*0.6505*	*0.0707*	*0.4845*
FCM	*0.0702*	*1.7287*	*0.0118*	*0.3229*
Agglomerative	*0.4386*	*0.8118*	*0.3538*	*0.3743*
K-Medoid	*0.0473*	*1.1102*	*0.0903*	*0.4845*
Gustafson Kessel	*0.0154*	*1.7038*	*0.0484*	*0.7409*
Ensemble Clustering	*0.7675*	*0.4454*	*0.7879*	*0.1748*

B. Clustering Quality Measures

It is mandatory for any clustering technique to evaluate the performance by using some standard quality measures. For this purpose, Dunn Index (DI) and Davies Bouldin Index (DBI) are used as clustering quality measures. These clustering quality measures have been computed from the resultant ensemble clustering results as well as the other algorithms results used for the ensemble. A high value of DI indicates better clustering whereas; the low value of DBI represents better clustering.

C. Scenario-I Performance Comparison of Proposed Ensemble Clustering at Textclus Dataset

In this section, first of all, k-means is used which is a part of partitioning methods on the specific dataset; k-means is mostly used for hard clustering. Cluster labels are obtained from k-means and stored it for comparison with other results, after that agglomerative algorithm is used on text data which is a part of hierarchical methods, clustering labels are also obtained from the agglomerative algorithm and then stored it for later use. Next, fuzzy c-Means is applied which is also a part of partitioning methods but mostly used for soft clustering. Similarly, this algorithm to the given data set is applied and then obtained clustering labels for future experiments, after that, k-medoid algorithm is used on the same dataset and obtained results for later use. At the end, the fifth algorithm which is Gustafson Kessel algorithm technique is used and obtained results, stored it for later use. After obtained clustering labels, from all five clustering labels, the ensemble clustering technique is applied to obtain better results, the ensemble cluster labels are obtained by used of majority voting. First, cluster labels are obtained and then applied quality parameters on all clustering labels which are obtained from, k-means, agglomerative, fuzzy c-means, k-medoid, Gustafson Kessel and ensemble clustering technique. In this research, Dunn Index and Devis Bouldin Index are used as quality parameters.

In figure 2, the graph represents the clustering quality of dataset "Textclus". A high value of DI indicates better clustering whereas; the low value of DBI represents better clustering. So Ensemble clustering results are compatible with Dunn Index and Davies Bouldin Index.

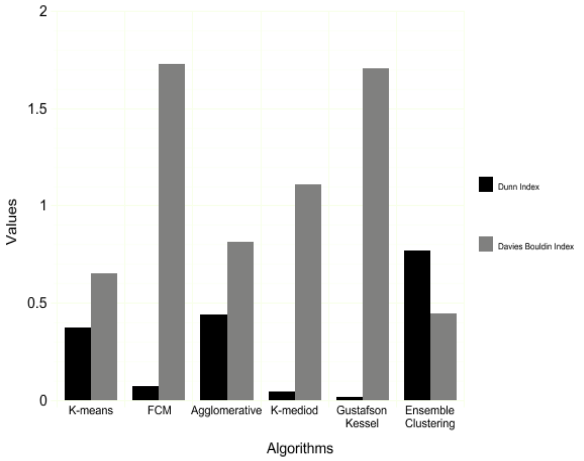

Fig. 2. Clustering quality of Dataset “Textclus”

D. Scenario-II Performance Comparison of Proposed Ensemble Clustering at 20newsgroups Dataset

For the experiment, another dataset "20newsgroups" is also used. Clustering results are obtained using "20newsgroups" and observed that ensemble clustering technique is better than all above techniques which are used before. The high value of DI indicates better clustering whereas; the low value of DBI represents better clustering. So it can be observed that ensemble clustering results are according to the standards of Dunn Index and Devis Bouldin Index which are defined below. After application of these parameters, different results are obtained from different clustering labels.

In this paper, a new clustering technique is introduced which is based on major voting. Various clustering techniques are applied including, agglomerative, fuzzy c–means, k-medoid, Gustafson Kessel and k-means. Inverse document frequency is calculated from given dataset, and used as an input for clustering algorithms, experimental results are obtained using above clustering algorithms then observed that all were different from each other; at the end, ensemble clustering technique is proposed for better results on the basis of majority voting. After that, Dunn Index and Davies Bouldin Index algorithms are applied for quality performance. The results can be observed by bar graph with both data sets.

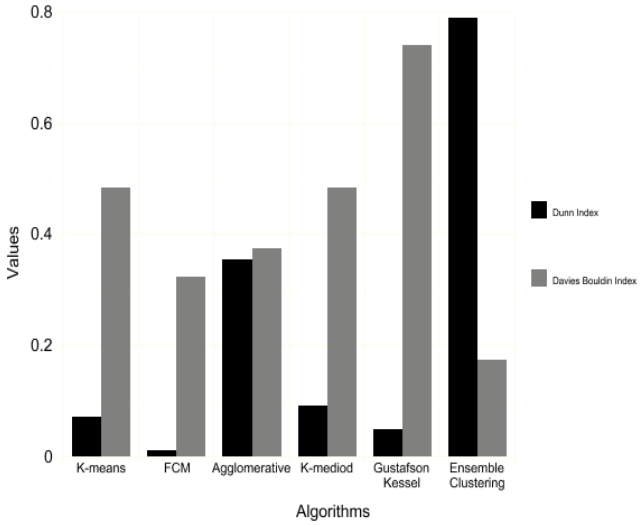

Fig. 3. Clustering quality of Dataset “20newsgroups”

In figure 3, the graph represents the clustering quality of dataset "Textclus". A high value of DI indicates better clustering whereas; the low value of DBI represents better clustering. So ensemble clustering results are compatible with Dunn Index and Davies Bouldin Index.

V. Conclusion and Future Work

In this paper, clustering is discussed that is a group of similar objects. Five clustering methods are applied on datasets, first on "Textclus" then on "20newsgroups" and obtained individual results. After that an ensemble clustering technique is proposed based on major voting, to enhance the performance of text clustering. Cluster quality parameters are applied named as Dunn Index and Devis Bouldin Index. A

high value of DI indicates better clustering whereas; the low value of DBI represents better clustering. During experiments on specified datasets, results of five clustering techniques represented in the above graphs, could not fulfill the requirements of DI and DBI, but ensemble clustering using majority voting technique proved fruitful for text clustering with better clustering results. Therefore, ensemble clustering is found better than five clustering techniques named as k-means, fuzzy c-means, agglomerative, k-medoid, and Gustafson Kessel. For future work, the ensemble clustering technique can be applied on text streams/web data clustering to separate contents and find the extremism content in it.

ACKNOWLEDGMENT

This research was supported by the Basic and Advanced Research Projects in Chongqing, China under Grant No.61672117.

REFERENCES

[1] Hau, C.C., Handbook of pattern recognition and computer vision. 2015: World Scientific.

[2] Cheng, G., J. Han, and X. Lu, Remote sensing image scene classification: benchmark and state of the art. Proceedings of the IEEE, 2017. 105(10): p. 1865-1883.

[3] Mather, P. and B. Tso, Classification methods for remotely sensed data. 2016: CRC press.

[4] Chen, F., et al., Data mining for the internet of things: literature review and challenges. International Journal of Distributed Sensor Networks, 2015. 11(8): p. 431047.

[5] Zuech, R., T.M. Khoshgoftaar, and R. Wald, Intrusion detection and big heterogeneous data: a survey. Journal of Big Data, 2015. 2(1): p. 3.

[6] Witten, I.H., et al., Data Mining: Practical machine learning tools and techniques. 2016: Morgan Kaufmann.

[7] Perez-Guaita, D., et al., Multispectral Atomic Force Microscopy-Infrared Nano-Imaging of Malaria Infected Red Blood Cells. Analytical chemistry, 2018. 90(5): p. 3140-3148.

[8] Nguyen, D.B., New methods for classification, prediction, and feature weighting in bioinformatics. 2015: University of Houston-Clear Lake.

[9] Rajinikanth, T. and G.S. Reddy, A SOFT SIMILARITY MEASURE FOR K-MEANS BASED HIGH DIMENSIONAL DOCUMENT CLUSTERING. IADIS International Journal on Computer Science & Information Systems, 2017. 12(1).

[10] Myhre, J.N., et al., Robust clustering using a kNN mode seeking ensemble. Pattern Recognition, 2018. 76: p. 491-505.

[11] Lu, J., Q. Zhu, and Q. Wu, A novel data clustering algorithm using heuristic rules based on k-nearest neighbors chain. Engineering Applications of Artificial Intelligence, 2018. 72: p. 213-227.

[12] Kang, M., J. Ahn, and K. Lee, Opinion mining using ensemble text hidden Markov models for text classification. Expert Systems with Applications, 2018. 94: p. 218-227.

[13] Liu, H., et al., Spectral ensemble clustering via weighted k-means: Theoretical and practical evidence. IEEE Transactions on Knowledge and Data Engineering, 2017. 29(5): p. 1129-1143.

[14] Huang, D., C.-D. Wang, and J.-H. Lai, Locally weighted ensemble clustering. IEEE transactions on cybernetics, 2018. 48(5): p. 1460-1473.

[15] Rathore, P., et al., Ensemble fuzzy clustering using cumulative aggregation on random projections. IEEE Transactions on Fuzzy Systems, 2018. 26(3): p. 1510-1524.

[16] Lin, S., B. Azarnoush, and G. Runger, CRAFTER: a Tree-ensemble Clustering Algorithm for Static Datasets with Mixed Attributes and High Dimensionality. IEEE Transactions on Knowledge and Data Engineering, 2018.

[17] Li, E., et al. Ensemble Clustering Using Maximum Relative Density Path. in Big Data and Smart Computing (BigComp), 2018 IEEE International Conference on. 2018: IEEE.

[18] Liu, H., et al. Infinite ensemble for image clustering. in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016: ACM.

[19] Zhao, H., et al., Ensemble subspace clustering of text data using two-level features. International Journal of Machine Learning and Cybernetics, 2017. 8(6): p. 1751-1766.

[20] Wazarkar, S. and B.N. Keshavamurthy, A survey on image data analysis through clustering techniques for real world applications. Journal of Visual Communication and Image Representation, 2018. 55: p. 596-626.

[21] Mohebi, A., et al., Iterative big data clustering algorithms: a review. Software: Practice and Experience, 2016. 46(1): p. 107-129.

[22] Learned-Miller, E., et al., Labeled faces in the wild: A survey, in Advances in face detection and facial image analysis. 2016, Springer. p. 189-248.

[23] Sengupta, S., S. Basak, and R.A. Peters. Data Clustering using a Hybrid of Fuzzy C-Means and Quantum-behaved Particle Swarm Optimization. in Computing and Communication Workshop and Conference (CCWC), 2018 IEEE 8th Annual. 2018: IEEE.

[24] Thulasiram, K., S. Ramakrishna, and M. Jayakameswaraiah, To Assess the Performance of EAHC Algorithm Using Sensor Discrimination Dataset for the Improvement of Data Mining System. 2018.

[25] Chen, M., Fuzzy reasoning based evolutionary algorithms applied to data mining. 2015, North Dakota State University.

[26] Masoudi, P., Application of hybrid uncertainty-clustering approach in pre-processing well-logs. 2017, Rennes 1.

[27] "Dheeru, D.a.K.T., Efi",, "{UCI} Machine Learning Repository",. "2017",