International Journal of Research in Advent Technology, Vol.4, No.6, June 2016
E-ISSN: 2321-9637
Available online at www.ijrat.org

Voting Approach for K-means Based Consensus Clustering

Nimesha M. Patil[1], Dipak V. Patil[2]
Department of Computer Engineering[1, 2],
GES's R. H. Sapat College of Engineering, Management Studies and Research, Nashik[1, 2]
Email: nimeshapatil2011@gmail.com[1], *dipakvpatil17@gmail.com*[2]

Abstract-In general, clustering algorithms perform grouping of data items with some limitations because of certain input assumptions. These assumptions made at different run instances of clustering algorithm may give different results. Such a set of multiple varying assumptions based clustering results for same input dataset are called as basic partitions and they may not agree with each other. These disagreements between basic partitions create confusions in deciding which the most promising results are. Consensus clustering considers each basic partitions and aggregate similarities among all of them. In the literature we found K-means algorithm has already been used for doing this type of aggregations. In this paper we used pair-wise similarity method for voting approach combined with k-means based consensus clustering known as KCC. The experiments are performed on well-known UCI Repository datasets. The presented method at its core uses iterative approach while doing aggregation and which we think is its success story.

Index Terms- Consensus clustering, KCC, pairwise similarity.

1. INTRODUCTION

Clustering process is essentially very important technique in decision making. The information items like vehicles information, patients information, shopping products information, movies information, cropping plants information etc. that are available with the user are simply given to clustering algorithms as input. By considering various attributes of the provided information items similarities and dissimilarities between them are calculated [1]. For example we can consider vehicles that are similar or dissimilar in their manufacturer, fuel type, dimensions, engines etc. The similar vehicles are put together in same group. The number of groups to be formed will directly indicate how much cohesiveness in terms of similarities you expect between items. Another issue is the selection of k starting points (initial centroids) for expected k clusters [2] which is very crucial in further progress of the clustering.

Clearly, these are nothing but input assumptions that this is the some assumed expected number of clusters k and these are the k starting points for those k clusters. Every time you change the assumptions and results are not guaranteed to be of same quality. K-means algorithm is a typical clustering algorithm based on distance. It uses distance as the similarity evaluation index, namely the closer the distance of the two objects is, the greater similarity they have, and then they are in the same group. K-means algorithm is very simple to implement and understand. It basically takes an input as number of clusters denoted as 'k' to be formed at the end result. Hence different k-values like1, 2, 3… n can be given by user. For any given k-value the algorithm itself will come-up with best stabilized k clusters. So though locally it is producing optimized result, we need globally optimized result which will be independent of any provided k-value. And selection of initial random centroids also causes to produce different results at different runs of k-means even if we keep same number of clusters.

We studied KCC [2] (K-means based consensus clustering) technique in the literature where a framework for utility functions was designed to become eligible as consensus function. They created basic partitions on UCI repository datasets by varying number of clusters and then applying consensus functions on that inspired us to get on the voting based KCC approach.

Voting approach for K-means based consensus clustering is the way to utilize the K-means algorithm for aggregating the basic partitions clustered results. This approach considers results from individual basic partitions. It looks it as individual votes and forms single meaningful partition of clusters for the information items. Iteratively, it starts with fixed two clusters and then catching the items in same cluster if maximum number of times they were bundled together. Then incrementing one more cluster with centroids as an item which is minimum number of times bundled together with previous centroids. This iterative approach makes it possible to aggregate results resulting into quality consensus partition of items.

In chapter 2, we have brief literature review on clustering techniques, consensus clustering, K-means and KCC. Then we have voting approach for KCC with algorithms in details in chapter 3. Experimental results on UCI datasets are given in chapter 4. Finally, we conclude our experiment and future scope for this work given at the end.

2. LITERATURE SURVEY

K-means has been enhanced from various perspectives by researchers. Some of them are presented below to highlight the noticeable enhancement trends.

Shi Na et al. [3] proposed a new algorithm that solves the need of calculating the distances between each data object and cluster centres in every iteration. For this they used a simple data structure that saves everything required in that iteration reducing the running time and hence computational complexity of standard k-means algorithm.

Iamon N. and Boongoen T. [4] proposed new linked based similarity measure with additional information available in network is included. According to authors this increases the quality of the measures, hence the resulting cluster decision. Compared to previous linked based cluster ensemble (LCE) this refinement performs better when experimented on synthetic and UCI benchmark datasets.

Chen-Chung Liu and Shao-Wei Chu [5] said that it is important to note that accuracy is always reduced because of presence of noisy data, outliers, and the data with quite different values within one cluster. To avoid this limitation in k-means authors proposed two-layer K -means algorithm whose goal is enhancing accuracy rather than the computing speed. When applied, the datasets directly are divided into K clusters that are selected to get sub-cluster centre by K -means algorithm in the 1st stage. The sub-cluster centre is separated into K groups in the 2nd stage. The two-layer K-means algorithm contains three steps: data normalization, Cluster centre initialization, and two-layer clustering. F-measure is the standard they used to evaluate the accuracy of algorithms in the experiment.

Bhatia S. [6] proposed a new technique that solves the need of initializing cluster centres of traditional k-means randomly and hence avoiding possible errors raised by this random nature. For this they used genetic algorithm to select the appropriate initial clusters converging quickly to local optimum. According to authors such proper selection of initial cluster simply puts the limit on the number of iterations required in traditional k-means algorithm.

Jie jhang and Jianrui Dong [7] proposed a new method called as P-partition. This method is useful to find cluster centres optimally. Two relative clustering algorithms are used for replacing the mean centre obtained by p-partition. Objective function here produces better values comparatively to standard k-means.

Juntaowang and Xiaolong Su [8] used noise data filters to identify noisy data based on their characteristics, which is referred as density based detection method. According to authors when done with such pre-processing of filtering operation the influence of noise is drastically decreased.

H. G. Ayad and M. S. Kamel [9] introduced cumulative voting` concept where probabilistic mapping is computed for aligning the cluster labels. Their methodology initially minimizes the average squared distance between the mapped partitions optimally represent the ensemble. As authors described that an efficient solution is obtained using an agglomerative algorithm that minimizes the average generalized divergence within the cluster.

Shaohong Zhang et al. [10] targeted ensembling problem by stating that selection of suitable cluster ensemble method for specific data in unsupervised manner becomes critical because of unavailability of true information at hand before clustering. According to authors consensus affinity of cluster ensemble helps significantly improvement for ensemble solution selection and even for partition selection. Carl Meyer et al. [11] proposed a methodology where cluster ensembling is used to determine the number of clusters. They defined graph on similarity matrix by using different k values and also using different algorithms. A random walk then performed on graph to determine number of clusters from Eigen values of respective transition probability matrix. Each iteration of consensus clustering refinement is done to remove noisy data.

Sadeghian A. H. and Nezam abadi-pour H. [12] presented idea of “Gravitational Ensemble Clustering (GEC)” to ensemble results of different but weak clustering algorithms for identification of true quality clusters. For this they have used theory of gravity. According to authors the proposed ensemble method proven to be robust and versatile while considering clusters of different shapes, sizes and densities overs are the individual and other ensembles clustering algorithms. Shi Yao Liu et al. [13] researched similarity-based methods of clustering. They adopted weights into those methods so that priorities can be assigned. Then they used all this integration for cluster ensembleing with experimenting on real world data sets. According to authors results are proven to be valid and advantageous than other approaches.

Abu-Jamous et al. [14] proposed binarization of consensus partition matrix to obtain a fuzzy based consensus partition. Here, the binarizations represent the truth that multiple clusters will be containing same genes and other clusters cannot have same genes at all. This enabled there to find out such genes that belongs to multiple clusters simultaneously. According to authors experimental results on periodic gene dataset successfully show gene clustering improvements.

Jain A.K. et al. [15] concentrated on one of the problem of consensus clustering that their inability in handling uncertain data pairs misleading in generation of final consensus partitions. For this proposed matrix completion method where data pairs agreed upon most

of the clustering algorithms are represented through similarity based matrix. The final data partition is computed by applying an efficient clustering algorithm to the completed matrix.

Abdullin and Nasraoui O. [16] worked on clustering of heterogeneous data. Data that comprises of multiple domains or modalities like categorical, numerical and transactional data are required to be converted into similar type format and then processed by traditional clustering algorithm. Another approach is ensemble clustering that achieves the same purpose for clustering heterogeneous data. Kuncheva L.I and Vetrov D. P. [17] proposed cluster ensembles based on k-means clusters where k values are randomly generated for multiple runs of k-means. Authors found that relationship between stability and accuracy with respect to the number of clusters depends on the data set, varying from almost perfect positive correlation to almost perfect negative correlation. In response to this authors proposed a new combined stability index to be the sum of the pair wise individual and ensembles.

Yazhou Ren et al. [18] said that most of clustering ensemble algorithms treats each clustering and each object as equally important and not much effort has been put towards incorporating weighted objects into the consensus process. In response to this problem authors proposed "Weighted-Object Ensemble Clustering". They determine the weights of the objects by looking at how difficult it would be to cluster an object by constructing the co-association matrix. Then presented three different consensus techniques reduce the ensemble clustering problem to a graph partitioning one.

Emmanuel Ramasso, Vincent Placet and Mohamed Lamine Boubakar [19] proposed another methodology for unsupervised example acknowledgment in acoustic outflow (AE) time-arrangement gave from compound materials. The innovation holds in the improvement of a clustering ensemble strategy ready to underline surprising developments of harms in mixes under requesting. The technique joins different allotments issued from various parameters, introductory conditions and calculations. A first stage consequently chooses different subsets of attributes in light of the entropy of groupings of harms distinguished by bunching. Unsupervised example acknowledgment in AE time-arrangement give from compound materials was handled by the utilization of different clustering's. A programmed highlight choice was proposed combined with an enhancement of the quantity of clusters.

Xing Xiaoxue and Guan Xiuli [20] proposed the use of the harsh Set hypothesis to predispose the information; persistent characteristic required is the fundamental and key step. Here, a discretization strategy depends on the k-means algorithm was built up. Utilizing this technique, the totally qualities could be arranged into the 2 sorts. Four sets data on UCI database were checked the presentation of the proposed strategy. In this analysis, the k-means algorithm was utilized to actualize the information discretization firstly then they are utilized to do characteristic lessening through unpleasant set.

It is a technique in view of unsupervised bunch. Every property will be unsupervised cluster into two classes, and afterward get less discrete breakpoints. The technique for discretization taking into account data entropy, the strategy for discretization in view of the trait significance and the strategy proposed in the work are recreated on UCI datasets. The results demonstrate that Classification precision rate of discretization strategy taking into account k-means enhanced the break point, diminish the season of the investigation and the multifaceted nature of the examination, enhanced the trial productivity, and got great results.

Nuwan Ganganath and Chi-Tsun Cheng [21] proposed a regularly utilized strategy as a part of account, software engineering, and building. In a large portion of the methodology, cluster sizes are either compelled to particular qualities or accessible as earlier information. Lamentably, typical Consensus techniques can't constrain confinements on group sizes. In this work, they propose some indispensable alterations to the standard k meansalgorithm such that it can incorporate size requirements for every cluster independently. The enhanced k-means algorithm can be utilized to acquire groups in usable sizes. A potential application would be obtaining clusters with equivalent group size. Recreation results on multidimensional information exhibit that the k-means algorithm with the present changes can satisfy group size limitations and help to more precise and strong results.

Data clustering techniques can't satisfy the size limitations on clusters. In this work, they present a powerful algorithm for data clustering with compelled group sizes. The proposed algorithm is created in light of the standard k-means algorithm. They changed the standard calculation such that it can consolidate bunch size requirements. In the introduction venture of the altered algorithm, it utilizes the former information to relegate information focuses as the beginning centroids of the groups, dissimilar to arbitrary information point task in a standard K-means calculation.A. Strehl and J. Ghosh [22] stated that, it is widely recognized that merge multiple classification or regression models typically gives better results compared to using a single. However, there are no well-known approaches to merge multiple non-hierarchical clustering. The idea of combining cluster labeling without accessing the original features leads the general knowledge reuse framework that call cluster ensembles. In this technique define the cluster ensemble problem is an optimization problem and to propose three successful and efficient combine for

solving it based on a hyper graph model. Result synthetic as well as real data sets are given to display that cluster ensembles can (i) improve quality and robustness and (ii) enable distributed clustering.

N. Nguyenand, R. Caruana [23] attended the problem of combining various clustering's without access to the primary features of the data. This process is well known in the literature as clustering ensembles, clustering aggregation, or consensus clustering. Consensus clustering provide a stable and robust final clustering that is in agreement with multiple clustering's. They find that an iterative EM-like method is remark ably productive for this problem. They presented an iterative algorithm and its variations for detecting clustering consensus. An extensive empirical study compares their proposed algorithms with eleven other consensus clustering methods on different four datasets using three different clustering performance metrics. The experiment all results shown that then ewensemble clustering methods produce clustering's that are as good as, and often better than, these other methods.

A. Topchy, A. Jain and W. Punch [24] presented that a dataset can be clustered in many ways depend on the clustering algorithm employed, parameter settings used and other factors. They addressed a question that can multiple clustering be combined so that the final partitioning of data provides better clustering? The answer best on the quality of clustering's to be combined as well as the properties of the fusion method. First, they introduce presentation for different clustering's and formulate the corresponding categorical clustering problem. A result, they presented that the consensus function is related to the classical intra-class variance standard using the generalized mutual information definition. Second, they showed thief of combining partitioning generated by poor clustering algorithms that use data projections and random data splits.

S. Vega-Ponsand, J. Ruiz-shulcloper [25] proposed that, cluster ensemble has proved to be a best alternative when facing cluster analysis problems. It is form of generating as of clustering's. From this a metadata stand combining them into anal clustering. The goal of this composition process is to improve the quality of individual data clustering. Due to the increasing appearance of new methods, their favorable results and the great number of applications, they consider that it is necessary to make a critical analysis of the be alive techniques and future projections an overview of clustering ensemble methods that can be very helpful for the community of clustering practitioners. The characteristics of several methods were discussed, which may use in the selection of the most arrogate on solve a problem at hand. They also presented taxonomy of these techniques and illustrated some important applications.

X. Wang, C. Yang and J. Zhou [26] proposed that, a large number of clustering algorithm exists; aggregating different clustered partition sin to a single consolidated one to obtain good results has become an important problem. In Fred and jain's accumulation algorithm, they construct a co- association matrix on original partition label's and then register minimum spanning tree to this matrix for the combined clustering.

K. Punera and J. Ghosh [27] observed the problem of obtaining a single consensus clustering solution from a ensemble of clustering's of a set of objects, has enhanced much interest recently because of its numerous practical applications. While a wide various types of approaches including graph partitioning, more possibly, genetic algorithms, and voting-merging have been presented so far to solve this problem, nearly all of them work on hard partitioning, i.e., where an object is a member of exactly one cluster in any individual solution.

V. Filkov and S. Steven [28] presented that, with the exploding volume of micro array experiments in the increasing interest in mining repositories of such data. Meaningfully merge results from diverged experiment so an equal basis is a challenging task. Here they proposed a general method for integrating heterogeneous datasets based on the consensus clustering formalism. Method analyzes source clustering's and identities a consensus set-partition which is as close as possible to all of them. They developed a general criterion to assess the potential of integrating multiple heterogeneous datasets, i.e. in case the integrated data is more instructive than the individual datasets.

B. Mirkin [29] proposed the category utility function is a partition quality score function applied in some clustering programs of machine learning. They are interpreted this function in terms of the data variance shows by a clustering or equivalently, in term of the square-error classical clustering criteria on those operators the K-Means and Ward methods. This analysis recommends extensions of the scoring function to situations with differently standardized and mixed scale data.

T. Li, M. M. Ogihara, and S.Ma [30] proposed that, many problems can be reduced to the problem of combining multiple clustering. In this work, they first encapsulate different application scenarios of combining multiple clustering and provided a new perspective of observing the problem as a categorical clustering problem. This is very crucial and very important technique they have used for improving the results. In consensus clustering, such techniques have been proved to be major role players. They directly affect on the quality of clusterings and gives proper analysis at hand for the decision making problems.

3. METHODOLOGY

Basic partition creation and voting approach for consensus clustering are the two major parts of our work. In first part basic partition creation varies number of desired clusters and initial centroids for obtaining disagreeing clustering results. For this the basic algorithm used at core is the traditional K-means algorithm. This K-means is run by varying K value from true clusters (N) to square root of number of instances $(\sqrt{n})$. For example in Iris dataset it is varied from 3 to 12. For every K-value 10 different initialization of centroids are done. Thus total of BPs created for Iris dataset are 100.

3.1. K-means Algorithm

Input: input instances, number of clusters (k)
Output: A partition of k clusters
Steps:

(1) Choose k initial centroids randomly from the input instances.
(2) Repeat following until stabilized clusters obtained
- Assign every instances to their closest cluster centroids using Euclidean distance
- Calculate new centroids in every clusters

(3) The final stabilized clusters are saving as a basic partition.

3.2. Iterative Pairwise Consensus Algorithm

This is second part which aggregates all basic BPs. The similarity measure between the data point x and a cluster of c data points $(x_1,x_2,x_3,\ldots.x_n)$ is defined as;

$$S(x,\{x_1,x_2,\ldots.,x_c\}) = \frac{\sum_{i=1}^{c} S(x, x_i)}{c}$$

Where, $S(x,x_i)$ is similarity count between a particular instance x and every other instance x_i in the input dataset. Similarity count indicates that in how many basic partitions that pair of instances is grouped together. Voting approach for KCC uses this algorithm.

Input: a set of basic partitions (∏)
Output: a consensus partition
Steps:

(1) Calculate similarity counts for every pair of instances in the given basic partitions.
(2) Choose 2 initial centroids randomly that are furthest apart instances.
(3) Repeat following until stabilized clusters obtained
- Assign every instance to their closest cluster centroids using Similarity Count.
- Create new cluster and assign new cluster centroid for it by finding instance that is furthest from previous centroids.

(4) The final stabilized cluster partition is saved as a consensus partition.

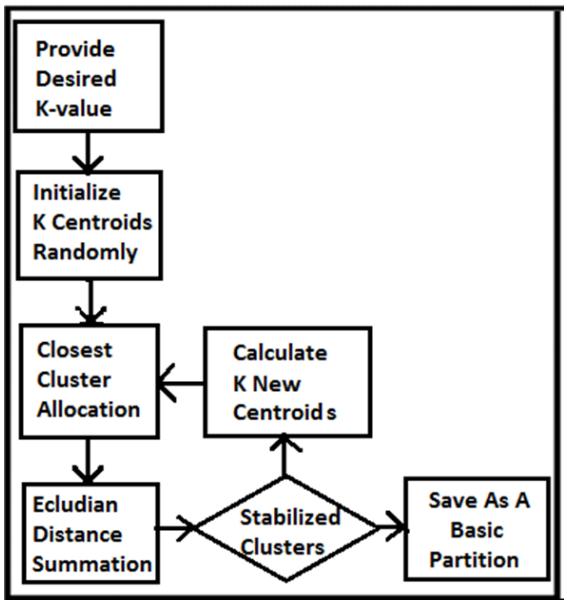

Fig. I. Flowchart of Creation of Basic Partitions

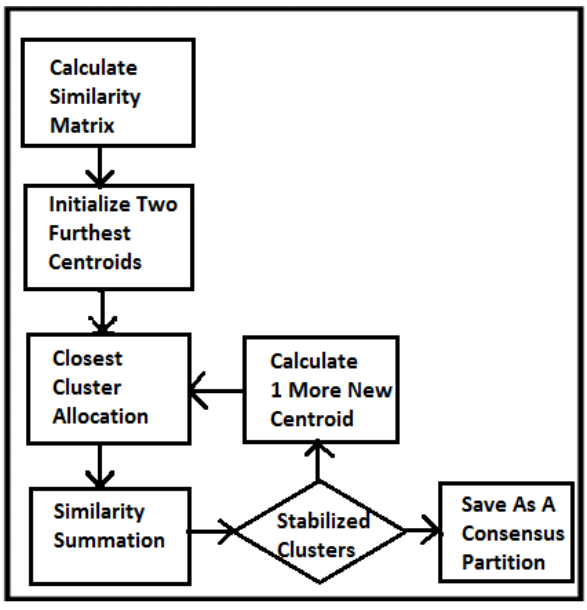

Fig. II. Flowchart of Creation of Consensus Partition

4. EXPERIMENTAL RESULTS

To check the performance of voting approach for k-means based consensus clustering we downloaded three most popular data sets from UCI database [31].The description of these datasets is as given in Table-I.

Table I: Dataset Description

Dataset	#Instances	#True Clusters	#Attributes
Iris	150	3	4
Ecoli	335	8	7
Wine	165	3	13

For measuring performances of basic partitions and voting approach for KCC we used Rand Index metric. Rand index (Rn) which is one of the external indexes metric has been used for this purpose. It has capability measure the inter-cluster and intra-cluster similarity for obtained basic partitions

$$\text{Rand Index (Rn)} = \frac{TP + TN}{TP + FP + FN + TN} \quad \text{Eq. (1)}$$

Where in Eq. (1),

- TP (true positive): Similar objects assigned to same cluster
- TN (true negative): Similar objects assigned to different clusters
- FP (false positive): Dissimilar objects assigned to same clusters
- FN(false negative) : Dissimilar objects assigned to different clusters

We generated 100 Basic partitions for every attribute of Each Datasets. We now count best rand index out of 100 and compare it with voting approach for KCC. We obtain the following comparison table II.

Table II: Quality Comparison Results

Dataset	Average Quality Basic Partitions			Quality Consensus Partition		
	TP	TN	RN	TP	TN	RN
Iris	546	7070	0.68	2526	5276	0.79
Ecoli	3804	34896	0.66	4125	34816	0.67
Wine	1150	9363	0.66	345	8240	0.71

The results show that quality of basic partitions varies as per input parameters and sometimes very poor to average clusters are obtained. We analyze that iterative problem solving of voting approach when combined with KCC basic partition creation technique produce improved quality clusters in consensus partition.

5. CONCLUSION AND FUTURE SCOPE

The decision making with the help of clustering techniques can be made easy and efficient by using consensus clustering methodology. KCC which uses k-means algorithm has been proven to be one of the best techniques for this task. We have used basic partition creation way from KCC and used them with voting approach. The results show the quality improvement after combining both of the techniques together. In future, new consensus functions can be designed to be helpful for consensus clustering. Parallel Computing can be introduced to create basic partitions at first and avails robust partitions as input to voting approach for KCC.

REFERENCES

[1] Nam Nguyen and Rich Caruana, “Consensus Clustering”, Seventh IEEE International Conference on Data Mining, 2010.
[2] Wu, Hongfu Liu, Hui Xiong, Senior Member, Jie Cao, and Jian Chen, “K-Means Based Consensus clustering: a unified view,” IEEE transactions on knowledge and data engineering, vol. 27, no. 1, January 2015.
[3] Shi Na, Liu Xumin and Guan Yong, “Research on k-means clustering algorithm: An improved K-means Clustering Algorithm,” Proceeding of 3rd IEEE International Symposium on Intelligent Information Technology and Security Informatics (IITSI), pp. 63-67, 2010.
[4] Iamon N. and Boongoen, “Improved link-based cluster ensembles,” Proceeding of International Joint Conference on Neural Networks (IJCNN), pp. 1-8, 2012.
[5] Chen Chung Liu and Shao-Wei Chu, “A Modified K-means Algorithm - Two-Layer K-means Algorithm,” Tenth International Conference on Intelligent Information Hiding and Multimedia Signal Processing, 2014.
[6] Bhatia S., “New Improved technique for initial cluster centers of k-means clustering using Genetic Algorithm,” Proceeding of IEEE 3rd International Conference for Convergence of Technology (I2CT), pp. 1-4, 2014.
[7] Jiejhang and Jianrui Dong, “A new method on finding optimal centers for improving k-means algorithm,” Proceeding of 27th IEEE Control and Decision Conference (CCDC), pp-1827-1832, May, 2015.

[8] Juntao wang and Xiaolong Su, "An improved K-means clustering algorithm," Proceeding of IEEE 3rd International Conference on Communication Software and Networks (ICCSN), pp-44-46, 2011.
[9] H. G. Ayad and M. S. Kamel, "Cumulative voting consensus method for partitions with variable number of clusters," IEEE Trans. Pattern Anal. Mach. Intell., vol. 30, no. 1, pp. 160–173, Jan. 2008.
[10] Shaohong Zhang, Liu Yang and DongqingXie, "Unsupervised evaluation of cluster ensemble solutions," Proceeding of IEEE 7th International Conference on Advanced Computational Intelligence (ICACI), pp. 101 – 106, March 2015.
[11] Carl Meyer, Shaina Race and Kevin Valakuzhy, "Determining the Number of Clusters via Iterative Consensus Clustering," August 6, 2014.
[12] Sadeghian, A.H. and Nezamabadi-pour H., "Gravitational ensemble clustering," Proceeding of IEEE Iranian Conference on Intelligent Systems (ICIS), pp. 1-6, 2014.
[13] ShiYao Liu, Qi Kang, Jing An and MengChu Zhou, "A weight-incorporated similarity-based clustering ensemble method," Proceeding of 2014 IEEE 11th International Conference on Networking, Sensing and Control (ICNSC), pp. 719 – 724, 2014.
[14] Abu-Jamous, B, RuiFa, Nandi A.K., Roberts D.J., "Binarization of Consensus Partition Matrix for ensemble clustering," Proceedings of the 20th European Signal Processing Conference (EUSIPCO), pp. 2193 – 2197, 2012.
[15] Jinfeng Yi, Tianbao Yang, Rong Jin and Jain, A.K., "Robust Ensemble Clustering by Matrix Completion," Proceeding of IEEE 12th International Conference on Data Mining (ICDM), pp. 1176 – 118, 2012.
[16] Abdullin and Nasraoui O., "Clustering Heterogeneous Data Sets," proceeding of IEEE Eighth Latin American conference on Web Congress (LA-WEB), pp. 1-8, Oct. 2012.
[17] Kuncheva L.I and Vetrov D. P., "Evaluation of Stability of k-Means Cluster Ensembles with Respect to Random Initialization," IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 1798 – 1808, Volume-28, Issue-11, 2006.
[18] Yazhou Ren, Domeniconi C., Guoji Zhang, Guoxian Yu, "Weighted-Object Ensemble Clustering," IEEE 13th International Conference on Data Mining (ICDM), pp. 627 – 636, 2013.
[19] Emmanuel Ramasso, Vincent Placet and Mohamed Lamine Boubakar, "Unsupervised Consensus Clustering of A Coustic Emission Time-Series for Robust Damage Sequence Estimation in Composites," IEEE., 2015.
[20] Xing Xiao xue and Guan Xiuli, "Continuous Attribute Discretization Algorithm of Rough Set Based on K-means," IEEE Work shop on Advanced Research and Technology in Industry, 2014.
[21] Nuwan Ganganath and Chi-Tsun Cheng Junjie , "Data Clustering with Cluster Size Constraints Using a Modified k-means Algorithm," IEEE DOI 10.1109/Cyber C., 2014.
[22] A. Strehl and J. Ghosh,, "Cluster ensembles A knowledge reuse framework for combining partitions," J.Mach. Learn. Res. , vol. 3, pp. 583–617, 2002.
[23] N. Nguyen and R. Caruana, "Consensus Clusterings," in Proc. Of IEEE Int. Conf. on Data Mining, pp. 607–612, 2007.
[24] A. Topchy, A. Jain and W. Punch, "Combining multiple weak clusterings," in Proc. of 3rd IEEE International .Conference on Data Mining, pp.331–338, 2003.
[25] S. Vega Ponsand J. Ruiz shul cloper, A survey of clustering ensemble algorithms,‖Int.J.PatternRecogn.Artif.Intell.,vol.25,no.3,pp.337–372,2011.
[26] X. Wang, C. Yang and J. Zhou, "Clustering aggregation by probability accumulation," PatternRecog., vol.42, no.5, pp. 668–675, 2009.
[27] K.Punera and J. Ghosh, "Consensus-based ensembles of soft clustering's," Appl. Artif. Intell., vol. 22,no.7-8, pp. 780–810, 2008.
[28] V.Filkov and S. Steven, "Integrating microarray data by consensus clustering", Int.J.Artif.Intell.Tools, vol. 13, no.4, pp. 863–880,2004.
[29] B.Mirkin, "Reinterpreting the category utility function," Machine Learning, vol. 45, no.2, pp. 219–228, Nov. 2001.
[30] T. Li, M. M. Ogihara and S. Ma, "On combining multiple clustering's: an overview and a new perspective," Appl. Intell., vol. 32, no.2, pp. 207–219, 2010.
[31] Blake C L Merz C J. UCI repository of machine learning databases [EB/OL], available:http://www.ics.uci.edu/mlearn/MLRepository.html.