Journal of Scientific & Industrial Research
Vol. 79, April 2020, pp. 323–327

A SimRank based Ensemble Method for Resolving Challenges of Partition Clustering Methods

R S M Lakshmi Patibandla* and N Veeranjaneyulu

Vignan's Foundation for Science, Technology & Research, Vadlamudi, Guntur, A P, India

Received 6 October 2018; revised 7 August 2019; accepted 4 February 2020

Traditional clustering techniques alone cannot resolve all challenges of partition-based clustering methods. In the partition based clustering, particularly in variants of K-means, initial cluster centre selection is a significant and crucial point. The dependency of final cluster is totally based on initial cluster centres; hence, this process is delineated to be most significant in the entire clustering operation. The random selection of initial cluster centres is unstable, since different cluster centre points are achieved during each run of the algorithm. Ensemble based clustering methods resolve challenges of partition-based methods. The clustering ensembles join several partitions generated by different clustering algorithms into a single clustering solution. The proposed ensemble methodology resolves initial centroid problems and improves the efficiency of cluster results. This method finds centroid selection through overall mean distance measure. The SimRank based similarity matrix find that the bipartite graph helps to ensemble.

Keywords: Partition clustering, Cluster ensemble, Similarity matrix

Introduction

Cluster analysis is a vital technique, which analyses the huge and multi-variant data of various fields like text data analysis, image analysis and spatial data analysis etc. There are many partition-based clustering methods, but there is no clustering method capable of finding number of clusters, appropriate initial centroid. The cluster analysis evolution is based on the various cluster validity measures, which are used to measure the quality of the clustering results.

Cluster Ensemble

There is a new methodology called cluster ensemble, helps to trust different clustering outcomes to increase the prominence of the cluster results. In the ensemble technique, there are two main steps namely Generation and Consensus.[1] Generation is foremost step in clustering ensemble method, where the set of clusters are combined. In the generation step, there is a mechanism of combining different clustering algorithms, different object representation, different clustering algorithms, and different parameter initialization or may be different subsets of objects.[2,3] The ensemble process may provide robustness, consistency, novelty and stability. The general notation of the cluster ensemble, let X= $\{x_1, x_2 \ldots\ldots, x_m\}$ be a set of m data points and E= $\{e_1, \ldots\ldots e_n\}$ be a cluster ensemble with n base clusters, each of which is referred to as an ensemble member. Every base cluster returns a set of clusters E_i= $\{c_1, c_2 \ldots c_k\}$is a partition of the set of objects X with k clusters. C_1 is the e^{th} cluster of i^{th} partition.

Generation Step

In this step different clustering algorithms are combine or same algorithm with different parameters initialization may be applied.[4] There are some methodologies for generation process homogenous ensembles, heterogeneous ensembles Random K ensemble, mixed heuristics etc.[5,6]

Consensus Function

The second step in the ensemble clustering is the consensus function. This function should be proficient of enlightening the solitary clustering algorithm. To obtain the cluster ensemble, different consensus functions have been developed for deriving efficient data[10] partitions. The basic idea in this is to determine, which must be cluster label connected to each object in the consensus partition. To do this each consensus function exploits an explicit form of information matrix, which reviews the improper cluster results.

The literature study exposed that there are numerous challenges in cluster analysis persisted, yet numerous investigators recommended many methods

*Author for Correspondence
E-mail: patibandla.lakshmi@gmail.com

to resolve. Still there is a chance to improve further.[4] In the partition based clustering in particularly variants of K-means, initial cluster centre selection is a significant and crucial point. The dependency of final cluster is totally based on initial cluster centres; hence, this process is delineated to be most significant in the entire clustering operation.[8,9] The simple K-means is a standard, simplest clustering technique to cluster numerical and unbalanced datasets. The random selection of initial cluster centres is unstable, since different cluster centre points are achieved during each run of the algorithm.

In this work, attentive to partition clustering algorithm is comprehensive and K-means algorithm is precise. The inadequate key challenges are recognized and find solutions to the partitioning methods as shown below:

- The selection of initial centroid data points as initial seed selection.
- Finding the correct number of cluster i.e. K value prior.

The proposed cluster ensemble, from Fig. 1 involves two generation methods namely Generation 1 and Generation 2. In the generation 1, homogeneous ensemble process of base clustering is produced consuming repetitive turns of a solitary K-means algorithm, through numerous sets of constraint initializations, namely different cluster centroids. The same K-means algorithm erratically choosing a number of K clusters for each ensemble member in generation 2. This data will be sent to Initial Centroid Algorithm. This work used link-based similarity method for consensus purpose. The co-association-based consensus function works for pair-wise similarity and makes it possible to find the co-occurrence relations between the data points. Various consensus functions used in the proposed model for getting efficient cluster results are Number of Clusters in each Partition (NCP), and consider the original Set of Objects (CSO). The amount of clusters in the consensus partition is a parameter of the consensus function (CPC).

Initial Centroid Selection Algorithm

The Initial centroid selection is based on variance, where the selection of centroid is based on the overall mean distance measurement. This proposed algorithm finds the pairwise distance between the clusters and sort the distance using mean distance measure.

Input: Data set, Number of clusters (K).

Step 1: Choose number of cluster (K) arbitrarily from the data set. Calculate the mean or median of all the input data sets.

Step 2: Find the pairwise distance between the data points

Step 3: For C=1 to K

The cluster centre is selected for all K- clusters

Return $(l_1, l_2, \ldots\ l_k)$

End for

Step 4: The clusters are formed through the normal run of K-means /K-medoid/K-modes. (If the data set is numerical use K-means, for other than numeric K-modes or K-medoid algorithm may applied). The initial centroid selection is based on the step 3 return values i.e. $(l_1, l_2, \ldots\ l_k)$, where $l_1, l_2, \ldots\ l_k$ are the initial centroids of the cluster. The total algorithm has two parts, in the first part is used to calculate the initial centroid selection procedure and it will be input for the second phase of the algorithm for efficient cluster results.

Link-Based Similarity Method

Pair Wise Similarity Matrix Generation

Generally, the cluster ensemble method is created on the data point pair wise resemblance amongst data points.[7] The data set X= $\{x_1, x_2..., x_N\}$, the clustering algorithm causes a cluster ensemble $\Pi = \{\pi_1, \pi_2.... \pi_M\}$ by smearing M base clustering to the data set X. Through this NxN similarity matrix is erected for each member of ensemble noted as Sm, m=1.... M. Every entry in the matrix represented the relationship between two data points. The entry in the matrix 1 indicates the same cluster otherwise 0. Similarity matrix (CO) is generated with equation (1).

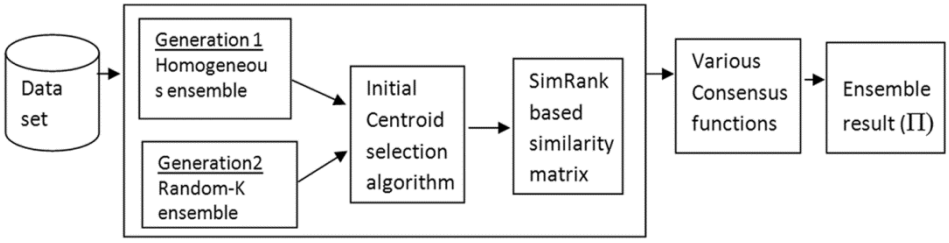

Fig.1 — The Proposed Ensemble Model

$$CO_{ij} = \frac{1}{M}\sum_{i=1}^{m} s_m(x_i\, y_j) \quad ...(1)$$

Where CO (i, j) Ɛ [0, 1] represents the similarity measure between x_i,y_j sample, which are belongs to the X. In this =1 indicates similarity among and otherwise. In this method generates co-association matrix, which is a similarity matrix.[7] The similarity matrix may apply to yield the final partitions Π.

Simrank Based Similarity (SRS) Matrix

In the similarity matrix, numerous matrix entries are unknown and mention them as '0'. The quality of the final clustering results may inefficient. Hence, the link-based method has overcome the problem of mysterious values and expands the exactness of the final clustering result. Simrank is considered as one of the standard technique for link-based similarity evolution. It covers the scope of similarity estimation out there the local context of adjacent neighbours. In a graph G= (V, E), where V and E are set of vertices and edges. In the Simrank based similarity measured method must find the similarity of any two vertices, v_i, v_j Ɛ V as per the following equation 2.

$$s(vi, vj) = \frac{DC}{|N_{vi}||N_{vj}|}\sum_{x=1}^{|N_{vi}|}\sum_{y=1}^{|N_{vj}|} s\left(N_{vi}^{x}, N_{vj}^{y}\right) \quad ...(2)$$

Where DC ∈[0,1] is a decay factor, $N_{vi} \subset V$ *and* $N_{vj} \subset V$ are neighbour sets and whose members are directly linked to vertices v_i, v_j. Individual neighbours are specified as N_{vi}^{x} *and* N_{vj}^{y}.In this process $s(vi, vj)$=0 when $N_{vi} = \emptyset$ *or* $N_{vj} = \emptyset$ and the optimal similarity values to fixed-point after iteration.

$$\lim_{t\to\infty} Rt(vi, vj) = s(vi, vj)\text{--} \quad ...(3)$$

This may be shortened as

$$Rt + 1(vi, vj) = \frac{DC}{|N_{vi}||N_{vj}|}\sum_{x=1}^{|N_{vi}|}\sum_{y=1}^{|N_{vj}|} Rt(N_{vi}^{x}, N_{vj}^{y}.) \quad ...(4)$$

The above-mentioned process is applied for the cluster ensemble as network of clusters.[7–9]The representation of bipartite shows the relationship among the objects. In this process as earlier mentioned cluster ensemble Π, a graph G= (V, E) constructed, and V represented a set of vertices of data points and clusters in the ensemble. E represents a set of edges among the data points and clusters, which they are assigned. Let us assume like SRS (a, b) is the entry in the SRS matrix, which represents the similarity between any pair of data points or similarity between any two clusters in the ensemble. Suppose for example a=b SRS (a,b)=1 Otherwise, the equation is mentioned equation(5).

$$SRS(a, b) = \frac{DC}{|N_a||N_b|}\sum_{x=1}^{N_a|}\sum_{y=1}^{|N_b|} s(N_a^{x}, N_b^{y}) \quad ...(5)$$

As mentioned in the equation (4) the iterative refinement of SimRank measure is used. This is shown in Fig. 2.

Experimental Results

Experimental evaluation is conducted over the bench mark data set, which is available in the UCI Machine Learning Repository. The experimentation is conducted for analysing the performance of the proposed methodologies in this work. The proposed methodology is compared with number of clustering algorithms. Each methodology in the clustering divides data set into K number of partitions, in the process of comparison against the actual cluster number along with the following mentioned evaluation parameters. Compactness, Davies-Bouldin (DB), Dunn and classification accuracy (CA) are the parameters used

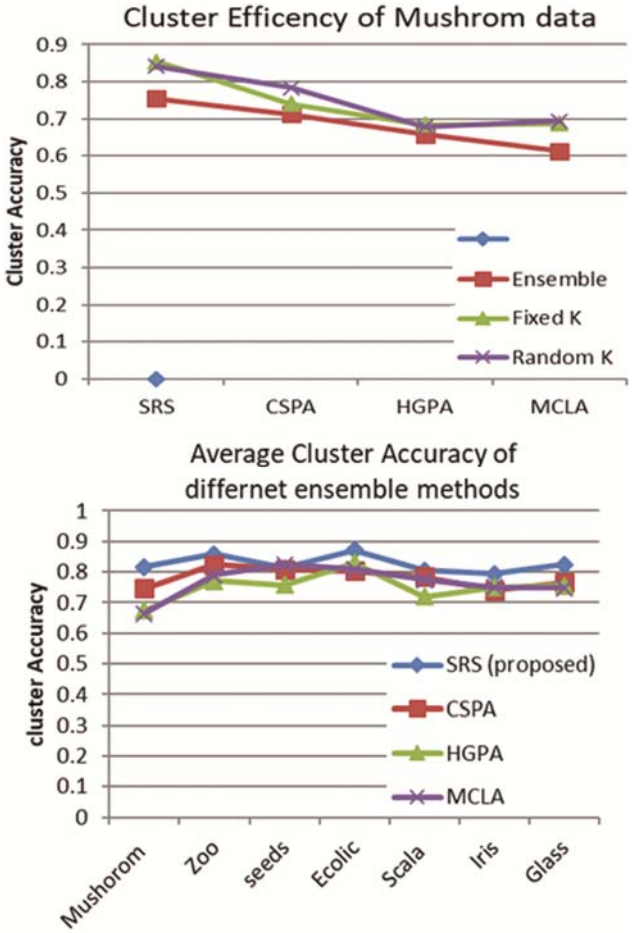

Fig. 2 — a) Cluster Efficiency of the Mushroom Data, b) Average Cluster Accuracy of Different Ensemble Methods

for comparing various ensemble methods.The cluster ensemble methods transform the cluster ensemble problem into a graph based partitioning problem.

As per the proposal of the researchers Strehl and Ghos[11], the heuristic based partition methods namely CSPA (Cluster based Similarity Partitioning Algorithm), HGPA (Hyper-Graph Partitioning Algorithm), and MCLA (Meta Clustering Algorithm) are solving consensus problems. The proposed work is compared with the above mentioned methods along with traditional methods K-means and K-modes. The proposed method potentiality is fully evaluated with base line methods. The evaluation is assessed against with three ensemble methods and two traditional methods as shown in the Table 1. The experimental results proved that the cluster accuracy is better than other existing methods namely CSPA, HGPA and MCLA.

Conclusions and Future Work

Cluster ensemble has been an emerging technique for cluster analysis and improves the efficiency of the cluster results. In the partition based clustering particularly, variants of K-means, initial cluster centre selection is a significant and crucial point. The dependency of final cluster is totally based on initial clustercentres; hence, this processis delineated to be most significant in the entire clustering operation. The simple K-means is a standard, simplest clustering technique to cluster numerical and unbalanced datasets. The random selection of initial cluster centres is unstable, since different cluster centre points are achieved during each run of the algorithm. The proposed ensemble methodology resolves initial centroid problems and improves the efficiency of cluster results. This method finds centroid selection through overall mean distance measure. The SimRank based similarity matrix finds a bipartite graph which helps to an ensemble. The experimental results proved that the proposed ensemble methodology resolve the mentioned partition-based clustering methods. In the future, extend this work towards ensemble various clustering methods to resolve different challenges

Table 1 — Comparison of Clustering Accuracy with Other Existing Methods

Data set	Ensemble type	SRS (proposed)	CSPA	HGPA	MCLA	K-means	K-modes
Mushroom	Ensemble	0.754	0.712	0.657	0.612	0.687	0.710
	Fixed K	0.854	0.741	0.684	0.687	--	--
	Random K	0.841	0.784	0.678	0.694	--	--
Zoo	Ensemble	0.897	0.784	0.658	0.711	0.756	0.810
	Fixed K	0.745	0.812	0.812	0.813	--	--
	Random K	0.932	0.875	0.841	0.845	--	--
Seeds	Ensemble	0.781	0.845	0.665	0.794	0.678	0.794
	Fixed K	0.812	0.798	0.800	0.781	--	--
	Random K	0.845	0.784	0.810	0.894	--	--
Ecoli	Ensemble	0.894	0.841	0.864	0.789	0.645	0.710
	Fixed K	0.845	0.779	0.813	0.801	--	--
	Random K	0.874	0.789	0.812	0.834	--	--
Scale	Ensemble	0.781	0.812	0.745	0.800	0.712	0.714
	Fixed K	0.845	0.794	0.657	0.784	--	--
	Random K	0.789	0.745	0.754	0.741	--	--
Iris	Ensemble	0.841	0.856	0.827	0.789	0.745	0.694
	Fixed K	0.799	0.674	0.710	0.783	--	--
	Random K	0.742	0.694	0.710	0.678	--	--
Glass	Ensemble	0.812	0.741	0.698	0.748	0.712	0.678
	Fixed K	0.845	0.754	0.789	0.810	--	--
	Random K	0.741	0.678	0.698	0.701	--	--
Wine	Ensemble	0.817	0.769	0.871	0.799	0.754	0.678
	Fixed K	0.789	0.657	0.741	0.687	--	--
	Random K	0.885	0.874	0.845	0.864	--	--
Fertility	Ensemble	0.840	0.800	0.784	0.810	0.741	0.678
	Fixed K	0.785	0.542	0.687	0.702	--	--
	Random K	0.812	0.801	0.782	0.687	--	--

like shapes of the clusters, a variety of data sets to apply, etc.

References

1 Sandrovega-pons & JoseRuiz-Shulcloper, A Survey of Clustering Ensemble Algorithm, *Int Jof Pat Rec and Art Int*, **25** (2011) 337–372.

2 Ahmad, A & Dey L A, K-Means Type Clustering Algorithm for Subspace Clustering of Mixed Numeric and Categorical Data Sets, *Pat Rec Let*, **32** (2011) 1062–1069.

3 Song G, Ye Y, Zhang H, Xu X, Lau R Y, & Liu F, Dynamic Clustering Forest: An Ensemble Framework To Efficiently Classify Textual Data Stream With Concept Drift, *Inf Sci*, **357** (2016) 125–143.

4 Madhuri R, Ramakrishna Murty M, Murthy J V R & Prasad Reddy P V G D, Cluster Analysis on Different Data sets using K-modes and K-prototype algorithms, *Adv Int Sys Com*, **249** (2014) 137–144.

5 Kuo R J, Mei C H, Zulvia F E, & Tsai, C Y, An Application of a Meta Heuristic Algorithm-Based Clustering Ensemble Method to APP Customer Segmentation, *Neucom*, **205** (2016) 116–129.

6 Li F J, Qian Y H, Wang J T, & Liang J Y, Multi-Granulation Information Fusion: A Dempster-Shafer Evidence Theory-Based Clustering Ensemble Method, In Mac Lea Cyb, **1** (2015) 58–63.

7 Ramakrishna Murty M, Murthy J V R & Prasad Reddy P V G D, Homogeneity Separateness: A New Validity Measure for Clustering Problems, *Adv Int Sys Com*, **248** (2014) 1–10.

8 Topchy A, Jain A K & Punch W, Combining Multiple Weak Clusterings, *IEEE Con Dat Min'03*, (2003) 331–338.

9 Ramakrishna Murty M, Murthy J V R, Prasad Reddy P V G D & Satapathy Suresh C, A Survey of cross-Domain Text Categorization Techniques, *Rec Adv Inf Tec* (2012) 499–504.

10 Huang D, Lai J-H, & Wang C-D, Robust Ensemble Clustering Using Probability Trajectories, *IEEETra Kno Dat Eng*, **28** (2016) 1312–1326.

11 Strehl A and Ghosh J, Cluster ensembles a knowledge reuse framework for combining multiple partitions, *J Mach Learn Res*, **3** (2003) 583–617.