© 2016 IJSRSET | Volume 2 | Issue 4 | Print ISSN: 2395-1990 | Online ISSN : 2394-4099
Themed Section: Engineering and Technology

A Survey : Clustering Ensemble Techniques with Consensus Function

M. Mekala, P. Elango

Department of Computer Science, Gobi Arts & Science College, Gobichettipalayam, Erode, Tamil Nadu, India

ABSTRACT

The clustering ensembles contains multiple partitions are divided by different clustering algorithms into a single clustering solutions. Clustering ensembles used for improving robustness, stability, and accuracy of unsupervised classification solutions. The major problem of clustering ensemble is the consensus function. Consensus functions in clustering ensembles including hyperactive graph partition, mutual information, co-association based functions, voting approach and finite machine. The characteristics of clustering ensembles algorithm are computational complexity, robustness, simplicity and accuracy on different datasets in previous techniques.

Keywords : Clustering ensembles, Consensus function, Unsupervised classification.

I. INTRODUCTION

Data clustering is the essential tools for perceptive structure of data set. It plays a initial role in data mining, information retrieval and machine learning. The traditional clustering algorithms are limited in managing datasets that have categorical attributes. Clustering ensembles have emerged as an effective solution that is able to overcome limitations and develop the strength of clustering results. Different clustering solutions are equally plausible without a prior knowledge about the underlying data distributions. Cluster ensemble is the method to merge numerous of dissimilar clustering to get common partition of the original dataset. Every clustering algorithm implicitly or explicitly uses certain data model and it produces erroneous or meaningless results.

II. METHODS AND MATERIAL

1. Literature Survey

Clustering ensembles have appeared as a dominant means for improving both the strength and stability of unproven classification solutions. Clustering analysis has widely applied in real world application domains such as data compression, data mining and pattern recognition.

Analoui and Sadeghian has proposed that using objective functions stable partitions and cluster Selections are produced using the genetic algorithm [6]. The probabilistic model of consensus using a finite mixture of multinomial distributions in a space of clustering, a combined partition is found as a soluition to be corresponding maximum problem using the genetic algorithm. The excellent scalability of the algorithm and comprehensible underlying model are particularly important for clustering of large datasets.They can be calculated a correction matrix that show correction between samples and found the best samples that in the center of cluster.

Ng et al. has proposed a popular multiwayspectral graph partitioning algotithm(SPEC) that searching to optimize the normalized criterion[5]. SPEC[8] can be simply described as a graph G=(V,W), it computes the degree matrix D,which is a diagonal matrix such as that D(i,i)=Pj W(I;j). Based on the matrix D, it computes a normalized weight matrix and find L's K largest vestors u1, u2…..uK to form matrix U= [u1….uK]. The rows of U are normalized to have unit length. SPEC produces the final clustering solution by clusterings the embedded points using K-means.

Comparing to HBGF, SPEC has low robust clustering performance.

Karypis and kumar has proposed a multilevel graph partitioning system named METIS, approaches the graph partitioning problem from a different angle. That coarsen the graph by collapsing vertices and edge in clustering and partition the coarsened graph refine the partitions[4]. In comparison to other graph partitioning algorithms METIS is highy efficient and competitive performance.

Strehl and Ghosh has proposed the knowledge reuse to influence a new cluster based on different set of futures, control size of partition, low computational cost of HGPA, improving the quality and robustness of the solution and allowing one to add a stage that selects the best concensus function without any supervisory information by objective function[7]. They assumed X={xi,x2….xn} denote a set of objects/samples/points. A partitioning of these n objects into k clusters can be represented as a set of k sets objects {Ct|ℓ=l…..k} or as a label vector. A cluster is a function that delivers a label vector given a tuple of objects.

Fern and brodley has proposed the low computational cost, high robust clustering performance against instance and cluster based approaches and comparing to IBGF and CBGF,the reduction of HBGF is lossless[3]. It constructs a bipartite graph from a set of partitioning to be combined, modeling objects and clusters simultaneously as vertices, and later partitioning the graph by a traditional graph partitioning technique. Their approach retained all of the information provided by a given ensemble, allowing the similarity among instances.

Fischer and buhmann[10],[11] presented path based clustering with automatic outlier detection that captures the empirical observation that group structures in embedding spaces might appear as considerable extension but are characterized by local homogeneity and connectivity. Path based clustering is applicable, some clustering in situation when the parametric form of such a transformation is unknown.

Dudoit and Fridlyand[12] proposed two bagged clustering procedures to improve and assess the accuracy of a partitioning clusterings method. The baggings are used to generate and aggregate multiple clustering to assess the confidence of cluster assignments for individual observation. In addition, bagged clustering procedures are more robust to the variable selection scheme, the accuracy is less sensitive to the number and type of variables used in the clustering.

2. Clustering Ensembles

Clustering ensemble can used in multiobjective clustering as a compromise between individual clustering with conflicting objective functions. Clustering ensemble system solves a clustering problem in two methods. The first method is a dataset as input and output as ensemble of clustering solutions. The second method is the cluster ensemble as input and combines the solutions to produce a single clustering as final output. There are many feasible approaches to improve the performance of clustering analysis.

Clustering ensembles can typically achieved by a single clustering algorithm in several aspects:

1. Robustness
2. Novelty
3. Stability and confidence estimation
4. Parallelization and scalability

The objective of the clustering process is to recover the quality of individual data clustering. Figure shows the general process of cluster ensemble, consists of generating a set of clusterings from the similar dataset and combining into an ultimate clustering [1].

There are two parts in cluster ensemble

1. Generation Mechanisms
2. Consensus Functions

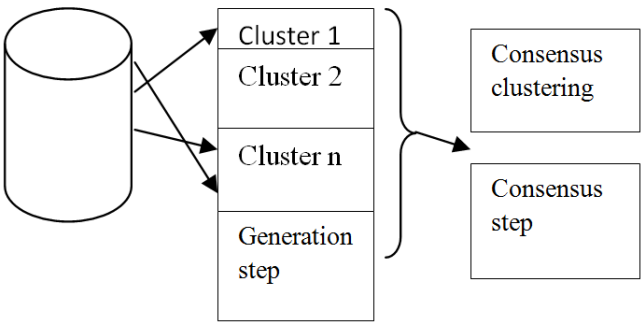


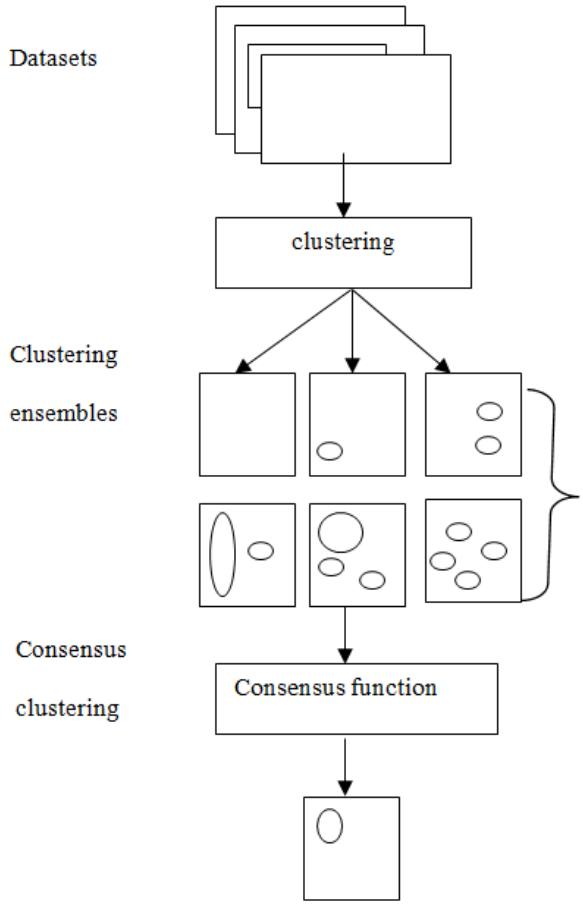
Figure 1. Process of cluster ensemble

2.1 Generation Mechanisms

Generation is the clustering ensemble methods,in which the set of clusterings is generated and combined. It generates collection of clustering solutions. Dataset of n instances X={X1,X2,…..Xn} an ensemble constructor generates a cluster ensemble. Clustering ensembles stores the results of some independent runs of K-means or other clustering algorithms.

2.2 Consensus Functions

The consensus function in clustering ensemble algorithm produces the final data partition or consensus partition, which is the result of any clustering ensemble is used [2].



Graph Based Algorithms

A weighted graph is denoted by G=(V,W), where V is a set of vertices and W is a nonnegative and symmetric |V|x|V| similarity matrix characteristics the similarity between each pair of vertices[8]. The contribution to graph partition crisis is a weighted graph G and a number K.The sum of the weights of these crossed edges is defined as the cut of a partition P: Cut(P;W)=PW(I;j), where vertices i and j do not belongs to the same cluster.

There are number of method for devising graphs from cluster ensemble. They are

A. Cluster Based Similarity Partition Algotithm(CSPA)
B. Hypergraph Partition Algotithm(HGPA)
C. Meta Clustering Algorithm(MCLA)
D. Hybrid Bipartite Graph Formulation(HBGF)
E. Spectral Graph Partitioning Algorithm(SPEC)

Cluster Based Similarity Partition Algotithm (CSPA)

In the Cluster-based Similarity Partitioning Algorithm(CSPA), from the hyper graph, a similarity matrix n x n (the co-association matrix) is constructed. This will be viewed as the adjacency matrix of a fully connected graph. The nodes are the elements of the set X and an edge between one or more objects has an associate weight equal to the number of times the objects are in the cluster[9]. Similarity between two objects are 1 in the same cluster and 0 otherwise. Each clustering, n x n binary similarity matrix is created by objects. The entry-wise average of r such matrices representing the r sets of grouping an overall similarity matrix. Alternatively this can be interpreted as using k binary clustering membership features and similarity as fraction of clusterings in two objects are in the same cluster[6].

$$S = \frac{1}{r} HH^{+}$$

Hypergraph Partition Algotithm (HGPA)

The clusters cloud be represented as hyperedges on a graph whose vertices match to the objects to be cluster, every hyper edges describes set of objects belongings to the identical clusters. The minimum-cut of a hypergraph is reduced by the process of consensus clustering. This hypergraph has a minimum k-cut into k components which gives the important consensus partitions are used[6]. Hypergraph partitioning is NP-hard problem. But it analyzed that it is very difficult to solve the k way min-cut partitioning problem with computational complexity. All hyperedges are considered to have same weight. Also, all vertices are similarity based.

$$k.max1\epsilon\{1 \ldots . k\}n\ 1/n \leq 1.05$$

Meta Clustering Algorithm(MCLA)

The Meta-Clusering Algorithm is based on clustering. It also confidence estimates of cluster membership. All the similarity between two clusters are defined in terms of the amount of objects grouped in both,using the index matrix. Then, a matrix of similarity between clusers is formed, represents the adjacency matrix of the graph built considering the clusters as nodes and assigning a weight to the edge between two nodes, equal to the clusters. The MCLA is to group and collapse related hyperedges is to be assign each object to the collapsed hyperedges in which it participates most of the matrices[6].

Hybrid Bipartite Graph Formulation (HBGF)

HGBF constructs a bipartite graph from a set of partitions to be combined, modeling objects and clusters simultaneously as vertices, and later partitioning technique. Their approach retined all of the information provided by a given ensemble, allowing the similarity among instances(IBGF) and the similarity among clusters(CBGF) to be considered collectively in forming the final clustering. HBGF has high robust clustering performance against IBGF and HBGF and reduction of HBGF is lossless.

Spectral Graph Partitioning Algorithm (SPEC)

Spectral graph partitioning chooses a different multi-way spectral graph partitioning algorithm, which tries to find to optimize the regulate criteria. SPEC described as a graph G=(V,W), it computes the degree matrix D, which is a diagonal matrix such that D(i,i)=Pj W(I;j). Based on D, it computes a normalized weight matrix and find L's K largest vestors u1,u2…..uK to form matrix U=[u1….uK]. The roes of U are normalized to have unit length.

III. CONCLUSION

Clustering ensemble is a technique emerged and acts as a major overcoming the drawbacks of individual clustering consequences. The major clustering ensemble approaches captivating into report of theoretical description and the mathematical computation. The paper describes the general process of cluster ensemble and different types of consensus function. The improving complexity in voting approach, finite mixture model and co-association based functions can be an investigation in future.some of the most important research works in clustering ensembles techniques have empirical results on different datasets. Most of the clustering ensembles technique need to improve their accuracy, therefore improving accuracy can be research in future.

IV. REFERENCES

[1] Sandro vega-pons and jose ruiz-shulcloper, "a survey of clustering ensemble algorithms", International journal of pattern recognition and artificial intelligence vol. 25, no.3(2011).

[2] Javd Azimi, Paul Cull and Xiaoli Fern, "Clustering Ensembles Using Ants Algorithm", EECS Department, Oregon State University, Corvallis, Oregon, 97330,USA.

[3] M. Analoui and N. Sadighian, "Solving cluster ensemble problems by correction's matrix & GA", IFIP International Federation for Information Processing, vol. 228, pp. 227-231,2006.

[4] A. Ng, M. Jordan and Y. Weiss,"On specral clustering: Analysis and an algorithm", NIPS 14,2002.

[5] G.Karypsis and V,kumar, "A fast and high quality multilevel scheme for partitioning irregular graphs", SIAM Journal on Scientific Computing, pp.359-392, 1998.

[6] A. Strehl and J.Ghosh, "Cluster ensembles – A knowledge reuse framework for combining multiple partitions", Journal of Machine Learning Research,pp.583-617, Feb. 2002.

[7] A.Topchy, B.Minaei Bidgoli, A.K.Jain and W.Punch, "Adaptive clustering ensembles", Proceedings of the International Conference on Machine Learning,Canada,2004.

[8] Xiaoli Zhang Fern and Carla E.Brodly,"Solving Cluster Ensemble Problems by Bipartite Graph Partitioning", International Conference on Machine Learning,Banff, Canada,(2004.)

[9] Joydeep Ghosh and Ayan Acharya, "Cluster ensembles" , Volume 1, July/August(2011).

[10] B.Fischer and J.M Buhmann, "path-based clustering for grouping of smooth curves and texure segmentati9on", IEEE Transaction on

pattern Analysis and Machine Intelligence, vol.25, no.4, Apr.2003.

[11] Y. Hong,s.Kwong, Y.Chang and Q.Ren. "Unsupervised feature selection using clustering ensembles and population based incremental learing algorithm", Pattern recognition society. vol.41, no.9, pp.119-123, Sep.2009.

[12] S.Dudoit and Ferd "Finding consistent cluster in data partition", Springer-verlag Berlin Heidelberg, Mcs,pp. 309-318, 2001.

[13] Y.C. Chiou and L.W. Lan. "Genetic clustering algorithms", EJOR European Journal of operational Research, vol .135, pp.413-427, Nov. 2001.