Imperial Journal of Interdisciplinary Research (IJIR)
Vol-2, Issue-5, 2016
ISSN: 2454-1362, http://www.onlinejournal.in

Survey of Single and Multi Objective Clustering Ensemble

Bhumi Patel[1] & Lokesh Gagnani[2]
[1]ME Student , Department of Information Technology, Kitrc-Kalol, Gandhinagar-382721.
[2]Assistant Professor, Department of Information Technology, Kitrc-Kalol, Gandhinagar-382721

Abstract: *Clustering is a popular data analysis and data mining technique. Most of the clustering methods use only one objective function to partition data items into the clusters. It seems that using more than one objective provide the ability for a clustering method to provide better performance. The most common purpose of an analysis is to choose the best trade-offs among all the defined and conflicting objectives. However, many clustering studies are formulated as a problem whose goal is to find the "best" solution, which corresponds to the minimum or maximum value of a single objective function that lumps all different objectives into one.*

1. Introduction

1.1 Data Mining

Data-mining is the process of extracting information from large amounts of data. Based on data that are processed, the extraction of data is useful for: Obtaining a model for future events; Identifying variables and attributes of the process which is studied; Prediction (forecasting) of future variation of variables[1]

Data mining can be classified into two high level categories, such as [1]

- Predictive Data Mining
- Descriptive Data Mining

1) Predictive Data Mining:

This model of data mining techniques creates a model to predict the future values based on the past and current data values. The various Predictive Data Mining techniques are

a) Classification
b) Regression Analysis
c) Time Series Analysis
d) Prediction

2) Descriptive Data Mining:

This model of data mining techniques organizes the data, based on their general properties and transforms it into human interpretable patterns, associations or correlations. The various Descriptive Data Mining techniques are

a) Clustering
b) Summarization
c) Association Rule Mining
d) Sequence Discovery

1.2 Clustering

Clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters).in other groups (clusters).

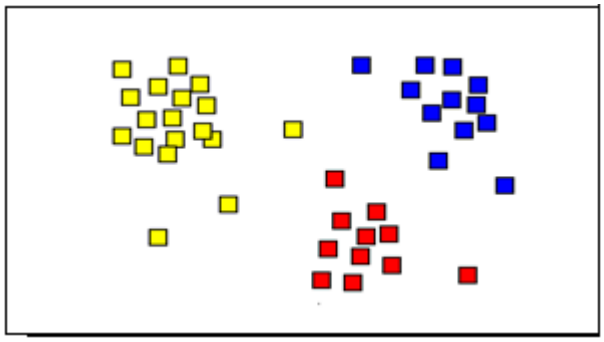

Figure 1: The result of a cluster analysis shown as the colouring of the squares into three clusters.

Data clustering is one of the essential tools for perceptive structure of a data set. It plays a vital and initial role in data mining, information retrieval and machine learning. The basic goal in cluster analysis is to discover natural groupings of objects in a dataset. The data set sometimes may be in mixed nature that it may consist of both numeric and categorical type of data and differ in their individuality.[2]

2. Clustering Ensemble

A cluster ensemble system solves a clustering problem in two steps. The first step takes a data set as input and outputs an ensemble of clustering solutions. The second step takes the cluster ensemble as input and combines the solutions to produce a single clustering as the final output. Figure 2 shows the general process of cluster ensemble, that consists of generating a set of clustering from the similar dataset and combining them into an ultimate clustering. The objective of this combination process

is to recover the quality of individual data clustering. The intend of combining dissimilar clustering results emerged as an unusual approach for improving the quality of the results of clustering algorithms.[3]

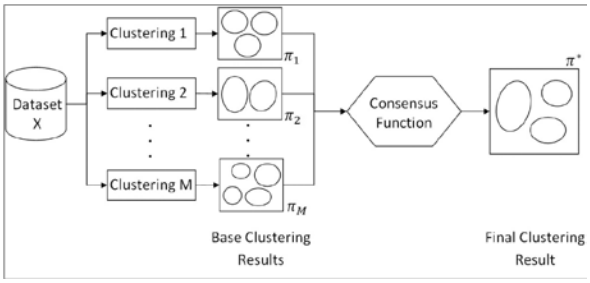

Figure 2: Basic Process of Cluster Ensembles[3]

There are two major parts in cluster ensemble
1) Generation mechanisms [3]
2) Consensus functions

A. Generation Mechanism

Generation is the first step in clustering ensemble methods, innwhich the set of clusterings is generated and combined. It generates a collection of clustering solutions i.e., a cluster ensemble.

Given a data set of n instances X = {X1,X2, · ··,Xn}, an ensemble constructor generates a cluster ensemble, represented asπ= {π1,...,πr} where r is the ensemble size(the number of clustering in the ensemble). Each clustering solution πi is simply a partition of the data set X into Ki disjoint clusters of instances, represented as πi =ci1.... cik.

B. Consensus Function

The consensus function is the main step in any clustering ensemble algorithm that produces the final data partition or consensus partition , which is the result of any clustering ensemble algorithm, is obtained.There are some types of consensus function such as:[8]

– Co-association based function
– Graph based methods
– Voting approaches
– Mixture model approaches
– Information theory approach

3. Multi objective clustering ensemble

The goal of multi-objective clustering is to find clusters dataset by applying several clustering algorithms corresponding to different objective functions. We propose a clustering approach that integrates the output of different clustering algorithms into a single partition. More precisely, given different clustering objective functions, we seek a partition that utilizes the appropriate objective functions for different parts of the data space. This framework can be viewed as a meta-level clustering since it operates on multiple clustering.

Multiobjective clustering is a two-step process: (i) independent or parallel discovery of clusters by different clustering algorithms, and (ii) construction of an "optimal" partition from the discovered clusters. The second step is a difficult conceptual problem, since clustering algorithms often are not accompanied by a measure of the goodness of the detected clusters. The objective function used by a clustering algorithm is not indicative of the quality of the partitions found by other clustering algorithms. The goodness of each cluster should be judged not only by the clustering algorithm that generated it, but also by an external assessment criteria.

4. Single vs Multi objective clustering

Many real-world decision making problems need to achieve several objectives: minimise risks, maximise reliability, minimise deviations from desired levels, minimise cost, etc. The main goal of single- objective (SO) Clustering is to find the "best" solution, which corresponds to the minimum or maximum value of a single objective function that lumps all different objectives into one. This type of clustering is useful as a tool which should provide decision makers with insights into the nature of the problem, but usually cannot provide a set of alternative solutions that trade different objectives against each other. On the contrary, in a multiobjective clustering with conflicting objectives, there is no single optimal solution. The interaction among different objectives gives rise to a set of compromised solutions, largely known as the trade-off, no dominated, noninferior or Pareto-optimal solutions. [12]

The consideration of many objectives in the design or planning stages provides three major improvements to the procedure that directly supports the decision-making process: [12]

(1) A wider range of alternatives is usually identified when a multiobjective methodology is employed.
(2) Consideration of multiple objectives promotes more appropriate roles for the participants in the planing and decision-making processes, i.e. "analyst" or "modeller"– who generates alternative solutions, and "decision maker" - who uses the solutions generated by the analyst to make informed decisions.
(3) Models of a problem will be more realistic if many objectives are considered.

Single-objective clustering identifies a single optimal alternative, however, it can be used within the multiobjective framework.

This does not involve aggregating different objectives into a single objective function, but, for example, entails setting all except one of them as constraints in the optimization process. Those objectives expressed as constraints are assigned different levels of attainment of their respective objective functions (e.g. minimum reliability levels) and several runs are performed to obtain solutions corresponding to different satisfaction of constraints. However, most design and planning problems are characterized by a large and often infinite number of alternatives. Thus, multiobjective methodologies are more likely to identify a wider range of these alternatives since they do not need to prespecify for which level of one objective a single optimal solution is obtained for another.[12]

A very simplified view of the decision-making process is that it involves two types of actors: analysts (modelers) and decision makers. This is a crude simplification of the process since many stakeholders and actors may be involved, but simple enough to demonstrate shortcomings of assuming that in general one person can assume both (or many more) roles. Analysts are technically capable people who provide information about a problem to decision makers who decide which course of action to take. Modeling and optimization techniques are tools which analysts may use to develop useful information for the decision makers. However, single-objective models require that all design objectives must be measurable in terms of a single fitness function. This in turn requires some a priori ordering of different objectives (i.e., a weighting scheme) to allow easy integration of them into a single function of same units. Thus, single-objective approaches place the burden of decision making squarely on the shoulders of the analyst. For example, it is the analyst who must decide the cost equivalent of a specific risk of failure. Even if the decision makers are technically capable and willing to provide some a priori preference information, the decision making role is taken away from them. By providing a trade-off curve between different objectives and alternative solutions corresponding to the points on this curve, multiobjective approaches allow for the responsibility of assigning relative values of the objectives to remain where it belongs: with the decision maker![12]

5. Conclusion

Clustering ensemble is a foremost technique emerged and acts as a major keystone for overcoming the
drawbacks of individual clustering consequences. Hence In this paper, we survey some of the major clustering ensemble approaches captivating into report their theoretical description used by all means. The paper describes the general process of cluster ensemble. Our future research effort will focus on achieving better consensus results in Multi objective clustering ensemble.

5. Acknowledgements

We would like to thank the support for the Help. We also are grateful to Julia Handl for her prompt assistance with MOCK.

References

[1]**Abdul Wahid, XiaoyingGao,PeterAndreae**" Multi-Objective Clustering Ensemble for High-Dimensional Data based on Strength Pareto Evolutionary Algorithm (SPEAII)" 978-1-4673-8273-1/15/$31.00 c 2015 Crown.

[2]**Jay Prakash · P. K. Singh**, "An effective multi objective approach for hard partitionalclustering" Springer-Verlag London Limited 2009

[3]**KattiFaceli, Marc´ılio C. P. de Souto**" Multi-Objective Clustering Ensemble"IEEE Proceedings of the Sixth International Conference on Hybrid Intelligent Systems (HIS'06)0-7695-2662-4/06 $20.00 © 2006.

[4] **KattiFaceli, Marcilio C.P. de Souto, Daniel S.A. de Araujo , Andre´ C.P.L.F. de Carvalho.**"Multi-objective clustering ensemble for gene expression data analysis" 0925-2312/$ - see front matter & 2009 Elsevier B.V.

[5]**Martin H. C. Law Alexander P. Topchy Anil K. Jain**, "Multiobjective Data Clustering" - To appear in IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004

[6]**Onur Can Sert , KayhanDursun, TanselÖzyer ,** "Ensemble of Multi-Objective Clustering Unified With H-Confidence Metric as Validity Metric, 2011 International Conference on Advances in Social Networks Analysis and Mining.2011 IEEE.

[7]**Reza Ghaemi · Nasir bin Sulaiman ,Hamidah Ibrahim · Norwati Mustapha**, "A review: accuracy optimization in clustering ensemblesusing genetic algorithms" - Springer Science+Busines Media B.V. 2010.

[8] **S.Sarumathi,N.ShanthiG.Santhiya"**A Survey of Cluster Ensemble**"-** International Journal of Computer Applications (0975 – 8887) Volume 65– No.9, March 2013

[9**]S.Sarumathi ,N.Shanthi, M.Sharmila**, "A Comparative Analysis of Different Categorical DataClustering Ensemble Methods in Data Mining" - International Journal of Computer Applications (0975 – 8887) Volume 81 – No.4, November 2013

[10]**SujoyChatterjee⁎, AnirbanMukhopadhyay**, "Clustering Ensemble: A Multiobjective Genetic Algorithm basedApproach"- ScienceDirec**t**, (CIMTA) 2013

[11]**Venkatadri. M,HanumatG. Sastry**" Genetic Programming in Data mining Tasks "© 2010, IJARCS All Rights Reserved
[12] **Dragan Savic**"Single-objective vs. Multiobjective Optimisation for Integrated Decision Support"University of Exeter, UnitedKingdom, E-mail: D.Savic@ex.ac.uk2013