ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 5, Issue 2, February 2018

A Survey of Cluster Ensembles and its Applications: Specific Problem Domains

Jyoti
MCA, M. D. University, Rohtak

Abstract **- Cluster ensembles have been shown to be better than any standard clustering algorithm at improving accuracy and robustness across different data collections. This meta-learning formalism also helps users to overcome the dilemma of selecting an appropriate technique and the corresponding parameters, given a set of data to be investigated. Almost two decades after the first publication of a kind, the method has proven effective for many problem domains, especially microarray data analysis and its down- streaming applications. Recently, it has been greatly extended both in terms of theoretical modelling and deployment to problem solving. The survey attempts to match this emerging attention with the provision of fundamental basis and theoretical details of state-of-the-art methods found in the present literature. It yields the ranges of ensemble generation strategies, summarization and representation of ensemble members, as well as the topic of consensus clustering. This review also includes different applications and extensions of cluster ensemble, with several research issues and challenges being highlighted.**

INTRODUCTION

Cluster ananlysis is usually employed in the initial stage of understanding a raw data, especially for new problems where prior knowledge is minimal. Also, in the pre-processing stage of super- vised learning, it is exploited to identify outliers and possible object classes for the following expert-directed labelling process. This is crucial when the complexity of modern-age information is generally overwhelming for a human investigation. The need to acquire knowledge or learn from the excessive amount of data is hence a major driving force for making clustering a highly active research subject. Data clustering is applied to a variety of problem domains such as biology [1], customer relationship management [2], in- formation retrieval [3,4], image processing [5,6], marketing [7,8], psychology [9] and recommender systems [10]. In addition, the recent development of clustering cancer gene expression data has attracted a lot of interests amongst computer scientists, biological and clinical researchers [11–13].Principally, the core of cluster analysis is the clustering process which divides data objects into groups or clusters such that objects in the same cluster are more similar to each other than to those belonging to different clusters [14]. Objects under examination are normally described in terms of object-specific (e.g., attribute values) or relative measurements (e.g., pairwise dissimilarity). Unlike supervised learning to which classification is categorized, clustering is 'unsupervised' and does not require class information, which is typically achieved through a manual tagging of category labels on data objects, by a domain expert (or through the con- sensus of multiple experts). Given its potential, a large number of research studies focus on several aspects of cluster analysis: for instance, clustering algorithms and extensions for particular data type [15], dissimilarity (or distance) metric [16], optimal cluster numbers [17], relevance of data attributes per cluster or subspace clustering [18], evaluation of clustering results [19], and cluster ensembles [20].

A solution to this dilemma remains an ultimate goal. In order to accomplish this, researchers invented the methodology of combining different clusterings into a single consensus clustering. This process which is widely known as 'cluster ensembles' can provide more robust and stable solutions across different domains and datasets [20,22,24]. However, modelling a mechanism (usually referred to as a 'consensus function') that is effective for integrating multiple data partitions in a cluster ensemble is far from trivial. This task is difficult since there is no well defined correspondence between the different clustering results. The further challenges arising from the need to combine data partitions and generate a better clustering result without prior knowledge are of high interest amongst researchers.

1.The problem of cluster ensembles

This paper first presents the fundamental concepts of data clustering including a number of benchmark algorithms that have been employed for various problem domains. Each of these con- ventional techniques are designed on a particular assumption(s), which is normally realized via input parameters. Generally, there is no clustering algorithm, or the algorithm with distinct parameter settings, that performs well for every set of data. To overcome the difficulty with identifying a proper alternative, the methodology of cluster ensemble which is the focus of this review has been continuously developed in the past decade. The second part of this section includes details of general framework and an overview of cluster ensemble methods found in the literature

II. UNTIL THE TERMINATION CRITERIA ARE MET.

The examples of termination criteria are: (i) no changes are made to the cluster centres (i.e., no reassignment of any data point from one cluster to another), (ii) the maximum number of iterations is exceeded, and (iii) there is no improvements in the objective function such as decrease in the square-error. The k- means algorithm is popular largely due to its efficiency, with time complexity of O(Nkr), where N is the number of data points, k is the number of clusters and r is the number of iterations. However, it is sensitive to the choices of initial cluster centres (i.e., different initial states can lead to different output partitions). One might have to run the algorithm multiple times with various initial partitions and chooses the resulting clustering that offers the minimum square- error. Yet, k-means does not work well on noisy data and non-convex cluster shapes.

III. CLUSTER ENSEMBLE METHODS

3.1. Direct approach

The first family of cluster ensemble methods is characterized by the use of a combination strategy such as 'voting', which has proven effective for classifier ensembles [68,69]. However, such practice is not directly applicable to the cluster ensemble problem where a priori class information is not available. The cluster labels in different data partition (i.e., base clustering) πg , g = 1 . . . M are arbitrary. As a result, a mechanism that finds 'label correspon- dence' and re-labels each partition in accordance with a reference partition, is necessary for developing such a voting model. Most methods in this category require the number of clusters in each base partition to be K , i.e., kg K, g 1 . . . M.

Simple Voting: Based on the analysis of Topchy et al. [70], the underlying re-labelling process is equivalent to the problem of maximum weight bipartite matching. This starts with the creation of a contingency matrix Ω RK ×K from the reference πr and to- be re-labelled πg partitions, where K is the number of clusters in each partition. Each entry Ω(l, l′) that denotes the co-occurrence statistics between labels l ∈ πr and l′ ∈ πg , is defined by

IV. RECENT EXTENSIONS AND APPLICATIONS

Soon after 2010, a large number of research studies have pub- lished new concepts and findings related to several issues of cluster ensemble. Some introduce theoretical improvement and exten- sions to the previous approaches to ensemble generation, repre- sentation and consensus clustering. Others focus on the application side, where existing methods are exploited for real problems and different data-mining tasks. The section is to provide details and a useful insight of these exciting developments.

4.1.1. Ensemble generation

It is known that the goodness of the ensemble decision is highly subjected to both diversity within the ensemble and accuracy of those ensemble members. Also, Fern and Lin [99] have recom- mended to form a smaller but better-performing cluster ensemble with a subset of members, than using all primary alternatives. In addition to the collection of general approaches to ensemble generation discussed in Section 2, this part provides details of more up-to-date methods to reach the aforementioned goal.

Validity index based generation: Similar to the study of Fern and Lin [99], Alizadeh et al. [100] introduce an en- semble clustering framework, which makes use of a subset of initial members in the ensemble, instead of employing all like before. As such, the quality metric of Normalized Mu- tual Information or NMI is exploited for the determination of these target clusterings. Of course, setting an appropri- ate NMI threshold is data dependent and requires domain knowledge. About the same time, Zhang et al. [101] make use of Adjusted Rand Index (ARI) to control the formation of cluster ensemble. In particular, this classical validity metric is generalized to new measures of ARImp and ARImm. The former compares the similarity between base clusterings and the consensus matrix that summarizes the entire en- semble, while the other computes the similarity between any pair of primary partitions. The NMI metric is also ex- ploited by Parvin and Minaei-Bidgoli [102] to determine a good subset of base clusterings, which are initially generated using the weighted locally adaptive clustering (WLAC) al- gorithm. Following that, a new asymmetric criterion named Alizadeh–Parvin–Moshki–Minaei (APMM) has been brought forward as the alternative to NMI to control the process of ensemble selection [103]. Likewise, the measure of cluster stability and dataset simplicity are coupled to assess the quality of subsets of base partitions [104]. Heuristic based generation: One of the recent extensions model the ensemble generation based on a concept called The Wisdom of Crowds [107]. It is a phenomenon founded in social science that suggests criteria applicable to group behaviour. Intuitively, with these criteria being satisfied, the group decisions may often be better than those of individual

4.2.Applications of cluster ensembles

In addition to the extensions elaborated earlier, this section looks into different applications of cluster ensemble, with respect to two viewpoints. The first part explores the applications to spe- cific problem domains such as time series analysis. The second emphasizes the use of cluster ensemble for other data-mining tasks, e.g., transfer learning and classification.

4.2.1. Application to other data mining tasks

Despite the fact that cluster ensemble has been established for unsupervised learning, the method has recently been exploited for other tasks related to data analysis. These include the followings.

Transfer learning: Conventional supervised learning usually assumes that both training and test data are from a common distribution. Thus a challenge arises in transfer learning, where training and test distributions may be mismatched. The problem is even worse when the test data is actually from a different domain and without labels. In order to resolve this, Acharya et al. [214] introduce an optimiza- tion framework, which takes as input one or more classifiers learned on the source domain as well as the results of a cluster ensemble operating solely on the target domain, and yields a consensus labelling of the data in the target domain. Detecting ambiguity in data: Another novel application scheme of cluster ensemble is to identify uncertain or am- biguous regions in the data under examination [215]. Fol- lowing the detection, two approaches have been suggested for the treatment of such uncertainty. Firstly, the simplest way is to ignore ambiguous patterns prior to the consensus clustering, thus preserving the non-ambiguous data as good prototypes for any further modelling. The other alternative is to use the ensemble solution obtained by the first to train a supervised model that is later applied to reallocate the ambiguous clusters.

Dimensionality reduction: With large amounts of data being generated in various domains such as bioinformatics and social networks, dimensionality reduction remains a challenging task for data-mining researchers. The concept of cluster ensemble is recently exploited for this problem with the use of genetic algorithm [216]. Based on the validation with conventional classification methods and benchmark data collections, its performance is promising with the ac- curacy on par with the latest approaches proposed in the literature.

Semi-supervised learning: For the analysis of gene expression data, Wang and Pan [217] introduce semi-supervised consensus clustering (SSCC) that integrate the LCE model [20] with semi-supervised clustering process. The clustering quality can be improved when prior knowledge (in terms of must-link and cannot-link constraints) is provided in addition to a typical proximity metric. This study follows the line of research initially brought about by Yu et al. [218] and Yang et al. [219]. As for the former, a new cluster ensemble method named knowledge based cluster ensemble (KCE) is proposed where prior knowledge of data is included into the cluster ensemble framework. Specific to this, pairwise constraints among data points are encoded as confidence factors between base clusterings. Later, these will be concluded in the form of consensus ma- trix from which the final result is generated. In the latter, an improved Cop-Kmeans (ICop-Kmeans) algorithm has been put forward to tackle the violation of pairwise constraints usually encountered with the original Cop-Kmeans model. Likewise, Zhang et al. [220] and Yu et al. [221] contribute to this subject by proposing the semi-supervised cluster- ing ensemble model based on collaborative training (SCET) and the incremental semi-supervised clustering ensemble framework (ISSCE), respectively. with data size. This is achieved through calculating spectral embedding of data with cluster-centre based representa- tion. Analysing the big data has become a major challenge, especially to those web-based organizations such as Google and Facebook. They commonly develop a customized vari- ation of non-relational systems not only to overcome the limitations of efficient storage and retrieval, but also pave the way for data analytics [235]. Some of the new ap- proaches to analysing big data gain a great deal of at- tention amongst commercial and academic researchers, e.g., Google's MapReduce framework, Hadoop and Hive. Ac- cording to the report of Dean and Ghemawat [236], MapRe- duce has been the most popular solution for parallel and analysis of large amount of data. Within the community of data mining, implementations of several techniques using MapReduce have been presented in the past few years. For instance, Liu et al. [237] introduce a MapReduce based parallel back-propagation neural network (MR-BPNN). As for data clustering, a MapReduce-based artificial bee colony (MR-ABC) is developed for a clustering method similar to k- means [238]. This ABC implementation helps to optimize the assignment of the large data objects to clusters. However, for cluster ensemble, such an implementation has rarely been reported in the literature. In fact, one recent publication kicks off this research direction, with the introduction of a new parallel k-means clustering based on MapReduce framework for aspect based summary generation [239]. Of course, an opportunity to coupling existing cluster ensemble methods with MapReduce or other big-data platforms is obvious. This may further boost its application that is in line with the new challenges encountered by big data scientists. Repository of tools: Ever since its introduction in the early 2000s, the scope of end users of cluster ensemble or consensus clustering is rather limited. As compared to conventional clustering algorithms like k-means or DBSCAN that are available in several well-known data mining tools (e.g., Weka2 and RapidMiner3), implementations of those ensemble models appear to be harder to obtain. Most of them are provided as a supplementary to the publication, which can disappear over time. Yet, this is typically not user friendly as it has been customized in a specific programming

environment. As a result, it is also significant to make this family of methods known to a wider public, perhaps as an extension to the well-established tools. This may help broaden the application domain to cover interesting problems in the new era of data intensive industry and society.

4.2. Applications of cluster ensembles

In addition to the extensions elaborated earlier, this section looks into different applications of cluster ensemble, with respect to two viewpoints. The first part explores the applications to spe- cific problem domains such as time series analysis. The second emphasizes the use of cluster ensemble for other data-mining tasks, e.g., transfer learning and classification.

4.2.2. Application to other data mining tasks

Despite the fact that cluster ensemble has been established for unsupervised learning, the method has recently been exploited for other tasks related to data analysis. These include the followings.

V. CHALLENGES AND CONCLUSION

This survey has presented classical and recently developed ap- proaches to cluster ensemble. It kicks off with the formal termi- nology by which the problem is defined. Four basic categories of consensus clustering methods are then discussed in depth with illustrative examples. After that, it provides details of extensions to three main components of a cluster ensemble framework: ensem- ble generation, representation and summarization, and consensus function, respectively. Given the superior capability to deliver accurate data partitions, many cluster ensemble techniques have been exploited for a wide range of applications and domain prob- lems. In addition, the use of this meta-learning approach for other data-mining tasks such as classification has been studied. The at- tention received by this subject has consistently increased over the years, especially after 2010 that is the focus of this survey. Based on the statistics shown in Fig. 4, the numbers of Google scholar search results for ‘‘cluster ensemble’’ or ‘‘consensus clustering’’ are 1240, 1660 and 2800 for the periods of 2011–12, 2013–14 and 2015– 16, respectively. It is clearly illustrated that these counts are much higher than those belonging to the intervals before 2010.

The aforementioned observation is greatly due to the maturity of basic practice to cluster ensemble and a flourish of its applica- tions. From the early period with most of the works relating only to bioinformatics, especially microarray data analysis, the application landscape has largely expanded over the past few years. It covers several new challenges to the modern age such as cybersecurity and time-series data analysis. The followings summarize potential challenges of cluster ensemble in the near future.

Heterogeneous data analysis: Despite the long history of development, most of cluster ensemble methods have been directed to numerical data analysis. Only a handful of publications report findings with other types of data. Specific to the work of Iam-On et al. [227], the link-based method or LCE is adopted for the clustering of biological samples. Each of these can be expressed by both continuous variables extracted from microarray data, and nominal variables obtained from clinical or pathological data of the samples under examination. This so-called integrative approach to biological data analysis has shown to improve the accuracy of prognostic outcome, as compared to those obtained by us- ing one of the aforementioned factors alone. However, given the fact that the aforementioned model is based simply on k-prototype algorithm, its performance is highly subjected to parameter setting (i.e., weights given to continuous and nominal variables).

A gap of improvement in terms of clustering quality and model robustness exists especially for implementing new inventions of mixed-type data clustering in the ensemble context. For instance, Blomstedt et al. [228] recently introduce a model-based algorithm for clustering attributes of mixed type, which is based a Bayesian predictive framework. Provided that clustering solutions represent random data partitions, the posterior probability for a partition can be determined using conjugate analysis. Another approach applies unsupervised feature learning (UFL) to mixed-type data in order to acquire a sparse representation. As a result, it becomes easier for clustering algorithms to disclose data partitions [229]. While conventional UFL techniques are designed for homogeneous data, the aforementioned works with the mixed-type data using fuzzy adaptive resonance theory (ART). In the biomedical domain, Abidin and Westhead [230] also point out the need for accurate cluster analysis of mixed type data. This commonly appears as a mixture of binary or nominal data (e.g. presence of mutations, binding and epigenetic marks) and continuous data (e.g. gene expression and metabolite levels). As such, a generic clustering method is proposed and evaluated with genetic regulation and the clustering of cancer samples.

Big data analysis: Common applications on office and social based platforms have facilitated the vast amount of data being generated daily. Analysing this so-called big data has been a major trend and challenge within the community of data mining. To better appreciate this, see the comparative study of Fahad et al. [231], where several classical clustering techniques are assessed against big datasets. In particular to an ensemble model, it may face the problem of scaling up, despite the quality it produces. In response, Su et al.

[232] introduce a novel cluster ensemble approach for fuzzy clustering, especially for big data. It first builds fuzzy base clusters with respect to each data feature. Then, it makes use of a fuzzy hierarchical graph to represent relationships between those base clusters. Based on this representation scheme, the final result is generated using hierarchical clus- tering as the consensus function. This work follows an initial attempt to mitigate the practice of cluster ensemble to large data [233]. In that, ECCA (Ensemble of Combined Cluster- ing Algorithms) is invented as a framework of ensemble of algorithms with fixed uniform grids. The final collective solution is based on pairwise classification of the elements of the grid structure. Another study attempts to deal with the curse of dimensionality in big data, especially for cluster ensemble [234]. In particular, a new fuzzy c-means (FCM) algorithm with random projection has been created as the basis of novel consensus clustering, which scales linearly Analysing the big data has become a major challenge, especially to those web-based organizations such as Google and Facebook. They commonly develop a customized vari- ation of non-relational systems not only to overcome the limitations of efficient storage and retrieval, but also pave the way for data analytics [235]. Some of the new ap- proaches to analysing big data gain a great deal of at- tention amongst commercial and academic researchers, e.g., Google’s MapReduce framework, Hadoop and Hive. Ac- cording to the report of Dean and Ghemawat [236], MapRe- duce has been the most popular solution for parallel and analysis of large amount of data. Within the community of data mining, implementations of several techniques using MapReduce have been presented in the past few years. For instance, Liu et al. [237] introduce a MapReduce based parallel back-propagation neural network (MR-BPNN). As for data clustering, a MapReduce-based artificial bee colony (MR-ABC) is developed for a clustering method similar to k-means [238]. This ABC implementation helps to optimize the assignment of the large data objects to clusters. However, for cluster ensemble, such an implementation has rarely been reported in the literature. In fact, one recent publication kicks off this research direction, with the introduction of a new parallel k-means clustering based on MapReduce framework for aspect based summary generation [239]. Of course, an opportunity to coupling existing cluster ensemble methods with MapReduce or other big-data platforms is obvious. This may further boost its application that is in line with the new challenges encountered by big data scientists. Repository of tools: Ever since its introduction in the early 2000s, the scope of end users of cluster ensemble or con- sensus clustering is rather limited. As compared to con- ventional clustering algorithms like k-means or DBSCAN that are available in several well-known data mining tools (e.g., Weka2 and RapidMiner3), implementations of those ensemble models appear to be harder to obtain. Most of them are provided as a supplementary to the publication, which can disappear over time. Yet, this is typically not user friendly as it has been customized in a specific programming

environment. As a result, it is also significant to make this family of methods known to a wider public, perhaps as an extension to the well-established tools. This may help broaden the application domain to cover interesting problems in the new era of data intensive industry and society.

REFERENCES

[1] D. Jiang, C. Tang, A. Zhang, Cluster analysis for gene expression data: A survey, IEEE Trans. Knowl. Data Eng. 16 (2004) 1370–1386.

[2] R.C. Wu, R.S. Chen, C.C. Chang, J.Y. Chen, Data mining application in customer relationship management of credit card business, in: Proceedings of interna- tional conference on Computer software and applications, 2005, pp. 39–40.

[3] S.K. Bhatia, J.S. Deogun, Conceptual clustering in information retrieval, IEEE Trans. Syst. Man Cybern. 28 (1998) 427–436.

[4] J. Zhang, J. Mostafa, H. Tripathy, Information retrieval by semantic analysis and visualisation of the concept space of D-Lib magazine, D-Lib Mag. 8 (2002).

[5] J.A.F. Costa, M. de Andrade Netto, Cluster analysis using self-organising maps and image processing techniques, Proc. IEEE Int. Conf. Syst. Man Cybern. 5 (1999) 367–372.

[6] H. Tao, T.S. Huang, Color image edge detection using cluster analysis, in: Proceedings of IEEE International Conference on Image Processing, 1997, pp. 834–836.

[7] G.S. Day, R.M. Heeler, Using cluster analysis to improve marketing experi- ments, J. Market. Res. 8 (1971) 340–347.

[8] A.G. Sheppard, The sequence of factor analysis and cluster analysis: Differ- ences in segmentation and dimensionality through the use of raw and factor scores, Tourism Anal. 1 (1996) 49–57.

[9] D.B. Henry, P.H. Tolan, D. Gorman-Smith, Cluster analysis in family psychol- ogy research, J. Family

Psychol. 19 (2005) 121–132.

[10] K. Kim, H. Ahn, A recommender system using GA K-means clustering in an online shopping market, Expert Syst. Appl. 34 (2008) 1200–1209.

[11] M. Bredel, C. Bredel, D. Juric, G. Harsh, H. Vogel, L. Recht, B. Sikic, Functional network analysis reveals extended gliomagenesis pathway maps and three novel MYC-interacting genes in human gliomas, Cancer Res. 65 (2005) 8679– 8689.

[12] E. Kim, S. Kim, D. Ashlock, D. Nam, MULTI-K: Accurate classification of microarray subtypes using ensemble k-means clustering, BMC Bioinform. 10 (2009) 260.

[13] T. Sorlie, R. Tibshirani, J. Parker, T. Hastie, J. Marron, A. Nobel, S. Deng, H. Johnsen, R. Pesich, S. Geisler, Repeated observation of breast tumor subtypes in independent gene expression data sets, Proc. Natl. Acad. Sci. USA 100 (2003) 8418–8423.

[14] A.K. Jain, M.N. Murty, P.J. Flynn, Data clustering: A review, ACM Comput. Survey 31 (1999) 264–323.

[15] A. Ahmad, L. Dey, A k-mean clustering algorithm for mixed numeric and categorical data, Data Knowl. Eng. 63 (2007) 503–527.

[16] Z. Huang, Claustering large data sets with mixed numeric and categorical values, in: Proceedings of the First Pacific Asia Knowledge Discovery and Data Mining Conference, 1997, pp. 21–34.

[17] S. Dudoit, J. Fridyand, A prediction-based resampling method for estimating the number of clusters in a dataset, Genome Biol. 3 (2002) RESEARCH0036.

[18] T. Boongoen, Q. Shen, Nearest-neighbour guided evaluation of data reliability and its applications, IEEE Trans. Syst. Man Cybern. B 40 (2010) 1622–1633.

[19] W.M. Rand, Objective criteria for the evaluation of clustering methods, J. Amer. Statist. Assoc. 66 (1971) 846–850.

[20] N. Iam-On, T. Boongoen, S. Garrett, LCE: A link-based cluster ensemble method for improved gene expression data analysis, Bioinformatics 26 (2010) 1513–1519.