Computer Science Review 28 (2018) 1–25

Contents lists available at ScienceDirect

Computer Science Review

journal homepage: www.elsevier.com/locate/cosrev

Review article

Cluster ensembles: A survey of approaches with recent extensions and applications

Tossapon Boongoen, Natthakan Iam-On *

IQ-D Research Unit, School of Information Technology, Mae Fah Luang University, Tasud, Muang District, Chiang Rai 57100, Thailand

ARTICLE INFO

Article history:
Received 10 April 2017
Received in revised form 26 December 2017
Accepted 29 January 2018

Keywords:
Data clustering
Cluster ensemble
Theoretical extension
Domain specific application

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.

Contents

1. Introduction

Cluster analysis is usually employed in the initial stage of understanding a raw data, especially for new problems where prior

* Corresponding author.
E-mail addresses: tossapon.boo@mfu.ac.th (T. Boongoen), natthakan@mfu.ac.th (N. Iam-On).

https://doi.org/10.1016/j.cosrev.2018.01.003
1574-0137/

knowledge is minimal. Also, in the pre-processing stage of supervised 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], information 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 consensus 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].

Specific to this survey, the practice of cluster ensembles is motivated by the fact that the performance of most clustering techniques are highly data dependent. A particular clustering model may produce an acceptable result for one dataset, but possibly become ineffective for others. Generally, there are two major challenges inherent to clustering algorithms. First, different techniques discover different structures (e.g., cluster size and shape) from the same set of data objects [21–23]. For example, *k*-means that is probably the best known technique is suitable for spherical-shape clusters, while single-linkage hierarchical clustering is effective to detect connected patterns. This is due to the fact that each individual algorithm is designed to optimize a specific criterion. Second, a single clustering algorithm with different parameter settings can also reveal various structures on the same dataset. A specific setting may be good for a few, but not all datasets. Users encounter these challenges, which consequently make the selection of a proper clustering technique very difficult.

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.

The rest of this survey is organized as follows. To set the scene for concepts and discussion presented here, Section 2 introduces the basis of cluster ensembles, including formal definition, framework and different ensemble generation strategies. Then, four major approaches to find a consensus clustering are illustrated in Section 3. In addition, Section 4 provides applications and recent theoretical extensions of those cluster ensemble techniques, especially the use of ensemble information as a data transformation approach for classification task. The survey is concluded in Section 5 with future research directions.

2. 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 conventional 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.

2.1. Data clustering and conventional techniques

Data clustering is one of the fundamental and effective tools for understanding the structure of a given dataset. It plays a crucial, foundational role in machine learning, data mining, information retrieval and pattern recognition. Clustering aims to categorize data into groups or clusters such that the data in the same cluster are more similar to each other than to those in different clusters. Similarity or proximity is measured using the attribute values that represent objects (data points) in the dataset [14]. Clustering is branded an unsupervised learning approach as the measurement of similarity is conducted without knowledge of class assignment. This knowledge-free scenario brings about a series of difficult decisions, hence the corresponding research studies, with respect to selecting appropriate algorithm, similarity measure, criterion function, and initial parameter condition [21,23]. Clustering is widely recognized as an ideal candidate for research and development [25], given its benefits and possible advances to be made in this field. There are a large number of clustering algorithms developed in the literature. Examples of well-known techniques are explained in this section.

***k*-means** is perhaps, the best known clustering technique that partitions data points into clusters. Its name comes from representing each of k clusters by the mean of its members or so-called 'centroid'. k-means is an iterative algorithm that exploits a square-error as a criterion function (i.e., the total distance between each data point and its cluster centre, [26]). It begins with initializing centroids randomly and then allocates data points to clusters such that the square-error is minimized. This criterion function tends to work well with separated and compact clusters. Given a dataset X, the square-error e^2 of a clustering $\pi = \{C_i, \ldots, C_k\}$ with k clusters is defined as

$$e^2(X, \pi) = \sum_{p=1}^{k} \sum_{\forall x \in C_p} \|x - \overline{c_p}\|^2, \tag{1}$$

where $\|.\|$ denotes the Euclidean norm and $\overline{c_p}$ is the centre of the pth cluster. A general description of the k-means algorithm is given as follows:

1. k data points are first randomly selected as initial cluster centres.
2. Repeat:
   (a) Assign each data point to its closest cluster centre. The Euclidean metric is commonly used to compute the distance between data points and centroids.