Information Sciences 393 (2017) 29–46

Contents lists available at ScienceDirect

Information Sciences

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

Using the stability of objects to determine the number of clusters in datasets

Etienne Lord[a,b], Matthieu Willems[a], François-Joseph Lapointe[b], Vladimir Makarenkov[a,*]

[a] *Département d'informatique, Université du Québec à Montréal, C.P. 8888, Succ. Centre-Ville, Montréal (QC) H3C 3P8 Canada*
[b] *Département de sciences biologiques, Université de Montréal, C.P. 6128, Succ. Centre-Ville, Montréal (QC) H3C 3J7 Canada*

ARTICLE INFO

Article history:
Received 21 February 2016
Revised 13 January 2017
Accepted 4 February 2017
Available online 6 February 2017

Keywords:
Cluster stability
Cluster validity indices
K-means
K-medoids
Number of clusters
Partitioning algorithms
Stability of individual objects in clustering

ABSTRACT

We introduce a novel method for assessing the robustness of clusters found by partitioning algorithms. First, we show how the stability of individual objects can be estimated based on repeated runs of the *K*-means and *K*-medoids algorithms. The quality of the resulting clusterings, expressed by the popular Calinski–Harabasz, Silhouette, Dunn and Davies–Bouldin cluster validity indices, is taken into account when computing the stability estimates of individual objects. Second, we explain how to assess the stability of individual clusters of objects and sets of clusters that are found by partitioning algorithms. Finally, we present a new and effective stability-based algorithm that improves the ability of traditional partitioning methods to determine the number of clusters in datasets. We compare our algorithm to some well-known cluster identification techniques, including *X*-means, Pvclust, Adegenet, Prediction Strength and Nselectboot. Our experiments with synthetic and benchmark data demonstrate the effectiveness of the proposed algorithm in different practical situations. The R package *ClusterStability* has been developed to provide applied researchers with new stability estimation tools presented in this paper. It is freely distributed through the Comprehensive R Archive Network (CRAN) and available at: https://cran.r-project.org/web/packages/ClusterStability.

1. Introduction

Clustering algorithms have been successively applied in many fields, including banking, bioinformatics, computer vision, marketing, and security, in order to extract the structure from a given dataset and to gain insight into its natural clusters [14,35]. There are two main clustering approaches that encompass hierarchical clustering and partitioning algorithms. In this article, we focus on the techniques for partitioning N objects into K clusters according to a specific similarity criterion. The total number of partitions of N objects into K non-empty and non-overlapping clusters is asymptotically equivalent to $K^N/K!$, as N tends to infinity [40]. Thus, heuristic algorithms such as K-means [29] and K-medoids [23] have been proposed to limit the number of possible solutions when searching for an optimal partition of objects. These heuristic algorithms are often preferred to more complex alternatives because of their simplicity and relatively good performances [41], as well as because of the availability of their parallel versions, which are scalable to large recognition problems [50]. Despite their popularity,

[*] Corresponding author.
E-mail addresses: lord.etienne@courrier.uqam.ca (E. Lord), willems.matthieu@courrier.uqam.ca (M. Willems), francois-joseph.lapointe@umontreal.ca (F.-J. Lapointe), makarenkov.vladimir@uqam.ca (V. Makarenkov).

http://dx.doi.org/10.1016/j.ins.2017.02.010
0020-0255/© 2017 Elsevier Inc. All rights reserved.

K-means and *K*-medoids usually provide solutions that are only local optima [35,41]. Moreover, these algorithms highly depend on the number of random starts [41], and the choice of starting partitions is crucial for them [32,41,43]. In addition, the *K*-means algorithm is very sensitive to the presence of noisy features in the data [18,19]. Usually, several hundred starts of *K*-means with different input random partitions are required in order to select an appropriate clustering [41]. Finally, like many partitioning algorithms, *K*-means and *K*-medoids also suffer from the need to specify the desired number of clusters [29,35].

Recently, there has been a renewed interest in assessing the robustness of clustering solutions that are provided by partitioning algorithms [16,28,43]. Furthermore, alternative methods, such as model-based evaluation [10] or bootstrapping [18,19], have been proposed to assess the reproducibility of clusterings. Despite this increased attention, the intriguing and challenging problem of estimating the stability of individual objects in clustering has not been fully addressed in the literature [25,28].

In this paper, we define a novel measure for assessing the stability of individual objects (i.e., individual *ST*-index) in clustering solutions provided by partitioning algorithms, based on their repeated runs. We also propose a cluster stability index, which reflects the stability of clusters, and a global stability index (i.e., global *ST*-index), which characterizes the robustness of entire clusterings (i.e., resulting partitions or clustering solutions found by partitioning algorithms). These new indices can help practitioners decide which individual objects and clusters should be kept in the dataset and which of them should be removed from it in order to improve the stability of a given clustering. Moreover, the results of our simulation study indicate that the stability of a clustering estimated by our stability indices is directly related to its quality, and that the global *ST*-index can be effectively used to improve the ability of traditional clustering algorithms to determine the true number of clusters in datasets. Our R package *ClusterStability* provides researchers with the new stability estimation tools that we describe in this paper.

2. Background and related work

2.1. Cluster validity indices

A variety of cluster validity indices are available to determine the number of clusters in a given dataset [2,10,34]. They can be defined as measures of partitioning quality. Most of these indices take into consideration the compactness of the objects in the same cluster and their separation in the distinct clusters [2,34]. In our experiments with real and synthetic data, we will use the Calinski–Harabasz [11], Silhouette [39], Dunn [9] and Davies–Bouldin [13,24] measures, which have been among the most recommended cluster validity indices according to several simulation studies [2,10,34].

The Calinski–Harabasz index is a normalized ratio of the overall inter-cluster variance and the overall intra-cluster variance [11]. The Silhouette width [39] of an individual object i is defined using its average intra-cluster distance, $a(i)$, and its average nearest-cluster distance, $b(i)$. It is calculated as follows: $(b(i) - a(i))/\max(a(i), b(i))$. The global Silhouette width is defined as the average of the individual Silhouette widths of all the objects. The Dunn index is a ratio-type coefficient in which the cluster separation is expressed through the maximum cluster diameter, and the cluster cohesion is expressed through the nearest neighbor distance [9]. While there are various versions of the Dunn coefficient, the most reliable are the generalized Dunn's indices [9]. The Davies–Bouldin index is also based on a ratio of intra-cluster and inter-cluster distances [13]. For a pair of clusters (C_1, C_2), the pairwise cluster distance $db(C_1, C_2)$ is first calculated as the sum of the average distances between the objects and centroids in both clusters, which is then divided by the distance between the cluster centroids. The Davies–Bouldin index is defined as the average of the largest $db(C_k, C_l)$'s $(l \neq k)$, computed over all available clusters C_k. An improved variant of this coefficient proposed by Kim and Ramakrishna [24] provides very good cluster recovery performances according to a recent comparative study of cluster validity indices conducted by Arbelaitz et al. [2].

2.2. Stability of clustering solution

A number of theoretical and empirical studies have addressed the problem of solution stability in clustering [6,16,18,19,26,33,41-43]. Milligan and Cheng [33] were first to investigate how the addition and removal of objects influence the quality of the resulting clusterings. Ben-Hur et al. [6] proposed to use, as a measure of cluster stability, the distribution of pairwise similarities between partitions obtained from clustering sub-samples of a given dataset. In order to determine the true number of clusters in a dataset, the authors suggested examining the clusters in which a transition from a stable to an unstable clustering state can occur. Lange et al. [26] introduced a measure that quantifies the reproducibility of clustering solutions and defined a function that minimizes the risk of misclassification. Ben-David et al. [4] provided a formal definition of cluster stability and concluded that, for large datasets, cluster stability is closely related to the behavior of the objective function of a given clustering algorithm.

Hennig [18,19] discussed several strategies for assessing the support of individual clusters in a clustering solution. One of these strategies relies on the use of the Jaccard coefficient and resampling techniques such as bootstrapping, jittering, and subsetting [18]. Hennig [19] showed how to determine the dissolution point and the isolation robustness of a cluster by adding to it new objects and outliers. Fang and Wang [16] proposed a different variant of data bootstrapping that allows for selecting the number of clusters in a dataset by examining randomness in the samples. While several papers discuss the stability of clustering methods with respect to changes in a given dataset, the work of de Mulder [37] focuses on cluster