Pattern Recognition 48 (2015) 1478–1489

Contents lists available at ScienceDirect

Pattern Recognition

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

Cluster validity measure and merging system for hierarchical clustering considering outliers

Frank de Morsier [a,1], Devis Tuia [b], Maurice Borgeaud [c], Volker Gass [d], Jean-Philippe Thiran [a]

[a] École Polytechnique Fédérale de Lausanne (EPFL), Signal Processing Laboratory (LTS5), Switzerland
[b] École Polytechnique Fédérale de Lausanne (EPFL), Laboratoire des Systèmes d'Information Géographique (LaSIG), Switzerland
[c] European Space Agency, ESRIN, Frascati, Italy
[d] École Polytechnique Fédérale de Lausanne (EPFL), Swiss Space Center, Switzerland

ARTICLE INFO

Article history:
Received 18 October 2013
Received in revised form
20 June 2014
Accepted 1 October 2014
Available online 14 October 2014

Keywords:
Clustering
Quality
Agglomerative clustering
Gaussian processes
Support vector clustering

ABSTRACT

Clustering algorithms have evolved to handle more and more complex structures. However, the measures that allow to qualify the quality of such clustering partitions are rare and have been developed only for specific algorithms. In this work, we propose a new cluster validity measure (CVM) to quantify the clustering performance of hierarchical algorithms that handle overlapping clusters of any shape and in the presence of outliers. This work also introduces a cluster merging system (CMS) to group clusters that share outliers. When located in regions of cluster overlap, these outliers may be issued by a mixture of nearby cores. The proposed CVM and CMS are applied to hierarchical extensions of the Support Vector and Gaussian Process Clustering algorithms both in synthetic and real experiments. These results show that the proposed metrics help to select the appropriate level of hierarchy and the appropriate hyperparameters.

1. Introduction

The problem of clustering, which consists of separating a dataset into different groups by optimizing a clustering criterion, has been widely studied in various fields where semantic information of the data is not available. Applications of clustering are broad and include grouping sequence of genes in molecular biology, segmentation of images, classification of diseases in medicine, and unsupervised classification of patterns.

Existing clustering algorithms are very diverse and can produce different solutions depending on the clustering criterion considered such as the number of samples used for learning or the choice of the free parameters involved (e.g., the number of clusters k in the k-means or the pruning level in the Ward hierarchical tree [1]). In order to optimally select these free parameters for a given application, it is beneficial to assess the performance of a clustering algorithm in an unsupervised way. This work addresses this need by proposing a validity criterion, which relies solely on the data.

Measures of clustering quality are usually based on the assumption that samples should be more similar to those inside a cluster than to those assigned to different clusters. Since the similarity and dissimilarity measures can be defined in various ways, there exists a large number of validity measures that we review in Section 2. However, most of these quality measures are valid only under specific assumptions about the data [2] and are not able to handle overlapping clusters of arbitrary shapes.

Recent developments in clustering algorithms have focused on handling arbitrary cluster shapes based on different types of criteria [3]: non-linear distances with the kernel k-means [4], neural-networks [5], Bregman distances [6] or graph-based algorithms [7–9], hierarchical representation with the agglomerative algorithms [10–13], based on density with DBScan [14,15], OPTICS [16], CHAMELEON [17], DenClue [18] or the mean shift algorithm [19]. Another main trend in clustering research is the detection or rejection of outliers. Algorithms such as CURE [20], ROCK [21], FLAME [22] or MITOSIS [23] are robust to outliers. However, they are not appropriate for overlapping clusters where the mixed samples between them are to be considered as outliers. Algorithms such as the support vector clustering (SVC) [24–26] and the Gaussian processes clustering [27] provide solutions with different levels of outliers rejection: they allow focusing on the part of the clusters not overlapped and ignoring the overlapped regions. The appropriate level of outlier rejection and the other hyperparameters are set based on heuristics, which are often difficult to tune and specific to the selected algorithm. These algorithms would strongly benefit from an unified cluster validity measure replacing these heuristics.

These validity measures are not taking into account the user needs and may provide clustering solutions far from user expectations.

E-mail address: frank.demorsier@gmail.com (F. de Morsier).
URL: http://lts5www.epfl.ch/ (F. de Morsier).
[1] Tel.: +41 21 693 26 01; fax: +41 21 693 76 00.

http://dx.doi.org/10.1016/j.patcog.2014.10.003
0031-3203/© 2014 Elsevier Ltd. All rights reserved.

The difference in the user needs and the nature of the data may become problematic and should be addressed by a merging criterion. As an example, the classes of interest of a user may be groups of several overlapping clusters difficult to retrieve directly by a clustering algorithm. The ideal algorithm should be able to find all the clusters and group the ones that overlap by merging them together. In [25], SVC clustering solutions are merged based on cluster overlap projected on the different dimensions. A Gaussian distribution is fitted to each cluster which limits its use to low dimensional data and simple cluster shapes. A non-parametric merging system would be needed to handle clusters of arbitrary shapes.

In the light of these shortcomings, we introduce here: (i) a cluster validity measure, which is valid for any clustering algorithm and provides a hierarchy of outlier rejection, (ii) A criterion for merging clusters based on the outlier structure, which allows to merge clusters into potential classes (natural groups of clusters). These two contributions are applied and tested on truly hierarchical clustering algorithms derived from support vector clustering [24,26] and Gaussian processes clustering [27], which are appropriate for our cluster validity measure and merging criterion.

The rest of the paper is organized as follows. Section 2 reviews the existing cluster validity measures and their limitations. Section 3 exposes our assumptions on the data structure. The proposed cluster validity measure (CVM) is presented in Section 4 and the cluster merging system (CMS) in Section 5. Finally, experiments on synthetic and real datasets are presented and discussed in Section 6.

2. Cluster validity measures

The Cluster validity measures (CVMs) or indices can be separated into two main categories: external and internal CVMs. The external CVMs, such as the Rand index or the Jaccard Coefficient [28], aim at comparing a clustering result with a pre-determined clustering partition, usually referred as the *golden* or *groundtruth* partition. However, in unsupervised settings a golden partition is usually unavailable. The internal CVMs aim at qualifying a partition based solely on the dataset and the cluster labels [29]. These CVMs have evolved together with the clustering algorithms: first CVMs assume spherical or elliptical clusters, while recent CVMs allow to cope with arbitrarily shaped clusters, even in the presence of outliers.

Most CVMs assume that clusters should be as compact and separated as possible. This allowed the development of validity measures, such as the *Davies–Bouldin* (*DB*) index [30], the *Xie–Beni* index [31] or Partition Coefficient (PC) [32] for fuzzy clustering, which exploit both the within cluster and the between cluster scattering. All these measures assume clusters as multivariate Gaussian distributions and can therefore only describe spherically shaped clusters.

Measures that allow arbitrarily shaped clusters have also been introduced: the Dunn index [33] or measures based on graphs [34]. Nonetheless, these CVMs are not particularly robust to outliers. More robust versions have been introduced to handle noise [35] and avoid solutions with an important number of small clusters [36–40], or overlapping cluster [41,42]. However, none of these CVMs consider clusters either overlapping or of arbitrary shapes. Some attempts to handle both have been pursued in the literature, either by first removing the overlap regions and then applying kernel *k*-means with the hyperparameters tuned by a *Fisher*-like criterion in the feature space [43]; or by a pseudo-hierarchical SVC algorithm that selects its optimal solutions based on a re-weighted inner scattering matrix [44]; or by selecting SVC hyperparameters based on heuristics related to the data [45,25].

Most CVMs in the literature cannot handle properly clusters that overlap without having recourse to algorithm-dependent heuristics. Therefore, we introduce in Section 4 a general CVM for hierarchical agglomerative clustering algorithms, which favours large homogeneous clusters separated by either empty regions or regions of outliers (due to overlaps).

3. Clustering with outlier hierarchy

In this section, we first present the concepts used to describe a dataset divided into clusters, so as to prepare the reader to understand the details about the validity measure and merging systems proposed in Sections 4 and 5, respectively.

Most of the data subject to clustering is composed of samples representative of the clusters and of additional samples being a mixture (often near-linear) of several clusters. This second type of samples corresponds to borderline patients in biomedical data, to pixels representing a mixture of sources (e.g. water+land) in remote sensing images, or to images containing several objects in problems of image categorization. Most of the methods that tackle this type of problems fall in the category of “unmixing techniques”, which try to find *pure* data samples by assuming linear or non-linear mixing of a certain number of sources [46]. Fig. 1 represents the typical “unmixing” setting and the more general one of clusters with mixtures.

In the specific problematic of remote sensing imagery, unmixing resorts to finding the materials that compose each pixels, where a pure pixel, representing a single material, is called an

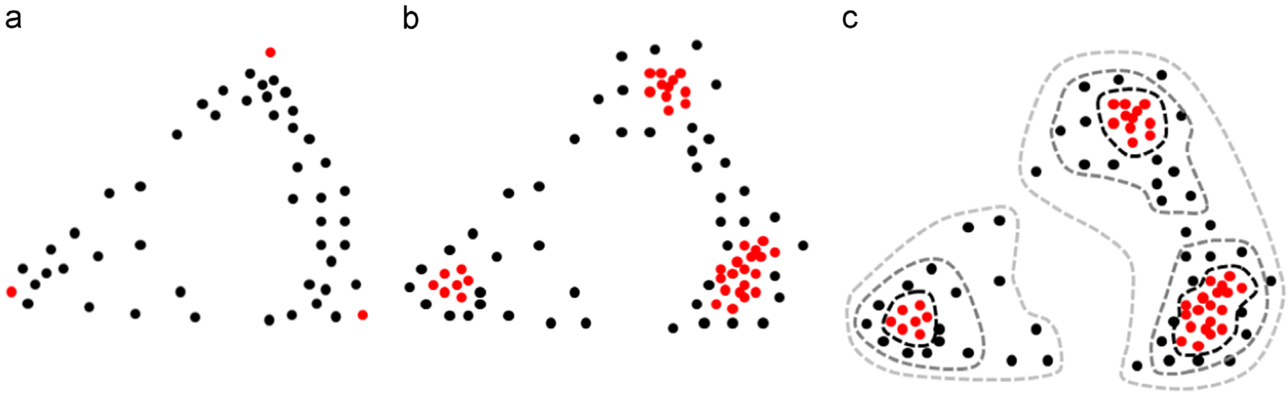

Fig. 1. Two-dimensional examples of (a) standard unmixing problem with 3 endmembers (red dots) and mixed samples (black dots), (b) the more general case with 3 clusters (red dots) surrounded by noisy and mixed samples (black dots) and (c) the corresponding outlier hierarchy levels. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this article.)