Computational Statistics and Data Analysis 73 (2014) 27–39

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

A non-parametric method to estimate the number of clusters

André Fujita [a,*], Daniel Y. Takahashi [b], Alexandre G. Patriota [c]

[a] Department of Computer Science, Institute of Mathematics and Statistics, University of São Paulo, Brazil
[b] Department of Psychology and Neuroscience Institute, Green Hall, Princeton University, USA
[c] Department of Statistics, Institute of Mathematics and Statistics, University of São Paulo, Brazil

ARTICLE INFO

Article history:
Received 6 February 2013
Received in revised form 19 November 2013
Accepted 20 November 2013
Available online 4 December 2013

Keywords:
Clustering
Silhouette method
k-means
Spectral clustering

ABSTRACT

An important and yet unsolved problem in unsupervised data clustering is how to determine the number of clusters. The proposed slope statistic is a non-parametric and data driven approach for estimating the number of clusters in a dataset. This technique uses the output of any clustering algorithm and identifies the maximum number of groups that breaks down the structure of the dataset. Intensive Monte Carlo simulation studies show that the slope statistic outperforms (for the considered examples) some popular methods that have been proposed in the literature. Applications in graph clustering, in iris and breast cancer datasets are shown.

1. Introduction

Cluster analyses are methods of classifying "similar" elements into clusters or groups. They are applied in a wide range of areas such as machine learning, pattern recognition, image analysis and bioinformatics. Several clustering methods have been proposed, namely, *k*-means, hierarchical clustering, expectation–maximization clustering, spectral clustering, and many others. By using clustering techniques, one important task is to estimate the proper number of clusters in actual datasets. For example, in cancer data analysis, the grade of tumorigenesis is determined by geometrical parameters such as the cell's shape, density, etc., and the estimation of the number of clusters using these characteristics is important to correctly classify the patients that will receive different treatments depending on the grade of the tumor. In neuroscience, functional magnetic resonance imaging (fMRI) data is clustered and the number of clusters is estimated in order to identify the cortical areas that are activated in a determined cognitive task (Sato et al., 2007). In machine learning and pattern recognition, the estimation of the number of clusters is important in image segmentation in order to identify different objects (Xiang and Gong, 2008), in real-time monitoring network to recognize emerging behavior of a physical system (Zang and Chen, 2010) and in the detection of the number of distinct facial poses under varying illuminations (He et al., 2010).

Although there are several proposals to determine the number of clusters, it is yet an unsolved and difficult problem due to the absence of a clear definition of cluster and especially because it is dependent on both the adopted clustering method and the characteristics of the data distribution (shape and scale, for instance). One difficulty for the majority of the methods is to correctly classify the dataset when the data points inside the same cluster are correlated or are not Gaussian, in high dimensional situations or when there is a dominant cluster (Sugar and James, 2003; Yin et al., 2008). In this paper we propose the slope statistic, a non-parametric and data-driven method for determining the number of clusters in a dataset. The slope statistic is free of reference distributions, has an intuitive interpretation and does not require

* Correspondence to: Rua do Matão, 1010 - Building C, Cidade Universitária São Paulo, SP, CEP 05508-090, Brazil. Tel.: +55 11 3091 5177.
E-mail address: andrefujita@gmail.com (A. Fujita).

0167-9473/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.csda.2013.11.012

intensive computations. Furthermore, it can handle situations when the dataset is not a mixture of Gaussian distributions, when there exists a dominant cluster and correlation in the dataset, and when the number of parameters is large. Our proposal is an extension of the silhouette method introduced by Rousseeuw (1987).

In intensive Monte Carlo simulations for determining the number of clusters on artificial datasets, we compare the proposed slope statistic to other seven methods: (a) Bayesian Information Criterion (BIC) (Celeux and Govaert, 1992) for a mixture of Gaussian distributions, (b) the Calinski and Harabasz (CH) index Calinski and Harabasz (1974), (c) the Krzanowski and Lai (KL) index Krzanowski and Lai (1985), (d) the silhouette method (Rousseeuw, 1987), (e) the gap statistic (Tibshirani et al., 2001), (f) the prediction strength (Tibshirani and Walther, 2005), and (g) the jump method (Sugar and James, 2003). In this article, we show that the slope heuristic performs significantly better than these seven methods when the data points inside the same cluster are correlated, non-Gaussian, in high dimensional situations, or when there is a dominant cluster. We also apply the slope statistic in graph clustering and actual biological datasets. We obtain results consistent with prior knowledge of the empirical datasets.

The paper unfolds as follows. Section 2 presents basic notations. Section 3 introduces the slope statistic. Section 4 provides a brief review of other common methods. Some simulations in items generated by different probability distributions and graph clustering are provided in Section 5 and finally, applications in actual datasets in Section 6.

2. Basic notation

Let $\mathcal{X} = \{x_1, \ldots, x_n\}$ be the data with n elements and let $d(x_i, x_j)$ denote the distance between x_i and x_j. The Euclidean distance is the most common choice but other metrics can also be considered. Suppose that we must classify each element of the data $\mathcal{X}$ in one of the following k clusters $C^1, C^2, \ldots, C^k$. The first difficulty is that the number of clusters k is usually unknown *a priori*, thus we have to estimate this value. Furthermore, we note that the definition of a cluster depends on the application and it is not always clear what should be the optimal number of clusters for a given problem even in theoretical grounds. The usual approach to solve this problem is to define a parametric model for the shape of clusters or to use a two-step procedure where a clustering algorithm is applied and then goodness of the classification determines the number of clusters. We follow the latter approach and use the silhouette statistic proposed by Rousseeuw (1987) as the goodness of classification measure. For the sake of completeness, we present a brief review of this measure in what follows.

Define

$$d(x_i, B) = \frac{1}{\#B} \sum_{x \in B} d(x_i, x), \tag{1}$$

as the average dissimilarity of x_i to all elements of cluster B, where $\#B$ is the number of elements of B. Denote by A the cluster to which x_i has been assigned by the clustering algorithm and by C any other cluster different of A. Define

$$a_i = d(x_i, A) \quad \text{and} \quad b_i = \min_{C \neq A} d(x_i, C).$$

The quantities a_i and b_i are the "within" dissimilarity and the smallest "between" dissimilarity, respectively. Then a proposal to measure how well object x_i has been clustered is given by Rousseeuw (1987)

$$s_i = \begin{cases} \dfrac{b_i - a_i}{\max\{b_i, a_i\}}, & \text{if } \#A > 1, \\ 0, & \text{if } \#A = 1. \end{cases} \tag{2}$$

Now, for each number of clusters $k = 2, 3, \ldots, n$ compute the silhouette statistic as

$$s(k) = \frac{1}{n} \sum_{i=1}^{n} s_i.$$

The choice of the silhouette statistic ($s(k)$) is interesting due to its interpretations. Notice that $-1 \leq s_i \leq 1$, therefore, there are three possible situations that must be analyzed. The first one is when $s_i \approx 1$. This implies that the "within" dissimilarity is much smaller than the smallest "between" dissimilarity ($a_i \ll b_i$). In other words, the object x_i has been assigned to an appropriate cluster since the second-best choice cluster is not nearly as close as the cluster the object is assigned. The second situation occurs when $s_i \approx 0$. Then $a_i \approx b_i$, and hence it is not clear whether i should have been assigned to the cluster the object is assigned or to the second-best choice cluster because object x_i lies equally far away from both. The third situation takes place when $s_i \approx -1$. Then $a_i \gg b_i$, so object x_i lies much closer to the second-best choice cluster than to the cluster the object is assigned. Therefore it is more natural to assign object x_i to the second-best choice cluster instead of the cluster the object is assigned because this object x_i has been "misclassified". Usually, the clustering algorithms (the k-means algorithm, for instance) find at least a local optimum solution, therefore this case where $s_i \approx -1$ rarely occurs. To conclude, s_i measures how well object x_i has been classified. Consequently, the silhouette statistic $s(k)$ ($-1 \leq s(k) \leq 1$) measures how well all the objects x_i for $i = 1, \ldots, n$ have been classified on average (Rousseeuw, 1987).

In Rousseeuw's original proposal, it is suggested to select the k such that $s(k)$ is maximum ($\hat{k} = \arg\max_{k \in \{2,\ldots,n\}} s(k)$). This procedure proceeds well if all clusters are homogeneous, i.e., they have approximately the same inner variability.