Information Sciences 324 (2015) 126–145

Contents lists available at ScienceDirect

Information Sciences

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

Recovering the number of clusters in data sets with noise features using feature rescaling factors

Renato Cordeiro de Amorim [a,*], Christian Hennig [b,1]

[a] School of Computer Science, University of Hertfordshire, College Lane Campus, Hatfield AL10 9AB, UK
[b] Department of Statistical Science, University College London, Torrington Place, London WC1E 6BT, UK

ARTICLE INFO

Article history:
Received 13 January 2015
Revised 22 April 2015
Accepted 22 June 2015
Available online 30 June 2015

Keywords:
Feature re-scaling
Clustering
K-Means
Cluster validity index
Feature weighting

ABSTRACT

In this paper we introduce three methods for re-scaling data sets aiming at improving the likelihood of clustering validity indexes to return the true number of spherical Gaussian clusters with additional noise features. Our method obtains feature re-scaling factors taking into account the structure of a given data set and the intuitive idea that different features may have different degrees of relevance at different clusters.

We experiment with the Silhouette (using squared Euclidean, Manhattan, and the *p*th power of the Minkowski distance), Dunn's, Calinski–Harabasz and Hartigan indexes on data sets with spherical Gaussian clusters with and without noise features. We conclude that our methods indeed increase the chances of estimating the true number of clusters in a data set.

1. Introduction

Clustering is one of the most popular tasks in data analysis. It aims to reveal a class structure in a data set by partitioning it in an unsupervised manner.

In this paper we address the fundamental issue of estimating the number of clusters K in a data set. This particular problem has raised considerable research interest over the years, but it is not without controversies. It is a very active field of research [1,7,22,29], but due to the lack of a generally accepted definition of what a "cluster" is there are no unified standards against which it can be assessed.

A cluster is a homogeneous group of entities. While entities in the same cluster are supposed to be homogeneous, according to some notion of similarity, entities in different clusters are expected to be heterogeneous. This is a rather loose definition for the term cluster, which does not help much defining the true number of clusters for a given data set. In order to make the problem more precise, in the present paper we are interested in finding clusters in the sense of K-Means but with additional non-informative ("noise") features, that is, clusters that are on the cluster-defining features approximately spherical and compact, with similar within-cluster variation, and that can be approximated well by Gaussian distributions. This clustering problem can be solved more precisely and in a more meaningful way if there is some degree of separation between clusters (i.e., distance between the cluster's high density areas), because otherwise, even if the number of clusters is correctly diagnosed, strong overlap between Gaussian distributions means that points cannot be reliably assigned to the Gaussian component that generated them. We agree with [17] that different clustering methods are appropriate for different clustering aims, and that when carrying

* Corresponding author. Tel.:+44 01707 286160; fax:+44 01707 284115.
E-mail addresses: r.amorim@herts.ac.uk (R.C. de Amorim), c.hennig@ucl.ac.uk (C. Hennig).
[1] The work of this author was supported by EPSRC grant EP/K033972/1.

http://dx.doi.org/10.1016/j.ins.2015.06.039
0020-0255/

out a cluster analysis, researchers need to define more precisely what kind of clusters they are interested in. In line with this thought, the above somewhat restrictive cluster definition can help researchers to decide whether the methods presented here are suitable, instead of claiming that we could solve the general problem of clustering and estimating the numbers of clusters.

This defines what we mean by "true" clusters in the following, acknowledging that it does not yield a general definition of the clustering problem, but rather a working definition for one of many possible ways to understand the term "cluster", at which the methods discussed here are aimed. The methods that we propose actually allow for more general than spherical cluster shapes as long as feature re-scaling transforms the cluster shapes into (approximately) spherical ones.

Various clustering algorithms, some explained in Sections 2 and 3, are unable to determine the number of clusters in a given data set, and in fact request this number to be specified beforehand. In scenarios in which this number is not known, a popular solution is to run a given clustering algorithm using different values for the number of clusters and then analyse the generated clusterings afterwards. The process of estimating how well a partition fits the structure underlying the data is often called "cluster validation" [1,15]. After all feasible possibilities are analysed the number of clusters that generated the best partition, according to a clustering validation index, is selected.

Note that it cannot be taken for granted that the problem of finding the true number of clusters coincides with finding the clustering solution that produces the best clustering in terms of the misclassification rate or the adjusted Rand index [20]. For example, it may be that if there are two true clusters, the clustering method splits the data set up incorrectly if indeed $K = 2$ is used as the number of clusters, whereas for $K = 3$ one cluster coincides perfectly with one true cluster and the other true cluster is split into two found clusters, which for many applications and in terms of the adjusted Rand index may be seen as the better solution. In this paper we aim to address both views, finding the true K, and finding the best clustering.

The quantity of noise features in a data set is an important concern. It is not uncommon to have data sets containing entities characterised by features, with some of the latter being irrelevant to the problem at hand. Generally speaking, noise features, together with the degree of overlap between clusters, are the factors with the greatest impact on clustering validation indexes performance [1,7], with a small inclusion of 10% noise features having already a considerable impact on such indexes [1].

In our experiments we simulate irrelevant features by adding features generated from uniform random values to our data sets. Difficulties in the estimation of the number of clusters raised by the presence of noise features in a data set have been considered before [12], however, there is still a view that the issue raised by noise features deserves more consideration [7].

The main contribution of this paper is to present three methods to re-scale data sets in such a way that cluster validity indexes become more likely to return the true number of clusters. Our experiments focus on versions of the most popular partitioning algorithm, K-Means, and the comparison of the performances of each index before and after re-scaling.

Section 2 reviews K-Means and a number of validation indexes. The methodological core of the paper is Section 3, in which we introduce a version of K-Means incorporating feature weighting and more general Minkowski metrics [11]. Different versions of feature re-scaling with and without re-clustering at the end are proposed for use with the validation indexes. In Section 4 we present our simulation study and discuss its results, followed by a conclusion.

2. Background and related work

2.1. K-Means

K-Means [2,24] is arguably the most popular partitioning clustering algorithm [21,29]. Given a data set Y of V-dimensional entities $y_i \in Y$, for $i = 1, 2, \ldots, N$, K-Means generates K non-empty disjoint clusters $S = \{S_1, S_2, \ldots, S_K\}$ around the centroids $C = \{c_1, c_2, \ldots, c_K\}$, by iteratively minimising the sum

$$W_K = W(S, C) = \sum_{k=1}^{K} \sum_{i \in S_k} d(y_i, c_k) \tag{1}$$

of the within-cluster distance between entities and centroids. Each centroid c_k uniquely represents a cluster S_k and is sometimes called its prototype. The K-Means criterion above returns an index representing how good a clustering is, the lower the better. $d(y_i, c_k)$ in (1) represents the distance between entity y_i and the centroid c_k. In the original K-Means, this distance measure is the squared Euclidean distance given by $d(y_i, c_k) = \sum_{v \in V} (y_{iv} - c_{kv})^2$, minimising the square error criterion. Other distance measures are possible, such as the Manhattan distance given by $d(y_i, c_k) = \sum_{v \in V} |y_{iv} - c_{kv}|$, although only with the squared Euclidean distance the cluster centroids minimising W are actually the within-cluster means. In the present paper, we will not only consider the Euclidean distance, but also the Manhattan distance and the pth power of the Minkowski distance $d_p(y_i, c_k) = \sum_v |y_{iv} - c_{kv}|^p$ for various values of p, because with a suitable choice of p this has been found to work well with noise features, see [11] and Section 3.

The pth power of the Minkowski distance is chosen here by analogy to the use of the squared Euclidean distance, rather than the Euclidean distance itself, in the original K-Means.

The minimisation of (1) has three simple steps, iterated until convergence.

1. Select the values of K entities $y_i \in Y$ as initial centroids $c_1, c_2, \ldots, c_K$. The initial entities may be chosen at random, but better strategies are available, see Section 2.6. Set $S = \{\emptyset\}$.
2. Assign each entity $y_i \in Y$ to the cluster S_k represented by c_k, the closest centroid to y_i.