Locally Scaled Density Based Clustering

Ergun Biçici and Deniz Yuret

Koç University
Rumelifeneri Yolu 34450
Sariyer Istanbul, Turkey
{ebicici,dyuret}@ku.edu.tr

Abstract. Density based clustering methods allow the identification of arbitrary, not necessarily convex regions of data points that are densely populated. The number of clusters does not need to be specified beforehand; a cluster is defined to be a connected region that exceeds a given density threshold. This paper introduces the notion of local scaling in density based clustering, which determines the density threshold based on the local statistics of the data. The local maxima of density are discovered using a k-nearest-neighbor density estimation and used as centers of potential clusters. Each cluster is grown until the density falls below a pre-specified ratio of the center point's density. The resulting clustering technique is able to identify clusters of arbitrary shape on noisy backgrounds that contain significant density gradients. The focus of this paper is to automate the process of clustering by making use of the local density information for arbitrarily sized, shaped, located, and numbered clusters. The performance of the new algorithm is promising as it is demonstrated on a number of synthetic datasets and images for a wide range of its parameters.

1 Introduction

Clustering is the process of allocating points in a given dataset into disjoint and meaningful clusters. Density based clustering methods allow the identification of arbitrary, not necessarily convex regions of data points that are densely populated. Density based clustering does not need the number of clusters beforehand but relies on a density-based notion of clusters such that for each point of a cluster the neighborhood of a given radius (ε) has to contain at least a minimum number of points ($\wp$). However, finding the correct parameters for standard density based clustering [1] is more of an art than science.

This paper introduces the locally scaled density based clustering (LSDBC) algorithm, which clusters points by connecting dense regions of space until the density falls below a threshold determined by the center of the cluster. LSDBC takes two input parameters: k, the order of nearest neighbor to consider for each data point for density calculation and α, which determines the boundary of the current cluster expansion based on its density. The algorithm is robust to background noise and density gradients for a wide range of its parameters.

Density based clustering in its original form, DBSCAN [1], is sensitive to minor changes in its parameters known as the neighborhood of a given radius (ε) and the minimum number of points that need to be contained within the neighborhood ($\wp$). We discuss density based clustering and identify some of its drawbacks in Sect. 2. Although

B. Beliczynski et al. (Eds.): ICANNGA 2007, Part I, LNCS 4431, pp. 739–748, 2007.
© Springer-Verlag Berlin Heidelberg 2007

using different parameters for the radius of the neighborhood and the number of points contained in it appear to give some flexibility, these two parameters are actually dependent on each other. Instead, the LSDBC technique employs the idea of local scaling. We order points according to their distance to their kth neighbor. This gives an approximate measure of how dense the region around each point is. Then, starting with higher density points, we cluster densely populated regions together. The resulting clustering technique does not require fine tuning of parameters and is more robust. OPTICS [2] also bases its clustering decisions on the local density by using *kNN* type density estimation (differences are explored in Sect. 6).

The local scaling technique, previously employed successfully by spectral clustering [3], makes use of the local statistics of points to separate the clusters within the dataset. The idea is to scale each point in the dataset with a factor proportional to its distance to its kth neighbor. Section 3 discusses local scaling and how it can be used for clustering purposes. We show that when local scaling is used in density based clustering, it creates more robust clusters and allows the automatic creation of clusters without any need for parameters other than k, the order of nearest neighbor to consider, and α, which decides when the drop in the density is necessary for the cluster change.

Density based clustering is important for knowledge discovery in databases. Its practical application aresas include biomedical image segmentation [4], molecular biology and geospatial data clustering [5], and earth science tasks [1].

The following lists the contributions of this paper. We introduce locally scaled density based clustering (Sect. 4), which correctly ignores background clutter and identifies clusters within background noise. LSDBC is also robust to changes in the parameters and produces stable clusters for a wide range of them. LSDBC makes the underlying structure of high-dimensional data accessible. The problems we deal with include: (1) finding appropriate parameter values, (2) handling data with different local statistics, (3) clustering in the presence of background clutter, and (4) reducing the number of parameters used. Our results show better performance than prominent clustering techniques such as DBSCAN, k-means, and spectral clustering with local scaling on synthetic datasets (Sect. 5). Our results on image segmentation tasks also show that LSDBC is able to handle image data and segment it into meaningful regions. Related work and density estimation are discussed in Sect. 6 and the last section concludes.

2 Density Based Clustering

Density based clustering differentiates regions which have higher density than its neighborhood and does not need the number of clusters as an input parameter. Regarding a termination condition, two parameters indicate when the expansion of clusters should be terminated: given the radius of the volume of data points to look for, ε, a minimum number of points for the density calculations, $\wp$, has to be exceeded.

Let $d(p, q)$ give the distance between two points p and q; we give the basic terminology of density based clustering below. ε neighborhood of a point p is denoted by $N_\varepsilon(p)$ and is defined by $N_\varepsilon(p) = \{q \in \textit{Points} \mid d(p, q) \leq \varepsilon\}$, where *Points* is the set of points in our dataset. A *core point* is defined as a point above the density threshold wrt. ε and $\wp$, i.e. $|N_\varepsilon(p)| \geq \wp$. A border point is defined as a point below the threshold but that belongs to the ε neighborhood of a core point.