Density-Based Clustering Based on Hierarchical Density Estimates

Ricardo J.G.B. Campello*, Davoud Moulavi, and Joerg Sander**

Dept. of Computing Science, University of Alberta, Edmonton, AB, Canada
{rcampell,moulavi,jsander}@ualberta.ca

Abstract. We propose a theoretically and practically improved density-based, hierarchical clustering method, providing a clustering hierarchy from which a simplified tree of significant clusters can be constructed. For obtaining a "flat" partition consisting of only the most significant clusters (possibly corresponding to different density thresholds), we propose a novel cluster stability measure, formalize the problem of maximizing the overall stability of selected clusters, and formulate an algorithm that computes an optimal solution to this problem. We demonstrate that our approach outperforms the current, state-of-the-art, density-based clustering methods on a wide variety of real world data.

1 Introduction

Density-based clustering [1,2] is a popular clustering paradigm. However, the existing methods have a number of limitations: *(i)* Some methods (e.g., DBSCAN [3] and DENCLUE [4]) can *only* provide a "flat" (i.e. non-hierarchical) labeling of the data objects, based on a global density threshold. Using a single density threshold can often not properly characterize common data sets with clusters of very different densities and/or nested clusters. *(ii)* Among the methods that provide a clustering hierarchy, some (e.g., gSkeletonClu [5]) are not able to automatically simplify the hierarchy into an easily interpretable representation involving only the most significant clusters. *(iii)* Many hierarchical methods, including OPTICS [6] and gSkeletonClu, suggest only how to extract a flat partition by using a global cut/density threshold, which may not result in the most significant clusters if these clusters are characterized by *different* density levels. *(iv)* Some methods are limited to specific classes of problems, such as networks (gSkeletonClu), and point sets in the real coordinate space (e.g., DECODE [7], and Generalized Single-Linkage [8]). *(v)* Most methods depend on multiple, often critical input parameters (e.g., [3], [4], [7], [8], [9]).

In this paper, we propose a clustering approach that, to the best of our knowledge, is unique in that it does not suffer from any of these drawbacks. In detail, we make the following contributions: *(i)* We introduce a hierarchical clustering

* Currently on a sabbatical leave, he is originally from SCC/ICMC/USP, University of São Paulo at São Carlos, Brazil. He acknowledges FAPESP and CNPq.
** This work has been partly supported by NSERC.

J. Pei et al. (Eds.): PAKDD 2013, Part II, LNAI 7819, pp. 160–172, 2013.

method, called HDBSCAN, which generates a complete density-based clustering hierarchy from which a simplified hierarchy composed only of the most significant clusters can be easily extracted. *(ii)* We propose a new measure of cluster stability for the purpose of extracting a set of significant clusters from possibly different levels of a simplified cluster tree produced by HDBSCAN. *(iii)* We formulate the task of extracting a set of significant clusters as an optimization problem in which the overall stability of the composing clusters is maximized. *(iv)* We propose an algorithm that finds the globally optimal solution to this problem. *(v)* We demonstrate the advancement in density-based clustering that our approach represents on a variety of real world data sets.

The remainder of this paper is organized as follows. We discuss related work in Section 2. In Section 3, we redefine DBSCAN, and we propose the algorithm HDBSCAN in Section 4. In Section 5, we introduce a new measure of cluster stability, propose the problem of extracting an optimal set of clusters from a cluster tree, and give an algorithm to solve this problem. Section 6 presents an extensive experimental evaluation, and Section 7 concludes the paper.

2 Related Work

Apart from methods aimed at getting approximate estimates of level sets and density-contour trees for continuous-valued p.d.f. — e.g., see [8] and references therein — not much attention has been given to hierarchical density-based clustering in more general data spaces. The works most related to ours are those in [6,9,5,10]. In [6], a post-processing procedure to extract a simplified cluster tree from the reachability plot produced by the OPTICS algorithm was proposed. This procedure did not become as popular as OPTICS itself, probably because it is very sensitive to the choice of a critical parameter that cannot easily be determined or understood. Moreover, no automatic method to extract a flat clustering solution based on local cuts in the obtained tree was described. In [9], an improved method to extract trees of significant clusters from reachability plots was proposed that is less sensitive to the user settings than the original method in [6]. However, this method is based on heuristics with embedded threshold values that can strongly affect the results, and the problem of extracting a flat solution from local cuts in the cluster tree was practically untouched; the only mentioned (ad-hoc) approach was to arbitrarily take all the leaf clusters and discard the others. In [5], the original findings from [6,9,11] were recompiled in the particular context of community discovery in complex networks. However, no mechanism to extract a simplified cluster tree from the resulting (single-linkage-like) clustering dendrogram was adopted, and only a method producing a global cut through the dendrogram was described. The algorithm AUTO-HDS [10] is, like our method, based on a principle used to simplify clustering hierarchies, which in part refers back to the work of [12]. The clustering hierarchy obtained by AUTO-HDS is typically a subset of the one obtained by our method HDBSCAN. Conceptually, it is equivalent to a sampling of the HDBSCAN hierarchical levels, from top to bottom, at a geometric rate controlled by a user-defined parameter, r_{shave}. Such