Asian Journal of Applied Science and Technology (AJAST)
(Quarterly International Journal) Volume 3, Issue 4, Pages 46-56, October-December 2019

Optics: A Density-Based Algorithm for Discovering Cluster in Large Databases with Noise

R.Nandhakumar[1] & Dr.Antony Selvadoss Thanamani[2]

[1]*Assistant Professor, Department of Computer Science, Nallamuthu Gounder Mahalingam College, Pollachi-642001, India.*
[2]*Associate Professor & Head, Department of Computer Science, Nallamuthu Gounder Mahalingam College, Pollachi-642001, India.*

Article Received: 13 April 2019 Article Accepted: 05 August 2019 Article Published: 16 October 2019

ABSTRACT

Cluster analysis is a primary method for database mining. It is either used as a stand-alone tool to get insight into the distribution of a data set, e.g. to focus further analysis and data processing, or as a preprocessing step for other algorithms operating on the detected clusters. Almost all of the well-known clustering algorithms require input parameters which are hard to determine but have a significant influence on the clustering result. Furthermore, for many real-data sets there does not even exist a global parameter setting for which the result of the clustering algorithm describes the intrinsic clustering structure accurately. We introduce a new algorithm for the purpose of cluster analysis which does not produce a clustering of a data set explicitly; but instead creates an augmented ordering of the database representing its density-based clustering structure. This cluster-ordering contains information which is equivalent to the density-based clustering's corresponding to a broad range of parameter settings. It is a versatile basis for both automatic and interactive cluster analysis. We show how to automatically and efficiently extract not only 'traditional' clustering information (e.g. representative points, arbitrary shaped clusters), but also the intrinsic clustering structure. For medium sized data sets, the cluster-ordering can be represented graphically and for very large data sets, we introduce an appropriate visualization technique. Both are suitable for inter- active exploration of the intrinsic clustering structure offering additional insights into the distribution and correlation of the data.

Keywords: *Cluster Analysis, OPTICS, DBSCAN, Reachability and Connectivity.*

1. INTRODUCTION

Larger and larger amounts of data are collected and stored in databases increasing the need for efficient and effective analysis methods to make use of the information contained implicitly in the data. One of the primary data analysis tasks is cluster analysis which is intended to help a user to understand the natural grouping or structure in a data set. Therefore, the development of improved clustering algorithms has received a lot of attention in the last few years. Roughly speaking, the goal of a clustering algorithm is to group the objects of a database into a set of meaningful subclasses. A clustering algorithm can be used either as a stand-alone tool to get insight into the distribution of a data set, e.g. in order to focus further analysis and data processing, or as a preprocessing step for other algorithms which operate on the detected clusters.

Applications of clustering are, for instance, the creation of thematic maps in geographic information systems by clustering feature spaces, the detection of clusters of objects in geographic information systems and to explain them by other objects in their neighborhood [17], or the clustering of a Web-log database to discover groups of similar access patterns which may correspond to different user profiles [7]. Most of the recent research related to the task of clustering has been directed towards efficiency. The more serious problem, however, is effectively, i.e. the quality or usefulness of the result. Although most traditional clustering algorithms do not scale well with the size and/or dimension of the data set, one way to overcome this problem is to use sampling in combination with a clustering algorithm (see e.g. [8]). This approach works well for many applications and clustering algorithms.

The idea is to apply a clustering algorithm A only to a subset of the whole database. From the result of A for the subset, we can then infer a clustering of the whole database which does not differ much from the result obtained by applying A to the whole data set. However, this does not ensure that the result of the clustering algorithm. A actually reflects the natural groupings in the data. There are three interconnected reasons why the affectivity of