Hierarchical constraints

Providing structural bias for hierarchical clustering

Korinna Bade · Andreas Nürnberger

Abstract Constrained clustering received a lot of attention in the last years. However, the widely used pairwise constraints are not generally applicable for hierarchical clustering, where the goal is to derive a cluster hierarchy instead of a flat partition. Therefore, we propose for the hierarchical setting—based on the ideas of pairwise constraints—the use of must-link-before (MLB) constraints. In this paper, we discuss their properties and present an algorithm that is able to create a hierarchy by considering these constraints directly. Furthermore, we propose an efficient data structure for its implementation and evaluate its effectiveness with different datasets in a text clustering scenario.

Keywords Constrained clustering · Hierarchical clustering · Semi-supervised learning

1 Introduction

Automatic structuring of data is a very helpful tool for improving a user's understanding of the data and providing an overview over the information contained therein. If the amount of data increases, a structuring into a hierarchy is especially useful since it can group the data on different levels of granularity. This gives a more detailed insight into the data and enables a user to concentrate on interesting parts of the data, analyzing details only when necessary. Hierarchical clustering methods were developed in the past to fulfill this task. The goal of these approaches is to group data only based on their features, i.e. information assigned to

Editor: Jennifer Dy.

K. Bade (✉)
Department of Computer Science and Languages, Anhalt University of Applied Sciences, Lohmannstraße 23, 06366 Köthen (Anhalt), Germany
e-mail: k.bade@inf.hs-anhalt.de

A. Nürnberger
Fakultät für Informatik, Institut für Technische und Betriebliche Informationssysteme, Otto-von-Guericke-University Magdeburg, 39106 Magdeburg, Germany
e-mail: andreas.nuernberger@ovgu.de