Data Clustering Analysis - From Simple Groupings to Scalable Clustering With Constraints

Osmar R. Zaïane, Andrew Foss, Chi-Hoon Lee, and Weinan Wang

University of Alberta, Edmonton, Alberta, Canada

Summary. Clustering is the problem of grouping data based on similarity and consists of maximizing the intra-group similarity while minimizing the iter-group similarity. While this problem has attracted the attention of many researchers for many years, we are witnessing a resurgence of interest in new clustering techniques in the data mining community. In this paper we discuss some very recent clustering approaches and recount our experience with some of these algorithms. We also present the problem of clustering in the presence of constraints and discuss the issue of cluster validation.

1 Introduction

Cluster analysis is the automatic identification of groups of similar objects. This analysis attempts to group data objects by maximizing inter-group similarity and minimizing intra-group similarity. Clustering is an unsupervised classification process that is fundamental to data mining. Many data mining queries are concerned either with how the data objects are grouped or which objects could be considered remote from natural groupings. There have been many works on cluster analysis but we are now witnessing a significant resurgence of interest in new clustering techniques. Indeed, discovering distributions of patterns in either numerical or categorical data is relevant for many application domains. Scalability and high dimensionality are not the only focus of the recent research in clustering analysis. Indeed, it is getting difficult to keep track of all the new clustering strategies, their advantages and shortcomings.

In data mining, people have been seeking for effective and efficient clustering techniques in the analysis of large databases. These following are the typical requirements for a good clustering technique in data mining [HK00]:

- **Scalability:** The cluster method should be applicable to huge databases. Techniques which can only be applied to small datasets are practically useless.
- **Ability to cluster different types of attributes:** Clustering objects could be of different types : numerical data, boolean data or categorical data. Ideally a clustering method should be suitable for all different types of data objects.