Chapter II

Approximating Proximity for Fast and Robust Distance-Based Clustering

Vladimir Estivill-Castro, University of Newcastle, Australia
Michael E. Houle, University of Sydney, Australia

Distance-based clustering results in optimization problems that typically are NP-hard or NP-complete and for which only approximate solutions are obtained. For the large instances emerging in data mining applications, the search for high-quality approximate solutions in the presence of noise and outliers is even more challenging. We exhibit fast and robust clustering methods that rely on the careful collection of proximity information for use by hill-climbing search strategies. The proximity information gathered approximates the nearest neighbor information produced using traditional, exact, but expensive methods. The proximity information is then used to produce fast approximations of robust objective optimization functions, and/or rapid comparison of two feasible solutions. These methods have been successfully applied for spatial and categorical data to surpass well-established methods such as k-MEANS in terms of the trade-off between quality and complexity.

INTRODUCTION

A central problem in data mining is that of automatically summarizing vast amounts of information into simpler, fewer and more comprehensible categories. The most common and well-studied way in which this categorizing is done is by

partitioning the data elements into groups called clusters, in such a way that members of the same cluster are as similar as possible, and points from different clusters are as dissimilar as possible. By examining the properties of elements from a common cluster, practitioners hope to discover rules and concepts that allow them to characterize and categorize the data.

The applications of clustering to knowledge discovery and data mining (KDDM) (Fayyad, Reina, & Bradley, 1998; Ng & Han, 1994; Wang, Yang, & Muntz, 1997) are recent developments in a history going back more than 30 years. In machine learning classical techniques for unsupervised learning are essentially those of clustering (Cheeseman et al, 1988; Fisher, 1987; Michalski & Stepp, 1983). In statistics, clustering arises in the analysis of mixture models, where the goal is to obtain statistical parameters of the individual populations (Titterington, Smith & Makov, 1985; Wallace & Freeman, 1987). Clustering methods appear in the literature of dimensionality reduction and vector quantization. Many textbooks have large sections devoted to clustering (Berry & Linoff, 1997; Berson & Smith, 1998; Cherkassky & Muller, 1998; Duda & Hart, 1973; Han & Kamber, 2000; Mitchell, 1997), and several are entirely devoted to the topic (Aldenderfer & Blashfield, 1984; Anderberg, 1973; Everitt, 1980; Jain & Dubes, 1998).

Although different contexts give rise to several clustering methods, there is a great deal of commonality among methods themselves. However, not all methods are appropriate for all contexts. Here, we will concentrate only on clustering methods that are suitable for the exploratory and early stages of a KDDM exercise. Such methods should be:

- **Generic**: Virtually every clustering method may be described as having two components: a search mechanism that generates candidate clusters, and an evaluation function that measures the quality of these candidates. In turn, an evaluation function may make use of a function that measures the similarity (or dissimilarity) between a pair of data points. Such methods can be considered generic if they can be applied in a variety of domains simply by substituting one measure of similarity for another.

- **Scalable**: In order to handle the huge data sets that arise in KDDM applications, clustering methods must be as efficient as possible in terms of their execution time and storage requirements. Given a data set consisting of \( n \) records on \( D \) attributes, the time and space complexity of any clustering method for the set should be sub-quadratic in \( n \), and as low as possible in \( D \) (ideally linear). In particular, the number of evaluations of the similarity function must be kept as small as possible. Clustering methods proposed in other areas are completely unsuitable for data mining applications, due to their quadratic time complexities.

- **Incremental**: Even if the chosen clustering method is scalable, long execution times must be expected when the data sets are very large. For this reason, it is desirable to use methods that attempt to improve their solutions in an incremental fashion. Incremental methods allow the user to monitor their progress, and to terminate the execution early whenever a clustering of
Graph-Based Modelling of Concurrent Sequential Patterns
www.igi-global.com/article/graph-based-modelling-concurrent-sequential/42151?camid=4v1a