Chapter 15

CLUSTERING METHODS

Lior Rokach
Department of Industrial Engineering
Tel-Aviv University
liorr@eng.tau.ac.il

Oded Maimon
Department of Industrial Engineering
Tel-Aviv University
maimon@eng.tau.ac.il

Abstract This chapter presents a tutorial overview of the main clustering methods used in Data Mining. The goal is to provide a self-contained review of the concepts and the mathematics underlying clustering techniques. The chapter begins by providing measures and criteria that are used for determining whether two objects are similar or dissimilar. Then the clustering methods are presented, divided into: hierarchical, partitioning, density-based, model-based, grid-based, and soft-computing methods. Following the methods, the challenges of performing clustering in large data sets are discussed. Finally, the chapter presents how to determine the number of clusters.

Keywords: Clustering, K-means, Intra-cluster homogeneity, Inter-cluster separability,

1. Introduction

Clustering and classification are both fundamental tasks in Data Mining. Classification is used mostly as a supervised learning method, clustering for unsupervised learning (some clustering models are for both). The goal of clustering is descriptive, that of classification is predictive (Veyssieres and Plant, 1998). Since the goal of clustering is to discover a new set of categories, the new groups are of interest in themselves, and their assessment is intrinsic. In classification tasks, however, an important part of the assessment is extrinsic, since the groups must reflect some reference set of classes. *"Understanding*

our world requires conceptualizing the similarities and differences between the entities that compose it" (Tyron and Bailey, 1970).

Clustering groups data instances into subsets in such a manner that similar instances are grouped together, while different instances belong to different groups. The instances are thereby organized into an efficient representation that characterizes the population being sampled. Formally, the clustering structure is represented as a set of subsets $C = C_1, \ldots, C_k$ of S, such that: $S = \bigcup_{i=1}^{k} C_i$ and $C_i \cap C_j = \emptyset$ for $i \neq j$. Consequently, any instance in S belongs to exactly one and only one subset.

Clustering of objects is as ancient as the human need for describing the salient characteristics of men and objects and identifying them with a type. Therefore, it embraces various scientific disciplines: from mathematics and statistics to biology and genetics, each of which uses different terms to describe the topologies formed using this analysis. From biological "taxonomies", to medical "syndromes" and genetic "genotypes" to manufacturing "group technology" — the problem is identical: forming categories of entities and assigning individuals to the proper groups within it.

2. Distance Measures

Since clustering is the grouping of similar instances/objects, some sort of measure that can determine whether two objects are similar or dissimilar is required. There are two main type of measures used to estimate this relation: distance measures and similarity measures.

Many clustering methods use distance measures to determine the similarity or dissimilarity between any pair of objects. It is useful to denote the distance between two instances x_i and x_j as: $\mathrm{d}(x_i,x_j)$. A valid distance measure should be symmetric and obtains its minimum value (usually zero) in case of identical vectors. The distance measure is called a metric distance measure if it also satisfies the following properties:

1. Triangle inequality $\mathrm{d}(x_i,x_k) \leq \mathrm{d}(x_i,x_j) + \mathrm{d}(x_j,x_k) \quad \forall x_i,x_j,x_k \in S$.

2. $\mathrm{d}(x_i,x_j) = 0 \Rightarrow x_i = x_j \quad \forall x_i, x_j \in S$.

2.1 Minkowski: Distance Measures for Numeric Attributes

Given two p-dimensional instances, $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ and $x_j = (x_{j1}, x_{j2}, \ldots, x_{jp})$, The distance between the two data instances can be calculated using the Minkowski metric (Han and Kamber, 2001):

$$d(x_i, x_j) = (|x_{i1} - x_{j1}|^g + |x_{i2} - x_{j2}|^g + \ldots + |x_{ip} - x_{jp}|^g)^{1/g}$$