Chapter 1
Exploring the Unknown Nature of Data:
Cluster Analysis and Applications

Rui Xu
Missouri University of Science and Technology, USA

Donald C. Wunsch II
Missouri University of Science and Technology, USA

ABSTRACT

To classify objects based on their features and characteristics is one of the most important and primitive activities of human beings. The task becomes even more challenging when there is no ground truth available. Cluster analysis allows new opportunities in exploring the unknown nature of data through its aim to separate a finite data set, with little or no prior information, into a finite and discrete set of "natural," hidden data structures. Here, the authors introduce and discuss clustering algorithms that are related to machine learning and computational intelligence, particularly those based on neural networks. Neural networks are well known for their good learning capabilities, adaptation, ease of implementation, parallelization, speed, and flexibility, and they have demonstrated many successful applications in cluster analysis. The applications of cluster analysis in real world problems are also illustrated. Portions of the chapter are taken from Xu and Wunsch (2008).

INTRODUCTION

To classify objects based on their features and characteristics is one of the most important and primitive activities of human beings. Objects displaying similar features and properties based on certain pre-specified criteria are classified into the same group or category. The properties of a specific new object can then be inferred using this classification information. For example, when we see a cat, we know immediately that it can climb trees and likes eating fish without really seeing it do so. This task becomes even more challenging when there is no ground truth available. Cluster analysis, also known as unsupervised classification or exploratory data analysis, aims to address this problem and explores

DOI: 10.4018/978-1-60566-766-9.ch001

the unknown nature of data through separating a finite data set, with little or no prior information, into a finite and discrete set of "natural," hidden data structures (Everitt et al., 2001; Hartigan, 1975; Jain and Dubes, 1988).

Clustering focuses on the partition of data objects (patterns, entities, instances, observances, units) into a certain number of clusters (groups, subsets, or categories). However, there is no universally agreed upon and precise definition of the term cluster. Most researchers describe a cluster in terms of internal homogeneity and external separation (Everitt et al., 2001; Hansen and Jaumard, 1997; Jain and Dubes, 1988). Data objects that are in the same cluster are required to be similar to each other, while data objects belonging to different clusters should display sufficient dissimilarity. Here, we provide simple mathematical descriptions of two types of clustering, known as partitional and hierarchical clustering, based on the discussion in Hansen and Jaumard (1997).

Given a set of input patterns $\mathbf{X}=\{\mathbf{x}_1,\ldots,\mathbf{x}_j,\ldots,\mathbf{x}_N\}$, where $\mathbf{x}_j=(x_{j1}, x_{j2,\ldots,} x_{jd})\in \mathrm{R}^d$, with each measure x_{ji} called a feature (attribute, dimension, or variable):

1. Hard partitional clustering attempts to seek a K-partition of $\mathbf{X}$, $C=\{C_1,\ldots,C_K\}$ ($K\leq N$), such that
   - $C_i \neq \phi,\ i=1,\cdots,K$ (1)
   - $\bigcup_{i=1}^{K} C_i = \mathbf{X}$ (2)
   - $C_i \cap C_j = \phi,\ i,j=1,\cdots,K \text{ and } i\neq j$ (3)
2. Hierarchical clustering attempts to construct a tree-like nested structure partition of $\mathbf{X}$, $H=\{H_1,\ldots,H_Q\}$ ($Q\leq N$), such that $C_i\in H_m$, $C_j\in H_l$, and $m>l$ imply $C_i\subset C_j$ or $C_i\cap C_j=\phi$ for all $i, j\neq i$, m, $l=1,\ldots,Q$.

As shown in Eq. 3, each data object is only associated with a single cluster. In some cases, it may also be possible that an object is related to several clusters. This can be achieved by allowing the object to belong to all K clusters with a degree of membership, $u_{i,j}\in[0,1]$, which represents the membership coefficient of the j^{th} object in the i^{th} cluster and satisfies the following two constraints (Zadeh, 1965):

$$\sum_{i=1}^{K} u_{i,j} = 1,\ \forall j\,, \tag{4}$$

and

$$\sum_{j=1}^{N} u_{i,j} < N,\ \forall i\,, \tag{5}$$

This is known as fuzzy clustering and is especially useful when the clusters are not well separated and the boundaries are ambiguous (Bezdek, 1981).

The basic procedure of cluster analysis is summarized in the following steps:

1. *Feature selection or extraction*. Feature selection chooses distinguishing features from a set of candidates, while feature extraction utilizes some transformations to generate useful and novel features from the original ones (Jain et al., 1999; Bishop, 1995). Effective selection or generation of salient features can greatly decrease the storage requirement and measurement cost, simplify the