Chapter 10

Introduction to Clustering: Algorithms and Applications

Raymond Greenlaw
Armstrong Atlantic State University, USA

Sanpawat Kantabutra
Chiang Mai University, Thailand

ABSTRACT

This chapter provides the reader with an introduction to clustering algorithms and applications. A number of important well-known clustering methods are surveyed. The authors present a brief history of the development of the field of clustering, discuss various types of clustering, and mention some of the current research directions in the field of clustering. Algorithms are described for top-down and bottom-up hierarchical clustering, as are algorithms for K-Means clustering and for K-Medians clustering. The technique of representative points is also presented. Given the large data sets involved with clustering, the need to apply parallel computing to clustering arises, so they discuss issues related to parallel clustering as well. Throughout the chapter references are provided to works that contain a large number of experimental results. A comparison of the various clustering methods is given in tabular format. They conclude the chapter with a summary and an extensive list of references.

BASICS OF CLUSTERING

Introduction

In today’s world, information has become increasingly indispensable within our daily lives. This information includes financial, medical, and scientific data, and even something as mundane as grocery-shopping information with which we are intimately familiar. Many people have been discussing and exposing about the information era for years. The widespread use of the Internet and web-based applications has further fueled the use of information worldwide. Hand-in-hand with the ever-growing use of information (both

DOI: 10.4018/978-1-60566-908-3.ch010
in terms of the sheer volume of information, but as well in the increased total number of users) comes significantly greater complexity of data and information management. The vast sea of data appears to be limitless not only in quantity but also in dimension. Real-life applications can generate high-dimensional data. In genomics, for instance, a single gene expression profile could yield a vector of measurements whose dimension is in the range between 5,000 and 25,000 (Bühlmann, 2006). Many people have tried to manage data, and some have been able to manage data effectively. But, others have met with less success, and so new methods have emerged to cope with these more-complex situations.

Preferably we want to keep, extract, and maintain knowledge, as opposed to the data itself. Hopefully, the knowledge will be in some useful forms, where one can utilize the knowledge to make informed and strategic decisions. In addition, we would like to deal with smaller quantities of information to simplify decision making. One way to manage a data set is to group data that have certain similar characteristics; this method is called clustering (Hartigan, 1975), (Spath, 1980), (Jain & Dubes, 1988), (Kaufman Rousseeuw, 1990), (Dubes, 1993), (Everitt, 1993), (Mirkin, 1996), (Jain, Murty & Flynn, 1999), (Fasulo, 1999), (Kolatch, 2001), (Han, Kamber, & Tung, 2001), (Ghosh, 2002), (Berkin, 2002). Clustering allows us to work with a small, manageable set of groups rather than a complete data set that is typically unwieldy. In this chapter we will explore the topic of clustering. We provide an introduction to this subject, as well as describe various algorithms and applications of clustering. We begin with a definition of clustering.

**Clustering Definition**

Intuitively, clustering is a grouping of “similar” objects, where similarity is some predetermined function. More formally, given a set of $n$ objects, the process of clustering partitions the set into different subsets of objects such that each subset shares specific common characteristics. The common characteristics are usually specified in terms of some mathematical relation. In geometric parlance, the objects can be viewed as points in some $d$-dimensional space. Clustering partitions these points into groups, where points in the same group are located near one another in space.

In Figure 1 we illustrate the partitioning of a set of points in two-dimensional space into four clusters. Each cluster is represented by a rectangle. Points in the same cluster are somewhat closer—more similar in terms of their Euclidean distance to one another than to points in other clusters. Notice in this case the number of points in each cluster is not perfectly balanced. The clusters have sizes from left-to-right, top-to-bottom of 33, 50, 23, and 16. However, rather than talking about the 122 points separately, we can talk about just four clusters. In base two this is almost a five-order magnitude decrease in the number of objects to consider. For large datasets the reduction in complexity of the number of items to discuss can be much greater still.

**Examples of Clustering**

Let us look at some of the far-reaching applications of clustering in this subsection by considering a few different domains in which clustering data are useful. In general, the field of clustering encompasses a number of different algorithms for grouping objects of similar kinds into respective categories. Clustering can also be viewed as an exploratory data-analysis tool to discover structural properties of data. Because typical objects can be translated into points in a $d$-dimensional space, examples of clustering abound in our daily lives. For example, a group of students majoring in computer science at a university may be
Related Content

Query Interaction Based Approach for Horizontal Data Partitioning
[www.igi-global.com/article/query-interaction-based-approach-for-horizontal-data-partitioning/125650?camid=4v1a](www.igi-global.com/article/query-interaction-based-approach-for-horizontal-data-partitioning/125650?camid=4v1a)

SpyNetMiner: An Outlier Analysis to Tag Elites in Clandestine Social Networks
[www.igi-global.com/article/spynetminer/106861?camid=4v1a](www.igi-global.com/article/spynetminer/106861?camid=4v1a)

Generational Differences Relative to Data-Based Wisdom
[www.igi-global.com/chapter/generational-differences-relative-to-data-based-wisdom/150232?camid=4v1a](www.igi-global.com/chapter/generational-differences-relative-to-data-based-wisdom/150232?camid=4v1a)

Optimal Features for Metamorphic Malware Detection
[www.igi-global.com/chapter/optimal-features-for-metamorphic-malware-detection/109973?camid=4v1a](www.igi-global.com/chapter/optimal-features-for-metamorphic-malware-detection/109973?camid=4v1a)