Chapter 27
Evaluation of Clustering Methods for Adaptive Learning Systems

Wilhelmiina Hämäläinen
University of Eastern Finland, Finland

Ville Kumpulainen
University of Eastern Finland, Finland

Maxim Mozgovoy
University of Aizu, Japan

ABSTRACT

Clustering student data is a central task in the educational data mining and design of intelligent learning tools. The problem is that there are thousands of clustering algorithms but no general guidelines about which method to choose. The optimal choice is of course problem- and data-dependent and can seldom be found without trying several methods. Still, the purposes of clustering students and the typical features of educational data make certain clustering methods more suitable or attractive. In this chapter, the authors evaluate the main clustering methods from this perspective. Based on the analysis, the authors suggest the most promising clustering methods for different situations.

INTRODUCTION

Clustering student data is a central task in the educational data mining and design of intelligent learning tools. Dividing data into natural groups gives a good summary how students are learning and helps to target teaching and tutoring. This is especially topical in the domain of online adaptive learning systems due to larger amount of students and their greater diversity. Clustering can also facilitate the design of predictive models, which are the heart of intelligent tutoring systems.

Indeed, a number of scholars report successful examples of clustering (for various purposes) in actual educational environments. However, the problem of selecting the most appropriate clustering method for student data is rarely addressed. There is a plenty of mainstream clustering methods

DOI: 10.4018/978-1-4666-9562-7.ch027

and literally thousands of specialized clustering algorithms available (Jain, 2010), and choosing the right method for the given task is not easy. In practice, researchers often just pick up the most popular k-means method without a second thought whether its underlying assumptions suit the data. In practice, this means that one may end up with an artificial partition of data instead of finding natural clusters.

The aim of the present work is to evaluate a variety of clustering methods from the perspective of clustering student data. We analyze the main approaches to clustering and see how useful models they produce and how well their underlying assumptions fit typical student data. We do not try to list as many algorithms as possible, but instead our emphasis is to describe the underlying clustering principles and evaluate their properties. Our main goal is to cover those clustering methods which are generally available in the existing data mining and statistical analysis tools, but we introduce also some promising "future methods". Based on this analysis, we suggest the most promising clustering methods for different situations.

The rest of the chapter is organized as follows: First, we give the basic definitions, analyze domain-specific requirements for clustering methods, and survey related research. Then, we introduce the main approaches for clustering and evaluate their suitability for typical student data. Finally, we discuss future research directions and draw the final conclusions. The basic notations used in this chapter are introduced in Table 1.

BACKGROUND

In this section, we define the basic concepts related to clustering, discuss the goals and special requirements of clustering educational data, and survey related research.

Basic Definitions

The main problem of clustering is how to define a cluster. There is no universally accepted precise definition of clustering. Intuitively, clustering means a grouping of data points, where points in one group are similar or close to each other but different or distant from points in the other groups. One may also describe clusters as denser regions of the data space separated by sparser regions or as homogeneous subgroups in a heterogeneous population. Here, we give only a very generic definition of clustering and then describe its different aspects.

Definition 1 (Clustering): Let $D = \{\mathbf{p_1}, \ldots, \mathbf{p_n}\}$ be a data set of n points, $C = \{C_1, \ldots, C_k\}$ a set of k clusters, and M some clustering criterion. A *hard clustering* assigns each point $\mathbf{p_j}$ into exactly one cluster C_i according to M. A *soft clustering* defines for each point-cluster pair $(\mathbf{p_j}, C_i)$ a degree of membership according to M.

Table 1. Basic notations

Notation	Meaning
m	Number of dimensions (variables)
n	Number of data points
k	Number of clusters
$\mathrm{p_i} = (p_{i1}, \ldots, p_{im})$	Data point in m-dimensional data space
$D = \{\mathrm{p_1}, \ldots, \mathrm{p_n}\}$	Data set of n points
C_i	Cluster
$\mathrm{c_i}$	Cluster centroid (representative point of cluster C_i)
$d(\mathrm{p_i}, \mathrm{p_j})$	Distance between points $\mathrm{p_i}$ and $\mathrm{p_j}$
$D(C_i, C_j)$	Distance between clusters C_i and C_j; inter-cluster distance
$\lvert C_i \rvert$	Size of cluster C_i; number of points belonging to C_i