A clustering algorithm to find groups with homogeneous preferences

J. Díez, J.J. del Coz, O. Luaces and A. Bahamonde

Centro de Inteligencia Artificial. Universidad de Oviedo at Gijón, Campus de Viesques,
E-33271 Gijón (Asturias), Spain
{jdiez, juanjo, oluaces, antonio}@aic.uniovi.es

Introduction

When we are learning people’s preferences the training material can be expressed as in regression problems: the description of each object is then followed by a number that assesses the degree of satisfaction. Alternatively, training examples can be represented by preference judgments: pairs of vectors \((v, u)\) where someone expresses that he or she prefers \(v\) to \(u\). Usually, obtaining preference information may be easier and more natural than obtaining the labels needed for a classification or regression approach. Moreover, this type of information is more accurate, since people tend to rate their preferences in a relative way, comparing objects with the other partners in the same batch.

Starting from a set of preference judgments we have to learn a real ranking function \(f\) in such a way that \(f(v) > f(u)\) whenever \(v\) is preferable to \(u\). If we accept linear ranking functions, \(f\) is represented by a hyperplane able to separate increasing or positive differences (like \(v - u\) when \(v > u\)) from decreasing or negative differences (like \(u - v\)). We will employ an SVM classifier: SVMlight [3].

Clustering of preference criteria

In this paper we present an algorithm for clustering preference criteria. This is a useful task. For instance, in [4], Joachims presents an information retrieval system equipped with a ranking function learned from click-through data collected from user interaction with a www search engine. To improve his proposal, the author acknowledges the need to obtain feasible training data. This requires developing clustering algorithms to find homogeneous groups of users.

An Adaptive Route Advisor is described in [2]; the system is able to recommend a route to lead users through a digitalized road map taking into account their preferences. An interesting extension discussed in the paper is to modify route recommendations depending on the time of the day or the purpose of the trip. The approach suggested includes an algorithm that clusters user preferences into contexts.

---

1 The research reported in this paper has been supported in part under MCyT and Feder grant TIC2001-3579.
Another important field of application is the analysis of sensory data used to test the quality (or the acceptability) of market products that are principally appreciated through sensory impressions.

Sensory data include the assessment of products provided by two different kinds of panels. The first one is a small group of expert, trained judges; these will describe each product by attribute-value pairs. This panel will play the role of a bundle of sophisticated sensors. To achieve this performance, 2-3 times as many panelists as those required are screened through a hard selection.

The second kind of panel is made up of untrained consumers; these are asked to rate their degree of acceptance of the tested products on a scale. The aim is to be able to relate sensory descriptions (human and mechanical) with consumer preferences in order to improve production decisions.

Both to select experts to the first kind of panel, and to discover groups of preferences in consumers, clustering algorithms are necessary.

The clustering algorithm

The algorithm has as input a family of preference judgment sets of a number of people. The key insight involved in the clustering algorithm is that ranking functions, learned from each preference judgment set, codify the rationale for these preferences. Therefore we will try to merge data sets with similar ranking functions. We learn the new ranking function from the merged data set, aggregating the data sets if the estimated accuracy is higher, to then adopt the new ranking function as the criteria of the group thus constituted. The algorithm stops when no more merges can be achieved.

To evaluate the algorithm we used a collection of preference judgments taken from EachMovie, a publicly available collaborative filtering database for movie ratings. The ratings were used only to build preference judgments pairs, so no kind of regression was used in our work.

References