Relational Clustering for the Analysis of Internet Newsgroups

T. A. Runkler¹, J. C. Bezdek²

¹Siemens AG, Corporate Technology, Information and Communications, D–81730 München, Germany
²University of West Florida, Computer Science Department, 11000 University Parkway, Pensacola, FL 32514, U.S.A.

Abstract: Clustering is used to determine partitions and prototypes from pattern sets. Sets of numerical patterns can be clustered by alternating optimization (AO) of clustering objective functions or by alternating cluster estimation (ACE). Sets of non-numerical patterns can often be represented numerically by (pairwise) relations. For text data, relational data can be automatically computed using the Levenshtein (or edit) distance. These relational data sets can be clustered by relational ACE (RACE). For text data, the RACE cluster centers can be used as keywords. In particular, the cluster centers extracted by RACE from internet newsgroup articles serve as keywords for those articles. These keywords can be used for automatic document classification.

1 Introduction

Clustering is an unsupervised learning method that partitions a set of patterns into groups (clusters) (Jain et al. (1999)). The partitions can be generated by objective function, cluster estimation, mixture resolving (e.g. expectation maximization), and other types of algorithms. Here we focus on objective function and cluster estimation algorithms. Objective function algorithms minimize square error criteria based on distances between patterns and clusters. This minimization is often done by alternatingly updating the partitions and the cluster prototypes. Alternating optimization is a clustering algorithm specified by an objective function. Alternating cluster estimation is a generalization of the alternating optimization scheme. It is specified by user-defined partition and prototype functions.

In this paper we focus on the application of clustering methods to text documents from the internet. An important focus in this area is clustering text documents (Steinbach et al. (2000)), i.e. to find groups of documents that are related to similar topics and to extract the most important keywords that are considered in this classification (Han and Karypis (2000)). In many articles, large word corpuses with hundreds of thousands of documents are used (Larsen and Aone (1999)) that allow

²Research supported by ONR grant # N00014-96-1-0642.
the use of probabilistic features like (relative) term frequencies (Yang and Liu (1999)) and allow to represent clusters as large groups of similar text documents. In this paper we focus on the analysis of much smaller sets of text documents (up to hundreds) that are collected from postings to internet newsgroups or from individual e-mail messages. Due to the relatively small size of the data sets, term frequencies are not applicable here, since most non-trivial terms occur only once. However, it can be observed that the significant keywords for individual texts often repeatedly appear in modified versions. For example, an article about electronics contains the words conduct, conductors, semiconductor, which will remain separate even after preprocessing with stemmer algorithms. Each of these words appears only once, but the complete cluster of similar words is apparently significant for the document. In this paper we try to find these clusters and use them for the analysis of the document contents. Thus, we break down the document clustering problem to a smaller scale: Instead of searching for clusters of similar documents in the whole data set, we are searching for clusters of similar words within individual documents.

Many clustering algorithms assume that the patterns are real vectors. However, the data in text documents are text strings that are non-numerical patterns but that can be represented numerically using (pairwise) relation matrices. Clusters in this kind of relational data sets can be found with relational alternating cluster estimation (Runkler et al. (1998, 2000)). For many text data sets it is difficult if at all possible to define semantically meaningful relations. In these cases relations can be computed based on comparing individual characters of pairs of strings, e.g. using the Levenshtein (or edit) distance (Levenshtein (1966)). In Runkler et al. (2000) it was found that cluster centers obtained from Levenshtein relations can be used as text keywords. In this article, we apply this method to automatically extract keywords for internet newsgroups.

2 Objective Function Clustering

Numerical clustering algorithms partition a data set \( X = \{x_1, \ldots, x_n\} \in \mathcal{R}^p \) into \( c \in \{2, \ldots, n - 1\} \) clusters. The clusters are described by a \( c \times n \) partition matrix \( U \) and a set \( V \) containing \( c \) cluster prototypes. Objective function clustering algorithms compute \( U \) and \( V \) from \( X \) by optimization of an objective function \( J(U, V; X) \). The \( c\)-means model (CM, Ball et al. (1965)), for example, is specified by

\[
J_{CM}(U, V; X) = \sum_{i=1}^{c} \sum_{k=1}^{n} u_{ik} \|x_k - v_i\|^2 ,
\]  
(1)