Identifying Single Clusters in Large Data Sets

Frank Klawonn
University of Applied Sciences Braunschweig/Wolfenbuettel, Germany

Olga Georgieva
Institute of Control and System Research, Bulgaria

INTRODUCTION

Most clustering methods have to face the problem of characterizing good clusters among noise data. The arbitrary noise points that just do not belong to any class being searched for are of a real concern. The outliers or noise data points are data that severely deviate from the pattern set by the majority of the data, and rounding and grouping errors result from the inherent inaccuracy in the collection and recording of data. In fact, a single outlier can completely spoil the least squares (LS) estimate and thus the results of most LS based clustering techniques such as the hard C-means (HCM) and the fuzzy C-means algorithm (FCM) (Bezdek, 1999).

For these reasons, a family of robust clustering techniques has emerged. There are two major families of robust clustering methods. The first includes techniques that are directly based on robust statistics. The second family, assuming a known number of clusters, is based on modifying the objective function of FCM in order to make the parameter estimates more resistant to the data noise. Among them one promising approach is the noise clustering (NC) technique (Dave, 1991; Klawonn, 2004). It maintains the principle of probabilistic clustering, but an additional noise cluster is introduced. NC was developed and investigated in the context of a variety of objective function-based clustering algorithms and it has demonstrated its reliable ability to detect clusters amongst noise data.

BACKGROUND

Objective function-based clustering aims at minimizing an objective function that indicates a kind of fitting error of the determined clusters to the given data. In this objective function, the number of clusters has to be fixed in advance. However, as the number of clusters is usually unknown, an additional scheme has to be applied to determine the number of clusters (Guo, 2002; Tao, 2002). The parameters to be optimized are the membership degrees that are values of belonging of each data point to every cluster, and the parameters, characterizing the cluster, which finally determine the distance values. In the simplest case, a single vector named cluster centre (prototype) represents each cluster. The distance of a data point to a cluster is simply the Euclidean distance between the cluster centre and the corresponding data point. More generally, one can use the squared inner-product distance norm, in which by a norm inducing symmetric and positive matrix different variances in the directions of the coordinate axes of the data space are accounted for. If the norm inducing matrix is the identity matrix we obtain the standard Euclidean distance that form spherical clusters. Clustering approaches that use more complex cluster prototypes than only the cluster centres, leading to adaptive distance measures, are for instance the Gustafson-Kessel (GK) algorithm (Gustafson, 1979), the volume adaptation strategy (Höppner, 1999; Keller, 2003) and the Gath-Geva (GG) algorithm (Gath, 1989). The latter one is not a proper objective function algorithm, but corresponds to a fuzzified expectation maximization strategy. No matter which kind cluster prototype is used, the assignment of the data to the clusters is based on the corresponding distance measure. In hard clustering, a data object is assigned to the closest cluster, whereas in fuzzy clustering a membership degree, that is, a value that belongs to the interval [0,1] is computed. The highest membership degree of a data corresponds to the closest cluster.

Noise clustering has a benefit of the collection of the noise points in one single cluster. A virtual noise prototype with no parameters to be adjusted is introduced that has always the same distance to all points in the data set. The remaining clusters are assumed to be the good clusters in the data set. The objective function that considers the noise cluster is defined in the same manner as the general scheme for the clustering minimization functional. The main problem of NC is the proper choice of the noise distance. If it is set too small, then most of the points will get classified as noise points, while for a large noise distance most of the points will be classified into clusters other than the noise cluster. A right selection of the distance will result in a classification where the points that are actu-