A Method to Enhance the 'Possibilistic C-Means with Repulsion' Algorithm based on Cluster Validity Index

Juan Wachs[1], Oren Shapira[1] and Helman Stern [1]

{juan|orensha|helman}@bgu.ac.il,
[1]Department of Industrial Engineering and Management, Intelligent Systems Division, Ben-Gurion University of the Negev, Be'er-Sheeva , Israel 84105

Abstract: In this paper, we examine the performance of fuzzy clustering algorithms as the major technique in pattern recognition. Both possibilistic and probabilistic approaches are explored. While the Possibilistic C-Means (PCM) has been shown to be advantageous over Fuzzy C-Means (FCM) in noisy environments, it has been reported that the PCM has an undesirable tendency to produce coincident clusters. Recently, an extension of the PCM has been presented by Timm et al., by introducing a repulsion term. This approach combines the partitioning property of the FCM with the robust noise insensibility of the PCM. We illustrate the advantages of both the possibilistic and probabilistic families of algorithms with several examples and discuss the PCM with cluster repulsion. We provide a cluster valid-ity function evaluation algorithm to solve the problem of parameter optimization. The algorithm is especially useful for the unsupervised case, when labeled data is unavailable.

Keywords: Possibilistic and probabilistic fuzzy clustering; Fuzzy C-Means; Cluster validity index, Robust methods.

1 Introduction

Cluster analysis is the process of classifying objects into subsets that have meaning in the context of a particular problem. The objects are thereby organized into an efficient representation that characterizes the population being sampled [1]. Hard clustering methods assume that each observation belongs to one class, how-

ever in practice clusters may overlap, and data vectors belong partially to several clusters. This scenario can be modeled properly using the fuzzy set theory [2], in which the membership degree of a vector x_k to the i-th cluster (u_{ik}) is a value from [0,1] interval. Bezdek [3] explicitly formulated this approach oriented to clustering by introducing the Fuzzy c-mean clustering algorithm. Unfortunately, this method showed the difficulty of high sensibility to noises and outliers in the data. To reduce this undesirable effect, a number of approaches have been proposed, but the most remarkable has been the possibilistic, introduced by [4], with their possibilistic c-means algorithm. In this algorithm the membership is interpreted as the compatibilities of the datum to the class prototypes (typicalities) which correspond to the intuitive concept of degree of belonging or compatibility. In the case of poor initializations, it is possible that the PCM will converge to a "worthless" partition where part or all the clusters are identical (coincident) while other clusters go undetected. Recently, a new scheme has been proposed, in order to overcome the problem of cluster mutual attraction forces, by introducing a supplementary term for cluster repulsion [5]. By use of cluster repulsion, as good separation between clusters is obtained, as with the FCM, while keeping the intuitive concept and the noise insensibility introduced by the PCM. The goal of this paper is to establish a connection between the possibilistic approach and cluster validation indices, such that the quantitative superiority of the 'PCM with repulsion' over other methods is tangible.

The organization of this paper is as follows. In Section 2, we analyze the possibilistic approach by [4] developed to cope with the problem of noise and concept of compatibility, but lacks of clusters discrimination. In Section 3, we review the recently proposed method by [5] based on repulsion between clusters, and reports the difficulty of choosing the proper value of the weighting factor γ. In Section 4, we compare four clustering techniques using several datasets and suggest a graphical method to obtain the optimal weighting factor γ; Finally, Section 5 presents our summary and conclusions.

2. Possibilistic Fuzzy Clustering

The most widely used prototype-based clustering method for data partition is probably the ISODATA or Fuzzy C-Means (FCM) algorithm [3]. Given a set of n data patterns, $X = x_1, ..., x_k, ..., x_n.$, the algorithm minimizes a weighted within group sum of squared error objective function. A constraint assures relative numbers for the membership, and therefore is not suitable for applications where memberships are supposed to represent typicalities or compatibilities. Thus, in the FCM the memberships in a given cluster of two points that are equidistant from the prototype of the cluster can be significantly different and memberships of two points in a given cluster can be equal even thought the two points are arbitrarily far away from each other [6]. This situation is illustrated in Figure 1.

In this example, there are two clusters and a pair of points x,y; which represents an outlier and a noise point respectively.