International Journal of Innovative Computing, Information and Control
Volume **1**, Number **1**, March **2005**

 ISSN 1349-4198
pp. **83–94**

FORMULATIONS OF FUZZY CLUSTERING FOR CATEGORICAL DATA

Kazutaka Umayahara

Graduate School of Knowledge Science
Japan Advanced Institute of Science and Technology
Tatsunokuchi, Ishikawa 923-1292, Japan
uma@jaist.ac.jp

Sadaaki Miyamoto

Faculty of Systems and Information Engineering
University of Tsukuba
Tennodai 1-1-1, Ibaraki 305-8573, Japan
miyamoto@esys.tsukuba.ac.jp

Yoshiteru Nakamori

Graduate School of Knowledge Science
Japan Advanced Institute of Science and Technology
Tatsunokuchi, Ishikawa 923-1292, Japan
nakamori@jaist.ac.jp

Received September 2004; revised December 2004

Abstract. *New formulations of the fuzzy clustering algorithm for categorical data are proposed in this paper. Although fuzzy c-means algorithm usually uses distances from cluster centers, the distances of the memberships weighted with categorical data from the unit vectors are used in our new formulations. Different types of metrics between the weighted membership vectors and unit vectors are considered for the objective functions. Optimal solutions for the objective functions are derived and illustrative examples are given to show the obtained theoretic results.*
Keywords: Fuzzy *c*-means, Categorical Data, Multiset

1. **Introduction.** In this paper, new formulations of fuzzy *c*-means algorithm for categorical data on the basis of multisets are proposed. Fuzzy *c*-means [1], which is the fuzzified algorithm of crisp *c*-means [2, 3], is one of the best known methods of fuzzy non-hierarchical clustering. It is mainly applied to continuous data, because it minimizes the sum of all distances between cluster centers and objects. Since fuzzy *c*-means algorithms use distances in the data space, it is difficult to use them for categorical data. First, the data space consists of different types of variables; also, since categorical data are given by natural numbers, distances become discrete thus fuzzy *c*-means algorithm is not suitable for categorical data.

In the proposed algorithm, the clustering is based on a different data space than in the original fuzzy *c*-means algorithm. New metrics proposed in this paper consider membership values as vectors. The vectors are weighted with the categorical data, and the metrics are defined between the weighted vectors and unit vectors. We will also discuss