International Journal of Innovative Computing, Information and Control
Volume **5**, Number **12(A)**, December **2009**

ICIC International ©2009 ISSN 1349-4198
pp. 4393–4405

CLUSTERING CATEGORICAL DATA BASED ON COMBINATIONS OF ATTRIBUTE VALUES

Hee-Jung Do and Jae Yearn Kim

Department of Industrial Engineering
Hanyang University
17 Hangdang-dong, Seongdong-ku, Seoul, 133-791, Republic of Korea
{ hyhjd4; jyk }@hanyang.ac.kr

Received October 2008; revised February 2009

Abstract. *Clustering is an important technique for exploratory data analysis. While most of the earlier clustering algorithms focused on numerical data, real-world problems and data mining applications frequently involve categorical data. Here, we propose a new clustering algorithm for categorical data that is based on the frequency of attribute value combinations. Our algorithm finds all the combinations of attribute values in an object, which represent a subset of all the attribute values, and then groups the object using the frequency of these combinations in each cluster. As our algorithm considers all the subsets of attribute values in an object, objects in a cluster have not only similar attribute value sets but also strongly associated attribute values. Also, the proposed algorithm is not the clustering method using the similarity between only two objects, but rather uses the similarity between an object and clusters. Therefore, it provides global information in clustering results. We conducted experiments with real and synthetic data sets to evaluate FAVC. We show that FAVC is more scalable and provides higher quality results than the previous method.*

Keywords: Data mining, Clustering, Categorical data

1. **Introduction.** Clustering is an important technique for exploratory data analysis [1]. Clustering problems occur in a wide range of applications and are an important component of data mining, statistical pattern recognition, machine learning, and information retrieval [2]. In data mining, clustering is widely used for grouping objects into classes or groups so that objects within one group are more similar to each other than those in other groups.

Most of the earlier clustering algorithms focused on numerical data, where the clustering was based on a distance measure such as the Euclidean distance. However, data in the real world or in data mining applications frequently involve categorical data. The clustering of categorical data is a difficult, yet important task [3].

Categorical data consist of attribute sets whose domain is not numeric, and each attribute takes on a set of values that is not ordered. The most important feature of categorical data is that there is no natural distance between two objects, as there is no obvious order in their values [4]. Therefore, the traditional approach of converting categorical data into numeric values does not necessarily produce meaningful results [5]. Also, transforming categorical values to numeric values leads to loss of semantics and waste of storage when the domain of the categorical attribute is large. In addition, Rock algorithm [6] showed that the Euclidean distance can be a poor measure of similarity for categorical attributes.

Therefore, similarity measures such as the Jaccard coefficient have often been used for clustering categorical data instead of the Euclidean distance. However, as the Jaccard coefficient computes the similarity between two objects only, it does not reflect the properties