*International Journal of Innovative Computing, Information and Control*
Volume **2**, Number **1**, February **2006**

ICIC International ©2006 ISSN 1349-4198
pp. **33–40**

# GENETIC DISTANCE MEASURE FOR K-MODES ALGORITHM

Ching-San Chiang

Chien-Kuo Technology University
No.1 Chieh Shou N. Rd., Changhua City 500, Taiwan

Shu-Chuan Chu

Department of Information Management
Cheng-Shiu University
840 Cheng-Cing Rd., NiaoSong Township, Kaohsiung County 833, Taiwan
scchu@csu.edu.tw

Yi-Chih Hsin

Department of Electronic Engineering
Kaohsiung University of Applied Sciences
415 Cheng Kung Rd., Kaohsiung City, Taiwan

Ming-Hui Wang

Department of Automatic Test and Control
Harbin Institute of Technology
Harbin 150001, P. R. China

Received February 2005; revised September 2005

Abstract. *K-means algorithm has been shown to be an effective and efficient algorithm for clustering. However, the k-means algorithm is developed for numerical data only. It is not suitable for the clustering of non-numerical data. K-modes algorithm has been developed for clustering categorical objects by extending from the k-means algorithm. However, no one applies this technique for classification of categorical data. In this paper, the k-modes algorithm is introduced for the classification of categorical objects based on Soybean and Nursery databases. Especially, a genetic algorithm is proposed for designing the dissimilarity measure termed Genetic Distance Measure (GDM) such that the performance of the K-modes algorithm may be improved by 10% and 76% for Soybean and Nursery databases compared with the conventional k-modes algorithm.*
**Keywords:** K-means, K-modes, Genetic algorithm, Categorical data, Numerical data

1. **Introduction.** In the clustering area, k-means algorithm [6,8] has been adopted for wide applications for the data sets with numerical values. The k-means algorithm is a centroid-based clustering technique. The representatives of the clusters are the centers of the clusters. The applications include the clustering of customers for supermarket, clustering of patients for hospital, codebook design for vector quantization [3,7], and the texture segmentation [10]. Huang proposed the k-modes algorithm for clustering the data sets with categorical values [4]. In fact, the original k-modes algorithm is extended from the k-means algorithm by applying the Hamming distance between objects for the categorical data. In k-modes algorithm, the modes are analogue to the means in the