International Journal of Innovative Computing, Information and Control
Volume **6**, Number **6**, June **2010**

 ISSN 1349-4198
pp. **2551–2566**

AN INTELLIGENT INITIALIZATION METHOD FOR THE K-MEANS CLUSTERING ALGORITHM

Jyh-Jian Sheu[1], Wei-Ming Chen[2], Wen-Bin Tsai[1] and Ko-Tsung Chu[3]

[1]Department of Information Management
National Dong Hwa University
1, Sec. 2, Dahsueh Road, Shoufeng, Hualien 97401, Taiwan
jjsheu@mail.ndhu.edu.tw

[2]Institute of Computer Science and Information Engineering
National Ilan University
NO. 1, Sec. 1, Shen-Lung Road, Yilan City 26047, Taiwan
wmchen@niu.edu.tw

[3]Department of Finance
Minghsin University of Science and Technology
1, Hsin Hsin Road, Hsin Feng, Hsinchu 30401, Taiwan
ktc1009@must.edu.tw

Received December 2008; revised June 2009

Abstract. *The K-Means algorithm is possessed of several advantages such as simple conception and stable efficiency for enormous data sets. While K-Means algorithm also has several shortcomings. The selection of initial clusters, decision of cluster number, and elimination of interference of outliers are the three important subjects for improving K-Means. However, most of the proposed methods of literatures treat only one of the three subjects mentioned above. In the paper, we propose a two-phase clustering method by modifying the initialization of K-Means algorithm, which can accomplish the following jobs simultaneously: (1) deciding the proper cluster number automatically, (2) choosing the better initial clusters, and (3) reducing the influence of outliers upon the result of clustering.*

Keywords: Clustering, Algorithms, K-Means, Partitioning method, Cluster number, Data mining

1. **Introduction.** The clustering problem has broad appeal and usefulness in exploratory data analysis, and the procedure of clustering is a basis process to human understanding [15]. The purpose of cluster analysis is to divide the disordered data into several clusters to make the data of the same cluster possess high similarity, and the data of different clusters have great differences [7,11,21,23,27,28,30,33].

The grouping of related objects can be found in various fields [5]. For example, recently, there has been great growth in the amount of commercial data of enterprises. Especially, the data of consumers' behavior are always accumulated rapidly. The cluster analysis can be applied to classify the consumers according to their behavior data. It could help the enterprises to efficiently analyze the customers' behavior so that they could understand more about the customers' demands which became the critical niche of corporate profits. The cluster analysis is contributive to the induction of enormous data, hence, the related clustering methods are extensively used in many fields, such as artificial intelligence, pattern recognition, statistics, compression, machine learning and market research and so on [3,7,10,14,15,17,22,29,34].