INFORMATION
Volume 26, Number 4, pp.201-231

ISSN 1343-4500 eISSN 1344-8994

Classification of Datasets Based on Combination Algorithm of Clustering and Neural Network

Yingfei Yang and Lei Li

Graduate School of Science and Engineering, Hosei University, Koganei, Tokyo 184-8584 Japan

E-mail: yingfei.yang.2i@stu.hosei.ac.jp, lilei@hosei.ac.jp

Abstract

How to get the needed information from the data accurately and effectively for analysis is a hot research topic nowadays. Accurate classification of data is the basis for smooth data analysis. In order to classify data effectively, scholars have proposed some classification algorithms, and the most frequently mentioned one is k-means algorithm. However, in previous studies, scholars have directly determined the number of classes to be classified for the data set to be classified. Therefore, in this paper, a combinatorial algorithm is proposed to improve the classification of data with unknown group classes. The clustering algorithm and neural network are also combined to improve classification accuracy. The main elements of the algorithm proposed in this paper are as follows.

First, one-third of a set of unknown group class data is selected as the sample data. In order to accurately assess the characteristics of a set of anonymous group class data, it is important first to choose a sample of the data. A sample of one-third of the total data set should be sufficient to provide a reliable representation of the entire population. This sample should be selected at random in order to ensure that the results of the assessment are as accurate as possible. The number of classes is determined by using hierarchical clustering method on the sample data. Then, the non-hierarchical clustering k-means method is used to classify the sample data. Finally, the classification results are trained as the training items of the neural network, and then the model generated after the training is used to classify the overall data.

This paper selects three datasets with different kinds, different numbers of variables, and different amounts of data for the experiments and analysis. This paper presents a comprehensive analysis of three distinct datasets. Each dataset has its own unique characteristics, such as its type, the number of variables, and the amount of data contained therein. By leveraging the properties of these datasets, the experiments, and analysis conducted in this paper will provide valuable insights into the data structures and trends contained within. Furthermore, the results from this analysis will serve as a foundation for further research and experimentation. The experimental results show that the combination of clustering algorithm and neural network algorithm will help to improve the accuracy of data classification and identification effectively. This research provides a new way to accurately and effectively perform data classification.

Keywords: hierarchical clustering; k-means; neural network; data classification.