International Journal of Software Innovation
Volume 8 • Issue 2 • April-June 2020

Incremental Hierarchical Clustering for Data Insertion and Its Evaluation

Kakeru Narita, Kyoto Institute of Technology, Kyoto, Japan
Teruhisa Hochin, Kyoto Institute of Technology, Kyoto, Japan
Yoshihiro Hayashi, Nitto Seiko Co., LTD., Kyoto, Japan
Hiroki Nomiya, Kyoto Institute of Technology, Kyoto, Japan

ABSTRACT

Clustering is employed in various fields. However, the conventional method does not consider changing data. Therefore, if the data is changed, the entire dataset must be re-clustered. This article proposes a clustering method to update the clustering result obtained by a hierarchical clustering method without re-clustering when a point is inserted. This article defines the center and the radius of a cluster and determine the cluster to be inserted. The insertion location is determined by similarity based on the conventional clustering method. this research introduces the concept of outliers and consider creating a cluster caused by the insertion. By examining the multimodality of a cluster, the cluster is divided. In addition, when the number of clusters increases, data points previously inserted are updated by re-insertion. Compared with the conventional method, the experimental results demonstrate that the execution time of the proposed method is significantly smaller and clustering accuracy is comparable for some data.

KEYWORDS

Cluster Division, Data Insertion, Hierarchical Clustering, Incremental Clustering, Re-insertion

INTRODUCTION

Advances in computer and network technologies allow us to obtain various information and services via the Internet, e.g., cloud computing (Agrawal, Das, & El Abbadi, 2011). Cloud computing is not dependent on fixed terminals, and is expected to enable handling of considerable data easily. In addition, the Internet of Things (IoT) technology, which can be used to control and obtain data from different devices, is expected to drive change (Gubbi, Buyya, Marusic, & Palaniswami, 2013; Lee & Lee, 2015). For example, IoT is changing manufacturing industries. In cloud manufacturing (He & Xu, 2015), manufacturing equipment is connected and controlled via the Internet, and a considerable amount of data is transmitted over the Internet (Agrawal, Das, & El Abbadi, 2011; Lee & Lee, 2015). Machine learning can be employed to derive useful information from such large amounts of data (Marsland, 2009; Zumel & Mount, 2014).

DOI: 10.4018/IJSI.2020040101

Machine learning can be classified as supervised and unsupervised (Marsland, 2009; Zumel & Mount, 2014). Supervised learning involves labeled input-output pairs. A supervised learning algorithm produces an inferring function to estimate an output from an input. In contrast, unsupervised learning attempts to find the underlying structure in a set of data points. Note that labeled data are not required for unsupervised learning; thus, unsupervised learning is useful when labeled training data are unavailable. In this paper, we focus on clustering, which is an important unsupervised learning technique.

Clustering, which is used for the analysis and classification of various data, involves generating groups, i.e., clusters, that have similar characteristics. However, the conventional clustering method does not consider changes in the data. Therefore, when new data are inserted, all data must be re-clustered, which is crucial in dynamic environments. When the amount of updating is small, the classification result does not change significantly. However, re-clustering requires considerable time; therefore, a method to partially change clusters without re-clustering when only a small amount of data is inserted is required.

Several incremental clustering methods have been proposed previously (Can, 1993; Charikar, Chekuri, Feder, & Motwani, 2004; Ester, Kriegel, Sander, Wimmer, & Xu, 1998; Gupta & Ujjwal, 2013), which attempt to update a single cluster or a few clusters locally when a data point is inserted, and such methods attempt to store the inserted data point in a cluster that has maximum similarity to it. Although these methods have achieved good performance, they are based on non-hierarchical clustering methods and their application is quite limited. Ribert et al. (1999) proposed an incremental hierarchical clustering method that attempts to find the best insertion point for a new data item. With this method, memory cost is very low; however, the number of computation of distances between clusters does not decrease significantly. Thus, a method that involves the computation of fewer distances is required. Gurrutxaga et al. (2009) proposed another incremental hierarchical method. This method reduced time complexity, but accuracy is worse than the conventional method in some cases. Therefore, improving clustering accuracy is required.

In a previous study, we proposed an incremental hierarchical clustering method (Narita, Hochin, & Nomiya, 2018). To avoid re-clustering of the entire data when a new point is inserted, the incremental hierarchical clustering method only updates some of the data points in an existing clustering result. Here, we define two features, the center and the radius of the cluster, and use these features to determine the cluster into which new data will be inserted. The location point is determined by calculating the similarity based on the original hierarchical clustering method. In addition, this previous method employs the concept of outliers and considers the creation of a new cluster due to data insertion. However, the concordance rate degrades due to the difference in the number of clusters.

In this study, to address this problem, we attempt to improve the previous incremental clustering method (Narita, Hochin, Hayashi, & Nomiya, 2019). In the improved method, we move some points in a given cluster to a different cluster by determining if the given cluster is multimodal. In addition, when the number of clusters increases, the previously inserted points are updated by re-insertion. We compare the proposed and conventional methods using seven datasets including five real datasets. The experimental results demonstrate that, compared to the existing method's re-clustering time, the proposed method can classify points faster when inserting new points.

The remainder of the paper is organized as follows. Section 2 describes work related to incremental clustering methods. Section 3 proposes an improved incremental clustering method. Section 4 compares the proposed method with Ward's method (Ward, 1963) relative to the execution time, concordance rate, and overall accuracy. Section 5 discusses the result. Finally, Section 6 provides the conclusions and suggestions for future work.