Handling Big Data with Fuzzy Based Classification Approach

Neha Bharill and Aruna Tiwari

Department of Computer Science and Engineering
Indian Institute of Technology
Indore, India
{phd12120103,artiwari}@iiti.ac.in

Abstract. Big data is a collection of very large and complex data that is difficult to load into the computer memory. The major challenges include searching, categorization and analysis of big data. In this paper, a fuzzy based supervised classifier is proposed to handle the searching, storage and categorization of big data. In this classifier, we proposed a Random Sampling Iterative Optimization Fuzzy c-Means (RSIO-FCM) clustering algorithm which partitions the big data into various subsets. These subsets adequately cover all the instances (object space) of big data. Then, clustering is performed on these subsets by feeding forward the centers of clustered subset to group remaining subsets. Further, the designed classifier based on Bayesian theory is used to assign the labels to these clusters and also used to predict labels of unknown instances. Thus, the proposed approach results in effective clusters formation which also eliminates the problem of overlapping cluster centers faced by algorithm discussed in [1] named as Simple Random Sampling plus Extension FCM (rseFCM). The effectiveness of proposed clustering algorithm over rseFCM clustering is evaluated on two very large benchmark datasets in terms of fuzzification parameter m, objective function, computational time and accuracy. Experimental results demonstrate that, the RSIO-FCM algorithm generates more appropriate cluster centers location due to which it achieves better classification accuracy as compared to the rseFCM algorithm. Thus, it observed that, cluster centers location will have significant impact over classification results.

1 Introduction

Big data is a collection of vast amount of data that is difficult to handle with existing computer memory [1]. The abundant amount of information generated from the various social networking sites, especially Facebook [1] alone logs produced 25 terabytes (TB) of data per day. Handling of such big data with existing resources is the major challenge. The challenges include categorization, searching and sharing and visualization of big data with limited resources. Clustering and classification are primary tasks used in pattern recognition to effectively handle the storage, searching and categorization of data from very large (VL) databases. Clustering [2] is a process of grouping data into manageable parts such that, the

M. Jamshidi et al. (eds.), *Advance Trends in Soft Computing WCSC 2013*,
Studies in Fuzziness and Soft Computing 312,
DOI: 10.1007/978-3-319-03674-8_21,

samples in each group share some similarity with each other. Classification is used for categorization or identification of samples, which determines the set of categories, the new observation or sample belongs to. Hence, both the approaches jointly handle these challenges associated with big data.

Many algorithms have been proposed to perform clustering of big data but, only few of them address the fuzzy clustering problems. The literal FCM, clusters the entire dataset and works well only for small range datasets. In contrast to the literal schemes, simple random sampling plus extension FCM (rseFCM)[1] approach is designed to perform clustering on VL data. It works on representative samples taken from very large data to perform clustering. However, the rseFCM suffers from overlapping cluster centers because the representative sample does not cover all the objects present in VL data, thus the big dataset cannot be grouped appropriately.

In this paper, the problem of overlapping cluster centers and challenges associated with big data are overcome with Random Sampling Iterative Optimization Fuzzy c-Means (RSIO-FCM) algorithm. The proposed algorithm generates various subsets of big data. These subsets covers all the objects present in big data. Then, it performs clustering on these subsets by feeding forward the cluster centers location of one subset to group remaining subsets. Thus, it generates non overlapping cluster centers location and works significantly well for big datasets. Finally, the designed classifier which is dependent on the cluster centers and based on the concept of Bayesian theory is used to predict the class labels of unknown samples. The designed classifier is dependent on the cluster centers therefore, cluster centers location will have significant impact over classification results.

Section 2 describes the Review of related work. Section 3, describes the proposed RSIO-FCM clustering algorithm for very large data. Then, we apply the clustering results of both the approaches on classification mechanism based on Bayesian theory. In section 4, we perform experimentation with two very large datasets for demonstrating the effectiveness of proposed approach. Finally, section 5 is presented with concluding remarks.

2 Review of Related Work

Many methods have been proposed by researchers for clustering very large data. Generally, these methods are based on various types of algorithms. Sampling methods, that compute cluster centers on sampled data which is randomly selected from huge dataset include CLARA [3], CURE [4], and the coresets algorithms [5]. These algorithms works well for crisp partitions. Methods that work well to produce fuzzy partition include the *fast* FCM (FFCM) [6], in which literal FCM as discussed in algorithm 1, is iteratively applied for larger nested samples till the change is reflected in the solution; and the multistage random FCM [7] which combines FFCM till the final run of FCM on the full dataset. These algorithm are based on extension of literal fuzzy c-means clustering as discussed in algorithm 1.