Available online at www.sciencedirect.com

ScienceDirect

Fuzzy Sets and Systems 258 (2015) 117–133

FUZZY sets and systems

www.elsevier.com/locate/fss

Parallel sampling from big data with uncertainty distribution

Qing He [a], Haocheng Wang [a,b,*], Fuzhen Zhuang [a], Tianfeng Shang [a,b], Zhongzhi Shi [a]

[a] *Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing 100190, China*

[b] *University of Chinese Academy of Sciences, Beijing 100049, China*

Available online 24 February 2014

Abstract

Data are inherently uncertain in most applications. Uncertainty is encountered when an experiment such as sampling is to proceed, the result of which is not known to us while leading to variety of potential outcomes. With the rapid developments of data collection and distribution storage technologies, big data have become a bigger-than-ever problem. And dealing with big data with uncertainty distribution is one of the most important issues of big data research. In this paper, we propose a Parallel Sampling method based on Hyper Surface for big data with uncertainty distribution, namely PSHS, which adopts a universal concept of Minimal Consistent Subset (MCS) of Hyper Surface Classification (HSC). Our inspiration for handling uncertainties in sampling from big data depends on (1) the inherent structure of the original sample set is uncertain for us, (2) boundary set formed of all the possible separating hyper surfaces is a fuzzy set and (3) the uncertainty of elements in MCS. PSHS is implemented based on MapReduce framework, which is a current and powerful parallel programming technique used in many fields. Experiments have been carried out on several data sets including real world data from UCI repository and synthetic data. The results show that our algorithm shrinks data sets while maintaining identical distribution, which is useful for obtaining the inherent structure of the data sets. Furthermore, the evaluation criterions of speedup, scaleup and sizeup validate its efficiency.

Keywords: Fuzzy boundary set; Uncertainty; Minimal consistent subset; Sampling; MapReduce

1. Introduction

In many applications, data contain inherent uncertainty. The uncertainty phenomenon emerges owing to the lack of knowledge about the occurrence of some event. It is encountered when an experiment (sampling, classification, etc.) is to proceed, the result of which is not known to us; it may also refer to variety of potential outcomes, ways of solution, etc. [1]. Uncertainty can also arise in categorical data, for example, the inherent structure of a given sample set is uncertain for us. Moreover, the role of each sample in the inherent structure of the sample set is uncertain.

[*] Corresponding author at: Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing, 100190, China.

E-mail addresses: heq@ics.ict.ac.cn (Q. He), wanghc@ics.ict.ac.cn (H. Wang), zhuangfz@ics.ict.ac.cn (F. Zhuang), shangtf@ics.ict.ac.cn (T. Shang), shizz@ics.ict.ac.cn (Z. Shi).

http://dx.doi.org/10.1016/j.fss.2014.01.016

0165-0114/© 2014 Elsevier B.V. All rights reserved.

Fuzzy set theory developed by Zadeh [2] is a suitable theory that proved its ability to work in many real applications. It is worth noticing that fuzzy sets are a reasonable mathematical tool for handling the uncertainty in data [3].

With the rapid developments of data collection and distribution storage technologies, big data have become a bigger-than-ever problem nowadays. Furthermore, there is a rapid growth in the hybrid study which connects the uncertainty and big data together. And dealing with big data with uncertainty distribution is one of the most important issues of big data research. Uncertainty in big data brings an interesting challenge as well as opportunity. Many state-of-the-art methods can only handle small scale of data sets, therefore, parallel process big data with uncertainty distribution is very important.

Sampling techniques, which play a very important role in all classification methods, have attracted amounts of research in the area of machine learning and data mining. Furthermore, parallel sampling from big data with uncertainty distribution becomes one of the most important tasks in the presence of the enormous amount of uncertain data produced these days.

Hyper Surface Classification (HSC), which is a general classification method based on Jordan Curve Theorem, is put forward by He et al. [4]. In this method, a model of hyper surface is obtained by adaptively dividing the samples space in the training process, and then the separating hyper surface is directly used to classify large database. The data are classified according to whether the number of intersections with the radial is odd or even. It is a novel approach which has no need of either mapping from lower-dimensional space to higher-dimensional space or considering kernel function. HSC can efficiently and accurately classify two and three dimensional data. Furthermore, it can be extended to deal with high dimensional data with dimension reduction [5] or ensemble techniques [6].

In order to enhance HSC performance and analyze its generalization ability, the notion of Minimal Consistent Subset (MCS) is applied to the HSC method [7]. MCS is defined as consistent subset with a minimum number of elements. For HSC method, the samples with the same category and falling into the same unit which covers at most samples from the same category make an equivalent class. The MCS of HSC is a sample subset combined by selecting one and only one representative sample from each unit included in the hyper surface. As a result, some samples in the MCS are replaceable, while others are not, leading to the uncertainty of elements in MCS. MCS includes the same number of elements, but the elements may be different samples. One of the most important features of MCS is that it has the same classification model as the entire sample set, and can almost reflect its classification ability. For a given data set, this feature is useful for obtaining the inherent structure which is uncertain for us. MCS is correspond to many real world problems, like classroom teaching. Specifically, the teacher explains some examples which is the Minimal Consistent Subset of various types of exercises at length to his students, then the students having been inspired will be able to solve the related exercises. However, the existing serial algorithm can only be performed on a single computer, and it is difficult for this algorithm to handle big data with uncertainty distribution. In this paper, we propose a Parallel Sampling method based on Hyper Surface (PSHS) for big data with uncertainty distribution to get the MCS of the original sample set whose inherent structure is uncertain for us. Experimental results in Section 4 show that PSHS can deal with large scale data sets effectively and efficiently.

Traditional sampling methods on huge amount of data consume too much time or even cannot be applied to big data due to memory limitation. MapReduce is developed by Google as a software framework for parallel computing in a distributed environment [8,9]. It is used to process large amounts of raw data such as documents crawled from web in parallel. In recent few years, many classical data preprocessing, classification and clustering algorithms have been developed on MapReduce framework. MapReduce framework is provided with dynamic flexibility support and fault tolerance by Google and Hadoop. In addition, Hadoop can be easily deployed on commodity hardware.

The remainder of the paper is organized as follows. In Section 2, preliminary knowledge is described, including the HSC method, MCS and MapReduce. Section 3 implements the PSHS algorithm based on MapReduce framework. In Section 4, we show our experimental results and evaluate our parallel algorithm in terms of effectiveness and efficiency. Finally, our conclusions are stated in Section 5.

2. Preliminaries

In this section we describe the preliminary knowledge, on which PSHS is based.