Computers, Environment and Urban Systems xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

journal homepage: www.elsevier.com/locate/compenvurbsys

Parallel implementation of Kaufman's initialization for clustering large remote sensing images on clouds

Huiyu Xia [a,b,*], Hassan A. Karimi [b], Lingkui Meng [a]

[a] School of Remote Sensing and Information Engineering, Wuhan University, 129 Luoyu Road, Wuhan 430079, China
[b] Geoinformatics Laboratory, School of Information Sciences, University of Pittsburgh, 135 N Bellefield Ave., Pittsburgh, PA 15260, United States

ARTICLE INFO

Article history:
Available online xxxx

Keywords:
Cloud computing
MapReduce
Remote sensing
K-Means clustering
Kaufman's initialization

ABSTRACT

Common clustering techniques, such as K-Means, for remote sensing images usually suffer from initial starting conditions effects. Kaufman's initialization can provide a set of initial centers of clusters to produce stable and accurate clustering results for remote sensing images. However, the most notable drawback of Kaufman's initialization is that it is computationally expensive and its performance is further challenged when it is applied to large remote sensing images. In this paper, we present a MapReduce-based Parallel Kaufman (MPK) implementation for accelerating the initialization step of clustering. As part of MPK, Grid-based Sequential Systematic Sampling (GS3), a new data partitioning method for remote sensing images, is also presented. GS3, unlike the conventional area-based data partitioning method, is designed specifically for parallel Kaufman implementation. MPK encompasses four key components and was implemented on the Hadoop cluster on a private cloud. Experiments, conducted on a number of remote sensing images with different sizes, show very promising results in terms of significant speedup.

© 2014 Elsevier Ltd. All rights reserved.

1. Introduction

Spatial data is one of the fastest growing types of scientific data and imposes significant computational challenges to scientists and researchers who must cope with processing and analyzing large spatial datasets. With the continuous improvement of sensor technology, the volume of remote sensing images has exploded in recent years and is expected to continually increase. Extracting hidden information from such large remote sensing images for scientific research in areas such as climate, weather, agriculture, ecosystem, biodiversity, and water analysis is computationally expensive (Chen, Chen, Yang, & Di, 2012).

Clustering is a technique widely used in processing large remote sensing images, such as image segmentation, image classification, and image compression. Clustering can be defined as grouping a number of objects in such a way that objects in the same group, called a cluster, have more similarity to each other than to those in other groups. Among existing clustering techniques, K-Means, due to its simplicity, automation and efficiency, remains as one of the most frequently-used techniques in the field of remote sensing data analysis. However, it is well known that K-Means suffers from initial starting conditions effects (Pena, Lozano, & Larranaga, 1999). Adverse effects of inappropriate initial conditions include empty clusters, many iterations, and a higher chance of getting stuck in local minima. As a result, much research has been focused on improving the initialization step of K-Means. Among the several initialization methods reported in literatures, a classical heuristic method called Kaufman's initialization (Kaufman & Rousseeuw, 1990) has been proven to provide, both for general datasets and for remote sensing images, appropriate initial centroids (centers of initial clusters) that produce stable and accurate clustering results with less iterations. However, the most notable drawback of Kaufman's initialization is that it is computationally expensive (He, Lan, Tan, Sung, & Low, 2004); the performance is further challenged when it is applied to large datasets. Although sampling data is considered as one strategy for reducing the running time of Kaufman's initialization (Pena et al., 1999; Zhong & Zhang, 2010), determining a suitable sample size is not straightforward.

Cloud computing is potential for addressing the computational challenge of extracting hidden information in large spatial datasets. It provides high storage, high performance, and flexibility to integrate the essential elements of geospatial sciences (Yang et al., 2011). Of the existing technologies, MapReduce, by Google (Dean & Ghemawat, 2008), is the most widely-used framework

* Corresponding author at: School of Remote Sensing and Information Engineering, Wuhan University, 129 Luoyu Road, Wuhan 430079, China. Tel.: +86 150 7235 1874.

E-mail address: xiahuiyu@whu.edu.cn (H. Xia).

http://dx.doi.org/10.1016/j.compenvurbsys.2014.06.002
0198-9715/© 2014 Elsevier Ltd. All rights reserved.

for data-intensive computing in clouds. Due to its high scalability, reliability and low cost, MapReduce is widely used in both industry (hadoop applications) and academia (Chu et al., 2006; Lee, Lee, Choi, Chung, & Moon, 2011) for processing large datasets.

To address the challenge of Kaufman's performance, in this paper we present an implementation of Kaufman's initialization on a cloud environment using MapReduce, hereafter referred to as MapReduce-based Parallel Kaufman (MPK). The main challenge is that the conventional area-based data partitioning method, which is widely used in parallel processing of remote sensing images (Dhodhi, Saghri, Ahmad, & Ul-Mustafa, 1999; Li, Zhao, & Lv, 2010; Maulik & Sarkar, 2012), is not suitable for Kaufman's initialization, due to its partial representation of sub-blocks. To overcome this problem, we introduce a new data partitioning method and take advantage of MapReduce in processing large remote sensing images.

The contributions of this paper are: (a) a parallel implementation of Kaufman's initialization (MPK) to improve its time performance and (b) a new remote sensing data partitioning method, called Grid-based Sequential Systematic Sampling (GS3). MPK, based on the GS3 method, was evaluated and the results show that it can significantly reduce the computation time as compared to the sequential Kaufman implementation and without loss of accuracy.

The remainder of this paper is organized as follows. Section 2 describes the K-Means initialization problems, Kaufman's initialization algorithm, and related works. Section 3 discusses different MapReduce-based parallelization strategies and the details of GS3. Section 4 describes the details of MPK. Section 5 evaluates the accuracy and time performance of MPK using real-world remote sensing images. Conclusions and future work are discussed in Section 6.

2. Background and related work

Managing, analyzing, and processing spatial data, due to the increasing volume of geospatial data with heterogeneous formats and various types including multi-dimensional geo-vector data and remote sensing images, pose new challenges. In particular, due to the incorporation of the newest generation of sensors to different earth observation platforms, high-resolution remote sensing images are produced over petabyte daily (Chen et al., 2012). Storing and managing such large remote sensing images efficiently and effectively require tackling new challenges. Moreover, remote sensing images are generally structured as georeferenced raster datasets with high spectrum. Analyzing and processing remote sensing images with such properties require more intensive disk I/O and computation compared with processing general unstructured data.

Clustering is widely used in remote sensing image processing such as image segmentation, image classification, and image compression. Among existing clustering techniques, K-Means clustering (Forgy, 1965; MacQueen, 1967) is by far the most popular clustering technique used in scientific and practical applications (Berkhin, 2006) including remote sensing imagery. K-Means begins with K stochastic centroids, typically chosen at random from the dataset. Each object in the dataset is then assigned to the nearest centroid, and each centroid is re-computed as the new centroid of mass of all objects assigned to it. The assignment and re-computing steps are iterated until the process stabilizes (Lloyd, 1982). However, due to variation in initial centroids, which are randomly chosen, K-Means cannot guarantee unique clusters. For this reason, sometimes the clusters obtained through K-Means cannot be relied on for further analysis. K-Means produces reasonable results only when the initial centroids are close to the actual cluster centroids. To produce reasonable results, many methods have been reported in literatures, one of which was proposed by Ball and Hall (1967), another by Tou and Gonzalez (1974); the latter method is called Simple Cluster Seeking (SCS) and is adopted in the FACTCLUS, which is a K-Means variance implemented in SAS.

Katsavounidis, Kuo, and Zhang (1994) suggested a parameterless method, called KKZ, which utilizes the sorted pairwise distances, generally using Euclidean distances, for initialization. KKZ chooses the centroids near the "edge" of the data, by choosing the object with the highest norm as the first centroid. Then, it chooses the next centroid, a point that is farthest from the nearest centroid.

AlDaoud and Roberts (1996) proposed a density-based clustering initialization method which partitions the data uniformly into N cells. From each of these cells, a number of centroids are chosen randomly until K centroids are obtained; the number of centroids is proportional to the number of objects in each cell.

Khan and Ahmad (2004) proposed an algorithm, called CCIA, to find the initial centroids by calculating mean and standard deviation of each dimension of the dataset, and then by separating the data with normal curve into certain partitions to observe the similarity and dissimilarity of objects.

Arthur and Vassilvitskii (2007) proposed the K-Means++ method, which is similar to KKZ, where instead of choosing the farthest point from the already chosen centroids, it chooses an object with a probability proportional to its distance from the already chosen centroids.

Among the several K-Means clustering initialization methods, Kaufman, a classical method introduced by Kaufman and Rousseeuw (1990), is focused in this work. Pena et al. (1999) conducted a comparative study of different initialization methods and found that Kaufman's initialization method outperformed the other methods in terms of effectiveness and robustness in K-Means clustering. Zhong and Zhang (2010) conducted a series of experiments in clustering remote sensing images to evaluate the performance of five different clustering initialization methods. They also found that Kaufman's initialization method helps K-Means obtain accurate and unique clustering results and used it in remote sensing classification projects (Zhong, Zhang, Huang, & Li, 2006).

Kaufman's initialization method heuristically obtains initial centroids by successive selection of representative objects until K objects have been found. The first centroid is chosen to be the most centrally located object in the dataset. The remaining centroids are selected according to the heuristic rule of choosing centroids that are surrounded with a high number of objects around them and are located as far as possible from the previously chosen centroids. The steps of Kaufman's initialization algorithm are:

```
1. Select the most centrally located object as the first centroid
2. For every non-selected object $w_i$
    2.1 For every non-selected instance object $w_j$, calculate
          $C_{ji} = \max(D_j - d_{ji}, 0)$
          $d_{ji}$ is the distance between $w_i$ and $w_j$
          $D_j$ is the distance between $w_j$ and its nearest centroid
    2.2 Calculate $\Sigma_j C_{ji}$
3. Select $w_i$ which maximizes $\Sigma_j C_{ji}$
4. If $K$ objects have been found, stop. Otherwise, return to Step 2.
```

Essentially, Kaufman's initialization is a density estimation method. It estimates the density of the input data through pairwise distance comparison of all contained objects. In the case of remote sensing imagery, without loss of generality, the distance between two pixels is defined as the Euclidean distance between color values of the two pixels. Take a three bands RGB image for example. Let p and q be the two pixels in this image; p has color value

Download English Version:

https://daneshyari.com/en/article/4965239

Download Persian Version:

https://daneshyari.com/article/4965239

Daneshyari.com