Astronomy and Computing 16 (2016) 166–173

Contents lists available at ScienceDirect

Astronomy and Computing

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

Full length article

A communication efficient and scalable distributed data mining for the astronomical data

A. Govada*, S.K. Sahay

Department of Computer Science & Information Systems, BITS-Pilani, K. K. Birla Goa Campus, Goa-403726, India

ARTICLE INFO

Article history:
Received 26 December 2015
Accepted 22 June 2016
Available online 30 June 2016

Keywords:
Distributed data mining
Astronomical data
Principal component analysis
Load balancing

ABSTRACT

In 2020, ∼60PB of archived data will be accessible to the astronomers. But to analyze such a paramount data will be a challenging task. This is basically due to the computational model used to download the data from complex geographically distributed archives to a central site and then analyzing it in the local systems. Because the data has to be downloaded to the central site, the network BW limitation will be a hindrance for the scientific discoveries. Also analyzing this PB-scale on local machines in a centralized manner is challenging. In this, virtual observatory is a step towards this problem, however, it does not provide the data mining model (Zhang et al., 2004). Adding the distributed data mining layer to the VO can be the solution in which the knowledge can be downloaded by the astronomers instead the raw data and thereafter astronomers can either reconstruct the data back from the downloaded knowledge or use the knowledge directly for further analysis. Therefore, in this paper, we present Distributed Load Balancing Principal Component Analysis for optimally distributing the computation among the available nodes to minimize the transmission cost and downloading cost for the end user. The experimental analysis is done with Fundamental Plane (FP) data, Gadotti data and complex Mfeat data. In terms of transmission cost, our approach performs better than Qi et al. and Yue et al. The analysis shows that with the complex Mfeat data ∼90% downloading cost can be reduced for the end user with the negligible loss in accuracy.

1. Introduction

Astronomy is afloat with data and an estimate shows that by 2020 more than 60 PB of archived data will be electronically accessible to astronomers. But the complete analysis of the whole accumulated data distributed globally is challenging, not only due to the volume of data but also the communication cost. In general, the computational model used in astronomy is to download the data from archives to a central site and is analyzed in the local machines. Hence, the network bandwidth limitations will be a hindrance for scientific discoveries and even analyzing this PB-scale on local machines in a centralized manner is challenging (Weske et al., 2007; Skoda, 2007; Anon, 0000; Berriman and Groom, 2011). Understanding the data collection rate (e.g. Large Synoptic Survey Telescope (LSST) will generate ∼30 tera bytes of data every night (Iveric et al., 2011)), the centralize technique will not suffice for comprehensive co-analysis to exploit the potential of distributed archived data (Bhaduri et al., 2009). In this Virtual Observatory (VO) is a step towards the problem, however, it does not provide the data mining model (Chilingarian et al., 2012). In this adding the distributed data mining layer to the VO can be the solution (Dutta and Kargupta, 2006) in which the knowledge can be downloaded by the astronomers instead the raw data, and thereafter astronomers can either reconstruct the data back from the downloaded knowledge or use the knowledge directly for further analysis.

Astronomical data are mostly high dimensions. Hence, reducing the dimensions of interrelated data will reduce the downloading cost of end users. To reduce the dimension of data, a technique called principal component analysis (PCA) is used in many fields (Joliffe, 1986). It reduces the dimensionality of the dataset of interrelated variables with retaining the variation present in the data Pearson (1901) and Hotelling (1933). The technique is linear as its components are linear combinations of the original variables, but non-linearity is preserved in the dataset. In this paper, we use Distributed Load Balancing Principal Component Analysis (DLPCA) which is a distributed version of normal PCA to reduce the transmission (cost incur for distributing the data among the computational nodes) and downloading cost significantly from globally distributed observatories. The algorithm is scalable and also optimally distribute the computational load among the available resources.

* Corresponding author.
E-mail addresses: garuna@goa.bits-pilani.ac.in (A. Govada), ssahay@goa.bits-pilani.ac.in (S.K. Sahay).

http://dx.doi.org/10.1016/j.ascom.2016.06.002
2213-1337/© 2016 Elsevier B.V. All rights reserved.

For the experimental analysis, we took the Fundamental plane (FP), Gadotti and complex Mfeat datasets and use Java Agent Development Framework (JADE), which simplifies the implementation of multi-agent systems through a middle-ware and complies the Foundation for Intelligent Physical Agents specifications. The experimental analysis of our approach is done by creating the multiple agents. The local principal components are computed and communicated among them for computing the global principal component.

This paper is organized as follows. In the next section, we discuss the related work on the analysis of globally distributed astronomical data. Section 3 briefly describe the mathematics of the PCA to provide an intuitive feeling of it. In Section 4 we present our algorithm for the communication efficient and scalable distributed data mining using DLPCA. In Section 5 we present the experimental results. Section 6 discusses the computational cost, load balancing and scalability of our approach. Finally, Section 7 contains the conclusion and future direction of the paper.

2. Related work

Big data analysis are primarily based on Distribute Data Mining (DDM) to process the heterogeneous data from databases located at different places. In literature, various DDM techniques have been proposed for the analysis of heterogeneous datasets. These techniques differ from the centralized data mining in which the analysis is done after downloading the data to one single location. Distributed Data Mining based on PCA can be done in two ways; either distributing the data horizontally or vertically (Dutta et al., 2007; Dutta and Kargupta, 2006). A distributed PCA algorithm based on the integration of local covariance matrices for the distributed databases which are horizontally partitioned is given by Qi et al. (2003). If the data is vertically distributed then it is necessary that all the considered sites are associated with a unique way of matching the distributed data (Dutta and Kargupta, 2006), for e.g in astronomy, it is done using right ascension and declination (RA, DEC) of the objects. A randomized PCA is discussed by Halko et al. (2011), for the datasets which are too large to store in the Random Access Memory (RAM).

Astronomical research communities do data mining for large datasets e.g. F-MASS (The ClassX Project, 2003), the Auton Astrostatistics Projects (The AUTON Project, 1993). However, this project does not fully based on Distribute Data Mining. A project called Grid Based Data Mining for Astronomy (GRIST) (Jacob et al., 2005) was the first attempt for large scale data mining in astronomy. Projects in Virtual Observatories such as Japanese Virtual Observatory (JVO), US National Virtual Observatory (NVO), European Virtual Observatory (EURO-VO) and International Virtual Observatory (IVOA), basically integrate and federate archive systems dispersed in a Grid by standardizing XML schema, data access layer, and query language of archival data (Jacob et al., 2005). In this NVO has developed an information technology infrastructure enabling easy and robust access to distributed astronomical archives, from which users can search and gather data from multiple archives with basic statistical analysis and visualization functions. Dutta and Kargupta (2006) describe the architecture of a system called Distributed Exploration of Massive Astronomy Catalogs (DEMAC) for distributed data mining of large astronomical catalogs. The system is designed to sit on top of the existing national virtual observatory environment to provide tools for distributed data mining without downloading the data to a centralized server. Daruru et al. (2010) proposed a distributed and multi-threaded Automated Hierarchical Density Data in Astronomy: from the Pipeline to the Virtual Observatory clustering algorithm to produce computationally efficient high-quality clusters and scalable from 1-1024 compute-cores. For massive astronomical data analysis distributed CPU/GPU architecture is proposed to handle the data in peta scales (Hassan et al., 2011). Recently a cloud based data mining system CANFAR+Skytree is proposed at Canadian Astronomy Data Centre (Ball, 2013). Kargupta et al. (2001) proposed the solutions for distributed clustering using collective principal component analysis. Their work is mainly focused to obtain a good estimation of the global covariance matrix with a trade-off between communication cost and information loss. Further, Yue et al. (2012) proposed a better DDM than Kargupta et al. (2001) in which one can achieve a better accuracy with the same communication cost. Our proposed algorithm is communication efficient and scalable DDM based on PCA to further reduce the communication cost with better accuracy which neither needs to send the local datasets to a central site nor require to reconstruct the local data for calculating the global principal components.

3. Principal component analysis

Principal component analysis is a simple non-parametric method to reduce the dimension of the datasets of possibly correlated variables into a smaller number of uncorrelated variables. The first component has maximum variability and each remaining component contains rest of the variabilities. It is abundantly used in many fields viz. astronomy, computer graphics etc. In this section, we briefly describe the mathematics of the PCA to provide an intuitive feeling of it. For detailed mathematical illustrations the springer series in statistics "Principal Component Analysis 2e" by I.T. Jollife is a good source (Jolliffe, 2002). Suppose we have a dataset as

$$[\mathbf{X}]_{n\times m} = (X_o, X_1, X_2, X_3, \ldots, X_{l-1})$$

where X_j is a $n \times m_j$ matrix and m_j is the number of columns in X_j.

The covariance matrix of the data can be computed as

$$cov_j^{pq} = \frac{\sum_{i=1}^{i=n}(X_{ji}^p - \mu_j^p)(X_{ji}^q - \mu_j^q)}{n-1}$$

where, μ_j^p, μ_j^q is the mean of the pth and qth column of the X_j matrix.

The obtained covariance matrix will be a square matrix and symmetric. If the two column data are completely uncorrelated, then the covariance will be zero. However, there may be a non-linear dependency between two variables that have zero covariance. As covariance matrix is symmetric, its eigenvalues and eigenvectors can be obtained by solving the equations

$$cov_j^{pq}E = \lambda; \qquad |cov_j^{pq} - \lambda I| = 0$$

where, E is the eigenvector of eigenvalue λ and I is the identity matrix of the same order of cov_j^{pq}. Now a matrix P can be made consists of eigenvectors of the covariances matrices. The computed eigenvectors are ordered according to their significance. To construct the reduced dataset the least significant eigenvectors are left out and computed as

Reduced dataset ($n \times l$ matrix) $=$ Original dataset (say $n \times m$ matrix) $-$ mean $\times$ Reduced eigenvector matrix (say $m \times l$ matrix).

Original dataset can be computed as follows

Original dataset ($n \times m$ matrix) $=$ Reduced dataset ($n \times l$ matrix) $\times$ (Reduced eigenvector matrix)^T ($l \times m$ matrix) $+$ original mean.

where n is the number of rows, m is the number of columns and l is the reduced number of columns.

4. Communication efficient and scalable DDM using distributed load balancing PCA

Our approach is basically a communication efficient and scalable DDM for the analysis of astronomical data. The algorithm is described below in eight steps.