Chapter 124
Distributed Based Serial Regression Multiple Imputation for High Dimensional Multivariate Data in Multicore Environment of Cloud

Lavanya K
Research Scholar, Department of Computer Science & Engineering, JNTUA College of Engineering, Anantapur, India

L.S.S. Reddy
Professor, Department of Computer Science & Engineering, KL University, Guntur, India

B. Eswara Reddy
Professor, Department of Computer Science & Engineering, JNTUA College of Engineering, Anantapur, India

ABSTRACT

Multiple imputations (MI) are predominantly applied in such processes that are involved in the transaction of huge chunks of missing data. Multivariate data that follow traditional statistical models undergoes great suffering for the inadequate availability of pertinent data. The field of distributed computing research faces the biggest hurdle in the form of insufficient high dimensional multivariate data. It mainly deals with the analysis of parallel input problems found in the cloud computing network in general and evaluation of high-performance computing in particular. In fact, it is a tough task to utilize parallel multiple input methods for accomplishing remarkable performance as well as allowing huge datasets achieves scale. In this regard, it is essential that a credible data system is developed and a decomposition strategy is used to partition workload in the entire process for minimum data dependence. Subsequently, a moderate synchronization and/or meager communication liability is followed for placing parallel impute methods for achieving scale as well as more processes. The present article proposes many novel applications for better efficiency. As the first step, this article suggests distributed-oriented serial regression multiple

DOI: 10.4018/978-1-7998-5339-8.ch124

imputation for enhancing the efficiency of imputation task in high dimensional multivariate normal data. As the next step, the processes done in three diverse with parallel back ends viz. Multiple imputation that used the socket method to serve serial regression and the Fork Method to distribute work over workers, and also same work experiments in dynamic structure with a load balance mechanism. In the end, the set of distributed MI methods are used to experimentally analyze amplitude of imputation scores spanning across three probable scenarios in the range of 1:500. Further, the study makes an important observation that due to the efficiency of numerous imputation methods, the data is arranged proportionately in a missing range of 10% to 50%, low to high, while dealing with data between 1000 and 100,000 samples. The experiments are done in a cloud environment and demonstrate that it is possible to generate a decent speed by lessening the repetitive communication between processors.

1. INTRODUCTION

In the realms of Big Data analysis, Medical analysis, network analysis, and image analysis, inadequate multivariate high dimensional data poses the biggest challenge (Chapman et al., 2000; Grosu & Chronopoulos, 2005). In general, these spheres contain data, mostly every sort of variables that are usually inadequate. In this backdrop, Multiple Imputations is considered offering the best possible practice that handles insufficiency in data (Ambler et al., 2007; Dempster et al., 1977; Dempster et al., 1977; Bu et al., 2016). However, the currently available multiple imputation algorithms suffer from issues such as high time complexity, often lacking the good properties that distributed and shared processing possess. Hence, we find them unsuitable to process data in a high-dimensional data ecosystem. Further, in view of the rapid movement of systems that run this task towards exascale, which demands communication and computation patterns needing high programmability (Li et al., 2014; Wang et al., 2015). However, such systems are quite complex, which the current communication models find difficult to locate and productively utilize for computation-communication overlap as High-Performance Computing (HPC) usually suffer from performance as well as programming related issues (Bu et al., 2016; Feng & Balaji, 2009; Humenay et al., 2007). The available communication models are found lacking close energy with multi-threaded programming models that happen in a rough or risky scenario in which communication and multi-threaded components of applications are synchronized. In fact, it is rather tough to program distributed memory systems having huge quantities of parallelism in every node. The available distributed memory models are intended to achieve scalability and communication. But in some particular applications, they are found unsuitable as programming models that leverage overt parallelism. But, shared memory task models are ideally suited to exploit overt parallelism. This paper proposes a set of Distributed algorithms for improving the efficiency of an MI task. A Cloud computing ecosystem is found appropriate for such a system that contains homogeneous as well heterogeneous resources. Moreover, multi-cluster resources have the ability to enhance computing capacity seen while executing a task among workers in data of higher dimensions. This ecosystem needs a suitable resource manage-