International Journal of Information Technologies and Systems Approach
Volume 15 • Issue 1

Evaluation Platform for DDM Algorithms With the Usage of Non-Uniform Data Distribution Strategies

Mikołaj Markiewicz, Warsaw University of Technology, Poland
Jakub Koperwas, Warsaw University of Technology, Poland

ABSTRACT

Huge amounts of data are collected in numerous independent data storage facilities around the world. However, how the data is distributed between physical locations remains unspecified. Downloading all of the data for the purpose of processing is undesirable and sometimes even impossible. Various methods have been proposed for performing data mining tasks, but the main problem is the lack of an objective strategy for comparing them. The authors present current research on a novel evaluation platform for distributed data mining (DDM) algorithms. The proposed platform opens up a new field to evaluate algorithms in terms of the quality of the results, transfer used, and speed, but also for the use of a non-uniform data distribution among independent nodes during algorithm evaluation. This work introduces a 'data partitioning strategy' term referring to a specific, not necessarily uniform data distribution. A brief evaluation for three clustering algorithms is also reported showing the usability and simplicity of identifying differences in processing with the use of the platform.

KEYWORDS

Algorithm Evaluation, benchmarking Platform, Classification, Clustering, Data Partitioning Strategies, Distributed Data Mining, Distributed Processing

INTRODUCTION

Various data mining methods have been presented in our world for many years. They play an important role in our lives, in both the business and scientific domains. The continuous growth of data collected in the world increases every day, even reaching a size on the order of zettabytes (where a ZB is 10^{21} B) (Chen et al., 2014). It is estimated to reach from 168 (Khattak et al., 2019) to 175 (Alnoukari, 2020) ZB by 2025. Moreover, countless devices that form the Internet of Things (IoT) produce quintillions of bytes (10^{18} B) (Dobre & Xhafa, 2014) of data every day as well. In fact, 500 ZB of data was generated from such IoT sources in 2020 (Mirza et al., 2021) due to their number. The number of such devices increases, and the data cannot be stored in one place because of hardware and network limitations. On the other hand, multiple independent data centers collect such data for different institutions and companies. Therefore, new algorithms and methods are required to enable new findings in this large and distributed amount of data. Distributed data mining (DDM) has been gaining importance in recent years, with various methods implemented (Gan et al., 2017). Depending on the domain of study, these approaches are called Collaborative Systems (e.g., Zhou et al., 2010) or DDM methods and frameworks. Such processing methods should provide both quality results and processing performance while preserving the privacy of transferred data.

DOI: 10.4018/IJITSA.290000

Much research on applications implementing various algorithms working in a Spark (Zaharia et al., 2012) cluster has been done. However, such processing requires a specific cluster environment and does not provide control over the processing and data transfer between nodes. Because of this lack of control, which will be discussed later, custom approaches and frameworks are usually implemented utilizing well-known worker-node communication architectures, such as centralized or peer-to-peer (P2P). On account of this expansion, for several years a great effort has been devoted to the study of distributed clustering, classification, and other data mining methods. One of the first examples is presented by Aouad et al. (2007), and a newer solution is presented by Bendechache and Kechadi (2015). Distributed classification methods have been studied by Navia-Vázquez et al. (2006) and Forero et al. (2010) and have even been studied explicitly in the network security domain (Hu et al., 2013). It is worth noting that recently, several authors (Forero et al., 2010; Xu et al., 2015; Jia et al., 2016) have pointed out the importance of preserving privacy in DDM processing. This processing concept, however, is not limited to the previously mentioned types of algorithms; it also applies to other methods, like frequent itemset mining, etc. Nevertheless, for the purposes of this work, we do not focus on these topics.

In recent years, research into what was described in the previous section—a real independent distributed approach working in a cooperative way—has become very popular. Due to data distribution, such an approach has gained strong interest, creating a new path of development alongside cluster-called applications like Spark. This problem is strongly related to orchestration challenges based on the chosen communication architecture rather than horizontal application scaling in order to increase the overall computational performance with data exchange between nodes. Several architectures are already used in which small independent computational nodes create the global processing result, not exposing all of their data in favor of partial results. In fact, in such an approach, the global model being prepared from multiple meta-models is also called, in the agent system literature, the agent and artifact paradigm (Omicini et al., 2008). In the extreme example of this approach, the agent node may contain knowledge only about a specific slice of the whole dataset. In cooperation with other independent nodes, it is able to deal with a complete set of previously unknown samples arriving into the system. Those samples might be unknown anomalies or clusters for clustering tasks, new classes of samples for classification, or even associated rules discovered by other nodes. Thanks to such cooperative work, every node can better respond to the new data that is applied. Depending on the architecture, there can be a centralized execution with one central coordinator (Januzaj et al., 2004; Markiewicz & Koperwas, 2019), fully distributed node cooperation, or a hierarchical method for results aggregation (Zhou et al., 2010; Bendechache et al., 2019). Our current research is focused on a centralized platform, and therefore, in this paper, this approach has been emphasized. This is due to the fact that this approach, with minimal effort and approximation, is adaptable to the hierarchical concept as well, which will be mentioned later.

Previously described methods and approaches are extensively used in subsequent studies. However, there is an ongoing discussion concerning how to verify the correctness of an algorithm in a standard way. The need arises to test it against other methods in similar conditions. It is also not easy to find a proper dataset to evaluate a given method. Nevertheless, when the dataset is present, one question that is still unanswered is how to scatter such data to show that the algorithm works without a uniform data distribution. That is why in this paper, a new technique that improves distributed data mining methods evaluation has been suggested. No analyst can be certain of a comprehensive data representation at every independent node in the real world. There are countless possibilities to spread an entire dataset among multiple nodes. A typical algorithm evaluation is performed using a random, uniform data distribution or a cross-validation method. For strongly unbalanced sets, the number of class samples during scattering is usually also considered to simulate an even class distribution. However, in a completely distributed environment, the entire dataset cannot be accessed or is deliberately restricted for privacy reasons. In order to fully examine the algorithm, the algorithm should also be evaluated on a specific data distribution in spite of typical scenarios. The non-typical