Simulation Modelling Practice and Theory 000 (2017) 1–13

Contents lists available at ScienceDirect

Simulation Modelling Practice and Theory

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

A new web-based solution for modelling data mining processes

Viktor Medvedev[a,*], Olga Kurasova[a], Jolita Bernatavičienė[a], Povilas Treigys[a,b], Virginijus Marcinkevičius[a], Gintautas Dzemyda[a]

[a] Vilnius University, Institute of Mathematics and Informatics, Akademijos str. 4, LT-08663 Vilnius, Lithuania
[b] Vilnius Gediminas Technical University, Faculty of Fundamental Science, Saulėtekio avn. 11, LT-10223 Vilnius, Lithuania

ARTICLE INFO

Article history:
Available online xxx

Keywords:
Data mining
Scientific workflow
Modelling data mining process
Dimensionality reduction
Cloud computing
High-performance computing

ABSTRACT

The conventional technologies and methods are not able to store and analyse recent data that come from different sources: various devices, sensors, networks, transactional applications, the web, and social media. Due to a complexity of data, data mining methods should be implemented using the capabilities of the Cloud technologies. In this paper, a new web-based solution named DAMIS, inspired by the Cloud, is proposed and implemented. It allows making massive data mining simpler, effective, and easily understandable for data scientists and business intelligence professionals by constructing scientific workflows for data mining using a drag and drop interface. The usage of scientific workflows allows composing convenient tools for modelling data mining processes and for simulation of real-world time- and resource-consuming data mining problems. The solution is useful to solve data classification, clustering, and dimensionality reduction problems. The DAMIS architecture is designed to ensure easy accessibility, usability, scalability, and portability of the solution. The proposed solution has a wide range of applications and allows to get deep insights into the data during the process of knowledge discovery.

1. Introduction

Data mining is an important part of the processes of knowledge discovery in medicine, economics, finance, telecommunication, and various scientific fields. Data mining helps to uncover hidden information from an enormous amount of data that are valuable for recognition of important facts, relationships, trends, and patterns. Moreover, data mining techniques are effective in the face of modelling and simulation tasks. For several decades, the attention was focused on new data mining methods, and software was developed to implement these methods [1–3]. However, most of the widely used software solutions were designed as standalone desktop applications. They include methods for data pre-processing, classification, clustering, regression, association, and dimensionality reduction [2]. The application of these data mining methods can uncover non-trivial knowledge from the simulated and real-world data.

Recently, the amount of data collected and stored across the world has been increasing at the exponential rate. The data are being produced by various devices, sensors, networks, transactional applications, the web, and social media. Usually, the data are large scale and heterogeneous. The conventional technologies and methods, available to store and analyse the

* Corresponding author.
E-mail address: viktor.medvedev@mii.vu.lt (V. Medvedev).

http://dx.doi.org/10.1016/j.simpat.2017.03.001
1569-190X/

data, cannot work efficiently with such an amount of them. Moreover, the Cloud gives new technological opportunities for these methods [2,4,5]. New technologies have boosted the ability to store, process and analyse the massive data. As Cloud-based technologies and platforms gain in popularity, new data mining and machine learning algorithms have been developed as the Cloud services. Moreover, another new trend known as big data brings new challenges to data mining due to large volumes and different varieties of data. The common methods and tools for data processing and analysis are unable to manage such data by conventional ways. Thus, the Cloud and big data not only yield new data storage and processing mechanisms but also introduce ways of the intelligent data analysis.

Due to the complexity of data, various data mining methods should be used jointly. The goal to make massive data mining simpler, effective, and easily understandable for data scientists and business intelligence professionals can be achieved by constructing scientific workflows for data mining process using a drag and drop interface. The usage of scientific workflows allows composing the convenient model of data mining process covering a number of different methods. Thus, the simulation and solving of real-world time- and resource-consuming data mining problems may be realised. Inspired by the aforementioned challenges facing data scientists and business intelligence professionals, a new web-based solution DAMIS has been developed.

The paper is structured as follows. In Section 2, the related works on the Cloud technologies and the state-of-the-art data mining solutions are reviewed. Section 3 introduces a new web-based data mining solution DAMIS. The capabilities of DAMIS to solve various data mining tasks are demonstrated in Section 4. The last section concludes the paper.

2. Related works

Data mining algorithms and processes of knowledge discovery are usually compute- and data-intensive [1,6–8]. The Cloud offers a computing and data management infrastructure to support a decentralised and parallel data analysis. The innovative Cloud solutions are aimed at making data mining and knowledge discovery process more attractive and straightforward.

Extracting useful knowledge from huge data requires intelligent and scalable analytics services, programming tools, and applications [9]. The Cloud allows the user to obtain various services from data storage to data mining without investing in the architecture. In the last years, many standard data mining algorithms have been migrated to the Cloud that make them high efficient and scalable [10]. The growing number of real-world applications, such as recommendation systems [11,12] and health-care systems[13,14], shows the high significance of this approach.

2.1. Cloud technologies

The Cloud technologies become a major tool for new solutions of data mining and innovation in various fields of science and business. There is a large number of distributed and data-intensive applications and data mining algorithms that cannot be fully exploited without the Cloud technologies [15–18]. The world's leading information technology (IT) companies conduct the research which becomes significantly related to the Cloud. The basic Cloud service types aggregate Infrastructure as a Service (IaaS), Platform as a Service (PaaS), Software as a Service (SaaS). As big data analytics becomes mainstream in the era of new technologies, Analytics as a Service (AaaS) is designed to help data scientists and business intelligence professionals to meet the growing demand for data analysis and research [19]. In order to use modelling and simulation on demand in the Cloud, a Modelling and Simulation as a Service (MSaaA) is introduced [20].

Cloud computing provides a possibility to access the distributed computing environments that can utilise computing resources on demand [2,21,22]. Cloud-based data mining allows distributing a compute-intensive data analysis among a large number of remote computing resources. Common software for Cloud computing has been based on a Service Oriented Architecture (SOA). It describes a set of principles allowing to build flexible, modular, and interoperable software applications. The implementation of SOA is represented by web services. A web service is a collection of functions that are packed and presented as a single entity published on the network and can be used by other applications through a standard network communication protocol [23]. The web service allows integrating heterogeneous platforms and applications. The services are running independently in the system, and external components do not know how the services implement the functionality. The components ensure that the services should return the expected results. So, web services are widely used for computing on demand [24]. WSDL (Web Service Definition Language), SOAP (Simple Object Access Protocol) and REST (REpresentational State Transfer) are concepts important to define web service oriented solutions.

The most known new products and services are Cloud-based solutions: Apache Hadoop and Spark, Amazon Web Services (AWS), Microsoft Azure, Google Cloud Platform, and others. Apache Hadoop[1] is one of open source Cloud computing environments [25] that implements the Google MapReduce framework. MapReduce is a programming model for processing large data sets as well as a running environment in large computer clusters. Due to HDFS (Hadoop Distributed File System) Hadoop enables to save the computing time needed for sending data from one computer to another. The Hadoop Mahout[2] library is developed for data mining, where classification, clustering, regression, and dimensionality reduction algorithms are

[1] http://hadoop.apache.org
[2] http://mahout.apache.org

Please cite this article as: V. Medvedev et al., A new web-based solution for modelling data mining processes, Simulation Modelling Practice and Theory (2017), http://dx.doi.org/10.1016/j.simpat.2017.03.001