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

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

implemented. However, there are only a few data mining algorithms implemented due to the MapReduce limitations, and not always the existing data mining algorithms can be used easily and efficiently[3].

Apache Spark[4] is another open source parallel processing framework that supports in-memory processing to boost the performance of big data analysis tasks. Spark is designed to work with HDFS to improve the MapReduce technology. Spark makes it easier for developers and data scientists to work with data and deliver advanced insights faster. For data mining purposes, data scientists can use scalable machine learning library MLlib, where traditional machine learning and statistical algorithms have been implemented. Spark is a lot faster than MapReduce because of the way it processes the data.

Amazon Web Services (AWS)[5] offer reliable, scalable, and inexpensive Cloud computing services: Amazon Elastic Compute Cloud (EC2), Amazon Simple Storage Service (S3). Amazon Elastic MapReduce (EMR) enables to process large amounts of data. It uses a Hadoop framework running on the web-scale infrastructure of EC2 and S3. It is possible to use different capacities to perform data-intensive tasks for applications in data mining, machine learning, scientific simulation, web indexing, and bioinformatics. Google has also introduced an on-line service to process large volumes of data[6]. The service supports ad hoc queries, reports, data mining, or even web-based applications. Microsoft Azure[7] is another Cloud computing platform and infrastructure to build and manage applications and services through a global network. It provides both PaaS and IaaS services and supports many different programming languages, tools, and frameworks.

2.2. Data mining solutions

At the beginning of solving of data mining problems, the methods were rapidly developed by adapting mathematical statistics methods and creating new ones inspired by modern applications. Later on, the data mining software was developed to facilitate solving the data mining problems. The majority of software is open sourced and available for free. Therefore it has become very popular among data scientists.

Recently, a scientific workflow paradigm becomes widely used in the software with the user-friendly interface [26,27]. The functionality of scientific workflows allows researchers to compose and execute a series of data analysis and computation procedures in scientific applications. Moreover, the scientific workflows are computational steps for scientific simulations and data analysis processes. The development of scientific workflows is under the influence of e-science technologies and applications [28–30]. The aim of e-science is to enable researchers to collaborate when carrying out a large scale of scientific experiments and knowledge discovery applications, using distributed systems of computing resources, devices, and data sets [31]. Scientific workflows play a major role to reach this aim. First of all, the scientific workflows can be extremely helpful to compose convenient platforms for experiments by retrieving data from databases and data warehouses and running data mining algorithms in the Cloud infrastructure. Secondly, web services can be easily imported as a new component of workflows. Thus, the scientific workflows provide an easy-to-use environment for researchers to design their workflows for individual applications, to execute the workflows and to view the results in real time.

The scientific workflows provide multiple benefits [32]. They are useful for sharing knowledge as services for collaborating scientists. The workflows are able to deal with big data problems and with modelling and simulation of data mining tasks. The workflows can be utilised to conduct scientific simulations in a parallel and automated manner.

The synergy of Cloud computing possibilities and scientific workflow paradigm allows developing new scalable, extensible, interoperable, modular, and easy-to-use data mining solutions. One of the most popular open source data mining software is Weka [33]. It contains tools for data pre-processing, classification, regression, clustering, association rules, and visualisation. Weka4WS is an extension of Weka to support distributed data mining [34]. It is a desktop application with a possibility to select remote computing resources.

Orange4WS [35] is an extension of another well-known data mining system – Orange [36]. Orange4WS includes some new interesting features comparing with Orange. There is a possibility to import external web services; only WSDL file location should be specified. The knowledge discovery ontology describes workflow components (data, knowledge and data mining services) in an abstract and machine-interpretable way.

The open source KNIME analytics platform [37] helps to discover the hidden information in data and to predict new values. There is a possibility to use a KNIME commercial extension, KNIME Cloud Server, to run analytics on more powerful hardware by offloading computationally intensive tasks to dedicated hardware.

RapidMiner Studio [38] offers a powerful, easy-to-use and intuitive graphical user interface for the design of analytic processes including data uploading from various sources, data pre-processing, model building and validation, as well as result visualisation. The commercial extensions RapidMiner Cloud and RapidMiner Radoop are available for executing high-performance, large scale predictive analytics on demand.

All the aforementioned data mining systems are standalone desktop applications. Nowadays web applications become more popular due to the ubiquity of web browsers. ClowdFlows is a web application based on a service-oriented data

[3] http://mahout.apache.org
[4] http://spark.apache.org
[5] http://aws.amazon.com
[6] https://cloud.google.com
[7] https://azure.microsoft.com

mining tool [39,40]. It is an open source Cloud-based platform for composition, execution, and sharing of interactive machine learning workflows. Here the data mining algorithms from Orange and Weka are implemented as local services.

DAME (DAta Mining & Exploration) is an innovative web-based, distributed data mining infrastructure[8], specialized in large data sets exploration [41]. DAME is organised as the Cloud of web services and applications. The idea of DAME is to provide a user-friendly and standardised scientific gateway to ease the access, exploration, processing, and understanding of large data sets. The DAME system includes not only web applications but also several web services, aimed at providing a wide range of facilities for different e-science communities.

The world's leading IT companies offer Cloud-based products and solutions for data mining. Microsoft Azure Machine Learning (ML)[9] is a Cloud service to build predictive analytics models and to easily deploy those models for consumption as the Cloud web services. A series of machine learning methods are implemented for data preparation, feature selection, anomaly detection, data classification, clustering, regression as well as for statistical functions and text analytics. With a browser at hand only, it is possible to upload data, and immediately start machine learning experiments. Azure ML enables to design machine learning workflows in the Cloud directly from the browser through a drag and drop interface. Such an approach makes the common machine learning tasks straightforward and quick.

SAS Enterprise Miner[10] streamlines the data mining process in order to create accurate predictive and descriptive analytical models using massive data. It offers the state-of-the-art predictive analytics and data mining capabilities that enable to analyse complex data and to find insights useful for decision making. IBM SPSS Modeler[11] is an extensive predictive analytics platform designed to bring predictive intelligence to decisions by providing a range of advanced algorithms and techniques that include text analytics, data mining, decision management and optimisation. Oracle Data Mining[12] is a component of the Oracle Advanced Analytics Database Option that provides powerful data mining algorithms and enables to discover insights and make predictions. Amazon Web Services offers a broad set of global computing, storage, database, analytics, application, and deployment services that help to manage large scale applications. Amazon Machine Learning[13] combines powerful machine learning algorithms together with interactive visual tools that guide easy creation, evaluation, deployment of machine learning models, and generation of predictions. It ensures a robust development, scalable and smart resulting applications.

The aforementioned commercial products can be not always suitable for the academic community despite the high functionality of these systems. Thus, free and open source products are gaining more and more popularity among data scientists due to not only their accessibility but also capability to extend, improve or even adapt the existing solution to the needs.

The comparison of web-based data mining solutions has been performed by the chosen criteria and the results are presented in [42]. Facing the complexity of data analysis, researchers need the data mining solutions that meet the following requirements:

- *Implementation of various data mining methods*. It helps to achieve the goal of a data mining problem that needs to be solved. In most cases, it is necessary to apply several different data mining methods to the same problem. Data preprocessing, classification, clustering, and dimensionality reduction are common data mining tasks.
- *Ability to design scientific workflows*. It assists to create a convenient environment for modelling and simulation of data mining experiments in the easy-to-use way.
- *Accessibility as a web application*. It does not require additional installations and any other tools. The system is used and controlled by a web browser at hand only, it is accessible from any place using any device connected to the web 24/7.
- *Accessibility to the latest version of the data mining algorithms*. Researchers work with the up-to-date realizations of the algorithms.
- *Usage of Cloud computing infrastructure*. It allows to solve time- and resource-consuming data mining problems.
- *Online data repository*. It allows to store the uploaded data in online repository and to use the data in different scientific workflows and experiments without a need to upload them each time.

Let us note that the existing open source data mining software does not meet all the requirements. Hence, it is necessary to design and implement a new data mining solution that should absorb all the merits of the existing data mining software and eliminate the shortcomings.

3. New data mining implementation inspired by the Cloud

Given the requirements, discussed in Section 2, we have developed a new web application as a Cloud solution to data mining, called DAMIS (DAta MIning Solution) (http://www.damis.lt). During the process of the architecture design of DAMIS, there is a need to solve several issues that are related to accessibility, portability, scalability, usability of the solution.

8 http://dame.dsf.unina.it
9 https://azure.microsoft.com/en-us/services/machine-learning/
10 http://www.sas.com/en_us/software/analytics/enterprise-miner.html
11 http://www-01.ibm.com/software/analytics/spss/products/modeler/
12 http://www.oracle.com/technetwork/database/options/advanced-analytics/odm/
13 https://aws.amazon.com/machine-learning/

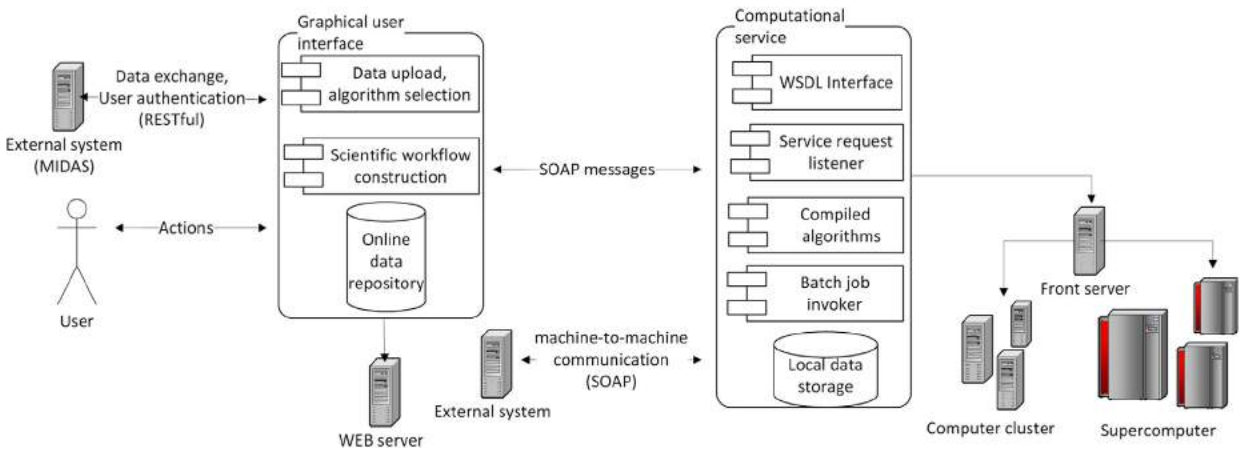


Fig. 1. DAMIS architecture.

The fact that the solution will implement a web service paradigm does not guarantee the desired features stated above. To solve the accessibility issue, the design of DAMIS is organised in such a manner that the DAMIS user would use the developed solution as a black box, i.e. a user has no need to know anything about web services, a Cloud infrastructure, a distributed computing paradigm, etc. To achieve that aim the architecture of the solution is split into several functional layers: a layer of the graphical user interface (GUI) and a layer of the computational services. These layers communicate by sending the SOAP messages according to the interface, described by WSDL (http://hpc.mii.vu.lt:8087/cgi-bin/DamisService.cgi?wsdl). The standardized interface allows to separate two layers depicted above and opens two possibilities: either to use computational service from GUI or directly in machine-to-machine communication scenario. In the case of machine-to-machine communication, the only thing to be implemented on the user's side is to construct a valid SOAP message that is compatible with the publicly available WSDL interface. Such an approach of functional layer separation ensures easy service accessibility. Moreover, login to DAMIS can be accomplished by using RESTful services, because not only GUI and computational service communicate through message passing, but also GUI can communicate in a machine-to-machine manner.

Further, by sending a SOAP message to the computational service, the desired data mining algorithm is invoked. It is important to note that every invocation has to be processed by computational resources such as a Cloud job scheduler. A scheduler is a software capable of organizing resource allocation and queues when receiving computational requests. Further processing of the inbound request might be delayed. It must wait till the required resources become available. Such an organization of job management actually indicates that the computational service layer has to be divided into two more independent parts: a service request listener and a batch job invoker. The service request listener as well as the service user depend on and implement the same interface, described by WSDL. The listener takes and stores locally the input data file, the name of the data mining algorithm to be executed, and all the other necessary parameters, related to the algorithm and computational infrastructure. If the input parameters and input data are valid, then it calls the batch job invoker that actually pushes the request to the job scheduler of specific computational infrastructure and waits till the scheduler processes the job. Afterwards, the service request listener sends back the SOAP message, containing the computed results, to the service user, otherwise, sends back a SOAP message with an error description immediately.

By implicating separation of the service request listener and the batch job invoker, it is possible to increase a portability of DAMIS. This is achieved because the batch job invoker is only a part, dependent on the computational infrastructure solution. The service request listener is a standard solution that implements Common Gateway Interface and can be deployed to any machine with the internet connection and HTTP server running on it. Due to the fact that both the user's side and the service request listener implement the same interface, easy scalability of the DAMIS solution was ensured. The DAMIS architecture is depicted in Fig. 1.

Finally, usability of DAMIS is ensured by GUI. DAMIS implements and presents the data mining solution as a service for the end user, and has friendly GUI that allows data scientists and business intelligence professionals to make data analysis more accessible. This solution allows investigating multidimensional data projection and data similarities as well as to identify the influence of individual features and their relationships using various data mining algorithms. All that is done by taking advantage of the Cloud infrastructure and the separation of computation invocation according to functional responsibilities.

The DAMIS user can benefit from:

- Individual account support;
- Use of the individual online data repository for data storage and easy management;
- Selection of a high-performance computing resource and its status monitoring;
- Use of the latest version of the data mining algorithms;

- Modelling and executing scientific workflows of data mining experiments on the selected Cloud-based infrastructure;
- Management of the accomplished experiments.

To analyze data by the implemented data mining algorithms, the DAMIS user initializes and manages the data mining experiment to model and simulate data analysis processes. Intending to get the analysis results, the experiment should be done. The user can select high-performance computing resource from the proposed alternatives. However, the DAMIS can be extended by different computing resources which location can be undefined. After uploading data files, they become a part of online data storage, and the user has a possibility to manage these files. All the performed experiments including the workflows and data analysis results are saved in the Cloud. Thus the management of the accomplished experiments can be accessible by the DAMIS user on demand.

Data mining experiments are initialized and managed by modelling scientific workflows. The DAMIS user has access to the components for uploading a data file, data pre-processing, computing statistical characteristics, dimensionality reduction, data clustering and classification, as well as for viewing the results. These connected components form a scientific workflow of the data mining experiment. DAMIS GUI with the scientific workflow for the data mining experiment is presented in Fig. 3. The user can modify the designed workflow by adding or removing components and reuse it for other data. The results of data mining experiment can be saved on a user's computer or another external system via the implementation of RESTful services.

Recently, the Cloud technologies and Internet of Things (IoT) gain in more and more popularity, data come from different sources - various devices, sensors, networks, transactional applications, web, and social media [43,44]. Typically, such data are high-dimensional, and the problem of knowledge discovery becomes evident, i.e. we have no means to investigate and fully understand the data in a high-dimensional space. Thus, the goal of dimensionality reduction methods is to extract a lower-dimensional structure from high-dimensional data by transforming (mapping) the data in the high-dimensional space to a space of fewer dimensions. Dimensionality reduction is extremely useful for data scientists and business intelligence professionals while analyzing data during the mining process [45]. Firstly, it is especially valuable in the case of a huge volume of data, since dimensionality reduction decreases such a volume, and then it becomes possible to analyze the data more effectively by other data mining methods. Secondly, often not all the features describing the data are significant because some of them do not characterize the data properly. This fact can influence the data mining results by making noise. Dimensionality reduction assists to eliminate non-informative properties from the data and to prevent from making a wrong decision.

Moreover, when the data dimensionality is reduced up to two, the obtained data can be visualized by presenting them on a 2D scatter plot. Visual representation allows to look inside the data and to see hidden relations that cannot be detected using the other conventional data analysis methods.

The focus of attention of DAMIS is dimensionality reduction. The DAMIS solution implements a series of dimensionality reduction methods:

- *Principal component analysis* (PCA). It is one of the wide-used methods for dimensionality reduction. It transforms the high-dimensional data to a lower-dimensional space so that it preserves the variance of data at best [46].
- *Multidimensional scaling* (MDS). It refers to a group of dimensionality reduction methods that are widely used for dimensionality reduction and visualization of high-dimensional data [3,47,48]. The aim of MDS is to transform the original high-dimensional data to lower-dimensional ones by using the information on the distances between the data items (points) in the original space so that the distances of the corresponding data points in a lower-dimensional space are preserved. The so-called Stress function (projection error) can be minimized using the SMACOF algorithm based on iterative majorization which guarantees a monotonic convergence of this function. The diagonal majorization algorithm (DMA) is a modification of the SMACOF algorithm which attains a slightly worse MDS projection error than SMACOF, but computations are faster and require less computer memory resources [49].
- *Relative* MDS. MDS is a topology preserving mapping but it does not offer a mapping of new data points. Relative MDS can be used for visualizing the new data points on the fixed mapping as well as for visualizing large data sets [50,51].
- SAMANN. The specific back-propagation-like learning rule SAMANN allows a feed-forward artificial neural network to learn one of the MDS group algorithms - Sammon's mapping in an unsupervised way [52,53]. After training the neural network gains a possibility of mapping previously unseen points [54].
- *Combination of the self-organizing map and multidimensional scaling* (SOM-MDS). SOM is another type of neural networks applied to both clustering and visualization of high-dimensional data [3,55,56]. The SOM result is a set of neurons characterized by high-dimensional points that can be mapped on a plane by MDS. The reason for combining SOM-MDS is to improve the visualization of SOM [3,57,58]. Moreover, such a combination allows decreasing the computation time of visualization as compared only with MDS, when the size of the analyzed data set is large enough.

The DAMIS solution also implements the well-known algorithms for data pre-processing, classification and clustering, such as data cleaning, feature selection, normalization, splitting, outliers' filtering, computation of data statistics, multi-layer perceptron (MLP), random decision forest (RDF), k-means, and SOM. To display the results of data mining, a matrix view or 2D scatter plot can be used. A possibility to download the results for local storage in a variety of formats is provided.

DAMIS is the open source data mining solution with multi-OS support which uses resources of a supercomputer and computer cluster (http://www.damis.lt). A possibility to extend a range of alternatives of computational resources is pro-

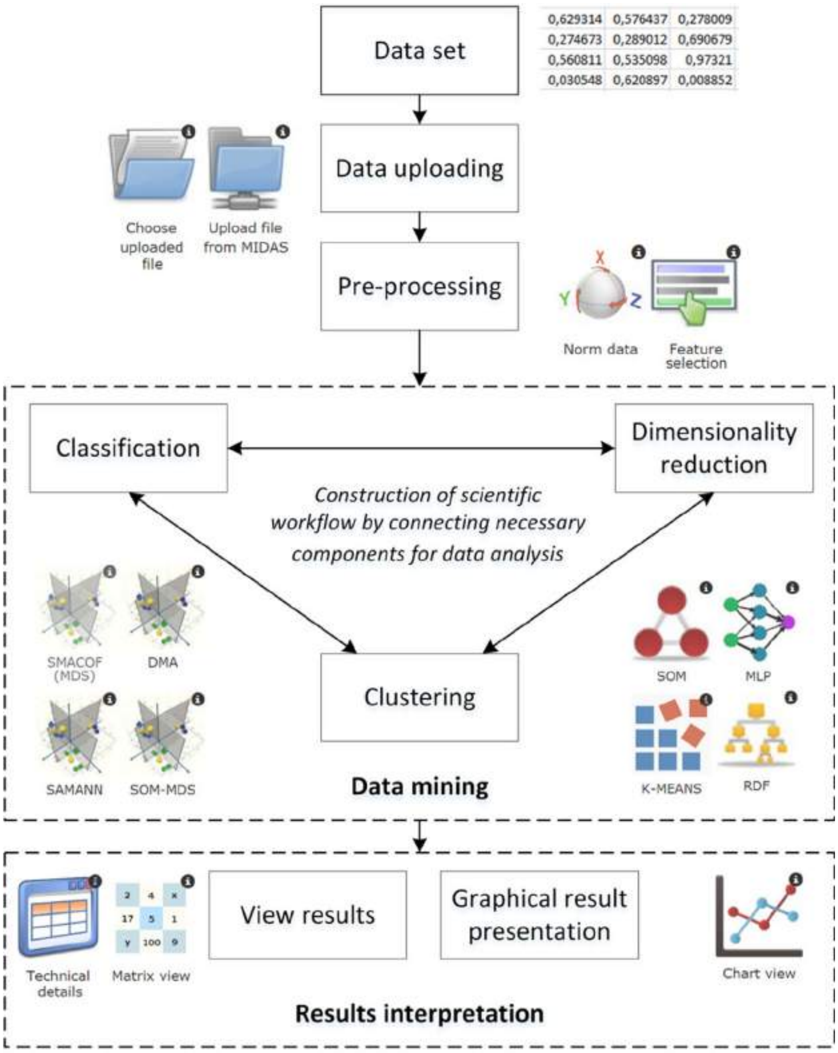


Fig. 2. Schematic presentation of the methodology of data mining for knowledge discovery using DAMIS.

vided. The source code is available on https://github.com/InScience/DAMIS. The solution is also available as the data analysis toolbox in the Lithuanian national open access scientific information archive MIDAS (https://www.midas.lt).

Given the aforementioned facts on DAMIS, a reasonable inference can be drawn. DAMIS gains an advantage over the competitive data mining software and can be an attractive solution for researchers to facilitate the processes of knowledge discovery and decision making.

4. Modelling data mining processes using DAMIS: applications

The capability of DAMIS for modelling data mining processes makes data mining more intelligent. Therefore, DAMIS has a wide range of applications. It is useful for data scientists to get deep insights into the data when solving the various data mining problems. A methodology of data mining for knowledge discovery using DAMIS is presented in Fig. 2. In order to illustrate DAMIS functionalities, a set of applications has been presented, where real-world and simulated data are used for modelling data mining processes.

The first data mining experiment with DAMIS is conducted using the Breast Cancer Wisconsin data set available in UCI Machine Learning Repository [59]. This benchmark data set is commonly used for testing and estimating data mining methods and techniques. 699 observations with nine features of the breast cancer are collected, where each instance is assigned to one of the two possible classes: benign or malignant. The empirical data mining experiment should be carried out to get meaningful information on these data. The experiment aims to demonstrate the capabilities of DAMIS to classify and reduce the dimensionality using the real-world data. The data mining process model is as follows:

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

Fig. 3. The example of data mining process model in DAMIS.

Table 1
Classification results of the Breast Cancer data using various data mining tools.

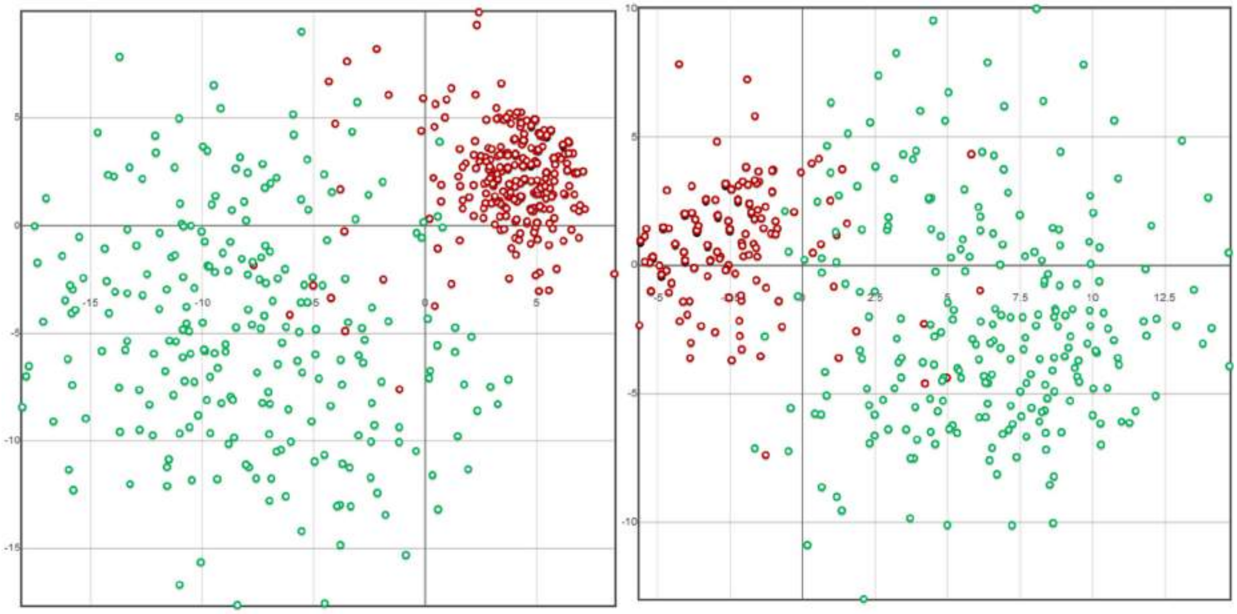
Data mining tool	Classifier	General classification accuracy (%)	Sensitivity (%)	Specificity (%)
Orange	RDF	0.9565	0.9778	0.9167
	MLP	0.9710	0.9778	0.9583
Weka	RDF	0.9855	0.9778	1.0000
	MLP	0.9565	0.9556	0.9583
MS Azure ML	RDF	0.9559	0.9512	0.9630
	MLP	0.9701	0.9756	0.9630
DAMIS	RDF	0.9710	0.9778	0.9583
	MLP	0.9710	0.9556	1.0000

- *Data file uploading*. To analyze the data with DAMIS, the data must be prepared in a compatible format, i.e., tab, txt, csv, xlsx, arff.
- *Data pre-processing*. It includes data cleaning in the case of missing data, data normalization to bring all feature values into the same interval as well as feature selection.
- *Data classification*. Random decision forest (RDF) and multilayer perceptron (MLP) are selected to classify these data.
- *Dimensionality reduction*. To present the complex data in a meaningful manner and easily understandable form, the dimensionality reduction-based visualization method SMACOF (MDS) is used.
- *Viewing the results*. To display the results, a matrix view or 2D scatter plot are selected.

Such a data mining model can be easily implemented in DAMIS by constructing a scientific workflow from the available components through a simple drag and drop interface (Fig. 3). The designed workflow is executed, and the obtained results can be viewed.

The obtained results of data classification by RDF and MLP in DAMIS are presented in Table 1. Other wide-used data mining tools provide similar classification results. This fact shows that DAMIS is a competitive solution for data classification as compared to other tools.

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



a) b)

Fig. 4. Visualization of the Breast Cancer data by SMACOF (MDS) in DAMIS: a) all the features, b) a part of the features are selected. (For interpretation of the references to colour in the text, the reader is referred to the web version of this article.)

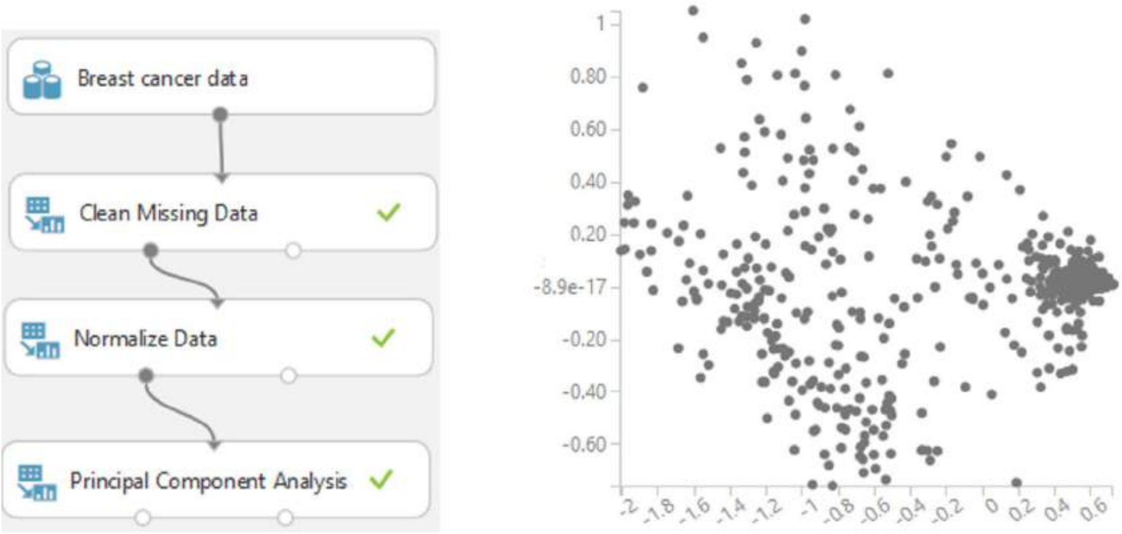
Fig. 4 presents 2D scatter plots of dimensionality reduction results, where the data dimensionality is reduced up to two by SMACOF (MDS). A large number of the points, corresponding to the benign tumor data (red points), are concentrated in one area, and the other points, corresponding to the malignant tumor data (green), are spread widely. This way of data visualization allows viewing the data as a whole in a meaningful manner and easily understandable form. Each of breast cancer data observations is described by nine features. A question arises whether all the features are essential and how the results of visualization change when only a part of features are selected (Fig. 4b). Comparing the obtained results in Fig. 4, we see that the feature selection does not significantly affect the visualization results, however, the obtained classification accuracy is worse. Thus, it can be concluded that all the features are essential.

The visualization results are compared with that obtained using powerful Microsoft Azure ML, where scientific workflows can also be designed for data mining experiments and executed in the Cloud. Unfortunately, only PCA and LDA (linear discriminant analysis) are implemented here as dimensionality reduction methods. A scientific workflow for the breast cancer data analysis in Azure ML is designed (Fig. 5a), and the dimensionality reduction method PCA is applied (Fig. 5b). The visualization results in Figs. 4 and 5 b point to the conclusions that DAMIS has an advantage in the case of data visualization, compared to Azure ML due to the possibility to represent data points in a more meaningful manner by using different colors and shapes. It allows separating data classes visually. Moreover, a wider range of dimensionality reduction methods is implemented in DAMIS.

As mentioned before, DAMIS provides a possibility to select high-performance computing resources from the proposed alternatives. It assists in the massive data analysis and can be helpful to execute time- and resource-consuming tasks. It is useful to solve data mining optimization problems when a multi-start strategy is required. To illustrate this possibility, another scientific workflow is designed, and the experiment is conducted by mining simulated data, i.e. ellipsoidal data set that contains ten overlapping ellipsoidal-type clusters, obtained by a generator[14]. 3140 50-dimensional data points are processed by SMACOF (MDS) and relative MDS, the results of which depend on the initial randomly generated values of lower-dimensional points. Thus, the multi-start strategy is effective in that cases. Using the high-performance computing resource in DAMIS, the same problem with different initial values can be solved on each computing node and a set of solutions is obtained. The solution with the smallest projection error is considered as a result of the dimensionality reduction method (Fig. 6). We can observe the data clusters clearly enough. Relative MDS manages to visualize the data clusters more precisely as compared to the SMACOF (MDS) algorithm. Moreover, the computation of relative MDS takes almost twice less time.

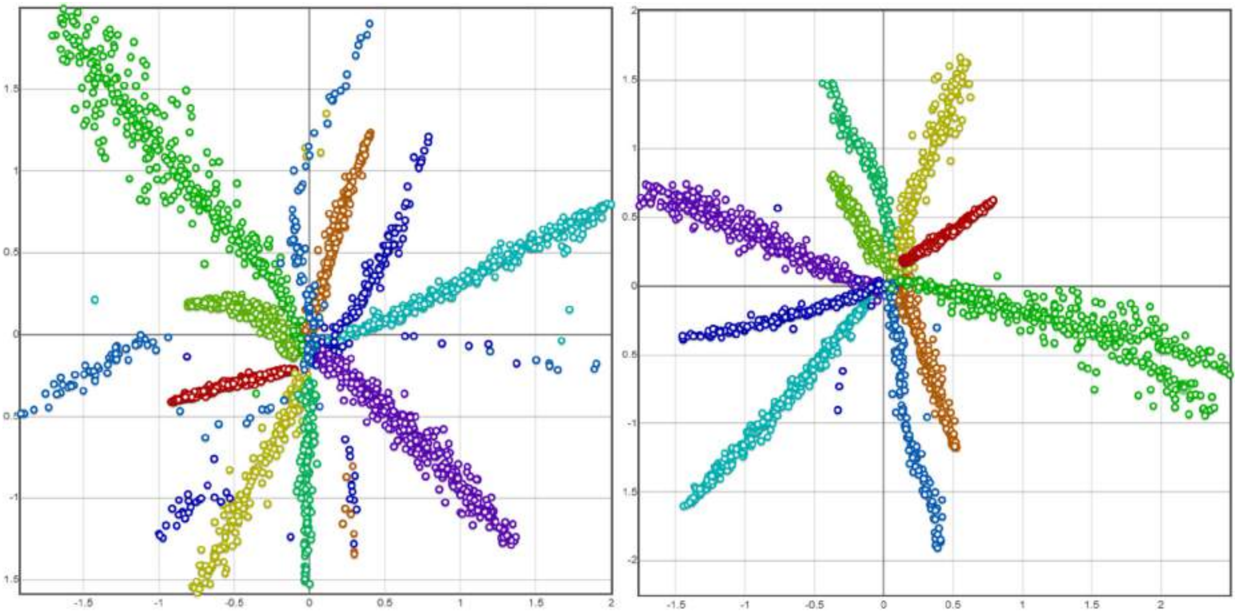
The aim of the third experiment is to show a possibility of DAMIS to present real world statistical data in a more comprehensive form by visualizing the data. To this end, we have chosen and analyzed some data from the EuroStat

[14] http://personalpages.manchester.ac.uk/mbs/julia.handl/generators.html



a) b)

Fig. 5. Data mining in Microsoft Azure ML: a) scientific workflow, b) visualization of the Breast Cancer Data by PCA.



a) b)

Fig. 6. Visualization of the ellipsoidal data in DAMIS: a) by SMACOF (MDS), b) by relative MDS.

portal[15]. The selected data set consists of information on the research and development (R&D) expenditure, expressed by the percentage of the gross domestic product (GDP) in the EU countries, USA, Russia, China, and Japan for the period from 2004 to 2014. Thus, in this experiment, the data set of 33 11-dimensional points is analyzed. The EU countries are grouped according to the year when they joined EU. The first group consists of six countries-founders (EU_1958), the second group – three countries joined till 2004 (EU_till_2004), the third group – ten countries joined in 2004 (EU_2004), the fourth group – nine countries joined from 2004 (EU_from_2004). Additionally, a data item (EU28) is introduced which consists of the

[15] http://ec.europa.eu/eurostat/

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

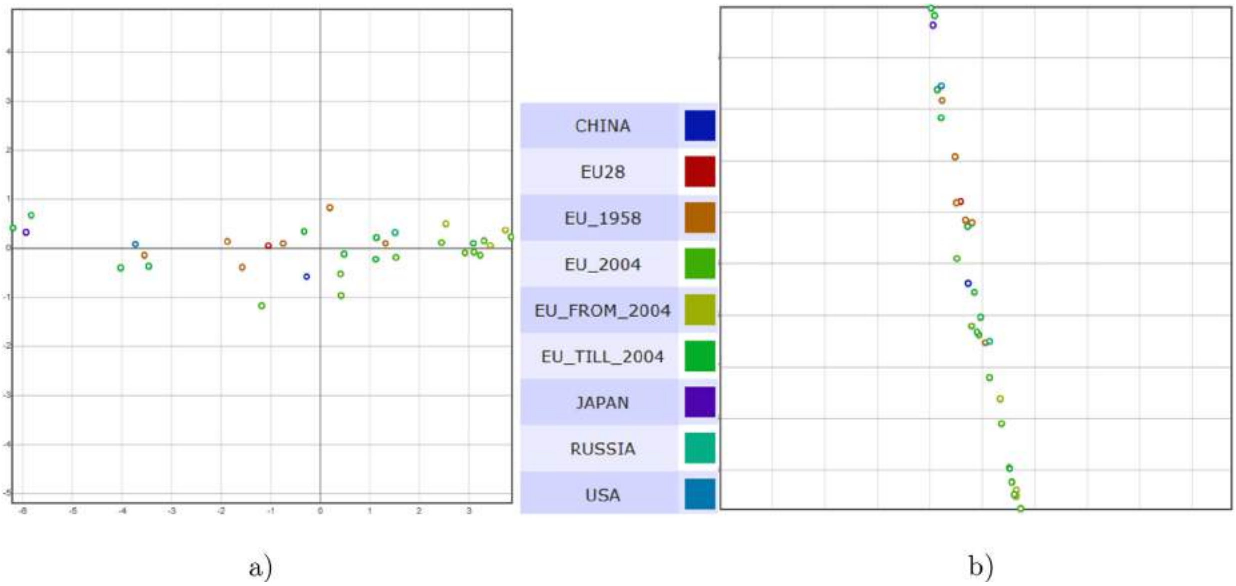


Fig. 7. Visualization of the country data in DAMIS: a) by PCA, b) by SAMANN.

averaged values of all the EU countries. It is interesting to get information on similarities and dissimilarities of the countries according to R&D expenditure in the easily understandable visual form. It is also interesting to see whether the countries form clusters as well as to compare the data of the USA, Russia, China and Japan with that of the EU countries. To this end, the dimensionality reduction-based methods PCA and SAMANN have been applied, and the results are presented in Fig. 7. Single point corresponds to a group of the EU countries or one of the non-EU countries. We can see that the point corresponding to Japan and the points, corresponding to two countries from the group EU_till_2004 are far away from the remaining points and form a separate cluster. The point corresponding to the USA and three points corresponding to the EU countries also form a cluster. However, the point corresponding to non-EU country Russia is among the points corresponding to the majority of the EU countries. The visualization results obtained by PCA and SAMANN are rather similar, which shows that the distribution of points corresponding to the countries is independent of the used dimensionality reduction method, in this case.

5. Conclusions

To discover useful knowledge from real-world and simulated big data, business intelligence professionals and data scientists face with new challenges. The conventional technologies and methods cannot store and analyze a large amount of data. Since the Cloud technologies gain in popularity, the attention is focused on the development of new Cloud-based data mining solutions with a possibility to access high-performance computing environments that can utilize remote computing resources on demand. Considering the review of the existing data mining software and solutions, an inference has been drawn. Unfortunately, the well-known open source data mining software does not meet all the requirements that the process of knowledge discovery would be more effective. A new open source web-based solution DAMIS implements these requirements.

DAMIS allows using various data mining methods jointly. The massive data mining becomes simpler, effective, and easily understandable by constructing scientific workflows for data mining process through a drag and drop interface. Here the scientific workflows allow composing the convenient model of data mining process covering a number of different methods.

The DAMIS architecture is designed to ensure easy accessibility, usability, scalability, and portability of this solution. DAMIS provides a possibility to cope with data classification, clustering, and dimensionality reduction tasks. Time- and resource-consuming data mining problems can be solved by selecting the high-performance computing resources from the proposed alternatives and utilizing them on demand. The experiments by modelling data mining processes with real-world and simulated data have proved the efficiency of the proposed solution. DAMIS represents the synergy of Cloud computing and data mining in solving data mining problems of different nature with a view to get deeper insights into the data.

References

[1] Data Mining Techniques in Grid Computing Environments, in: W. Dubitzky (Ed.), John Wiley and Sons, Ltd, 2009, doi:10.1002/9780470699904.ch1.
[2] D. Talia, P. Trunfio, Service-oriented Distributed Knowledge Discovery, Chapman and Hall/CRC, 2012, doi:10.1201/b12990-4.
[3] G. Dzemyda, O. Kurasova, J. Žilinskas, Multidimensional Data Visualization: Methods and Applications, Springer Optimization and its Applications, 75, Springer, 2013, doi:10.1007/978-1-4419-0236-8.

[4] M.D. Assuno, R.N. Calheiros, S. Bianchi, M.A. Netto, R. Buyya, Big data computing and clouds: trends and future directions, J. Parallel Distrib. Comput. 79–80 (2015) 3–15, doi:10.1016/j.jpdc.2014.08.003. Special Issue on Scalable Systems for Big Data Management and Analytics.
[5] I. Kholod, M. Kuprianov, I. Petukhov, Parallel and distributed data mining in cloud, in: Advances in Data Mining. Applications and Theoretical Aspects: Proceedings of 16th Industrial Conference, ICDM 2016, Springer, 2016, pp. 349–362, doi:10.1007/978-3-319-41561-1_26.
[6] A.D. Barrachina, A. O'Driscoll, A big data methodology for categorising technical support requests using Hadoop and Mahout, J. Big Data 1 (2014), doi:10.1186/2196-1115-1-1.
[7] C.P. Chen, C.-Y. Zhang, Data-intensive applications, challenges, techniques and technologies: a survey on big data, Inf. Sci. 275 (2014) 314–347, doi:10.1016/j.ins.2014.01.015.
[8] D.T. Larose, C.D. Larose, Data Mining and Predictive Analytics, John Wiley & Sons, 2015.
[9] D. Talia, Toward cloud-based big-data analytics, IEEE Comput. Sci. (2013) 98–101.
[10] A. Fernández, S. del Río, F. Herrera, J.M. Benítez, An overview on the structure and applications for business intelligence and data mining in cloud computing, in: 7th International Conference on Knowledge Management in Organizations: Service and Cloud Computing, Springer, 2013, pp. 559–570.
[11] N.S. Bhosale, S.S. Pande, A survey on recommendation system for big data applications, Data Mining Knowl. Eng. 7 (1) (2015) 42–44.
[12] J.P. Verma, B. Patel, A. Patel, Big data analysis: recommendation system with hadoop framework, in: IEEE International Conference on Computational Intelligence & Communication Technology (CICT 2015), IEEE, 2015, pp. 92–97.
[13] M. Parekh, B. Saleena, Designing a cloud based framework for healthcare system and applying clustering techniques for region wise diagnosis, Procedia Comput. Sci. 50 (2015) 537–542.
[14] A. Castiglione, R. Pizzolante, A. De Santis, B. Carpentieri, A. Castiglione, F. Palmieri, Cloud-based adaptive compression and secure management services for 3d healthcare data, Future Gener. Comput. Syst. 43 (2015) 120–134.
[15] G. Dzemyda, V. Marcinkevičius, V. Medvedev, Large-scale multidimensional data visualization: a web service for data mining, in: Proceedings of the 4th European Conference on Towards a Service-Based Internet, in: Lecture Notes in Computer Science, 6994, Springer Berlin Heidelberg, 2011, pp. 14–25, doi:10.1007/978-3-642-24755-2_2.
[16] G. Dzemyda, V. Marcinkevičius, V. Medvedev, Web application for large-scale multidimensional data visualization, Math. Model. Anal. 16 (2) (2011) 273–285, doi:10.3846/13926292.2011.580381.
[17] V. Kravtsov, T. Niessen, V. Stankovski, A. Schuster, Service-based resource brokering for grid-based data mining, in: Proceedings of Int Conference on Grid Computing and Applications, 2006, pp. 163–169.
[18] D. Talia, P. Trunfio, How distributed data mining tasks can thrive as knowledge services, Commun. ACM 53 (2010) 132–137, doi:10.1145/1785414.1785451.
[19] H. Demirkan, D. Delen, Leveraging the capabilities of service-oriented decision support systems: putting analytics and big data in cloud, Decis. Support Syst. 55 (1) (2013) 412–421, doi:10.1016/j.dss.2012.05.048.
[20] S. Kothari, T. Peck, J. Zeng, F. Oblea, A.E. Votaw, G. Dispoto, Simulation as a cloud service for short-run high throughput industrial print production using a service broker architecture, Simul. Modell. Pract. Theory 58 (Part 2) (2015) 115–139, doi:10.1016/j.simpat.2015.05.003. Special issue on Cloud Simulation.
[21] J. Byrne, C. Heavey, P. Byrne, A review of web-based simulation and supporting tools, Simul. Modell. Pract. Theory 18 (3) (2010) 253–276, doi:10.1016/j.simpat.2009.09.013.
[22] I.A.T. Hashem, I. Yaqoob, N.B. Anuar, S. Mokhtar, A. Gani, S.U. Khan, The rise of “big data” on cloud computing: review and open research issues, Inf. Syst. 47 (2015) 98–115, doi:10.1016/j.is.2014.07.006.
[23] H. Kreger, Web services conceptual architecture (WSCA 1.0), Architecture 5 (2001) 6–7.
[24] D. Birant, Service-oriented data mining, in: K. Funatsu (Ed.), New Fundamental Technologies in Data Mining, 2011, pp. 3–18, doi:10.5772/14066.
[25] T. White, Hadoop: The Definitive Guide, 54, O'Reilly Media, 2012.
[26] S. Bowers, Scientific workflow, provenance, and data modeling challenges and approaches, J. Data Semant. 1 (1) (2012) 19–30, doi:10.1007/s13740-012-0004-y.
[27] J. Liu, E. Pacitti, P. Valduriez, M. Mattoso, A survey of data-intensive scientific workflow management, J. Grid Comput. 13 (4) (2015) 457–493.
[28] N. Cerezo, J. Montagnat, M. Blay-Fornarino, Computer-assisted scientific workflow design, J. Grid Comput. 11 (3) (2013) 585–612. 0.1007/s10723-013-9264-5.
[29] J. Liu, E. Pacitti, P. Valduriez, M. Mattoso, A survey of data-intensive scientific workflow management, J. Grid Comput. 13 (2015) 457–493, doi:10.1007/s10723-015-9329-8.
[30] Y. Zhao, Y. Li, S. Lu, I. Raicu, C. Lin, Devising a cloud scientific workflow platform for big data, in: 2014 IEEE World Congress on Services, IEEE Computer Society, 2014, pp. 393–401, doi:10.1109/SERVICES.2014.75.
[31] A. Barker, J.I. Van Hemert, Scientific workflow: a survey and research directions, Parallel Process. Appl. Math. 4967 (2007) 746–753, doi:10.1007/978-3-540-68111-3_78.
[32] K. Görlach, M. Sonntag, D. Karastoyanova, F. Leymann, M. Reiter, Conventional workflow technology for scientific simulation, in: Guide to e-Science, Springer, 2011, pp. 323–352.
[33] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I.H. Witten, The WEKA data mining software: an update, SIGKDD Explor. Newslett. 11 (1) (2009) 10–18, doi:10.1145/1656274.1656278.
[34] D. Talia, P. Trunfio, O. Verta, The Weka4WS framework for distributed data mining in service-oriented grids, Concurr. Comput. Pract. E. 20 (2008) 1933–1951, doi:10.1002/cpe.1311.
[35] V. Podpečan, M. Zemenova, N. Lavrač, Orange4WS environment for service-oriented data mining, Comput. J. 55 (2012) 82–98, doi:10.1093/comjnl/bxr077.
[36] J. Demšar, T. Curk, A. Erjavec, Črt Gorup, T. Hočevar, M. Milutinovič, M. Možina, M. Polajnar, M. Toplak, A. Starič, M. Štajdohar, L. Umek, L. Žagar, J. Žbontar, M. Žitnik, B. Zupan, Orange: data mining toolbox in python, J. Mach. Learn. Res. 14 (2013) 2349–2353.
[37] M.R. Berthold, N. Cebron, F. Dill, T.R. Gabriel, T. Kötter, T. Meinl, P. Ohl, C. Sieb, K. Thiel, B. Wiswedel, KNIME: the Konstanz information miner, Studies in Classification, Data Analysis, and Knowledge Organization (GfKL 2007), Springer, 2007, doi:10.1145/1656274.1656280.
[38] M. Hofmann, R. Klinkenberg, RapidMiner: Data Mining Use Cases and Business Analytics Applications, Chapman & Hall/CRC, 2013.
[39] J. Kranjc, V. Podpečan, N. Lavrač, Clowdflows: a cloud based scientific workflow platform, in: Machine Learning and Knowledge Discovery in Databases, in: Lecture Notes in Computer Science, 7524, Springer Berlin Heidelberg, 2012, pp. 816–819.
[40] J. Kranjc, J. Smailovič, V. Podpečan, M. Grčar, M. Žnidaršič, N. Lavrač, Active learning for sentiment analysis on data streams: methodology and workflow implementation in the clowdflows platform, Inf. Process. Manag. 51 (2) (2014) 187–203.
[41] M. Brescia, G. Longo, M. Castellani, S. Cavuoti, R. D'Abrusco, O. Laurino, DAME: A distributed web based framework for knowledge discovery in databases, Metnorie della Societa Astronomica Italiana Supplement 19 (2012) 324–329.
[42] V. Medvedev, O. Kurasova, Cloud technologies: a new level for big data mining, in: F. Pop, J. Kolodziej, B.D. Martino (Eds.), Resource Management for Big Data Platforms: Algorithms, Modelling, and High-performance Computing Techniques, Computer Communications and Networks, Springer, 2016, pp. 55–67.
[43] C.C. Aggarwal, N. Ashish, A. Sheth, The internet of things: a survey from the data-centric perspective, in: Managing and Mining Sensor Data, Springer US, 2013, pp. 383–428, doi:10.1007/978-1-4614-6309-2_12.
[44] R. Stackowiak, A. Licht, V. Mantha, L. Nagode, Big Data and the Internet of Things, Springer Optimization and its Applications, Apress, 2015, doi:10.1007/978-1-4842-0986-8.
[45] R. Karbauskaitė, G. Dzemyda, Fractal-based methods as a technique for estimating the intrinsic dimensionality of high-dimensional data: a survey, Informatica 27 (2) (2016) 257–281, doi:10.15388/Informatica.2016.84.

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

[46] I. Jolliffe, Principal Component Analysis, Springer, Berlin, 1986.
[47] I. Borg, P. Groenen, Modern Multidimensional Scaling: Theory and Applications, Springer, 2005, doi:10.1007/0-387-28981-X.
[48] G. Dzemyda, O. Kurasova, V. Medvedev, Dimension reduction and data visualization using neural networks, in: Emerging Artificial Intelligence Applications in Computer Engineering : Real Word AI Systems with Applications in eHealth, HCI, in: Frontiers in artificial intelligence and applications, 19, 2007, pp. 25–49.
[49] J. Bernatavičienė, G. Dzemyda, V. Marcinkevičius, Diagonal majorization algorithm: properties and efficiency, Inf. Technol. Control 36 (2007) 353–358.
[50] J. Bernatavičienė, G. Dzemyda, V. Marcinkevičius, Conditions for optimal efficiency of relative MDS, Informatica 18 (2) (2007) 187–202.
[51] A. Naud, Visualization of high-dimensional data using an association of multidimensional scaling to clustering, in: 2004 IEEE Conference on Cybernetics and Intelligent Systems, 1, 2004, pp. 252–255, doi:10.1109/ICCIS.2004.1460421.
[52] J. Mao, A.K. Jain, Artificial neural networks for feature extraction and multivariate data projection., IEEE Trans. Neural Networks 6 (2) (1995) 296–317, doi:10.1109/72.363467.
[53] V. Medvedev, G. Dzemyda, O. Kurasova, V. Marcinkevičius, Efficient data projection for visual analysis of large data sets using neural networks, Informatica 22 (4) (2011) 507–520.
[54] S. Ivanikovas, G. Dzemyda, V. Medvedev, Large datasets visualization with neural network using clustered training data, in: Proceedings of the 12th East European Conference on Advances in Databases and Information Systems, in: ADBIS '08, Springer-Verlag, 2008, pp. 143–152, doi:10.1007/978-3-540-85713-6_11.
[55] T. Kohonen, Self-Organizing Maps, Springer Series in Information Science, 3rd, Springer, Berlin, 2001, doi:10.1007/978-3-642-56927-2.
[56] P. Stefanovič, O. Kurasova, Visual analysis of self-organizing maps, Int. J. Nonlinear Anal. 16 (4) (2011) 488–504.
[57] O. Kurasova, A. Molytė, Integration of the self-organizing map and neural gas with multidimensional scaling, Inf. Technol. Control 40 (1) (2011) 12–20.
[58] O. Kurasova, A. Molytė, Quality of quantization and visualization of vectors obtained by neural gas and self-organizing map, Informatica 22 (1) (2011) 115–134.
[59] K. Bache, M. Lichman, UCI Machine Learning Repository, 2013. http://archive.ics.uci.edu/ml.