Hindawi Publishing Corporation
The Scientific World Journal
Volume 2014, Article ID 859279, 10 pages
http://dx.doi.org/10.1155/2014/859279

Research Article

Cloud Based Metalearning System for Predictive Modeling of Biomedical Data

Milan Vukićević, Sandro Radovanović, Miloš Milovanović, and Miroslav Minović

Faculty of Organizational Sciences, University of Belgrade, Jove Ilića 154, 11000 Belgrade, Serbia

Correspondence should be addressed to Miroslav Minović; miroslav.minovic@fon.bg.ac.rs

Received 20 December 2013; Accepted 21 January 2014; Published 14 April 2014

Academic Editors: R. Colomo-Palacios and V. Stantchev

Rapid growth and storage of biomedical data enabled many opportunities for predictive modeling and improvement of healthcare processes. On the other side analysis of such large amounts of data is a difficult and computationally intensive task for most existing data mining algorithms. This problem is addressed by proposing a cloud based system that integrates metalearning framework for ranking and selection of best predictive algorithms for data at hand and open source big data technologies for analysis of biomedical data.

1. Introduction

Data mining can be defined as the process of finding previously unknown patterns and trends in databases and using that information to build predictive models [1]. Due to increasing amount of data generated in healthcare systems (medical records, gene expression data, medical image data, etc.), analysis became too complex and voluminous for traditional methods and this is why data mining is becoming increasingly important [2].

In the last decade data mining techniques (like clustering, classification, or association) were successfully applied on different medical and biomedical problems like prediction of heart attacks [3], diagnostics based on gene expression microarray data [4], classification of Parkinson's disease [5], identification of liver cancer signature [6], and so forth.

Special area of medical data mining is biomedical data mining that seeks to connect phenotypic data to biomarker profiles and therapeutic treatments, with the goal of creating predictive models of disease detection, progression, and therapeutic response. This area includes mining genomic data (and data from other high-throughput technologies such as DNA sequencing and RNA expression), text mining of the biological literature, medical records, and so forth, and image mining across a number of modalities, including X-rays, functional MRI, and new types of scanning microscopes [7].

Even though many algorithms were specially designed for application in this area [8], the exponential increase of genomic data brought by the advent of the third generation sequencing (NGS) technologies and the dramatic drop in sequencing cost have posed many challenges in terms of data transfer, storage, computation, and analysis of big biomedical data [2, 7, 9, 10]. These authors emphasize the lack of computing power and storage space, as a major hurdle in achieving research goals. They propose cloud computing as a service model sharing a pool of configurable resources, which is a suitable workbench to address these challenges (Figure 1).

One of the "soft" approaches for reducing the need for computer power for data analysis is introduction of metalearning systems for selection and ranking of the best suited algorithms for different problems (datasets). These systems store historical experimental records (descriptions of datasets and algorithm performances) and, based on these records, evolve models for prediction of algorithm performances on a new dataset. By using these systems, analyst does not have to evaluate large number of algorithms on a big data (only ones with the best predicted performance) and this way saves computational and time resources. Even though specialized metalearning systems are developed for many application areas like electricity load forecasting [11], gold market forecasting [12], choosing metaheuristic optimization algorithm for traveling salesman problem [13], and so forth,

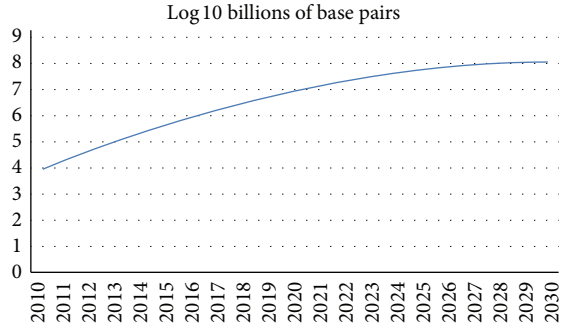

Figure 1: Projected growth of DNA sequence data in the 21st century [7].

there are not many researches that utilize metalearning in medicine [14] and in this paper we will propose such approach. Similar efforts have been made in the field of continuous improvement of business performance with big data [15], which lets users analyze business performance in distributed environments with a short response time, which can be analogous with biomedical systems.

Researchers in this area suggest that the main problem of exponential data growth is to provide adequate computing infrastructure that has the possibility to assemble, manage, and mine the enormous and rapidly growing data [2, 7, 9, 10]. They emphasize that intersection point between genome technology, cloud computing, and biological data mining provides a launch pad for developing a globally applicable cloud computing platform capable of supporting a new paradigm of data intensive, cloud-enabled predictive medicine.

In this paper we propose an extension of cloud based systems [16, 17] with data and model driven services based on metalearning approach. Additionally, this system includes open source data mining environments as a platform service for users. System is based on open source technologies and this is very important since they enable collaborative collection of data and fast development of new algorithms [18].

2. State of the Art

In this section a brief overview of cloud based healthcare systems and metalearning systems which are correlated with proposed system is discussed. Cloud systems emerged as a technology breakthrough in the last decade and impacted a wide range of business like SME [19], education [20], e-government [21], data mining [22], and so forth.

Ahuja et al. [23] exhaustively reviewed usage and consideration points in implementing cloud healthcare system. They identified that the most important points are infrastructure and number of facilities. Infrastructure has great influence since most of the healthcare facilities and office locations were built years ago and cannot use cloud systems. Number of facilities is important on operation of health organization and whether their IT infrastructure is distributed between facilities or is in a single datacenter. Moving to the cloud would help communication, application, and collaboration between health organizations. Cloud computing reduces operating costs, because the need for IT staff in each facility is lower and overall IT budget is reduced.

The advantages of cloud computing and big data technologies, like Hadoop and related software, increased their popularity in medicine and bioinformatics. Dai et al. [16] identified four bioinformatics cloud services. Those are DaaS (data as a service), SaaS (software as a service), PaaS (platform as a service), and IaaS (infrastructure as a service). Bioinformatics generates huge amount of raw data and they should be available for data analysis through DaaS. Additionally, a large diversity of software tools is necessary for data analysis and SaaS is provided as an option in regarding this problem. Platform as a service provides programmable platform for development, testing, and deploying solutions online. IaaS offers a complete computer infrastructure for bioinformatics analysis.

Lack of support for complex and large scale healthcare application of electronic medical records was identified by Li et al. [24]. Therefore, XML was used as a model for managing medical data while Hadoop infrastructure and MapReduce framework were used for data analysis. Their system, called XBase, is doing various data mining tasks like classification of heart valvular disease, detecting association rules, diagnosis assistance, and treatment recommendation.

As Schatz et al. [25] stated, sequencing of DNA chain is improving at a rate of about 5-fold per year, while computer performance is doubling only every 18 or 24 months. Therefore, addressing the issue of designing data analysis arises as a question. A practical solution for solving this problem is to concentrate on developing methods that make better use of multiple computers and processors, where cloud computing emerges with promising results. They stated that Hadoop/MapReduce technology is particularly well suited, from genomic point of view, for analysis of DNA sequence. The Crossbow genotyping program leverages Hadoop/MapReduce to launch many copies of the short read in parallel leveraging of Hadoop/MapReduce and Crossbow for greater results. In their benchmark test on the Amazon cloud, Crossbow Hadoop/MapReduce analyzed 2.7 billion data points in about 4 hours, which included the time required for uploading the raw data, for a total cost of $85 USD. Beside this, they described obstacles which can pose significant barrier in analysis of DNA sequence.

An interesting approach to design of biomedical cloud system was described by Taverna [26]. It is a workflow management system that allows uploading data to the cloud from web application, creating data flow and run analysis from multiple computing units. While analysis is running, user can monitor the progress. This application also allows sharing of data flow, enabling many researchers from the same or other projects to influence and give contribution to research process.

Frameworks for cloud based genome data analysis, such as Galaxy [27], offer generalized tools and libraries as components in workflow editor. Galaxy enables users to define pipelines that through specifically developed visualization show progression of the workflow. It is extensible and, therefore, a community built around it contributes in developing

various tools for genome analysis. It is important to notice that Galaxy is integrated with biomedical databases, such as UCSC table data, BioMart Central, and modENCODE server.

Hadoop and MapReduce distributed computing paradigm has been implemented in CloudBurst [28]. It is used for mapping short reads to reference genomes in a parallel fashion in cloud environment. Essentially, it provides a parallel read-mapping algorithm optimized for mapping sequence data to the human genome and other reference genomes, intended for use in a biological analysis including SNP discovery, genotyping, and personal genomics.

Another large biological extensible workbench is SeqWire [29]. Users are allowed to write and share pipeline modules. It provides massive parallel processing using sequencing technologies, such as ABI SOLID and Illumina, web application, pipeline for processing and annotating sequenced data, query engine, and a MetaDB.

Web based cloud system, such as FX [30], provides high usability for users that are not familiar with programming techniques. It is developed for various biomedical data analyses such as estimating gene expression level and genomic variant calling from the RNA sequence using transcriptome-based references. User uploads data and configures data analysis settings on Amazon Web Service (AWS). Since this application is domain specific, it does not require manual arrangement of pipelines.

Critical cloud services in biomedicine are infrastructure services. Therefore, CloVR [31] offers a virtual operating system with preinstalled packages and libraries required for biomedical data analysis, such as large-scale BLAST searches, whole-genome assembly, gene finding, and RNA sequence analysis. It is implemented as an online application but does not provide GUI. Instead, command-line based automated analysis pipelines with preconfigured software packages for composing workflows are implemented.

Chae et al. [9] focus on two emerging problems in bioinformatics data analysis. Those are computation power and big data analysis for the biomedical data. Biomedical analysis requires very big computing power with huge storage space. They proposed BioVLab as an affordable infrastructure on the cloud, with a graphical workflow creator which provides an efficient way to deal with these problems. BioVLab consists of three layers. The first layer is a graphical workflow engine, called XBaya, which enables the composition and management of scientific workflows on a desktop. The second layer, gateway, is a web-based analysis tool for the integrated analysis of microRNA and mRNA expression data. Analysis is done on Amazon S3 Interface, which presents third layer of architecture. Data and commands from gateway are transferred to cloud, which analyze data and return results to user on desktop. They emphasized that analysis of big medical data requires use of appropriate tools and databases from a vast number of tools and databases; therefore using cloud would not solve problems of computational power and big data analysis.

Cloud healthcare application could have great impact on society, but security, privacy, and government regulation issues limit its usage. Zhang and Liu [32] defined security model for cloud healthcare system. First part of the model is secure collection of data. Data collection module is created and maintained independently by care delivery organizations. Second part is secure storage. Storage system must be encrypted and it can allow only authorized access. Third part is secure usage, which consists of medicine staff signature and verification sections. Their model deals with problem of information ownership, authenticity and authentication, nonrepudiation, patient consent and authorization, integrity and confidentiality of data, and availability of system.

Since security is identified as a major challenge in cloud systems, Wooten et al. [33] designed and implemented secure healthcare cloud system. This was achieved with a trust-aware role-based access control and a tag system. System was implemented on top of Amazon Web Services (AWS) Elastic Compute Cloud (EC2) and Linux, Apache, MySQL, and PHP (LAMP) solution stack.

Liu and Park [34] focused on challenges and adaptation of e-healthcare cloud systems. This system extends the cloud paradigm in order to satisfy global demands in digital healthcare applications. Therefore, technology, healthcare process, and service are identified as the main characteristics of healthcare cloud systems. Similarly, new challenges arose by the unique requirements of the e-healthcare industry for using cloud services for regulation, security issues, access, intercloud connectivity, and resource distribution.

IBM Watson is also used in healthcare as cloud service. Giles and Wilcox [35] used Watson ability to use natural language processing and combine it with content analysis in order to help medical staff in diagnostic analysis. This application of Watson identifies diseases, symptoms, right medications, and modifiers directly from medical records from different medical facilities stored on cloud.

Knowledge cloud based systems in medicine, as Lai et al. [17] stated, are one of the major government's strategic plans to drive the healthcare services which are identified as public concerns in China. They highlighted some successful criteria for establishment of a private knowledge network for business network collaboration and the knowledge cloud system for radiotherapy dynamic treatment service in China, such as innovation outsourcing, marketing opportunity, economy of scales, leverage existing resources, and service on demand. Three parties are identified in the KaaS service model. The first party is the *knowledge user* (patients, hospitals, and doctors), the one who pays for the knowledge service on demand. The second party is the *knowledge expert* (external consultants), the one who provides the knowledge service on demand. The third party is the *knowledge agent* who links together the knowledge user and the knowledge expert on demand.

Great potential of cloud services in the area of biomedicine is identified by Grossman and White [7], who made a vision of biomedical cloud in the future. Since amount of data which hospitals and medical institutions are dealing with is growing rapidly, big data technologies will have indispensable role in data analysis. Consequently, managing and processing data will fundamentally change, and new data mining and machine learning algorithms will be developed to deal with these changes. Explosion of data is expected to be in genomic, proteomic, and other

"omic" data and molecular and system biology. Probes that collect data from tissue are more complex and allow simultaneous tracking and collect more data than few years ago. Authors have also considered several issues such as security, scalability of storage, scalability of analysis, peer with other private clouds, and peer with public clouds.

Metalearning presents powerful methodology which enables learning on its past knowledge of solving different tasks. This methodology is used in fraud detection [36], time series forecasting [37], load forecasting [11], and others.

3. Metalearning Framework for Clustering Biomedical Data

Exponential growth of the data and rapid development of large number of complex and computationally intensive data mining algorithms led to one of the major problems in modern data mining: selection of the best algorithm for data at hand [38]. Namely, analyst often does not have enough time or resources for creating models and evaluating them with all available algorithms.

One of the most promising approaches for dealing with this problem is metalearning [39, 40]. Metalearning is methodology which solves different data mining tasks based on past knowledge. The main idea is to store history of experimental results with descriptions (meta-attributes) of datasets (e.g., dataset characteristics, algorithm, and classification accuracies for classification problems) and, based on this, to create metamodel (classification or regression) that will predict the performances of each algorithm on new dataset. Creating of such metamodel (with good performance) would reduce the need for brute force evaluation (evaluation of every algorithm).

Metalearning system is built on set of algorithm or combination of algorithms (ensembles). Therefore, every algorithm or combination of algorithms (ensembles) is simpler. Theoretically, metalearning system can be infinitely large by putting metalearning as component of other metalearning systems. An advantage of its using is that it can address new types of tasks that have not been seen but are similar to already defined problems.

Metalearning, by Smith-Miles [39, 40], is defined with the following aspects:

(i) the problem space, P, which represents set of instances (datasets) of a given problem class;

(ii) the meta-attribute space, M, which contains characteristics that describe existing problems (e.g., number of attributes, entropy, normality, etc.);

(iii) the algorithm space, A, which represents the set of candidate algorithms which can be used to solve the problems defined in problem space P;

(iv) a performance metric, Y, which represents measures of performance of an algorithm on a problem (e.g., classification accuracy (for classification problems) or root mean square error (for regression problems)).

General procedure for metalearning is done in several steps: first, datasets from problem space are evaluated by algorithms from algorithm space. Further, metafeatures of the datasets are related to algorithm performance, forming the database of metaexamples. Then, regression or classification models are created (and evaluated by performance metric). Finally, when new problem (dataset) arrives, meta-attributes are extracted and performance prediction is made. In this way analyst does not have to evaluate each algorithm on each dataset but only ones with the best predicted performance of the problem and algorithm spaces. While technologies for data collection enabled cheap and fast accumulation of data and extension of problem space, development and collection of data mining algorithms is a more difficult problem since development of new algorithms demands a lot of time and effort, and also different algorithms are implemented on different platforms.

The most important issue for good performance of metalearning systems is the size of problem and algorithm space because the accuracy of metamodels is directly dependant on these spaces [39, 40]. This means that cloud based system and service oriented architecture should be natural environment for this kind of systems because it would enable community based extension models to be created and evaluated by performance metric in order to capture relations between meta-attributes and algorithm performance.

With metalearning approach in solving problems time for choosing appropriate algorithm for problem is greatly reduced but requires time for creating and updating metamodels, especially if data and algorithms are gathered from community. This is one of the main motivations for integration of such a system in cloud based environment and integration with big data technologies (like Hadoop, Hive, and Mahout) for storing and aggregation of data and predictive modeling.

3.1. Component Based Metalearning System for Biomedical Data. One of the promising approaches for tackling these problems (existence of large algorithm space and existence of efficient procedure for selection of the best algorithm) is component based data mining algorithm design [41–43]. This approach divides algorithms with similar structure (in this case representative based algorithms) into parts with the same functionality called subproblems. Every subproblem has standardized I/O structure and can be solved with one or more reusable components (RCs), presented in Table 1. This approach combination of RCs, which originates from different algorithms, can be used to design large number (thousands) of new "hybrid" algorithms. This approach gave very promising results in the area of clustering biomedical (gene expression) data [8, 44, 45].

Combining RCs is used for reproducing or creation of cluster algorithms. For example, K-means algorithms can be reconstructed as RANDOM-EUCLIDEAN-MEAN-COMPACT. However, a new hybrid algorithm can be constructed using DIANA-CORREL-MEDIAN-CONN, where DIANA is used to initialize representatives, CORREL to measure distance, MEDIAN to update representatives, and CONN to evaluate clusters.

Table 1: Sub-problems and RCs for generic clustering algorithm design.

Sub-problem	Reusable components
Initialize representatives	DIANA, RANDOM, XMEANS, GMEANS, PCA, KMEANS++, SPSS
Measure distance	EUCLIDEAN, CITY, CORREL, COSINE
Update representatives	MEAN, MEDIAN, ONLINE
Evaluate clusters	AIC, BIC, SILHOU, COMPACT, XB, CONN

Extended metalearning system, shown in Figure 2, is used in this research. Problem space is presented in upper left cloud. Every problem (dataset) from problem space P has its task (clustering, classification, regression, etc.) denounced x. Based on problem, function f extracts meta-attributes. For selected problem, based on meta-attributes, function S selects algorithm from algorithm space A. Every algorithm is constructed from reusable components (RCs), from which additional meta-attributes were derived (algorithm descriptions). Also, as a result of clustering algorithm on specific dataset internal evaluation measures (additional meta-attributes) are calculated and saved as meta-attributes. Central cloud is the most important part of metalearning system. It is responsible for ranking and selection of algorithms. Inputs in this cloud are task x, algorithm a, and performance metric π. Meta-attributes created for each task x are input for ranking and selection, as they are a basis for learning on metalevel. For most tasks performance of meta-learning system is calculated earlier, as output label, and it is available on metalevel.

3.2. Initial Evaluation of Component Based Metalearning System for Clustering Biomedical Data. In this section we will describe the data and the procedure for initial evaluation of the proposed system. 30 datasets gathered from original metalearning system [46] were used (http://bioinformatics.rutgers.edu/Static/Supplements/CompCancer/datasets.htm).

For the construction of the metaexamples a set of 13 meta-attributes for dataset description proposed by Nascimiento et al. [46] are used (detailed description of datasets and metafeatures can be found in Nascimiento et al. [46]).

Additionally, meta-attribute space is extended with descriptions of algorithms (four components of algorithm and normalization type described in Table 1) and internal cluster evaluation measures including compactness, global silhouette index, AIC, BIC, XB-index, and connectivity. These three types of meta-attributes (dataset descriptions, reusable components, and internal evaluation measures) form the space of 24 meta-attributes.

Component based clustering algorithms were used to define algorithm space. For that purpose 504 RC-based cluster algorithms were designed for experimental evaluation. These algorithms were built by combining already described RCs (Table 1) with 4 different normalization techniques, which lead to total of 2016 clustering experiments.

Table 2: Meta-algorithm performance.

Algorithm/Error	RMSE	MAE
RBFN	0.143	0.109 (±0.092)
LR	0.111	0.086 (±0.070)
LMSR	0.265	0.094 (±0.248)
NN	0.101	0.064 (±0.078)
SVM	**0.050**	**0.034 (±0.036)**

For validation of clustering models AMI (adjusted mutual information) index was used since it is recently recommended as a "general purpose" measure for clustering validation, comparison, and algorithm design [47], after exhaustive comparison between a number of information theoretic and pair counting measures. Even more, this measure is thoroughly evaluated on gene expression microarray data.

After validation of component based clustering algorithms on 30 datasets, 55326 valid results were gathered, which represent metaexample repository. Next step was identification of the best algorithm for ranking and selection of algorithms for clustering gene expression microarray data.

A procedure for ranking and selection of the best clustering algorithms is based on regression algorithms, called meta-algorithms, which predict (regression task) AMI values based on a dataset metafeatures, algorithm components, and internal evaluation measures.

In this research five meta-algorithms were used. Those are radial basis function network (RBFN), linear regression (LR), least median square regression (LMSR), neural network (NN), and support vector machine (SVM).

Estimate of quality of regression algorithms is done using mean absolute error (MAE) and root mean squared error (RMSE). Validation of results is done using 70% of dataset for training the model and the remaining 30% for testing.

Performance of each algorithm, in terms of MAE and RMSE, and best values are shown in bold (Table 2).

Although all five algorithms showed good results, SVM, as in Table 2, gave the best performance, and this model should be used for prediction of algorithm performances for new datasets. RMSE of 0.05 and MAE of 0.034, with the smallest variance (numbers in brackets), indicate that this metamodel is applicable to the new problems since AMI measure takes values from 0 to 1 where 1 is the best. Note that with an extension of algorithm space and problem space these results could be changed and so continuous evaluation of available meta-algorithms should be done. Because of this it is important to have adequate computing infrastructure for processing big data and updating the models. Process for creating and updating the models is presented in Figure 3. Creation of model contains several steps. First, microarray metaexamples are loaded, from which only important variables are selected. After that, data preparation phase was conducted where only those attributes that are important for model building were selected, label attribute was set on AMI attribute, missing values were replaced with average value, and nominal values were transformed to numerical values using dummy coding. Modeling phase is conducted

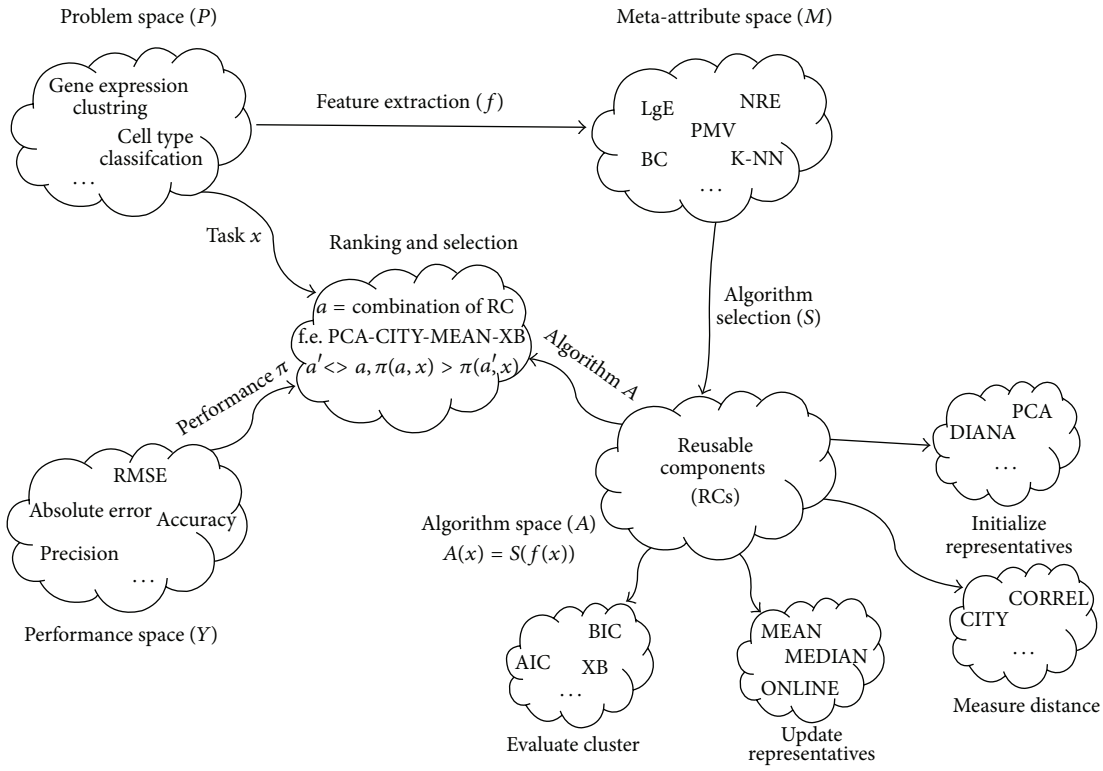

Figure 2: Extended metalearning system for clustering biomedical data.

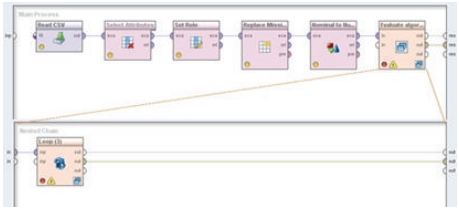

Figure 3: Main process for finding the best model for prediction of AMI.

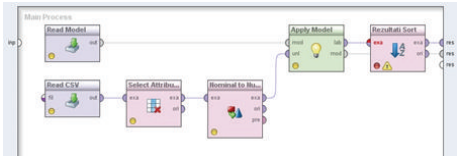

Figure 4: Stream for application of metalearning system on new cases.

using 10-fold cross validation where the above-mentioned five algorithms were used. Every trained model is saved on hard disk, which allows its reusability.

Automatic application of the selected (in this case SVM) model is presented in Figure 4. After every update of the model or after inserting new dataset, this process needs to be updated. Saved model, in this case SVM, is loaded and applied on new dataset. Results gathered are sorted and exported. Detailed information can be found in Radovanovic et al. [48].

4. Extended Cloud Based Model for Big Data Analysis

Dai et al. [16] addressed the problems of storing and analysis of biomedical data by proposing a cloud based model for analysis of biomedical data and it is composed of four service categories:

(i) data as a service (DaaS),

(ii) software as a service (SaaS),

(iii) platform as a service (PaaS),

(iv) infrastructure as a service (IaaS).

DaaS is group of cloud services which enables on demand data access and provides up-to-date data that are accessible by a wide range of devices that are connected over the web. In case of bioinformatics, Amazon Web Services (AWS) provides repository of public archives (data sets), including GenBank, Unigene, Ensembl, and Influenza Virus, which can be accessed from cloud based applications.

Since bioinformatics requires a large diversity of software tools for data analyses, the task of SaaS in bioinformatics is to deliver and enable remote access to software services online. Thus, installation of software tools on desktop computer is no longer required. Another one advantage of using SaaS is enabling much easier collaboration between dispersed groups of users.

Platform as a service (PaaS) should offer a programmable environment for users in order to develop, test, and deploy cloud applications. Computer resources scale automatically

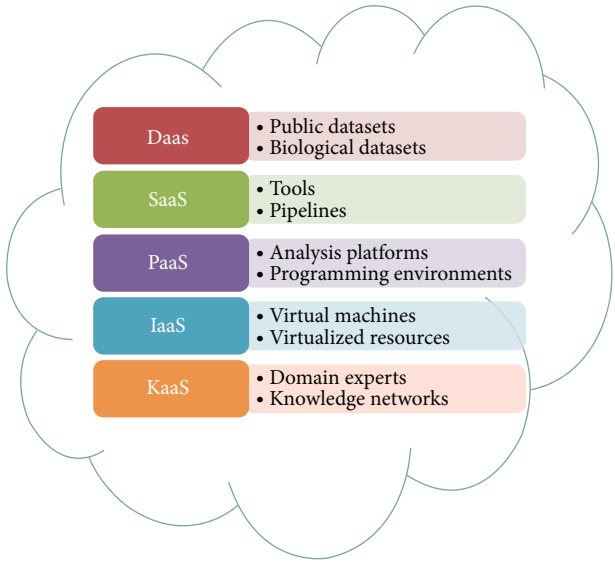

Figure 5: Cloud based system [16] integrated with KaaS extension [17] for analysis of biomedical data.

without user interference. In most PaaS services, besides programming environment, database and web server are available.

Since most medical institutions do not have computing resources, such as CPU, IaaS offers a full computer infrastructure delivering virtualized resources. Those virtualized resources can be operating systems, RAM, CPU, or other computer resource.

Lai et al. [17] introduced a new service model in cloud computing—the knowledge as a service (KaaS) that facilitates the interoperations among members in a knowledge network. Extended model is depicted on Figure 5.

This new service model relies on data created in a collaboration process of domain experts. It was recognized as a new form of cloud service and categorized as KaaS. In a nutshell, such approach is data driven and lacks higher order structure in order to provide knowledge as a service. This implies that KaaS in this form is provided by human, domain experts. We extend this class of services with data and model approach. By combining data driven models and expert knowledge that is stored in unstructured data like documents, notes, and collaborations, knowledge is created and offered to cloud users. Specifically, we extend models of [16, 17] by including big data technologies, platforms for data mining, and metamodels for ranking and selection of the best algorithms for biomedical data mining. System components are displayed on a diagram (Figure 6) and classified according to service type directed towards end user. Big data engine provides data storage and access and it is a basis of each class of services (diagram center).

Key component of KaaS is metalearning algorithms and selected algorithms for clustering biomedical data. Both are represented by algorithm space component, and both are accompanied by their describing metamodels. These metamodels are used in runtime for ranking and selection of best algorithms for the new problem (dataset). Experimental results provide knowledge on performances of each algorithm with meta-attributes on each dataset. Data flows are kept using big data engine such as Hadoop. Segment of experimental results component lies on DaaS since these meta-attributes can be provided as a data service. DaaS provides biomedical data. It is divided into protected and public segments.

Cloud approach provides data accumulation and higher availability of data to interested parties (with rights of access implied). Interested parties can be found among not only medical employees and researchers but also community using different software tools to access data using DaaS.

Central part of the circle (Figure 6) contains components of a big data engine (HDFS, Hive, and Mahout) where all the data are centralized (medical data, metadata, and algorithms performances). Additionally, big data engine provides interfaces for data manipulation and analysis. Apache Hive [49] is data warehouse software built on top of Hadoop, used for querying and managing large datasets residing in distributed storage. Apache Mahout [50] is also built on top of Hadoop and is used as a scalable machine learning library for classification, clustering, recommender systems, and dimension reduction.

Our solution provides software for data analysis in a service form (SaaS) with respect of SOA dependability [51]. Additionally, third party software can easily become an integrated part of SaaS (RapidMiner [52], R [53], or others).

These software solutions are recommended because they have a direct interface for access and analysis of big data (e.g., Radoop [54] allows using visual RapidMiner interface and has operators that run distributed algorithms based on Hadoop, Hive [49], and Mahout [50] without writing any code).

Most medical institutions require not only computing resources, such as CPUs, but also communication infrastructure; these components are offered as a part of IaaS, in a form of virtualized resources. Higher level of service rests on using platform such as variety of operating systems. But platform can also provide tools for development of algorithms and applications (Eclipse, Netbeans, RapidMiner, R. . .). The circle is closed by development of new algorithms and also for algorithm deployment and execution. For this reason algorithms space component is partly situated in PaaS space.

5. Conclusion and Future Research

In this paper we proposed a cloud based architecture for storing, analysis, and predictive modeling of biomedical big data. Existing service based cloud architecture is extended by including metalearning system as a data and model driven knowledge service. As a part of the proposed architecture, we provided a support for community based gathering of data and algorithms that is an important precondition for quality of metalearning. Advancement of this research area and adding new value are enabled through platform for development and execution of distributed data mining processes

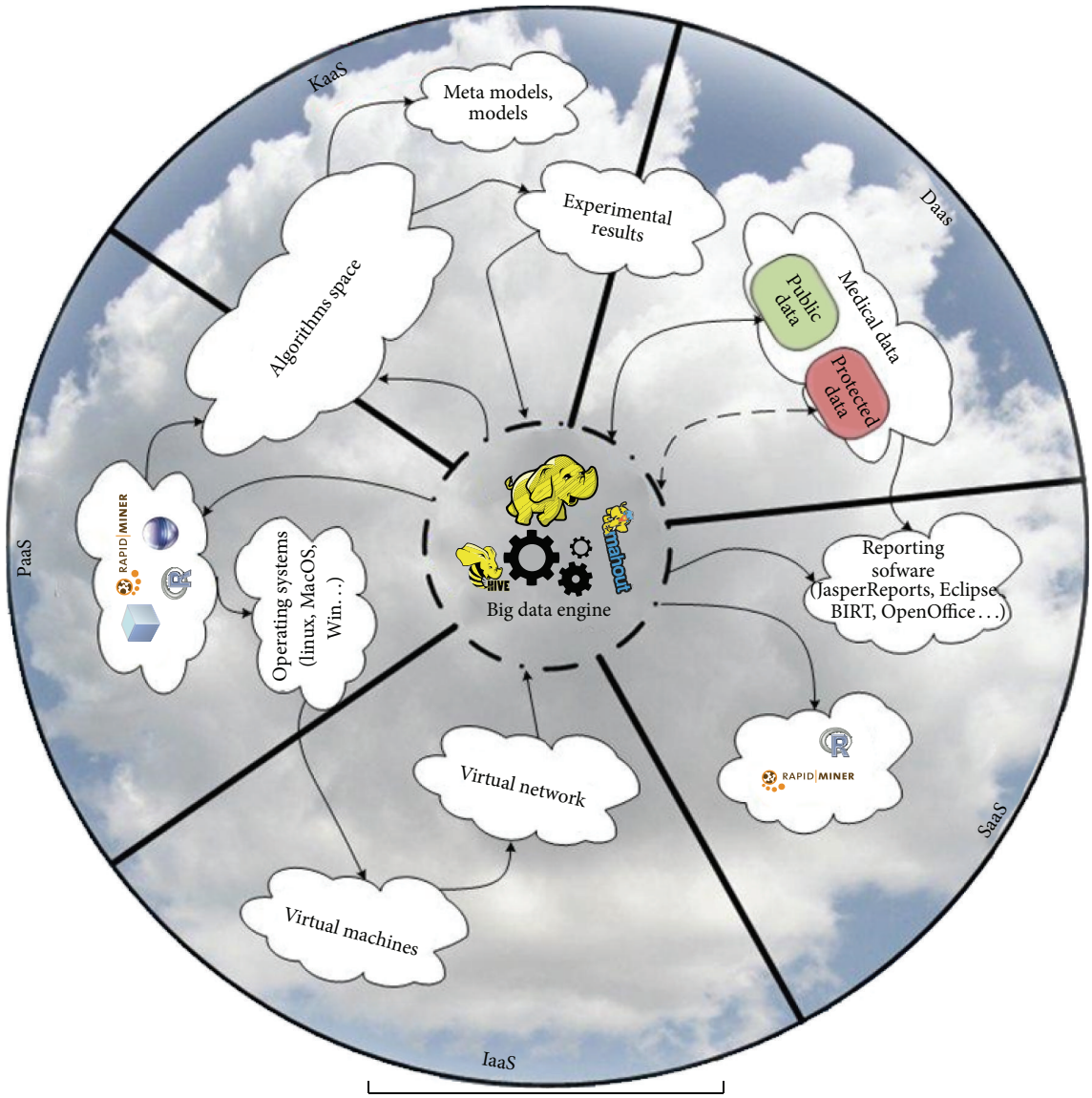

Figure 6: Cloud based system for predictive modeling of biomedical data.

and algorithms. Finally, we provided data and model driven decision support on selecting best algorithms for working with biomedical data.

Retrospectively, proposed solution focuses on a specific type of biomedical data, while other types still remain to be included and evaluated. Data security and privacy still remains a concern to be taken into a more serious account. In order to provide even further impact in research community, additional work is necessary on providing interoperability among potential open source components.

System was tested on microarray gene expression data, with specific meta-attributes for this data type (e.g., chip type). Further efforts will be made to include other types of biomedical data. This will be done by identifying specific meta-attributes that fit newly included data types. Additionally, as a further work, integration with OpenML [55] platform, used for storing and gathering datasets and clustering algorithm runs, is planned. This platform provides a base for a community to share experiments, algorithms, and data. Significant clustering algorithm meta-attributes can be extracted and used for updating our metalearning system.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgment

This research is partially funded by Grants from the Serbian Ministry of Science and Technological Development, Contract nos. III 41008 and TR 32013.

References

[1] K. Kincade, "Data mining: digging for healthcare gold," *Insurance and Technology*, vol. 23, no. 2, pp. IM2–IM7, 1998.

[2] H. C. Koh and G. Tan, "Data mining applications in healthcare," *Journal of Healthcare Information Management*, vol. 19, no. 2, pp. 64–72, 2011.

[3] K. Srinivas, B. K. Rani, and A. Govrdhan, "Applications of data mining techniques in healthcare and prediction of heart attacks," *International Journal on Computer Science and Engineering*, vol. 2, no. 2, pp. 250–255, 2010.

[4] P. Maji, "Mutual information-based supervised attribute clustering for microarray sample classification," *IEEE Transactions on Knowledge and Data Engineering*, vol. 24, no. 1, pp. 127–140, 2012.

[5] R. Das, "A comparison of multiple classification methods for diagnosis of Parkinson disease," *Expert Systems with Applications*, vol. 37, no. 2, pp. 1568–1572, 2010.

[6] A. Thomas, N. H. Patterson, M. M. Marcinkiewicz, A. Lazaris, P. Metrakos, and P. Chaurand, "Histology-driven data mining of lipid signatures from multiple imaging mass spectrometry analyses: application to human colorectal cancer liver metastasis biopsies," *Analytical Chemistry*, vol. 85, no. 5, pp. 2860–2866, 2013.

[7] R. L. Grossman and K. P. White, "A vision for a biomedical cloud," *Journal of Internal Medicine*, vol. 271, no. 2, pp. 122–130, 2012.

[8] M. Vukićević, B. Delibasic, Z. Obradovic, M. Jovanović, and M. Suknović, "A method for design of data-tailored partitioning algorithms for optimizing the number of clusters in microarray analysis," in *Proceedings of the IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB '12)*, pp. 252–259, IEEE, San Diego, Calif, USA, 2012.

[9] H. Chae, I. Jung, H. Lee, S. Marru, S. W. Lee, and S. Kim, "Bio and health informatics meets cloud: BioVLab as an example," *Health Information Science and Systems*, vol. 1, no. 1, article 6, 2013.

[10] J. Sun and C. K. Reddy, "Big data analytics for healthcare," in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 1525–1525, ACM, August 2013.

[11] M. Matijaš, J. A. Suykens, and S. Krajcar, "Load forecasting using a multivariate meta-learning system," *Expert Systems With Applications*, vol. 40, no. 11, pp. 4427–4437, 2013.

[12] S. Zhou, K. K. Lai, and J. Yen, "A dynamic meta-learning rate-based model for gold market forecasting," *Expert Systems with Applications*, vol. 39, no. 6, pp. 6168–6173, 2012.

[13] J. Kanda, C. Soares, E. Hruschka, and A. de Carvalho, "A meta-learning approach to select meta-heuristics for the traveling salesman problem using MLP-based label ranking," in *Neural Information Processing*, pp. 488–495, Springer, Berlin, Germany, 2012.

[14] A. Attig and P. Perner, "Meta-learning for image processing based on case-based reasoning," in *Computational Intelligence in Healthcare*, vol. 4, pp. 229–264, Springer, Berlin, Germany, 2010.

[15] A. Vera-Baquero, R. Colomo-Palacios, and O. Molloy, "Business process analytics using a big data approach," *IT Professional*, vol. 15, no. 6, pp. 29–35, 2013.

[16] L. Dai, X. Gao, Y. Guo, J. Xiao, and Z. Zhang, "Bioinformatics clouds for big data manipulation," *Biology Direct*, vol. 7, no. 1, article 43, 2012.

[17] I. K. Lai, S. K. Tam, and M. F. Chan, "Knowledge cloud system for network collaboration: a case study in medical service industry in China," *Expert Systems with Applications*, vol. 39, no. 15, pp. 12205–12212, 2012.

[18] S. Sonnenburg, M. L. Braun, S. O. Cheng et al., "The need for open source software in machine learning," *Journal of Machine Learning Research*, vol. 8, pp. 2443–2466, 2007.

[19] S. Marston, Z. Li, S. Bandyopadhyay, J. Zhang, and A. Ghalsasi, "Cloud computing: the business perspective," *Decision Support Systems*, vol. 51, no. 1, pp. 176–189, 2011.

[20] N. Sultan, "Cloud computing for education: a new dawn?" *International Journal of Information Management*, vol. 30, no. 2, pp. 109–116, 2010.

[21] W. Zhang and Q. Chen, "From E-government to C-government via cloud computing," in *Proceedings of the 1st International Conference on E-Business and E-Government (ICEE '10)*, pp. 679–682, IEEE, May 2010.

[22] A. Fernández, S. del Río, F. Herrera, and J. M. Benítez, "An overview on the structure and applications for business intelligence and data mining in cloud computing," in *Proceedings of the 7th International Conference on Knowledge Management in Organizations: Service and Cloud Computing*, pp. 559–570, Springer, Berlin, Germany, January 2013.

[23] S. P. Ahuja, S. Mani, and J. Zambrano, "A survey of the state of cloud computing in healthcare," *Network and Communication Technologies*, vol. 1, no. 2, pp. 12–19, 2012.

[24] W. Li, J. Yan, Y. Yan, and J. Zhang, "Xbase: cloud-enabled information appliance for healthcare," in *Proceedings of the 13th International Conference on Extending Database Technology: Advances in Database Technology (EDBT '10)*, pp. 675–680, March 2010.

[25] M. C. Schatz, B. Langmead, and S. L. Salzberg, "Cloud computing and the DNA data race," *Nature Biotechnology*, vol. 28, no. 7, pp. 691–693, 2010.

[26] T. Oinn, M. Addis, J. Ferris et al., "Taverna: a tool for the composition and enactment of bioinformatics workflows," *Bioinformatics*, vol. 20, no. 17, pp. 3045–3054, 2004.

[27] B. Giardine, C. Riemer, R. C. Hardison et al., "Galaxy: a platform for interactive large-scale genome analysis," *Genome Research*, vol. 15, no. 10, pp. 1451–1455, 2005.

[28] M. C. Schatz, "CloudBurst: highly sensitive read mapping with MapReduce," *Bioinformatics*, vol. 25, no. 11, pp. 1363–1369, 2009.

[29] B. D. O'Connor, B. Merriman, and S. F. Nelson, "SeqWare query engine: storing and searching sequence data in the cloud," *BMC Bioinformatics*, vol. 11, supplement 12, article S2, 2010.

[30] D. Hong, A. Rhie, S. Park et al., "FX: an RNA-seq analysis tool on the cloud," *Bioinformatics*, vol. 28, no. 5, pp. 721–723, 2012.

[31] S. V. Angiuoli, M. Matalka, A. Gussman et al., "CloVR: a virtual machine for automated and portable sequence analysis from the desktop using cloud computing," *BMC Bioinformatics*, vol. 12, no. 1, article 356, 2011.

[32] R. Zhang and L. Liu, "Security models and requirements for healthcare application clouds," in *Proceedings of the 3rd IEEE International Conference on Cloud Computing (CLOUD '10)*, pp. 268–275, IEEE, July 2010.

[33] R. Wooten, R. Klink, F. Sinek, Y. Bai, and M. Sharma, "Design and implementation of a secure healthcare social cloud system," in *Proceedings of the 12th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid '12)*, pp. 805–810, IEEE, May 2012.

[34] W. Liu and E. K. Park, "E-healthcare cloud computing application solutions: cloud-enabling characteristics, challenges and adaptations," in *Proceedings of the International Conference on Computing, Networking and Communications (ICNC '13)*, pp. 437–443, IEEE, January 2013.

[35] T. Giles and R. Wilcox, "IBM Watson and medical records text analytics," in *Proceedings of the 2011 Healthcare Information and Management Systems Society Meeting*, 2011.

[36] A. Abbasi, C. Albrecht, A. Vance, and J. Hansen, "Metafraud: a meta-learning framework for detecting financial fraud," *MIS Quarterly*, vol. 36, no. 4, pp. 1293–1327, 2012.

[37] C. Lemke and B. Gabrys, "Meta-learning for time series forecasting and forecast combination," *Neurocomputing*, vol. 73, no. 10-12, pp. 2006–2016, 2010.

[38] N. Iam-On, T. Boongoen, and S. Garrett, "LCE: a link-based cluster ensemble method for improved gene expression data analysis," *Bioinformatics*, vol. 26, no. 12, pp. 1513–1519, 2010.

[39] K. A. Smith-Miles, "Towards insightful algorithm selection for optimisation using meta-learning concepts," in *Proceedings of the (IEEE World Congress on Computational Intelligence) IEEE International Joint Conference on Neural Networks (IJCNN '08)*, pp. 4118–4124, IEEE, Hong Kong, June 2008.

[40] K. A. Smith-Miles, "Cross-disciplinary perspectives on meta-learning for algorithm selection," *ACM Computing Surveys*, vol. 41, no. 1, article 6, 2008.

[41] B. Delibašić, K. Kirchner, J. Ruhland, M. Jovanović, and M. Vukićević, "Reusable components for partitioning clustering algorithms," *Artificial Intelligence Review*, vol. 32, no. 1–4, pp. 59–75, 2009.

[42] B. Delibašić, M. Vukićević, M. Jovanović, K. Kirchner, J. Ruhland, and M. Suknović, "An architecture for component-based design of representative-based clustering algorithms," *Data and Knowledge Engineering*, vol. 75, pp. 78–98, 2012.

[43] M. Suknović, B. Delibašić, M. Jovanović, M. Vukićević, D. Becejski-Vujaklija, and Z. Obradovic, "Reusable components in decision tree induction algorithms," *Computational Statistics*, vol. 27, no. 1, pp. 127–148, 2012.

[44] M. Vukićević, B. Delibašić, M. Jovanović, M. Suknović, and Z. Obradović, "Internal evaluation measures as proxies for external indices in clustering gene expression data," in *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM '11)*, pp. 574–577, IEEE Computer Society, 2011.

[45] M. Vukićević, K. Kirchner, B. Delibašić, M. Jovanović, J. Ruhland, and M. Suknović, "Finding best algorithmic components for clustering microarray data," *Knowledge and Information Systems*, vol. 35, no. 1, pp. 111–130, 2013.

[46] A. Nascimiento, R. Prudencio, M. de Souto, and I. Costa, "Mining rules for the automatic selection process of clustering methods applied to cancer gene expression data," in *Proceedings of the 19th International Conference on Artificial Neural Networks: Part II*, pp. 20–29, Springer, Limassol, Cyprus, 2009.

[47] N. X. Vinh, J. Epps, and J. Bailey, "Information theoretic measures for clusterings comparison: variants, properties, normalization and correction for chance," *Journal of Machine Learning Research*, vol. 11, pp. 2837–2854, 2010.

[48] S. Radovanović, M. Vukićević, M. Jovanović, B. Delibasic, and M. Suknović, "Meta-learning system for clustering gene expression microarray data," in *Proceedings of the 4th RapidMiner Community Meeting and Conference (RCOMM '13)*, pp. 97–111, Shaker, Porto, Portugal, 2013.

[49] A. Thusoo, J. S. Sarma, N. Jain, Z. Shao, P. Chakka, and S. Anthony, "Hive: a warehousing solution over a map-reduce framework," *Proceedings of the VLDB Endowment*, vol. 2, no. 2, pp. 1626–1629, 2009.

[50] R. Anil, T. Dunning, and E. Friedman, *Mahout in Action*, Manning, 2011.

[51] V. Stantchev and M. Malek, "Addressing dependability throughout the SOA life cycle," *IEEE Transactions on Services Computing*, vol. 4, no. 2, pp. 85–95, 2011.

[52] I. Mierswa, M. Wurst, R. Klinkenberg, M. Scholz, and T. Euler, "YALE: rapid prototyping for complex data mining tasks," in *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '06)*, pp. 935–940, ACM, August 2006.

[53] Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2005.

[54] Z. Prekopcsák, G. Makrai, T. Henk, and C. Gáspár-Papanek, "Radoop: analyzing big data with rapidminer and hadoop," in *Proceedings of the 2nd RapidMiner Community Meeting and Conference (RCOMM '11)*, June 2011.

[55] J. N. van Rijn, B. Bischl, L. Torgo, B. Gao, V. Umaashankar, and S. Fischer, "OpenML: a collaborative science platform," in *Machine Learning and Knowledge Discovery in Databases*, pp. 645–649, Springer, Berlin, Germany, 2013.