The Proposal of Service Oriented Data Mining System for Solving Real-Life Classification and Regression Problems

Agnieszka Prusiewicz[1] and Maciej Zięba[1],

[1] Institute of Informatics, Faculty of Computer Science and Management, Wroclaw University of Technology, Wybrzeże Wyspiańskiego 27, 50-370 Wrocław, Poland
{Agnieszka.Prusiewicz, Maciej.Zięba}@pwr.wroc.pl

Abstract. In this work we propose an innovative approach to data mining problem. We propose very flexible data mining system based on service-oriented architecture. Developing applications according to SOA paradigm emerges from the rapid development of the new technology direct known as sustainability science. Each of data mining functionalities is delivered by the execution of the proper Web service. The Web services, described by input and output parameters and the semantic description of its functionalities, are accessible for all applications that are integrated via Enterprise Service Bus.

Keywords: Service Oriented Data Mining, Sustainable design, SOA, Classification Services

1 Introduction

With the rising necessity of mining huge data volumes and knowledge discovery from many distributed resources there is a natural interest of using grid solutions. Execution machine learning algorithms in distributed environments allow organizations to execute computationally expensive algorithms on large databases, with relatively inexpensive hardware. Additionally an opportunity to merge data and discovered knowledge from many geographically distributed resources are created by the Internet. These elements favour of approving of a new field named as Distributed Data Mining (DDM) [11]. The survey of some Grid-based data mining systems is given in [1]. The other type of distributed computing, that rapidly develops in last decade is based on Service Oriented Architecture (SOA) paradigm [8]. The main idea of SOA is to treat applications, systems and processes as encapsulated components, which are called services. These services are represented by input and output parameters and the semantic description of its functionalities. Combining distributed data mining techniques with Web services has a lot of advantages. Web services are currently seen as a solution for integration of the heterogeneous resources and making heterogeneous systems interoperable. They are self-contained, self-describing and modular applications that can be published and invoked across the Web [15]. As an example of some Web services based data

mining applications we can indicate Web based system for metadata learning (WebDiscC) [13] or Association Rule Mining software called DisDaMin [2]. But the area of Web services-based data mining systems is still not well recognised. Taking into account current tendencies toward computer systems development there is a need for elaboration data mining distributed systems compatible with SOA paradigm.

In this work we propose Service Oriented Data Mining System (SODM System) for solving chosen data mining tasks i.e. classification tasks, equipped with mechanisms for advising which classifier should be used as the best for a given user request (data and requirements that must be classified). The advantage of our solution that is a consequence of compatibility with SOA paradigm, is that the users may use the model of classifiers and classifiers that have be created by the others. It is caused by implementing the functionalities of building models of classifiers and classifiers as Web services that are published and accessible via Internet.

2 Contribution to Sustainability

Presented in this paper Data Mining System is innovative approach to well-known machine learning solutions. Developing applications according to SOA paradigm emerges from the rapid development of the new technology direct known as sustainability science. Representing machine learning solutions as Data Mining Services has a significant contribution to sustainable design of new technological solutions. Data mining solutions are successively used for supporting decision-making processes in nature-society systems. They can be applied for extracting long-term trends in environments or to predict survival period in modern society. One of the drawbacks of data mining approaches is the problem with integration between such solutions and life-support systems from different fields of science. For instance, if there is a need to classify credit holders in bank system additional functionalities for customer modelling must be implemented in such system. SODM System solves the problem of integration and accessibility. In data mining solutions are fully accessible for all applications, which are integrated via ESB. For example, each application has an ability to invoke services responsible for creating clusters of objects simply by sending the objects in SOAP message. There is no need to create additional components of the application responsible for solving data mining problems (that is time consuming and demands data mining abilities), because these solutions are available as Data Mining Services in SODM System. Life-support systems, which communicate, with Data Mining Services are becoming interdisciplinary systems. Additionally, different systems representing various disciplines can invoke the same Data Mining Services. For instance, the same model can be used to predict survival period by medical and governmental system.

Every new solution developed by economists, biologists or mathematicians can be easily added to the SODM System. According to the basic sustainable design principle i.e. durability new data mining functionalities may be add as the new services without a need of rebuilding the overall system. This solution can be easily evaluated by data mining researchers, or simply used by all applications, which has access to ESB.

4 Service Oriented Architecture And Web Services

The basic concept of Service Oriented Architecture (SOA) approach is a semantically described service. In this context, SOA is the application framework that enables organizations to build, deploy and integrate these services independent of the technology systems on which they run. In SOA, applications and infrastructure can be managed as a set of reusable assets and services. The main idea about this architecture was that businesses that use SOA could respond faster to market opportunities and get more value from their existing technology assets [12]. According to SOA paradigm services are published and then access via Enterprise Service Bus (ESB) and used by Web applications. ESB is an implementation technology supporting SOA approach. ESB as an enterprise-wide extendable middleware infrastructure provides virtualization and management of service interactions, including support for the communication, mediation, transformation, and integration technologies required by services [10]. Web services are technologies that are based on XML (Extensible Markup Language) for messaging, service description and discovery. [8]. Web services use such standards as: SOAP, WSDL, and UDDI. SOAP (Simple Object Access Protocol) is a protocol for application communication and exchanging information over HTTP. WSDL (Web Services Description Language) is an XML-based language for describing Web services and how to access them. And finally UDDI (Universal Description, Discovery and Integration) is open standard for services discovery and invoking. UDDI being interrogated by SOAP messages provides access to WSDL documents that describe the protocol bindings and message formats required to interact with the Web services. In other words specifications define a way to publish and discover information about Web services. The relations between services provider and requester using xml standards are as follows. The provider gives an access to a Web service by publishing a WSDL description of its Web service, and the requester accesses the description using a UDDI or other type of registry that contains register of discovered services, and requests the execution of the provider's service by sending a SOAP message [8](Fig.1).

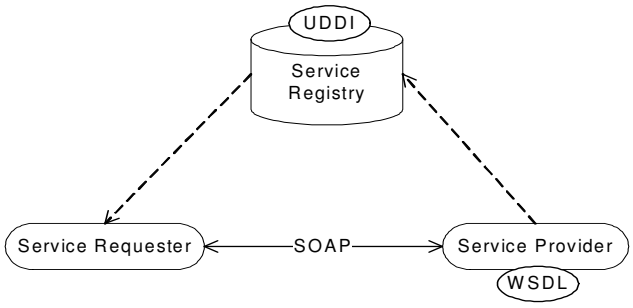

Fig. 1. The schema of interaction between Web services components

5 The Architecture of SODM System

The problems of classification are one of the common issues of data mining field [5,6,7,17]. There are plenty of dissertations, which touch mainly following problems: developing ensemble models for classification accuracy improvement, dealing with missing values of attributes, or building algorithms for incremental learning [3,4,14]. In most cases authors concentrate mostly on creating complex, difficult to understand methods, which can be used by data mining experts. Considering a classification task connected for example with labelling of web clients for marketing purposes there is a need for building a suitable model of classifier using past observations of the clients and their behaviours. The built model is further used to classify the new clients. Such model should be easily updated for a new set of data. To solve the classification tasks the application must be equipped with the mechanisms of building, updating and finding the most suitable model of the classifier. We propose SODM System that is based on Service Oriented Architecture (SOA) and uses encapsulation of data mining solutions in services. Hence each of the models of classifiers is represented by a service. In each service it is possible to distinguish operations responsible for creating and testing the model or classifying object using created model. The communication between applications (service clients) and service providers is made by Enterprise Service Bus (ESB) using Simple Object Access Protocol (SOAP). SODM System consists of the following components: User Interfaces, Service Usage Repository, Data Mining Services and Data Mining Services Manager (Fig. 2). All the components are integrated via ESB. User Interface is a component responsible for communication between users and services. It can be default application, which has ability to communicate with other Data Mining Components. More then one User Interface can be integrated with ESB. For instance using one interface it is possible to invoke Data Mining Services responsible for building and updating the model of classifier and the other can be used on mobile phone to classify unclassified object using this model. Service Usage Repository is responsible monitoring the usage of Data Mining Services. Data Mining Services are services responsible for solving classification and regression problems, clustering objects, data filtering or extracting associative rules. To improve the quality of usage of those services Data Service Manager is considered in the system. This service is responsible for filtering the most suitable services for the problem stated by the user in a request.

Data mining solutions presented in above architecture have couple of advantages. First of all, Data Mining Services are easily accessible by every service client, which is integrated via ESB. The client can be a mobile application, educational system, or some process in the system.

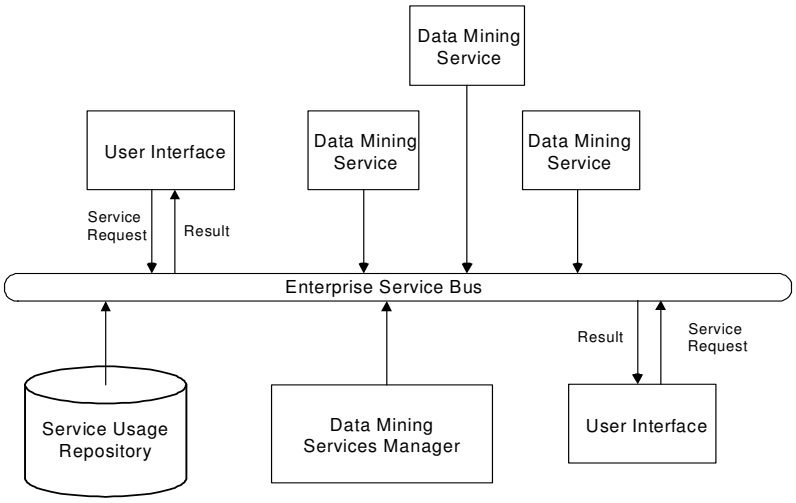

Fig. 2. The Components of SODM System.

To invoke a Data Mining Service it is sufficient to create proper SOAP message, send it using ESB and interpret the response message. There is no need to implement data mining solutions locally because all such mechanisms are implemented and covered in the service. Moreover, the new data mining functionalities can be easily included in the system as a new service with specified input and output parameters. Presented architecture includes also Data Mining Services Manager, which helps to find the best solutions for users without data mining abilities.

5.1 The functionality of SODM System

Following functionalities can be distinguished in the SODM: building and updating classification and regression models, grouping objects using various clustering algorithms, filtering data (missing values of attributes replacement, features selection, etc.) and associative rules extraction. In this paper we concentrate only on functionalities related with solving classification and regression problems. In this field we can distinguish: building the model of classifier (training the classifier), instances classification, testing performance of the model of classifier, printing the model of classifier and updating the model. The problem of building the classifier is one of key issues in pattern recognition field. The model of classifier takes on the input vector of features values of the object to be classified and returns the class label (continuous value if regression problem is considered). This model can be given by an expert (as a set of rules) or it can be extracted using data composed of past observation of the objects (training dataset). In SODM System each of classifiers types (decision trees, decision rules, neural networks) is represented by one Data Mining Service (Classification Service) and building the classifier functionality is fulfilled by an operation which takes training dataset and returns the label of built model. Other functionalities are realized by other operations included in Classification Service. Instances classification is made using operation, which

takes set of unlabeled objects on the input and returns their labels on the output by making classification using existing model. The performance of the model of classifier can be evaluated using testing operation. The dataset is given on the input of the service and testing methodology is defined and the operation returns testing rates values (accuracy of classification, Kappa statistic, etc.). Some of models of classifiers (rules, decision trees) have got understandable structure so there is an operation in each of Classification Service, which returns the structure of the model. There are couples of models of classifiers (Naïve Bayes), which can be easily updated, in incremental learning process. If the model of classifier is updatable the corresponding Classification Service contains an operation responsible for updating the model.

SOSM System has additional functionalities related with filtering Classification Services. These functionalities are fulfilled by operations included in Classification Services Manager. Classification Service Manager is able to choose the models of classifier which are accurate for the given on the input dataset by fining corresponding Classification Services. There is also possibility for defining additional requirements for the model like updatability, which cannot be extracted from the data.

5.2 The Classification Components of SODM System

Classification Services

Following Classification Services are distinguished in our system: NBService, J48Service, JRipService, MLPService and LRService. These services are respectively represents following types of classifiers: Naïve Bayes, J48 (decision tree), JRip (decision rules), MLP (neural network) and LRService (Logistic Regression) [17]. Each of Classification Services contains operations described in previous subsection: buildClassifier (building the model of classifier), classifyInstances (classifying objects), getCapabilities (getting capabilities of the classifier), printClassifier (printing the structure of classifier), testingClassifier (testing the performance of classifier). Additionally, NBService contains the operation, which enables to update the model of classifier [14]. Initially SODM System contains only five Classification Services, but additional models of classifiers can be easily included in the system as service with defined standard operations and parameters.

Classification Services Manager

SODM System includes also managing service responsible for finding the best models according to request given by the client. For instance, the client can invoke the service by sending only the set of past observations, which should be used to build the model of classifier, and the response message will provide the most suitable Classification Services. Classification Service Manager contains three filtering operations. In first operation only dataset is taken on the input and the operation is responsible for finding the types of classifiers (Classification Services),

which can be trained using this dataset. For instance, dataset can contain missing values and only those classifiers can be filtered, which are able to deal with missing values. In second operation some other requirements can be specified, for instance the classifier must be updatable. Third operation requires only model capabilities without putting the dataset on the output.

6 Implementation and Use Case Example

The implementation of SODM System is compatible with SOA standards. Each of SODM services is represented using WSDL language. Services are implemented in Java using Weka library [16]. To present functionalities of the SODM system we consider Educational System. On of the problems, which occurs in such systems, is to dived user (students) into priority groups [9]. The priorities of the users can be used to divide them to early and late registration on courses groups, or to filter students for scholarships awards. Assume that students are described with following three attributes: Average mark form whole studying period (AM), Number of uncompleted courses (NoUC), Number of Awards obtained outside the university (NoA). Educational System collects the past observation of students and their priorities. The goal is to classify the students to priority groups using the knowledge extracted from past observations (training data). Implementing new data mining components in the system can solve the problem but it is much simpler to use Data Mining Services from SODM System. To solve the problem it is necessary to find the type of classifier suitable for the data. To do this Classification Service Manager service can by invoke by putting on the input on the operation representative portion of the data. It is recommended to put the data, which is going to be further used to build the classifier (training data). As a response a specification of Classification Services which corresponds to classifiers types which can be build basing on given on the input dataset are returned. Next, one of the Classification Services can be used to create the model of classifier by invoking buildClassfifer operation of these services. This operation takes on the input training datasets (past observations of students and their priorities in the considered example) and returns the key to the created model. The model can be further used to classify objects (unlabeled students) by invoking classifyInstances operation.

7. Final Remarks and Future Works

In this paper we presented basic concepts of SODM System. The system is based on SOA paradigm so the components are fully accessible via ESB. We presented functionalities of Classification and Regression component of SODM System. In future works other data mining components must be developed to create complete service oriented data mining tool. In particular we will apply ensemble classifiers to improve the accuracy of classification.

Acknowledgements

The research presented in this work has been partially supported by the European Union within the European Regional Development Fund program no. POIG.01.03.01-00-008/08.

References

1. Cannataro M., Congiusta A., Pugliese A., Talia D., Trunfio P, Distributed data mining on grids: Services, tools, and applications, IEEE Transactions on Systems, Man, and Cybernetics, vol. 34 (6), pp. 2451–2465, (2004)
2. Fiolet V., Olejnik R., Lefait G., Toursel B., Optimal grid exploitation algorithms for data mining, pp. 246-252, (2006)
3. Garcia-Laencina P. J., Sancho-Gomez J. L., Figuerias-Vidal A. R., Pattern Classification with Missing Data: a Review, Neural Comput. & Applic. 19:263-282, (2010)
4. Kuncheva L., *Combining Pattern Classifiers: Methods And Algorithms*. A JOHN WILEY & SONS, INC., PUBLICATION (2004)
5. Kurzyński M., *Pattern recognition: statistical methods.* Oficyna Wydawnicza PWr Wrocław (1997)
6. Marques De Sa J. P., *Pattern Recognition – Concepts, Methods and Applications*, Springer, Oporto University, Portugal (2001)
7. Mitchel T.M., *Machine Learning*, McGraw-Hill Science, (1997)
8. Newcomer, E., Lomow, G., *Understanding SOA with Web Services*, Addison Wesley Professional, (2004)
9. Prusiewicz A., Zięba M., *Services recommendation in systems based on Service Oriented Architecture by applying modified ROCK algorithm*, Communications in Computer and Information Science, p. 226-238, (2010).
10. Rosen, M., Lublinsky, B., Smith, K.T., Balcer, M.J., Service-Oriented Architecture and Design Strategies, Wiley Publishing, Inc., (2008)
11. Secretan J., Georgiopoulos M., Koufakou A., Cardona K., *APHID: An architecture for private, high-performance integrated data mining*, Future Generation Computer Systems 26, pp. 891-904, (2010)
12. SOA Reference Model Technical Committee. A Reference Model for Service Oriented Architecture, OASIS, (2006)
13. Tsoumakas G., Bassiliades N., Vlahavas I., *A knowledge-based web information system for the fusion of distributed classifiers*, IDEA Group, pp. 271-308, (2004)
14. Tomczak J., Świątek J., Brzostowski K., Bayesian Classifiers with Incremental Learning for Nonstationary Datastreams, Advances in System Science 251-261, (2010)
15. WEB SERVICES OVERVIEW, http://publib.boulder.ibm.com/infocenter/rtnlhelp/v6r0m0/index.jsp?topic=/com.ibm.etools.webservice.doc/concepts/cws.html.
16. Weka, http://www.cs.waikato.ac.nz/ml/weka/
17. Witten I. H., Frank E., *Data Mining. Practical Machine Learning Tools and Techniques*, Elsevier, San Francisco, (2005)