Regression model for Quality of Web Services dataset with WEKA

Shalini Gambhir\textsuperscript{1} , Puneet Arora\textsuperscript{2} , Jatin Gambhir\textsuperscript{3}

\textsuperscript{1}Department Information Technology
\textsuperscript{2}BVCOE, New Delhi
\textsuperscript{2}Infosys, Hyderabad
\textsuperscript{3}TCS, New Delhi

\textsuperscript{1}Shalinichawla2006@gmail.com
\textsuperscript{2}arora.puneet777@gmail.com
\textsuperscript{3}er.jatin@gmail.com

Abstract- The Waikato Environment for Knowledge Analysis (WEKA) came about through the perceived need for a unified workbench that would allow researchers easy access to state-of-the-art techniques in machine learning algorithms for data mining tasks. It provides a general-purpose environment for automatic classification, regression, clustering, and feature selection etc. in various research areas. This paper provides an introduction to the WEKA workbench and briefly discusses regression model for some of the quality of web service parameters.

Keywords – WEKA, Data Preparation, Data Analysis, QWS Dataset, Linear Regression Model.

I. INTRODUCTION

More than twelve years have elapsed since the first public release of WEKA. In that time, the software has been rewritten entirely from scratch, evolved substantially and now accompanies a text on data mining [5]. Nowadays, WEKA is recognized as a landmark system in data mining and machine learning [3]. The software is written in the Java™ language and contains a GUI for interacting with data files and producing visual results. It also has a general API, to embed WEKA, like any other library, in other applications to such things as automated server-side data-mining tasks. In this paper we will discuss the first technique for data mining: regression, using WEKA which transforms existing data into a numerical prediction for future data.

Regression model\textsuperscript{1} is one of the data mining methods used to predict the result of an unknown dependent variable, given the values of the independent variables. It is the easiest to perform and the least powerful method of data mining, but it served a good purpose as an introduction to WEKA and provided a good example of how raw data can be transformed into meaningful information. The rest of the paper is organized as follows. The regression model and other modeling techniques for data mining are briefly described in section II. QWS (quality of web services) dataset considered to rank web services are explained in section III. Experimental results and interpretation are presented in section IV. Concluding remarks are given in section V.

II. LINEAR REGRESSION MODEL IN WEKA

Many different data mining, query model, processing model, and data collection techniques are available. This section covers briefly about various techniques currently being used:

A. Link Analysis: Association Rules –

It is a technique developed specifically for data mining and is used when we are provided with a dataset of customer transactions (collection of items) and we need to find Correlations between items as rules. Examples: Supermarket baskets or Attached mailing in direct marketing.
Regression model for Quality of Web Services dataset with WEKA

B. Predictive Modeling: Classifications –
The main objective of classification is to build structures from examples of past decisions that can be used to make decisions for unseen cases. It is often referred to as supervised learning. Some of the techniques used are Decision Tree Rule induction and Neural Networks.

C. Predictive Modeling: Regression –
Regression is a technique in which unknown values of a discrete variable are predicted based on known values of one or more continuous and/or discrete variables. It has many applications in trend analysis, business planning, marketing, financial forecasting, time series prediction, biomedical and drug response modeling, and environmental modeling.

D. Data Base Segmentation: Clustering –
Clustering is a process of partitioning a set of data (or objects) into a set of meaningful sub-classes, called clusters. It helps users understand the natural grouping or structure in a data set. It is often referred to as unsupervised learning and is used either as a stand-alone tool to get insight into data distribution or as a preprocessing step for other algorithms. The purpose is to find “natural” grouping of instances given un-labeled data.

The predictive modeling method being presented here for data mining is — Regression — which allows to predict a numerical value for a given set of input values and also provide a good example of transforming raw data into meaningful information.

III. WEB SERVICES AND QWS DATASET

Web services are emerging as a new paradigm for developing and deploying business processes within and across enterprises. Enterprises encapsulate their internal business processes as web services and publish them into public directories such as UDDI [2] so that other enterprises can invoke these business processes through well-defined service interface in their business processes. Generally, a business process contains a number of tasks that describes the required business functions. At runtime, appropriate services need to be chosen to perform these tasks.

With the speedy arising of web services technology in a number of research areas, it also appears to be very practical and profitable for many business-to-business (B2B) applications and enterprise computing arenas.

A Web service (also Web Service, Web service) is defined by the W3C as "a software system designed to support interoperable machine-to-machine interaction over a network"[6]. Web services are frequently just Internet application programming interfaces (API) that can be accessed over a network, such as the Internet, and executed on a remote system hosting the requested services.

The W3C Web service definition encompasses many different systems, but in common usage the term refers to clients and servers that communicate over the HTTP protocol used on the Web.

Web Services is special and different from the rest technologies as it is based on the already existing and well-known HTTP protocol, and uses XML as the base language. This makes it a very developer-friendly service system. However, most of the technologies such as RMI, COM, and CORBA involve a whole learning curve. New technologies and languages have to be learnt to implement these services.

Also, Web Services is based on a set of standardized rules and specifications, making it more portable. This was not the case with the technologies such as RMI, COM, and CORBA.
The QWS data set[4] used for regression model example will focus on parameters to rank the web services. The Web services Relevancy Function -WsRF (the dependent variable) is used to measure the quality ranking of a Web service based on many independent variables —Response Time, Availability, Throughput, Successability, Reliability, Compliance, Best Practices, Latency, Documentation etc. So, to rank Web Service Quality, we created a regression model to predict WsRF. The model is based on some comparable attributes in the training set and what WsRF value is obtained for (the model), then put the values of attributes in the test set to produce an expected WsRF value.. The attributes or the independent variables in this data set are:

Table -1 QWS Parameters and Units

<table>
<thead>
<tr>
<th>ID</th>
<th>Parameter Name</th>
<th>Description</th>
<th>Units</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Response Time</td>
<td>Time taken to send a request and receive a response</td>
<td>Ms</td>
</tr>
<tr>
<td>2</td>
<td>Availability</td>
<td>Number of successful invocations/total invocations</td>
<td>%</td>
</tr>
<tr>
<td>3</td>
<td>Throughput</td>
<td>Total Number of invocations for a given period of time</td>
<td>invokes/second</td>
</tr>
<tr>
<td>4</td>
<td>Successability</td>
<td>Number of response / number of request messages</td>
<td>%</td>
</tr>
<tr>
<td>5</td>
<td>Reliability</td>
<td>Ratio of the number of error messages to total messages</td>
<td>%</td>
</tr>
<tr>
<td>6</td>
<td>Compliance</td>
<td>The extent to which a WSDL document follows WSDL specification</td>
<td>%</td>
</tr>
<tr>
<td>7</td>
<td>Best Practices</td>
<td>The extent to which a Web service follows WS-I Basic Profile</td>
<td>%</td>
</tr>
<tr>
<td>8</td>
<td>Latency</td>
<td>Time taken for the server to process a given request</td>
<td>Ms</td>
</tr>
</tbody>
</table>
Regression model for Quality of Web Services dataset with WEKA

| 9 | Documentation | Measure of documentation (i.e. description tags) in WSDL | % |

IV. **CREATING THE REGRESSION MODEL WITH WEKA**

The QWS data set is loaded in WEKA using Attribute-Relation File Format (ARFF), method by supplying each row of dataset in a comma-delimited format and then the Linear Regression Model is built on it. The ARFF file used to create the model appears below.

![Figure 2. WEKA File Format](image)

![Figure 3. WEKA with QWS data loaded](image)
V. CONCLUSION

The Linear Regression Model obtained for QWS dataset using WEKA can be interpreted as:

Web Service Relevancy Function (WsRF) = -0.0012 * ResponseTime + 0.0892 * Availability + 0.4723 * Throughput + 0.1629 * Successability + 0.2563 * Reliability + 0.0792 * Compliance + 0.0009 * Latency + 0.1359 * Documentation + 16.7025

As Data Mining techniques are often used for identifying patterns and rules similarly here Linear Regression Model is created that detect patterns, predict output, and come up with conclusions backed by the inputted dataset. Besides just a WsRF value following pattern has also been obtained by the model:

Best Practices doesn’t matter — WEKA will only use columns that statistically contribute to the accuracy of the model (measured in R-squared, but beyond the scope of this article). It will throw out and ignore columns that don’t help in creating a good model. So this regression model is telling us that Best Practices parameter doesn’t affect the WsRF value.

REFERENCE

Regression model for Quality of Web Services dataset with WEKA


