Turnitin Originality Report

Paper by Baskar A

From Paper (CSE Papers)

Processed on 28-Jan-2015 6:51 PM
SGT
ID: 499074360
Word Count: 3218

Similarity Index	Similarity by Source	
8%	Internet Sources:	8%
	Publications:	3%
	Student Papers:	5%

sources:

1 2% match (Internet from 27-Jul-2014)
http://www.slideshare.net/dataminingtools/rapidminer-introduction-to-datamining

2 2% match (Internet from 13-Nov-2013)
http://orange.biolab.si/addons/

3 1% match (Internet from 05-May-2014)
http://www.ijser.org/researchpaper%5CA-survey-on-Data-Mining-Tools-Techniques-Applications-Trends-and-Issues.pdf

4 1% match (Internet from 19-Dec-2014)
http://ijctonline.com/ojs/index.php/ijct/article/view/991

5 1% match (Internet from 12-Jan-2015)
http://en.wikipedia.org/wiki/Orange_(software)

6 1% match (publications)
Tiago Mota. "Identificação e Quantificação de Células Oncocíticas em Imagens Microscópicas", Repositório Aberto da Universidade do Porto, 2014.

7 1% match (student papers from 30-Jan-2012)
Submitted to MCC Training Institute on 2012-01-30

paper text:

A STUDY OF OPEN SOURCE DATAMINING TOOLS AND ITS APPLICATIONS Subathra P, Deepika R , YaminiK, Arunprasad P, Shriram K Vasudevan Department of Computer Science and Engineering, Amrita School of Engineering, Amrita Vishwa Vidyapeetham (University), Coimbatore, India. ---------- ---------- ABSTRACT: Data Miningis a technology that is used for the process of analyzing and summarizing useful information from different perspectives of data.The importance of choosing data mining software tools for the developing applications using miningalgorithms has led to the analysis of the commercially available open source data mining tools. . The paper discusses about the following likeness of the data mining tools - KNIME, WEKA, ORANGE, R Tool and RAPID MINER. The Historical Development and state -of- art ? The Applications supported. ? Data mining algorithms are supported by each tool. ? The pre-requisites and the procedures

to install a tool. ? The input file format supported by each tool. To extract useful information from these dataeffectively and efficiently, data mining tools are used. The availability of many open source data mining tools, there is an increasing challenge in deciding upon the apace-updated tools for a given application.This paper has provided abrief study about theopen source knowledge discovery tools with their installation process, algorithms support and input file formats support. Key Words:Data Mining, KNIME, R-Tool, Rapid Miner, ORANGE, and WEKA. ---------- 1. INTRODUCTION Now-a-days database and data repository storage size ranges from tera-bytes to zetta- bytes within which many useful strategic informationare hidden which provide immenseknowledge in the process of decision making. Data Mining plays an active part in discovering strategic patterns which is used to find relationship among the data using various data analytics tools. Therefore Data mining can be used in various areas such as Biological Data Analysis, Financial data analysis, Fraud detection, Telecommunication Industry, Retail Industry, Prediction of future Stock exchange values etc. This paper addresses the five data mining tools - KNIME, R-Tool, WEKA, ORANGE and RAPIDMINER with their installation procedures, supported algorithms and supported input file formats. Materials and Methods 2. HISTORICAL DEVELOPMENT AND STATE-OF-ART 2.1 KNIME Konstanz Information Miner (KNIME) is one of the open source data mining tool. This acme software was developed at the University of Konstanz headed by Michael Berthold from Silicon Valley in January 2014. FEATURES KNIME is purely based on the nodal work which incorporates more than 100 nodes for analyzing the data. The modular approach of KNIME helps in documenting and storing the order in which the analysis processes are conceived and implemented ensuring that the intermediate results are always available to the user. KNIME achieves scalability through sophisticated data handling (catching the data automatically in the background while increasing the throughput performance).We can also import/export workflows in order to exchange the work modules with other KNIME users. It can also incorporate WEKA analysis modules and R scripts through additional plugins.It supportsParallel execution on multi-core systems and Command line version. VERSIONS The entry open source version of KNIME is the KNIME Desktop which provides limited features. Paid versions of KNIME is available with more of additional features and with great supporting factors. It is based on the Eclipse IDE which makes it as a development platform similar to data mining platform. The following table shows the available versions of KNIME tool and its supported platforms Table 1.Knime available versions and platforms 2.2 R TOOL R is best known as a statistical tool for analyzing data. It is a simple programming language based on S programming language .It was initially created by Ross Ilaka and Robert gentleman and was recently developed by R development core team at the University of Auckland, New Zealand. FEARTURES R is an interpreted language (effective programming language) that is,in most computer languages like C, FORTRAN, and Pascal requires built option in executing a command but R differs from them by executing commands directly. R provides an easy programming environment for its users and includes options for data summary, data exploration, graphical presentation and data modelling. It can also incorporate other R functions that can often be incredibly useful for analyzing and visualizing data.The facilities like data manipulation, calculation, graphical facilities and displayare provided by this software. It contains some intermediate tasks flexibility to the users. Table 2.R Tool available versions and platforms 2.3 WEKA Waikato Environment for knowledge analysis (WEKA) is popular machine learning toolkit developed at the University of Waikato, New Zealand and is mainly use in academia. FEATURES Waikato Environment for Knowledge Analysis is developed at the University of Waikato, New Zealand. It provides toolkits for machine learning. WEKA provides a comprehensive collection of algorithms. WEKA provides itself with the GUI, so algorithms can be easily used to the dataset from GUI directly or it can be called from the Java code. It also supports working in the command prompt. Tasks like preprocessing,feature selection, clustering, can be done using WEKA.It is written in Java so it is platform independent and can run in almost any platform.It also supportsvisualization tasks and many machine learning applications. WEKA is

freeware available under the general public license agreement (GNU). PREREQUIREMENTS The following table shows Java version necessary to run a specific WEKA version Table 3. Weka version and the supported Java version NOTE: A workaround for the problem of Look 'n' Feel due to the combination of Java 5.0 and later versions with the Gnome/Linux, was introduced with version 3.4.5/3.5.0. From WEKA m 3.6.5/3.7.4 Mac OS X users will need Java packageversion updated. VERSIONS WEKA 3.6 is the latest stable version of WEKA. Table 4.Weka available platform NOTE: If you already have Java 1.6 (or later) on your system then go for the version without java VM. 2.4 ORANGE: ORANGE [10] is a component-based suite which provides a programming front-end for visualization and analysis of data. Its user interface builds upon Qt framework which is a cross-platform. Orange is distributed free under theGPL. It uses python libraries to do scripting works.Shortest script for doing training, cross validation.The easiest tool to learn.It is implemented in C++ and Python. The invention of this software was done at University of Ljubljana, Slovenia. FEATURES Tasks like data preprocessing, feature scoring and filtering are done using this software.

5**Unlike its** competitor **scikit-learn and mlpy, Orange does not tie into NumPy and its ecosystem of tools; it focuses on traditional, symbolic algorithms, more than numeric ones.**

It also supports model evaluation, and exploration techniques, algorithms comparison and prediction.Works both as a script and interface. VERSIONS: Table 5.Orange versions and platforms available 2.5 RAPID MINER Rapid Miner, formally known as YALE [8](Yet another Learning Environment)was developed by Ralf Klinkenberg, Ingo Mierswa Simon Fischer in 2001 in the technical University of Dort mung is freeware suite. It is one of the comprehensive, flexible and the most widespread-used due to the combination of its leading-edge technologies and its functional range. . Rapid Miner can be used as an intuitive graphical user interface (GUI) or as a command line version for extracting patterns. Like other software processes like Visualization, optimization, modelling, construction is supported.

1**Software versions: Community edition (open – source) Enterprise edition (Community Edition + More Features + Services + Guarantees).**

Table 6. Rapid miner versions and available platforms specify which model is more accurate for different modelling techniques so that it can be validated to ensure the accuracy rate. The output of the workflow is in the PMML format so that it can be used for different applications. Manufacturing: Energy usage prediction: This KNIME workflow is used to predict the future values of the timeseries with respect to the past values. This workflow helps in removing the seasonality from the time series and also trains an auto-regressive model for time series prediction. 3.APPLICATIONS SUPPORTED Pharma: Virtual High Throughput Screening: KNIME is designed to handle huge data sets so the results of predictive analysis of the untested data can be easily visualized with the help of the enrichment plotter. 3.1 KNIME Customer Intelligence: KNIME [1] is used in the Telecommunication industry for the identification of classes oftelecommunication customers with the help of K-Means clustering methods. It allows the users to give the different number of clusters for the K- Means to calculate since the number of clusters is unknown beforehand. Chemical library Enumeration: This KNIME workflow helps the chemists to create a virtual library of amides based on the amines and acids. Further some of the molecular properties of the

enumerated products are calculated and filtered based upon the Lipinski “rule of 5”.This workflow uses RDKit, CDK and indigo integration nodes to demonstrate interoperability. Finance: Credit scoring –KNIME workflow deals with the creation of credit scoring model which based on the historical data that uses Decision tree, neural networks and svm at the same time to Retail: Social Media Music Recommendation: We can use KNIME’s predictive analysis algorithm to recommend music preferences for the in the social media. It also helps in transforming the data to make it more suitable for applying association and advanced association algorithms to list the artists and recommendations of interest. Cross Industry: Address duplication-Typos are common mistakes of any user, but restaurant names, postal services and many other services depends on the correct string pattern. Therefore string matching is an important part of data analytics. KNIME uses string matching algorithm to get rid of typo errors and finds out the closest string to what the user wrote. By taking the string distance between our current misspelled data and the reference data set. Government: Network Traffic Reporting: this workflow of KNIME is used for generating a report of the connection statistics of the IP addresses of the hosts in a network .Dataset is obtained by capturing a package of data using WIRESHOCK and it is converted in CSV file format. The CSV file is used KNIME for further analysis to generate the report on connection statistics. This workflow of KNIME is available under its Server package. Cell Miner: KNIME is used for analyzing cell images. A new data cell is integrated for images and picture file reader node is added to the repository. A segmenter node is also used to identify the nodes in the images. Multiple feature extraction nodes are also used to extract the data in order to input it to the classifier algorithm. 3.2 R tool: Data Exploration and visualization R includes all kinds of data manipulation, statistical model which is needed for Data Analysis technique [5]. It supports Unique Data Visualization for representing the complex data into charts and graphs.R provides flexibility to mix-and-match the models which provides best performance. Outlier Detection The Extreme values package of R provides outlier detection and plot functions for univariate data.The parameters for a model distribution are estimated using regression of the sorted data obtained from the subset of data on their QQ-plot positions. Text mining The tm package of R provides the text mining framework which has methods for data import, corpus handling, preprocessing, metadata management, and creation of term-document matrices. R is also used for Association rules mining, Social network analysis, Multimedia scaling, Parallel computing etc. 3.3 WEKA Machine learning WEKA is used in solving a variety of real-world problems like agricultural and horticultural domains. Data visualization - The 3D visualization perspective is a plugin Spoon perspective for WEKA version 4.0 and above provides a Java based 3D scatter plot visualization and a histogram matrix. Time series and analysis- WEKA has a dedicated time series analysis framework that allows forecasting models to be developed, evaluated and visualized. . This environment helps in modeling time series by transforming the data into a form that can be processed by the standard propositional learning algorithms. WEKA is also used in the field of Text mining and Fraud detection 3.4 ORANGE: Bioinformatics Orange Bioinformatics add-on provides a way for analytics that is useful in

2**gene selection, quality control, scoring distances between experiments**

.It also provides public access to PIPAx database ,GEO data sets, Biomart, GO, Atlas, ArrayExpress, , KEGG. These

2**features can be combined with powerful visualization, network exploration and data mining techniques from the Orange data mining framework.[** 10]

Model Maps **Orange**

Model Maps add-on

2**extends Orange by providing modules for building model maps** and **also provides widgets for Orange Canvas**

that enables exploration of model maps by the users. Network analysis Orange Network add-on that the users with provides network visualization and analysis tools. Text Mining Orange Text Mining add-on

2**extends Orange by providing common functionality for basic** text mining **tasks**

that makes ease for the users. 3.5 RAPID MINER:

1**Different levels of analysis that are available: Artificial neural networks – Non-linear predictive models that resemble biological neural networks in structure. Genetic algorithms Optimization techniques that use processes such as genetic combination, mutation, and natural selection in a design based on the concepts of natural evolution.**

Other applications: ?

1**Automatic abstraction ? Financial forecasting ? Targeted marketing ? Medical diagnosis ? Credit card fraud detection ? Weather forecasting**

4. INPUT DATA FORMATS The following table shows the input data formats that are accepted by each tool under study. Table 7. Input File Format supported by Knime,Rtool,Weka ,Orange and Rapid Miner. ? ? ? ? ? ? ? Java Sript Option Notation (.json) Extensible Markup language (.xml) Attribute Relation file format (.arff) Comma separated values (.csv) Atom Binary file format (.bin) C4.5 file format (.c45) Tab-separated values file Format (.TSV) ? .xrff – extended attribute relation file format 3. INSTALLATION STEPS 5.1 KNIME Instructions: Download the package according to your Windows version and system type. Step1:Download the KNIME Software Package from http://www.knime.Org/ downlaods /knime/ Step 2:Run the setup.exe file and follow the below instructions. Click next to view the license agreement. Step 3: Carefully read the license agreement and accede the agreement as above. Step 4:Give the installation path location for KNIME Analytics Platform Step 5:Create the path location for shortcut files of KNIME Analytics Platform and click next Step 7:Click install to proceed. Step 6:Click the check box icon in order to create a desktop icon and click next. Wait until the installation process is over .It may take 4-6 minutes to get installed. Step 2:Click next and proceed. Now the KNIME Analytics Package will be successfully installed. 5.2 R TOOL Step 1:Select the language to use during the installation and click Ok button. Step 3:Read the terms and conditions to proceed Step 4:Select the destination location to save the installation. Step 5:Select the appropriate system file and proceed further. Step 8:Choose the available options to

customize else Check all the check boxes for default settings And click next to proceed. Step 6:Set the startup configuration if required. Step 7:Select the path location to place the program shortcuts. Step 9:Click finish. Package will successfully installed Step 4: Select the installation path and click next button tocontinue 5.3 WEKA Step1:Run setup.exe . Step 5:Give a name for the shortcut and specify the folder in which the shortcut should appear. Click next button to continue. Step 2:Read the instructions carefully and accede it. Step 6:Install java if it is not already installed in your system.Click install button and continue. Step 3:Check all the packages

> 7**and click next button** to continue. **Step** 9 **:Check** Start WEKA **and click** on finish **button.**

Step 7:Click close button and proceed. Now the WEKAwill be successfully installed. Step 8:Click next and continue. 5.4 Orange Step 1:Run setup.exe. Read the instructions carefully and accede it. Step 3:Toinstall python Click yes button and continue. Step 2:Click yes button and continue. Step 4:Click next button and continue. Step 5:Click next button and continue. Step 7:PyQt setup wizard appears

> 7.**click next button** and **continue. Step 6:Click next button** and **continue. Step**

8:Read the liscense agreement carefully and click on ‘I Agree’ button to proceed. Step 9:Select the Desired components to be installed. For default settings select check all Step 11:Click finish button the check boxes. Step 10:Select the installation path and click next button to continue. The PyQt package will be successfully Step 12:PyQwt setup wizard appears .click next button and continue Step 13:Read the liscense agreement carefully and click on ‘I Agree’ button to proceed. Step 15:Select the installation path and click next button to continue. Step 14:Select the Desired components to be installed. For default settings select check all the check boxes. Step 16:Click finish button The PyQwt package will be successfully Step 17:Click next button and continue. Step 18:Click finish Button. 5.5 RAPID MINER Step 1:Run Setup.exe file to start installation Step 2: Read the instructions carefully and accede it Step 3: Choose the installation path for the package Orange tool will be successfully installed in your system. Step 4:Installation complete Step 5:Click FINISH button to complete installation. 4. ALGORITHM SUPPORT CLUSTERING ALGORITHMS Clustering is a process of partitioning a set of abstract data/objects into classes of similar objects by assigning the labels to the groups. These sub-classes are known a clusters .the following table shows the various clustering algorithms supported by the data mining tools Table. 8 Classification algorithms supported by Knime , R tool, Weka, Orange and Rapid Miner CLASSIFICATION ALGORITHM In finding and retrieving data in effective and efficient manner, Data classification plays a vital role. It organizes data into categories. The following table shows the classification algorithms that are supported by datamining tools. Table. 9Clustering algorithms supported by Knime , R tool, Weka, Orange and Rapid Miner Results and Conclusion This study has given a brief introduction on the five different data mining tools and their applications –KNIME, ORANGE, R Tool, WEKA and RAPID MINER. Each tool has its own pros and cons. Be that as it may, KNIME, ORANGE, R Tool, WEKA and RAPID MINER have most of the desired characteristics and functions for a fully- functional Data Mining platform and thereby these tools can be used for most of the Data Mining tasks. As a future work, we are going to study and compare the performance of various data mining classification and clustering algorithms for various data mining tools. [5] http://www.revolutionanalytics.com/what-r [6] Paradis, Emmanuel. "R for Beginners." (2002). [7] http://www.cs.waikato.ac.nz/ml/weka/- Main Reference. [8]http://it.toolbox.com/wiki/index.php/Rapid Miner

[9]Wahbeh,

6**Abdullah H., et al. "A comparison study between data mining tools over some classification methods." International Journal of** Advanced **Computer Science and Applications, Special Issue (2011):**

18-26. [10] http://orange.biolab.si/ [11]http://www.cs.waikato.ac.nz/ml/weka/requ irements.html-referenece [12

3**]Y. Ramamohan, K.Vasantharao, C.** KalyanaChakravarti, **A.S.K.Ratnam. A Study of Data Mining Tools in Knowledge Discovery Process, International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2, Issue-3, July 2012** [13] The WEKA **data mining**

4**software: An update, Mark Hall, Eibe Frank, G. Holmes, B. P** fahringer, P. **Reutemann , IH Witten, ACM SIGKDD Explorations, Newsletter, Pages 10- 18, volume 11 issue 1, june 2009-**

for applications. References [1] https://www.knime.org/ [2] https://tech.knime.org/ [3] Spector, Phil. "Introduction to R." University of California, Berkeley (2004). [4] http://www.RDataMining.com