Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 3, Issue. 3, March 2014, pg.445 – 452

RESEARCH ARTICLE

Categorization of Data Mining Tools Based on Their Types

J.Mary Dallfin Bruxella[1], S.Sadhana[2], S.Geetha[3]

[1]Asst.prof.Department of Computer Science, KSR College of arts & science, Tiruchengode
[2, 3] M.Phil, Department of Computer Science, Sri Vijay Vidyalaya College of Arts & Science, Dharmapuri
[1] dallfin.j@gmail.com; [2] sadhanasaravanan@gmail.com; [3]geethapmp@yahoo.com

Abstract— ***Data mining the extraction of hidden predictive information from large databases is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. The development and application of data mining algorithms requires the use of powerful software tools. As the number of available tools continues to grow, the choice of the most suitable tool becomes increasingly difficult. This paper attempts to support the decision-making process by discussing the historical development and presenting a range of existing state-of-the-art data mining and related tools. This paper is organized as follows: the first section Introduction about Data Mining. The criteria to compare data mining software are explained in the second section Review of Data Mining Tools. The last section Categorization of Data Mining Software into Different Types proposes a categorization of data mining software and introduces typical software tools for the different types.***

Keywords— ***Data mining tools; Rapid miner; Decision making; MATLAB; WEKA; SPSS***

I. INTRODUCTION

Data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use.

The most important tasks in data mining are

- Supervised learning, with a known output variable in the dataset, including
    a. **Classification** is learning a function that maps (classifies) a data item into one of several predefined classes.
    b. **Fuzzy classification** with gradual memberships with values in between 0 and 1 applied to the different classes;
    c. **Regression** is learning a function which maps a data item to a real-valued prediction variable.
- Unsupervised learning, without a known output variable in the dataset, including

a. **Clustering** is a common descriptive task where one seeks to identify a finite set of categories or clusters to describe the data.
b. **Summarization** involves methods for finding a compact description for a subset of data.
c. **Association learning** finds typical groups of items that occur frequently together in examples;

- Semi-supervised learning, whereby the output variable is known only for some examples.
- Each of these tasks consists of a chain of low level tasks. Furthermore, some low-level tasks can act as stand-alone tasks; for example, by identifying in a large dataset elements that possess a high similarity to a given example. Examples of such low-level tasks are:
- Data cleaning (e.g., outlier detection);
- Data filtering (e.g., smoothing of time series);
- Feature extraction from time series, images, videos, and graphs (e.g., consisting of segmentation and segment description for images, characteristic values such as community structures in graphs);
- Feature transformation (e.g., mathematical operations, including logarithms, dimension reduction by linear or nonlinear combinations by a principal component analysis, factor analysis or independent component analysis);
- Feature evaluation and selection (e.g., by filter or wrapper methods);
- Computation of similarities and detection of the most similar elements in terms of examples or features (e.g., by *k*-nearest-neighbor methods and correlation analysis);
- Model validation (cross validation, boot strapping, statistical relevance tests and complexity measures);
- Model fusion (mixture of experts); and
- Model optimization (e.g., by evolutionary algorithms).

For almost all of these tasks, a large variety of classical statistical methods—including classifiers using estimated probability density functions, factor analysis and others, and newer machine learning methods—such as artificial neural networks, fuzzy models, rough sets, support vector machines, decision trees, and random forests, are available. In addition, optimization models such as evolutionary algorithms can assist with the identification of model structures and parameters.

II. REVIEW OF DATA MINING TOOLS

Data mining tools can purchase mining programs designed for existing software and hardware platforms, which can be integrated into new products and systems as they are brought online, or they can build their own custom mining solution. Data mining has a long history, with strong roots in statistics, artificial intelligence, machine learning, and database research.1, 2 The word ‘data mining’ can be found relatively early, as in the article of Lovell,3 published in the 1980s. Advancements in this field were accompanied by development of related software tools, starting with mainframe programs for statistical analysis in the early 1950s, and leading to a large variety of standalone, client/server, and web based software as today’s service solution.

Following the original definition given in Ref 1,data mining is a step in the knowledge discovery from databases (KDD) process that consists of applying data analysis and discovery algorithms to produce a particular enumeration of patterns (or models) across the data. In that same article, KDD is defined as the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data. Sometimes, the wider KDD definition is used synonymously for data mining. This wider interpretation is especially popular in the context of software tools because most such tools sup-port the complete KDD process and not just a single step.

Another group of methods stemmed from artificial intelligence- like decision trees, rule-based systems, and others. The term ‘machine learning’ includes methods such as support vector machines and artificial neural networks. There are several different and sometimes overlapping categorizations; for example, fuzzy logic, artificial neural networks, and evolutionary algorithms, which are summarized as computational intelligence. The typical life cycle of new data mining methods begins with theoretical papers based on in house software prototypes, followed by public or on-demand software distribution of successful algorithms as research prototypes. Then, either special commercial or open source packages containing a family of similar algorithms are developed or the algorithms are integrated into existing open source or commercial packages.

Many companies have tried to promote their own standalone packages, but only few have reached notable market shares. The life cycle of some data mining tools is remarkably short. Typical reasons include internal marketing decisions and acquisitions of specialized companies by larger ones, leading to a renaming and integration of product lines. The largest commercial success stories resulted from the step-wise integration of data mining methods into established commercial statistical tools. Companies such as SPSS, founded in 1975 with precursors from 1968, or SAS, founded in 1976, have been offering statistical tools for mainframe computers since the 1970s.

These tools were later adapted to personal computers and client/server solutions for larger customers. With the increasing popularity of data mining, algorithms such as artificial neural networks or decision trees were integrated in the main

products and specialized data mining companies such as Integrated Solutions Ltd. (acquired in 1998 by SPSS) were acquired to obtain access to data mining tools such as Clementine. During these periods, renaming of tools and company mergers played an important role in history; for example, the tool Clementine (SPSS) was renamed as PASW Modeler, and is now available as IBM SPSS Modeler after the acquisition of SPSS by IBM in 2009. In general, tools of this statistical branch are now very popular for the user groups in business application and applied research.

Concurrently, many companies offering business intelligence products have integrated data mining solutions into their database products; one example is Oracle Data Mining (established in 2002). Many of these products are also a product of the acquisition and integration of specialized data mining companies. In 2008, the worldwide market for business intelligence(i.e., software and maintenance fees) was7.8 billion USD, including 1.5 billion USD in so called 'advanced analytics', containing data mining and statistics.7 This sector has grown 12.1% between2007 and 2008, with large players including companies such as SAS (33.2%, tool: SAS Enterprise Miner), SPSS (14.3%, since 2009, an IBM company; tool: IBMSPSS Modeler), Microsoft (1.7%, tool: SQL Server Analysis Services), Tera data (1.5%, tool: Tera data Database, former name Tera Miner), and TIBCO(1.4%, tool: TIBCO Spot fire).

Open-source libraries have also become very popular since the 1990s. The most prominent example is Waikato Environment for Knowledge Analysis(WEKA), see Ref 8. WEKA started in 1994 as a C++ library, with its first public release in 1996.

In 1999, it was completely rebuilt as a JAVA package; since that time, it has been regularly updated. In addition, WEKA components have been integrated in many other open-source tools such as Pentaho, RapidMiner, and KNIME.A large group of research prototypes are based on script-oriented mathematical programs such as MATLAB (commercial) and R (open source). Such mathematical programs were not originally focused on data mining, but contain many useful mathematical and visualization functions that support the implementation of data mining algorithms. Recently, graphical user interfaces such as those utilized for R(e.g., Rattle) and Matlab e.g., Gait-CAD, Established in 2006) can be used as integration packages (INT)for many single, open-source algorithms.

In the past 10–15 years, data mining has become a technology in its own right, is well established also in business intelligence (BI), and continues to exhibit steadily increasing importance in technology and life sciences sectors. For example, data mining was a key factor supporting methodological breakthroughs in genetics.16 It is a promising technology for future fields such as text mining and semantic searchengines,17 learning in autonomous systems—as with humanoid robots18 and cars, chemoinformatics19 and others .Various standardization initiatives have been introduced for data mining processes, data and model interfaces—as with Cross Industry Standard Process for Data Mining for industrial data mining, and approaches focused on clinical and biological applications. New methods, especially for data streams,23 extremely large datasets, graphmining,24, 25 text mining,17 and others have been proposed in the last few years. In the near future, methods for high-dimensional problems such as imageretrieval26 and video mining27 will also be optimized and embedded into powerful tools.

Different types of data mining tools are available in the marketplace, each with their own strengths and weaknesses.

III. CATEGORIZATION OF DATA MINING SOFTWARE INTO DIFFERENT TYPES

Following the criteria from the previous section, different types of similar data mining tools can be found. The typical characteristics of these types are explained in this section.
In this paper, the following nine types are proposed:

A. Data mining suites (DMS)

DMS focus largely on data mining and include numerous methods. They support feature tables and time series, while additional tools for text mining are sometime available. The application focus is wide and not restricted to a special application field, such as business applications; however, coupling to business solutions, import and export of models, reporting, and a variety of different platforms are nonetheless supported. In addition, the producers provide services for adaptation of the tools to the workflows and data structures of the customer. DMS is mostly commercial and rather expensive, but some open-source tools such as Rapid Miner exist. Typical examples include IBM SPSS Modeler, SAS Enterprise Miner, Alice Isoft, DataEngine, DataDetective, GhostMiner, Knowledge Studio, KXEN, NAG Data Mining Components, Partek Discovery Suite, STATISTICA, and TIBCO Spot fire.

B. Business intelligence packages (BIs)

BIs have no special focus to data mining, but include basic data mining functionality, especially for statistical methods in business applications. BIs are often restricted to feature tables and time series, large feature tables are supported. They have a highly developed reporting functionality and good support for education, handling, and adaptation to the workflows of the customer. They are characterized by a strong focus on database coupling, and are implemented via client/server architecture. Most BI software's are commercial (IBM Cognos 8 BI,Oracle DataMining, SAPNetweaver Business Warehouse, Teradata Database, DB2 Data Warehouse from IBM, and PolyVista), but a few open-source solutions exist (Pentaho).

C. Mathematical packages

MATs have no special focus on data mining, but provide a large and extendable set of algorithms and visualization routines. They support feature tables, time series, and have at least import formats for images. The user interaction often requires programming skills in a scripting language. MATs are attractive to users in algorithm development and applied research because data mining algorithms can be rapidly implemented, mostly in the form of extensions (EXT) and research prototypes (RES). MAT packages exist as commercial (MATLAB and R-PLUS) or open-source tools (R, Kepler). In principle, table calculation software such as Excel may also be categorized here, but it is not included in this paper. Most tools are available for different platforms but have weaknesses in database coupling.

D. Integration packages

INTs are extendable bundles of many different open-source algorithms, either as stand-alone software (mostly based on Java; as KNIME, the GUI-version of WEKA, KEEL, and TANAGRA) or as a kind of larger extension package for tools from the MAT type (such as Gait-CAD, PRTools for MATLAB, and RWEKA for R). Import and export support standard formats, but database support is quite weak. Most tools are available for different platforms and include a GUI. Mixtures of license models occur if open-source integration packages are based on commercial tools from the MAT type. With these characteristics, the tools are attractive to algorithm developers and users in applied research due to expandability and rapid comparison with alternative tools, and due to easy integration of application-specific methods and import options.

E. EXT

EXT are smaller add-ons for other tools such as Excel,Matlab, R, and so forth, with limited but quite useful functionality. Here, only a few data mining algorithms are implemented such as artificial neural networks for Excel (Forecaster XL and XLMiner) or MATLAB (Matlab Neural Networks Toolbox). There are commercial or open-source versions, but licenses for the basic tools must also be available. The user interaction is the same as for the basic tool, for example, by using a programming language (MATLAB) or by embedding the extension in the menu (Excel).

F. Data mining libraries

LIBs implement data mining methods as a bundle of functions. These functions can be embedded in other software tools using an Application Programming Interface (API) for the interaction between the software tool and the data mining functions. A graphical user interface is missing, but some functions can support the integration of specific visualization tools. They are often written in JAVA or C++ and the solutions are platform independent. Open source examples are WEKA (Java-based), MLC++ (C++ based), JAVA Data Mining Package, and LibSVM (C++ and JAVAbased) for support vector machines. A commercial example is Neurofusion for C++, whereas XELOPES (Java, C++, and C_) uses different license models. LIB tools are mainly attractive to users in algorithm development and applied research, for embedding data mining software into larger data mining software tools or specific solutions for narrow applications.

G. Specialties

SPECs are similar to DMS tools, but implement only one special family of methods such as artificial neural networks. They contain many elaborate visualization techniques for such methods. SPECs are rather simple to handle as compared with other tools, which eases the use of such tools in education. Examples are CART for decision trees, Bayesia Lab for Bayesian

networks, C5.0, WizRule, RuleDiscovery System for rule-based systems, MagnumOpus for association analysis, and JavaNNS, Neuroshell, NeuralWorks Predict, RapAnalyst for artificial neural networks.

H. RES

RES are usually the first—and not always stable—implementations of new and innovative algorithms. They contain only one or a few algorithms with restricted graphical support and without automation support. Import and export functionality is rather restricted and database coupling is missing or weak. RES tools are mostly open source. They are mainly attractive to users in algorithm development and applied research, specifically in very innovative fields. Examples are GIFT for content-based image retrieval, Himalaya for mining maximal frequent item sets, sequential pattern mining and scalable linear regression trees, Rseslibs for rough sets, and Pegasus for graph mining. Early versions of today's popular tools such as WEKA and RapidMiner started in this category and shifted later to other categories as DMS.

I. Solutions

SOLs describe a group of tools that are customized to narrow application fields such as text mining (GATE), image processing (ITK, ImageJ), drug discovery (Molegro Data Modeler), image analysis in microscopy (CellProfilerAnalyst), or mining gene expression profiles (Partek Genomics Suite, MEGA). The advantage of these solutions is the excellent support of domain specific feature extraction techniques, evaluation measures, visualizations, and import formats. The level of data mining methods ranges from rather weak support (particularly in image processing) to highly developed algorithms. In some cases, more general tools from types DMS or INT also support specific domains (KNIME, Gait-CAD for peptide chemo informatics). There are many commercial and open-source solutions. A large variety of tools actually requires a fuzzy categorization with gradual memberships to different types. Examples are tools including a set of different algorithms (LIB) with an additional GUI acting as an INT, DMS, including special methods for narrow application fields and others. In these cases, a main type was assigned and the other fuzzy memberships are discussed in the Excel table in the additional material section.

IV. SUMMARY OF DATA MINING TOOLS AND THEIR TYPES

Tool name	Type	Remarks
11 Ants	DMS	family of data mining tools with a focus on business applications
Alice (d'Isoft)	DMS	focused on decision trees, but other methods available
Bayesia Lab	SPEC	Bayesian networks
CART	SPEC	specializing in decision tree, non-parametric regression, and logistic regression methods
Coheris SPAD Data Mining	DMS	company provides also solutions for text mining, former company SPAD
Data Applied	DMS	web service for Data Analysis, SAAS
DataDetective	DMS	with tools for fuzzy matching, applications on CRM, crime analysis, fraud detection
DB2 Data Warehouse	BI	component of DB2 OLAP Server, implements an opportunity discovery algorithm for multidimensional data, part of InfoSphere Warehouse (?)
DeltaMaster	BI	enhanced reporting
Forecaster XL	EXT	providing neural networks based software for forecasting, e.g. forecasting Excel add-in, based on neural networks
IBM Cognos 8 BI	BI	180 days trial version
IBM SPSS Modeler	DMS	former Clementine, now in cooperation with IBM, Predictive Analytics Software (PASW), SPSS is an IBM company since 2009
IBM SPSS Statistics	MAT	originally oriented on statistical questions, but with many methods for data mining
InfoSphere Warehouse	BI	complete business solution incl. data mining
JMP	DMS	free trial, additional special tools for genomics
KnowledgeMiner	SPEC	includes ANNs, fuzzy rule induction and fuzzy clustering, Excel support
KnowledgeStudio	DMS	PMML support and code generation
MARS	SPEC	specialized on regression splines

MATLAB	MAT	one of the world largest tools for mathematical computations, large variety of commercial and open source toolboxes, script-based, but powerful functions for visualization
MATLAB Neural Network Toolbox	EXT	extension toolbox for MATLAB
Model Builder	DMS	company's former name Fair Isaac Corporation
Molegro Data Modeler	SOL	data mining, some chemistry specials, integration into Knime possible
NAG Data Mining Components	LIB	components in C++
Neurofusion	LIB	is a general-purpose ANN C++ library that can be used to create, train and apply constructive neural networks for solving both regression and classification problems
Neuroshell	SPEC	restricted to ANN
Oracle Data Mining (ODM)	DMS	provides GUI, PL/SQL-interface, and Java-interface to Attribute Importance, Bayes Classification, Association Rules, Clustering, SVM
Partek Discovery Suite	DMS	additional special solutions for genomics, free demos
Partek Genomics Suite	SOL	for microarrays etc.
PolyAnalyst	DMS	from Goebel99, support for text mining
PolyVista	BI	text mining support, in addition tools for different solutions, e.g. quality analytics etc.
Predixion Enterprise Insight	DMS	data mining suite with a focus to standard worksflows, big data support, cloud options, OEM options possible
Random Forests	SPEC	for random forests
R-PLUS	MAT	commercial interface for R
Salford Predictive Modeling Suite (SPM)	DMS	includes former separate tools CART, MARS, TreeNet, Random Forests
SAP Netweaver Business Warehouse	BI	data warehouse with analytical capabilities, decision trees, clustering, association analysis, scoring, classification
SAS	MAT	statistic suite that is also useful for data mining
SAS Enterprise Miner	DMS	one of the world's leading tools, enterprise oriented
SQL Server Analysis Services	DMS	special coupling to SAP software
Stata	DMS	actually coming from statistics, many methods included
STATISTICA	DMS	additional tools for text mining
Teradata Database	BI	data warehouse solution with data mining components
Think Enterprise Data Miner (EDM)	DMS	massively scalable, embeddable, Java-based real-time data-mining platform, former name K.wiz
TIBCO Spotfire Miner	DMS	coupling to S-Plus, R
TreeNet	SPEC	specializing in decision trees

Table: 1 Commercial Tools

Tool name	**Type**	**Remarks**
Apache Drill	SPEC	open source framework for the interactive analysis of large data sets, includes Google'S Dremel
CellProfilerAnalyst	SOL	find rules from cell images, etc.
Fast Artificial Neural Network Library	LIB	multilayer artificial neural networks in C
Fiji	SOL	JAVA software for image processing and analysis, related to ImageJ with a variety of organized libraries, plugins for simple analysis tasks, far away from data mining, focus on bioinformatics
Frida	INT	JAVA package with different algorithms
GATE	SOL	text mining

GeNIe	SPEC	specialized to Bayesian Networks, graphical interface to SMILE
GIFT	RES	open-source package for Content-based image retrieval (CBIR) using the Query-by-Example (QBE) paradigm.
Himalaya	RES	focused on Frequent Item Sets and regression trees, subtools MAFIA, SPAM, SECRET
ImageJ	SOL	JAVA software for image processing and analysis, plugins for simple analysis tasks, far away from data mining
ITK	SOL	library for image processing and analysis, first elements of a support for simple data mining tasks like classification, additional tool VTK to build GUIs
JAVA Data Mining Package	LIB	JAVA based, alpha version, no update since 2009
KEEL	INT	many algorithms, EA focus
Kepler	MAT	workflow-oriented, R support, JAVA based, BSD license
KH Coder	SPEC	text mining, specialized also to Japanese, Spanish, Portuguese, Italian and German language
KNIME	INT	many algorithms, image processing, text mining, WEKA and R integration, special solutions for bioinformatics, see Chen07
LibSVM	LIB	for support vector classification and regression, C++, JAVA-based
MLC++	LIB	C++ library for supervised learning, included in SGI's MineSet
Mlpy	SPEC	Python framework for Machine Learning
MOA	SPEC	MOA stands for massive online analysis, focus on real-time analytics of data streams, JAVA based
Neuroph	SPEC	JAVA framework for Neural Networks
OpenNN	LIB	open ANN library, multilayer perceptron neural network in the C++, former name Flood
OpenPR	LIB	library for image processing, pattern reognition, computer vision and natural language processing, based on C++, Scilab support
Orange	LIB	Python scripts, extensions for text mining and bioinformatics, see Chen07, Alcala09
Pentaho	BI	open-source community edition, licensed full version with additional support, includes WEKA functionality
PRTools	EXT	collection of MATLAB functions for data mining
PSPP	MAT	program for statistical analysis with a synthax compatible to IBM SPSS
R	MAT	complete statistical suite, script-based, GNU-GPL
RapidMiner	DMS	formerly YALE, more than 1000 algorithms and operators for data mining, text mining, web mining, time series analysis and forecasting, audio mining, image mining, predictive analytics, ETL, reporting, integrates Weka and R and Hadoop (Radoop), repository under sourceforge.net/projects/rapidminer/
Rattle	INT	data mining GUI for R, PMML support
ROOT	LIB	C++ support, LPGL license, general parallel processing framework
RWEKA	INT	integration of WEKA for R
Scilab	MAT	software for numerical computation
SciPy	MAT	scientific Python package including statistical tools
SMILE	LIB	specialized to Bayesian Networks, developed since 1998
TANAGRA	INT	software for educational and research purposes, includes data preparation and experimental analysis. see Chen07, Alcala09

Waffles	LIB	C++ library, additional command line functionality, some exotic methods
WEKA	LIB, DMS	most well-known software, integrated in many other tools, different extensions, e.g. for human genetics WEKA-CG
MEGA	SOL	only for genetic data, free for academic and commercial use, no redistribution

Table: 2 Open Source Tools

V. CONCLUSION

Many advanced for data mining are available either as open-source or commercial software. They cover a wide range of software products, from comfortable problem- independent data mining suites, to business-centered data warehouses with integrated data mining capabilities, to early research prototypes for newly developed methods. In this paper, nine different types of tools are presented: DMS, BIs, MATs, INT, EXT, SPECs, RES, IBs, and SOLs. They vary in many different characteristics, such as intended user groups, possible data structures, implemented tasks and methods, interaction styles, import and export capabilities, platforms and license policies are variable. Recent tools are able to handle large datasets with single features, time series, and even unstructured data like texts; however, there is a lack of powerful and generalized mining tools for multidimensional datasets such as images and videos.

References

[1]. Fayyad U, Piatetsky-Shapiro G, Smyth P. From data mining to knowledge discovery in databases. *AI Mag*1996, 17:37–54.

[2]. Smyth P. Data mining: Data analysis on a grand scale?*Stat Methods Med Res* 2000, 9:309–327.

[3]. Lovell MC. Data mining. *Rev Econ Stat* 1983, 65:1–11.

[4]. Han J, Kamber M. *Data Mining: Concepts and Techniques*.San Francisco: Morgan Kaufmann; 2006.

[5]. Hastie T, Tibshirani R, Friedman J. *The Elements ofStatistical Learning: Data Mining, Inference, and Prediction*.New York: Springer; 2008.

[6]. Engelbrecht AP. *Computational Intelligence - An Introduction*.Chichester: John Wiley; 2007.

[7]. Vesset D, McDonough B. Worldwide business intelligence tools 2008 vendor shares, IDC Competitive Analysis Report (2009).

[8]. Frank E, Hall M, Holmes G, Kirkby R, Pfahringer B,Witten I. Weka: A machine learning workbench for data mining. *Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers*. New York: Springer; 2005, 1305–1314.

[9]. Goebel M. A survey of data mining and knowledge discovery software tools, ACM SIGKDD Explorations. *Newsletter* 1999, 1:20–33.

[10]. Wang J, Hu X, Hollister K, Zhu D. A comparison andscenario analysis of leading data mining software. *IntJ Knowl Manage* 2008, 4:17–34.

[11]. Wang J, Chen Q, Yao J. Data mining software. In:Tomei L, ed., *Encyclopedia of Information TechnologyCurriculum Integration*. Hershey, PA: InformationScience Publishing; 2008, 173–178.

[12]. Giraud-Carrier C, Povel O. Characterising data miningsoftware. *Intell Data Anal* 2003, 7:181–192.

[13]. Chen X, Ye Y, Williams G, Xu X. A survey of opensource data mining systems, Lecture Notes in ComputerScience 2007, 4819:3–14.

[14]. Alcal ´a-Fdez J, S´anchez L, Garc´ıa S,del Jesus M,Ventura S, Garrell J, Otero J, Romero C, Bacardit J,Rivas V, et al. KEEL: A software tool to assess evolutionaryalgorithms for data mining problems. *SoftComput* 2009, 13:307–318.

