Mining Big Data in WEKA

RENATA JANOŠCOVÁ
Vysoká škola manažmentu, Trenčín, Slovakia

Abstract: Recent versions of WEKA 3.8 also provide access to new packages for distributed data mining. The first new package is called distributed WEKA Base. It provides base "map" and "reduce" tasks that are not tied to any specific distributed platform. A second, called distributed WEKA Hadoop, provides Hadoop-specific wrappers and jobs for these base tasks. A third, called distributed WEKA Spark, provides Spark-specific wrappers.

Keywords: Data Mining, Big Data, WEKA.

1 Introduction to Data Mining and Big Data

Now is an age of Data Mining (DM). Popularity of data mining can be proven by the fact that the result of searching the expression “data mining” in the Google (November 2016) - was more than 60 million pages.

1.1 Data Mining

Data Mining is multidisciplinary in nature (statistics, informatics, machine learning, artificial intelligence, enterprise information systems). “Data mining is the discovery of interesting, unexpected or valuable structures in large datasets...” [11]. An interesting definition of data mining is: “Data mining is a process of discovering various models, summaries and derived values from a given collection of data” [14]. Currently DM not be focused only on the database as we know them. Ordinary applications are DM in text files or on the web. Real DM applications based on multimedia content (hadoop) they are already among us.

1.2 Big Data

The term "Big Data" is currently even more popular as the term "data mining". Result of searching the expression “Big Data” in the Google (November 2016) - was more than 300 million pages. Big data is a long time among us, but only now they are in the spotlight the lay and professional public. They they are most frequent characterized with the aid of 3 - 5 “**V**”:

- **„Volume“** – amount of data is n*TB;
- **„Velocity“** – high speed of arrival of data and speed data processing;
- **„Variety“** – different types of data (the web content text, sound, video, digits, semi-structured databases and other ...).

These additional characteristics are sometimes appended:

- **„Veracity“** – informative value;
- **„Value“** – the value of information and knowledge.

We can be regarded as big data those entity which can be characterized by at **least 3 “V”** (Volume, Velocity and Variety).

1.3 Mining Big Data

We understand “Data Mining” as the an analogy as the "extracting valuable nuggets of large quantities of clay." Then we can understand the term “Mining Big Data” as a "extracting valuable nuggets of massive amounts of clay, stone and mud."

2 Software for Big Data

Commercial software systems for data mining and Big Data are mostly very expensive. Thankfully the market offers a lot of open source tools (software systems), which make the data mining more affordable for a lot businesses.

Big data, data that can’t maintain relationships. The problems with processing big data are varied, and no one tool can handle it all. Portal InfoWorld compiled 12 of the best open source tools for Big Data [13], processed in Tab. 1.

Tab. 1 The best open source big data tools by Bossie Awards 2016

No	Software system	Logo	Description	Special characteristics
1	Apache Spark™		In-memory distributed processing framework, which use batched processing of RDDs to a concept of a DataFrame without bounds [4].	Structured Streaming.
2	Apache Beam		To not rewrite code every time our processing engine changes. Extended features and performance of Google’s DataFlow [1].	Doesn’t support developer features like REPL.
3	TensorFlow™		Character recognition, image recognition, natural language processing, or some other kind of complicated machine learning application [20].	Run both distributed code and optimized parallel code on GPUs and CPUs.
4	Apache Solr™		Brings trusted and mature search engine technology to the enterprise. Able to deal with the scale of handling high query volumes across billions of documents [5].	Reliability.
5	Elasticsearch		Distributed search engine that focuses on modern concepts like REST APIs and JSON documents [8].	Able to detection and domain-specific business analytics.
6	SlamData		SlamData has a SQL-based engine that talks natively to MongoDB. Not sucking all the data into PostgreSQL and calling it a BI connector [9].	Basic analytics on MongoDB data store.

No	Software system	Logo	Description	Special characteristics
7	Apache Impala		It is Cloudera's engine for SQL on Hadoop. A row-based, distributed, massively parallel processing system [12].	If you need SQL over some files that you have on HDFS.
8	Apache Kylin™		It is designed to provide SQL interface and multi-dimensional analysis (OLAP) on Hadoop supporting extremely large datasets [3].	Lets you query massive data set at sub-second latency in 3 steps.
9	Apache Kafka™		Is the standard for distributed publish and subscribe. It is used for building real-time data pipelines and streaming apps [2].	Easy to install and configure
10	StreamSets™		Easily develop and reliably operate any-to-any data flows that connect a variety of sources to your Big Data platforms [19].	Robust and growing list of connectors.
11	Titan		Scalable graph database optimized for storing and querying graphs containing hundreds of billions of vertices and edges distributed across a multi-machine cluster [21].	Support graph traversals in real time.
12	Apache Zeppelin		Apache Zeppelin interpreter concept allows any language/data-processing-backend to be plugged into Zeppelin [25].	Supports Apache Spark, Python, JDBC, Markdown and Shell.

These software systems often implement methods other open source systems at its core. These software systems often implement methods of other free tools at their core such as R, KNIME, WEKA, Python and others.

The current trend is the integration of specialized software applications of DM in BIS[1]. An example for all can be *Pentaho Corporation* and the acquisition of highly specialized software application WEKA. In 2006 *Pentaho Corporation* acquired an exclusive license to use WEKA for BI[2] [16]. It forms a sub-module for data mining and predictive analytics system for management decisions *Pentaho BI Suite*.

[1] BIS - Business Information Systems.
[2] Business Intelligence.

3 Mining Big Data with WEKA

The **W**aikato **E**nvironment for **K**nowledge **A**nalysis (**WEKA**) project dates from the 1992 in University of Waikato, Hamilton, New Zealand [10]. WEKA is landmark system in data mining and machine learning, it is widely used tool for data mining research [16] and supports process models of data mining such as CRISP-DM [18]. This system had been released as open source software system. WEKA giving users free access to the source code and it has enabled a thriving community to develop and facilitated the creation of many projects that incorporate or extend WEKA [10].

3.1 Data Mining with WEKA

Basic functionality of WEKA:

- *Data preprocessing* – processing of format ARFF[3] and support of various other formats, filters);
- *Classification* - more than 100 classification methods, interface for classifiers implemented in Groovy and Jython;
- *Clustering* - unsupervised learning is supported by several clustering schemes;
- *Attribute selection* - the set of attributes used is essential for classification performance;
- *Data visualization* (tree viewer, Bayes network viewer with automatic layout, and a dendrogram, viewer for hierarchical clustering, 2D, 3D).

Furthermore Weka also includes *Association* and *Forecasting in time series.*

3.2 Graphical User Interfaces (GUI)

Design of graphical user interface began in 1999. In 2005 appeared the interactive versions of WEKA GUI - the Explorer, Experimenter, and Knowledge Flow interface (Fig. 1).

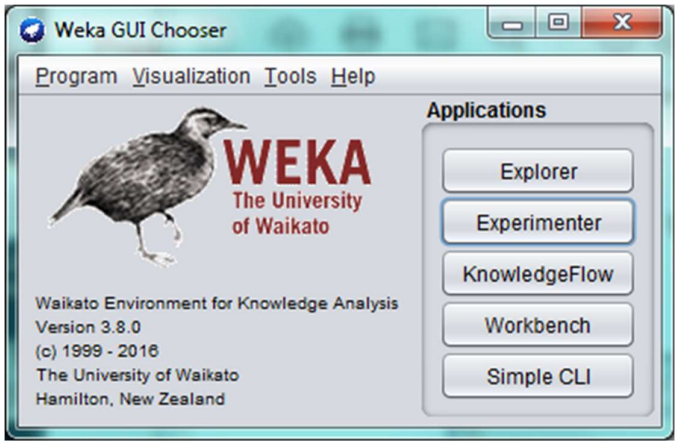

Fig. 1 The GUI Chooser

Displaying of attributes in the WEKA Explorer user interface:

[3] ARFF - **A**ttribute **R**elation **F**ile **F**ormat.

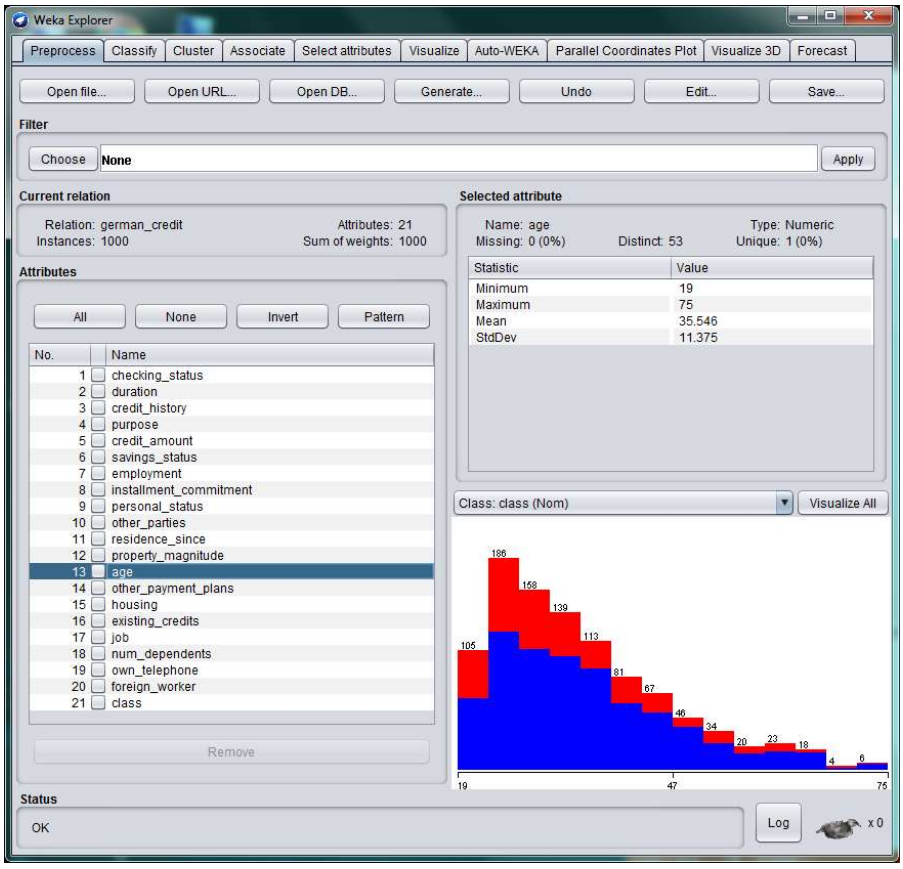

Fig. 2 The WEKA Explorer user interface, selected atribut "*age*"

Example of a decision tree – classifier with *J48:*

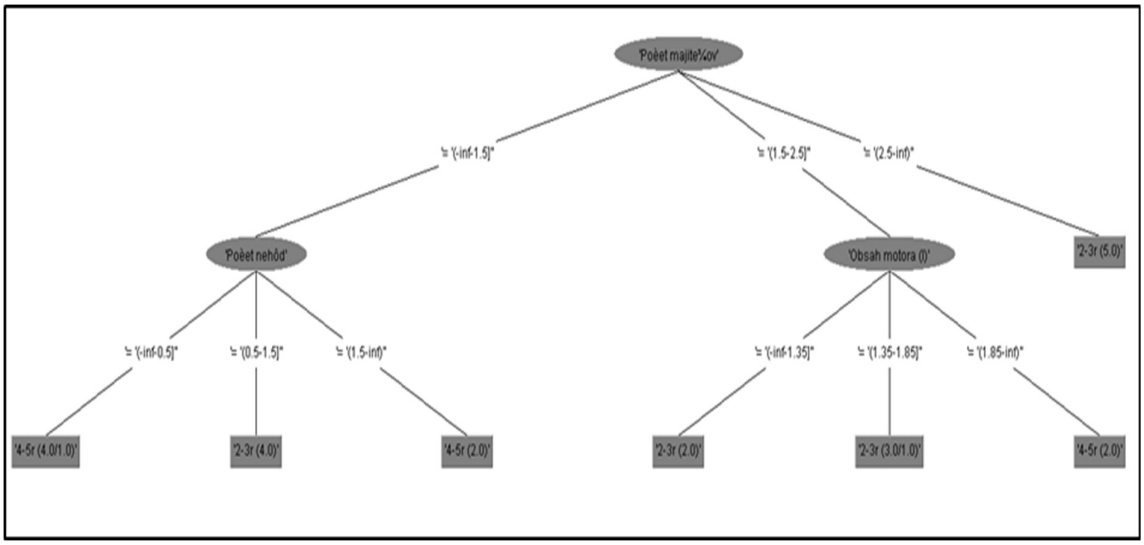

Fig. 3 Decision tree (J48)

The example mining of associations with *Apriori* associator:

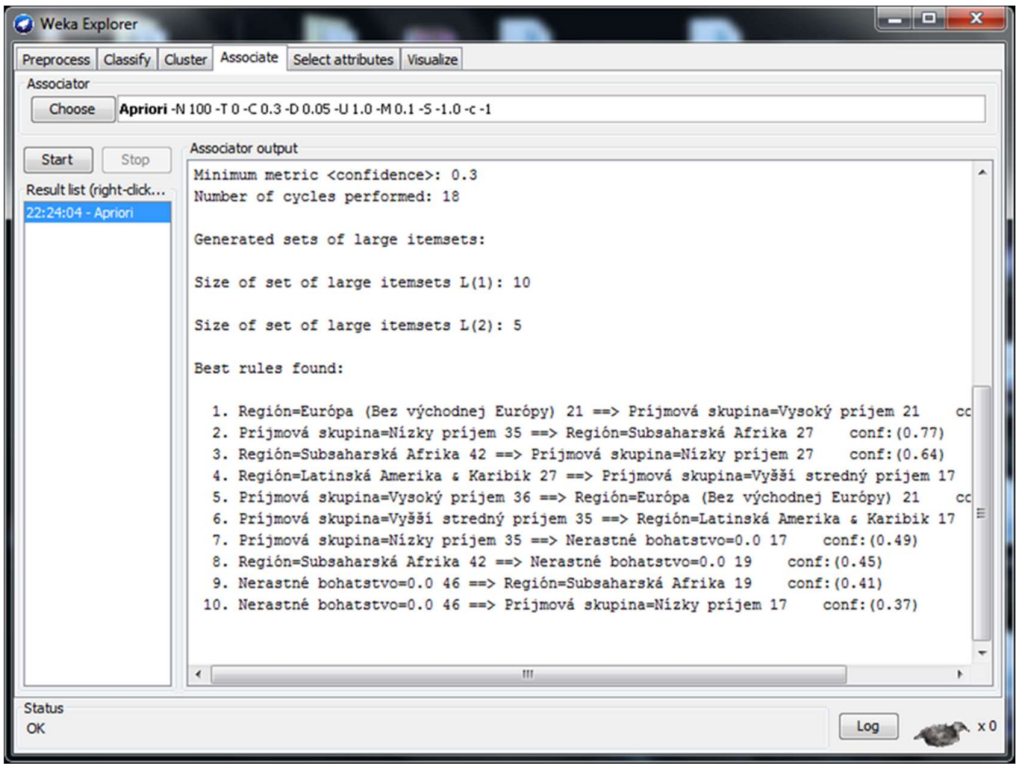

Fig. 4 Associator output (Apriori)

Example of prognosis by means of neural network (*Multilayer Perceptron*):

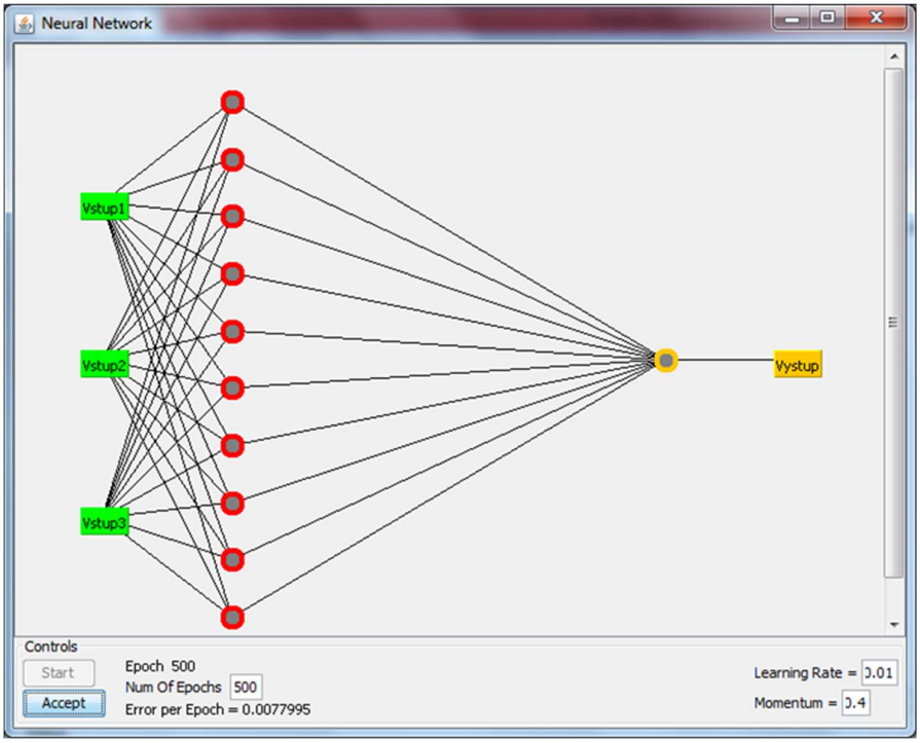

Fig. 5 Neural network vizualization (topology 3-10-1)

Most tasks that can be tackled with the *Explorer* can also be handled by the *Knowledge Flow*. In addition to batch-based training, its data flow model enables incremental updates with processing nodes that can load and preprocess individual instances before feeding them into appropriate incremental learning algorithms. It also provides nodes for visualization and evaluation [7] (Fig. 6).

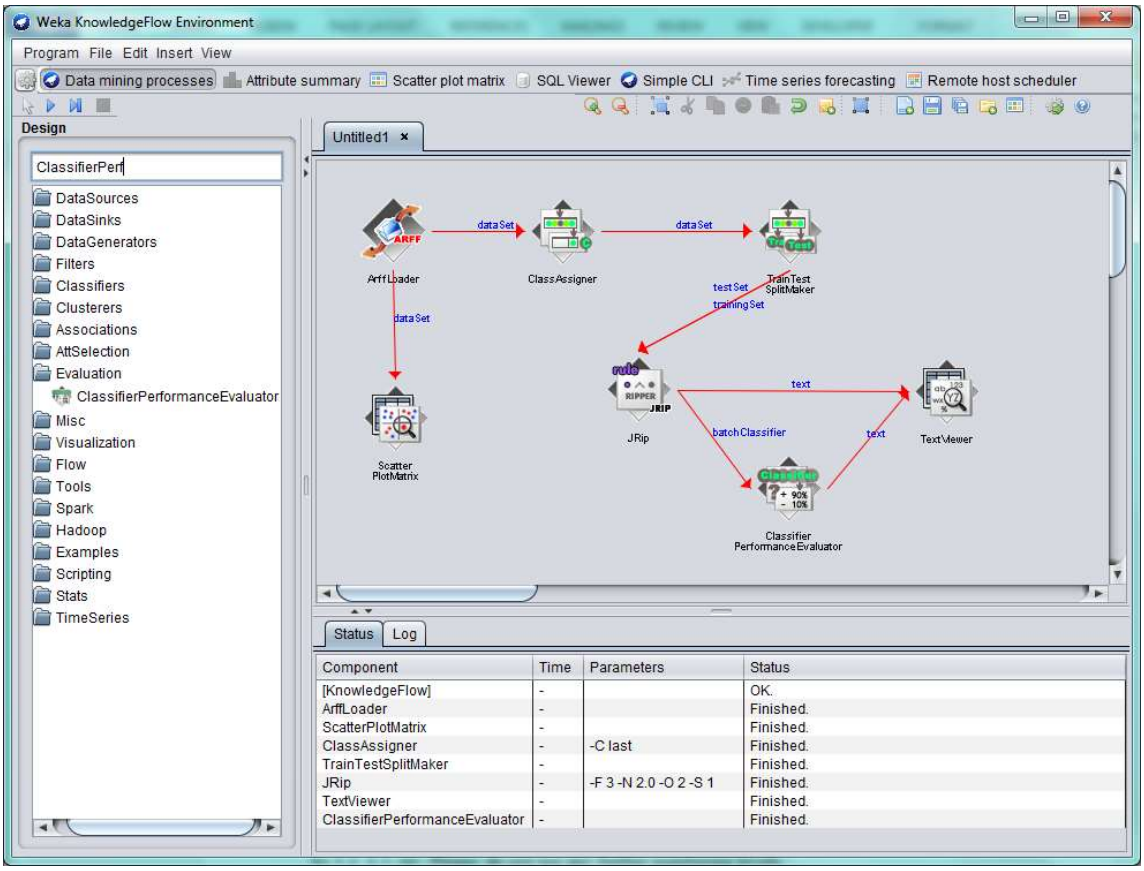

Fig. 6 The WEKA Knowledge Flow user interface

Once a set-up of interconnected processing nodes has been configured, it can be saved for later re-use.

3.3 Mining Big Data with WEKA

From version 3.7.2 WEKA has the concept of a package as a bundle of additional functionality, separate from that supplied in the main weka.jar file. This simplifies the core system and allows users to install just what they need or are interested in. WEKA includes a facility for the management of packages and a mechanism to load them dynamically at runtime. There is GUI package manager also (Fig. 7).

This "Package Manager" adds support for running WEKA in Hadoop and in Spark. With the support of these packages system can can predict in real time in a very challenging real-world applications with almost all models [22].

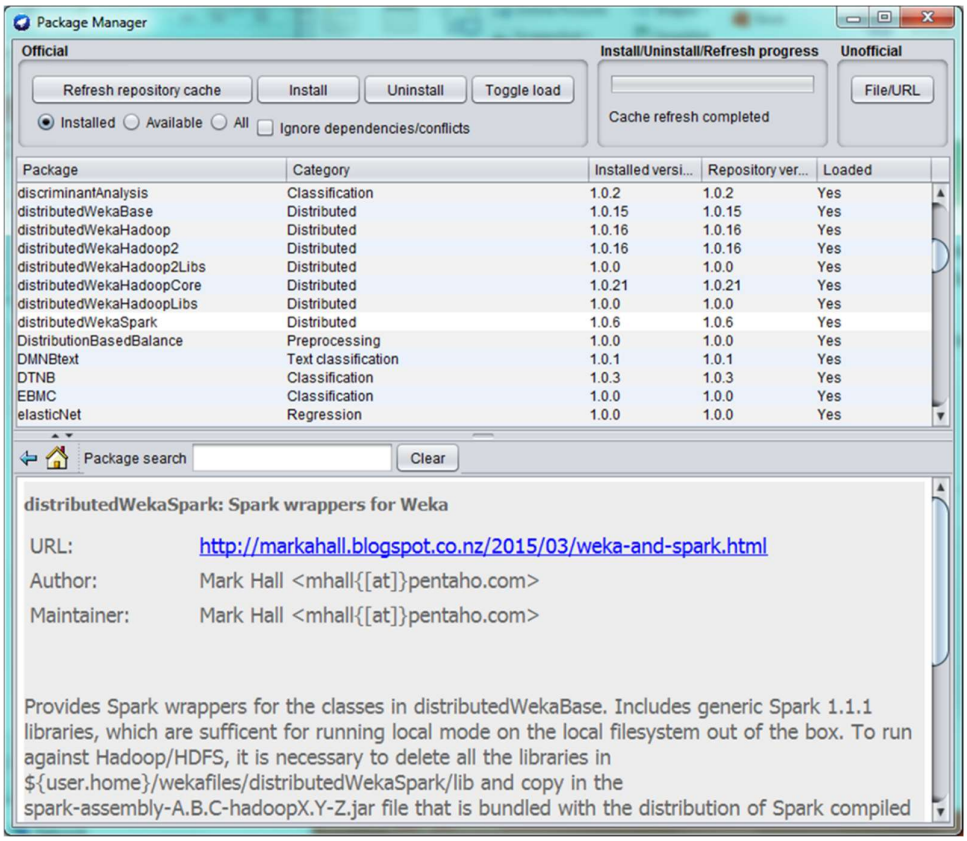

Fig. 7 Package Manager

3.3.1 MOA

In version 3.8 is package for **MOA** (**M**assive **O**nline **A**nalysis) data streams. Nowadays, data is generated at an increasing rate from sensor applications, measurements in network monitoring, log records, manufacturing processes and others.

All this data generated can be considered as streaming data since it is obtained from an interval of time. In the data stream model, data arrive at high speed, and an algorithm must process them under very strict constraints of space and time. MOA is an open-source framework for dealing with massive, potentially infinite, evolving data streams, it permits evaluation of data stream learning algorithms on large streams, in the order of tens of millions of examples [6].

The implemented classifier methods currently include: *Naive Bayes, Decision Stump, Hoeffiding Tree, Hoeffiding Option Tree, Bagging, Boosting*, and other (Fig. 8). MOA contains also an experimental framework for clustering data streams.

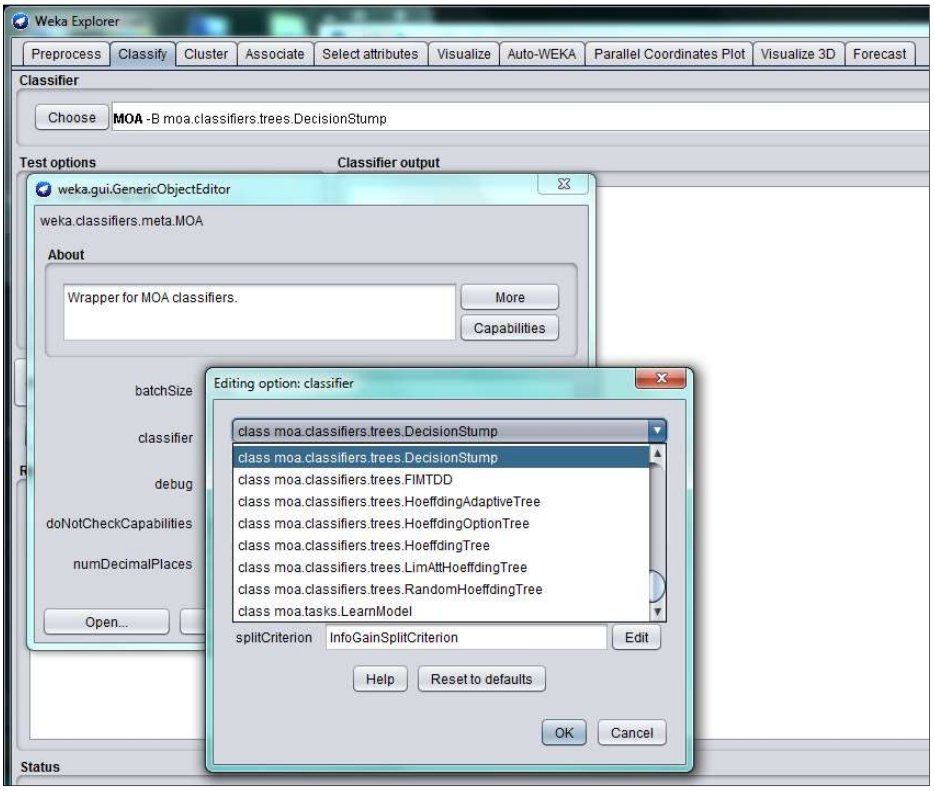

Fig. 8 MOA classifiers

Though the current focus in MOA is on clustering and classification, the autors plan to extend the framework to include regression, and frequent pattern learning [6].

3.3.2 Distributed WEKA

Distributed WEKA is a plugin (since version 3.7) that allows WEKA algorithms to run on a cluster of machines. This would use this when dataset is too large to load into main RAM, or use an algorithm that would take too long to run on a single machine. Distributed WEKA works with distributed processing frameworks that use map-reduce. It is more suited to large, offline, batch-based processing scenarios. Data is divided up over the nodes in the processing cluster (the machines in a processing cluster) and is conquered. Each piece is conquered independently of the other pieces [23]. The distributed WEKA plugin is made up of two packages: *distributedWekaBase*, that provides general map-reduce style tasks for machine learning that are not tied to any particular map-reduce framework implementation. A second package is *distributedWekaSpark*. This is a wrapper for the base tasks that works on the *Spark* platform (Fig. 7).

In addition to these basic packages, there is also a packages for work with Hadoop, depending on which version or flavor of *Hadoop* that you have installed [23]. The design goals of distributed Weka is to provide a similar experience to using standalone desktop Weka. It enables to use any classification or regression learner in Weka and also has support for clustering.

4 Conclusions

A common opinion is that the WEKA machine learning software cannot be applied to large datasets and Big Data. Main problem lies in training models from large datasets, not prediction for large datasets. Weka is being used to make predictions in real time in very demanding real-world applications. It is correct that it may be impossible to train models from large datasets using the WEKA *Explorer* GUI (even when the Java heap size has already been increased), because the Explorer always loads the entire dataset into the computer's main memory [23]. When dealing with large datasets, it is best to use a command-line interface (CLI) or the *Simple CLI* included in WEKA, use *Knowledge Flow* GUI, or write code directly in Java or a Java-based scripting language such as Groovy or Jython [23]. For new version WEKA 3.8, there is a library that provides access to the MOA data stream software containing state-of-the-art algorithms for large datasets or data streams. WEKA 3.8 also provides access to new packages for distributed data mining. The models that are output from distributed WEKA are normal WEKA models. That means they can be saved to your file system, loaded into desktop WEKA at a later stage, and used for making predictions, just like any other Weka model [23]. At the end, it must be stressed that the most important factor in the DM human resources.

It is not enough just to obtain information, but is equally important to choose the most important information, correctly interpret the information obtained and then make the necessary and effective action.

Literature

1. Apache Beam (incubating), 2016. [online]. Available at: http://beam.incubator.apache.org/ [Accessed 5 August 2016].
2. Apache Kafka™, 2016. [online]. Available at: https://kafka.apache.org/ [Accessed 1 June 2016].
3. Apache Kylin, 2015. [online]. Available at: https://kylin.apache.org/ [Accessed 4 July 2016].
4. Apache Spark™ - Lightning-Fast Cluster Computing. [online]. Available at: http://spark.apache.org/ [Accessed 2 August 2016].
5. Apache Solr, 2016. [online]. Available at: https://lucene.apache.org/solr/ [Accessed 9 June 2016].
6. Bifet, A. et al., 2010. MOA: Massive Online Analysis, a Framework for Stream Classification and Clustering. *Journal of Machine Learning Research (JMLR)*, 11, pp. 44-50.
7. Bouckaert, R., Frank, E., Hall, M. et al. WEKA - Experiences with a Java Open-Source Project. *Journal of Machine Learning Research*, 11, pp. 2533-2541.
8. Elasticsearch: REST ful, Distributed Search & Analytics, 2016. [online]. Available at: https://www.elastic.co/products/elasticsearch [Accessed 6 August 2016].
9. Enterprise Analytics for Modern Data | SlamData, 2015. [online]. Available at: https://slamdata.com/ [Accessed 7 July 2016].
10. Hall, M., Frank, E., Holmes, G. et al. 2009. The WEKA data mining software: An update. *SIGKDD Explorations*, 11(1), pp. 10–18.

11. Hand, D. J. 2007. Principles of Data Mining. *Drug safety* [online]. 2007, 30(7), pp. 621-622. [cit. 2016-02-02]. ISSN 01145916. Aviallable on: http://ezproxy.cvtisr.sk:2271/journal/40264/30/7/page/1 [Accessed 1 July 2016].
12. Impala. [online]. Available at: https://impala.apache.org/index.html [Accessed 1 June 2016].
13. InfoWorld - Technology insight for the enterprise, ©1994-2016. [online]. Available at: http://www.infoworld.com/ [Accessed 9 July 2016].
14. Kantardzic, M. 2011. *Data Mining. Concepts, Models, Methods, and Algorithms*. 2nd Edition. IEEE/John Wiley & Sons, 2011. ISBN: 978-0-470-89045-5.
15. Oliver, A. C. & Pointer, I. Bossie Awards 2016: The best open source big data tools, 2016. [online]. Available at: http://www.infoworld.com/article/3120856/open-source-tools/bossie-awards-2016-the-best-open-source-big-data-tools.html#slide1 [Accessed 5 July 2016].
16. Pentaho Acquires WEKA Project. ©2005-2016. [online]. Available from http://www.pentaho.com/pentaho-acquires-weka-project. [Accessed 6 February 2016].
17. Piatetsky-Shapiro, G. 2005. *KDnuggets news on SIGKDD service award*. [online]. Available at: http://www.kdnuggets.com/news/2005/n13/2i.html [Accessed 9 July 2016].
18. Shearer, C. 2000. The CRISP-DM model: The new blueprint for data mining. *Journal of Data Warehousing*, 5(4).
19. StreamSets - Performance Management of Data Flows, 2016. [online]. Available at: https://streamsets.com/ [Accessed 9 June 2016].
20. TensorFlow - an Open Source Software Library for Machine Intelligence. [online]. Available at: https://www.tensorflow.org/ [Accessed 3 July 2016].
21. Titan: Distributed Graph Database. [online]. Available at: http://titan.thinkaurelius.com/ [Accessed 4 June 2016].
22. WEKA – home. [online]. Available at: http://weka.wikispaces.com/ [Accessed 6 July 2016].
23. WekaMOOC: Advanced Data Mining with Weka. University of Waikato, 2016. [online courses]. Available at: https://weka.waikato.ac.nz/advanceddataminingwithweka [Accessed 5 June 2016].
24. Weka 3 - Mining Big Data with Open Source Machine Learning Software in Java. University of Waikato. [online]. Available at: http://www.cs.waikato.ac.nz/ml/weka/bigdata.html [Accessed 2 July 2016].
25. Zeppelin. [online]. Available at: https://zeppelin.apache.org/ [Accessed 8 June 2016].

Contact data:

Renata Janoščová, Ing., PhD.

Vysoká škola manažmentu v Trenčíne

Bezručova 64

911 01 Trenčín, Slovakia

rjanoscova@vsm.sk