Joshua Hirtz
May 6, 2008

Weka

Executive Summary

Commercial machine learning programs can be expensive and out of reach for those with a tight budget. Therefore, I have chosen to take a look at an alternative to commercial programs in the form of the Waikato Environment for Knowledge Analysis (a.k.a. WEKA)

WEKA is free and protected under the General Public License, GNU, and it comes with a large library of algorithms that can be accessed through one of its four interfaces, which include:

1. A command line interface known as Simple CLI.
2. A GUI interface known as Explorer.
3. Another GUI interface with the ability to do multiple algorithms in a row on the same data known as Experimenter.
4. A data flow inspired interface known as KnowledgeFlow.

In order for data to be run through any of these however, they much follow a compatible format for WEKA. One of which is the .arff file format which consists of four parts. These are:

1. @RELATION *name* which gives the filename of the dataset.
2. @ATTRIBUTE *name type*, which gives the name of the attribute, and the type (numeric, nominal, string, and date).
3. @DATA which says that the following information is the data to form the algorithms on.
4. *data*,*data*,*data* which is the data of each attribute being separated by a comma and each instance being on a line of its own.

For example here is a clip of an iris.arff dataset that comes with the WEKA program:

```
@RELATION iris

@ATTRIBUTE sepallength  REAL
@ATTRIBUTE sepalwidth   REAL
@ATTRIBUTE petallength  REAL
@ATTRIBUTE petalwidth   REAL
@ATTRIBUTE class  {Iris-setosa,Iris-versicolor,Iris-virginica}

@DATA
5.1,3.5,1.4,0.2,Iris-setosa
4.9,3.0,1.4,0.2,Iris-setosa
4.7,3.2,1.3,0.2,Iris-setosa
4.6,3.1,1.5,0.2,Iris-setosa
5.0,3.6,1.4,0.2,Iris-setosa
5.4,3.9,1.7,0.4,Iris-setosa
4.6,3.4,1.4,0.3,Iris-setosa
5.0,3.4,1.5,0.2,Iris-setosa
```

Finally, thanks to WEKA's being open source, it allows the user to personally go into the coding and even try to expand upon it if they wish. However, even with WEKA's benefits it still seems to lack some of the flexibility, power, and support that a commercial product like SPSS's Clementine offers.

Thus, I would suggest WEKA as an alternative product for users with a tight budget that are planning to deal with small datasets and/or interested in learning how one can implement data

mining algorithms into a program. However, if they are looking for something that can easily go through large datasets, then a commercial product cannot be beat.

Problem Description

Data mining algorithms can be complex and time consuming to calculate out by hand even before adding on the massive amounts of data that might need to be placed through them. This has made them quite unusable in the past. However, now that we have computers that can do these same calculations in a mere fraction of the time that it would have taken us before, data mining has become a large part of the world today and likely will continue to be in the future. While that in itself is not so much of a problem, finding a program that is both affordable and capable of accomplishing what one needs it to can be. I have chosen to take a look at a machine learning program called Waikato Environment for Knowledge Analysis, or also known as WEKA, to see if it could help me solve this problem.

Analysis Technique

Waikato Environment for Knowledge Analysis, or WEKA, is a free open source machine learning program available under the GNU General Public License. It was created in New Zealand by the Computer Science Department of the University of Waikato. Its goals, as stated from the WEKA website, include:

- make ML techniques generally available;
- apply them to practical problems that matter to New Zealand industry;
- develop new machine learning algorithms and give them to the world;
- contribute to a theoretical framework for the field.

On the technical side, it comes with four different interfaces that include Simple CLI, Explorer, Experimenter, and KnowledgeFlow.

Simple CLI, as its name gives away, is a command line interface that allows one to insert lines of code that then tells the program what to do. Next, Explorer uses a GUI to make the program a bit friendlier for the user by implementing the use of buttons and text fields so that the user does not have to have an extensive knowledge of the programs code to do as they wish. Then, you have the Experimenter that is another GUI interface that has the capabilities of allowing the user to create an experiment that will perform multiple algorithms in a row on the same data so that the hassle of performing each one individually is taken away. And finally we come to the KnowledgeFlow which is a data-flow inspired interface that allows the user to drag icons which represent different functions of the program into a field and connect them to form experiments of their own. Unfortunately this feature is still being worked out, but it is capable of performing some functions at this time.

Each of these features gives access to part, if not all, of the algorithms in WEKA's extensive library, however for each to perform on the data, the data must be put into a format that WEKA can use. For mine, I focused on their .arff data type which consists of the following:

1. @RELATION *name* which gives the filename of the dataset.
2. @ATTRIBUTE *name type* which gives the name of the attribute and the type (numeric, nominal, string, and date).
3. @DATA which says that the following information is the data to form the algorithms on.
4. *data*,*data*,*data* which is the data of each attribute being separated by a comma and each instance being on a line of its own.

For example, the following picture shows the beginning of an iris.arff dataset that comes with the program.

```
@RELATION iris

@ATTRIBUTE sepallength  REAL
@ATTRIBUTE sepalwidth   REAL
@ATTRIBUTE petallength  REAL
@ATTRIBUTE petalwidth   REAL
@ATTRIBUTE class  {Iris-setosa,Iris-versicolor,Iris-virginica}

@DATA
5.1,3.5,1.4,0.2,Iris-setosa
4.9,3.0,1.4,0.2,Iris-setosa
4.7,3.2,1.3,0.2,Iris-setosa
4.6,3.1,1.5,0.2,Iris-setosa
5.0,3.6,1.4,0.2,Iris-setosa
5.4,3.9,1.7,0.4,Iris-setosa
4.6,3.4,1.4,0.3,Iris-setosa
5.0,3.4,1.5,0.2,Iris-setosa
```

Finally, WEKA, thanks to being open source, gives the user the ability to go into the program's code and see how it works and/or even try to expand upon it unlike a commercial product. However, despite WEKA's benefits, it still lacks a bit concerning commercial products such as SPSS's Clementine. Three such things include:

1. WEKA comes as a product that has everything incorporated into one program and thusly it lacks the flexibility Clementine has to be able to be setup appropriately for different environments like servers.
2. WEKA chokes on large datasets while Clementine is designed to be able to deal with them efficiently.
3. WEKA seems to lack the support SPSS gives to its Clementine users.

Assumptions

Data is preformatted to a style that WEKA can read.

Results

WEKA is an affordable and flexible program worth your time if you are planning to deal with smaller datasets, but lacks the power of commercial products like Clementine. However, thanks to its GNU General Public License, it gives the user the chance to look behind the scenes at the programs coding and even attempt to improve upon them if they so wish. I would suggest WEKA for those that are looking for a free program capable of dealing with small datasets, interested in learning how data mining algorithms can be implemented into a program to deal with data, and/or trying their luck at creating an algorithm of their own.

Issues

There were many sources about the program. However none of the ones I looked at seemed to cover all of the questions I had.

The KnowledgeFlow interface is not fully functional and is still being worked on.

References

Collaborated with John Aleshunas.

Weka Machine Learning Project. (N.A.). Retrieved May 6, 2008, from http://www.cs.waikato.ac.nz/~ml/index.html

WEKA (Machine Learning). (May 3, 2008). Retrieved May 6, 2008, from http://en.wikipedia.org/wiki/WEKA

Frank, Eibe. (N.A.). Machine Learning with WEKA. Retrieved May 6, 2008, from http://www.cs.waikato.ac.nz/ml/weka/

Pfahringer, Bernhard. (N.A.). Machine Learning with WEKA. Retrieved May 6, 2008, from http://www.cs.waikato.ac.nz/ml/weka/

N.A. (N.A.). WEKA: Machine Learning and Data Mining as ClickandPlay. Retrieved May 6, 2008, from http://www.google.com/search?q=cache:MeH2vRZYZ5EJ:www.informatik.uni-freiburg.de/~mlpult/slides/WEKA-12-01.pdf+weka+pros+and+cons&hl=en&ct=clnk&cd=7&gl=us

Jakob, Michal. (N.A.). WEKA: Machine Learning & Softcomputing. Retrieved May 6, 2008, from http://www.google.com/search?q=cache:dSWjfexxIDcJ:cyber.felk.cvut.cz/gerstner/teaching/ppdm/weka_lecture.ppt+weka+pros+and+cons&hl=en&ct=clnk&cd=6&gl=us

Scope Creep. (April 28, 2008). Retrieved May 6, 2008, from http://en.wikipedia.org/wiki/Functionality_creep

Assessing Student Proficiency in a Reading Tutor that Listens. (N.A.). Retrieved May 6, 2008, from http://www.cs.cmu.edu/~listen/pdfs/UM2003_paper_test_prediction.pdf

en:SimpleCLI (3.5.6). (June 4, 2007). Retrieved May 6, 2008, from http://weka.sourceforge.net/wekadoc/index.php/en:Simple_CLI_%283.5.6%29

en:Explorer (3.5.6). (June 4, 2007). Retrieved May 6, 2008, from http://weka.sourceforge.net/wekadoc/index.php/en:Explorer_%283.5.6%29

en:Experimenter – Standard Experiments (3.5.6). (February 25, 2008). Retrieved May 6, 2008, from http://weka.sourceforge.net/wekadoc/index.php/en:Experimenter_-_Standard_Experiments_%283.5.6%29

en:KnowledgeFlow (3.5.7). (February 21, 2008). Retrieved May 6, 2008, from http://weka.sourceforge.net/wekadoc/index.php/en:Knowledge_Flow_%283.5.7%29