Gnome Data Mine Tools Evaluation Report

CMPUT695 Assignment 2

Haobin Li, Junfeng Wu
Thursday, November 04, 2004

Overview

The gnome-data-mine-tools (GDataMine) is an open source data mining tool set which collects some free data mining programs. GDataMine contains two parts. One part is the underlying command-line data mining applications, and the other is the GUIs for these applications.

Currently, GDataMine includes three data mining tools:

- Apriori based association rules mining tool (Apriori)
- Decision tree classifier (DT)
- Bayes classifier (Bayes)

We evaluated GDataMine from several aspects and gave a rating for each aspect. The meaning of our rating system is described below:

- Unacceptable (✱)
- Poor (✱✱)
- Acceptable (✱✱✱)
- Good (✱✱✱✱)
- Excellent (✱✱✱✱✱)

We also compared GDataMine with another open source data mining tool Weka.

1. Structure (✱✱✱✱)

GDataMine separates the underlying applications and the user interfaces. The underlying applications are implemented by Christian Borgelt (http://fuzzy.cs.uni-magdeburg.de /~borgelt/software.html) and written in C. GDataMine provides GUIs, which are written in Python, on top of these applications to achieve ease of use and better human-machine interaction.

The separation of algorithms and GUIs also helps to grow and improve GDataMine overtime. Borgelt's implementations have been improving steadily over the past ten years. GDataMine can easily integrate newer versions of the underlying applications and become better.

One drawback of the separation is that there is no guarantee that the GUIs will work with the future versions of underlying applications. The authors of underlying applications may change the command line parameters in future versions without informing the authors of GUIs. The end users may have to suffer the delay of the new GUIs.

2. Correctness and Performance (✱✱)

Data mining tasks typically involve crunching huge amount of data and performing intensive computation, thus both runtime and memory requirement are crucial criteria in evaluating data mining software.

Thanks to the highly efficient programming language C and many years optimization, GDataMine excels in runtime performance. In fact, the Apriori in GDataMine is one of the fastest implementations.

However, there are some errors/limitations in the Apriori implementation. It can only generate the rules with one item as consequence, for example, "264 <- 234, 300, 350, 8, 145 (2%, 100%)".

As the comparison, Weka is limited by the performance of Java VM. It also has some memory issues when it is dealing with big dataset (see Figure 1).

Figure 1

3. User Interface (✱✱✱)

GDataMine provides a nice GUI for each underlying applications. Figure 2 is an example of the Bayes classifier's interface.

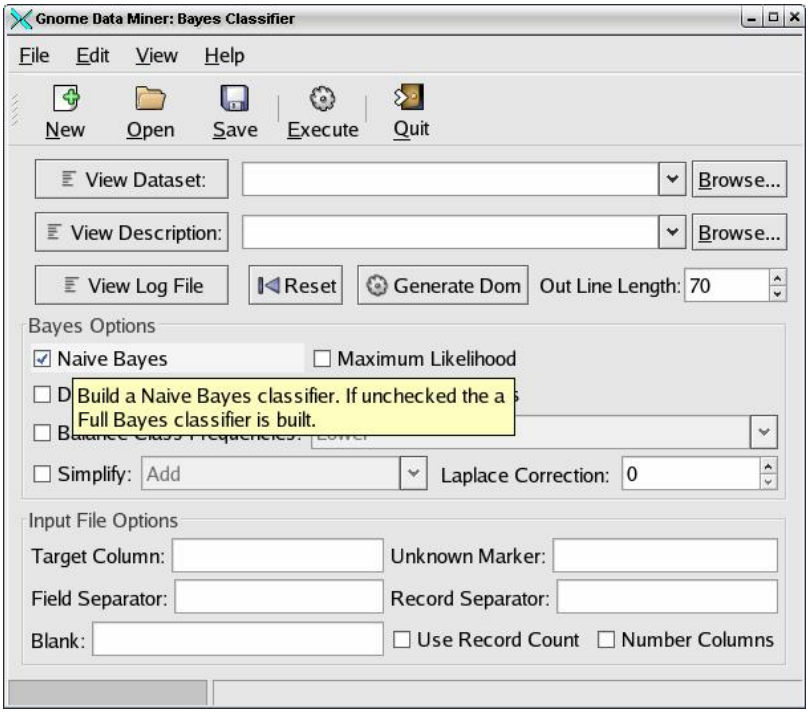

Figure 2

The GUI also provides tool tip help for each item. For example, when a user moves the mouse over the checkbox "Naïve Bayes", a hint will pop up. Thus, the user does not have

to remember all the parameters for the underlying application and can use the application easily.

However, the GUIs have not been totally implemented yet. For instance, clicking on some menu or toolbar item will open a warning window, which says "This function is not yet implemented" (see Figure 3).

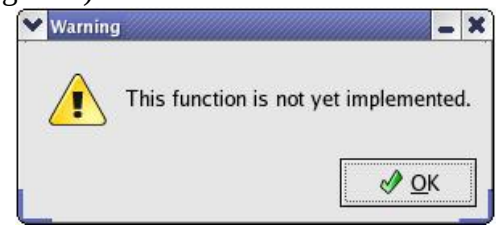

Figure 3

GDataMine also lacks visualization features. It only shows the result in plaint text. On the other hand, other data mining tools (e.g. Weka) provide many features to represent the result in graphics.

Installation and Execution (✱✱)

For a user with administrator privilege, there is no problem for installation and execution. But for other users, it is a totally different story.

GDataMine provides a "Makefile" to do the installation. However, it does not give the chance to change the default installation folder. The default setting is "/usr/local", which makes that a normal user (without administrator privilege) can not install GDataMine.

The default setting also causes the execution problem. Even if an administrator installed GDataMine, a normal user may still not be able to successfully execute it. GDataMine needs to write temporary file under "/tmp" folder where a normal user may not have the permission to write there, for example, in our department's Linux box.

With these problems, the GUIs of GDataMine is almost useless for normal users. To fix these problems, the user needs to know the programming language Python and how to set up the "Makefile".

The "Makefile" does not give the option to uninstall the tool. The administrator has to manually delete every file created during the installation process.

Weka has two ways to get installed. One can use the installer version to install/uninstall weka or simply unzip the package to any folder and delete the folder if do not need it anymore.

Documentations (✱)

There are only several web pages telling the user how to extract the files and how to execute each tool.

There is also no information about which input format each tool can take. The users have to look at the sample data to figure it out.

Weka has very detailed documents about each algorithm, parameter and data format.

API and Source Code (✱)

GDataMine is an open source tool set under the GNU General Public License (GNU/GPL). Weka is also an open source tool set under GNU/GPL.

GDataMine does not provide any API for programmers to extend the functions. Programmers have to read the source code and extract the parts by themselves.

On the other hand, Weka has the API documents about each Java class to help programmers to reuse or extend the functions.

Cost (✱✱✱✱✱)

Free

Multiple Platform Support (✱)

The underlying applications can be recompiled under other platforms. However, the GUIs only work under Linux with Gnome 2.

As the comparison, Weka can run under any platform with Java VM.

Conclusion

GDataMine is not a tool worth acquiring. Although it is a fast tool, it has some errors or limitations with the results. In addition, it lacks user manual and API documents; the GUIs only work under Linux with gnome; only an administrator can install and use it.

Comparing to other data mining tools, GDataMine is more like in alpha version status.

In our opinion, the ideal data mining tool should at least have the following features:

- Providing correct result
- Capacity of handling large dataset
- Linear running time with respect to the dataset size
- Reasonable memory usage
- Complete parameter control through GUI
- Full documentation including user manual and API documents
- Visualization of results

The following features are also worth to have:

- Capacity of taking different data formats
- Multiple platform
- Low cost

Appendix

Feature Table about GDataMine and Weka

	GDataMine	**Weka**
Free	Yes	Yes
Programming Language	C/Python	Java
Source code	Open source	Open source
Platform	Linux with gnome	Multiple
GUI	Yes but not fully implemented	Yes
Associate Rules	Yes (1 algorithm)	Yes (3 algorithms)
Classification	Yes (2 algorithms)	Yes (7 algorithms)
Clustering	No	Yes (5 algorithms)
Visualization	No	Yes
Documentation	Brief	Detail
Installer	Yes but not work well	Yes