A Survey of Open Source Data Mining Systems

Xiaojun Chen[1], Graham Williams[2], and Xiaofei Xu[1]

[1] Department of Computer Science and Technology
Harbin Institute of Technology, Harbin 150001, China
xjchen_hitsz@yahoo.com.cn
[2] School of Information Sciences and Engineering
University of Canberra, Australia
graham.williams@togaware.com

Abstract. Open source data mining software represents a new trend in data mining research, education and industrial applications, especially in small and medium enterprises (SMEs). With open source software an enterprise can easily initiate a data mining project using the most current technology. Often the software is available at no cost, allowing the enterprise to instead focus on ensuring their staff can freely learn the data mining techniques and methods. Open source ensures that staff can understand exactly how the algorithms work by examining the source code, if they so desire, and can also fine tune the algorithms to suit the specific purposes of the enterprise. However, diversity, instability, scalability and poor documentation can be major concerns in using open source data mining systems. In this paper, we survey open source data mining systems currently available on the Internet. We compare 12 open source systems against several aspects such as general characteristics, data source accessibility, data mining functionality, and usability. We discuss advantages and disadvantages of these open source data mining systems.

Keywords: Open source software, data mining, FLOSS

1 Introduction

Open source software has a solid foundation with, for example, the GNU project [1] dating from 1984. Open source software (also referred to as free, librè, open source software, or FLOSS) is well known through the GNU software suite upon which GNU/Linux is based, but also through the widely used MySQL, Apache, JBoss, and Eclipse software, just to highlight a few. Open source host site `sourceforge.net`, for example, lists over 100,000 open source projects.

Open source for business intelligence (BI) has also been gathering momentum in recent years. The open source database, MySQL [2], is widely used in building data warehouses and data marts to support BI applications. The open source data mining platform, Weka [3], has been a popular platform for sharing

*This paper was supported by the National Natural Science Foundation of China (NSFC) under grants No.60603066.

algorithms amongst researchers. In the past 6 months, for example, there have been between 16,353 and 29,950 downloads per month [3].

Open source data mining is particularly important and effective for small and medium enterprises (SMEs) wishing to adopt business intelligence solutions for marketing, customer service, e-business, and risk management. Due to the high cost of commercial software and the uncertainty associated with bringing data mining into an enterprise, many SMEs look to adopt a low cost approach to experimenting data mining solutions and in gaining data mining expertise. With open source software an enterprise can easily initiate a data mining project using the most current technology. Often the software is available at no cost, allowing the enterprise to instead focus on ensuring their staff can freely learn the data mining techniques and methods. Open source ensures that staff can understand exactly how the algorithms work by examining the source code, if they so desire, and can also fine tune the algorithms to suit the specific purposes of the enterprise. Therefore, SMEs no longer need be trailing the beneficial data mining technology.

However, issues such as diversity, stability, scalability, usability, documentation and support hinder the wide adoption of open source data mining in business. Such issues present challenges and incur costs for open source data mining projects, whereby effort must be expended by the business user on dealing with these deficiencies rather than solving business problems.

In this paper, we present a survey of 12 popular open source data mining systems available on the Internet. We evaluate the comparative characteristics of these open source systems, their data access functionality, their data mining functionality, and their usability, including user interfaces and extensibility. Based on the analysis, We discuss their advantages and disadvantages.

The rest of this paper is organized as follows. Section 2 provides a brief introduction to data mining and data mining systems. Section 3 surveys and evaluates 12 commonly used open source data mining systems, discussing their advantages and disadvantages.

2 Data Mining and Data Mining Systems

Data mining refers to the process of extracting new and useful knowledge from large amounts of data [4]. Data mining is widely used to solve many business problems, such as customer profiling [5], customer behavior modeling [6], credit scoring, product recommendation, direct marketing [7], cross selling, fraud detection [8, 9]. Data mining is adopted in many industries, e.g., retail [10], bank, finance [11], and so on [12, 13].

The solution to a data mining problem is carried out in the data mining process, which varies depending on the application domain. In general, a data mining process consists of the following seven steps [14]:

[3] `http://sourceforge.net/project/stats/?group_id=5091&ugn=weka&type=&mode=year`

1. Identify the business problems.
2. Identify and study data sources, and select data.
3. Extract and preprocess data.
4. Mine the data, e.g., discover association rules or build predictive models.
5. Verify the mining results.
6. Deploy models in the business process.
7. Measure the return on investment (ROI).

Each data mining process is composed of a sequence of data mining operations, each implementing a data mining function or algorithm. We can categorize data mining operations into the following groups:

1. **Data Understanding Operation**: Access data from various sources and explore the data to become familiar with it and to "discover" the first insights.
2. **Data Preprocessing Operation**: Generally involves data filtering, cleaning, and transformation, to construct the final dataset for the modeling operations.
3. **Data Modeling Operation**: Implements the data mining algorithms, such as k-means clustering. These operations are used to build data mining models. The common modeling operations include classification, prediction, clustering, association rule, and interactive exploration such as link analysis.
4. **Evaluation Operation**: Used to compare and select data mining models by choosing the best one. Common operations include confusion matrix, lift chart, gain chart, cluster validation and visualization.
5. **Deployment Operation**: Involves deploying a data mining model to make decisions, such as using a predictive model to predict potential customer churn or a campaign model to score customers for a target campaign.

A *data mining system* is a software system that integrates many operations, and provides an easy-to-use (and often graphical) user interface to effectively perform the data mining process. Goebel and Gruenwald [15] previously investigated 43 open source and closed source data mining systems and presented several important features for data mining to accommodate its users effectively. We use some of the same characteristics for our comparison, but also limit our study to just open source software. Also, their study was conducted over 7 years ago, and most of the open sources systems considered here did not exist then.

In the next section, we survey the commonly used open source data mining systems and evaluate their characteristics.

3 Survey of Open Source Data Mining Systems

In this section we survey 12 commonly used open source data mining systems, evaluate them, and discuss their advantages and disadvantages.

3.1 Open Source Systems

Open source software provides users with the freedom to run, copy, distribute, study, change and improve the software (see [16] for a detailed definition). To adopt open source software (or in fact any software) we must understand its license and the limitations that it places on us. Unlike closed source licenses which aim to limit your rights, open source software aims to give you the right to do whatever you please with the software. The common open sources licenses include the GPL, LGPL, BSD, NPL, and MPL. Bruce has discussed each license with suggestions for choosing a proper license [17].

In the past decades, open source products such as GNU/Linux, Apache, BSD, MySQL, and OpenOffice have achieved great success, clearly demonstrating that open source software can be as robust, or even more robust, than commercial and closed source software. Using open source software generally also means saving on the software costs, and allowing an enterprise to instead invest in skilling its people. Wang and Wang [18] have suggested many criteria for adopting open source software.

In the next subsection we discuss some important features to categorize open source data mining systems.

3.2 Important Features of Open Source Data Mining Systems

Open source systems are diverse in design and implementation. Although developed for data mining, they are very different in many aspects. To understand the characteristics of these diverse open source data mining systems and to evaluate them, we look into the following important features:

- **Ability to access various data sources**. Data comes from databases, data warehouses, and flat files in different formats. A good system will easily access different data sources.
- **Data preprocessing capability**. Preprocessing occupies a large proportion of the time in a data mining process [19]. Data preparation is often the key to solving the problem. A good system should provide various data preprocessing functions tasks easily and efficiently.
- **Integration of different techniques**. There is no single best technique suitable for all data mining problems. A good data mining system will integrate different techniques (preprocessing functions and modelling algorithms), providing easy access to a wide range of different techniques for different problems.
- **Ability to operate on large datasets**. Commercial data mining system, such as SAS Enterprise, can operate on very large datasets. This is also very important for open source data mining systems, so scalability is a key characteristic.
- **Good data and model visualization**. Experts and novices alike need to investigate the data and understand the models created.

- **Extensibility**. With new techniques and algorithms it is very important for open source data mining systems to provide an architecture that allows incorporation of new methods with little effort. Good extensibility means easy integration of new methods.
- **Interoperability with other systems**. Open standards means that systems (whether open or closed source) can interoperate. Interoperability includes data and model exchange. A good system will provide support of canonical standards, such as CWM [20] and PMML [21].
- **Active development community**. An active development community will make sure that the system is maintained and updated regularly.

These features categorize open source data mining systems. In the following, we investigate commonly used open source data mining systems with respect to these features.

3.3 Survey of Open Source Data Mining Systems

In this work we investigated 12 open source data mining systems accessible from the Internet. We studied these systems from four aspects: general characteristics, data source accessibility, data mining functionality, and usability. The survey results are summarized in the following four tables.

General Characteristics
We consider some general system features, including Activity, License, Programming Language and Operating Systems. The results are listed in Table 1.

The activity is measured by the frequency of updates and time of latest update. In terms of the listed operating systems, it is noteworthy that open source data mining systems that run on GNU/Linux will also run on Unix in general, and specific other Unix variances such as Solaris and HPUX.

Data Source Aspect
In real world applications data comes from different sources in different formats. The ability to access different data formats is important in selecting an open source system. In Table 2 we list 11 commonly encountered data sources and identify which systems are able to access which data sources. An indication of the data size that the system is know to be able to deal with is also provided.

Functionality Aspect
To be able to solve different data mining problems, the functionality of an open source data mining system is an important feature. Table 3 lists the summary of functionality of the 12 systems. The functions can be divided into six groups: data preprocessing, classification and prediction, clustering, association rules, evaluation, and visualization. For data preprocessing, we score each according to their capability, with 3 as the highest score, and 0 meaning not supported. We include the basic data mining tools that are offered by the commercial closed source systems, decision trees, neural networks, kmeans clustering, and association rules, but also the more modern techniques of support vector machines

Table 1. General characteristics of the 12 systems.

Product	**Activity**	**License**	**Language**	***Linux***	***Mac***	***Windows***
ADAM [22]	Medium	Unknown	Python	x	x	x
AlphaMiner [23]	High	GPL	Java	x	x	x
Databionic ESOM [24]	High	GPL	Java	x	x	x
Gnome Data Miner [25]	Low	GPL	C++ Python	x	x	x
KNIME [26]	High	Other	Java	x	x	x
Mining Mart [27]	High	Unknown	Java	x	x	x
MLC++[28]	Low	Other	C++	x	.	x
Orange [29]	High	GPL	C++ Python	x	x	x
Rattle [30]	High	GPL	R	x	x	x
TANAGRA [31]	High	Other	C++	.	.	x
Weka [3]	High	GPL	Java	x	x	x
YALE [32]	High	GPL	Java	x	x	x

Table 2. Data source characteristics of 12 systems.

Product	***Oracle***	***Sybase***	***SQLServ***	***MySQL***	***Access***	***ODBC***	***JDBC***	***ARFF***	***CSV***	***Excel***	**Data Size**
ADAM [22]	.	.	.	.	.	.	.	x	.	.	Large
AlphaMiner [23]	.	.	.	.	x	x	.	x	x	x	Medium
Databionic ESOM [24]	.	.	.	.	.	.	.	.	.	.	Medium
Gnome Data Miner [25]	.	.	.	.	.	.	.	.	x	.	Medium
KNIME [26]	.	.	.	x	x	x	x	x	x	.	Medium
Mining Mart [27]	x	.	.	.	x	.	.	.	.	.	Unknown
MLC++ [28]	.	.	.	.	.	.	.	.	.	.	Large
Orange [29]	.	.	.	x	.	.	.	.	.	.	Medium
Rattle [30]	.	.	.	x	x	x	.	.	x	x	Large
TANAGRA [31]	.	.	.	.	.	.	.	x	.	x	Medium
Weka [3]	.	.	.	.	.	.	x	x	x	.	Medium
YALE [32]	x	x	x	x	.	.	x	x	x	x	Medium

(SVM), boosting, and random forests. It is clear that the current open source data mining systems already include the commonly used data mining functions. The difference lies primarily in visualization capabilities.

Table 3. Functionality of the 12 systems.

Product	***Preprocess***	***Bayes***	***DTree***	***NNet***	***SVM***	***Boosting***	***Forests***	***KMeans***	***Associations***	***Evaluation***	***Data Vis***	***Model Vis***
ADAM [22]	3	x	x	x	.	.	.	x	x	x	3	3
AlphaMiner [23]	3	x	x	x	x	x	x	x	x	x	3	3
Databionic ESOM [24]	1	.	.	.	.	.	.	.	.	x	3	3
Gnome Data Miner [25]	0	x	x	.	.	.	.	.	.	x	1	1
KNIME [26]	3	x	x	x	x	x	x	x	x	x	3	3
Mining Mart [27]	3	.	.	.	.	.	.	.	.	x	3	3
MLC++ [28]	3	x	x	x	.	.	.	.	.	x	3	3
Orange [29]	3	x	x	.	x	x	x	.	x	x	3	3
Rattle [30]	2	x	x	x	x	x	x	x	x	x	3	3
TANAGRA [31]	3	x	x	x	x	x	.	x	x	x	1	1
Weka [3]	3	x	x	x	x	x	x	x	x	x	3	3
YALE [32]	3	x	x	x	x	x	x	x	x	x	3	3

Usability Aspect

The usability aspect describes how easy an open source data mining system can be used in solving real world business problems in different data and system environments. Here, we consider human interaction, interoperability and extensibility.

Human Interaction indicates how much interaction is required with the discovery process. Autonomous indicates that the system requires no tuning, you simply load the data and it builds the model. Guided indicates that the system provides significant assistance with the process, and Manual indicates very little guidance in the data mining process.

Interoperability essentially indicates whether PMML (the Predictive Modelling Markup Language) is supported, or whether the system is only internally "interoperable".

Clearly, there are many differences in usability in the current open source data mining systems.

Table 4. Usability aspect of 12 systems

Product	Human Interaction	Interoperability	Extensibility
ADAM [22]	Autonomous	Self	Simple
AlphaMiner [23]	Manual	PMML	Excellent
Databionic ESOM [24]	Manual	Self	None
Gnome Data Miner [25]	Guided	Self	Simple
KNIME [26]	Manual	PMML	Excellent
Mining Mart [27]	Manual	Self	Simple
MLC++ [28]	Guided	Self	Simple
Orange [29]	Manual	Self	Excellent
Rattle [30]	Guided	PMML	Simple
TANAGRA [31]	Manual	Self	Simple
Weka [3]	Manual	Self	Excellent
YALE [32]	Manual	Self	Excellent

3.4 Evaluation

From the surveyed characteristics of current open source data mining systems, we are able to evaluate these systems. We have selected 14 characteristics from the 4 aspects and assigned a weight to each characteristic as given in Table 5. We remark that the weight was assigned subjectively according to our own belief and experience with data mining in practise, and specifically from a business point of view.

According to the weight value in each characteristic and the surveyed results of the 12 systems, we calculated a score for each system in each aspect. Figures 1(a), 1(b), 1(c) and 1(d) plot the scores of the 12 systems, and the overall scores of the 12 systems are shown in Figure 1(e).

It must be noted here that we lack some of the background information for Mining Mart and MLC++ which has an impact on figures 1(b) and 1(e).

From the overall scores in figure 1(e), we can see that KNIME, AlphaMiner, Weka, Rattle and YALE are good open source data mining systems. A Google search also confirms that Weka, YALE, AlphaMiner and Rattle are popular (primarily because of their longer history). KNIME is a new entry focused on bioinformatics and has good performance.

Based on the above analysis, we summarize the advantages of open source data mining systems as follows:

1. **Supporting multi-platform**. GNU/Linux, Mac, and MS/Windows are supported by almost all systems.
2. **Good human interaction**. Almost all these systems support some degree of human guided discovery process, and more than half of them offer workflow style processes. AlphaMiner, KNIME, Mining Mart, Orange, Weka, and YALE provide drag-and-drop style case constructing.

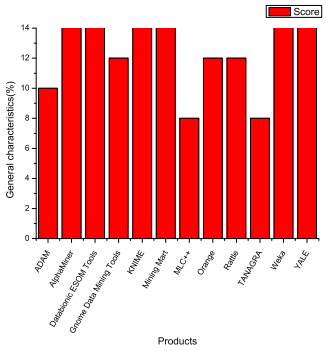
Table 5. The weighted distributions of characteristics for evaluation.

Perspective	**Characteristic**	**Weight**(%)
General	Activity	9
	Operating Systems	6
Data Source	Data sources	7
	Size of Data	12
Functionality	Preprocessing	12
	Classification/Prediction	9
	Clustering	3
	Associations	3
	Evaluation	4
	Data Visuals	3
	Model Visuals	3
Usability	Human Interaction	9
	Interoperability	10
	Extensibility	10

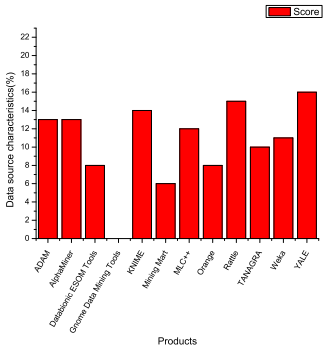
3. **Plentiful algorithms**. Each system has integrated plentiful algorithms for data preprocessing and modeling. Some contain more than 100 algorithms. However, this often actually works against the usefulness of the tool, as it confuses the business user.
4. **Reuse of cases**. All of them can reuse the cases. AlphaMiner, KNIME, and Rattle can export models in PMML format, which means that they can share models with other PMML compliant applications.
5. **Simple extensibility**. Most systems can integrate new operations in the form of plug-ins, some of which can be called in scripts.

There are also some serious disadvantages in these systems:

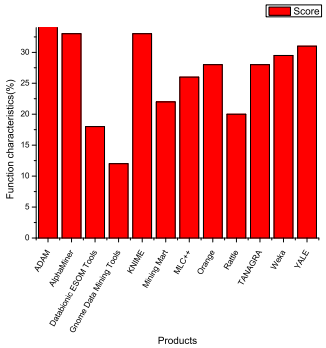
1. **Lack of support for diverse data resources**. YALE, KNIME, AlphaMiner, and Rattle support most file formats and relational databases through ODBC and JDBC . Some systems only focus on a smaller number of file formats.
2. **Difficulty with large volumes of data**. Real world applications often deliver very large datasets, with millions of rows. Most open source data mining systems are great for demonstrating and comparing the range of algorithms on smaller datasets but there has not been a focus on dealing with very large datasets. This limitation hinders the use of open source data mining systems in real applications, especially in business.
3. **Poor documentations and services**. During the investigation, we found that most of these systems have poor documentations, of limited use for a novice user wanting to quickly master the systems. Also support services is a serious issue, relying on the good will of a community of users.



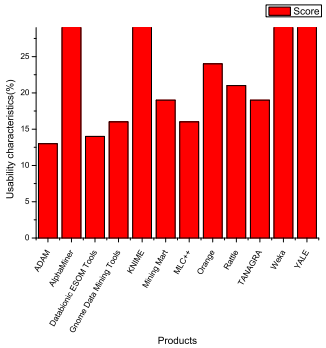
(a) General characteristics



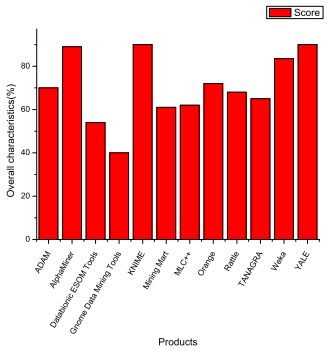
(b) Data source characteristics



(c) Functionality characteristics



(d) Usability characteristics



(e) Overall characteristics

Fig. 1. Score of 12 systems.

4 Conclusions

In this paper we have presented important features for open source data mining systems, and use these criteria to investigate 12 of the most commonly used open source data mining systems. Based on the analysis, we find that most systems have excellent functionality, and offer powerful tools for users in education and researchers. Bur for commercial use, there is still quite some effort in making the tools accessible.

We present the four following important points for open source data mining systems to gain greater success in deployment:

1. **Supporting various data sources**
   A good open source data mining system should support most commonly used data sources, such as the open source and commercial databases, csv files, and user defined formats.
2. **Providing high performance data mining**
   Most open source data mining system can't operate on large volumes of data. To offer high performance data mining we need to either rewrite the algorithms (e.g. parallel and distributed algorithms) or more simply to improve the hardware on which the software is running.
3. **Proving more domain-specific techniques**
   Most data mining system integrate many algorithms at the whim of the researchers, rather than for the benefit of business. We need to better identify algorithms that match the data to be processed. One approach will be to provide domain-specific techniques based on a generic platform.
4. **Better support for business application**
   Real business application are complex, placing many demands on the data mining system. Open source data mining systems need to in improve scalability, reliability, recoverability, and security [33].

References

1. Free Software Foundation: The GNU project (2007) Website: `http://www.gnu.org`.
2. DuBois, P.: MySQL. Sams (2005)
3. University of Waikato, New Zealand: Weka 3.4.9 (2006) Website: `http://www.cs.waikato.ac.nz/ml/Weka/index.html`.
4. Han, J., Kamber, M.: Data Mining: Concepts and Techniques. Morgan Kaufmann (2000)
5. Adomavicius, G., Tuzhilin, A.: Using data mining methods to build customer profiles. Computer (2001)
6. Bounsaythip, C., Rinta, E.: Overview of data mining for customer behavior modeling. Technical report, VTT Information Technology (2001)
7. Ling, C.X., Li, C.: Data mining for direct marketing: Problems and solutions. American Association for Artificial Intelligence (1998)
8. Rygielski, C., Wang, J.C., C, D.: Data mining techniques for customer relationship management. Technology in Society **24** (2002) 483–502

9. Apte, C., Liu, B., Pednault, E.P.D., Smyth, P.: Business applications of data mining. Communications of the ACM **45** (2002) 49–53
10. Ahmed, S.R.: Applications of data mining in retail business. In: Proceedings of the International Conference on Information Technology: Coding and Computing. (2004)
11. Kovalerchuk, B., Vityaev, E.: Data Mining in finance: Advances in Relational and Hybrid Methods. Kluwer Academic Publishers (2000)
12. Han, J., Altman, R.B., Kumar, V., Mannila, H., Pregibon, D.: Emerging scientific applications in data mining. Communications of the ACM **45** (2002) 54–58
13. Grossman, R., Kamath, C., Kegelmeyer, P., Kumar, V., Namburu, R.: Data Mining for Scientific and Engineering Applications. Kluwer Academic Publishers (2001)
14. Huang, J.: Data mining overview. Technical report, E-Business Technology Institute (2006)
15. Goebel, M., Gruenwald, L.: A survey of data mining and knowledge discovery software tools. In: SIGKDD Explorations. Volume 1., ACM SIGKDD (1999) 20–33
16. Open Source Initiative: The open source definition (2007) Website: `http://www.opensource.org/docs/definition_plain.html`.
17. Perens, B.: The open source definition (2007) Website: `Onlineathttp://perens.com/Articles/OSD.html`.
18. Wang, H., Wang, C.: Open source software adoption: A status report. IEEE SOFTWARE (2001)
19. D, P.: Data Preparation for Data Mining. Morgan Kaufman (1999)
20. Object Management Group: Common warehouse metamodel (cwm) (2007) Website: `http://www.omg.org/cwm/`.
21. Data Mining Group: Predictive model markup language (pmml) (2005)
22. Information Technology and Systems Center (ITSC) at the University of Alabama in Huntsville: Algorithm development and mining system (2005) Website: `http://datamining.itsc.uah.edu/adam/`.
23. HIT-HKU BI Lab: Alphaminer 2.0 (2006) Website: `http://bi.hitsz.edu.cn/AlphaMiner/`.
24. Data Bionics Research Group, University of Marburg: Databionic esom tools (2006) Website: `Onlineathttp://databionic-esom.sourceforge.net/`.
25. Williams, G.J.: Gnome data mining tools (2006) Website: `http://www.togaware.com/datamining/gdatamine/`.
26. Chair for Bioinformatics and Information Mining, University of Konstanz, Germany: Knime 1.1.2 (2006) Website: `http://www.knime.org/`.
27. MiningMartResearch Team: Mining mart 1.1 (2006) Website: `http://mmart.cs.uni-dortmund.de/`.
28. Stanford: Mlc++ (1997) Website: `http://www.sgi.com/tech/mlc/`.
29. Artificial Intelligence Laboratory, University of Ljubljana, Slovenia: Orange 0.9.64 (2007) Website: `http://www.ailab.si/orange/`.
30. Williams, G.J.: Rattle 2.1.116 (2006) Website: `http://Rattle.togaware.com/`.
31. Ricco RAKOTOMALALA, University Lyon, France: Tanagra 1.4.12 (2006) Website: `http://chirouble.univ-lyon2.fr/~ricco/tanagra/en/tanagra.html`.
32. Artificial Intelligence Unit, University of Dortmund, Germany: Yale 3.4 (2006) Website: `http://rapid-i.com/`.
33. Kleissner, C.: Data mining for the enterprise. In: In Proceeding of the 31st Annual Hawaii International Conference on System Science. (1998) 295–304