Global Advanced Research Journals

Global Advanced Research Journal of Engineering, Technology and Innovation (ISSN: 2315-5124) Vol. 1(6) pp. 131-137, September, 2012
Available online http://garj.org/garjeti/index.htm

Full Length Research Paper

Methodology and tool selection criteria in data mining

Ogwueleka, Francisca Nonyelum and Okeke, Georgina Nkolika

Department of Computer Science, Federal University of Technology, Minna, Niger State.
Email: nonnyraymond@yahoo.co.uk

Accepted 20 August 2012

The application of data mining algorithms requires the use of powerful software tools. Data mining and decision support software is expensive and selection of the wrong tools can be costly both in terms of wasted money and time lost. One of the most difficult tasks in the whole data mining process is to choose the right data mining tool (software), as data mining is evolving and maturing and many organizations are incorporating this technology into their business practices; the number of available tools continues to grow, the selection of the most suitable tool becomes increasingly difficult. This paper proposes a methodology for selecting the best among the assortment of commercially available data mining software tools.

Keywords: Data mining, algorithm, methodology, data mining tools, knowledge discovery

INTRODUCTION

Data mining has a long history, with strong roots in statistics, artificial intelligence, machine learning, and database research (Fayyad, 1996; Smyth, 2000). 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 (Lovell, 1983). 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. KDD is defined as the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data (Fayyad, 1996). 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 support the complete KDD process and not just a single step (Lovell, 1983). As the KDD field matures, it is relevant to question which data mining software vendors are positioned to dominate the market. Meanwhile business users face the daunting task of deciding which tool best suits their needs and budgets. However, the cost of selecting an improper data-mining tool for a particular application is even more costly in terms of personnel resources, wasted time, and the potential for acting on spurious results.

The Center for Data Insight (CDI) at Northern Arizona University (NAU), United State, has proposed a methodology for selecting from among the assortment of commercially available data mining software tools. This methodology was based on firsthand experiences in data mining using commercial data sets from a variety of industries. The KDD studies that took place in the CDI involve a wide-variety of commercial software tools applied to business data from a variety of industries as well as scientific data (Ken, et al, 1999).

The remaining sections of this paper present a framework for the selection criteria, a methodology within

Table1. Performance Criteria table

Criteria	Description
Platform Variety	Does the software run on a wide-variety of computer platforms? More importantly, does it run on typical business user platforms?
Software Architecture	Does the software use client-server architecture or a stand-alone architecture? Does the user have a choice of architectures?
Heterogeneous Data Access	How well does the software interface with a variety of data sources (Relational Database Management System (RDBMS), Open Database Connectivity (ODBC), Common Object Request Broker Architecture (CORBA), etc)? Does it require any auxiliary software to do so? Is the interface seamless?
Data Size	How well does the software scale to large data sets? Is performance linear or exponential?
Efficiency	Does the software produce results in a reasonable amount of time relative to the data size, the limitations of the algorithm, and other variables?
Interoperability	Does the tool interface with other KDD support tools easily? If so, does it use a standard architecture such as CORBA or some other proprietary API?
Robustness	Does the tool run consistently without crashing? If the tool cannot handle a data mining analysis, does it fail early or when the analysis appears to be nearly complete? Does the tool require monitoring and intervention or can it be left to run on its own?

Source: Ken, et al (1999: 2)

which the selection is to be applied and review of case studies of the CDI experience with this methodology.

REVIEW OF RELATED STUDIES

There are not many works in the literature addressing the issue of choosing the right tool for data mining tasks, one of the available works was done by Charest and Delisle (2006). Karina et al (2010) was of the opinion that two parameters are essential for choosing the right tool with the right techniques which include main goal of the problem to be solved and the structure of the available data. Ralf and Markus (2011) proposed criteria for tool categorization based on different user groups, data structures, data mining tasks and methods, visualization and interaction styles, import and export options for data and models, platforms, and license policies. These criteria are then used to classify data mining tools into nine different types.

Abdullah et al (2011) has conducted a comparative study between a number of some of the free available data mining and knowledge discovery tools and software packages. The results showed that the performance of the tools for the classification task is affected by the kind of dataset used and by the way the classification algorithms were implemented within the toolkits. For the applicability issue, the WEKA toolkit achieved the highest applicability followed by Orange, Tanagra, and KNIME respectively. Finally; WEKA toolkit achieved the highest improvement in classification performance; when moving from the percentage split test mode to the Cross Validation test mode, followed by Orange, KNIME and finally Tanagra respectively.

Nakhaeizadeh and Schnabl (1997) suggested a multi-criterion based metrics that was used as comparators for an objective evaluation of data mining algorithms (DM-algorithms). Each DM-algorithm was characterized generally by some positive and negative properties when it was applied to certain domains. Some of these properties are the accuracy rate, understandability, interpretability of the generated results and stability. Space and time complexity and maintenance costs were considered as negative properties. By then, there was no methodology to consider all of these properties, simultaneously, and use them for a comprehensive evaluation of DM-algorithms. Most of available studies in literature use only the accuracy rate as a unique criterion to compare the performance of DM algorithms and ignore the other properties. This suggested approach, however, took into account all available positive and negative characteristics of DM-algorithms and combined them to construct a unique evaluation metric.

SELECTION CRITERIA FRAME WORK

The CDI experience suggests four categories of criteria

Table 2. Functionality Criteria table

Criteria	Description
Algorithmic Variety	Does the software provide an adequate variety of mining techniques and algorithms including neural networks, rule induction, decision trees, clustering, etc.?
Prescribed Methodology	Does the software aid the user by presenting a sound, step-by-step mining methodology to help avoid spurious results?
Model Validation	Does the tool support model validation in addition to model creation? Does the tool encourage validation as part of the methodology?
Data Type Flexibility	Does the implementation of the supported algorithms handle a wide-variety of data types, continuous data without binning, etc.?
Algorithm Modifiability	Does the user have the ability to modify and fine-tune the modeling algorithms?
Data Sampling	Does the tool allow random sampling of data for predictive modeling?
Reporting	Are the results of a mining analysis reported in a variety of ways? Does the tool provide summary results as well as detailed results? Does the tool select actual data records that fit a target profile?
Model Exporting	After a model is validated does the tool provide a variety of ways to export the tool for ongoing use (e.g., C program, SQL, etc.)?

Source: Ken, et al (1999: 2)

for evaluating data mining tools: performance, functionality, usability, and support of ancillary activities. This experience is supported in the relatively small body of literature related to software tool evaluation (Morril, 1998; Andriaans and Zantinge, 1996).

Performance

Hardware configuration has a major impact on tool performance from a computational perspective. Furthermore, some data mining algorithms are inherently more efficient than others (Nakhaeizadeh and Schnabl, 1997). This category focuses on the qualitative aspects of a tool's ability to easily handle data under a variety of circumstances rather than on performance variables that are driven by hardware configurations and/or inherent algorithmic characteristics. The performance criteria table is shown in table 1.

Functionality

Software functionality helps assess how well the tool will adapt to different data mining problem domains as discussed in table 2

Usability

Is accommodation of different levels and types of users without loss of functionality or usefulness. One problem with easy-to use mining tools is their potential misuse. Not only should a tool be easily learned, it should help guide the user toward proper data mining rather than "data dredging". KDD is a highly iterative process. Practitioners typically adjust modeling variables to generate more valid models. A good tool will provide meaningful diagnostics to help debug problems and improve the output. The usability criteria table description is shown in table 3.

Ancillary Task Support

Allows the user to perform the variety of data cleansing, manipulation, transformation, visualization and other tasks that support data mining. These tasks include data selection, cleansing, enrichment, value substitution, data filtering, binning of continuous data, generating derived variables, randomizing, deleting records, as shown in table 4. Since it is rare that a data set is truly clean and ready for mining, the practitioner must be able to easily fine-tune the data for the model building phase of the KDD process.

METHODOLOGICAL STEPS

This criteria work best within a larger assessment model. Using standard decision matrix concepts (Urich and Eppinger, 1995), the methodology consists of the following steps:

1) Tool prescreening

Table 3. Usability Criteria table

Criteria	Description
User Interface	Is the user interface easy to navigate and uncomplicated? Does the interface present results in a meaningful way?
Learning Curve	Is the tool easy to learn? Is the tool easy to use correctly?
User Types	Is the tool designed for beginning, intermediate, advanced users or a combination of user types? How well suited is the tool for its target user type? How easy is the tool for analysts to use? How easy is the tool for business (end) users to use?
Data Visualization	How well does the tool present the data? How well does the tool present the modeling results? Are there a variety of graphical methods used to communicate information?
Error Reporting	How meaningful is the error reporting? How well do error messages help the user debug problems? How well does the tool accommodate errors or spurious model building?
Action History	Does the tool maintain a history of actions taken in the mining process? Can the user modify parts of this history and re-execute the script?
Domain Variety	Can the tool be used in a variety of different industries to help solve a variety of different kinds of business problems? How well does the tool focus on one problem domain? How well does it focus on a variety of domains?

Source: Ken, et al (1999: 3)

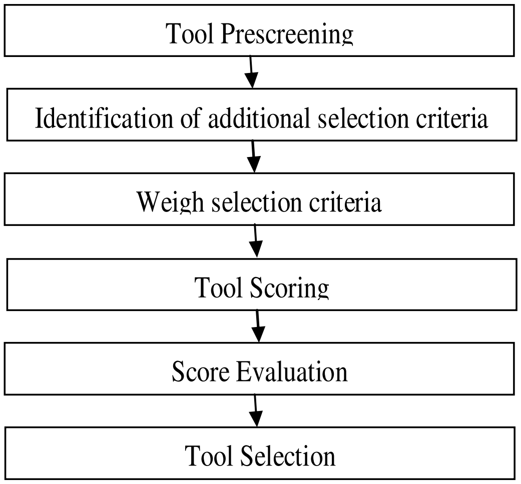

Figure 1. Methodological Steps

2) Identification of Additional Selection Criteria
3) Weight Selection Criteria
4) Tool Scoring
5) Score Evaluation
6) Tool Selection

The methodology steps are shown on figure 1

The methods applied at each phase are discussed under the six steps and showed how the criteria were

Table 4. Ancillary Task Support Criteria table

Criteria	Description
Data Cleansing	How well does the tool allow the user to modify spurious values in the data set or perform other data cleansing operations?
Value Substitution	Does the tool allow global substitution of one data value with another (e.g., replacing ‘M’ or ‘F’ with 1 or 0 for uniformity)?
Data Filtering	Does the tool allow the selection of subsets of the data based on user-defined selection criteria?
Binning	Does the tool allow the binning of continuous data to improve modeling efficiency? Does the tool require continuous data to be binned or is this decision left to user discretion?
Deriving Attributes	Does the tool allow the creation of derived attributes based on the inherent attributes? Is there a wide-variety of methods available for deriving attributes (e.g. statistical functions, mathematical functions, boolean functions, etc.)?
Randomization	Does the tool allow randomization of data prior to model building? How effective is the randomization? How efficient is the randomization?
Record Deletion	Does the tool allow the deletion of entire records that may be incomplete or may bias the modeling results in some way? Does the tool allow the deletion of records from entire segments of the population? If so, does the tool allow these records to be easily reintroduced later if necessary?
Handling Blanks	Does the tool handle blanks well? Does the tool allow blanks to be substituted with a variety of derived values (e.g., mean, median, etc.)? Does the tool allow blanks to be substituted with a user-defined value? If so, can this be done globally as well as value-by-value?
Metadata Manipulation	Does the tool present the user with data descriptions, types, categorical codes, formulae for deriving attributes, etc.? If so, does the tool allow the user to manipulate this metadata?
Result Feedback	Does the tool allow the results from a mining analysis to be fed back into another analysis for further model building?

Source: Ken, et al (1999: 3)

used to support the methodology. During these steps, a selection matrix was developed to aid in scoring and selecting the best tool.

Step 1: Tool Prescreening

This step is aimed at reducing the set of tools being considered to a manageable number. Eliminating tools that clearly will not be selected due to rigid constraints of the organization or the tool vendor does this. As an example, if the organization has already made the decision that decision support software must run on a Unix server, then any non-Unix tools can be eliminated (Ken et al, 1999).

Step 2: Identification of Additional Selection Criteria

Unfortunately evaluating data mining tools is not simply a matter of selecting the best tool for all purposes. Instead an organization must consider the tools with respect to their particular environment, and analysis needs. While the evaluation framework provides most of the technical criteria for selection, the aim of this step is to identify any additional criteria that are specific to a particular organization. Software cost is usually considered during this step in addition to such things as platform restrictions, end-user abilities, specific data mining projects, etc. Additionally, it is during this step that framework criteria are examined and irrelevant items are discarded if necessary (Ken et al, 1999).

Table 5. Weighting selection

Criteria	Weight	Tool A	Tool B	Tool C
Performance Platform Variety	.05			
Software Architecture	.05			
Heterogeneous Data Access	.10			
Data Size	.40			
Efficiency	.15			
Interoperability	.05			
Robustness	.20			

Source: Ken, et al (1999:2)

Table 6. Tool Scoring/Selection table

Criteria	Weight	Tool A (reference)		Tool B		Tool C	
Performance (.30)		Rating	Score	Rating	Score	Rating	Score
Platform Variety.	0.5	3	.15	3	.15	4	.20
Software Architecture	0.5	3	.15	3	.15	5	.25
Heterogeneous Data Access	.10	3	.30	4	.40	4	.40
Data Size	.40	3	1.2	2	.80	4	1.6
Efficiency	.15	3	.45	2	.30	3	.45
Interoperability	.05	3	.15	3	.15	4	.20
Robustness	.20	3	.60	1	.20	5	1.00
Performance Score		**3.0**		**2.15**		**4.1**	
Functionality (.20)		Rating	Score	Rating	Score	Rating	Score
Mining Techniques	.15	3	.45	4	.60	3	.45
…							
Model Exporting	.00	3	.00	1	.00	2	.00
Functionality Score		**3.0**		**3.8**		**1.85**	
Usability (.30)		Rating	Score	Rating	Score	Rating	Score
User Interface	.00	3	.00	2	.00	3	.00
…							
Domain Variety	.25	3	.75	3	.75	5	1.25
Usability Score		**3.0**		**1.8**		**3.95**	
Ancilary Task Support(.10)		Rating	Score	Rating	Score	Rating	Score
Data Cleansing	.15	3	.45	4	.60	5	.75
…							
Result feed back	.05	3	.15	3	.15	4	.20
Ancilary Task Score		**3.0**		**4.7**		**4.25**	
Other Criteria (.10)							
…							
Weighted Average		**3.0**		**4.52**		**3.51**	

Source: Ken, et al (1999:4)

Step 3: Weight Selection Criteria

Following step 2 the evaluator has five categories of selection criteria. These include the four groups represented by the framework (performance, functionality, usability, and ancillary task support) plus an additional group of organization specific criteria identified in step 2. During step 3 the criteria within each category are assigned weights so that the total weight within each category equals 1.00 or 100%. An example of this is provided for the Performance category in Table 5. This weighting must be conducted with respect to the intended use of the software. Consider an organization whose data warehouse is centrally located on a Windows NT server, and whose local area network consists exclusively of Windows NT workstations. Such an organization will probably assign a low weight to platform variety since any other platforms on which the tool is supported do not matter (Ken et al, 1999).

Step 4: Tool Scoring

Once the criteria have been weighted with respect to a set of targeted needs, the tools can now be scored for comparison. Rather than scoring on some artificially absolute scale, scoring is done relative to a reference tool. Generally the evaluator is predisposed toward a favorite tool for a variety of subjective reasons. This "favorite" should be selected as the *reference tool*. Any tool may be selected in the absence of a favorite. The reference tool receives a score of 3 for each criterion (Ken et al, 1999). Other tools are then rated against the reference tool for each criterion using the following discrete rating scale:

Relative Performance	**Rating**
Much worse than the reference tool	1
Worse than the reference tool	2
Same as the reference tool	3
Better than the reference tool	4
Much better than the reference tool	5

Using this scheme a score is calculated for every criterion for each tool. These scores are then totaled to produce a score for each category. Finally, the categorical scores are combined in a weighted-average to calculate an overall tool score. By default each criteria category receives a weight of .20. However, adjusting these weights allows the evaluator to emphasize or deemphasize particular categories of criteria. Table 6 is a selection table for a partial example of tool scoring.

CONCLUSION

Experience with a variety of commercial tools and data sets has led to a data mining tool assessment framework and methodology for using the framework. The framework considers performance, functionality, usability, and ancillary task support to evaluate data mining tools. The assessment methodology takes advantage of decision matrix concepts to objectify an inherently subjective process. Furthermore, using a standard spreadsheet application this framework is easily automatable, thus rendering it easy and feasible to employ.

Data mining software is costly and generally accompanied by moderately steep learning curves. Selection of the wrong tool is expensive both in terms of wasted money and lost time. The methodology presented in this paper provides a means of avoiding the selection of an inappropriate tool. This framework should help practitioners avoid spending needless money only to discover that a particular tool does not provide the necessary solution. Furthermore, this methodology provides a method for publishing tool comparisons and evaluations in the literature.

REFERENCES

Abdullah H, Qasem A , Mohammed N, Emad M. (2011). " A Comparison Study between Data Mining Tools over some Classification Methods) International Journal of Advanced Computer Science and Applications" (IJACSA), Special Issue on Artificial Intelligence, Pg. 18 - 26

Adriaans P, Zantinge D (1996). " *Data Mining"*, Addison- Wesley Longman, Harlow England.

Charest M, Delisle S (2006). "Ontology-guided intelligent data mining assistance:

Combining declarative and procedural knowledge", Artificial Intelligence and Soft Computing, pg. 9-14

Elder JF, Abbott DW (1998) "A Comparison of Leading Data Mining Tools", *Tutorial KDD-98*, AAAI Press, New York, Pg. 1-68

Fayyad U, Piatetsky-Shapiro G, Smyth S (1996). "Knowledge Discovery and Data Mining: Towards a Unifying Framework", *Proceedings KDD-96*, AAAI Press, Portland, Oregon, Pg. 82-88

Fayyad U, Piatetsky-Shapiro G, and Smyth P (1996) "From data mining to knowledge discovery in databases". pp 17: 37–54.

Karina G, Miquel S, Victor C (2010). "Choosing the Right Data Mining Technique: Classification of Methods and Intelligent Recommendation", a paper presented at *2010 International Congress on Environmental Modelling and Software Modelling for Environment's Sake, Fifth Biennial Meeting, Ottawa, Canada. pp*

Lovell MC (1983). "Data mining", *Rev Econ Stat* 1983, 65:1– 11

Morrill L (1998). "Enterprise Mining: More than a Tool", *Database Programming and Design*, 11(2): 1-10.

Morrill L (1998). "Intelligence in the Mix", *Database Programming and Design.* 11(3): 1- 15.

Nakhaeizadeh G, Schnabl A (1997) "Development of Multi-Criteria Metrics for Evaluation of Data Mining Algorithms", *Proceedings KDD-97*, AAAI Press, Newport Beach, California, Pg. 37-42

Piatetsky-Shapiro G, Brachman R, Khabaza T, Kloesgen W, and Simoudis E (1996). "An Overview of Issues in Developing Industrial Data Mining and Knowledge Discovery Applications", *Proceedings KDD-96*, AAAI Press, Portland, Oregon, Pg. 89-95

Ralf M, Markus R (2011). "Data Mining Tools", WIREs Data Mining and Knowledge Discovery, John Wiley & Sons, Inc, Germany, 1: 431 – 443.

Smyth P (2000). " Data mining: Data analysis on a grand scale?" *Stat Methods Med Res* 2000, 9:309–327.

Ulrich K, Eppinger S (1995). "Product Design and Development", McGraw-Hill, United States,