

Methodology for Selection of a Data Mining Tool

Author: Vikas Verma¹; Sunil Dhawan²;

Department of Computer Applications, Chandigarh Business School, Landran ¹

Research Scholar, Department of Computer Science, NIMS University, Jaipur²

ABSTRACT

In this paper, we describe the procedure for selection of a data mining tool for a software project. These days it is impossible or very difficult for any software developer to ignore the usage of data mining tools. Even some developers are using data mining data or information as the basis for the development of the software. Thus there is a high relevance factor for using the interpreted or mined data using these tools. So, it becomes necessary for any developer to select an appropriate tool for this purpose. As there are several tools available for the same. It involves many factors such as architecture, platform, usage etc. for the selection of the data mining tool. This paper presents a framework or methodology using which a data mining tool can be selected based on a set of characteristics.

General Terms

Data Mining Tool, Selection of Tools, Guidelines for Data Mining, Procedure for Selection.

Keywords

Data Mining Tools, Selection procedures, Methodology for selection of a tool.

1. INTRODUCTION

Software development is very complex process and involves many sub domains. The main architecture suggests that there are four primary phases. Analysis, Design, Code and Test. One of the most important phases is the Analysis phase as the requirements or the preliminary application architecture is created or designed. Requirement Analysis is done by analysing various factors using data from various sources. There is a common but critical problem is detecting and finding out that the tool deleted is inappropriate to do the objective of business. If you are doing this, you are wasting both time and money. We can understand that this process of selection isn't a common one and the organisation expects the return of the investment in a prudential time. The methodology proposed in this paper tries to organise the process of selection, so that the organisation would select the best tool based on its requirements. Both performance and economic metrics are used for decision making points. The selection procedure starts by analysing the requirements and risk factors associated.[3] Then we analyse the economic factors. Combing the best of the two scenarios we choose the

most appropriate tool. We design a framework for evaluating data mining tools based on multiple criteria.

2. Framework Evaluation

There are four primary concerned areas or categories using which we can evaluate and design a framework for selection of a data mining tool. Performance, Functionality, Usability and Support are four primary areas using which we could establish a framework. [3]

Performance: This is the first and most important area. It 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 or algorithm characteristics. The evaluation must be performed keeping in mind the hardware configurations as the configuration of the test bed has major impact over tool's score of performance. In an overall prospective this area of concern deals with the computational perspective .

Functionality: The second area is functionality which address the concern of capabilities of the tool. It includes a variety of capabilities, techniques, and methodologies for data mining. Software functionality helps assess how well the tool with adapt to different data mining problem, domains.

Usability: The third primary area discussed while building this framework is usability. Usability means that how many users can use without any loss in functionality or usefulness. Easy to use and misuse go together. A tool even if provides quick and easy understanding but should also focus on quality of data it is modelling.

Support: The fourth criteria is to perform various secondary functions required by Data mining process. These tasks include data selection, cleansing, enrichment, filtering, randomising and deletion etc. The process of normalisation of data can be considered for evaluation of the tool for support concerns because it is highly inappropriate to expect a clean dataset which has high normalised degree.[5]

3. Establishing Criteria

Based on the available resources and literature following are the criteria for the major or primary areas.

3.1. Measured Performance Criteria

This criteria deals with the computational performance and asks whether a tool is appropriate in terms of platform, architecture, data size and efficiency etc. This is mentioned in Table 1.

Criteria	Description
Platform	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?
Data Access	What software interface does it require?
Data Size	Does the software scale to large data sets?
Efficiency	Does the software produce results in a reasonable amount of time?
Interoperability	Does the tool interface with other tools or set of tools
Robustness	What is consistency of the tool? How often does it crash?

Table 1

3.2. Functionality Criteria

This criteria deals with various factors such as different capabilities, techniques and methodologies. All of these factors test the tool against the data mining problem so that we could know how well the tool would adapt under different circumstances. It also allows us to test the critical functionality of the tool with regards to the critical path for construction of algorithm of the data mining problem. The functionality criteria is mentioned in Table 2.

Criteria	Description
Variety	Does the software provide variety of mining techniques and algorithms to support decisions?
Methodology	Does the software aid the user by presenting a step by step mining methodology?
Validation	Does the tool support model validation in addition to model creation?
Data Type	Does the implementation of the supported algorithms handle variety of data types?
Modifiability	Does the user have the ability to modify and fine tune the algorithms?
Data Sampling	Does the tool allow random sampling of data for predictive modeling?
Reporting	Does the results of a mining analysis reported in different ways?
Model Exporting	Is it possible to export the model into different tools format such as excel or sql?

Table 2

3.3. Support Criteria

Support Criteria is used by us to follow the support criteria and allocation of support resources are measured using this criteria.[2] We use this criteria to build up attributes that contributes towards after support of a system. Attributes include Data Cleansing, Substitution, Filtering, Deletion etc. These attributes or criteria are mentioned in Table 3

Criteria	Description
Data Cleansing	Does the tool allow the user to modify spurious values in the data set or perform other operations designed for data cleansing?
Value Substitution	Does the tool allow global substitution of one or more data values?

Data Filtering	Does the tool allow the selection of subsets of the data based on user-selection criteria?
Randomisation	Does the tool allow randomisation of data prior to model building?
Deletion of Records	Does the tool allow the deletion of the entire or specific records?
Handling Blanks	Does the tool handle blanks to avoid data corruption?
Metadata Manipulation	Does the tool present the user with data descriptions, types?
Result Feedback	Does the tool allow the results from alining analysis?

Table 3

3.4 Usability Criteria

Usability Criteria is used by us to follow the usage criteria and ease of usage is measured using this criteria. We use this criteria to build up attributes that contributes towards usage of a system. Attributes include User Interface Learning Curve, User Types, Data Visualisation etc. These attributes or criteria are mentioned in Table 4.

Criteria	Description
User Interface	Does the interface presents results in a meaningful way?
Learning Curve	Is the tool easy to learn?
User Types	Is the tool designed for beginners intermediate, advanced users or a combination of user types?
Data Visualisation	How well does the tool present the data?
Error Reporting	How meaningful is the error reporting?
Action History	Does the tool maintain a history of actions taken in the mining process?

Domain vAriety	Can the tool be used in a variety of applications and industries to solve different business problems?
----------------	--

Table 4

4. Application of Methodology

For application of the methodology a large assessment model is used as a base. We have used simple decision making concepts such a decision trees to formulate selection. This methodology consists of following phases:

- a. **Tool Prescreening:** This step is used to reduce the no. of tools to a number using which we can easily manage. The tools that don't satisfy the constraints set by the organisation. e.g. If the organisation has taken a decision about using mac tools then we should eliminate the non-mac tools. It is a simple yet valuable phase that helps us in wasting less amount of time.
- b. **Identification of Additional Selection Criteria:** It is not possible for a common tool to do take care of all the things related to data mining. The organisation must use a combination of the tools. Most of the time evaluation framework provides the technical criteria for selection, the aim of this phase is to provide additional criteria that is specific to a particular scenario. Attributes such as software cost, platform restrictions, end user abilities are used in this phase as additional criteria.
- c. **Weight for Criteria:** There are four different criteria used but all the criteria cannot have same value and usage. In this phase the criteria within each category are assigned weights so that the total weight is 100%. The assignment of weights must be consistent with industry practices or organisation's policy. An organisation whose data warehouse is centrally located on a Windows Server and clients consists of Windows Machines would assign a low weight to platform variety as they won't be using the multiple platforms.[4] The support for multiple platform doesn't matter in this scenario.
- d. **Tool Scoring:** Once the weights have been assigned to a set of needs, the tools can now be put for testing and recording their scores. The scoring must be done relative to a reference tool rather than using an absolute scale. A reference tool must be set for a variety of subjective reasons. According to the industry practices, the reference tool must have a score of 3 and the evaluations can have maximum score of 5. If the relative performance is low we can use 1 or 2 and if it equal to the reference tool we would give a score of 3. If the tool exceeds the performance constraints of reference tool it can be 4 or 5. Using this scheme scores are calculated for every criteria for each tool. These scores are

then totalled to produce a score for each category. The scores are then multiplied by the weight factor attached.

- e. **Scoring Evaluation:** The evaluation methodology is designed to objectify a subjective process. The intuition should be followed in order to get a clear picture in some circumstances. Discrepancies between scores and intuition are generally due to incorrect weightings of criteria. If such a case exists then the weights must be reassigned to the criteria which is considered to be inconsistent in providing the scores.

Thus this methodology provides an iterative approach and results are more consistent once you use the methodology again and again it becomes much more consistent.

5. Conclusion and Future work

A variety of tools exists in the data mining category and assessment of various tools led to creation of a methodology for selection of a tool. The assessment methodology takes the advantage of decision incepts to achieve the necessary precision scores.

In future we would work on the automation of the methodology using a spreadsheet package. Using a spreadsheet package makes it easy to deploy and would save a lot of money.

5. References

- [1] U. Fayyad, G. Piatetsky-Shapiro, and S. Smyth, "Knowledge Discovery and Data Mining: Towards a Unifying Framework", Proceedings KDD-96, AAAI Press, Portland, Oregon, August 2-4, 1996.
- [2] L. Morrill, "Enterprise Mining: More than a Tool", Database Programming and Design, vol. 11 no.2, February 1998.
- [3] L. Morrill, "Intelligence in the Mix", Database Programming and Design, vol. 11 no. 3, March 1998.
- [4] G. Nakhaeizadeh, and A. Schnabl, "Development of Multi-Criteria Metrics for Evaluation of Data Mining Algorithms", Proceedings KDD-97, AAAI Press, Newport Beach, California, August 14-17, 1997.
- [5] G. Piatetsky-Shapiro, R. Brachman, T. Khabaza, W. Kloesgen, and E. Simoudis, "An Overview of Issues in Developing Industrial Data Mining and Knowledge Discovery Applications", Proceedings KDD-96, AAAI Press, Portland, Oregon August 2-4, 1996.