Proceedings
V Ibero-American Symposium on Software Engineering.
Puebla (Mexico) Pp. 85-90

Tool Selection Methodology in Data Mining

Paola Britos, Hernán Merlino, Enrique Fernández, María Alejandra Ochoa, Eduardo Diez and Ramón García-Martinez

Software & Knowledge Engineering Center. Graduate School. Buenos Aires Institute of Technology
Intelligent Systems Laboratory. School of Engineering. University of Buenos Aires.
pbitos@itba.edu.ar

Abstract

A very important problem in the data mining process is detecting too late that the tool selected is inappropriate to do the objective of business. If this happens, we are wasting time and money. This paper presents a methodology that permits to select a tool for the process of data mining from a set of characteristics.

1. Introduction

The impact of a data-mining tool in the organization and, the investment of money involved in this process selection is important. As well we can understand that this process of selection isn't a common one and the organization expects the return of the investment in a prudential time. The methodology proposed in this paper tries to organize the process of selection, so the organization would select the best tool for its requirements, not only does this methodology select the tool from the economic point of view but it also analyzes the business requirements up. We have worked in, cycle of life selection [5], the activities selection of map activities in project software [2], expert system tool selection methodology [3], ERP selection methodology [1], but we don't have a methodology to select a tool of data mining. The existence of a lot of suppliers and some open source tools for the data mining process, without a methodology to evaluate the process of selection, can probably mean in the selection of an inappropriate tool, the consequences could be: (a) Time and money spend and (b) the increment of risk to achieve the business objectives.

2. Problem resolution

The following methodology proposed would permit the selection of the tool for data mining. This methodology has these phases: 1- documentation, 2- Requirements Analyzed, 3- Market Searching; 4- Suppliers Contact, 5- Suppliers candidate meeting and information recollection, 6- Report Built, 7- Suppliers Evaluation, 8- Product demonstration, 9- Final evaluation.

2.1. Phase 1: Documentation

First we define a frame of reference for the selection tool. The aspects to define are:

- Departments and function of the organization.
- Define the objective to be achieved with the tool.

2.2. Phase 2: Requirement Analyze

We need to document the business needs to be satisfied by the tool in relation to: the organization department, the business process and the budget. The objective of this phase is to obtain a preliminary group of suppliers. This phase only tries to describe the best data-mining tool for this organization, in other words we can say that, pay the right price and use it completely

2.3. Phase 3: Market Searching

The objective of this phase is to get the all the possible information from the suppliers in our market, we suggest to search in Internet, software expositions, magazines, bibliography and to contact with expert, from other companies. At the end of this phase we will provide a report of the supplies found.

2.4. Phase 4: Suppliers Contact

We will contact the suppliers selected from the list. First it is not necessary to generate a meeting; we only need get more information about the tool and suppliers. We suggest a list not longer than 5 suppliers, because we will do a deep study of each one. Some of the activities involved are: product demo, the suppliers meeting and report of each product.

2.5. Phase 5: Suppliers candidate meeting and information recollection

The objective of this phase is to arrange a meeting with each suppliers candidate and complete the missing information, so as to analyze the tools in the right way; we can compare component modules, functionality, and other implementations. The last step in this phase is information verification to make a comparison. We will provide a report with the supplier organization supplier's characteristic and tool characteristics.

2.6. Phase 6: Report writing

The objective of this phase is to make a report considering the main points of the business. This report is the base for future reports. We need to define a set of criteria and common comparison points to make a comparison and selection of tool. The reports with criteria can be used as model (see template in section 2.9 and 2.10), this should be adapted to the organization requirements and data miners. The report criterions are grouped in four categories:

- Technique and functional characteristics: we need to group all the tool characteristics.
- Suppliers characteristics: we include, supplies organization, growing and evolution, annual invoicing, location other clients and experience.
- Service characteristics: we include the specific aspects of supplier service.
- Economic characteristics: license cost and maintenance.

The steps to make the report are:

- Write the most important criteria for the organization.
- Divide the criterion in the above groups.
- Estimate the characteristics in function of impact in the group. The sum of the group is equal to 100, this is the sum of all criterions equal to 400 (see question proposed in section 2.9)

2.7. Phase: Suppliers evaluation

In this phase the team should arrange a new meeting with the suppliers, the team will request to the suppliers a technical and economic proposal with all the characteristics explained. We recommend to visit the to supplier's office. The team gives a grade between 1 and 4, in the column "Y" to make the report (the classification are 1=Bad, 2=Poor, 3=Good, 4=Excellent). Each value in the column "Y" should be multiplied by the factor in the column "Pond X" and the result should be write in column "W X*Y". The column "W X*Y" should be multiplied by the group value and divide by 100, with this the team would obtain the general ratio of the group. The team repeats this calculation with the other group in the template (see the example model). When finish the phase is finished the team would be able to make a product demo.

2.8. Phase 8: Product demonstration

In this phase the supplier demos her/his product and completes the questions from the team (see template evaluation product proposed section 2.10). The users in the demo meeting qualify each aspect in column "P" with a value between 0 and 5 (this values are explained in the header of template.). When the team finishes all the ratios, it should compare the rations between tools. The team will attach an evaluation to each one. When this step is finished the organization will have a complete report of each supplier, with information about supplier organization, products demonstrations, advantages and disadvantages and any important information to evaluate the supplier.

2.9. Phase 9: Final evaluation

The team compares the result (ratios + demo) and selects the product with the best evaluation.

2.10. Templates proposed

Header

Tool Name:	
Supplier:	
Evaluation: 1 = Bad, 2 = Poor, 3 = Good, 4 = Excellent	1=No, 4 = Yes

Questionnaire

Selection criterion	Description		Weigh X	Value Y	Pond X* Y
1. Functional-Techniques characteristics					
Methodology / Cycle of life reported	Methodology / Cycle of life supported by the tool for the data-mining process (CRISP-DM, SEMMA, etc.)		3		
Adaptability and flexibility to get data	From Database	Total format supported	8		
	From other source (word, excel, etc.)	Total format supported	8		
Integration	Support different technique of data mining		5		
Multi-language	Work with different languages		2		
Techniques supported	Quantity of techniques (neural networks, Bayesians networks, induction algorithmic, etc.)		18		
Report and visualization tools	Report and graphics generation		12		
Multiplatform	Run in multiplatform		5		
Remote installation	The administration and maintenance is remote o in site		5		
Multiples users	It has user profiles.		2		
Security	Profiles of security by users		2		
Backup	Methodology of backup		2		
Friendly	Users interface		10		
Configurations	Support profile configurations		8		
Documentation	Help and service support		5		
Connectivity	Support connectivity with: Internet, EDI, FTP, ERPs		2		
Message system support	Support sends information by e-mail, etc.		3		
TOTAL			100%	Z= ∑	
	Group weight		40%	P1 = Z * 0,40	

2.- Suppliers characteristics				
Suppliers characteristics	Background	30		
Growing	Future perspectives.	10		
Geographic location	Office location	30		
Implementations	Others implementation of the same tool	5		
	Contacts with other clients	5		
Confidence	No quantify criteria	20		
TOTAL		100%	Z = ∑	
	Group weight	25%	P2= Z * 0,25	

3.- Service characteristics				
Product guaranty	Duration and reaches	30		
Upgrade	It obligatory, supports old version.	20		
License	Reaches, post sell support, cost.	30		
Support	Helpdesk, time response, availability	20		
TOTAL		100%	Z = ∑	
	Group weight	20%	P3 = Z * 0,20	

4.- Economics characteristics				
Software Cost	Tool cost	30		
Hardware Cost	Hardware new or upgrade to run the tool.	20		
Other Software Cost	Other software (backup, web servers, database, etc)	20		
Licenses	License, policies	10		
Financing	Exist	10		
Upgrade	Average Cost	10		
TOTAL		100%	Z = ∑	
	Group weight	-15%	P4= Z * (- ,15)	

The economic factor is negative, in this way the products with low cost have an advantage to the high cost in the final result.

Aspects weigh		
P1: Technical and Functional	40%	
P2: Suppliers	25%	
P3: Service	20%	
P4: Price	- 15%	
TOTAL	100%	

Advantages and disadvantages	
Advantages	We should reserve an item to write possible advantage not contemplated
Disadvantages	We should reserve an item to write possible disadvantage not contemplated

2.11. Evaluation product templates

Header

Data miner name:
Date:
Supplier:
Weighing: 0 = No evaluate item 1 = Evaluate Item but it doesn't support 2 = Evaluate Item support, but not complete 3 = Evaluate Item supported but with modification 4 = Evaluate Item supported 5 = Evaluate Item supported and added value

Questionnaire

Criterion	W
Branches	
Multiplatform	
Multilanguage	
Help in company language	
Documentation in company languages	
Import data form external source	
Quantity of data mining techniques	
Integration	
Methodology and cycle of life	
Visualization and reports	
Security	
Global product appreciation	
Confidence	
Knowledge of product by supplier	
Quality service	
Helpdesk support	
General presentation	
TOTAL	

3. Use Case

3.1. Phase 1: Documentation

The organization is a meteorological station that pretends to calculate the velocity of the wind in a zone and determinate its reasons, in function of this the station has the capacity of calculate the quantity of eolic energy.

3.2. Phase 2: Requirements Analyze

The data-mining tool has to:

- Predict the speed of the wind; you can use backpropagation neural nets.
- Determine the rules of behaviour of the wind; you can use C5 algorithmic.
- Predict the quantity of eolic energy generating; you can use backpropagation neural nets.

3.3. Phase 3: Market Searching

The candidates are:

- Clementine, by SPSS.
- MatLab, by Mathworks.
- Weka, open source, by Department of Computer Science, University of Waikato, New Zealand.

3.4. Phase 4: Suppliers Contact

We get information about the suppliers:

- Clementine, by SPSS, www.spss.com.ar.
- MatLab, by Mathworks, www. mathworks. com.
- Weka, open source, by Department of Computer Science, University of Waikato, New Zealand, http://www.cs. waikato.ac.nz/ ~ml/weka/index.html.

3.5. Phase 5: Candidate suppliers meeting and information recollection

We could only arrange and interview with Clementine (SPSS), because this is the only supplier located near us. The others were only contacted by e-mail.

3.6. Phase 6: Report built

The criteria of evaluation is the same determine in the model.

3.7. Phase 7: Suppliers evaluate

Selection criteria		Description	W X	Clementine		MatLab		Weka	
				Valor Y	Wd X* Y	Valor Y	W X*Y	Valor Y	W X* Y
1. Functional-Techniques characteristics									
Methodology / Cycle of life reported		Methodology / Cycle of life supported by the tool for the data-mining process (CRISP-DM, SEMMA, etc.)	3	4	12	--	--	--	--
Adaptability and flexibility to get data	From Database	Total format supported	8	5	40	4	32	3	24
	From other source (word, excel, etc.)	Total format supported	8	5	40	4	32	3	24
Integration		Support deferent technique of data mining	5	3	15	3	15	1	5
Multi-language		Work with different idioms	2	--	--	--	--	--	--
Techniques supported		Quantity of techniques (neural networks, Bayesians networks, induction algorithmic, etc.)	18	5	90	5	90	5	90
Report and visualization Tools		Generate report and graphics	12	5	60	5	60	3	36
Multiplatform		Run in multiplatform	5	--	--	2	10	2	10
Remote installation		The administration and maintainability is remote o in site	5	3	15	NE	--	--	--
Multiples users		It has users profiles.	2	5	10	NE	--	--	--
Security		Profiles security by users	2	2	4	NE	--	--	--
Backup		Methodology of backup	2	--	--	NE		--	--
Friendly		Users interface	10	5	50	5	60	3	36
Configurations		Support profile configurations	8	--	--	--	--	--	--
Documentation		Help and service support	5	5	25	5	25	4	20
Connectivity		Support connectivity with: Internet, EDI, FTP, ERPs	2	5	10	NE	--	--	--
Message system support		Support sends information by e-mail, , etc.	3	--	--	NE	--	--	--
TOTAL			100%		371		324		245
		Group weight	40%	P1	148,4		129,6		98

Selection criteria	Description	W X	Clementine		MatLab		Weka	
			Valor Y	W X*Y	Valor Y	W X*Y	Valor Y	W X*Y
2.- Suppliers characteristics								
Suppliers characteristics	History, earns, employs.	30	5	150	5	150	--	--
Growing	Future perspectives.	10	5	50	5	50	3	30
Geographic location	Office location, near us	30	5	150	1	30	1	30
Implementations	Others implementation of the same tool	5	1	5	1	5	--	--
	Contacts with other clients	5	1	5	1	5	--	--
Confidence	No quantify criterion	20	5	100	5	100	3	60
TOTAL		100%		470		350		120
	Group weight	25%	P2	117,5		87,5		30

Selection criteria	Description	W X	Clementine		MatLab		Weka	
			Valor Y	W X*Y	Valor Y	W X*Y	Valor Y	W X*Y
3.- Service characteristics								
Product guaranty	Duration and reaches	30	4	120	NE	--	1	30
Upgrade	It obligatory, supports old version.	20	3	60	NE	--	2	40
License	Reaches, has post sell support, cost.	30	3	90	NE	--	--	--
Support	Helpdesk, time response, availability	20	4	80	NE	--	--	--
TOTAL		100%		350		NC		70
	Group weigh	20%	P3	70		NC		14

Selection criteria	Description	W X	Clementine		MatLab		Weka	
			Valor Y	W X*Y	Valor Y	W X*Y	Valor Y	W X*Y
4.- Economics characteristics								
Software Cost	Tool cost	30	1	30	NE	--	--	0
Hardware Cost	Hardware new or upgrade to run the tool	20	--	0	5	100	--	0
Other Software Cost	Other software (backup, web servers, database, etc)	20	--	0	5	100	--	0
Licenses	License politic	10	2	20	NE	--	--	0
Financing	Exist	10	3	30	NE	--	--	0
Upgrade	Average Cost	10	3	30	NE	--	--	0
TOTAL		100%		110		200		0
	Group weigh	- 15%		P4	-16,5	- 30		0

Aspects weigh				
		Clementine	MatLab	Weka
P1: Technical and Functional	40%	148,4	129,6	98,0
P2: Suppliers	25%	117,5	87,5	30,0
P3: Service	20%	70,0	NE	14,0
P4: Economic	- 15%	- 16,5	- 30,0	0,0
TOTAL	**100%**	**319,4**	**187,1**	**142**

Note: NE is "No evaluate".

Advantages and Disadvantages			
	Clementine	MatLab	Weka
Advantages	Local office. Excellent interface.	Excellent interface.	Open source.
Disadvantages	Cost	Not local office	Without support

3.8. Phase 8: Product demonstration

Data miner: Paola Britos			
Date: April de 2005			
Weighing: 0 = No evaluate item 1 = Evaluate Item but it doesn't support 2 = Evaluate Item support, but not complete 3 = Evaluate Item supported but with modification 4 = Evaluate Item supported 5 = Evaluate Item supported and added value			
Critera	Clementine	MatLab	Weka
Branches	4	4	1
Multiplatform	1	1	1
Multilanguage	1	1	1
Help in company idiom	1	1	1
Documentation in company idiom	1	1	1
Import data form external source	4	4	2
Quantity of data mining techniques	5	5	4
Integration	2	2	2
Methodology and cycle of life	2	1	1
Visualization and reports	2	5	3
Security	5	5	1
Global product appreciation	5	4	3
Confidence	4	4	3
Knowledge of product by supplier	5	5	4
Quality service	5	0	0
Helpdesk support	5	0	0
General presentation	5	0	0
TOTAL	57	43	29

3.9. Phase 9: Final evaluation

Product evaluation:

Criteria weigh:

Aspects weigh		Clementine	MatLab	Weka
P1: Technical and Functional	40%	148,4	129,6	98,0
P2: Suppliers	25%	117,5	87,5	30,0
P3: Service	20%	70,0	NE	14,0
P4: Economic	- 15%	- 16,5	- 30,0	0,0
TOTAL	100%	319,4	187,1	142

Tool Demo:

Clementine	MatLab	Weka
57	43	29

Total:

weigh	Clementine	MatLab	Weka
Criteria	319,4	187,1	142,0
Demo	57,0	43,0	29,0
TOTAL	**376,4**	**230,1**	**171**

4. Conclusions

We can say that, the methodology has:

- A framework to evaluate a tool.
- Generate metrics.

- It is flexible and adaptable
- We can evaluate in the right way a tool without losing the business objectives.

5. Biography

[1] Chiesa, F. 2004. *Metodología para Selección de Sistemas ERP*. Reportes Técnicos en Ingeniería del Software. (6)1:17-37.

[2] Diez, E., Britos, P., Rossi, B. y García-Martínez, R. 2003. *Generación Asistida del Mapa de Actividades de Proyectos de Desarrollo de Software*. Reportes Técnicos en Ingeniería del Software. (5)1:13-18.

[3] García Martínez, R. y Britos, P. 2004. *Ingeniería de Sistemas Expertos*. ISBN 987-1104-15-4. Editorial Nueva Librería. Buenos Aires.

[4] Métrica, 1990. *Metodología de planificación y desarrollo de sistemas de información MÉTRICA Versión 2*.1990.

[5] Rossi, B. 2001. *Sistema Experto de Ayuda para la Selección del Modelo de Ciclo de Vida.*Tesis de Master en Ingeniería del Conocimiento. Facultad de Informática. Universidad Politécnica de Madrid. http://www.itba.edu.ar/capis/webcapis/tesisdemagisterterminadas.htm. Pagina web vigente al 03/07/05.