

Cost Drivers of a Parametric Cost Estimation Model for Data Mining Projects (DMCOMO)

Oscar Marbán, Antonio de Amescua, Juan J. Cuadrado, Luis García
Universidad Carlos III de Madrid (UC3M)

Abstract

Data Mining is a research line that began in 1980 in order to find the knowledge that is hidden in the data that organizations are storing in a daily basis. This knowledge supports the decision-making processes in organizations. As a consequence companies of every kind have been developing data mining projects since the term appeared. However, there is no way to estimate this kind of projects.

Although there are many references to Data Mining algorithms in the bibliography, not many authors have dealt the problem from Software Engineering point of view.

CRISP-DM is a model process, from Software Engineering point of view, that appeared in 2000. CRISP-DM is the first standard of Data Mining projects development.

In the standard of software development model process, e.g. ISO 12207 and IEEE 1074, processes and tasks are proposed similar to those in CRISP-DM model. Nevertheless, in software development a lot of methods are described to estimate the costs of project development (SLIM, SEER-SEM, PRICE-S and COCOMO). These methods are not appropriate in the case of Data Mining projects because in Data Mining software development is not the first goal.

Some methods have been proposed to estimate some phases of a Data Mining project but there is no method to estimate the global cost of a generic Data Mining project. As a consequence, in this paper we propose the cost driver of a parametric estimation method for Data Mining projects.

Keywords

Data Mining, Software Engineering, Cost Estimation, Parametric Models

Introduction

In the last few years, the volume of data stored by the companies has grown until unsuspected level at high speed.

In that quantity of data we can discover hidden information impossible to detect by simple observation, however this hidden knowledge situated into data can be obtained with the help of data mining techniques [1], [2]. This is known as Knowledge Discovery in Databases (KDD).

These two facts, the new big databases and the development of KDD, have been that more and more data mining projects has been developed by different organisations.

However until the development of CRISP-DM a methodology to develop data mining projects not exists and only generic proposals over the phases of KDD were proposed.

CRISP-DM is a model which not only define the phases for a data-mining project, but that define the processes to by carry out develop each one of them, with its own inputs and outputs.

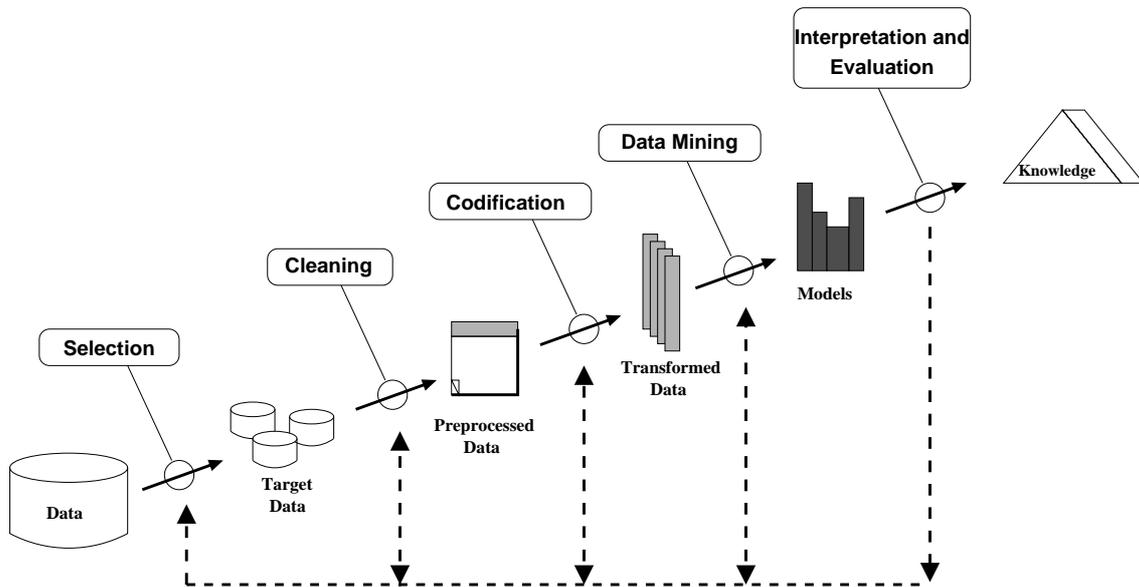


Figure 1: Data Mining phases [1]

The CRISP-DM standard proposes a process model for data-mining systems development similar to the standards ISO 12207 [7] and IEEE 1074 [8].

These process models propose a project management phase into which ever exists a task to be completed that consists in the estimation of the effort and time that the project will consume.

We could think that this estimation could be carried out apply different software cost estimation models like COCOMO II [9], SLIM [10], PRICE-S [11], or other. But at once we could find many difficulties when we try to apply them, because they are not design for other kind of projects. Projects whose products are software applications and not the generation of models with the knowledge extracted from the input data.

Some proposals of data mining projects cost estimation models have been published in the literature [12], [13], [14], [15], but either are built only for a specific kind of data mining technique, or only deal with a certain project phase.

The authors are working on the development of a new parametric mathematical model, Data Mining Cost Model (DMCOMO) that allows the cost estimation of an overall data mining project how a part of a research project on software metrics found by the Spanish ministry of science and technology.

In this paper part of the results obtained in this research, specifically a list of cost drivers for data mining projects is presented.

The next section discusses with some detail about some of the main proposal about data mining cost estimation models existing today.

The third section presents the cost drivers selected for the DMCOMO, its description and classification.

And finally, in the four section conclusions and future work are presented

Data Mining estimation methods

Current data mining projects estimation models are very specific and only deal with a certain project phase and once the project is in progress.

In [17] a taxonomy of the different cost types for inductive learning is presented. Turney distinguish nine kinds of cost:

- Cost of misclassification errors
- Cost of test
- Cost of teacher
- Cost of intervention
- Cost of unwanted achievements
- Cost of computation
- Cost of cases
- Human-Computer Interaction Cost

In [12], Domingos present a estimation model for data mining projects based on NPV (Net Present Value) [18]. That model can be used to go from the data mining model to its application in a real environment, but this estimation model only can be used with machine learning applications.

Moreover, this model can be only applied in an advanced phase of the project, and at that point a large sum of the project resources has been wasted. For this reason it can't be used in an early project phase.

In [13] an estimation method is proposed for classification problems. It is based on items misclassified and in the cost of the misclassification.

In [14] an estimation method is proposed for predictive models. The method is based on a business model and on the predicted behaviour.

In [15] a comparison of estimation algorithms based on the cost of misclassification is presented.

In [16], data-mining are described from the micro-economic point of view. In this work, games theory is used to maximize the value of patterns extracted form data.

So, we can concluded that we are using data-mining for 20 years, nowadays there is not an estimation method for data-mining projects. In this paper we propose the cost drivers for a parametric estimation method for generic data-mining projects. We call the method DMCOMO (Data Mining Cost Model).

DMCOMO cost drivers

Parametric mathematical cost estimation models use mathematical equations to obtain the value of a set of output dependent variables starting from the values given to another set of independent ones, and. The estimations with this kind of models usually have a two step process:

1. To do a first approximation or estimation which depends on the value of a reduced set of parameters whose weight in the final result is considered greater than the rest and is not normally related to the features of the project but the product.
2. Once finished this first step, the final result is determined using another set of variables that allow the estimation to be refined by introducing the specific characteristics of the application and development environment.

Mainly four factors affect the accuracy of the estimations done with them:

1. A precise definition of the equations to be used.
2. Constant refining of the parameters used. This involves not only adding or removing them to reflect changes in technology but also a thorough understanding of those selected.
3. An accurate calibration of the numerical values for each parameter rating levels [19].
4. Wise selection of the rating level for each parameter used for the selected model in order to calculate the estimations for the specific project.

Here we present the result of the research carried out to select the cost drivers that must be used by a parametric mathematical cost estimation model for data mining projects.

This selection has been done using a two round Delphi technique as was used in [20]

Taking into account the aspect of the project that describe, each cost driver has been classified in one of six groups:

1. Data

Group of cost drivers that take into account the data volume to handle in the data-mining project.

Likewise it is taking into account the attribute dispersion, that is, the number of different values that can be taken by each one of the attributes.

Also is considered the percentage of unknown or null values in the data and if the data models are available to be consulted by the participants in the data mining projects.

2. Models

Group of cost drivers that take into account features related with the data-mining model that will be used to carry out the project.

3. Platform

These cost drivers introduce into the model the features of the platform used to develop the project.

4. Tools

The tool driver allows the consideration of the role that in the use of CASE tools have in the amount of effort spent in the project.

5. Projects

This group has four drivers that describe different aspects related with the organizational structure used to develop the project. Aspects like the development site or the number of departments implied.

6. Staff

These cost drivers introduce into the model the influence that personal aspects like continuity, aptitude, experience or knowledge will have in the project effort.

Conclusions and Future Works

In this paper we propose a set of cost drivers for data mining projects cost estimation and a classification of the same in six groups.

These cost drivers has been selected by a consult to the experts with the help of a Delphi Technique and its main feature is that can be used by a parametric mathematical model to the overall data mining project cost estimation.

Future work includes the develop of a mathematical equation that use these cost drivers to estimate the cost of the projects, the selection of the cost drivers input values and the study of the possible correlation between some cost drivers in order to group them.

References

- [1] G. Piatetsky-Shapiro, W. Frawley, "Knowledge Discovery in Databases", AAAI/MIT Press, 1991
- [2] U. Fayyad, G. Piatetsky-Shapiro, P. Smith, R. Uthurusamy, "Advances un Knowledge Discovey and Data Mining", AAAI/MIT Press, 1996

- [3] G. Piatetsky-Shapiro, "An Overview of Knowledge Discovery in Databases: Recent Progress and Challenges", *Journal of Rough Sets, Fuzzy Sets and Knowledge Discovery*, p.p. 1-11, 1994
- [4] G. H. John, "SIPping from Data Firehose", *The Third Conference of Knowledge Discovery and Data Mining (KDD97)*, 1997
- [5] P. Cabena, P. Hadjinian, R. Stadler, J. Verhees, A. Zanasi, "Discovery Data Mining. From Concept to Implementation", Prentice Hall, 1997
- [6] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, R. Wirth "CRISP-DM 1.0 Step-by-Step Data Mining Guide", 2000
- [7] International Organization for Standardization, "ISO/IEC Standard 12207:1995.
- [8] IEEE Computer Society, "Standard for Developing Software Life Cycle Processes. {IEEE} Std. 1074-1991", 1991
- [9] B. W. Boehm, C. Abts, A. W. Brown, S. Chulani, B. K. Clark, E. Horowitz, R. Madachy, D. Reifer, B. Steece, "Software Cost Estimation with COCOMO II", Prentice Hall, 2000
- [10] L.H. Putnam Sr., D. T. Putnam, L.H. Putnam Jr., M. A. Ross, "Software Lifecycle Management (SLIM) Training. SLIM Estimate Exercises with Answers, Quantitative Software Management", Mc Lean, 2000
- [11] PRICE Systems, LLC, "PRICE-H Reference Manual Version 3.0", 1998
- [12] P. Domingos, "How to Get a Free Lunch: A Simple Cost Model for Machine Learning Applications", *Proceedings of AAAI-98/ICML-98 Workshop on the Methodology of Applying Machine Learning*, 1998
- [13] M. Pazzani, C. Merz, P. Murphy, K. Ali, T. Hume, C. Brunk, "Reducing Misclassification Costs", *Eleventh International Conference on Machine Learning*, p.p. 217-225, Morgan Kaufmann, 1994
- [14] B. Masand, G. Piatetsky-Shapiro, "A Comparison of Approaches for Maximizing Business Payoff of Prediction Models", *Second International Conference on Knowledge Discovery and Data Mining*, p.p. 195-201, 1996
- [15] P. Domingos, "MetaCost: A General Method for Making Classifiers Cost-Sensitive", *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, p.p. 155-164, 1999
- [16] J. Kleinberg, C. Papadimitriou, P. Raghavan, "A Microeconomic View of Data Mining", *Journal of Data Mining and Knowledge Discovery*, Kluwer Academic Publishers, 1999
- [17] P. D. Turney, "Cost-Sensitive Classification: Empirical Evaluation of a Hybrid Genetic Decision Tree Induction Algorithm", *Journal of Artificial Intelligence Research*, p.p. 369-409, 1995

- [18] R.A. Brealey, S.C. Myers, "Principles of Corporate Finance", McGraw-Hill, 1996
- [19] S. Chulani, B. Clark, B. Boehm, "Calibration Approach and Results of COCOMOII.1997", 22nd Software Engineering Workshop, 1997
- [20] Chulani, S., B. Boehm, and B. Steece. 1999. "From Multiple Regression to Bayesian Analysis for COCOMO II." 21st Annual Conference of the International Society of Parametric Analysts (ISPA) and the 9th Annual Conference of the Society of Cost Estimating and Analysis (SCEA).