Available online at www.sciencedirect.com
ScienceDirect
Information Systems 33 (2008) 133–150
www.elsevier.com/locate/infosys

A cost model to estimate the effort of data mining projects (DMCoMo)

Oscar Marbán*, Ernestina Menasalvas, Covadonga Fernández-Baizán

Facultad de Informática, Universidad Politécnica de Madrid (U.P.M.), Campus de Montegancedo s/n., 28660 Boadilla del Monte, Madrid, Spain

Received 26 February 2007; accepted 7 July 2007
Recommended by N. Koudas

Abstract

CRISP-DM is the standard to develop Data Mining projects. CRISP-DM proposes processes and tasks that you have to carry out to develop a Data Mining project. A task proposed by CRISP-DM is the cost estimation of the Data Mining project.

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. The lack of Data Mining project estimation methods is because of many real-life project failures due to the non-realistic estimation at the beginning of the projects.

Consequently, in this paper we propose to design and validate a parametric cost estimation model, similar to COCOMO or SLIM in software development, for Data Mining projects (DMCoMo[1]). The drivers of the model will be proposed first and later the equation of the model will be proposed.

Keywords: Data Mining; Knowledge discovery; Cost estimation; Parametric model

1. Introduction

The concept of CRM (*Customer Relationship Management*) evolved when the man of the caverns could choose if he wanted to trade with Og or Thag. However, the CRM was used for the first time in the middle of the 1990s. CRM can be defined as: "to give to the client what he wants, when he wants and where he wants" [1]. The main objective of the CRM projects is to recover a one-to-one relationship with the client. The one-to-one relationship has been lost as a consequence of the competitive environment in which modern companies work. For this reason, companies have been developing CRM systems, in order to retain their clients, for the past 10 years. In CRM systems we can distinguish between three areas: operational CRM, collaborative

*Corresponding author. Tel.: +34913367388; fax: +34913367393.
E-mail addresses: omarban@fi.upm.es (O. Marbán), emenasalvas@fi.upm.es (E. Menasalvas), cfbaizan@fi.upm.es (C. Fernández-Baizán).
[1]The work presented in this paper has been partially supported by UPM project ERDM ref.14589.

0306-4379/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.is.2007.07.004

CRM and analytic CRM. Analytic CRM analyzes the operational data to optimize the relationship with the client. Due to the great volume of data that must be analyzed Data Mining techniques must be used [2,3].

Therefore, Data Mining researches have been increasing in the last years [4,5]. This growth has been motivated because companies need to find the knowledge that is hidden in their data. This knowledge allows companies to compete against other companies. For this reason, companies are investing more resources in Data Mining projects [6].

The need of efficient methods to search knowledge in data has caused that a lot of Data Mining algorithms and Data Mining tools have been developing [7–10]. However, due to the complexity of the Data Mining process, a Data Mining methodology is needed. The Data Mining methodology is CRISP-DM [11].

CRISP-DM evolved to solve the problems that companies had in the development of Data Mining projects. CRISP-DM is a process model to develop Data Mining projects and was proposed by a consortium of companies (Teradata, SPSS (ISL), Daimler-Chrysler and OHRA). CRISP-DM defines the processes and tasks that you have to do in order to develop a successful Data Mining project. For each task proposed by CRISP-DM, the inputs and outputs of the task are also proposed. Hence, CRISP-DM proposed a process model to develop Data Mining projects such as ISO 12207 [12] and IEEE 1074 [13] for developing software projects.

In the "*Business understanding*" phase CRISP-DM proposes a task to make the project plan. In this task you have to budget the project, and you have to calculate the cost of the project taking into account the time and the personnel that are needed to develop the Data Mining project. But, CRISP-DM does not propose how to carry out this task.

If we wish to rate the success or failure of a Data Mining project, we need a method to calculate the goodness of the knowledge extracted by the model, the time used to obtain the knowledge, the cost of the personnel and resources used in the project, etc. However, it is needed to estimate the cost of the project too, because if the cost of the knowledge is not accessible to the company the project is non-viable.

Some researches have been done to estimate the goodness of the knowledge extracted from the data. Thus, in [14] a framework to estimate the goodness of knowledge after the Data Mining phase in CRM projects is proposed. This framework tries to maximize the value of the knowledge extracted. In [15] the value of customers is taken into account to maximize the benefit of the predictive Data Mining models.

About the cost estimation of the Data Mining projects, in [16] a cost estimation model for classification problems is proposed, which can be used in any moment along the project. This model is based on NPVs (*Net Present Values*) [17]. NPV is calculated as the difference between the money invested in the project and the recovered money from that investment. In the model presented in [16] the NPV is used to decide whether the project will continue. NPV is calculated at any point in the project, and the project will continue only if NPV has a positive value.

All previous estimation methods do not allow to establish the effort, time and cost at the beginning of the project. But we can try to use the software estimation tools such as COCOMO II [18], SLIM [19] or PRICE-S [20] to estimate the cost of Data Mining projects. If we take a look at these tools, we can conclude that they are not useful for estimating Data Mining projects, because they use the size of the software as the main input, in lines of code, to be developed. Other factors used in the estimation of software are experience of the development team, use of tools, features of the development platform and so forth. These features must also be used to estimate the cost of Data Mining projects. Nevertheless, if we wish to estimate Data Mining projects, we allow for other features of Data Mining projects such as characteristics of data sources, data integration level, the kind of Data Mining problem to be solved and the number of models to build inter alia. Software estimation methods do not consider those features of Data Mining projects. Hence, software estimation methods are not useful for estimating Data Mining projects.

Consequently, we can say that nowadays there is no cost estimation method for Data Mining projects, although Data Mining projects have been developing for the past 20 years. Therefore, in this paper we propose a parametric estimation model for Data Mining projects. The model is named DMCoMo (Data Mining Cost Model). DMCoMo is based on a parametric cost estimation model such as COCOMO family. DMCoMo allows to estimate the effort (*men* × *month*) that is needed to develop a Data Mining project since its conception until its deployment.