Available online at www.sciencedirect.com
ScienceDirect
Expert Systems with Applications 34 (2008) 522–529
Expert Systems with Applications
www.elsevier.com/locate/eswa

An association rule mining method for estimating the impact of project management policies on software quality, development time and effort

María N. Moreno García [a,*], Isabel Ramos Román [b], Francisco J. García Peñalvo [a], Miguel Toro Bonilla [b]

[a] Dept. Informática y Automática, University of Salamanca, Plaza Merced s/n, 37008 Salamanca, Spain
[b] Dept. Lenguajes y Sistemas Informáticos, University of Sevilla, Avda. Reina Mercedes s/n, 41012 Sevilla, Spain

Abstract

Accurate and early estimations are essential for effective decision making in software project management. Nowadays, classical estimation models are being replaced by data mining models due to their application simplicity and the rapid production of profitable results. In this work, a method for mining association rules that relate project attributes is proposed. It deals with the problem of discretizing continuous data in order to generate a manageable number of high confident association rules. The method was validated by applying it to data from a Software Project Simulator. The association model obtained allows us to estimate the influence of certain management policy factors on various software project attributes simultaneously.

Keywords: Association rules; Data mining; Software estimation; Project management; Simulation

1. Introduction

Software quality, project duration and development effort are important factors to be kept under control in the software development process. They are interrelated and influenced by many other factors which complicate the monitoring tasks. When managers have to take decisions about a project, they must consider a great number of variables and the complex relations between them. The simulation of software development projects by using dynamic models has contributed to a better knowledge of the influence of these variables and their relations. The *Software Project Simulators* (SPS), based on dynamic models, enable us to simulate the project's behavior and to evaluate the impact of different management policies and other factors. Nevertheless, they have important drawbacks: first, the number of input parameters needed for the simulation, and second, the number of possible combinations of factors influencing the development process that make it difficult to choose the best combination for the desired objectives. An important improvement in SPS is the treatment of the data generated by the simulator by using machine learning and evolutionary algorithms in order to facilitate their use (Aguilar-Ruiz, Ramos, Riquelme, & Toro, 2001; Ramos, Riquelme, & Aroba, 2001). The combination of factors for achieving specific objectives can be learned through the application of these supervised techniques. Such information allows managers to establish the correct management policy taking as reference the model generated by these algorithms.

When machine learning techniques are used, only one output variable is the target of the prediction. In classification problems this variable, named the class attribute, must be discrete. Classification algorithms use classified historical data for inducing a relation model between the class attribute and the other attributes (descriptive attributes).

* Corresponding author. Tel.: +34 923 294400.
E-mail address: mmg@usal.es (M.N. Moreno García).

0957-4174/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2006.09.022

Later, the model can make predictions about new, unclassified data.

The aim of this work is to present a method for estimating three variables simultaneously. We propose an association rule mining algorithm for building a model that relates management policy attributes to the output attributes quality, time and effort. All the available attributes are continuous, they must thus be split into intervals of values in order to generate the rules. The applicability and interest of the discovered associations depend mostly on how the data is discretized. The success of our method is mainly due to the supervised, multivariable procedure used for discretization. The result is an association model comprised of a manageable number of high confidence rules representing relevant patterns between project attributes. Those patterns provide managers with important information for decision making.

The rest of the paper is organized as follows: next section contains the fundamentals and main works concerning SPS, association rules and data discretization. Section 3 describes the experimental data provided by a dynamic simulation environment. The following section deals with the stage of data preprocessing. The proposed method for association rule mining and its results are presented in Section 5. The evaluation of the associative model obtained is given in Section 6 and, finally, we draw some conclusions.

2. Background

2.1. Software project simulators

The simulation of software project behavior by means of an SPS provides managers with a valuable tool for trying out several policies in order to take better decisions.

A dynamic model consists of a collection of parameters and functions needed for building the simulation environment. It is articulated in mathematical terms by means of a set of differential equations which express restrictions between variables that change over time. These restrictions enable us to analyze cause–effect relations among several project factors, such as management policies and technological, product and process factors. The analysis of the project with SPS can be done before the project start (a priori analysis), during the development (project monitoring) and when the project has finished (post-mortem analysis).

Since the publication of a dynamic model for software projects by Abdel-Hamid and Madnick in 1991 (Abdel-Hamid & Madnick, 1991), many other models and simulation environments for diverse application domains have appeared. In recent years the research in this field has increased, producing significant advances (Kellner, Madachy, & Raffo, 1999; Rodrigues & Williams, 1997). The most complete models are very complex and manage a large number of parameters which should be known previously. However, the numerous combination possibilities of such parameters make it difficult to find the best combination for a specific situation or purpose. These are the main drawbacks for the use of an SPS. In order to solve this problem, reduced dynamic models have been proposed for specific phases of the project (Ramos & Ruiz, 1998; Ruiz, Ramos, & Toro, 2001). Another alternative is the construction of data mining models from SPS data in order to analyze separately the influence of some of the factors related to the management policy on some of the project attributes. A number of data mining techniques, such as machine learning and evolutionary algorithms, can be used for building the models (Ramos et al., 2001) (Aguilar-Ruiz et al., 2001). The main drawback of a model generated by one of these supervised algorithms is that it can be used to estimate the repercussion of different management policies on just one project attribute. In this work we propose an algorithm for mining association rules that provide an associative model that will let managers know simultaneously the influence of policy factors on several project attributes.

2.2. Association rules

Data mining models can be obtained by employing supervised and unsupervised algorithms. Supervised methods require a learning stage for building a predictive model. The target of the prediction is a special attribute called the "label". The model is built from historical labeled data records by encoding the relation between the label and the other attributes; then, the model can be used for making predictions about new, unlabeled data. The two most common supervised modeling methods are classification and regression. If the label is discrete, it is named the class label and the task is called classification; if the label is continuous, the task is called regression. Unsupervised algorithms belong to knowledge discovery modeling. This task is descriptive instead of predictive and the objective is to detect patterns in present data without need of previous learning.

Traditionally, association analysis is considered an unsupervised technique, so it has been applied in knowledge discovery modeling. Recent studies have shown that knowledge discovery algorithms, such as association rule mining, can be successfully used for prediction in classification problems (Hu, Chen, & Tzeng, 2002; Li, Shen, & Topor, 2001; Moreno, García, & Polo, 2004; Wang & Wong, 2003). Patterns that have been extracted from historical data can serve to predict upcoming behaviours.

Since Agrawal et al. introduced the concept of association between items (Agrawal, Imielinski, & Swami, 1993a; Agrawal, Imielinski, & Swami, 1993b) and proposed the Apriori algorithm (Agrawal & Srikant, 1994), many other authors have studied better ways for obtaining association rules from transactional databases. Below, we introduce the foundations of association rules and some concepts used for quantifying the statistical significance and goodness of the generated rules (Padmanabhan & Tuzhilin, 2002).