Chapter 6.7
A Two-Stage Zone Regression Method for Global Characterization of a Project Database

J. J. Dolado
University of the Basque Country, Spain

D. Rodríguez
University of Reading, UK

J. Riquelme
University of Seville, Spain

F. Ferrer-Troyano
University of Seville, Spain

J. J. Cuadrado
University of Alcalá de Henares, Spain

ABSTRACT

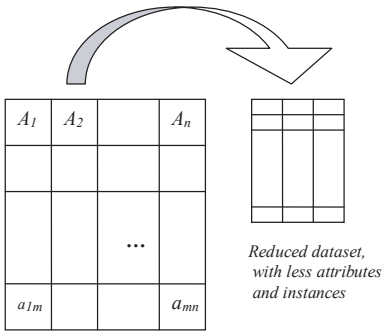
One of the problems found in generic project databases, where the data is collected from different organizations, is the large disparity of its instances. In this chapter, we characterize the database selecting both attributes and instances so that project managers can have a better global vision of the data they manage. To achieve that, we first make use of data mining algorithms to create clusters. From each cluster, instances are selected to obtain a final subset of the database. The result of the process is a smaller database which maintains the prediction capability and has a lower number of instances and attributes than the original, yet allow us to produce better predictions.

INTRODUCTION

Successful software engineering projects need to estimate and make use of past data since the inception of the project. In the last decade, several organizations have started to collect data so that companies without historical datasets can use these generic databases for estimation. In some cases, project databases are used to compare data from the organization with other industries, that is, benchmarking. Examples of such organizations collecting data include the International Software Benchmarking Standards Group (ISBSG, 2005) and the Software Technology Transfer Finland (STTF, 2004).

One problem faced by project managers when using these datasets is that the large number of attributes and instances needs to be carefully selected before estimation or benchmarking in a specific organization. For example, the latest release of the ISBSG (2005) has more than 50 attributes and 3,000 instances collected from a large variety of organizations. The project manager has the problem of interpreting and selecting the most adequate instances. In this chapter, we propose an approach to reduce (characterize) such repositories using data mining as shown in Figure 1. The number of attributes is reduced mainly using expert knowledge although the data mining algorithms can help us to identify the most relevant attributes in relation to the output parameter, that is, the attribute that wants to be estimated (e.g., *work effort*). The number of instances or samples in the dataset is reduced by selecting those that contribute to a better accuracy of the estimates after applying a version of the M5 (Quinlan, 1992) algorithm, called M5P, implemented in the Weka toolkit (Witten & Frank, 1999) to four datasets generated from the ISBSG repository. We compare the outputs before and after, characterizing the database using two algorithms provided by Weka, multivariate linear regression (MLR), and least median squares (LMS).

Figure 1. Characterizing dataset for producing better estimates



This chapter is organized as follows: the *Techniques Applied* section presents the data mining algorithm; *The Datasets* section describes the datasets used; and the *Evaluation of the Techniques and Characterization of Software Engineering Datasets* section discusses the approach to characterize the database followed by an evaluation of the results. Finally, the *Conclusions* section ends the chapter.

TECHNIQUES APPLIED

Many software engineering problems like cost estimation and forecasting can be viewed as *classification* problems. A classifier resembles a function in the sense that it attaches a value (or a range or a description), named the *class, C,* to a set of attribute values A_1, A_2,... A_n, that is, a classification function will assign a class to a set of descriptions based on the characteristics of the instances for each attribute. For example, as shown in Table 1, given the attributes *size*, *complexity*, and so forth, a classifier can be used to predict the *effort*.

In this chapter, we have applied data mining, that is, computational techniques and tools designed to support the extraction, in an automatic way, of the information useful for decision support or exploration of the data source (Fayyad,