Chapter 12

Software Effort Estimation: Harmonizing Algorithms and Domain Knowledge in an Integrated Data Mining Approach

Jeremiah D. Deng
University of Otago, New Zealand

Martin K. Purvis
University of Otago, New Zealand

Maryam A. Purvis
University of Otago, New Zealand

ABSTRACT

Software development effort estimation is important for quality management in the software development industry, yet its automation still remains a challenging issue. Applying machine learning algorithms alone often cannot achieve satisfactory results. This paper presents an integrated data mining framework that incorporates domain knowledge into a series of data analysis and modeling processes, including visualization, feature selection, and model validation. An empirical study on the software effort estimation problem using a benchmark dataset shows the necessity and effectiveness of the proposed approach.

INTRODUCTION

Data mining research has made prominent advances in recent years, with various computational methods contributed from fields such as statistics and machine learning being explored and applied in many application domains. On the other hand, researchers have also realized the lack of incorporating domain knowledge in data mining solutions (Pohle, 2003; Sinha & Zhao, 2008). Without sufficient and proper use of domain knowledge, data mining applications may risk the danger of taking the wrong approach, or choosing inappropriate or suboptimal algorithms or models. Apart from the lacking of domain knowledge incorporation, an automated data mining process is prone to lead to
the misinterpretation of data analysis outcomes, and therefore compromise our confidence in using the employed computational methods.

Despite these widely accepted notions, integrating domain knowledge into data mining solutions remains challenging. Some domain knowledge is difficult to put into explicit form, such as rules. Often, there are inconsistencies between what computer algorithms suggest and what human experts understand.

As a promising application area for data mining, making realistic estimates of software development effort has been desired by the software industry for a long time. An accurate effort prediction can benefit project planning and management, and better guarantee the service quality of software development. The importance of software effort modeling is obvious and people have spent considerable effort in collecting project development data in large quantities. While expert judgment is widely used, there is also increasing interest in applying statistics and machine learning techniques to predict software project effort. In recent years, with the release of software development data into the public domain such as those from ISBSG (Chulani et al., 1999), the prospects of building powerful, practical estimation tools seem more favourable, even though formidable challenges still remain.

In this article, we present a case study of software effort estimation that combines the use of data mining algorithms and domain knowledge. While various regression techniques have previously been proposed in software effort estimation, few have paid attention to such important issues as feature selection, model selection and validation. In this paper, we have adopted an integrated data mining approach, where data visualization, feature selection, and modeling have been conducted, and all these processes interact with relevant domain knowledge. Through such interaction, data analysis processes can often be assisted by domain knowledge, and the results get confirmed. When there are discrepancies between algorithmic outcome and domain knowledge, this can help either to draw attention to the corresponding issues (which can be further analyzed in latter data processing), or even correct possible misunderstandings on the human side. In the end, domain knowledge also helps to validate the prediction model as well as assist in the interpretation of its final outcome, making the entire data analysis and modeling process both relevant and effective.

The remainder of the paper is organized as follows. First we briefly review a few relevant works on using data mining techniques for software effort estimation. Then, we present the basic computational methods we have adopted. The case study is presented and finally, we conclude the paper with a discussion on some future directions.

**RELEVANT WORK**

Machine learning techniques have long found to be useful for software effort development. Two machine learning methods were used in Srinivasan and Fisher (1995) to build estimators from historical data and the model sensitivity of these methods was discussed. A notable work by Shepperd and Schofield (1997) investigated the use of case-based reasoning (CBR). A number of machine learning approaches were investigated by Mair et al. (2000), including CBR, artificial neural networks (ANNs) and rule induction. Olivieira employed support vector regression on the NASA software project data and obtained significantly improved results over linear regression and radial basis functions. Genetic algorithms were used for feature weighting in CBR-based estimation methods (Huang & Chiu, 2006; Li et al., 2009). Elish (2009) studied using multiple additive regression trees to improve the estimation of project effort. Mendes et al. (2002) looked at the comparison of some