Chapter 14

Connectionist and Evolutionary Models for Learning, Discovering and Forecasting Software Effort

Parag C. Pendharkar and Girish Subramanian
Pennsylvania State University at Harrisburg, USA

ABSTRACT

Mining information and knowledge from very large databases is recognized as a key research area in machine learning and expert systems. In the current research, we use connectionist and evolutionary models for learning software effort. Specifically, we use these models to learn the software effort from a set of training data set containing information on software projects and test the performance of the model on a holdout sample. The design issues of developing connectionist and evolutionary models for mining software effort patterns on a data set are described. Our research indicates that connectionist and evolutionary models, whenever carefully designed, hold a great promise for knowledge discovery and forecasting software effort.

INTRODUCTION

Data mining (DM), which is also referred to as knowledge discovery, is a systematic approach to finding hidden patterns, trends and relationships in data
Recently, DM has attracted tremendous attention from both researchers and practitioners because of its applicability to information management, decision support, fraud detection, marketing strategy, financial forecasting, process control and many other applications (Chen et al., 1996; Cheung, Ng, Fu, & Fu, 1996).

Many data mining algorithms and methodologies have been proposed (Agrawal, Ghosh, Imielinski, & Swami, 1992; Agrawal, Imielinski, & Swami, 1993a, 1993b; Agrawal, Link, Sawhney, & Shim, 1995; Agrawal, Mehta, Shafer, Srikant, Arning, & Bollinger, 1996; Agrawal & Srikant, 1994; Agrawal & Srikant, 1995; Cheeseman & Stutz, 1996; Chen, Han, & Yu, 1996; Chen et al., 1996; Cheung et al., 1996). Different researchers classified the data mining algorithms based on the kinds of databases to work on, kind of knowledge to be mined, and kind of techniques to be utilized (Chen et al., 1996; Cheung et al., 1996). Recently, Pass (1997) classified approaches to data mining into two categories: methodologies and technologies. According to Pass (1997), the methodologies to data mining consist of cluster analysis, linkage analysis, visualization, and categorization analysis. The technologies were connectionist models/neural networks, decision trees, genetic algorithms, fuzzy logic, statistical approaches and time series approaches.

Software development and maintenance involve significant cost for the organizations (Banker & Slaughter, 1997). The productivity improvements in software development and maintenance can enable IS departments to reduce long-term costs and increase profitability. The previous research (Banker & Slaughter, 1997; Subramanian & Zarnich, 1996) suggests that systems development methodology, project size, developer’s experience, and system design tools (e.g., CASE tools) all significantly explain the software effort and productivity. While most of the studies focused on the study of the significance of the predictor variables with the predicted software effort, very few studies focused on applying data mining techniques to the historical data to improve the software effort estimation in the future. There is evidence that the relationships between the predictor and predicted variables is nonlinear. In a recent study, Banker and Slaughter (1997) used a data envelopment analysis approach to illustrate that scale economies exist in software maintenance. The existence of the scale economies in software productivity indicates that better estimates can be obtained if non-linear forecasting models were used to learn and predict software effort. Further, lack of a priori knowledge about the exact non-linear relationships between the multivariate independent variables with the dependent variable precludes the use of any predetermined non-linear statistical technique.

In the current research, we use connectionist (also called artificial neural networks) and evolutionary (also called genetic algorithms) models to discover and learn the non-linear relationships between the predictor variables in the training
Related Content

A New Similarity Metric for Sequential Data
www.igi-global.com/chapter/new-similarity-metric-sequential-data/61178?camid=4v1a

Ranking Potential Customers Based on Group-Ensemble
Zhi-Zhuo Zhang, Qiong Chen, Shang-Fu Ke, Yi-Jun Wu, Fei Qi and Ying-Peng Zhang (2008). International Journal of Data Warehousing and Mining (pp. 79-89).
www.igi-global.com/article/ranking-potential-customers-based-group/1809?camid=4v1a

Frequent Pattern Discovery and Association Rule Mining of XML Data
Qin Ding and Gnanasekaran Sundarraj (2012). XML Data Mining: Models, Methods, and Applications (pp. 243-263).
www.igi-global.com/chapter/frequent-pattern-discovery-association-rule/60912?camid=4v1a
A Survey of Parallel and Distributed Data Warehouses
www.igi-global.com/chapter/survey-parallel-distributed-data-warehouses/53075?camid=4v1a