Chapter XI

Applying Rule Induction in Software Prediction

Bheki sipho Twala, Brunel University, UK
Michelle Cartwright, Brunel University, UK
Martin Shepperd, Brunel University, UK

Abstract

Recently, the use of machine learning (ML) algorithms has proven to be of great practical value in solving a variety of software engineering problems including software prediction, for example, cost and defect processes. An important advantage of machine learning over statistical analysis as a modelling technique lies in the fact that the interpretation of production rules is more straightforward and intelligible to human beings than, say, principal components and patterns with numbers that represent their meaning. The main focus of this chapter is upon rule induction (RI): providing some background and key issues on RI and further examining how RI has been utilised to handle uncertainties in data. Application of RI in prediction and other software engineering tasks is considered. The chapter concludes by identifying future research work when applying rule induction in software prediction. Such future research work might also help solve new problems related to rule induction and prediction.
Introduction

Machine learning (ML), which has been making great progress in many directions, is a hallmark of machine intelligence just as human learning is the hallmark of human intelligence. The ability to learn and reason from observations and experience seems to be crucial for any intelligent being (Forsyth & Rada, 1986; Holland, 1975; Winston, 1992).

One major problem for applying ML algorithms in software engineering is the unavailability and scarcity of software data, that is, data for training the model. Surveys for collecting software engineering data are usually small but difficult and expensive to conduct. This is due to the lack of expertise with required knowledge to carry out and maintain high quality information and the nature of software development which cannot be collected in an environmental setting. Also, the lack of data could arise from the need for confidentiality—industrial companies are often reluctant to allow access to data on software failures because of the possibility that people might think less highly of their products. Another problem as far as ML application is concerned is the form that the dataset takes, that is, the dataset characteristics. Most learning techniques assume that the data are presented in a simple attribute-value format. Another important feature of a problem domain is the quality of the data available. Most real data is imperfect: incomplete; irrelevant; redundant; noisy; and erroneous. The aim of a learning system is to discover a set of decision rules that is complete, in that it describes all of the data and predicts the data accurately.

In this chapter, we explore what RI can do in the software engineering domain to increase the awareness of learning methods when building software prediction models. However, our main focus is on the application of rule induction to software prediction.

The topic is significant because:

1. RI is an emerging technology that can aid in the discovery of rules and patterns in sets of data.
2. RI has an advantage over statistical analysis as a modelling technique due to the fact that the interpretation of production rules is more straightforward and intelligible to human beings than, say, principal component and patterns with numbers that represent their meaning.
3. Due to the lack of adequate tools to evaluate and estimate software project estimation, RL strategies have been used to tackle such problems, including software prediction.

Rule induction is one of the most established and effective data mining technology in use today that has been applied successfully in several disciplines and to real-world domains. These include: preventing breakdowns in electric transformers (Riese, 1984); increasing yield in chemical process control (Leech, 1986); improving separation of gas from oil (Guilfoyle, 1986); making credit card decisions (Michie, 1989); diagnosis of mechanical devices (Giordana, Neri, & Saitta, 1996); monitoring quality of rolling emulsions (Karba & Drole, 1990); coronary heart disease diagnosis and risk group delivery (Gamberger, Lavrac, & Krstacic, 2002); categorising text documents (Johnson, Oles, Zhang, & Goetz, 2002); and among other areas.
Related Content

Object-Oriented Technology for Expert System Generation
[www.igi-global.com/chapter/object-oriented-technology-expert-system/38488?camid=4v1a](www.igi-global.com/chapter/object-oriented-technology-expert-system/38488?camid=4v1a)

Case-Based Reasoning and Some Typical Applications
[www.igi-global.com/chapter/case-based-reasoning-and-some-typical-applications/97060?camid=4v1a](www.igi-global.com/chapter/case-based-reasoning-and-some-typical-applications/97060?camid=4v1a)

A Recovery-Oriented Approach for Software Fault Diagnosis in Complex Critical Systems
[www.igi-global.com/chapter/recovery-oriented-approach-software-fault/56153?camid=4v1a](www.igi-global.com/chapter/recovery-oriented-approach-software-fault/56153?camid=4v1a)
Unexplored Hypotheses on Potency-Magnitude Relations of eWOM Messages with Intensified Comparative Expressions


[www.igi-global.com/article/unexplored-hypotheses-on-potency-magnitude-relations-of-ewom-messages-with-intensified-comparative-expressions/101316?camid=4v1a](www.igi-global.com/article/unexplored-hypotheses-on-potency-magnitude-relations-of-ewom-messages-with-intensified-comparative-expressions/101316?camid=4v1a)