AGILE MINING

- A NOVEL DATA MINING PROCESS FOR INDUSTRY PRACTICE BASED ON AGILE METHODS AND VISUALIZATION

Xiao Zhu
Advanced Analytics Institute
School of Software
Faculty of Engineering and Information Technology
University of Technology Sydney
Publication List during Master Degree Study:


DECLARATION

This dissertation is the result of my work and includes nothing, which is the outcome of work done in collaboration except where specifically indicated in the text. It has not been previously submitted, in part or whole, to any university or institution for any degree, diploma, or other qualification.

Signed: ________________________________________________________________

Date: 15/11/2017

Xiao Zhu
ABSTRACT

Current standard data mining processes like CRoss-Industry Standard Process for Data mining (CRISP-DM) are vulnerable to frequent change of customer requirement. Meanwhile, Stakeholders might not acquire sufficient understanding to generate business value from analytic results due to a lack of intelligible explanatory stage. These two cases repeatedly happen on those companies which are inexperienced in data mining practice. Towards this issue, Agile Mining, a refined CRISP-DM based data mining (DM) process, is proposed to address these two friction points between current data mining processes and inexperienced industry practitioners. By merging agile methods into CRISP-DM, Agile Mining process achieves a requirement changing friendly data mining environment for inexperienced companies. Moreover, this Agile Mining transforms traditional analytic-oriented evaluation to business-oriented visualization-based evaluation. In the case study, two industrial data mining projects are used to illustrate the application of this new data mining process and its advantages.

Keyword: Data Mining, Data Mining Standard Process, CRISP-DM, Visualization.
CONTENTS

1 INTRODUCTION ........................................................................................................ 4
  1.1 RESEARCH BACKGROUND .............................................................................. 4
  1.2 RELATED WORK .................................................................................................. 5
    1.2.1 Data Mining Process .................................................................................. 5
    1.2.2 Agile Methods ............................................................................................ 6
    1.2.3 Visualization in Data Mining ................................................................. 6
  1.3 RESEARCH ISSUE ............................................................................................... 7
  1.4 RESEARCH NOVELTY .......................................................................................... 8
  1.5 STRUCTURE OF THIS PAPER .............................................................................. 8

2 LITERATURE REVIEW .......................................................................................... 10
  2.1 DATA MINING PROCESS ................................................................................ 10
    2.1.1 CRISP-DM .............................................................................................. 11
    2.1.2 SEMMA ................................................................................................. 14
    2.1.3 KDD ........................................................................................................ 15
    2.1.4 Comparison of KDD, SEMMA and CRISP-DM ....................................... 16
  2.2 AGILE METHOD FOR SOFTWARE DEVELOPMENT .......................................... 18
    2.2.1 Character of Agile Method ......................................................................... 19
    2.2.2 Extreme Programming ............................................................................. 21
    2.2.3 Scrum Method .......................................................................................... 23
    2.2.4 Comparison of Scrum and XP .................................................................. 28
  2.3 VISUALIZATION IN DATA MINING ................................................................... 29
    2.3.1 Definition of Data Mining Visualization ................................................... 30
    2.3.2 Objective of Data Mining Visualization .................................................... 31
  2.4 SUMMARY OF LITERATURE REVIEW ............................................................... 33

3 A NOVEL DATA MINING PROCESS BASED ON AGILE METHODS AND
VISUALIZATION ........................................................................................................ 34
  3.1 INTRODUCTION TO AGILE MINING PROCESS ............................................... 34
  3.2 REQUIREMENT RECONFIRMING ..................................................................... 36
LIST OF ABBREVIATIONS AND ACRONYMS

CRISP-DM: Cross Industry Standard Process for Data Mining
AM: Agile Mining
RR: Requirement Reconfirming
PR: Performance Refining
RF: Random Forest
LR: Logistic Regression
SVM: Support Vector Machine
NB: Naïve Bayes