LOCALIZATION OF SOFTWARE DEVELOPMENT EFFORT ESTIMATION BASED ON CLASSIFICATION OF PROJECTS

VAHID KHATIBI BARDSIRI

A thesis submitted in fulfilment of the
requirements for the award of the degree of
Doctor of Philosophy (Computer Science)

Faculty of Computing
Universiti Teknologi Malaysia

JUNE 2013

I declare that this thesis entitled: "Localization of Software Development Effort Estimation based on Classification of Projects" is the result of my own research except as cited in references. The thesis has not been accepted for any degree and is not concurrently submitted in candidature of any other degree.

Signature	:	
Name	:	VAHID KHATIBI BARDSIRI
Date	:	

“To my beloved wife and daughter”

ACKNOWLEDGEMENT

I would like to express my special appreciation and thanks to my supervisor Assoc.Prof.Dr. Dayang Norhayati Abang Jawawi, you have been a tremendous mentor for me. I would like to thank you for encouraging my research and for allowing me to grow as a research scientist. Your advice on both research as well as on my career have been invaluable. My sincere appreciation also extends to Assoc.Prof.Dr. Siti Zaiton mohd Hashim who has provided assistance at various occasions. Her views and tips have been really useful throughout the current research. At the same time, I also appreciate Universiti Teknologi Malaysia for International Doctorate Fellowship (IDF) award.

A special thanks to my family. Words cannot express how grateful I am to my family. Your prayer for me was what sustained me thus far.

I would also like to thank my beloved wife, Elham. Thank you for supporting me, and especially I can’t thank you enough for encouraging me throughout this experience. To my beloved daughter, Bita, I would like to express my thanks for being such a good girl always cheering me up. I will keep on trusting You for my future.

ABSTRACT

Accurate estimation of software development effort is strongly associated with the success or failure of software projects. Since the existing estimation models are confronted by special characteristics of software projects such as non-normality of project attributes, intangible product, and lack of well-defined measurement techniques, the accuracy of estimates is not convincing and the estimation models are not sufficiently adaptable to support a wide range of software projects. The main issue is that prior studies have considered software projects to be similar to other types of projects regardless of the fact that software projects are entirely different. Furthermore, an accurate estimation of development effort in software projects is unreachable in global space, thus this research proposed that the estimation process be localized based on the characteristics of the project being estimated. The aim of the study is to alleviate the inconsistency and heterogeneous nature of software projects through the ideas of localization and classification. This research focused on a widely accepted estimation model called analogy based estimation (ABE). Four different estimation models were proposed to overcome the problems of: low quality in construction of machine-learning based estimation models, difficulty of attribute weighting, limited support of software projects, and biased comparisons. Each proposed model was a hybrid model constructed using a classification method, ABE, and a complimentary method (soft computing or statistical techniques). Performance evaluations of the models were carried out using five real software project datasets and the estimates were compared to those achieved by other well-known estimation models. A comparative study was also conducted for further verification of the models. The evaluation results confirmed the ability of the proposed models to improve the accuracy of estimates, quality of training, efficiency of attribute weighting, adaptability of estimation process, correctness of comparisons, and support domain of software projects.

ABSTRAK

Ketepatan anggaran untuk usaha pembangunan perisian adalah sangat berkait rapat dengan kejayaan atau kegagalan projek perisian. Oleh kerana model anggaran usaha yang sedia ada berhadapan dengan ciri-ciri khas projek perisian seperti ketidak-normalan atribut projek, ketidak-ketaraan produk, dan kekurangan teknik pengukuran yang jelas, ketepatan anggaran tidak begitu menyakinkan dan model anggaran tersebut tidak mudah disesuaikan untuk menyokong pelbagai projek perisian. Isu utama ialah kajian sebelum ini menganggap projek perisian adalah sama seperti jenis projek lain tanpa mengambilkira kenyataan yang projek perisian berbeza secara keseluruhannya. Tambahan pula, anggaran yang tepat pada projek perisian tidak tercapai pada ruang global, oleh itu penyelidikan ini mencadangkan proses anggaran yang disetempatkan berdasarkan ciri-ciri projek yang dianggarkan. Tujuan kajian ini adalah untuk mengurangkan ketidak-konsistenan dan tabii berheterogen projek perisian menerusi idea penyetempatan dan pengelasan. Penyelidikan ini memberi fokus kepada model anggaran penganggaran usaha yang telah diterima pakai secara meluas yang dinamakan model berasaskan analogi (ABE). Empat model anggaran yang berbeza telah dicadangkan untuk mengatasi: masalah kualiti rendah pada pembinaan model anggaran berasaskan pembelajaran-mesin, kesukaran penentuan pemberat atribut, sokongan terhad kepada projek perisian dan perbandingan berat sebelah. Setiap model yang telah dicadangkan adalah model gabungan yang dibangunkan dengan menggunakan kaedah pengelasan, ABE, dan kaedah pelengkap (pengkomputeran lembut atau teknik statistik). Penilaian-penilaian prestasi model-model telah dilakukan terhadap lima set data sebenar projek perisian dan anggaran dibandingkan dengan pencapaian model-model anggaran yang sedia ada. Kajian perbandingan telah dilaksanakan untuk pengesahan lanjut model. Keputusan penilaian mengesahkan keupayaan model yang dicadangkan memperbaiki ketepatan anggaran, kualiti latihan pembelajaran, kecekapan pemberat atribut, kebolehsesuaian proses anggaran, ketepatan perbandingan dan sokongan bidang projek perisian.

TABLE OF CONTENTS

CHAPTER	TITLE	PAGE
	DECLARATION	ii
	DEDICATION	iii
	ACKNOWLEDGEMENT	iv
	ABSTRACT	v
	ABSTRAK	vi
	TABLE OF CONTENTS	vii
	LIST OF TABLES	xv
	LIST OF FIGURES	xvii
	LIST OF ABBREVIATIONS	xx
	LIST OF APPENDICES	xxii
1	**INTRODUCTION**	**1**
	1.1 Overview	1
	1.2 Research Background	3
	1.3 Problem Statement	8
	1.3.1 Domain Problems	8
	1.3.2 Technical Problems	9
	1.4 Research Question	12
	1.5 Research Goal	13
	1.6 Research Objectives	13
	1.7 Research Scope	14
	1.8 Research Justification	15
	1.8.1 Importance of Software Development Effort Estimation	15
	1.8.2 Importance of Localization	16

1.9 Thesis Outline 18
1.10 Chapter Summary 19

2 LITERATURE REVIEW 20
2.1 Introduction 20
2.2 Algorithmic Models 22
2.2.1 Size Metrics 22
2.2.1.1 Source Lines of Code 22
2.2.1.2 Function Point Size Estimates 23
2.2.1.3 Object Oriented Metrics 24
2.2.2 Estimation Models 25
2.2.2.1 Putman’s Model 25
2.2.2.2 SEER-SEM 25
2.2.2.3 COCOMO 26
2.2.3.4 COCOMO II 29
2.2.3 Models Overview 30
2.2.4 Regression Model 31
2.3 Non-Algorithmic Models 32
2.3.1 Analogy Based Estimation 32
2.3.1.1 Similarity Function 33
2.3.1.2 Solution Function 34
2.3.1.3 K-Nearest Neighborhood 35
2.3.1.4 Previous Studies on ABE 36
2.3.2 Classification And Regression Tree 44
2.3.3 Expert Judgement 44
2.3.4 Soft Computing based Models 46
2.3.4.1 Artificial Neural Networks 47
2.3.4.2 Fuzzy Method 51
2.3.4.3 Neuro-Fuzzy Method 53
2.3.4.4 Optimization Algorithms 55
2.4 Analytical Studies 55
2.5 The Idea of Segmentation, Clustering and Localization 57
2.6 Chapter Summary 60

3 RESEARCH METHODOLOGY 61

3.1 Introduction 61

3.2 Research Phases 61

3.3 Research Framework 63

3.4 Research Process 65

3.5 Phase A: Primary Studies and Initial Planning 68

3.5.1 A1: Analysis of Existing Literature 68

3.5.2 A2: Problem Discovery 69

3.5.3 A3: Description of Datasets 72

3.5.3.1 DPS Dataset 72

3.5.3.2 CF Dataset 73

3.5.3.3 Desharnais Dataset 74

3.5.3.4 Maxwell Dataset 74

3.5.3.5 ISBSG Dataset 75

3.5.4 A4: Determining the Performance Metrics 77

3.5.4.1 Leave-One-Out 78

3.5.4.2 Cross Validation 78

3.6 Phase B: Design and Implementation of FCNNEE 79

3.6.1 B1: Describe the Problem and Associated Solution 79

3.6.2 B2: Determine the Clustering Method 80

3.6.2.1 C-Means Clustering Algorithm 81

3.7 Phase C: Design and Implementation of RCOEE 81

3.7.1 C1: Describe the Problem and Associated Solution 81

3.7.2 C2: Determine the Optimization Algorithm 82

3.7.2.1 Optimization Process 83

3.8 Phase D: Investigate the Effect of Selective Classification 84

3.8.1 D1: Data Filtering and Validating 85

3.8.2 D2: Classification of Projects 86

3.9 Phase E: Design and Implementation of SCOEE 87

3.9.1 E1: Describe the Problem and Associated Solution 87

3.9.2 E2: Select the Method of Classification 87
3.9.3 E3: Select the Key Attributes 88
3.10 Phase F: Design and Implementation of SCSAEE 89
3.10.1 F1: Describe the Problem and Associated Solution 89
3.10.2 F2: Determine the Statistical Methods 90
3.10.2.1 ANOVA 90
3.10.2.2 ANCOVA 91
3.11 Phase G: Comparison and Evaluation 91
3.11.1 G1: Comparing the Proposed Models 91
3.11.2 G2: Determining the overall Comparison Scope 92
3.11.2.1 Model Selection 92
3.11.2.2 Model Implementation 93
3.11.3 G3: Conducting a Comparative Study 94
3.11.4 G4: Providing a Set of Recommendations 95
3.12 Chapter Summary 96

4 FUZZY CLUSTERING AND NEURAL NETWORK BASED EFFORT ESTIMATION MODEL 97
4.1 Introduction 97
4.2 Design the Training Stage 97
4.2.1 Determining the Number of Clusters 98
4.2.2 Marking the Clusters 100
4.3 Design the Testing Stage 101
4.4 Experimental Results 102
4.4.1 Primary Adjustments 102
4.4.1.1 C-Means Clustering 102
4.4.1.2 Analogy Based Estimation 103
4.4.1.3 Artificial Neural Network 103
4.4.2 Comparison Scope 103
4.4.3 Numerical Results 104
4.4.3.1 Numerical Results on Desharnais Dataset 104

4.4.3.2 Numerical Results on Maxwell Dataset 107
4.4.4 Improvement Analysis 111
4.5 A Comparative Study 112
4.6 Chapter Summary 113

5 RANDOM CLUSTERING AND OPTIMIZATION BASED EFFORT ESTIMATION MODEL 115
5.1 Introduction 115
5.2 Design the Training Stage 115
5.3 Design the Testing Stage 118
5.4 Experimental Results 119
5.4.1 Primary Adjustments 119
5.4.2 Cross Validation Customization 120
5.4.3 Results on DPS Dataset 121
5.4.4 Results on CF Dataset 123
5.4.5 RCOEE versus other Estimation Models 124
5.4.6 Improvement Analysis 125
5.5 A Comparative Study 127
5.6 Chapter Summary 128

6 THE EFFECT OF SELECTIVE CLASSIFICATION 130
6.1 Introduction 130
6.2 Design the Evaluation Framework 130
6.3 Experimental Results 132
6.3.1 Classification based on Development Type 132
6.3.2 Classification based on Organization Type 135
6.3.3 Classification based on Development Platform 137
6.3.4 Efficiency Analysis 140
6.3.4.1 Development Type 140
6.3.4.2 Organization Type 141
6.3.4.3 Development Platform 142
6.3.4.4 Overall Evaluation 142
6.4 Chapter Summary 143

7 SELECTIVE CLASSIFICATION AND OPTIMIZATION BASED EFFORT ESTIMATION MODEL 144

7.1 Introduction 144

7.2 Design the Training Stage 144

7.2.1 Localization of Comparisons and Weighting Process 144

7.2.2 Combination Process 145

7.3 Design the Testing Stage 148

7.4 Experimental Results 149

7.4.1 Primary Adjustments 149

7.4.2 Results on ISBSG Dataset 150

7.4.3 Scope of Comparisons 150

7.4.4 SCOEE versus other Estimation Models 151

7.4.5 Percentage of Improvement 154

7.4.6 Analysis and Discussion 155

7.5 A Comparative Study 156

7.6 Chapter Summary 158

8 SELECTIVE CLASSIFICATION AND STATISTICAL ANALYSIS BASED EFFORT ESTIMATION MODEL 159

8.1 Introduction 159

8.2 Design the Training Stage 159

8.3 Design the Testing Stage 162

8.4 Experimental Results 164

8.4.1 Primary Adjustments 164

8.4.2 Cross Validation Results 164

8.4.3 Scope of Comparisons 165

8.4.4 SCSAEE versus other Estimation Models 165

8.4.5 Percentage of Improvement 168

8.4.6 Analysis and Discussion 169

8.5 A Comparative Study 170

8.6 Chapter Summary 172

9 EVALUATION OF THE PROPOSED MODELS 173

9.1 Introduction 173

9.2 A Comparison among the Proposed Models 174

9.2.1 Cross Validation Results for the Proposed Models 174

9.2.2 Comparison based on MMRE 175

9.2.3 Comparison based on PRED(0.25) 175

9.2.4 The Range of MRE for the Proposed Models 176

9.3 Evaluation and Discussion 177

9.3.1 Performance Analysis of the Proposed Models 177

9.3.2 Weaknesses and Strengths of the Proposed Models 178

9.4 Chapter Summary 181

10 CONCLUSION AND FUTURE WORK 182

10.1 Introduction 182

10.2 Recommendations to Use the Models 183

10.2.1 Important Issues on FCNNEE 183

10.2.2 Important Issues on RCOEE 184

10.2.3 Important Issues on SCOEE 185

10.2.4 Important Issues on SCSAEE 186

10.3 Threats to Validity 187

10.3.1 Internal Threats 187

10.3.1.1 Evaluation Validity 187

10.3.1.2 Measurement Metrics Validity 188

10.3.1.3 Construct Validity 188

10.3.1.4 Methods Selection Validity 189

10.3.2 External Threats 189

10.4 Thesis Contributions 191

10.4.1 Major Contributions 191

10.4.2 Minor Contributions 193

10.5 Future Work 194

10.5.1 Improving the Performance of FCNNEE 195

10.5.2 Combining SCOEE and SCSAEE 196

10.5.3 Supporting the Software Projects in a Wider Domain 197

10.5.4 Proposing an Advanced Multiple Estimator Model 198

10.6 Chapter Summary 199

REFERENCES **200**

Appendices A - F 218-229