

Contents

List of Figures	xv
List of Tables	xvii
Abbreviations	xix
Symbols	xxi
Preface	xxiii
1 Introduction	1
1.1 Research Background	1
1.1.1 The Existence and Harm of Software Bugs	1
1.1.2 Software Bugs	3
1.1.3 Software Testing	3
1.2 Software Bug Prediction	4
1.2.1 Types of SBP Based on Labeled Software Modules	6
1.2.2 Category of SBP Based on Bug Values	7
1.3 Brief Introduction of studied SBP Scenarios	8
1.3.1 Classification-Based Supervised SBP	8
1.3.2 Classification-Based Unsupervised SBP	9
1.3.3 Regression-Based Unsupervised SBP	10
1.3.4 Classification-Based Unsupervised SBP in Functional Paradigm (FP)	11
1.4 Major Challenges Faced in SBP Scenario	12
1.5 Motivation for the Thesis	14
1.5.1 Motivation for Classification-Based Supervised SBP	14
1.5.2 Motivation for Classification-Based Unsupervised SBP	15
1.5.3 Motivation for Regression-Based Unsupervised SBP Prediction	17
1.5.4 Motivation for Classification-Based Unsupervised SBP in FP	19
1.6 Thesis Objective	21
1.7 Contributions to the Thesis	22
1.7.1 WMV: Classification-Based Supervised SBP	23
1.7.2 TCL/TCLP: Classification-Based Unsupervised SBP	24

1.7.3	MTB/MTBP: Regression-Based Unsupervised SBP	26
1.7.4	UMV: Classification-Based Unsupervised SBP in FP	27
1.8	Thesis Organization	29
2	Related Work and Preliminaries	31
2.1	Related Work on Supervised SBP	31
2.1.1	Classification-Based Supervised SBP	32
2.1.2	Regression-Based Supervised SBP	43
2.2	Related Work on Unsupervised SBP	46
2.2.1	Classification-Based Unsupervised SBP	46
2.2.1.1	Threshold-Based Unsupervised SBP	47
2.2.2	Regression-based Unsupervised SBP	48
2.3	Studies on Classification-Based SBP in FP	49
2.4	Preliminary	50
2.4.1	Software Metrics	50
2.4.2	Datasets Description	52
2.4.2.1	Benchmark Datasets	52
2.4.3	Performance Measures	56
2.4.3.1	Classification-Based Performance Measure	56
2.4.3.2	Regression-Based Performance Measure	58
2.4.3.3	Boxplot Presentation	59
2.4.4	Statistical Test	59
2.4.5	Baseline Methods	60
2.4.6	Hardware and Software Used	62
3	A Supervised SBP Technique Using Reward-Based WMV Ensemble Approach	63
3.1	Introduction	63
3.2	Proposed Approach: WMV	64
3.2.1	Simple Majority Voting Ensemble Method (SMV)	66
3.2.2	Weighted Majority Voting (WMV) Ensemble Method	67
3.2.2.1	Reward-Based Weight Evaluation Scheme	68
3.3	Experimental Design	72
3.3.1	Research Questions (RQ)	72
3.3.2	Datasets (DSs)	74
3.3.3	Experimental Baseline	74
3.3.4	Performance Measures	75
3.4	Experimental Results	75
3.4.1	Results with Respect to Accuracy	79
3.4.2	Results with Respect to FM	81
3.4.3	Results with Respect to MCC	83
3.4.4	Response of Research Questions	85
3.4.5	Result Discussion	92
3.5	Threats to Validity	93

3.6	Conclusion and Future Work	94
4	An Unsupervised SBP Technique Using Threshold Derivation	97
4.1	Introduction	97
4.2	Proposed Approach: TCL/TCLP	99
4.2.1	Threshold Derivation	100
4.2.1.1	Descriptive Statistics and Distribution	100
4.2.1.2	Software Metrics Threshold Calculation	101
4.2.2	Clustering and Labelling	101
4.2.3	Metric Selection and Instance Selection	103
4.2.4	Learning and Prediction	110
4.3	Experimental Design	112
4.3.1	Research Questions	112
4.3.2	Datasets (DSs)	113
4.3.3	Experimental Baselines	113
4.3.4	Performance Measures	114
4.4	Experimental Results	115
4.4.1	Results in Terms of Accuracy	120
4.4.2	Results in Terms of FM	122
4.4.3	Results in Terms of MCC	123
4.4.4	Results After Applying Sampling Technique	123
4.4.5	Answer of Research Questions	124
4.5	Explainability and Interpretability of TCL/TCLP	127
4.5.1	Explanation on Number of Metrics and Instances Selected for Training	127
4.5.2	Number of Data Points Correctly Labeled by Proposed Algorithm (TCLP) Differ from the Rest of the Methods within Each Group	128
4.5.3	Automated Labeling Versus Expert Labeling	130
4.5.4	Significance of Variation in Number of Metrics Selected Across Each DS Group Using MVR	131
4.5.5	Significance of Variation in Percentage of Instances Selected Across Each DS Group Using IVR	131
4.5.6	Execution Time of Different Models	133
4.6	Threats to Validity	133
4.7	Conclusion and Future Work	136
5	An Unsupervised SBCV Prediction Technique Based on Selected Software Metrics	137
5.1	Proposed Approach: MTB/MTBP	139
5.1.1	Unsupervised Metric Selection (UMS) Method	140
5.1.2	Threshold Calculation	143
5.1.3	Software Bug Count Vector (SBCV) Generation	144
5.1.4	Machine Learning Model Building	146

5.1.5	Software Bug Count Vector (SBCV) Prediction	147
5.2	Experimental Design	148
5.2.1	Research Questions	148
5.2.2	Datasets	149
5.2.3	Performance Measures	150
5.3	Experimental Results	150
5.3.1	Result Analysis for RQ1	151
5.3.2	Result Analysis for RQ2	156
5.4	Explainability and Interpretability of SBCV Prediction Models	159
5.4.1	Software Metrics Selected	159
5.4.2	Performance Analysis of MTB/MTBP and Baseline Models with respect to Bug %	159
5.4.3	Performance Analysis of MTB/MTBP and Baseline Models with respect to Software Projects Size	161
5.4.4	Usage Scenario and Limitations of MTB/MTBP	162
5.5	Threat to Validity	163
5.6	Conclusion and Future Work	164
6	An Unsupervised SBP Technique in Functional Paradigm	167
6.1	Introduction	167
6.2	Datasets Description	170
6.2.1	Dataset Collection	172
6.2.2	Software Metric Values Extraction:	174
6.3	Proposed Approach: UMV	175
6.4	Experimental Design	181
6.4.1	Research Questions	181
6.4.2	Performance Measure	183
6.4.2.1	Average Silhouette Width Coefficient	183
6.5	Experimental Results	184
6.5.1	Answer of Research Questions	191
6.5.2	Experimental Findings	191
6.5.3	Discussion	193
6.6	Threat to Validity	194
6.7	Conclusion and Future Work	195
7	Discussion	197
8	Conclusion and Future Work	203
8.1	Conclusions	203
8.2	Future Work	206
A	List of Publications	211
A.1	Journal Papers	211
A.2	Conference Papers	212

B Sample Datasets and Other Results	213
B.1 Distinguishing properties of functional paradigms	220
 Bibliography	 223