

THESIS

**ABSOLUTE CORRELATION BASED WEIGHTED NAÏVE BAYES
FOR SOFTWARE DEFECT PREDICTION**

By

RIZKY TRI ASMONO

P31.2012.01293



A Thesis Submitted in Partial Fulfillment of

The Requirement for the Degree of

Masters of Informatics Engineering

POSTGRADUATE PROGRAM

MASTERS OF INFORMATICS ENGINEERING

DIAN NUSWANTORO UNIVERSITY

SEMARANG

2014



DIAN NUSWANTORO UNIVERSITY

LEGALIZATION STATUS OF THE THESIS

TITLE : ABSOLUTE CORRELATION BASED WEIGHTED NAÏVE
BAYES FOR SOFTWARE DEFECT PREDICTION
NAME : RIZKY TRI ASMONO
STUDENT ID : P31.2012.01293

Permit this Thesis of Masters of Informatics Engineering is stored in Library of Dian Nuswantoro University with terms and usability as follows:

1. The thesis is the property of the Dian Nuswantoro University.
2. Library of Nuswantoro Dian University permitted to make copies for reference purposes only.
3. The library is also permitted to make copies of this thesis as the material exchange between institutions.
4. Provide a $\sqrt{\quad}$ mark accordance with the category of Thesis

☐ Top-secret

☐ Secret

☒ Common

Authenticated by:

Rizky Tri Asmono

Dr. Abdul Syukur, MM

Address:

Kompleks GPM Blok T no. 10 Kendal

Date:

Date:



DIAN NUSWANTORO UNIVERSITY

STATEMENT OF THE AUTHOR

TITLE : ABSOLUTE CORRELATION BASED WEIGHTED NAÏVE
BAYES FOR SOFTWARE DEFECT PREDICTION

NAME : RIZKY TRI ASMONO

STUDENT ID : P31.2012.01293

“I declare and responsible that this thesis is my own work except for excerpts and summaries that I have explained the source respectively. If in the further, there are other parties that claim this thesis, accompanied by sufficient evidence, then I am willing to cancel my Master degree of Informatics Engineering with all the rights and obligations are attached to the title”

Semarang, August 2014

Rizky Tri Asmono

Author



DIAN NUSWANTORO UNIVERSITY

THESIS APPROVAL

TITLE : ABSOLUTE CORRELATION BASED WEIGHTED NAÏVE
BAYES FOR SOFTWARE DEFECT PREDICTION

NAME : RIZKY TRI ASMONO

STUDENT ID : P31.2012.01293

This thesis has been examined and approved,

Semarang, August 2014

Dr. Abdul Syukur, MM

Main Supervisor

Romi Satria Wahono, M.Eng

Co-Supervisor



DIAN NUSWANTORO UNIVERSITY

THESIS LEGALIZATION

TITLE : ABSOLUTE CORRELATION BASED WEIGHTED NAÏVE
BAYES FOR SOFTWARE DEFECT PREDICTION

NAME : RIZKY TRI ASMONO

STUDENT ID : P31.2012.01293

This thesis has been tested and presented to the Board of Examiners on Thesis Trial on July 1, 2014. Based on our perspective, this thesis is sufficient in terms of quality for the purpose of conferment the degree of Masters of Informatics Engineering

Semarang, August 2014

M. Arief Soeleman, M.Kom

Examiner

Consent

Affandy, M.Kom

Examiner

Dr. Abdul Syukur, MM

Director of Postgraduate Program

Dr. Stefanus Santosa, M.Kom

Chairman of the Examiner

ABSTRACT

Develop a flawless software is difficult and often times there are some errors or unknown bugs or unexpected defects, although the software-development methodology has been applied with cautious. The maintenance phase of the software project become very expensive for the developer team and harmful to the users because some flawed software modules. It can be avoided by detecting defects during the development phase of the software project. Software defect prediction will provide an opportunity for the developer team to test modules or files that have a high probability defect. Naïve Bayes can be used to classifying the defect-prone of software, especially defect and non-defect label on software defect prediction. It is one of two best methods that can be used to predict software defects. However, Naïve Bayes assumes all attributes are equally important and are not related each other while, in fact, this assumption is not true in many cases. The absolute value of correlation coefficient has been proposed as weighting method to overcome Naïve Bayes assumptions. The absolute value of the correlation coefficient can be used to measure the relevance between attributes. The relevance between attributes shows the significant of these attributes in the probability, more influential an attribute in probabilities, higher the weight value. In this study, Absolute Correlation Weighted Naïve Bayes and Absolute Correlation Weighted Attribute-class Naïve Bayes have been proposed. The proposed methods have been evaluated on NASA MDP Dataset and used area under the curve (AUC) as evaluation method. The AUC values are compared with parametric test using t-test and non-parametric test using Friedman test. P-value of t-test on Naïve Bayes and Absolute Correlation Weighted Naïve Bayes is 0.008, whereas p-value of t-test on Naïve Bayes and Absolute Correlation Weighted Attribute-class Naïve Bayes is 0.019. P-value of Friedman test is 0.02. The results of parametric test and non-parametric show that the proposed method can improve the performance of Naïve Bayes for classifying defect-prone on software defect prediction.

Keywords : Software Defect Prediction, Classification, Naïve Bayes, Weighted Naïve Bayes, Weighted Attribute-class Naïve Bayes, Correlation Coefficient, Absolute Correlation

xvii + 147 pages; 44 figures; 71 tables; 27 attachments

Bibliography: 76 (1940-2013)

ABSTRAK

Mengembangkan perangkat lunak yang sempurna itu sulit dan sering kali ada beberapa kesalahan atau *bug* yang tidak diketahui atau cacat yang tidak terduga, meskipun metodologi pengembangan perangkat lunak telah diterapkan dengan hati-hati. Tahap pemeliharaan pada proyek perangkat lunak menjadi sangat mahal bagi tim pengembang dan merugikan bagi pengguna karena beberapa modul perangkat lunak yang cacat. Hal ini dapat dihindari dengan mendeteksi cacat selama tahap pengembangan pada proyek perangkat lunak. Prediksi cacat perangkat lunak akan memberikan kesempatan bagi tim pengembang untuk menguji modul atau *file* yang memiliki probabilitas cacat tinggi. Naïve Bayes dapat digunakan untuk mengklasifikasikan kerawanan cacat perangkat lunak, terutama label cacat dan tidak cacat pada prediksi cacat perangkat lunak. Naïve Bayes adalah salah satu dari dua metode terbaik yang dapat digunakan untuk memprediksi cacat perangkat lunak. Namun, Naïve Bayes menganggap semua atribut sama pentingnya dan tidak berhubungan satu sama lain. Padahal kenyataannya, asumsi ini tidak benar dalam banyak kasus. Nilai absolut dari koefisien korelasi diusulkan sebagai metode pembobotan untuk mengatasi asumsi Naïve Bayes. Nilai absolut dari koefisien korelasi dapat digunakan untuk mengukur relevansi antara atribut. Relevansi antara atribut menunjukkan signifikan dari atribut-atribut ini dalam probabilitas, lebih berpengaruh atribut dalam probabilitas, semakin tinggi nilai bobot. Pada penelitian ini, *Absolute Correlation Weighted Naïve Bayes* dan *Absolute Correlation Weighted Attribute-class Naïve Bayes* diusulkan. Metode yang diusulkan tersebut dievaluasi pada NASA MDP Dataset dan menggunakan *area under the curve* (AUC) sebagai metode evaluasi. Nilai AUC dibandingkan dengan uji parametrik menggunakan t-test dan uji non-parametrik menggunakan Friedman Test. P-value dari t-test pada Naïve Bayes dan Absolute Correlation Weighted Naïve Bayes adalah 0.008, sedangkan p-value dari t-test pada Naïve Bayes dan Absolute Correlation Weighted Attribute-class Naïve Bayes pada 0.019. P-value dari Friedman test adalah 0.02. Hasil uji parametrik dan non-parametrik menunjukkan bahwa metode yang diusulkan dapat meningkatkan kinerja Naïve Bayes untuk mengklasifikasikan kerawanan cacat pada prediksi cacat perangkat lunak.

Kata Kunci : Prediksi Cacat Perangkat Lunak, Klasifikasi, Naïve Bayes, Weighted Naïve Bayes, Weighted Attribute-class Naïve Bayes, Koefisien Korelasi, Koefisien Korelasi Absolut

xvii + 147 halaman; 44 gambar; 71 tabel; 27 lampiran

Referensi: 76 (1940-2013)

ACKNOWLEDGMENTS

Thesis with title “Absolute Correlation based Weighted Naïve Bayes for Software Defect Prediction” unable to be finished without the support, counseling and motivation directly or indirectly from all sides. Therefore, I would like to convey gratitude to:

1. Allah SWT for giving health and strength to complete this thesis.
2. Beloved parents, for moral and material support.
3. Dr. Ir. Edi Noersasongko, M.Kom as Rector of Dian Nuswantoro University.
4. Dr. Abdul Syukur, MM as Director of Postgraduate Program of Dian Nuswantoro University, at the same time as the supervisor who has provided knowledge, counseling and encouragement in drafting this thesis.
5. Romi Satria Wahono, M.Eng as the supervisor who has provided knowledge, counseling and encouragement in drafting this thesis.
6. The entire lecturers and administrative staff in the Postgraduate Program at Dian Nuswantoro University who have assisted during the study.
7. Tyas Setiyorini who gives a special motivation.
8. Group Intelligent Systems, who have struggled together.
9. MTI XXII friends who unable to mention individually.

This thesis is still a lot of deficiencies. The advice and constructive criticism are needed in future studies. Hopefully, this thesis can be beneficial for knowledge development.

Semarang, August 2014

Author

TABLE OF CONTENTS

LEGALIZATION STATUS OF THE THESIS	i
STATEMENT OF THE AUTHOR	ii
THESIS APPROVAL	iii
THESIS LEGALIZATION	iv
ABSTRACT.....	v
ABSTRAK	vi
ACKNOWLEDGMENTS	vii
TABLE OF CONTENTS	viii
LIST OF FIGURES	xiii
LIST OF TABLES	xv
CHAPTER 1 INTRODUCTION	1
1.1 Research Background	1
1.2 Research Problems	3
1.3 Research Question	3
1.4 Research Objectives	4
1.5 Relationship between Research Problems, Questions and Objectives	4
1.6 Research Contributions	5
CHAPTER 2 LITERATURE REVIEW	7
2.1 Software Engineering	7
2.1.1 Introduction to Software Engineering	7
2.1.2 Software-Development Life Cycle	8
2.1.3 Software Testing	10
2.2 Software Defect Prediction	11
2.2.1 Software Defect	11

2.2.2	Software Defect Prediction	12
2.2.3	Software Defect Prediction Methods	14
2.3	Bayesian Network.....	15
2.3.1	Introduction to Bayesian Network	15
2.3.2	Naïve Bayes	16
2.3.3	Weighted Naïve Bayes	17
2.4	Correlation Coefficient	17
2.5	Evaluation Methods of Classification Algorithm	18
2.5.1	Confusion Matrix	18
2.5.2	Area of Under Curve (AUC).....	19
2.5.3	Significant Test	20
2.5.3.1	Parametric Test using T-test	20
2.5.3.2	Non-Parametric Test using Friedman Test	21
2.6	Related Research.....	22
2.6.1	Lin et al. [67] Model	22
2.6.2	Taheri et al. [17] Model	24
2.6.3	Wu et al. [6] Model	26
2.6.4	Summary of the Related Research	28
2.7	Research Framework	31
CHAPTER 3 RESEARCH METHOD		34
3.1	Research Phases	34
3.2	Research Scope	35
3.3	Research Design.....	36
3.4	Data Gathering	38
3.5	Data Pre-processing	42
3.6	Proposed Model	43

3.6.1	Absolute Correlation Weighted Naïve Bayes Model.....	45
3.6.2	Absolute Correlation Weighted Attribute-class Naïve Bayes Model ..	48
3.7	Model Testing and Experiment.....	51
3.8	Result Evaluation	52
CHAPTER 4	RESULTS AND ANALYSIS	53
4.1	Results.....	53
4.1.1	Training Calculation Results of Naïve Bayes	53
4.1.2	Experiment Results on Naïve Bayes	59
4.1.3	Training Calculation Results of Absolute Correlation based Weighted Naïve Bayes	69
4.1.4	Experiment Results on Absolute Correlation Based Weighted Naïve Bayes.....	74
4.1.5	Training Calculation Results of Absolute Correlation based Weighted Attribute-class Naïve Bayes.....	102
4.1.6	Experiment Results on Absolute Correlation Based Weighted Attribute-class Naïve Bayes	106
4.2	Analysis.....	135
4.2.1	Parametric Test.....	137
4.2.2	Non-Parametric Test	140
4.2.3	Summary of Analysis	143
CHAPTER 5	CONCLUSION	145
5.1	Conclusion	145
5.2	Future Works	146
Reference.....		148
APPENDICES		154
A.	Naïve Bayes Model.....	154
A1.	Dataset CM1	154

A2.	Dataset JM1.....	155
A3.	Dataset KC1	156
A4.	Dataset KC3	156
A5.	Dataset MC1	158
A6.	Dataset MC2	159
A7.	Dataset PC1	160
A8.	Dataset PC3.....	162
A9.	Dataset PC4.....	163
A10.	Dataset PC5.....	165
B.	Absolute Correlation based Weighted Naïve Bayes Model and Absolute Correlation based Weighted Attribute-class Naïve Bayes	166
B1.	Dataset CM1	166
B2.	Dataset JM1.....	168
B3.	Dataset KC1	169
B4.	Dataset KC3	170
B5.	Dataset MC1	171
B6.	Dataset MC2	172
B7.	Dataset PC1	174
B8.	Dataset PC3.....	175
B9.	Dataset PC4.....	177
B10.	Dataset PC5.....	179
C.	Source Code	180
C1.	MainClass.java.....	180
C2.	CrossValidation.java	180
C3.	BubbleSort.java.....	182
C4.	AUC.java.....	182

C5.	NaiveBayes.java.....	183
C6.	AcWnb.java.....	185
C7.	AcWacnb.java	187

LIST OF FIGURES

Figure 1.1 Relationship between Research Problems (RP), Research Questions (RQ) and Research Contributions (RC)	6
Figure 2.1 Block Diagram of Lin et al. [67] Model	24
Figure 2.2 Block Diagram of Taheri et al. [17] Model	26
Figure 2.3 Block Diagram of Wu et al. [6] Model	28
Figure 2.4 Research Framework	31
Figure 3.1 Research Phases	34
Figure 3.2 Research Scope	36
Figure 3.3 Research Design	37
Figure 3.4 Block Diagram of Proposed Model	44
Figure 3.5 Flow Chart of AC-WNB Method	47
Figure 3.6 Flow Chart of AC-WACNB Method	50
Figure 4.1 ROC Curve of Naïve Bayes on Dataset CM1	60
Figure 4.2 ROC Curve of Naïve Bayes on Dataset JM1	61
Figure 4.3 ROC Curve of Naïve Bayes on Dataset KC1	62
Figure 4.4 ROC Curve of Naïve Bayes on Dataset KC3	63
Figure 4.5 ROC Curve of Naïve Bayes on Dataset MC1	64
Figure 4.6 ROC Curve of Naïve Bayes on Dataset MC2	65
Figure 4.7 ROC Curve of Naïve Bayes on Dataset PC1	66
Figure 4.8 ROC Curve of Naïve Bayes on Dataset PC3	67
Figure 4.9 ROC Curve of Naïve Bayes on Dataset PC4	68
Figure 4.10 ROC Curve of Naïve Bayes on Dataset PC5	69
Figure 4.11 ROC Curve of AC-WNB on Dataset CM1	74
Figure 4.12 ROC Curve of AC-WNB on Dataset JM1	77
Figure 4.13 ROC Curve of AC-WNB on Dataset KC1	79
Figure 4.14 ROC Curve of AC-WNB on Dataset KC3	81
Figure 4.15 ROC Curve of AC-WNB on Dataset MC1	84
Figure 4.16 ROC Curve of AC-WNB on Dataset MC2	87
Figure 4.17 ROC Curve of AC-WNB on Dataset PC1	90
Figure 4.18 ROC Curve of AC-WNB on Dataset PC3	93

Figure 4.19 ROC Curve of AC-WNB on Dataset PC4.....	96
Figure 4.20 ROC Curve of AC-WNB on Dataset PC5.....	99
Figure 4.21 ROC Curve of AC-WACNB on Dataset CM1	107
Figure 4.22 ROC Curve of AC-WACNB on Dataset JM1	110
Figure 4.23 ROC Curve of AC-WACNB on Dataset KC1.....	112
Figure 4.24 ROC Curve of AC-WACNB on Dataset KC3.....	114
Figure 4.25 ROC Curve of AC-WACNB on Dataset MC1	117
Figure 4.26 ROC Curve of AC-WACNB on Dataset MC2.....	120
Figure 4.27 ROC Curve of AC-WACNB on Dataset PC1	123
Figure 4.28 ROC Curve of AC-WACNB on Dataset PC3	126
Figure 4.29 ROC Curve of AC-WACNB on Dataset PC4	129
Figure 4.30 ROC Curve of AC-WACNB on Dataset PC5	132
Figure 4.31 Performance (Accuracy) of the Models.....	136
Figure 4.32 Performance (AUC) of the Models.....	136
Figure 4.33 Average Rank Based on AUC	141

LIST OF TABLES

Table 1.1 Research Problems (RP) - Research Questions (RQ) - Research Objectives (RO)	5
Table 2.1 Confusion Matrix	19
Table 2.2 Summary of the Related Research on Weighted Naïve Bayes Classification	30
Table 3.1 Dataset Specifications	39
Table 3.2 . The Independent Variables of NASA MDP Dataset	40
Table 3.3 The Difference with Related Research	45
Table 3.4 The specification of The Computer in Experiments	52
Table 4.1 Example Dataset of JM1	54
Table 4.2 Naive Bayes Model on Example Dataset.....	58
Table 4.3 Experiment Result of Naïve Bayes with Dataset CM1	59
Table 4.4 Experiment Result of Naïve Bayes with Dataset JM1	60
Table 4.5 Experiment Result of Naïve Bayes with Dataset KC1	61
Table 4.6 Experiment Result of Naïve Bayes with Dataset KC3	62
Table 4.7 Experiment Result of Naïve Bayes with Dataset MC1	63
Table 4.8 Experiment Result of Naïve Bayes with Dataset MC2.....	64
Table 4.9 Experiment Result of Naïve Bayes with Dataset PC1	65
Table 4.10 Experiment Result of Naïve Bayes with Dataset PC3	66
Table 4.11 Experiment Result of Naïve Bayes with Dataset PC4	67
Table 4.12 Experiment Result of Naïve Bayes with Dataset PC5	68
Table 4.13 Absolute Correlation based Weighted Naïve Bayes Model on Example Dataset.....	72
Table 4.14 Experiment Result of AC-WNB with Dataset CM1	74
Table 4.15 Correlation Values (CV) and Weight (W) of AC-WNB on Dataset CM1	75
Table 4.16 Experiment Result of AC-WNB with Dataset JM1	77
Table 4.17 Correlation Values (CV) and Weight (W) of AC-WNB on Dataset JM1	78
Table 4.18 Experiment Result of AC-WNB with Dataset KC1	79
Table 4.19 Correlation Values (CV) and Weight (W) of AC-WNB on Dataset KC180	

Table 4.20 Experiment Result of AC-WNB with Dataset KC3	81
Table 4.21 Correlation Values (CV) and Weight (W) of AC-WNB on Dataset KC382	
Table 4.22 Experiment Result of AC-WNB with Dataset MC1	84
Table 4.23 Correlation Values (CV) and Weight (W) of AC-WNB on Dataset MC1	85
Table 4.24 Experiment Result of AC-WNB with Dataset MC2.....	87
Table 4.25 Correlation Values (CV) and Weight (W) of AC-WNB on Dataset MC2	88
Table 4.26 Experiment Result of AC-WNB with Dataset PC1	90
Table 4.27 Correlation Values (CV) and Weight (W) of AC-WNB on Dataset PC1	91
Table 4.28 Experiment Result of AC-WNB with Dataset PC3	93
Table 4.29 Correlation Values (CV) and Weight (W) of AC-WNB on Dataset PC3	94
Table 4.30 Experiment Result of AC-WNB with Dataset PC4	96
Table 4.31 Correlation Values (CV) and Weight (W) of AC-WNB on Dataset PC4	97
Table 4.32 Experiment Result of AC-WNB with Dataset PC5	99
Table 4.33 Correlation Values (CV) and Weight (W) of AC-WNB on Dataset PC5	100
Table 4.34 Absolute Correlation based Weighted Attribute-class Naive Bayes Model on Example Dataset.....	105
Table 4.35 Experiment Result of AC-WACNB with Dataset CM1	106
Table 4.36 Weight on Class “N” (W (N)) and Weight on Class “Y” (W (Y)) of AC-WACNB on Dataset CM1	108
Table 4.37 Experiment Result of AC-WACNB with Dataset JM1	110
Table 4.38 Weight on Class “N” (W (N)) and Weight on Class “Y” (W (Y)) of AC-WACNB on Dataset JM1	111
Table 4.39 Experiment Result of AC-WACNB with Dataset KC1	112
Table 4.40 Weight on Class “N” (W (N)) and Weight on Class “Y” (W (Y)) of AC-WACNB on Dataset KC1	113
Table 4.41 Experiment Result of AC-WACNB with Dataset KC3	114
Table 4.42 Weight on Class “N” (W (N)) and Weight on Class “Y” (W (Y)) of AC-WACNB on Dataset KC3	115
Table 4.43 Experiment Result of AC-WACNB with Dataset MC1	117

Table 4.44 Weight on Class “N” (W (N)) and Weight on Class “Y” (W (Y)) of AC-WACNB on Dataset MC1.....	118
Table 4.45 Experiment Result of AC-WACNB with Dataset MC2	120
Table 4.46 Weight on Class “N” (W (N)) and Weight on Class “Y” (W (Y)) of AC-WACNB on Dataset MC2.....	121
Table 4.47 Experiment Result of AC-WACNB with Dataset PC1.....	123
Table 4.48 Weight on Class “N” (W (N)) and Weight on Class “Y” (W (Y)) of AC-WACNB on Dataset PC1	124
Table 4.49 Experiment Result of AC-WACNB with Dataset PC3.....	126
Table 4.50 Weight on Class “N” (W (N)) and Weight on Class “Y” (W (Y)) of AC-WACNB on Dataset PC3	127
Table 4.51 Experiment Result of AC-WACNB with Dataset PC4.....	129
Table 4.52 Weight on Class “N” (W (N)) and Weight on Class “Y” (W (Y)) of AC-WACNB on Dataset PC4	130
Table 4.53 Experiment Result of AC-WACNB with Dataset PC5.....	132
Table 4.54 Weight on Class “N” (W (N)) and Weight on Class “Y” (W (Y)) of AC-WACNB on Dataset PC5	133
Table 4.55 Summary of Performance of the NB, AC-WNB and AC-WACNB Model	135
Table 4.56 T-Test: Paired Two Samples for Means of AUC of NB and AC-WNB	137
Table 4.57 T-Test: Paired Two Samples for Means of AUC of NB and AC-WACNB	138
Table 4.58 T-Test: Paired Two Samples for Means of AUC of AC-WNB and AC-WACNB.....	139
Table 4.59 P Values of AUC Comparison by Using T-test.....	139
Table 4.60 Average Rank of Each Model Based on its AUC.....	140
Table 4.61 Friedman Test of AUC.....	141
Table 4.62 Multiple Pairwise Comparisons Using Nemenyi Procedure.....	142
Table 4.63 Table of Pairwise Differences.....	142
Table 4.64 Significant Difference	142

CHAPTER 1

INTRODUCTION

1.1 Research Background

Software is computer programs and related documentation. Software products can be expanded for a specific customer or may be developed for the common marketplace in accordance with the functions and needs. Develop a flawless software is difficult and often times there are some errors or bugs, unknown or unexpected defects, although the software-development methodology has been applied with cautious [1]. The maintenance phase of the software project become very expensive for the developer team and harmful to the users because some flawed software modules. Surely, it can be avoided by detecting defects during the development phase of software projects. Defect prediction will provide an opportunity for the developer team to test modules or files that have a high probability defect. The completion of defect prediction problems currently focusing on 1) estimating the number of defects in the existing software systems, 2) discovering defect associations and 3) classifying the defect-prone of software, specially defect and non-defect label [2]. The things that detrimental to users and developer team can be avoided as early as possible with a software defect prediction.

For classifying defect-prone, Hall conducted an investigation on software defect prediction [3]. Hall compared C4.5, Decision Tree, Logistic Regression, Naïve Bayes, Neural Network, etc. The results of the investigation showed the two best methods that can be used to predict software defects are Naive Bayes (NB) and Logistic Regression. Logistic Regression is a statistical probabilistic classification method. The advantages of logistic regression are computationally inexpensive, slight to implement and mild to interpret knowledge representation. The disadvantages of logistic regression are prone to under fitting and may have a low accuracy [4]. Naïve Bayes is a modest probabilistic classifier. It is very comfortable because it does not require any complicated parameter estimation. Therefore, NB ready to be used for large amounts of data. Moreover, NB is also very facile to explain so the users who do not have the

technological classification capability can understand the reason why the classification was made [5]. However, NB assumes all attributes are equally important and are not related each other while, in fact, this assumption is not true in many cases [6] [7] [8]. The assumption made by NB can be detrimental to its performance in real data mining applications.

Naïve Bayes assumes that all the attributes are not dependent on each other, in fact, the class depends on others attribute. Naïve Bayes also assumes the relationship between class and one attribute is as strong as the relationship between class and other attribute [7]. This case clearly unrealistic. For example, data set for evaluate risk of loan application, it does not seem fair to assume that between income, age and education levels are equally important. The assumption made by Naïve Bayes harming the performance of classification in reality [9]. This assumption can cause the unwanted error increase.

Many methods have been developed to cover this attribute independence assumption. Jiang [10] [11] categorized solutions to these problems into five: 1) Attribute selection, 2) Local Learning, 3) Attribute Weighting, 4) Instance Weighting and 5) Structure Extension. Previous researchers have proposed many useful methods to evaluate the important attributes. Ratanamahatana used Decision Tree as feature selection on Naïve Bayes [12]. Zhang used Gain Ratio to determine attribute weight on Naïve Bayes [13]. Wu used Differential Evolution Algorithm to weighting attribute [6]. Decision Tree-based attribute weighting for Naïve Bayes proposed by Hall [14]. Averaged n-Dependence Estimators (AnDE) was proposed by Webb [15]. AnDE was developed from Averaged One-Dependence estimators (Aode) which reduce the NB independence assumption [9]. Zaidi proposed Weighting attributes to Alleviate Naive Bayes Independence Assumption (WANBIA) by set all weights to a single value [16]. Taheri proposed Attribute Weighted Naive Bayes (AWNB) which define more than one weight for each attribute [17]. AWINB limited to binary classification.

Because of the attribute does not has the same role, some of them more important than the others, one of the ways to develop Naïve Bayes is set a different weight value of each attributes. It is becoming the main idea of the new algorithm called Weighting