

**ENHANCEMENT OF NEW SMOOTH SUPPORT VECTOR MACHINES  
FOR CLASSIFICATION PROBLEMS**

SANTI WULAN PURNAMI

Thesis submitted in fulfillment of the requirements  
for the award of the degree of  
Doctor of Philosophy in Computer Science

Faculty of Computer System and Software Engineering  
UNIVERSITI MALAYSIA PAHANG

JUNE 2011

### **SUPERVISOR'S DECLARATION**

I hereby declare that I have checked this thesis and in my opinion, this thesis is adequate in terms of scope and quality for the award of the degree of Doctor of Philosophy in Computer Science.

Signature :

Name of Supervisor : PROF. DR. JASNI MOHAMAD ZAIN

Position : DEAN, FACULTY OF COMPUTER SYSTEM AND  
SOFTWARE ENGINEERING UNIVERSITI MALAYSIA PAHANG

Date : 21 JUNE 2011

## STUDENT'S DECLARATION

I hereby declare that the work in this thesis is my own except for quotations and summaries which have been duly acknowledged. The thesis has not been accepted for any degree and is concurrently submitted for award of other degree.

Signature :  
Name : SANTI WULAN PURNAMI  
ID Number : PCC 07001  
Date : 21 JUNE 2011

## PREAMBLE

Say: Though the ocean became ink for the words of my Lord, verily the sea would be used up before the words of my Lord were exhausted, even if we added another ocean like it, for its aid (*Al Qur'an, Al Kahfi 18:109*)

And if all the trees on the earth were pens and the ocean (were ink), with seven oceans behind it to add to its (supply), yet would not the words of Allah be exhausted (in the writing). For Allah is exalted in power, full of wisdom. (*Al Qur'an, Lukman 31:27*)

Behold! In the creation of the heavens and the earth, and the alternation of night and day, there are indeed signs for men of understanding. Men who celebrate the praises of Allah, standing, sitting and lying down on their sides, and contemplate the (wonders of) creation in the heavens and the earth, (with the thought): "Our lord! Not for naught hast thou created (all) this! Glory to thee! Give us salvation from the penalty of the fire (*Al Qur'an, Ali Imron 3:190-191*)

Is one who worships devoutly during the hours of the night prostrating himself or standing (in adoration), who takes heed of the hereafter, and who places his hope in the mercy of his Lord- (like one who does not)? Say: "Are those equal, those who know and those who do not know? It is those who are endued with understanding that receive admonition. (*Al Qur'an, Az Zumar 39:9*)

**Dedicated to my parents and my family**

## **ACKNOWLEDGMENTS**

Bismillah Ar-Rahman Ar-Rahim. I would like to thank to ALLAH S.W.T for His power to make this research possible. I would like to express my deep gratitude to PM. Jasni Mohamad Zain for his guidance, advice and invaluable assistance during my studies at the University Malaysia Pahang. I also like to express my thanks to Prof. Abdullah Embong for having served as my advisor during 2007-2009. His expertise and guidance were it great importance for this research.

I would like to thank to Institut Teknologi Sepuluh Nopember (ITS) Surabaya for giving me a chance to study in Malaysia. I would also like to acknowledge the support of University Malaysia Pahang under GRS no. 070147.

Finally, I would like to thank my parents, my parents in law and all of my family for their strong support of my study, for their courage, wisdom and love. Special thanks to my husband, Yoyok Setyo Hadiwidodo and my children Fauzan, Niswah, Syahida and Humaida for their understanding and strong support during my study.

## TABLE OF CONTENTS

	<b>Page</b>
<b>SUPERVISOR’S DECLARATION</b>	ii
<b>STUDENTS’S DECLARATION</b>	iii
<b>PREAMBLE</b>	iv
<b>DEDICATION</b>	v
<b>ACKNOWLEDGMENTS</b>	vi
<b>ABSTRACT</b>	vii
<b>ABSTRAK</b>	viii
<b>TABLE OF CONTENTS</b>	ix
<b>LIST OF TABLES</b>	xiii
<b>LIST OF FIGURES</b>	xvi
<b>LIST OF SYMBOLS</b>	xviii
<b>LIST OF ABBREVIATIONS</b>	xx
<b>CHAPTER 1      INTRODUCTION</b>	
1.1      Research Background	1
1.2      Problem Statement	7
1.3      Research Objectives	8
1.4      Research Scopes	8
1.5      Outline of the Thesis	9
<b>CHAPTER 2      LITERATURE REVIEW</b>	

2.1	Introduction	12
2.2	Support Vector Machines	12
	2.2.1 Linear Support Vector Machines	13
	2.2.2 Nonlinear Support Vector Machines	16
	2.2.3 Parameter Selection	19
2.3	Smooth Support Vector Machines and its Variants	24
	2.3.1 Polynomial Smooth Support Vector Machines	29
	2.3.2 Spline Smooth Support Vector Machines	31
2.4	Reduced Support Vector Machines (RSVM)	32
	2.4.1 Systematic sampling RSVM	34
	2.4.2 Clustering RSVM	35
2.5	Multi-Class Support Vector Machines	35
	2.5.1 One-against-all (OAA) Method	35
	2.5.2 One-against-One (OAO) Method	38
2.6	Summary	40

### **CHAPTER 3 RESEARCH METHODOLOGY**

3.1	Introduction	41
3.2	Developing SSVM Using a New Smooth Function	41
	3.2.1 Get a New Smooth Function	41
	3.2.2 Construct the New SSVM	42
3.3	Extending New SSVM to Multiclass Classification	50
	3.3.1 Implementation	50
	3.3.2 Description of Dataset	51
3.4	Extending SSVM to RSVM	52
	3.4.1 Implementation	52
	3.4.2 Description of Dataset	55
3.5	Summary	60

### **CHAPTER 4 A NEW SMOOTH SUPPORT VECTOR MACHINES USING MULTIPLE KNOT SPLINE FUNCTION**

4.1	Introduction	61
-----	--------------	----



4.2	Multiple Knot Spline (MKS) Function	61
4.2.1	Formulation of Multiple Knot Spline (MKS) Function	62
4.2.2	Performance Analysis of MKS Function	65
4.2.3	Comparison with other Smoothing Function	68
4.3	A New Smooth Support Vector Machines using MKS Function	72
4.3.1	Multiple Knot Spline Smooth Support Vector Machines (MKS-SSVM)	72
4.3.2	Implementation of MKS-SSVM	73
4.4	Experiment and Analysis	78
4.4.1	Diabetes Disease Dataset	78
4.4.2	Heart Disease Dataset	81
4.4.3	Breast Cancer Diagnosis Dataset	84
4.4.4	Breast Cancer Prognosis Dataset	87
4.4.5	Discussion	89
4.5	Extension of MKS-SSVM to Multiclass Problem	90
4.5.1	One against All (OAA) Method	91
4.5.2	One against One (OAO) Method	92
4.5.3	Implementation	94
4.6	Summary	95

## **CHAPTER 5           ALTERNATIVE REDUCED SMOOTH SUPPORT VECTOR MACHINE BASED ON CLUSTERING TECHNIQUE**

5.1	Introduction	97
5.2	Overview Clustering	98
5.3	Reduced Support Vector Machines (RSVM) based on $k$ -Modes Clustering for Categorical Data	101
5.3.1	$K$ -Modes Clustering	101
5.3.2	RSVM based on $k$ -Modes Clustering	104
5.3.3	Experiments and Results	105
5.3.4	Comparison KMo-RSVM and RSVM	109
5.4	RSVM based on $k$ -Prototype Clustering for Mixed Data	112
5.4.1	$K$ -Prototype Clustering	112
5.4.2	RSVM based on $K$ -Prototype Clustering	114
5.4.3	Experiment and Analysis	114
5.4.4	Comparison KPro-RSVM and RSVM	117
5.5	Summary	120

## **CHAPTER 6      CONCLUSION**

6.1	Conclusion of the Study	121
6.2	Contribution and Achievements	123
6.3	Recommendations for Future Research	124

<b>REFERENCES</b>		125
-------------------	--	-----

<b>APPENDICES</b>		133
-------------------	--	-----

A.	List of Publications	133
B1	Matlab Code for MKS-SSVM Training	135
B2	Matlab Code for MKS-SSVM Prediction	141
B3	Matlab Code For Multiclass MKS-SSVM	142
B4	Matlab Code for KMo-RSVM	144
B5	Matlab Code for KPRo-RSVM	149
C1	List of Datasets	151
C2	Sample of Medical Dataset (Diabetes Disease Dataset)	152
C3	Sample of Multiclass Dataset (Iris Dataset)	167
C4	Sample of Mixed Attributes Dataset (German Credit Dataset)	171
C5	Attributes of German Credit Dataset	193
C6	Attributes of Census Dataset	196

## LIST OF TABLES

<b>Table No.</b>	<b>Title</b>	<b>Page</b>
3.1	Representation of confusion matrix	43
3.2	Description of four medical dataset	47
3.3	Statistical analysis of diabetes disease dataset	48
3.4	Brief statistical analysis of heart disease dataset	49
3.5	A WDBC cell nuclei characteristic attributes	50
3.6	Descriptions of multiclass dataset	52
3.7	Descriptions of four categorical dataset	55
3.8	Descriptions of three mixed attributes dataset	59
4.1	The points and functions at interval $\left[\frac{-2}{5k}, \frac{1}{5k}\right]$	63
4.2	Comparison of four smoothing functions	70
4.3	Comparison of previous functions and MKS function	71
4.4	Ten-fold cross validation accuracy of diabetes disease	79
4.5	Confusion matrix for 90-10% training-test partition and 80-20% training-test partition of diabetes disease	79
4.6	Obtained classification accuracy, sensitivity and specificity for 90-10% training-test partition and 80-20% training-test partition of diabetes disease	80
4.7	Classification accuracies obtained with our method and other Classifiers for diabetes disease	81
4.8	Ten-fold cross validation accuracy of heart disease	82
4.9	Confusion matrix for 90-10% training-test partition and 80-20% training-test partition of heart disease	82
4.10	Obtained classification accuracy, sensitivity and specificity for 90-10% training-test partition and 80-20% training-test partition of heart disease	83

4.11	Classification accuracies obtained with our method and other classifiers of heart disease	84
4.12	Ten-fold cross validation accuracy of breast cancer diagnosis	85
4.13	Confusion matrix for 90-10% training-test partition and 80-20% training-test partition of breast cancer diagnosis	85
4.14	Obtained classification accuracy, sensitivity and specificity breast cancer diagnosis	86
4.15	Classification accuracies obtained with our method and other classifiers of breast cancer diagnosis dataset	86
4.16	Ten-fold cross validation accuracy of breast cancer prognosis dataset	87
4.17	Confusion matrix for 90-10% training-test partition and 80-20% training-test partition of breast cancer prognosis dataset	87
4.18	Obtained classification accuracy, sensitivity and specificity of breast cancer prognosis dataset	88
4.19	Classification accuracies obtained with our method and other classifiers of breast cancer prognosis dataset	88
4.20	Comparison classification accuracy between other classifiers, SSVM and MKS-SSVM on four medical dataset	89
4.21	Classification accuracy and computation time of MKS-SSVM on multiclass dataset using OAA and OAO	95
5.1	Testing accuracy and computational time of KMo-RSVM and SSVM on breast cancer dataset	106
5.2	Testing accuracy and computational time of KMo-RSVM and SSVM on tic-tac-toe dataset	107
5.3	Testing accuracy and computational time of KMo-RSVM and SSVM on chess dataset	108
5.4	Testing accuracy and computational time of KMo-RSVM and SSVM on mushroom dataset	109
5.5	Testing accuracy of RSVM on four categorical attributes dataset	110
5.6	Computational time of RSVM on four categorical attributes dataset	110
5.7	Comparisons accuracy, time and reduced set ( $\bar{m}$ ) of variants	

	Methods (SSVM, RSVM, KMo-RSVM) on four categorical attributes dataset	111
5.8	Testing accuracy and computational time of KPro-RSVM and SSVM on credit dataset	115
5.9	Testing accuracy and computational time of KPro-RSVM and SSVM on german dataset	116
5.10	Testing accuracy and computational time of KPro-RSVM and SSVM on census dataset	117
5.11	Testing accuracy of RSVM on three mixed attributes dataset	118
5.12	Computational time of RSVM on three mixed attributes dataset	119
5.13	Comparisons accuracy, time and reduced set (m) of variants methods (SSVM, RSVM, KPro-RSVM) on three mixed attributes dataset	119

## LIST OF FIGURES

<b>Figure No.</b>	<b>Title</b>	<b>Page</b>
1.1	Development process model	9
1.2	Research structure	11
2.1	Linearly separable case	14
2.2	Nonlinearly separable case	15
2.3	Mapping the input space into a high dimensional feature space	17
2.4	The DOE sampling	21
2.5	The UD sampling patterns	22
2.6	The nested UD based model selection method (13 points in First stage and 9 points in second stage)	23
2.7	Smooth approximation to the plus function	26
2.8	The main idea of Systematic Sampling RSVM Algorithm	34
2.9	Discrete decision function	37
2.10	Continuous decision function	38
2.11	One-Against-One Algorithm	39
3.1	Implementation of new SSVM	44
3.2	Steps of UD method	45
3.3	Flowchart of Newton Armijo Algorithm	46
3.4	Steps of KMo-RSVM	53
3.5	Steps of KPro-RSVM	54
4.1	The three order spline function at $k = 10$	62
4.2	Comparison figure of the integral of sigmoid function $p(x, k)$ , three order spline function $t(x, k)$ and MKS function $m(x, k)$	65
4.3	Comparison of previous functions and MKS function at $k = 10$	71
4.4	Implementation of MKS-SSVM	77

4.5	Comparison classification accuracy between other classifiers, SVM and MKS-SSVM on four medical dataset	90
5.1	Flowchart of $k$ -means algorithm	99
5.2	An example of the $k$ -means clustering	100
5.3	Comparison Testing accuracy and computational time of KMo-RSVM and SSVM on breast cancer dataset	106
5.4	Comparison Testing accuracy and computational time of KMo-RSVM and SSVM on tic-tac-toe dataset	107

## LIST OF SYMBOLS

$A$	Data points
$A_+$	Class positive
$A_-$	Class negative
$D$	Diagonal matrix with +1 or -1 along its diagonal to specify the membership of each point.
$w$	Normal vector of the hyperplane
$\gamma$	Distance hyperplane to the origin
$y$	Slack variable
$e$	Column vector of one of arbitrary dimension
$\nu$	Regularization parameter of SVM
$u_i$	The dual variable corresponds to a training point $A_i$
$\mu$	Parameter of RBF kernel
$(x)_+$	Plus function
$p(x, \alpha)$	Integral of sigmoid function
$\Phi_\alpha(w, \gamma)$	Objective function of SSVM
$\nabla \Phi_\alpha(w, \gamma)$	Gradient of objective function
$\nabla^2 \Phi_\alpha(w, \gamma)$	Hessian of objective function
$d$	Newton direction
$\lambda$	Armijo stepsize
$K(A, A')$	Full kernel matrix
$K(A, \bar{A}')$	Reduced kernel matrix
$\bar{A}$	Reduced set from $A$
$T(x, k)$	Three order spline function with parameter $k$



$m(x,k)$	Multiple knot spline function with parameter $k$
$C$	Classifier
$S_k(x)$	First derivative of $m(x,k)$
$d(x,y)$	Number of mismatches
$P(W,Q)$	Cost function
$d_2(x,y)$	The dissimilarity between two mixed type object X and Y

## LIST OF ABBREVIATIONS

ANFIS	Adaptive neuro fuzzy inference system
ANNs	Artificial neural network
CRSVM	Clustering Reduced Support Vector Machine
CV	Cross validation
DAGSVM	Directed Acyclic Graph Support Vector Machine
DOE	Design of experiment
ERM	Empirical risk minimization
FN	False negative
FNA	Fine needle aspirate
FP	False positive
FPSSVM	Forth polynomial Smooth Support Vector Machine
GA-AWAIS	Genetic algorithm weighted artificial immune system
GRNN	General regression neural network
GSVM	Generalized Support Vector Machine
K	Kernel
LOO	Leave one out
LS-SVM	Least square Support Vector Machine
MCLP	Multiple criteria linear programming
MKS	Multiple knot spline
MKS-SSVM	Multiple knot spline Smooth Support Vector Machine
MLNN	Multilayer neural network
OAA	One against all
OAo	One against one

PCA	Principal Component Analysis
PNN	Probabilistic neural network
PSSVM	Polynomial Smooth Support Vector Machine
QP	Quadratic programming
QPSSVM	Quadratic polynomial Smooth Support Vector Machine
RBF	Radial basis function
RS-MCLP	Rough set multiple criteria linear programming
RSVM	Reduced Support Vector Machine
SMO	Sequential Minimal Optimization
SRM	Structural risk minimization
SSRSVM	Systematic Sampling Reduced Support Vector Machine
SSVM	Smooth Support Vector Machine
SVM	Support Vector Machine
TN	True negative
TP	True positive
TSSVM	Three order spline Smooth Support Vector Machine
UD	Uniform design
VC	Vapnik-Chervonenkis
WDBC	Wisconsin diagnostic breast cancer
WPBC	Wisconsin prognostic breast cancer