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 MALAYSA 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

		Page
SUPERVISOR’S DECLARATION		ii
STUDENTS’S DECLARATION		iii
PREAMBLE		iv
DEDICATION		v
ACKNOWLEDGMENTS		vi
ABSTRACT		vii
ABSTRAK		viii
TABLE OF CONTENTS		ix
LIST OF TABLES		xiii
LIST OF FIGURES		xvi
LIST OF SYMBOLS		xviii
LIST OF ABBREVIATIONS		xx
CHAPTER 1	**INTRODUCTION**	
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
CHAPTER 2	**LITERATURE REVIEW**	

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

REFERENCES		125
APPENDICES		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

Table No.	Title	Page
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

Figure No.	Title	Page
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
γ	Distance hyperplane to the origin
y	Slack variable
e	Column vector of one of arbitrary dimension
v	Regularization parameter of SVM
u_i	The dual variable corresponds to a training point A_i
μ	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
λ	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

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 MALAYSA PAHANG

JUNE 2011

ABSTRACT

Research on Smooth Support Vector Machine (SSVM) for classification problem is an active field in data mining. SSVM is reformulation of standard Support Vector Machines (SVM). In SSVM, smoothing technique must be applied to convert constraint optimization to the unconstraint optimization problem since the objective function of this unconstraint optimization is not twice differentiable. The smooth function is used to replace the plus function to obtain a smooth support vector machine (SSVM). To get more accuracy performance, Multiple Knot Spline SSVM (MKS-SSVM) is proposed. MKS-SSVM is a new SSVM which used multiple knot spline function to approximate the plus function instead the integral sigmoid function in SSVM. To obtain optimal accuracy results, Uniform Design method is used to select parameter. The performance of the method is evaluated using 10-fold cross validation accuracy, confusion matrix, sensitivity and specificity. To evaluate the effectiveness of our method, an experiment is carried out on four medical dataset, i.e. Pima Indian diabetes dataset, heart disease, breast cancer prognosis, and breast cancer diagnosis. The results of this study showed that MKS-SSVM was effective to diagnose medical dataset and this is promising results compared to the previously reported results. SSVM algorithms are developed for binary classification. However, in many real problems data points are discriminated into multiple categories. Hence, MKS-SSVM is extended for multiclass classification. Two popular multiclass classification methods One against All (OAA) and One against One (OAO)) were used to extend MKS-SSVM. Numerical experiments show that the classification accuracy of OAA and OAO method are competitive with each other and there is no clear superiority of one method over another. While the computation time, the OAO method is lower than the OAA method on three dataset. This indicated that the OAO method is usually more efficient than the OAA. In the final part, the reduced support vector machine (RSVM) was proposed to solve computational difficulties of SSVM in large dataset. To generate representative reduce set for RSVM, clustering reduced support vector machine (CRSVM) had been proposed. However, CRSVM is restricted to solve classification problems for large dataset with numeric attributes. In this research, an alternative algorithm, *k*-mode RSVM (KMo-RSVM) that combines RSVM and *k*-mode clustering technique to handle classification problems on categorical large dataset and *k*-prototype RSVM (KPro-RSVM) which combine *k*-prototype and RSVM to classify large dataset with mixed attributes were proposed. In our experiments, the effectiveness of KMo-RSVM is tested on four public available dataset. It turns out that KMo-RSVM can improve speed of running time significantly than SSVM and still obtained a high accuracy. Comparison with RSVM indicates that KMo-RSVM is faster, gets smaller reduced set and comparable testing accuracy than RSVM. From experiments on three public dataset also show that KPro-RSVM can tremendously reduces the computational time and can handling classification for large mixed dataset, when the SSVM method ran out of memory (in case: census dataset). The comparison with RSVM indicate that the computational time of KPro-

RSVM less than RSVM method, and obtained testing accuracy of KPro-RSVM a little decrease than RSVM.

ABSTRAK

Penyelidikan Mesin Vector Sokongan Licin ((SSVM) adalah bidang yang aktif dalam pelombongan data. SSVM adalah perumusan semula dari Mesin Vektor Sokongan (SVM). Dalam SSVM, teknik pelicinan diterapkan untuk menukarkan pengoptimuman berkekangan dengan masalah pengoptimuman tidak berkekangan karena fungsi tujuan dari pengoptimuman tidak berkekangan tidak dibezakan. Fungsi *pelicinan* digunakan untuk menggantikan fungsi *plus (plus function)* sehingga disebut Mesin Vector Sokongan Licin (SSVM). Untuk mendapatkan ketepatan yang lebih baik, *Multiple Knot Spline* SSVM (MKS-SSVM) dicadangkan untuk masalah pengkelasan. MKS-SSVM adalah SSVM baru yang menggunakan *Multiple Knot Spline* untuk menganggarkan fungsi *plus* menggantikan fungsi integral sigmoid dalam SSVM. Untuk mendapatkan hasil ketepatan yang optimum, kaedah Uniform Design digunakan untuk memilih parameter. Prestasi MKS-SSVM dinilai menggunakan *10-fold cross validation*, *confusion matrix*, *sensitivity* dan *specitivity*. Untuk menilai keberkesanan kaedah ini, percubaan dilakukan pada empat dataset perubatan, iaitu dataset diabetes, penyakit jantung, prognosis kanser payudara, dan diagnosis kanser payudara. Keputusan kajian ini menunjukkan bahawa MKS-SSVM berkesan untuk mendiagnosis dataset perubatan dan ini sangat menjanjikan hasil berbanding dengan keputusan yang dilaporkan sebelum ini. Algoritma SSVM dibangunkan untuk pengkelasan perduaan. Walau bagaimanapun, dalam banyak masalah sebenar data didiskriminasi ke dalam berbilang kategori. Oleh itu, MKS-SSVM dilanjutkan untuk pengkelasan berbilang kategori. Dua kaedah pengkelasan yang popular iaitu *One Agains All* (OAA) dan *One Against One* (OAO) digunakan untuk membangun MKS-SSVM. Dari eksperimen menunjukkan bahawa kaedah ketepatan klasifikasi OAA dan OAO bersaing antara satu sama lain dan tidak ada keunggulan yang jelas dari satu kaedah di atas yang lain. Dalam bahagian akhir, Reduced Support Vector Machines (RSVM) telah dicadangkan untuk menyelesaikan masalah pengiraan SSVM dalam dataset yang besar. Untuk menjana *reduce set* untuk RSVM, clustering reduced support vector machine (CRSVM) telah dicadangkan. Walau bagaimanapun, CRSVM adalah terhad untuk menyelesaikan masalah pengelasan untuk dataset besar dengan sifat-sifat angka. Dalam kajian ini, algoritma alternatif, *k*-mode RSVM (KMo-RSVM) yang menggabungkan RSVM dan *k-modes* clustering teknik untuk menangani masalah pengelasan pada dataset kategori yang besar dan *k*-prototaip RSVM (KPro-RSVM) yang menggabungkan *k*-prototaip dan RSVM untuk mengelaskan dataset besar dengan sifat-sifat campuran telah dicadangkan. Dalam percubaan kami, keberkesanan KMo-RSVM diuji pada empat dataset. Ternyata KMo-RSVM dapat meningkatkan kelajuan masa secara signifikan dari SSVM dan masih memperoleh ketepatan yang tinggi. Perbandingan dengan RSVM menunjukkan bahawa KMo-RSVM lebih cepat, mendapatkan set yang lebih kecil dan mengurangkan ketepatan ujian setanding dari RSVM. Dari percubaan pada tiga dataset awam juga menunjukkan bahawa KPro-RSVM dapat mengurangkan masa pengkomputeran secara signifikan dan dapat menangani pengkelasan untuk dataset campuran, ketika kaedah SSVM kehabisan memori (dalam hal: dataset *census.*

REFERENCES

Abe, S. 2010. *Support vector machines for pattern classification*. 2nd ed. New York: Springer-Verlag.

Al Harbi, S.H. and Smith, V.J.R. 2006. Adapting *k*-means for supervised clustering. *Appl Intell*. **24**: 219–226.

Amir, A. and Lipika, D. 2005. Algorithm for fuzzy clustering of mixed data with numeric and categorical attributes. *Lecture Notes in Computer Science*. 561-572.

Aranganayagi, S and Thangavel, K. 2009. Improved *k*-modes for categorical clustering using weighted dissimilarity measure. *International Journal of Computational Intelligence*. **5** (2): 182-188.

Armijo, L. 1966. Minimization of functions having Lipshitz-continuous first partial derivatives. *Pacific Journal of Mathematics*. **16**: 1-3.

Bradley, P.S. ,Mangasarian, O.L. and Musicant, D.R. 1999. Optimization methods in massive datasets. Technical Report 99-01. Data Mining Institute. Department of Computer Sciences. University of Wisconsin. Madison.

Bradley, P.S. and Mangasarian, O.L. 2000. Massive data discrimination via linear support vector machines. *Optimization Method and Sofware*. **13**: 1-10.

Burgers, C.J.C. 1998. A tutorial on support vector machines for pattern recognition. *Data Mining Knowledge Discover*. **2**(2): 1- 47.

Chang, C.C. and Lee, Y.J. 2004. Generating the reduced set by systematic sampling. *Lecture Notes in Computer Science*. **3177**: 720-725.

Chaturvedi, A., Green, P., and Carrol, J. 2001. K-modes clustering. *Journal of Classification*. **18**: 35-55.

Chen, B. and Harker, P.T. 1993. A non interior point continuation method for linear complementary problems. *SIAM Journal on Matrix Analysis and Applications*. **14:** 1168-1190.

Chen, C. 1995. *Smoothing methods in mathematical programming*. PhD Thesis. Department of Computer Sciences. University of Wisconsin Madison.

Chien, L.J., Chang, C.C. and Lee, Y.J. 2010. Variant methods of reduced set selection for reduced support vector machines, *Journal of Information Science and Engineering*. **26** (1): 183-196

Cochran, W.G. 2007. *Sampling techniques*. 3rd edition. New York: John Wiley & Sons, Inc.

Das, R., Turkoglu, I. and Sengur, A. 2009. Effective diagnosis of heart disease through neural networks ensembles. *Expert Systems with Applications*. **36**: 7675–7680

Diaz, G.S. and Schulcloper, J.R. 2001. A clustering method for very large mixed data sets. *Proceedings of International Conference on Data Mining,* 638-644.

Duan, K.B. and Keerthi, S.S. 2005. Which is the best multiclass SVM method? an empirical study. *Lecture Notes of Computer Science.* 278-285.

Feng, F.Y., Xian, Z.D. and Can, H.H. 2007. Smooth SVM research: A polynomial-based approach. *The 6th International Conference on Information, Communications and Signal Processing.*

Fung, G and Mangasarian, O.L. 2001. Proximal Support Vector Machine classifier. *Proceedings KDD-2001 : Knowledge Discovery and Data Mining*, 77-86.

Gan, G., Yang, Z., Wu, J. 2005. A genetic *k*-modes algorithm for clustering categorical data. *Lecture Notes of Computer Science*. 195-202.

Goutte, C., Hansen, L. K., Liptrot, M. G. and Rostrup, E. 2001. Feature-space clustering for fmri meta-analysis. *Human Brain Mapping*. **13** (3): 165–183.

Goyal, M. .2007. *Computer based numerical and statistical techniques*. Hingham: Infinity Science Press LLC.

Gunn, S.R. 1998. Support vector machines for classification and regression. Technical Report. Faculty of Engineering and Applied Science. Department of Electronics and Computer Sciences.

Hastie, T., Tibshirani, R., and Friedman, J. 2009. *The elements of statistical learning: Data mining, inference, and prediction*. Springer Series in Statistics. 2nd. USA: Springer.

Hsu, C. W. and Lin, C. J. 2002. A simple decomposition method for support vector machines. *Machine Learning*, **46**: 291-314.

Hsu, C.W. and Lin, C.J. 2002. A comparison of methods for multi-class support vector machines. *IEEE Transactions on Neural Networks.* **13:** 415-425.

Hsu, C.W., Chang, C.C., and Lin, C.J. 2003. A practical guide to support vector classification. Department of Computer Science and Information Engineering. National Taiwan University.

Huang, C.M., Lee, Y.J., Lin, D.K.J., Huang, S.Y. 2007. Model selection for support vector machines via uniform design. *A Special issue on Machine Learning and Robust Data Mining of Computational Statistics and Data Analysis*, **52**: 335-346

Huang, Z. 1997. Clustering large data sets with mixed numeric and categorical values. *Proceedings of The First Pacific Asia Knowledge Discovery and Data Mining Conference*, Singapore.

Huang, Z. 1998. Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data Mining and Knowledge Discovery*. **2**: 283-304.

Huang, Z. 2003. A note on *k*-modes clustering. *Journal of Classification.* **20:** 257-261.

Huang, Z. and Ng, M. K. 1999. A fuzzy *k*-modes algorithm for clustering categorical data. *IEEE Transactions On Fuzzy Systems*. **7** (4): 446-452.

Jen, L.R. and Lee, Y.J. 2004. Clustering model selection for reduced support vector machines, *Lecture Notes in Computer Science*. **3177**: 714-719.

Joachims, T. 1999. Making large scale support vector machine learning practical. In Bernhard Scholkopf, Christopher J. C. Burges, and Alexander J. Smola, editors, *Advances in Kernel Method – Support Vector Learning*, 185-208. MIT Press.

Kahramanli, H and Allahverdi, N. 2008. Design of a hybrid system for the diabetes and heart diseases. *Expert Systems with Applications*. **35**: 82-89.

Kayaer, K. and T. Yildirim, 2003 Medical diagnosis on pima indian diabetes using general regression neural networks. *Proceedings of the International Conference on Artificial Neural Networks and Neural Information Processing*, pp. 181-184.

Ketchen, D.J. and Shook, C.L. 1996. The application of cluster analysis in Strategic Management Research: An analysis and critique. *Strategic Management Journal*. **17** (6): 441–458.

Khan, S. S. and Kant, S. 2007. Computation of initial modes for *k*-modes clustering algorithm using evidence accumulation. *Proceeding of 20th International Joint Conference on Artificial Intelligence*, pp. 2784-2789.

Kohavi, R., and Provost, F. 1998. Glossary of terms. *Editorial for the Special Issue on Applications of Machine Learning and the Knowledge Discovery Process*. **30**: 2-3.

KreBel, U. 1999. Pairwise classification and support vector machines. In : *Advances in Kernel Method : Support Vector Learning*. Cambridge, MA: MIT Press, 255-268.

Lee, S.G. and Yun, D.K., 2003. Clustering categorical and numerical data: A new procedure using multidimensional scaling. *International Journal of Information Technology and Decision Making*. **2** (1): 135-160.

Lee, Y. J. 2001. *Support vector machines in data mining*. PhD thesis. University of Wisconsin-Madison, USA.

Lee, Y.J. and Huang, S.Y. 2007. Reduced support vector machines: A Statistical Theory. *IEEE Trans.Neural Network*. **18**(1).

Lee, Y.J. and Mangasarian, O.L. 2001a. A smooth support vector machine. *Journal of Computational Optimization and applications.* **20**: 5 – 22.

Lee, Y.J. and Mangasarian, O.L. 2001b. RSVM: reduced support vector machines. *In Proceedings of the First SIAM International Conference on Data Mining.*

Lee, Y.J., Hsieh, W. and Huang, C. 2005. ε-SSVR: A smooth support vector machine for ε-insensitive regression, *IEEE Trans. Knowledge and Data Engineering*.**17**(5): 678-685.

Lei, H. and Govindaraju, V. *Half against half support vector machines*. USA: Center for Biometrics and Sensors, Computer Science and Engineering Department.

Liang, Y.Z, Fang, K.T, and Xu, Q.S. 2001. Uniform design and its applications in chemistry and chemical engineering. *Chemometrics and Intelligent Laboratory System*. **58**: 43-57.

Liao, Z. Fan, X., Zhou, Y., Liu, K. 2008. A clustering algorithm for mixed data based on lattice theory. *Proceedings of the 9th International Conference for Young Computer Scientists*.

Lin, K.M and Lin, C.J. 2003. A Study on Reduced Support Vector Machines. *IEEE Trans.Neural Network*. **14**(6): 1449-1459.

Lin, Y. 2000. *On the support vector machine*. Technical Report No. 1029, Department of Statistics, University of Wisconsin, Madison.

Lin, Y., Wahba, G., Zhang, H., and Lee, Y. 2002. Statistical properties and adaptive tuning of support vector machines. *Machine Learning.* **48:** 115-135.

Liu, B., Yang, X., and Hao .Z. 2006. Binary tree support vector machine based on kernel fisher discrimanant for multi-classification. In: *J.Wang et al (Eds) : ISNN 2006, LNCS* 3971, pp. 997-1003.

Liu, H., Dai, B., He, H., Fan, Y. 2006. The *k*-prototype algorithm of clustering high dimensional and large scale mixed data. *Proceedings of the International Computer Conference,* 738-743.

Lieti, R., Ortiz, M.C., Sarabia, L.A, Sánchez, M.S. (2004). Selecting variables for k-means cluster analysis by using a genetic algorithm that optimizes the silhouettes. *Analytica Chimica Acta.* **515**: 87–100.

Luo, L., Lin, C., Peng, H. and Zhou, Q. 2006. A study on piecewise polynomial smooth approximation to the plus function. *ICARCV.*

MacQueen, J.B. 1967. Some methods for classification and analysis of multivariate observations. *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, 281-297.

Maglogiannis, I., Zafiropoulos, E., and Anagnostopoulos, I. 2009. An Intelligent system for automated breast cancer diagnosis and prognosis using SVM based classifiers. *Applied Intelligent.* **30**: 24-36.

Mangasarian, O.L. 2000. Generalized support vector machines. In Smola, A., Bartlett, P., Scholkopf, B., and Schurrmans, D. editors, *Advances in large Margin Classifiers*, pages 135 – 146, Cambridge, MA. MIT Press.

Mangasarian, O.L. 2001. Data mining via support vector machines. *IFIP Conference on System Modelling and Optimization*, Trier, Germany.

Mangasarian, O.L. and Musicant, D.R. 1999. Successive over relaxation for support vector machines, *IEEE Transactions on Neural Networks*. **10**: 1032-1037.

Mangasarian, O.L. and Musicant, D.R. 2001. Lagrangian support vector machines. *Journal of Machine Learning Research.* **1**: 161-177.

Mardia, K.V., Kent, J.T., and Bibby, J.M. 1979. *Multivariate Analysis*. Academic Press.

Milenova, B.L. and Campos, M.M. 2003. Clustering large bases with numeric and nominal values using orthogonal projection, *Proceeding of 29th VLDB Conference*, Berlin, Germany.

Newman, D.J., Hettich, S., Blake, C. L. S., & Merz, C. J., 1998. UCI repository of machine learning database, Irvine, CA: University of California, Dept. of Information and Computer Science. http://www.ics.uci.edu/~mlearn/ ~MLRepository.html, Ng, M. K. and Wong, J. C. 2002. Clustering categorical data sets using tabu search techniques. *Pattern Recognition*. **35** (12): 2783-2790.

Ozsen, S. and Gunes, S. 2009. Attribute weighting via genetic algorithms for attribute weighted artificial immune system (AWAIS) and its application to heart disease and liver disorders problems. *Expert Systems with Applications*. **36**: 386-392.

Pinar, M.C. and Zenios, S.A. 1994. On smoothing exact penalty functions for convex constrained optimization. SIAM Journal on Optimization. **4**: 486-511.

Platt, J. 1999. Sequential minimal optimization: a fast algorithm for training support vector machines. In Bernhard Scholkopf, Christopher J. C. Burges, and Alexander J. Smola, editors, *Advances in Kernel Method – Support Vector Learning*, pp. 185-208. MIT Press.

Platt, J.C. 1999. Probabilistic outputs for support vector machines and comparison to regularized likelihoods methods. In : Smola, A., Bartlett, P., Scholkopf, B., and Schurrmans, D. editors, *Advances in large Margin Classifiers*, pages 61 – 74, Cambridge, MA. MIT Press.

Platt, J.C., Cristianini, N. and Taylor, J.S. (2000). Large margin dags for multiclass classification. In : *Advances in Neural Information Processing Systems.* **12**: 547-553.

Polat, K. and Gunes, S. 2007. An expert system approach based on principal component analysis and adaptive neuro-fuzzy inference system to diagnosis of diabetes disease. *Digital Signal Processing*. **17:** 702-710.

Polat, K., Gunes, S. and Aslan, A. 2008. A cascade learning system for classification of diabetes disease: Generalized discriminant analysis and least square support vector machine. *Expert System with Applications*. **34**: 214-22.

Polat, K., Sahan, S. and Gunes, G. 2007. Automatic detection of heart disease using an artificial immune recognition system (AIRS) with fuzzy resource allocation mechanism and k-nn (nearest neighbour) based weighting preprocessing. *Expert Systems with Applications.* **32**: 625–631.

Polat,K., Sahan, S., Kodaz, H. and Günes, S. 2005. A new classification method to diagnosis heart disease: supervised artificial immune system (AIRS), *Proceedings of the Turkish Symposium on Artificial Intelligence and Neural Networks (TAINN).*

Ralambondrainy, H. 1995. A conceptual version of the K-Means algorithm. *Pattern Recognition Letters.* **16:** 1147-1157.

Rousseuw, P.J., 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Computational and Applied Mathematics.* **20**: 53–65.

San, O.M., Huynh, V.N., Nakamori, Y. 2004. An alternative extension of the k-means algorithm for clustering categorical data, *International Journal Applied Mathematic Computing Science*. **14** (2): 241-247.

Santosa, B. 2005. *Learning from data with uncertainty: Robus multiclass kernel based classifiers and regressors*. PhD thesis. University of Oklahoma, USA.

Scholkopf, B., Sung, K.K., Burges, C., Girosi, F., Niyogi, P., Poggio, T. And Vapnik, V. 1997. Comparing support vector machines with gaussian kernels to radial basis function classifiers. *IEEE Transactions on Signal Processing* **45.**

Sugar, C.A. and James, G.M. 2003. Finding the number of clusters in a data set: An information theoretic approach. *Journal of the American Statistical Association.* **98** (January): 750–763.

Suykens, J.A.K., and Vandewalle, J. 1999. Least squares support vector Machines, *Neural Processing Letters.* **9(3).**

Suykens, J.A.K., and Vandewalle, J. multiclass least squares support vector machines, Technical Report, Katholieke Universiteit Leuven, Dept. of Electr. Eng., ESAT-SISTA, Kardinaal Mercierlaan 94, B-3001 Leuven (Heverlee), Belgium.

Teknomo, K. 2006. K-means clustering tutorials. http:\\people.revoledu.com\ kardi\ tutorial\kMean\.

Temurtas, H. et al. 2009. A comparative study on diabetes disease using neural networks. *Expert System with Applications*. **36**: 8610-8615.

Vapnik, V. 1995. *The nature of statistical learning theory*. New York: Springer-Verlag.

Vapnik, V. 1998. *Statistical learning theory*. New York: Wiley.

Wang, J., Wu, X. and Zhang, C. 2005. Support vector machines based on k-means clustering for real time business intelligence systems. *International Journal Business Intelligence and Data Mining*. **1**(1): 54-64.

Wang, L. 2005. *The support vector machines: Theory and Applications*. Berlin, Heidelberg: Springer-Verlag.

Weston, J.A.E. 1999. *Extensions to the support vector method*. PhD Thesis. Department of Computer Science, England.

Weston, J.A.E. and Watkins, C. 1998. Multi-class support vector machines. Technical Report, Department of Computer Science, England.

Wu, M., Scholkopf, B. and Bakir, G. 2006. A direct method for building sparse kernel learning algorithms. *Journal of Machine Learning Research*. **7**: 603-624.

Wu, T.F., Lin, C.J., and Weng, R.C. 2004. Probability estimates for multiclass classification by pairwise coupling. *Journal of Machine Learning Research.* **4:** 975-1005.

Xiong, J., Hu, T., Jinlian Hu, Li, G., and Peng, H. 2006b. Smoothing support vector machines for insensitive regression. *Proceedings of the 6th International Conference on Intelligent systems design and application (ISDA), IEEE.*

Xiong, J., Hu, T., Li, G., and Peng, H. 2006a. A comparative study of three smooth svm classifiers. *Proceedings of the 6th World Congress on Intelligent Control and Automation*, June 21-23, Dalian, Cina.

Yin, J., Tan, Z. F., Ren, J. T., Chen, Y. Q. 2005a. An efficient clustering algorithm for mixed type attributes in large dataset. *Proceedings of the fourth International Conference on Machine Learning and Cybernetics*, Guangzhou, 1611-1614.

Yin, J., Tan, Z. F., Ren, J. T., Chen, Y. Q. 2005b. Clustering mixed type attributes in large dataset. *Lecture Notes in Computer Science.* 655-661.

Yuan, Y. and Li, C. 2005. *A new smooth support vector machine*. School of Applied Mathematics, University of Electronic Science and Technology of China, China.

Yuan, Y., Fan, W., and Pu, D. 2007. Spline function smooth support vector machine for classification. *Journal of Industrial and Management Optimization.* **3(3)**: 529-542.

Yuan, Y., Yan J., and Xu C. 2005. Polynomial smooth support vector machine (PSSVM). *Chinese Journal of Computers.* **28:** 9-17.

Zhang, C., Pu Han, Guiji Tang, and Guori Ji. 2007. Simulation of time series prediction based on smooth support vector regression. *In D.Liu et al (eds), ISNN part III, LNCS* 4493, 545-552, Springer-Verlag Berlin Heideberg.

Zhang, Z., Shi, Y., and Gao, G. 2009. A rough set-based multiple criteria linear programming approach for the medical diagnosis and prognosis. *Expert System with Applications.*, **36:** 8932-8937.