

CONTENTS

	Page
Acknowledgment	c
Abstract in Thai	d
Abstract in English	f
List of Tables	k
List of Figures	m
List of Symbols	p
Statements of Originality in Thai	q
Statements of Originality in English	r
Chapter 1 Introduction	1
1.1 Background and motivation	1
1.2 Literature review	5
1.3 Research objective	10
1.4 Research scope	10
1.5 Education advantages	10
1.6 Research location	11
1.7 Thesis organization	11
Chapter 2 Principles and Theories of the Study	12
2.1 Output-constrained clustering	12
2.2 Prototype selection method	13
2.2.1 Direction of search	13
2.2.2 Type of selection	14
2.2.3 Evaluation of search	15
2.2.4 Criteria to compare prototype selection methods	16

2.3 Fuzzy c-means clustering algorithm	17
2.4 Cluster validity indices	19
2.5 Technique for order preference by similarity to ideal solution	20
2.6 Supervised learning	22
2.6.1 Support vector machine	22
2.6.2 Nearest neighbor rule	24
2.6.3 Artificial neural network	24
Chapter 3 Research Designs and Methods	26
3.1 Instance reduction for regression problems	26
3.1.1 Training data split	26
3.1.2 Output data quantization	29
3.1.3 Input data clustering	30
3.1.4 Data combination	31
3.2 Instance reduction for classification problems	31
3.2.1 Noise and overlapping data removal	32
3.2.2 Class boundary approximation	32
3.2.3 Reduced data set (\mathcal{RS}) initialization	33
3.2.4 Extracting instances in \mathcal{T} to \mathcal{RS}	33
3.2.5 Data pruning	34
Chapter 4 Results and Discussion	36
4.1 Result of data reduction on regression problems	36
4.1.1 One-dimensional synthesized data set	37
4.1.2 Two-dimensional synthesized data set	41
4.1.3 Real world data set 1	44
4.1.4 Real world data set 2	47
4.1.5 Real world data set 3	50
4.1.6 Concerning issues of the proposed method for regression problems	56
4.1.7 Setting the number of small parts and quantization levels	56

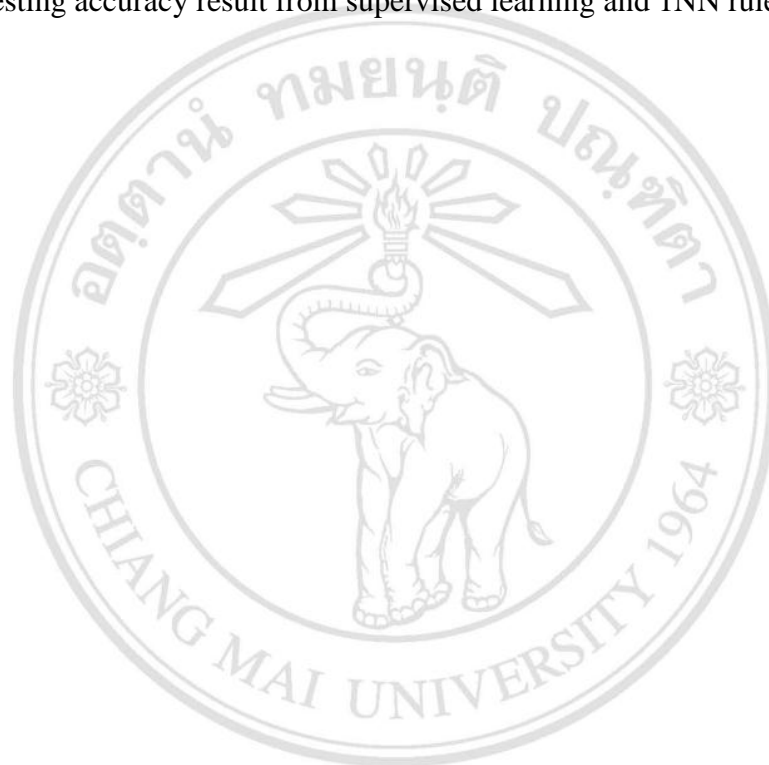
4.2 Result of instance reduction of classification problems	56
4.2.1 Two-dimensional problem : two-class synthesized data set	56
4.2.2 Two-dimensional problem : multiclass synthesized data set	59
4.2.3 Public data set from sci2s.ugr.es : small data sets	61
4.2.4 Public data set from sci2s.ugr.es : medium data sets	69
4.2.5 Public data set from sci2s.ugr.es : large data sets	73
4.2.6 Evaluation of reduced data sets in model-based classifiers	76
4.2.7 Concerning issues of the proposed method for classification problems	77
Chapter 5 Conclusion	78
References	80
Appendix	89
Curriculum Vitae	100

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
 Copyright© by Chiang Mai University
 All rights reserved

LIST OF TABLES

	Page
Table 4.1 Comparison between training and testing RMSE of proposed algorithm and two others	41
Table 4.2 Four-fold cross validation result (The number of input samples = 50)	42
Table 4.3 Four-fold cross validation result (The number of input samples = 400)	42
Table 4.4 Mean absolute errors from SVR for two-dimensional data set using different methods	42
Table 4.5 Root-mean-square errors obtained by using original data and our proposed method	47
Table 4.6 Root-mean-square errors from SVR for real data set 1 using different methods	49
Table 4.7 V2 ECG signal reconstructed result	51
Table 4.8 V3 ECG signal reconstructed result	52
Table 4.9 V4 ECG signal reconstructed result	53
Table 4.10 V5 ECG signal reconstructed result	55
Table 4.11 List of prototype selection methods	62
Table 4.12 Groups of prototype selection methods	63
Table 4.13 List of small data sets	64
Table 4.14 Result of small data sets from different methods	65
Table 4.15 Result of small data sets from different methods without hybrid-wrapper algorithm	66
Table 4.16 Percentage improvement between the proposed method and other methods using small data sets	67
Table 4.17 List of medium data sets	69
Table 4.18 Result of medium data sets from different methods	70
Table 4.19 Result of medium data sets from different methods without hybrid-wrapper algorithm	71

Table 4.20 Percentage improvement between the proposed method and other methods using medium data sets	72
Table 4.21 List of large data sets	73
Table 4.22 Result of large data sets from different methods	74
Table 4.23 Percentage improvement between the proposed method and other methods using large data sets	74
Table 4.24 Testing accuracy result from supervised learning and 1NN rule	76



ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
 Copyright© by Chiang Mai University
 All rights reserved

LIST OF FIGURES

	Page
Figure 1.1 (a)-(d) ECG input signals, (e) ECG output signal	4
Figure 1.2 Electrical load series	5
Figure 1.3 Banana data set (a) Original data set, (b) Class boundary data	5
Figure 2.1 Soft margin ϵ -insensitive loss setting in a linear SVR	23
Figure 2.2 Support vector regression model	24
Figure 2.3 Feed-forward neural network model	25
Figure 3.1 Procedure for the input-output clustering method	26
Figure 3.2 Original training data and sorted training data	28
Figure 3.3 (Top) Histogram of original data, (Bottom) Histogram of each small part	29
Figure 3.4 Procedure for data reduction for classification problems	32
Figure 3.5 Demonstration of class boundary approximation	33
Figure 3.6 Proposed data reduction algorithm for classification problems	34
Figure 4.1 Relationship between data reduction rate and the number of quantization levels	38
Figure 4.2 Relationship between RMSE and the number of quantization levels	38
Figure 4.3 Relationship between computation time for the reduction process and the number of quantization levels	39
Figure 4.4 Plot of (a) original data, (b)-(e) reduction results from various quantization levels ($q = 10, 50, 100, 200$)	39
Figure 4.5 Surface area of two-dimensional input function	41
Figure 4.6 Regression result between the reduced and original data	43
Figure 4.7 Comparison between the desired output and regression result of each high dimensional function	44

Figure 4.8	Relationship of number of quantization levels, q and data reduction rate	45
Figure 4.9	Minimum RMSE from four-fold cross validation at various quantization levels.	46
Figure 4.10	Average RMSE from four-fold cross validation at various quantization levels	46
Figure 4.11	Histogram of (a) original data and (b - e) four small parts	48
Figure 4.12	Data reduction rate versus the number of small parts	48
Figure 4.13	Plot between RMSE and the number of small parts	49
Figure 4.14	V2 ECG signal reconstructed result	51
Figure 4.15	(Top) Original V2 chest lead signal, (Bottom) Reconstructed V2 chest lead signal.	51
Figure 4.16	V3 ECG signal reconstructed result	52
Figure 4.17	(Top) Original V3 chest lead signal, (Bottom) Reconstructed V3 chest lead signal.	53
Figure 4.18	V4 ECG signal reconstructed result	54
Figure 4.19	(Top) Original V4 chest lead signal, (Bottom) Reconstructed V4 chest lead signal.	54
Figure 4.20	V5 ECG signal reconstructed result	55
Figure 4.21	(Top) Original V5 chest lead signal, (Bottom) Reconstructed V5 chest lead signal	55
Figure 4.22	Scatter plots of two-dimensional data for two class problem (a) The original data, (b) Preprocessed with removal of noise and overlap data, (c) Initialized the reduced data set (\mathcal{RS}), (d) Extracted the instance data near the decision boundary, (e) The results from pruning the \mathcal{RS}	58
Figure 4.23	Voronoi diagram of the original data set (Left), Voronoi diagram of the reduced data set (Right)	59
Figure 4.24	Scatter plots of two-dimensional data for 4-class problem (a) The original data, (b) Preprocessed with removal of noise and overlapping data, (c) Initialized the reduced data set (\mathcal{RS}),	

	(d) Extracted the instance data near the decision boundary,	
	(e) The results from pruning the RS	60
Figure 4.25	Voronoi diagram of the original data set : four-classes (Left), Voronoi diagram of the reduced data set : four-classes (Right)	61
Figure 4.26	Average of normalized accuracy and reduction ratio of small data sets	68
Figure 4.27	Range of average accuracy of each data in small data sets	69
Figure 4.28	Average of normalized accuracy and reduction ratio of medium data set	72
Figure 4.29	Range of average accuracy of each data in medium data sets	73
Figure 4.30	Average of normalized accuracy and reduction ratio of large data set	75
Figure 4.31	Range of average accuracy of each data in large data sets	75

LIST OF SYMBOLS

c	Number of clusters
C	Number of classes
ℓ	Number of sample points of the data set
N	Number of attributes of the data set
\mathbf{x}_i	An input column vector of the output y_i
\mathbf{T}	The training data set matrix
\mathbf{X}	The input matrix
\mathbf{y}	The output vector
\mathcal{T}	The set of training data
\mathbf{x}	A sample point in \mathcal{T}
\mathcal{B}	The approximated class boundary set
\mathcal{RS}	The reduced data set
$\xi(\mathbf{x})$	The nearest enemy function of \mathbf{x}
\mathcal{X}	The temporary set

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

ข้อความแห่งการริเริ่ม

- 1) วิทยานิพนธ์นี้นำเสนอวิธีการลดข้อมูลแบบใหม่ บนพื้นฐานของการจัดกลุ่มอินพุต-เอาต์พุตที่สามารถควบคุมความถูกต้องและอัตราการลดข้อมูล โดยกำหนดค่าเริ่มต้นให้กับการแบ่งส่วนข้อมูลและจำนวนของการแบ่งนับเพื่อใช้สอนระบบการทำนาย
- 2) วิทยานิพนธ์นี้ยังได้นำเสนอวิธีการลดข้อมูลแบบใหม่ บนพื้นฐานของวิธีการลดจำนวนและการแก้ไขชุดข้อมูล เพื่อใช้สอนระบบการจำแนก วิธีการที่นำเสนอมีความซับซ้อนในการคำนวณน้อยกว่าวิธีการที่นำมาเปรียบเทียบ เนื่องจากการประมวลผลเบื้องต้นเพื่อหาข้อมูลบริเวณขอบของกลุ่มข้อมูลอื่นมาใช้ประมวลผลแทนการประมวลผลกับข้อมูลทั้งหมด โดยวิธีการนี้สามารถเลือกและคงข้อมูลที่สำคัญที่อยู่บริเวณขอบของกลุ่มข้อมูลไว้ได้



ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

STATEMENTS OF ORIGINALITY

1. This thesis proposed a new instance reduction method for regression data based on input-output clustering with the abilities to control the accuracy and reduction ratio via setting the number of separation parts and quantization levels.
2. This thesis also proposed a new instance reduction method based on condensation and edition methods for classification. The computational complexity of the method is lower than other methods because the class boundary of the enemy class is preprocessed by calculating the class boundary instead of all enemy sample points. The proposed method can select and retain the important data which are located at the class boundary.



ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved