

Improved Clustering and Soft Computing Algorithms

by

Shu-Chuan Chu, B.Sc. (Hons)
School of Informatics and Engineering,
Faculty of Science and Engineering

January 1, 2004

A thesis presented to the Flinders University of South Australia in total fulfillment of the requirements for the degree of Doctor of Philosophy

Adelaide, South Australia, 2004 © (Shu-Chuan Chu, 2004)

Contents

\mathbf{A}	Abstract				
\mathbf{C}	Certification				
\mathbf{A}	ckno	wledge	ements	xvi	
1	Inti	oduct	ion	1	
	1.1	Relate	ed Existing Clustering Algorithms	5	
	1.2	Soft C	Computing	8	
		1.2.1	Fuzzy Theorem	8	
		1.2.2	Artificial Neural Networks	10	
		1.2.3	Simulated Annealing	11	
		1.2.4	Tabu Search Approach	12	
		1.2.5	Genetic Algorithms	12	
		1.2.6	Particle Swarm Optimization	14	
		1.2.7	Ant Systems and Ant Colony Systems	15	
	1.3	Thesis	s Structure	16	
2	Imp	oroved	K-medoids Algorithms	19	
	2.1	Relate	ed existing K -Medoids Algorithms	19	
		2.1.1	PAM - Partitioning Around Medoids	21	
		2.1.2	CLARA – Clustering LARge Applications	22	
		2.1.3	CLARANS – Clustering Large Applications Based on Randomized Search	24	
		2.1.4	Fuzzy K -Medoids Algorithms	26	
		2.1.5	Genetic K -Medoids Algorithm	28	

CONTENTS

	2.2		A - Clustering Large Applications Based on Simulated Angressian Science - Comparison of Comparison - Comparis	28
		2.2.1	What's CLASA	29
		2.2.2	Experiments	31
	2.3	Efficier	nt Search Based K -Medoids Algorithms	35
		2.3.1	Revisiting Swap-Comparison of PAM	37
		2.3.2	VQ-Based Techniques	41
		2.3.3	Partial Distance Search	43
		2.3.4	Triangular Inequality Elimination	44
		2.3.5	Previous Medoid Index	48
	2.4	Memor	ry Utilization Based K -Medoids Algorithm	50
	2.5	Experi	mental Results	52
		2.5.1	Gaussian Source	53
		2.5.2	Gauss-Markov Source	54
		2.5.3	Rectangular Clusters	55
		2.5.4	Elliptic Clusters	56
		2.5.5	Curved Clusters	57
		2.5.6	Lena Dataset	57
		2.5.7	Real-World Dataset	58
		2.5.8	Summary	59
3	Sam	nling (Schemes for K -Medoids Algorithm	74
)	3.1		uction	7 4
	3.1		Centroid, Multi-Run Sampling Scheme $(MCMRS)$	76
	5.4	3.2.1	Motivation	76
			Experimental Results	
	3.3	3.2.2 Increm	nental Multi-Centroid, Multi-Run Sampling Scheme (IMCMRS)	78 84
	ა.ა	3.3.1	Motivation	84
		3.3.2	Experimental Results	85

CONTENTS

4	Imp	oroved	Centroid-Based Clustering Algorithms	94
	4.1	Relate	ed Existing Centroid-Based Clustering Algorithm	95
		4.1.1	K-Means (GLA)	96
		4.1.2	Simulated Annealing for Clustering	97
		4.1.3	Tabu Search Approach for Clustering	98
		4.1.4	Clustering Using Stochastic Relaxation Approach	101
	4.2	Cluste	ering using Tabu Search with Simulated Annealing	102
		4.2.1	Motivation	102
		4.2.2	Experimental Results	103
	4.3	Genet	ic Clustering for Mean-Residual Vector Quantization	111
		4.3.1	Motivation	111
		4.3.2	Experiments	118
	4.4	Incren	mental Splitting Clustering	125
		4.4.1	Introduction	125
		4.4.2	Related Works	127
		4.4.3	Proposed Algorithm – Incremental Splitting	128
		4.4.4	Experimental Results	129
	4.5	Label	led Bisecting K -Means Clustering Algorithm for Watermarking	
		4.5.1	Introduction	132
		4.5.2	Proposed Algorithm	133
		4.5.3	Experimental Results	138
	4.6		mard Transform Based Fast Codeword Search Algorithm for Dimensional VQ Encoding	143
		4.6.1	Introduction	
		4.6.2	Related Existing Nearest Neighbour Codeword Search Al-	
			gorithms	146
		4.6.3	Basic Definitions And Properties	153
		4.6.4	Proposed Algorithm	156
		4.6.5	Experimental Results	160
5	Par	allel P	Particle Swarm Optimization	162
	5.1	Histor	ry of Particle Swarm Optimization	162
	5.2	Partic	ele Swarm Optimization with Communication Strategies	167
		5.2.1	Motivation and Description	167
		5.2.2	Experiments	173

CONTENTS

6	Par	allel ar	nd Constrained Ant Colony Optimizations	177
	6.1	History	y of Ant System and Ant Colony System	178
	6.2	Ant Co	olony System with Communication Strategies	183
		6.2.1	Description	183
		6.2.2	Experimental Results	188
	6.3	Adapti	ive Ant Colony System for Data Clustering	195
		6.3.1	Ant Colony Optimization with Different Favor ($ACODF$) .	195
		6.3.2	The Constrained Ant Colony Optimization $(CACO)$	196
		6.3.3	Experiments and Results	201
7	Con	clusion	ns and Future Work	212
	7.1	Summ	ary	212
	7.2	Conclu	isions	214
		7.2.1	Efficient and Effective K -medoids Algorithms	214
		7.2.2	$K\operatorname{\!-medoids}\nolimits$ Algorithms Based on Sampling Schemes	215
		7.2.3	Centroid-Based Clustering Algorithms	215
		7.2.4	Labeled Bisecting K -means Clustering	216
		7.2.5	Hadamard Transform Based Inequalities for Efficient Clustering	217
		7.2.6	PPSO with Communication Strategies	217
		7.2.7	PACS with Communication Strategies	218
		7.2.8	Constrained Ant Colony Optimization for Data Clustering	218
	7.3	Future	e Work	219
		7.3.1	Transform Domain Based K -medoids Algorithm	219
		7.3.2	PSO for Clustering of Objects	219
		7.3.3	Sampling Scheme and Tree Structure for CACO	220
		7.3.4	Application of CACO for Texture Segmentation	220
		7.3.5	Application of CACO for Clustering of Categorical Objects	220
\mathbf{A}	Pub	olicatio	ns arising from this thesis	221
Bi	bliog	raphy		22 4

List of Figures

1.1	Problem datasets for some clustering algorithms	3
1.2	The hybridization of soft computing	S
2.1	Clustering LARge Applications ($CLARA$)	22
2.2	Clustering Large Applications Based on RANdomized Search Algorithm $(CLARANS)$	25
2.3	Clustering Large Applications Based on Simulated Annealing Algorithm $(CLASA)$	30
2.4	Data Generation Program	32
2.5	Elliptic Clusters	33
2.6	Performance Comparison of <i>CLARA</i> , <i>CLARANS</i> and <i>CLASA</i> for Gauss-Markov Sources	36
2.7	Performance Comparison of <i>CLARA</i> , <i>CLARANS</i> and <i>CLASA</i> for Elliptic Clusters	36
2.8	Four Cases When Medoid o_{old} Is Replaced as Representative Object by o_{new}	39
2.9	Fifth Case When Medoid o_{old} Becomes The Non-Medoid Object $% \left(1\right) =\left(1\right) \left(1\right) +\left(1\right) \left(1\right) \left(1\right) +\left(1\right) \left(1\right) \left($	40
2.10	The First Criterion of TIE	45
2.11	The Second Criterion of TIE	46
2.12	The Third Criterion of TIE	47
2.13	The Combination of Criterion 2 and 3 of TIE	49
2.14	The Combination of Criterion 1, 2 and 3 of TIE \dots	49
2.15	The Three Categories of Distance Calculation	50
2.16	Usage of Memory for Distance Calculations	51
2.17	Twelve Elliptic Clusters	56
2.18	Curved Clusters	57
2.19	Results of Gauss-Markov Experiment	70

LIST OF FIGURES vii

2.20	Results of Elliptic Clusters Experiment	70
2.21	Results of Curved Clusters Experiment	71
2.22	Results of Experiments by Number of Calculations	72
2.23	Results of Experiments by Number of Calculations	73
3.1	Compact Clusters with Noise	78
3.2	Performance comparison of $CLARA$, $CLARANS$, $CLASA$, $MCMRS$ $CLASA$ and $MCMRS$ for Four Elliptic Clusters	S- 80
3.3	Performance Comparison of $CLARA$, $CLARANS$, $MCMRS$ and $MCMRS$ - $CLASA$ for Twelve Elliptic Clusters	81
3.4	Performance Comparison of <i>CLARA</i> , <i>CLARANS</i> , <i>MCMRS</i> and <i>MCMRS-CLASA</i> for Compact Clusters with Noise	82
3.5	The flowchart of Incremental Multi-Centroid, Multi-Run Sampling Scheme (IMCMRS)	86
3.6	Performance Comparison of <i>CLARA</i> , <i>CLARANS</i> , <i>MCMRS</i> and <i>IMCMRS</i> for Four Elliptic Clusters	88
3.7	Performance Comparison of <i>CLARA</i> , <i>CLARANS</i> , <i>MCMRS</i> and <i>IMCMRS</i> for Twelve Elliptic Clusters	89
3.8	Performance comparison of $CLARA$, $CLARANS$, $MCMRS$ and $IMCMRS$ for Five Compact Clusters	90
3.9	Performance Comparison of $CLARA$, $CLARANS$, $MCMRS$ and $IMCMRS$ for Gauss-Markov Source	92
3.10	Performance Comparison of $CLARA$, $CLARANS$, $MCMRS$ and $IMCMRS$ for $Lena$ Image	93
4.1	Flowchart of The Tabu Search Approach for Clustering Patterns .	96
4.2	Flowchart of The Tabu Search Approach with GLA Algorithm	100
4.3	Flowchart of The Tabu Search Approach with Simulated Annealing 1	104
4.4	Performance Comparison Using $LENA$ Image	107
4.5	Performance Comparison Using $Baboo$ Image	107
4.6	Performance Comparison Using Pepper Image	107
4.7	Comparison of $Lena$ Image Recovered by Different Algorithm	108
4.8	Comparison of $Baboo$ Image Recovered by Different Algorithm	108
4.9	Comparison of $Pepper$ Image Recovered by Different Algorithm . 1	110
4.10	Block Diagram of VQ Compression	112
4.11	Schematic for Mean-Residual Vector Quantization $(M/R VQ)$ 1	114

LIST OF FIGURES viii

4.12	Flowchart of Genetic Clustering Algorithm for $(M/R\ VQ)$	117
4.13	Experiment for Mutation Rate	118
4.14	Experiment for Survival Rate	119
4.15	Experiment for Population Size	119
4.16	Experiment for Maximum Number of Generation	120
4.17	Experiment for the Number of Iteration for GLA	121
4.18	Pseudo Code for Local-Descent Methods	127
4.19	Pseudo Code for K -Means Method	128
4.20	Pseudo Code for Incremental Splitting Algorithm	129
4.21	Distribution of 10,000 Sample Data	130
4.22	Distribution of 256 Clusters Using Different Heuristics	131
4.23	A Concrete Example to Describe The Embedding Process for Each Input Vector	135
4.24	Flowchart of labelled bisecting k-means clustering algorithm $$	136
4.25	Original Image and Watermark	139
4.26	JPEG Compressed Watermarked Image and Corresponding Extracted watermark	140
4.27	Median Filtered Watermarked Image and Corresponding Extracted Watermark	141
4.28	Blurred Watermarked Image and Corresponding Extracted Watermark	141
4.29	Sharpened Watermarked Image and Corresponding Extracted Watermark	142
4.30	Attacked Watermarked Image by Cropping and Corresponding Extracted Watermark	142
5.1	Object function F_2	165
5.2	Progress of PSO on object function F_2	165
5.3	The distribution of particles at different iterations	166
5.4	Communication Strategy for Loosely Correlated Parameters	170
5.5	Communication Strategy for strongly correlated parameters	171
5.6	A General Communication Strategy for Unknown Correlation Between Parameters	172
6.1	A traveling salesman problem with 12 cities	179

LIST OF FIGURES ix

6.2	The shortest route found by the ant system
6.3	The snapshot of pheromone intensities after 20 episodes 183
6.4	Update the pheromone level according to the best route of all groups 185
6.5	Update the pheromone level between each pair of groups 186
6.6	Update the pheromone level according to the ring structure 186
6.7	Update the Pheromone level to the neighbours according to the group number j differs by one bit
6.8	Performance comparison among AS , ACS and two arbitrarily chosen strategies for EIL101 data set
6.9	Performance comparison among AS , ACS and two arbitrarily chosen strategies for ST70 data set
6.10	Performance comparison among AS , ACS and two arbitrarily chosen strategies for TSP225 data set
6.11	Ant tends moving toward the object with dense cluster 19
6.12	Conventional search route using Equation 6.1 199
6.13	Shrinking search route using Equation 6.14
6.14	Clustering result of CACO $(N_1 = \frac{1}{55})$
6.15	Clustering result of CACO $(N_1 = \frac{1}{20})$
6.16	Clustering result of CACO $(\gamma = 1)$
6.17	Clustering result of CACO $(\gamma = 5)$
6.18	Clustering results of Four-Cluster by ACODF algorithm 20^4
6.19	Clustering results of Four-Cluster by CACO algorithm 204
6.20	Clustering results of Four-Cluster by DBSCAN algorithm 208
6.21	Clustering results of Four-Cluster by CURE algorithm 208
6.22	Clustering results of Four-Bridge by ACODF algorithm 200 $$
6.23	Clustering results of Four-Bridge by DBSCAN algorithm 200 $$
6.24	Clustering results of Four-Bridge by CURE algorithm 20
6.25	Clustering results of Four-Bridge by CACO algorithm 20
6.26	Clustering results of Smile-Face by ACODF algorithm 208 $$
6.27	Clustering results of Smile-Face by DBSCAN algorithm 208
6.28	Clustering results of Smile-Face by CURE algorithm 209
6.29	Clustering results of Smile-Face by CACO algorithm 209
6.30	Clustering results of Shape-Outliers by ACODF algorithm 210

LIST OF FIGURES	X

6.31	Clustering results of Shape-Outliers by DBSCAN algorithm	210
6.32	Clustering results of Shape-Outliers by CURE algorithm	211
6.33	Clustering results of Shape-Outliers by CACO algorithm	211

List of Tables

2.1	Results of Experiment for Gaussian Source	34
2.2	Results of Experiment for Rectangular Clusters	35
2.3	Results of Experiment for Gauss-Markov Source	37
2.4	Results of Experiment for Elliptic Clusters	38
2.5	Summarisation of Cases Regarding Cost of Replacing o_{old} with o_{new} as medoid	39
2.6	Results of Experiment for Gaussian Source	60
2.7	Results of Experiment for Gauss-Markov Source	61
2.8	Results of Experiment for Gauss-Markov Source–Compared with $CLARANS$ and Extended $CLASA$ Algorithms	62
2.9	Results of Experiment for Rectangular Clusters	63
2.10	Results of Experiment for Rectangular Clusters–Compared with $CLARANS$ and Extended $CLASA$ Algorithms	64
2.11	Results of Experiment for Elliptic Clusters	65
2.12	Results of Experiment for Curved Clusters	66
2.13	Experimental Results for Ten Runs of <i>CLARANS</i> , <i>CLARANS</i> - <i>ITP</i> , <i>CLARANS</i> - <i>M</i> and <i>CLARANS</i> - <i>MITP</i> Algorithms for 8 Medoids of <i>Lena</i> Dataset	67
2.14	Experimental Results for Ten Runs of <i>CLARANS</i> , <i>CLARANS</i> - <i>ITP</i> , <i>CLARANS</i> - <i>M</i> and <i>CLARANS</i> - <i>MITP</i> Algorithms for 16 Medoids of <i>Lena</i> Dataset	68
2.15	Results of Experiment for <i>Lena</i> Dataset	68
2.16	Results of Experiment for Co-Occurence Texture of Corel Image Collection	69
3.1	Results of Experiment for Four Elliptic Clusters	79
3.2	Results of Experiment for Twelve Elliptic Clusters	81
3.3	Results of Experiment for Compact Clusters	82

LIST OF TABLES xii

3.4	Results of Experiment for <i>Lena</i> Image	83
3.5	Results of Experiment for Four Elliptic Clusters	88
3.6	Results of Experiment for Twelve Elliptic Clusters	89
3.7	Results of Experiment for Compact Clusters	90
3.8	Results of Experiment for Gauss-Markov Source	91
3.9	Results of Experiment for <i>Lena</i> Image	93
4.1	Performance Comparison Using $Lena$ Image as The Training Patterns	105
4.2	Performance Comparison Using $Baboo$ Image as The Training Patterns	106
4.3	Performance Comparison Using Pepper Image as The Training Pattern	106
4.4	Performance Comparison for M/R VQ with Mean Codebook Size 64 and Residual Codebook Size 64 Using $Pepper$ Image	121
4.5	Performance Comparison for M/R VQ with Mean Codebook Size 64 and Residual Codebook Size 64 Using $Lena$ Image	122
4.6	Performance Comparison for M/R VQ with Mean Codebook Size 128 and Residual Codebook Size 128 Using $Pepper$ Image	122
4.7	Performance Comparison for M/R VQ with Mean Codebook Size 128 and Residual Codebook Size 128 Using $Lena$ Image	123
4.8	Performance Comparison for M/R VQ with mean Codebook Size 256 and Residual Codebook Size 256 Using $Pepper$ Image	123
4.9	Performance Comparison for M/R VQ with Mean Codebook Size 256 and Residual Codebook Size 256 Using $Lena$ Image	124
4.10	Performance Comparison for M/R VQ with Mean Codebook Size 64 and Residual Codebook Size 64 Using F16, $Pepper$ and $Lena$	
	Image	125
	Error for Different Sample Size (Number of Cluster=256)	130
	Error for Different Number of Clusters (Sample Size = 10000)	131
4.13	The relationship between the $PSNR$ of the watermarked image and the number of clusters	139
4.14	NHS Value for Various Attacks	140
4.15	Comparisons of various fast search algorithms for 'LENA' image in the training set	161
4.16	Comparisons of various fast search algorithms for 'BABOO' image in the training set	161

LIST OF TABLES xiii

5.1	Asymmetric initialization ranges and V_{max} values	174
5.2	Performance Comparison of PSO and $PPSO$ with The First Communication Strategy for Rosenbrock Function	175
5.3	Performance Comparison of PSO and $PPSO$ with The First Communication Strategy for Rastrigin Function	175
5.4	Performance Comparison of <i>PSO</i> and <i>PPSO</i> with The Second Communication Strategy for Griewank Function	175
5.5	Performance Comparison of PSO and $PPSO$ with The Third Communication Strategy	176
6.1	The performance of ACS with communication strategies (strategy $1 \sim 7$) obtained in comparison with AS and ACS for EIL101 data set on TSP problem	192
6.2	The performance of ACS with communication strategies (strategy $1 \sim 7$) obtained in comparison with AS and ACS for ST70 data set on TSP problem	193
6.3	The performance of ACS with communication strategies (strategy $1 \sim 7$) obtained in comparison with AS and ACS for TSP225 data set on TSP problem	194

Abstract

Clustering algorithms have been widely applied and different approaches have been developed for different domains of application. This thesis investigates efficient and effective clustering and soft computing algorithms. In the investigation of k-medoids algorithms, several improved algorithms are proposed, such as the Clustering Large Applications Based on Simulated Annealing (CLASA) algorithm, the Multi-Centroids with Multi-Runs Sampling Scheme (MCMRS) and Incremental Multi-Centroid, Multi-Run Sampling Scheme (IMCMRS) algorithms. The Partial Distance Search (PDS), Triangular Inequality Elimination (TIE) and Previous Medoid Index are also presented to improve the clustering speed of k-medoids based algorithms. In addition, a new memory utilization scheme is derived and applied to efficient k-medoids algorithms.

In the investigation of centroid-based clustering algorithms, the tabu search with simulated annealing algorithm is proposed and applied to codebook design for vector quantization. Genetic clustering is also presented for mean-residual vector quantization. Several theorems based on Hadamard Transform for nearest neighbour search are presented and applied to efficient cluster (codeword) search for vector quantization. A label bisecting clustering algorithm is proposed and applied to create a robust watermarking technique.

Parallel particle swarm optimization based on three communication strategies are proposed to solve the problems in which the relationship between parameters are either independent, loosely correlated, strongly correlated or unknown. Seven communication strategies for $Ant\ Colony\ Systems\ (ACS)$ are proposed to improve the ACS for the traveller salesman problem and the $Constrained\ Ant\ Colony\ Optimization\ (CACO)$ based on the quadratic metric, sum of k nearest neighbour distance, constrained addition of pheromone and a shrinking range strategy is also proposed and demonstrated to be better than the $Ant\ Colony\ Optimization\ with\ Different\ Favor\ (ACODF)$.

Certification

I certify that this thesis does not incorporate without acknowledgement any material previously submitted for a degree or diploma in any university; and that to the best of my knowledge and belief it does not contain any material previously published or written by another person except where due reference is made in the text.

As requested under Clause 14 of Appendix D of the *Flinders University Research Higher Degree Student Information Manual* I hereby agree to waive the conditions referred to in Clause 13(b) and (c), and thus

- Flinders University may lend this thesis to other institutions or individuals for the purpose of scholarly research;
- Flinders University may reproduce this thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

Signed	Dated	

Shu-Chuan Chu

Acknowledgements

I would like to first thank my supervisor, Professor John F. Roddick, for his constant guidance, patience and encouragement.

I would like to thank the principal of National Kaohsiung University of Applied Sciences, Professor Jen-Yi Lin, and the former principal, Professor Kuang-Chih Huang without whose high level of support and help, I cannot contentedly continue to finish this work in my academic career. Further, I would like to extend my thanks to other colleagues and friends who have provided help and advice during my PhD study. Without particular order, thanks to: Professor Wen-Kuei Cheng, Professor Bin-Yi Liao, Chin-Shiuh Shieh and Chin-Yen Chang.

Finally, I wish to dedicate this thesis to my husband, my mother and my parents in law who have guided me all through my life. I would like to thank them for all the love, suggestions and encouragements.

Shu-Chuan Chu January 2004 Adelaide.