ABSTRACT

Clustering is a discipline devoted to find and describe cohesive or homogeneous chunks in data, the clusters. An example of clustering problem is the automatic revealing of meaningful parts in a digitalized image. The motivation for the focus on data clustering is the fact that data clustering is an important process in pattern recognition and machine learning. Clustering algorithms are used in many applications such as image segmentation, vector and color image quantization, compression, etc. Therefore, finding an efficient clustering algorithm is very important for researchers in many different disciplines.

The primary objective of this thesis is to utilize G*enetic Algorithm* (GA) as a clustering tool for the unsupervised classification of grayscale image data. It presents two variants of GA: the first variant is based on the canonical GA while the second variant is based on *compact* GA, cGA. The characteristics components of each algorithm are presented in term of individual and population representation, fitness function evaluation, evolution (selection, crossover, and update) operators, and stopping condition. These two genetic algorithms are then coupled with one popular local-search cluster algorithm, known as K-means algorithm. By coupling, the objective is to harness the power of each algorithm: GA search exploration power and K-means search exploitation power. Moreover, the canonical mechanism of perturbation operators symbolized by both crossover and mutation is imitated in a modified version of cGA in an attempt to improve its search power.

To show the applicability of the presented clustering algorithms, Human medical MRI and land sat images, are used in the experiments. Also, the experiments considered different number of clusters. Comparison results are reported in qualitative terms (i.e. visually) and in quantitative terms using quantization error, weighted error (sum of cluster compactness, clusters separation, and quantization error), and compactness-separation ratio. Results demonstrate that cross-fertilization between the two algorithms is of being benefit in image data clustering, and it outperforms K-means and genetic-based algorithms when they operated individually. Additionally and more interestingly, the modified cGA outperforms the traditional cGA, which leverage the influence of the added perturbation operators including two-point crossover and binary mutation.

Acknowledgments

I would like to express my sincere appreciation to my supervisors, Dr. Bara'a Ali and Dr. Sawsan Kamal, for giving me the major steps to go on to explore the subject, sharing with me the ideas in my research "Clustering Approach for Unsupervised Image Classification using Genetic Algorithm" and discuss the points that I felt they are important.

Grateful thanks for the Head of Department of Computer Science, Dr. Taha S. Bashaga,

Also ,I would like to say "thank you" to my Parents and my Husband for their supporting and encouragement and I want to thank my faithful friends for supporting and giving me advises.

Supervisor Certification

We certify that this thesis was prepared under our supervision at the Department of Computer Science / College of Science / Al-Nahrain University, by **Kholood Jamal Moulod** as partial fulfillment of the requirements for the degree of Master of Science in Computer Science.

Supervisors

Signature: Baraa
Name : **Bara'a Ali**
Title : **Assist. Prof.**
Date : / /**2008**

Signature:
Name : **Sawsan Kamal**
Title : **Lecturer**
Date : / /**2008**

The Head of the Department Certification

In View of the available recommendations, I forward this thesis for debate by the examination committee.

Signature: T Bashaga
Name : **Dr. Taha S. Bashaga**
Title : **Head of the department of Computer Science,**
Al-Nahrain University.
Date : / /**2008**

Supervisor Certification

We certify that this thesis was prepared under our supervision at the Department of Computer Science/College of Science/Al-Nahrain University, by **Wurood Saad Ibraheem Al-Obaidi** as partial fulfillment of the requirements for the degree of Master of Science in Computer Science.

Supervisors

Signature:	Signature:
Name: **Lamia H. Khalid**	Name: **Ban N. Al-Kallak**
Title: **Assist. Prof.**	Title: **Lecturer**
Date: / / **2007**	Date: / / **2007**

The Head of the Department Certification

In view of the available recommendations, I forward this thesis for debate by the examination committee.

Signature:
Name: **Dr. Taha S. Bashaga**
Title: **Head of the department of Computer Science, Al-Nahrain University.**
Date: / / **2007**

Examining Committee Certification

We certify that we have read this thesis and as an examining committee, examined the student in its content and what is related to it, and that in our opinion it meets the standard of a thesis for the degree of Master of Science in Computer Science.

Examining Committee Certification

Signature:

Name: **Dr. radhi A. Zboon**

Title: **Assistance Professor (Chairman)**
Date: / / **2008**

Signature:	Signature:
Name: **Dr. Amer S. Al-Malah**	Name: **Dr. Ban N. Thanon**
Title: **Assistance Professor (Member)**	Title: **Lecturer (Member)**
Date: / / 2008	Date: / / 2008

Supervisors Certification

Signature:	Signature:
Name: **Dr.Bara'a A. Attea**	Name: **Dr. Sawsan K. Thamer**
Title: **Assistance Professor**	Title: **Lecturer**
Date: / / **2008**	Date: / / **2008**

The Dean of the College Certification

Approved by the Council of the College of Science

Signature:

Name: **Dr. LAITH ABDUL AZIZ AL - ANI**

Title: The Dean of College of Science**, Al-Nahrain University.**

Date: / / **2008**

Chapter One

Overview

Chapter Two

Data Clustering: Mathematical Formulation

Chapter Three

Genetic Algorithm:
Canonical and Compact Variants

Chapter Four

GA-based Clustering Algorithms

Chapter Five

Experimental Results

Chapter Six

Conclusions
and
Future Suggestions

References

DEDICATED TO

My Parents…
My Husband and Son…
My Sisters and Brothers…
My Friends…

Kholood

List of Algorithms

List of Denotations

GA	Genetic Algorithm.
cGA	compact Genetic Algorithm.
MRI	Magnetic reasoning image.
AI	Artificial Intelligent.
PSO	Particle Swarm Optimization.
DCPSO	Dynamic Clustering Particle Swarm Optimization.
NP	Non-deterministic Polynomial problem.
EA	Evolutionary Algorithm.
Γ	cGA chromosome decoding function, $\Gamma : \{0,1\}^{8*K} \to \{0,255\}^{K}$.
δ	Positive scaling function.
Φ_{cGA}	Fitness function of a cGA chromosome, $\Phi_{cGA} : \{0,255\}^{K} \to R$.
Φ_{GA}	Fitness function of a GA chromosome, $\Phi_{GA} : \{0,255\}^{l_{GA}} \to R$.
$a, \; ab, \; b$	cGA chromosome, $a, \; ab, \; b \in \{0,1\}^{l_{cGA}}$.
$\{S_1,\ldots,S_K\}$	Partition of image Z in K disjoint clusters $S_k \subset Z, k = 1,\ldots,K$.
$chrom$	Canonical chromosome definition.
$chrom_i$	i^{th} chromosome (individual) in the GA population, $chrom_i = \{c_{i,k}\} = (c_{i,1}, c_{i,2},\ldots,c_{i,k})$.
c_k	Centroid of cluster $S_k, k = 1,\ldots,K$.
$c_{i,k}$	k^{th} cluster centroid of the i^{th} GA chromosome.
$d(x, y)$	Distance between two pixel points.
$d_{man}()$	Manhattan distance.
$d_{eucl}()$	Euclidean distance.

$d_{\max}$	Intra-distance (compactness error).
$d_{\min}$	Inter-distance (separation error).
g	GA or cGA generation number.
g_{epoch}	An epoch of GA generations.
$g_{\max}$	Maximum number of GA generations.
ι	Termination criterion.
J_e	Quantization error.
K	Number of clusters in an image I.
l	Canonical chromosome length.
l_{cGA}	cGA chromosome length, $l_{cGA} = 8 * K$.
l_{GA}	GA chromosome length, $l_{GA} = K$.
l_x	Length of chromosome segment.
m	Number of chromosome segments.
P_m	Mutation rate.
n	GA and cGA population size.
n_k	Number of pixels in cluster $S_k, k = 1,...,K$.
N_p	Number of patterns i.e. pixel in image Z, $N_p = ImageHeight * ImageWidth$.
p	l_{cGA}-*dimensional* probability vector, $p = [0,1.0]^{l_{cGA}}$.
POP	GA population, $Pop = \{chrom_1,...,chrom_n\}$.
P_c	Crossover rate.
$U()$	Uniformly distributed random number.
T	GA population subset
w_1	Priority of quantization error, $0 \le w_1 \le 1.0$.
w_2	Priority of compactness error, $0 \le w_1 \le 1.0$.

w_3	Priority of separation error, $0 \leq w_1 \leq 1.0$.
Z	Data set of grayscale image pixels $Z = \{z_1, z_2, ..., z_{Np}\}$.

List of Tables

List of Figures

بسم الله الرحمن الرحيم

(الْحَمْدُ لِلَّهِ الَّذِي هَدَانَا لِهَٰذَا وَمَا كُنَّا

لِنَهْتَدِيَ لَوْلَا أَنْ هَدَانَا اللَّهُ)

صدق الله العظيم

الأعراف ٤٣

References

[Ahn03]	C. W. Ahn, and R. S. Ramakrishna, “***Elitism-based Compact Genetic Algorithm***”, IEEE Trans. on Evolutionary Computation, Vol. 7, No. 4, pp. 368-385, 2003.
[Ami02]	Y. Amit, "***2D Object Detection and Recognition Models, Algorithms, and Networks***", The MIT Press, Cambridge, Massachusetts, London, England, 2002.
[Apo01]	C. Aporntewan, and P. Chongstitvatana, “***A Hardware Implementation of the Compact Genetic Algorithm***”, In Proc. of the 2001, IEEE Congresson Evolutionary Computation, Korea, pp.624-629, 2001.
[Ban01]	S. Bandyopadhyay, and S. K. Pal, “***Pixel Classification Using Variable String Genetic Algorithms with Chromosome Differentiation***”, IEEE Trans. on Geoscience and Remote Sensing, vol. 39, no. 2, Feb. 2001.
[Bäc93]	T. Bäck, and H.-P. Schwefel, “***An Overview of Evolutionary Algorithms for Parameter Optimizations***”, Evolutionary Computation, Cambridge, Vol.1, No.1, pp. 1-23, 1993.
[Bäc95]	T. Bäck, and H.-P. Schwefel, “***Evolution Strategies I: Variants and their Computational Implementation***”, Genetic Algorithms in Engineering and Computer Science, pp.11-126, 1995.
[Ban07]	S. Bandyopadhyay, U. Maulik, and A. Mukhopadyay, “***Multiobjective Genetic Clustering for Pixel Classification in Remote Sensing Imagery***”, IEEE Trans. on Geoscience and Remote Sensing, Vol. 45, No. 5, pp. 1506-1511, 2007.

[Ber02]	P. Berkhin, "***Survey of Clustering Data Mining Techniques***", www.ee.ucr.edu/~barth/EE242/Clusterung_Survey.pdf, 2002.
[Car96]	C. Carpineto, and G. Romano, "***A Lattic Conceptual Clustering System and its Application to Browsing Retrieval***", Machine Learning, Vol. 24, No. 2, pp. 95-122, 1996.
[Cui05]	X. Cui, T. E. potok, and P. Palathingal, "***Document Clustering using Particle Swarm Optimization***", 2005.
[Deb91]	K. Deb, "***Binary and Floating-point Function Optimization Using Messy Genetic Algorithms***", PhD dissertation, Dept. Engineering Mechanics in the Graduate School of the University of Alabama, Tuscaloosa, Alabama, 1991.
[Gol89]	D.E. Goldberg, "***Genetic Algorithm in Search, optimization, and Machine learning***", Addison Wesley, 1989.
[Hal01]	M. Halkidi, and M. Vazirgiannis, "***On Clustering Validation Techniques***", Intelligent Information Systems Journal, Kluwer Publishers, vol. 17, no. 2-3, pp.107-145, 2001.
[Ham02]	G. Hammerly and C. Elkan, "***An Alternative to the K-means Algorithm that Find Better Clusterings***", In Proceedings of the ACM Conference on Information and Knowledge Management (CIKM-2002), pp. 600-607, 2002.
[Har97]	G. R. Harik, F.G. Lobo, and D. E. Goldberg, "***The Compact Genetic Algorithm***", IlliGAL Report No.97006, Urbana: University of Illinois at Urbana-Champaign, Illinois Genetic Algorithms Laboratory. 1997.

[Hau04]	R. L. Haupt, and S. E. Haupt, "***Practical Genetic Algorithms***", 2nd edition, John Wiley and Sons, INC., Publication, 2004.
[Hol00]	J. H. Holland, "***Building Blocks, Cohort Genetic Algorithms, and Hyperplane-Defined functions***", Evolutionary Computation, Vol. 8, No. 4, pp. 373-391, 2000.
[Jai99]	A.K Jain, M.N.Murty, and P. J. Flynn, "***Data clustering: a Review***", ACM Computing Surveys, Vol. 31, No. 3, pp. 264-323, 1999.
[Jud98]	D. Judd, P.K. McKinley, and A.K. Jain, "***Large-scale Parallel Data Clustering***", IEEE Transactions on Volume 20, Issue 8, Aug. 1998 Page(s):871 - 876
[Kar95]	H. Kargupta, "***Search, Polynomial Complexity, and the Fast messy Genetic Algorithms***", PhD Dissertation, Tuscaloosa, Alabama, 1995.
[Kon00]	A. Konar, "***Artificial Intelligence and Soft Computing Behavioral and Cognitive Modeling of the Human Brain***", CRC Press, 2000.
[Koz92]	J. R. Koza, "***Genetic Programming, On the Programming of Computers by Means of Natural Selection***", A Bradford Book, The MIT Press, Cambridge, Massachusetts, 1992 first edition, 1992.
[Kun98]	L. I. Kuncheva and J. C. Bezdek , "***Nearest Prototype Classification: Clustering, Genetic Algorithms, or Random Search***", IEEE Trans. on Systems, Man, And Cybernetics—Part C: Applications And Reviews, Vol. 28, No. 1, pp.160-164, Feb. 1998.

[Loo05]	M. Looks, “***Learning Computer Programs with the Bayesian Optimization Algorithm***”, M. Sc. Thesis, University of Washington, School of Engineering and Applied Science, 2005.
[Mau03]	U. Maulik, and S. Bandyopadhyay “***Fuzzy Partitioning Using Real Coded Variable Length Genetic Algorithm for Pixel Classification***”, IEEE Trans. on Geoscienlce and Remote Sensing, vol. 41, no. 5, pp. 1075-1081, 2003.
[Men04]	A. Menon, "***Frontiers of Evolutionary Computation***", ProductSoft Inc., Pittsburgh, Pennsylvania, Kluwer Academic Publishers, USA, 2004.
[Mir05]	B. Mirkin, "***Clustering for Data Mining A Data Recovery Approach***", Taylor and Francis Group, LLC, 2005.
[Mit96]	itchell, "***An Introduction to Genetic Algorithms***", A Bradford Book, MIT Press, 1996.
[Muk06]	A. Mukhopadyay, S. Bandyopadhyay , and, U. Maulik, “***Clustering Using Multi-objective Genetic Algorithm and its Application to Image segmentation***”, IEEE International Conf. on Systems, Man, and Cybernetics, pp. 2678-2683, Oct. 2006.
[Omr04]	M. G. H. Omran, “***Particle Swarm Optimization Methods for Pattern Recognition and Image Processing***”, Ph.D. thesis, University of Pretoria, Pertoria, 2004.

[Omr05]	M. G. H. Omran, A. P. Engelbrecht, and A. Salman, "***Dynamic Clustering using Particle Swarm Optimization with Application in Unsupervised Image Classification***", Trans. on Engineering, Computing and Technology, Vol.9, pp. 199-204, Nov. 2005.
[Per02]	S. W. Perry, H. Wong, and L. Guan, "***Adaptive Image Processing A Computational Intelligence Perspective***", CRC Press LCC, 2002.
[Sas07]	K. Sastry, and A. Orriols-Puig, "***Extended Compact Genetic Algorithm in Matlab***", IlliGAL Report No. 2007009, 2007.
[Sch81]	H.-P. Schwefel, "***Numerical Optimization of Computer Models***", New York: John Wiley & Sons, 1981.
[Sch97]	H.-P Schwefel, and T. Bäck, "***Artificial Evolution: How and Why***", Genetic Algorithms and Evolution Strategies in Engineering and Computer Science. Recent Advances and Industrial Applications. John Wiley and Sons Ltd. pp. 1-19, 1997.
[Spe92]	W. M. Spears, "***Crossover or Mutation?***", Proceedings of the 2nd Foundations of Genetic Algorithms Workshop, ed. Whitly, D. San Mateo, CA: Morgan Kaufmann, pp.230-236, 1992.
[Spe93]	W. M. Spears, K. A., De Jong, T. Bäck, ,D. B.Fogel, and H. de Garis, "***An Overview of Evolutionary Computation***", European Conference on machine learning, ed. Brazdil, P. Berlin: Springer Verlag, 667: pp.442-459, 1993.

[Tek07]	K. Teknomo, “***K-means Clustering Tutorial***”, http;//people.revoledu.com/kardi/tutorial/kMean/index.html , 2007.
[Tur01]	R. H. Turi, “***Clustering-based Color Image Segmentation***”, Ph.D. Thesis, Monash, University of Australia, 2001.
[Vee03]	C. Veenman, M. Reinders, and E. Backer, “***A Cellular Coevolutionary Algorithm for Image Segmentation***”, IEEE Trans. on Image Processing, Vol. 12, No. 3, pp. 304-316, 2003.
[Wei07]	T. Weise, "***Global Optimization Algorithms- Theory and Applications***", www.it-Weise.de/projects/book.pdf , 2007.
[Wit05]	I. H. Witten and E. Frank, "***Data Mining Practical Machine Learning Tools and Techniques***", 2nd edition, Morgan Kufmann Publishers, 2005.
[Wu01]	M. Wu, M. Fuller, and R. Wilkinson, “***Using Clustering and Classification Approaches in Interactive Retrieval***”, information processing & management 37 459-484, 2001.

Contents

Chapter 1: Overview

Chapter 2: Data Clustering: Mathematical Formulation

Chapter 3: Genetic Algorithm: Canonical and Compact Variants

Chapter 4: GA-based Clustering Algorithms

Chapter 5: Experimental Results

Chapter 6: Conclusions and Future Suggestions

Republic of Iraq
Ministry of Higher Education
and Scientific Research
Al-Nahrain University
College of Science

Clustering Approach for Unsupervised Image Classification using Genetic Algorithm

A Thesis
Submitted to the College of Science, Al-Nahrain University in Partial Fulfillment of the Requirements for The Degree of Master of Science in Computer Science

By
Kholood Jamal Moulod
(B.Sc. 2004)

Supervised by
Dr. Bara'a Ali Attea **Dr. Sawsan Kamal Thamer**

April 2008 **Rabe'e El-Thani 1429**

الخلاصة

العنقدة هي نظام خصص لإيجاد ووصف قطع متماسكة ومتجانسة في البيانات، هي العناقيد. مثال عن مسألة العنقدة هو الإظهار أو الكشف الأوتوماتيكي لأجزاء ذات معنى في الصورة الرقمية. الحث على التركيز على عنقدة البيانات بسبب إن عنقدة البيانات هي عملية مهمة في تمييز الأنماط و التعليم الالي. خوارزميات العنقدة تستخدم في عدة تطبيقات مثل تقطيع الصور، حساب متجه ولون الصورة، ضغط البيانات... الخ. لذلك، إيجاد خوارزمية عنقدة كفوءة هو شيء مهم جدا للباحثين في عدة أنظمة مختلفة.

الهدف الرئيسي لهذه الأطروحة هو استخدام الخوارزمية الجينية كأداة عنقدة للتصنيف غير الموجه للصورة ذات تدرج رمادي. هذه الأطروحة تقدم نوعين مختلفين من الخوارزمية الجينية: النوع الأول يعتمد على خوارزمية جينية بسيطة، بينما النوع الثاني يعتمد على الخوارزمية الجينية المضغوطة. المكونات الخصائصية لكل خوارزمية تعرض بصيغة تمثيل الفرد والمجموعة، حساب دالة الملائمة، عمليات التطور (الانتقاء، التقاطع، والتحديث)، و شرط التوقف. هاتان الخوارزميتان الجينيتان سوف تربطان كل على حدة مع خوارزمية عنقدة بحث محلي تعرف ب(*K means algorithm*). الهدف من الربط هو تسخير قوة كل خوارزمية: قوة الخوارزمية الجينية و قوة (*K means*). بالإضافة الى ذلك، الآلية القانونية لعمليات التشويش التي يرمز لها بالتقاطع والطفرة المفاجئة تحاكي بنسخة محدثة من الخوارزمية الجينية المضغوطة في محاولة لتحسين قوة بحثها.

لعرض قابلية تطبيق خوارزميات العنقدة، تم استخدام صور أشعة رنين مغناطيسي لدماغ الإنسان، صور أقمار صناعية. أيضا تم استخدام أعداد عناقيد مختلفة في الاختبارات العملية. نتائج المقارنة سجلت بصيغة نوعية (مرئية) و صيغة كمية باستخدام (*Quantization Error*)و (*Weighted Sum of cluster compactness, cluster separation, quantization error*) والنسبة بين (*cluster compactness and cluster separation*).

النتائج أظهرت أن (*cross-fertilization*) بين الخوارزميتين يعتبر فائدة في عنقدة الصورة، و يعمل بشكل أفضل من خوارزمية (*K-means*) و الخوارزميات الجينية عندما يعملان بشكل منفرد. بالإضافة إلى ذلك وهو الأكثر أهمية، الخوارزمية الجينية المضغوطة المحدثة تعمل بشكل أفضل من الخوارزمية الجينية المضغوطة العادية، لأنها زادت من قوة تأثير معاملات التشويش المضافة والتي تتضمن التقاطع ذو نقطتين و الطفرة الثنائية.

جمهورية العراق
وزارة التعليم العالي والبحث العلمي
جامعة النهرين
كلية العلوم

طريقة عنقدة للتصنيف غير الموجه للصورة باستخدام الخوارزمية الجينية

رسالة
مقدمة إلى كلية العلوم في جامعة النهرين كجزء من متطلبات
نيل درجة الماجستير في علوم الحاسبات

من قبل
خلود جمال مولود
(بكالوريوس جامعة النهرين ٢٠٠٤)

اشراف
د. براء علي   د.سوسن كمال

ابريل ٢٠٠٨   ربيع الثاني ١٤٢٩