EFFICIENT GENETIC PARTITIONING-AROUND-MEDOID ALGORITHM FOR CLUSTERING

SARMAD MAKKI MOHAMMED GARIB

FSKTM 2019 50

EFFICIENT GENETIC PARTITIONING-AROUND-MEDOID ALGORITHM FOR CLUSTERING

By

SARMAD MAKKI MOHAMMED GARIB

Thesis Submitted to the School of Graduate Studies, Universiti Putra Malaysia, In Fulfilment of the Requirements for the Degree of Doctor of Philosophy

May 2019

DEDICATION

To my father, mother and my beloved family.

Sarmad Makki

Abstract of thesis presented to Senate of Universiti Putra Malaysia in fulfilment of the requirement for the degree of Doctor of Philosophy

EFFICIENT GENETIC PARTITIONING-AROUND-MEDOID ALGORITHM FOR CLUSTERING

By

SARMAD MAKKI MOHAMMED GARIB

May 2019

Chairma : Associate Professor Razali Yaakob, PhD
Faculty : Computer Science and Information Technology

Throughout the years, considerable efforts made to tackle the clustering problem. Yet, because of the nature of the clustering problem, finding an efficient clustering optimization algorithm with reasonable performance is still an open challenge. In general, genetic based clustering algorithms showed the ability to reach near global optimal solution. These algorithms mostly built upon the partitioning *k*-means clustering algorithm. Nevertheless, these algorithms either dealt with part of the issues inherited in the *k*-means or in turn produce some deficiency by itself.

One of the main issues in genetic *k*-means based algorithms is their sensitivity to outliers and unevenly distributed clusters due to the mean compromised computations. Besides, these algorithms more frequently implemented with the lengthy and redundant fixed *N* length integer encoding. Additionally, they used either random or non-mathematically proved population initializations methods. Adopting the medoid instead of the mean can enhance the efficiency. However, the complexity of the *k*-medoid based algorithms in general is more than the complexity of the *k*-means based algorithms. Lastly, in order to externally judge the validity of these types of clustering algorithms, there is a need for a method to correctly and efficiently evaluate their variant multiclass clustering results.

This study utilizes genetic algorithms based upon the medoid rather than the mean as a centroid-selection schema to improve the clustering efficiency. It uses the compact and unique variable *K* length real encoding. Accordingly, the corresponding genetic operators are adapted to suite the medoid and to incorporate much clustering-specific domain knowledge. The algorithm is also preceded with careful seeding using mathematically proved to converge *k*-means++ algorithm. Secondly, by utilizing the medoid property as being an actual data item in the dataset and with the aid of the

proposed indexing method, the complexity of the computation is notably decreased. Finally, an algorithm is developed for automatic evaluation of external validity measures on the generated variant multiclass clustering results.

The experiments are divided into four interdependent sets. The first set showed the efficiency and performance of the proposed composing techniques including: *k*-mean++ algorithm for initialization, fitness scaling for fitness selection, the proposed split and merge mutation operator, and choosing the medoid instead of the mean as a centroid-selection schema. The rest sets of experiments carried out to evaluate the proposed algorithms. Precisely, the second set revealed that the proposed genetic medoid based algorithms with both *DB* and *VRC* fitness functions produced more accurate results compared with the genetic means based algorithms in terms of the *F*-score. The third set affirmed that the enhancement on the proposed algorithm, which made use of indexing method that suits the medoids, could boost the performance to about 9 to 27 times in terms of execution time depending on the complexity of the dataset. Finally, the fourth set dealt with the developed method for externally evaluating the variant multi class clusters. The experiments acknowledged that the proposed relabeling algorithm could perform better in quality than the external F-score measure. Also it outperform in term of its complexity $O(N^2)$ compared with the complexity $O(2^N)$ of the Adjusted Rand Index (*ARI*).

Abstrak tesis yang dikemukakan kepada Senat Universiti Putra Malaysia sebagai memenuhi keperluan untuk ijazah Doktor Falsafah

ALGORITMA PEMETAKAN GENETIK BERDASARKAN MEDOID YANG EFISIEN UNTUK PENGELOMPOKAN

Oleh

SARMAD MAKKI MOHAMMED GARIB

Mei 2019

Pengerusi : Profesor Madya Razali Yaakob, PhD
Fakulti : Sains Komputer dan Teknologi Maklumat

Di sepanjang tahun, usaha yang agak banyak telah dibuat bagi mengatasi masalah pengelompokan. Namun, disebabkan sifat masalah pengelompokan, pencarian algoritma optimisasi pengelompokan yang efisien dengan prestasi yang munasabah merupakan cabaran yang masih terbuka. Secara amnya, algoritma pengelompokan berdasarkan genetic telah menunjukkan keupayaan untuk hampir mencapai penyelesaian optimal global. Algoritma tersebut kebanyakannya telah dibina berasaskan pemetakan algoritma pengelompokan *K*-min. Walau bagaimanapun, algoritma tersebut sama ada berkaitan dengan sebahagian isu yang wujud dalam *K*-min atau sebaliknya menghasilkan beberapa kekurangan sendiri.

Salah satu isu utama dalam genetik *K*-min berdasarkan algoritma ialah sensitiviti mereka pada unsur luar dan kluster yang tidak tersebar rata disebabkan min komputasi yang dikompromi. Tambahan pula, algoritma tersebut lebih kerap diimplementasikan dengan pengekodan integer *N* panjang yang tetap bertindan dan berpanjangan. Di samping itu, mereka sama ada menggunakan kaedah pememulaan populasi terbukti secara rawak atau bukan matematik. Menerima pakai medoid dan bukan min dapat meningkatkan keberkesanan. Walau bagaimanapun, kekompleksitian algoritma berdasarkan *K*-medoid secara amnya adalah lebih daripada kekompleksitian algoritma berdasarkan *K*-min. Akhirnya, bagi menilai validiti jenis algoritma pengelompokan tersebut secara luaran , terdapat keperluan untuk suatu kaedah bagi membetulkan dan secara efisien menilai keputusan pengelompokan multikelas varian mereka..

Kajian ini menggunakan algoritma genetik bagi memperbaiki keberkesanan pengelompokan data berdasarkan medoid dan bukan min sebagai suatu skema pemilihan sentroid. Ia menggunakan pengekodan sebenar variabel *K* panjang yang unik dan kompak. Dengan itu, operator genetik yang sepadan telah disesuaikan bagi

memadankan medoid dan bagi menginkorporasikan pengetahuan domain pengelompokan khusus yang banyak. Algoritma tersebut juga didahului dengan pembenihan teliti menggunakan algoritma *K*-min++ bertumpu terbukti secara matematik. Kedua, dengan menggunakan sifat medoid sebagai item data sebenar dalam set data dan dengan bantuan kaedah pengindeksan yang disyorkan, kekompleksitian komputasi telah menurun secara jelas. Akhirnya, suatu algoritma bagi penilaian automatik bagi pengukuran validiti luaran ke atas keputusan pengelompokan multikelas varian tersebut telah dibangunkan.

Akhirnya, algoritma berdasarkan medoid genetik yang disyorkan menghasilkan dapatan yang lebih baik berbanding dengan algoritma berdasarkan min genetik dari segi skor F. Dengan kaedah pengindeksian, ia juga lebih baik dari segi tempoh pelaksanaan. Akhir sekali, kaedah yang dibangunkan tersebut bagi menilai secara luaran dapatan pengelompokan menunjukkan prestasi yang lebih baik berbanding dengan pengukuran skor F luaran dari segi kualiti dan lebih baik daripada kaedah Indeks Rand Terlaras (ARI) dari segi kekompleksitian masa..

ACKNOWLEDGEMENTS

First and foremost, I am eternally thankful to Allah for His blessings, strength and perseverance bestowed on me, enabling me to complete this thesis.

I would like to take this opportunity to thank my supervisor, Associate Prof. Dr. Razali Yaakob for his support, guidance, and understanding. Special appreciation goes to him for his mentorship and constant support throughout this research. His comments and suggestions for further development as well as his assistance during writing this thesis are invaluable to me. His patience, humility, tutorship, interest, teaching and research style have provided for me an exceptional opportunity to learn and become a better researcher.

I would also like to thank the committee members, Professor Dr. Hamidah Ibrahim and Associate Professor Dr. Norwati Mustapha for their help and valuable suggestions.

Above all I am very grateful to my father and mother for their encouragement, my loving and caring family for their support and love in my efforts to successfully complete this work, especially during my severe sickness. For those others who have either directly or indirectly helped me in carrying out my work, I thank you all. Lastly, my heartfelt thanks to the Universiti Putra Malaysia and kind Malaysia for their support, and make me feel at home during the entire PhD journey.

Sarmad Makki, August 2019

Declaration by graduate student

I hereby confirm that:

- this thesis is my original work;
- quotations, illustrations and citations have been duly referenced;
- this thesis has not been submitted previously or concurrently for any other degree at any institutions;
- intellectual property from the thesis and copyright of thesis are fully-owned by Universiti Putra Malaysia, as according to the Universiti Putra Malaysia (Research) Rules 2012;
- written permission must be obtained from supervisor and the office of Deputy Vice-Chancellor (Research and innovation) before thesis is published (in the form of written, printed or in electronic form) including books, journals, modules, proceedings, popular writings, seminar papers, manuscripts, posters, reports, lecture notes, learning modules or any other materials as stated in the Universiti Putra Malaysia (Research) Rules 2012;
- there is no plagiarism or data falsification/fabrication in the thesis, and scholarly integrity is upheld as according to the Universiti Putra Malaysia (Graduate Studies) Rules 2003 (Revision 2012-2013) and the Universiti Putra Malaysia (Research) Rules 2012. The thesis has undergone plagiarism detection software

Signature: _____ Date: _____

Name and Matric No: Sarmad Makki Mohammed Garib, GS27325

Declaration by Members of Supervisory Committee

This is to confirm that:

- the research conducted and the writing of this thesis was under our supervision;
- supervision responsibilities as stated in the Universiti Putra Malaysia (Graduate Studies) Rules 2003 (Revision 2012-2013) were adhered to.

Signature: _____
Name of Chairman of Supervisory Committee: Associate Professor Dr. Razali Yaakob

Signature: _____
Name of Member of Supervisory Committee: Professor Dr. Hamidah Ibrahim

Signature: _____
Name of Member of Supervisory Committee: Associate Professor Dr. Norwati Mustapha

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

Term	Description
PAM	Partition Around Medoid
GA	Genetic Algorithm
CVI	Cluster Validity Index
GPAMC	Genetic Partitioning Around Medoid Clustering
EGPAMC	Efficient Genetic Partitioning Around Medoid Clustering
DBI	Davies-Bouldin Index
CH	Calinski-Harabasz index
SSE	the Sum of Squared Error
WSS	Within Sum Square
BSS	Between Sum Square
ARI	Adjusted Rand Index
CVI	Cluster Validity Index
TP	True Positive
FP	False Positive
FN	False Negative
TN	True Negative
P	Precision of clustering
R	Recall of clustering
F	F_1-score of clustering
CM	Confusion Matrix

Notation	Description
$S(N,k)$	Stirling number
$B(N,k)$	Bell number
D	Data set
d_i	An item i in the dataset
C_i	The i^{th} class label
P_j	The j^{th} Partition or cluster
X_j	The j^{th} Center
n_j	Number of items in the j^{th} cluster
N	Number of data items in the dataset
K	The known number of clusters in the dataset
M	Number of attributes in the data set
d_{ih}	The h^{th} attribute of the i^{th}data item

CHAPTER 1

INTRODUCTION

1.1 Overview

Clustering is one of the popular tasks in Data Mining and Knowledge Discovery. It aims to partition the data items into groups or clusters. The resulted groups should possess two properties: (1) homogeneity within the group; i.e. the data items that belongs to the same group should be as similar as possible in some term of proximity measures and (2) heterogeneity among the groups; i.e. the data items belongs to different groups should be as different as possible in some other term of proximity. In some traditional fields such as pattern recognition and machine learning, clustering might referred to as "unsupervised classification", as it does not use external information to aid in the partitioning process.

Genetic Algorithms (*GA*) are powerful randomized search and optimization techniques guided by the principles of evolution and natural selection. Genetic algorithms can be modeled to solve various optimization problems; one of such is the clustering problems. In lectures, several genetic based algorithms were used for clustering in diverse application domains such as those in the domain of *Data Clustering*, where the *M* dimensional data partitioned in the absence of specific labeling information (Anusha & Sathiaseelan, 2014; Beg & Islam, 2015, 2016c; Gwal & Choudhary, 2015; Patil, Thakare, & Dhote, 2015; Pizzuti & Procopio, 2017; Rahman & Islam, 2014; Terence & Singh, 2015). Another domain that applied the genetic based algorithms is the *Document Clustering*, where each preprocessed and unique team represented by an axis in the *M* dimensional terms space (Abualigah, Khader, & Al-betar, 2016; Ben et al., 2016; L. M. Carlantonio, Maria, & Costa, 2009; Jian-xiang, Huai, Yue-hong, & Xin-ning, 2009; Shi et al., 2011). *Satellite Image Segmentation* is another typical example of genetic based clustering (Bandyopadhyay & Maulik, 2002; Pare, Bhandari, Kumar, Singh, & Khare, 2015). Medical image such as Magnetic Resonance Imaging (*MRI*), Ultrasound, Computed Axial Tomography (*CAT*) and *X*-Ray are often stored in archiving systems. Researches conducted to mine information from these images including the genetic *based Medical Image Segmentation* (Halder, Pradhan, Dutta, & Bhattacharya, 2016; J. Wang, Zhang, & Li, 2015). It is to be noted that genetic clustering had been applied to several other domains for the clustering task, the above mentioned domain are some of these examples of the applications in the partitional clustering domains.

1.2 Problem Statement

Because of its simplicity and scalability, *k*-means algorithm ranked as one of the top ten most efficient data mining algorithms (Wu et al., 2007). However, this deterministic iterative algorithm has several drawbacks. Moreover, the GA based on the *k*-means either inherited part of its drawbacks or in turn produce some deficiency

by itself. This study addresses the drawbacks of the *k*-means algorithms and the improper GA modeling to the clustering problem as follows:

Mean calculation considers all the data items in the cluster, including outliers and unevenly distributed data items. Thus, *k*-means based algorithms are sensitive to outliers due to this 'compromise' centroid-selection which in turn increase the possibilities to converge to local optima as in the previous studies of (Anusha & Sathiaseelan, 2014; Sunanda Das & Chaudhuri, 2016;Pizzuti & Procopio, 2017). Moreover, the previous genetic-based clustering algorithms assumed fixed *N* length integer chromosomes encoding schema, where *N* is the number of data items in the dataset (Abualigah et al., 2016; Sunanda Das & Chaudhuri, 2016; Najeeb et al., 2016; Pizzuti & Procopio, 2017). This representation schema is lengthy, generates redundant chromosomes and uses the standard genetic operators (Swagatam Das et al., 2009; Falkenauer, 1998). Additionally, the previous genetic-based clustering algorithms either used random or non-mathematically proved initializations methods. A Poor choice of initial population may lead to a local optimum or immature convergence.

Other effective centroid-selection methods might be used such as the *k*-medoid (Han, 2008; Rui Xu & Wunsch, 2009). But, the complexity of the *k*-medoid based algorithms in general is more than the complexity of the *k*-means based algorithms (D. Xu & Tian, 2015). It is known that the centroid-based algorithms repeatedly calculate the proximity (distance/similarity) measures among each data items and all centers. Moreover, in the genetic based algorithms, these calculations are rapidly increasing with inter- and intra-clustering measurement of the fitness function.

Lastly, to externally judge the validity of the clustering algorithm there is a need to compare the resulted variant multiclass clusters at each run with the actual ground truth partitioning. The previous clustering algorithms either evaluate their results with internal validity measures or with the misleading external 'accuracy' measure. Some algorithms evaluate its results with the deterministic Adjusted Rand Index (*ARI*) which works on the matrix of the data itself and took a large amount of computation dataset (Abualigah et al., 2016; Sunanda Das & Chaudhuri, 2016; Najeeb et al., 2016; Pizzuti & Procopio, 2017).

1.3 Research Objectives

The main objective of this research is to develop a genetic-based clustering algorithm that can efficiently and appropriately finds the right clustering for *N* data items, along with evolving to the proper number of clusters *K*. The specific objective of this research is as follows:

i. To propose a partitional genetic-based algorithm with automatic number of clusters in which it is effectively able to cluster datasets without sensitivity to outliers or unevenly distributed clusters.
ii. To propose an efficient method that reduces the number of computation during both assigning data items to medoids phase and fitness estimation phase.
iii. To propose an algorithm that is capable of facilitates and accurately estimates the external validity indexes for the clustering of variant multiclass clusters given the ground truth actual clusters.

1.4 Research Scope

The scope of this research study is presented in the following points:

- This research study used the Genetic Algorithm among the other evolutionary based algorithms as it is the most dominant model among the others (Beg & Islam, 2016a, 2016b; Ben et al., 2016; L. Carlantonio & Costa, 2009; Choi, Lee, & Park, 2011; Sunanda Das & Chaudhuri, 2016; Hruschka, Campello, Freitas, & De Carvalho, 2009; Jian-xiang et al., 2009; Lavangnananda & Poolphol, 2014; Liu, Wu, & Shen, 2011; Najeeb et al., 2016; Premalatha & Natarajan, 2009; Sheikh, Raghuwanshi, & Jaiswal, 2008; Shi et al., 2011; Song & Park, 2006; Suhaimi & Kamaliah, 2015; Verma, Kandpal, Pandey, & Dhar, 2010; Wei, Liu, Sun, & Su, 2009; Zhengyu, Ping, Chunlei, & Lipei, 2010).
- The type of clustering which is considered in this research is the partitional clustering algorithms and limited to *k*-means and partition around medoid algorithms as it is the most frequently applied algorithms (Anusha & Sathiaseelan, 2014; Beg & Islam, 2015, 2016c; Feng & Wang, 2011; He & Tan, 2012; Kumar, Ranjan, & Dhar, 2012; Lavangnananda & Poolphol, 2014; Pizzuti & Procopio, 2017; Rahman & Islam, 2014; Jianxin Wang, Zhang, Dong, Xu, & Mei, 2010).
- This research study concentrates on two types of crisp datasets, namely: synthetic and real datasets. The synthetic dataset including are *Fixed15*, *Scattered*, *Unbalanced*, *Leicester*, *AD_3_2*, *AD_5_2*, *Syn3*, *Syn4* and *Syn6* (Bandyopadhyay & Maulik, 2002); Pizzuti & Procopio, 2017). While the real datasets include *Iris*, *Cancer*, *Glass* and *Ecoli* (Pizzuti & Procopio, 2017). These are the most popular types of datasets that have been used in this area (Anusha & Sathiaseelan, 2014; Bandyopadhyay & Maulik, 2002; Choi et al., 2011; Feng & Wang, 2011; Jagannath & Panda, 2014; José-garcía & Gómez-flores, 2016; Patil et al., 2015; Pizzuti & Procopio, 2017; Song & Park, 2006).
- This research work assumes that the outlier data items or unevenly distributed clusters might be occurred in one or more datasets.

1.5 Contributions of the Study

The main contributions of this thesis can be explained as follows:

i. A partitional genetic based algorithm with automatic number of clusters is proposed. The algorithm is able to evolve a population of chromosomes, each representing a clustering solution without the need to preset the number of clusters in advance. To this extent, each individual chromosome is represented with unique and compact variable-length encoding schema. Accordingly, a novel genetic group-based crossover operator is proposed to achieve population diversity and to avoid generating twin genes which lead to improper clustering. Moreover, a novel mutation operator is proposed to achieve the ability to converge to the proper number of clusters. To speed up the convergence and to get global optima efficient results, a population initialization approach is proposed based on the *k*-means++ algorithm.
ii. An efficient version of the algorithm is proposed with suitable indexing method. This algorithm aims at reducing the complexity of computation in both assigning data items to clusters and genetic fitness estimation phases.
iii. An algorithm is proposed to evaluate the results of these types of clustering algorithms that generates variable and multiple numbers of clusters; given the ground truth actual class labels.

1.6 Organization of the Thesis

This thesis is organized as follows:

Chapter 1 is an introductory chapter that discusses the problem statement, the objective, the scope and the contributions of the research.

Chapter 2 is the literature review chapter that explains the previous existing concepts and algorithms related to the clustering and genetic algorithms. It also reviews related works proposed by previous researchers of genetic clustering specifically. The chapter presents the features of these algorithms with their pros and cons.

Chapter 3 presents the detail description of the proposed approaches for clustering using genetic algorithms. This chapter explains and discusses the different phases of this research work and the methodology followed during each phase with running examples. It's Also presents the proposed approach for externally evaluating the clustering process.

Chapter 4 presents the results of the experiments conducted to evaluate the quality and performance of the proposed approaches compared with the most relevant existing approaches. Along with the result of comparison with other algorithms, this chapter

discusses in details the effects of different techniques affect on the proposed algorithms.

Chapter 5 reflects the conclusions and the contributions of this research. Besides, the recommendations of the future works are presented in this chapter.

REFERENCES

Abualigah, L. M., Khader, A. T., & Al-betar, M. A. (2016). Unsupervised Feature Selection Technique Based on Genetic Algorithm for Improving the Text Clustering. *IEEE*, 3-8.

Aggarwal, C., & Zhai, C. (2012). A Survey of Text Clustering Algorithms. In C. C. Aggarwal & C. Zhai (Eds.), *Mining Text Data* (Vol. 1, pp. 77-128). US: Springer.

Aggarwal, C. C., Hinneburg, A., & Keim, D. A. (2001). *On the surprising behavior of distance metrics in high dimensional spaces.* Paper presented at the ICDT.

Aggarwal, C. C., & Reddy, C. K. (2014). *Data Clustering: Algorithms and Applications*: Chapman \& Hall/CRC.

Anusha, M., & Sathiaseelan, J. G. R. (2014). An Enhanced *K*-Means Genetic Algorithms for Optimal Clustering. *IEEE*.

Arbelaitz, O., Gurrutxaga, I., Muguerza, J., Jes, #250, P, S. M., . . . Perona, I. (2013). An extensive comparative study of cluster validity indices. *Pattern Recogn., 46*(1), 243-256. doi: 10.1016/j.patcog.2012.07.021

Arthur, D., & Vassilvitskii, S. (2007). *k-means++: the advantages of careful seeding*. Paper presented at the Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms, New Orleans, Louisiana.

Bandyopadhyay, S., & Maulik, U. (2001). NonParametric Genetic Clustering: Comparison of Validity Indices. *IEEE Transaction on systems, Man, and Cybernetics, 31*(1), 62-66.

Bandyopadhyay, S., & Maulik, U. (2002). Genetic clustering for automatic evolution of clusters and application to image classfication. *Patten Recognition, 35*, 1197-1208.

Bandyopadhyay, S., & Pal, S. K. (2007). *Classification and learning using genetic algorithms: applications in bioinformatics and web intelligence*: Springer Science & Business Media.

Beg, A. H., & Islam, M. Z. (2015). Clustering by Genetic Algorithm- High Quality Chromosome Selection for Initial Population. *IEEE*, 129-134.

Beg, A. H., & Islam, M. Z. (2016a). Advantages and Limitations of Genetic Algorithms for Clustering Records. *IEEE*, 2478-2483.

Beg, A. H., & Islam, M. Z. (2016b). Branches of Evolutionary Algorithms and their Effectiveness for Clustering Records. *IEEE*, 2484-2489.

Beg, A. H., & Islam, M. Z. (2016c). Novel Crossover and Mutation Operation in Genetic Algorithm for Clustering. *IEEE*, 2114-2121.

Ben, W., Karaa, A., Ashour, A. S., Sassi, D. B., Roy, P., Kausar, N., & Dey, N. (2016). MEDLINE Text Mining : An Enhancement Genetic Algorithm Based Approach for Document Clustering. 267-287. doi: 10.1007/978-3-319-21212-8

Berkhin, P. (2006). A Survey of Clustering Data Mining Techniques *Grouping Multidimensional Data - Recent Advances in Clustering* (pp. 25-71).

Bezdek, J. C., & Pal, N. R. (1998). Some new indexes of cluster validity. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 28*(3), 301-315.

Breaban, M. E. (2011). *Clustering: evolutionary approaches.*

C.A. Murthy, & Das, S. (2014). *K*-Medoids and DBSCAN. [Chivukula Anjaneya Murthy].

Caliński, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics, 3*(1), 1-27. doi: 10.1080/03610927408827101

Carlantonio, L., & Costa, R. M. (2009). Exploring a Genetic Algorithm for Hypertext Documents Clustering. In N. Nedjah, L. Macedo Mourelle, J. Kacprzyk, F. G. França & A. De Souza (Eds.), *Intelligent Text Categorization and Clustering* (Vol. 164, pp. 95-117). Berlin Heidelberg: Springer

Carlantonio, L. M., Maria, R., & Costa, E. M. (2009). Exploring a Genetic Algorithm for Hypertext Documents Clustering. *Itelligent Text Categorization and Clustering*, 95-117.

Cha, S.-H. (2007). Comprehensive survey on distance/similarity measures between probability density functions. *City, 1*(2), 1.

Choi, L. C., Lee, J. S., & Park, S. C. (2011). Double Layered Genetic Algorithm for Document Clustering. In T.-h. Kim, H. Adeli, H.-k. Kim, H.-j. Kang, K. Kim, A. Kiumi & B.-H. Kang (Eds.), *Software Engineering, Business Continuity and Education* (Vol. 257, pp. 212-218): Springer Berlin Heidelberg.

Cowgill, M. C., Harvey, R. J., & Watson, L. T. (1999). A genetic algorithm approach to cluster analysis. *Computers & Mathematics with Applications, 37*(7), 99-108.

Das, S., Abraham, A., & Konar, A. (2009). Metaheuristic Pattern Clustering–An Overview. In S. Das, A. Abraham & A. Konar (Eds.), *Metaheuristic Clustering* (pp. 1-62). Berlin Heidelberg: Springer.

Das, S., & Chaudhuri, S. (2016). Cluster Analysis for overlapping Clusters using Genetic Algorithm. *IEEE*, 6-11.

Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*(2), 224-227.

Desgraupes, B. (2013). Clustering indices.

Deza, M. M., & Deza, E. (2009). Encyclopedia of distances *Encyclopedia of Distances* (pp. 1-583): Springer.

Duda, R. O., Hart, P. E., & Stork, D. G. (2000). *Pattern Classification (2nd Edition)*: Wiley-Interscience.

Dunn, J. C. (1973). A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters.

Elavarasi, S. A., Akilandeswari, J., & Sathiyabhama, B. (2011). A survey on partition clustering algorithms. *International Journal of Enterprise Computing and Business Systems, 1*(1).

Falkenauer, E. (1998). *Genetic Algorithms and Grouping Problems*: Wiley.

Feng, M., & Wang, Z. (2011). A Genetic *K*-means Clustering Algorithm Based on the Optimized Initial Centers. *Computer and Information Science, 4*(3), 88-94. doi: 10.5539/cis.v4n3p88

Gan, G., Ma, C., & Wu, J. (2007). *Data clustering: theory, algorithms, and applications*: SIAM.

Ganesan, K. (2014). Computing Precision and Recall for Multi-Class Classification Problems, from http://text-analytics101.rxnlp.com/2014/10/computing-precision-and-recall-for.html

Ghosh, J., & Strehl, A. (2006). Similarity-Based Text Clustering: A Comparative Study: Springer.

Govardhan, P., Wagh, K., & Chatur, P. (2013). Survey on Similarity Measure for Clustering. *International Journal, 3*(12).

Gupta, D., & Ghafir, S. (2012). *An Overview of methods maintaining Diversity in Genetic Algorithms* (Vol. 2).

Gwal, N., & Choudhary, T. (2015). A High Dimensional Clustering Scheme for Data Classification. *Int. Journal of Engineering research and Applications, 5*(9), 101-106.

Halder, A., Pradhan, A., Dutta, S. K., & Bhattacharya, P. (2016, 6-8 April 2016). *Tumor extraction from MRI images using dynamic genetic algorithm based image segmentation and morphological operation.* Paper presented at the 2016 International Conference on Communication and Signal Processing (ICCSP).

Halkidi, M., Batistakis, Y., & Vazirgiannis, M. (2001). On Clustering Validation Techniques. *Journal of Intelligent Information Systems, 17*(2-3), 107-145.

Han, J. (2008). *Data Mining: Concepts and Techniques*: Morgan Kaufmann Publishers Inc.

He, H., & Tan, Y. (2012). Neurocomputing A two-stage genetic algorithm for automatic clustering. *Neurocomputing, 81*, 49-59. doi: 10.1016/j.neucom.2011.11.001

Hruschka, E. R., Campello, R. J. G. B., Freitas, A. A., & De Carvalho, A. P. L. F. (2009). A Survey of Evolutionary Algorithms for Clustering. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, 39*(2), 133-155.

Huang, A. (2008). *Similarity Measures for Text Document Clustering.* Paper presented at the Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand.

Jagannath, S., & Panda, G. (2014). A survey on nature inspired metaheuristic algorithms for partitional clustering. *Swarm and Evolutionary Computation, 16*, 1-18. doi: 10.1016/j.swevo.2013.11.003

Jain, A. K., & Dubes, R. C. (1988). Algorithms for Clustering Data.

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data Clustering: A Review. *ACM computing surveys (CSUR), 31*(3), 264-323.

Jian-xiang, W., Huai, L., Yue-hong, S., & Xin-ning, S. (2009). Application of Genetic Algorithm in Document Clustering. *IEEE, 2*(40771163), 146-149. doi: 10.1109/itcs.2009.269

Jones, W. P., & Furnas, G. W. (1987). Pictures of relevance: A geometric analysis of similarity measures. *Journal of the American Society for Information Science, 38*(6), 420.

José-garcía, A., & Gómez-flores, W. (2016). Automatic clustering using nature-inspired metaheuristics : A survey. *Applied Soft Computing Journal, 41*, 192-213. doi: 10.1016/j.asoc.2015.12.001

Kaur, S., & Kaur, U. (2013). A survey on various clustering techniques with *k*-means clustering algorithm in detail. *Int. J. Comput. Sci. Mob. Comput, 2*, 155-159.

Krishna, K., & Murty, M. N. (1999). Genetic *K*-Means Algorithms. *IEEE Transactions on System, Man, And Cybernetics, 29*(3), 433-439.

Kumar, R., Ranjan, A., & Dhar, J. (2012). A Fast and Effective Partitioning Algorithm. 264-271.

Lavangnananda, K., & Poolphol, R. (2014). A Genetic Algorithm Approach to Partitioning Clustering : A case study on M . Sc . applicants. *13th International Conference on Machine Learning and Applications*. doi: 10.1109/icmla.2014.93

Lawrence Hubert, P. A. (1985). Comparing partitions. *Journal of Classification, 2*(1), 193.

Lin, Y.-S., Jiang, J.-Y., & Shie-Jue, L. (2013). A Similarity Measure for Text Classification and Clustering. *IEEE Transactions on Knowledge and Data Engineering, 99*(PrePrints), 1-1.

Liping, J. (2005). Survey of Text Clustering (pp. 7695-1754). Hong Kong, China: Department of Mathematics, The University of Hong Kong.

Liu, Y., Li, Z., Xiong, H., Gao, X., & Wu, J. (2010). *Understanding of internal clustering validation measures.* Paper presented at the Data Mining (ICDM), 2010 IEEE 10th International Conference on.

Liu, Y., Li, Z., Xiong, H., Gao, X., & Wu, J. (2010, 13-17 Dec. 2010). *Understanding of Internal Clustering Validation Measures.* Paper presented at the 2010 IEEE International Conference on Data Mining.

Liu, Y., Wu, X., & Shen, Y. (2011). Automatic clustering using genetic algorithms. *Applied Mathematics and Computation, 218*(4), 1267-1279. doi: 10.1016/j.amc.2011.06.007

Lu, Y., Lu, S., Fotouhi, F., Deng, Y., & Brown, S. J. (2004). FGKA : A Fast Genetic *K*-means Clustering Algorithm. *ACM Symposium on Applied Computing*, 622-623.

Lucasius, C. B., & Kateman, G. (1993). Understanding and using genetic algorithms Part 1. Concepts, properties and context. *Chemometrics and Intelligent Laboratory Systems, 19*(1), 1-33. doi: https://doi.org/10.1016/0169-7439(93)80079-W

Makki , S., Yaakob, R., Mustapha, N., & Ibrahim, H. (2015). Advances in Document Clustering with Evolutionary-Based Algorithms. *American Journal of Applied Sciences - Science Publications, 12*(10), 689-708. doi: 10.3844/ajassp.2015.689.708

Manning, C. D., Raghavan, P., & Schütze, H. (2009). *Introduction to Information Retrieval*: Cambridge University Press.

Matej, #268, repin, #353, ek, Liu, S.-H., & Mernik, M. (2013). Exploration and exploitation in evolutionary algorithms: A survey. *ACM Comput. Surv., 45*(3), 1-33. doi: 10.1145/2480741.2480752

Maulik, U., & Bandyopadhyay, S. (2000). Genetic Algorithm-Based Clustering Technique. *Pattern recognition, 33*(9), 1455-1465.

Michalewicz, Z. (1996). *Genetic algorithms + data structures = evolution programs (3rd ed.)*: Springer-Verlag.

Milligan, G. W., & Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika, 50*(2), 159-179. doi: 10.1007/bf02294245

Mirkes, E. (2016). Kmeans and Kmedoids, from https://github.com/Mirkes/Data_Mining_Softbook/wiki/*k*-means-and-*k*-medoids

Najeeb, A. R., Aibinu, A. M., Nwohu, M. N., Salami, M. J. E., & Salau, H. B. (2016). Performance Analysis of Clustering Based Genetic Algorithm. *IEEE*. doi: 10.1109/iccce.2016.76

Pare, S., Bhandari, A. K., Kumar, A., Singh, G. K., & Khare, S. (2015, 21-24 July 2015). *Satellite image segmentation based on different objective functions using genetic algorithm: A comparative study*. Paper presented at the 2015 IEEE International Conference on Digital Signal Processing (DSP).

Patil, S. P., Thakare, A. D., & Dhote, C. A. (2015). An efficient hybrid data clustering method based on Candidate Group Search and Genetic Algorithm. *IEEE*.

Peña, J. M., Lozano, J. A., & Larrañaga, P. (1999). An empirical comparison of four initialization methods for the *K*-Means algorithm. *Pattern Recognition Letters, 20*(10), 1027-1040. doi: https://doi.org/10.1016/S0167-8655(99)00069-0

Peng, P., Addam, O., Elzohbi, M., Özyer, S. T., Elhajj, A., Gao, S., . . . Alhajj, R. (2014). Reporting and analyzing alternative clustering solutions by employing multi-objective genetic algorithm and conducting experiments on cancer data. *KNOWLEDGE-BASED SYSTEMS, 56*, 108-122. doi: https://doi.org/10.1016/j.knosys.2013.11.003

Pizzuti, C., & Procopio, N. (2017). A *K*-means Based Genetic Algorithm for Data Clustering A *K*-means based Genetic Algorithm for Data Clustering. (October). doi: 10.1007/978-3-319-47364-2

Premalatha, K., & Natarajan, A. M. (2009). Genetic Algorithm for Document Clustering with Simultaneous and Ranked Mutation. *Modern Applied Science, 3*(2), 75-82.

Rahman, A., & Islam, Z. (2014). A Hybrid Clustering Technique Combining a Novel Genetic Algorithm with *K*-Means. *KNOWLEDGE-BASED SYSTEMS*(August). doi: 10.1016/j.knosys.2014.08.011

Rendon, E., Abundez, I. M., Gutierrez, C., Zagal, S. D., Arizmendi, A., Quiroz, E. M., & Arzate, H. E. (2011). *A Comparison of Internal and External Cluster Validation Indexes*. Paper presented at the Proceedings of the 2011 American conference on applied mathematics and the 5th WSEAS international conference on Computer engineering and applications, Puerto Morelos, Mexico.

Rokach, L. (2010). A survey of Clustering Algorithms. In O. Maimon & L. Rokach (Eds.), *Data Mining and Knowledge Discovery Handbook* (pp. 269-298). Boston, MA: Springer US.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics, 20*, 53-65. doi: https://doi.org/10.1016/0377-0427(87)90125-7

Sadjadi, F. (2004). *Comparison of fitness scaling functions in genetic algorithms with applications to optical processing* (Vol. 5557): SPIE.

Saitta, S., Raphael, B., & Smith, I. F. C. (2007). A Bounded Index for Cluster Validity. In P. Perner (Ed.), *Machine Learning and Data Mining in Pattern Recognition: 5th International Conference, MLDM 2007, Leipzig, Germany, July 18-20, 2007. Proceedings* (pp. 174-187). Berlin, Heidelberg: Springer Berlin Heidelberg.

Saitta, S., Raphael, B., & Smith, I. F. C. (2008). A comprehensive validity index for clustering. *Intell. Data Anal., 12*(6), 529-548.

Sathiyakumari, K., Preamsudha, V., Manimekalai, G., & Scholar, M. P. (2011). A Survey on Various Approaches in Document Clustering. *International Journal of Computer Technology, 2*(5), 1534-1539.

Saxena, A., Prasad, M., Gupta, A., Bharill, N., Patel, O. P., Tiwari, A., . . . Lin, C.-T. (2017). A review of clustering techniques and developments. *Neurocomputing, 267*, 664-681. doi: http://dx.doi.org/10.1016/j.neucom.2017.06.053

Sheikh, R. H., Raghuwanshi, M. M., & Jaiswal, A. N. (2008). *Genetic Algorithm Based Clustering: A Survey*. Paper presented at the First International Conference on Emerging Trends in Engineering and Technology.

Shi, K., Li, L., He, J., Zhang, N., Liu, H., & Song, W. (2011). Improved Genetic Algorithm-Based Text Clustering Algorithm *Proceeding of IEEE IC-BNMT2011*, 1-5.

Sivanandam, S. N., & Deepa, S. N. (2007). *Introduction to Genetic Algorithms*: Springer Publishing Company, Incorporated.

Song, W., & Park, S. C. (2006). Genetic Algorithm-based Text Clustering Technique : Automatic Evolution of Clusters with High Efficiency. *Procedding of the Seventh Inernational Conference on Web-Age Information Managment Workshops*.

Strehl, A., Ghosh, J., & Mooney, R. (2000). *Impact of similarity measures on web-page clustering.* Paper presented at the Workshop on artificial intelligence for web search (AAAI 2000).

Suhaimi, N. S., & Kamaliah, S. N. (2015). Optimizing Cluster of Questions by Using Dynamic Mutation in Genetic Algorithm. *3rd International Conference on Atrificial Intelligence, Modelling and Simulation*, 15-18. doi: 10.1109/aims.2015.81

Tan, P.-N., Steinbach, M., & Kumar, V. (2005). *Introduction to Data Mining, (First Edition)*: Addison-Wesley Longman Publishing Co., Inc.

Tan, P.-N., Steinbach, M., & Kumar, V. (2006). *Introduction to Data Mining, (First Edition)*: Addison-Wesley Longman Publishing Co., Inc.

Taya, A. (2015). Performance Improvement of Genetic Algorithm by Fitness Scaling

and Diversity Maintenance. *International Journal for Research in Technological Studies, 2*(7).

Terence, J., & Singh, S. K. (2015). Genetic Algorithms based Enhanced K Strange Points Clustering Algorithm. *Intl. conference on Computing and Network Communications*, 737-741.

Verma, H., Kandpal, E., Pandey, B., & Dhar, J. (2010). A Novel Document Clustering Algorithm Using Squared Distance Optimization Through Genetic Algorithms. *International Journal on Computer Science and Engineering, 02*(05), 1875-1879.

Wagner, S., & Wagner, D. (2007). *Comparing clusterings: an overview*: Universität Karlsruhe, Fakultät für Informatik Karlsruhe.

Wang, J., Zhang, F., & Li, P. (2015, 14-16 Oct. 2015). *Medical image segmentation based on 2D maximum fuzzy entropy and improved genetic algorithm.* Paper presented at the 2015 8th International Congress on Image and Signal Processing (CISP).

Wang, J., Zhang, H., Dong, X., Xu, B., & Mei, B. (2010). An Effective Hybrid Crossover Operator for Genetic Algorithms to Solve *K*-means Clustering Problem. *IEEE*(Icnc), 2271-2275.

Wei, J.-X., Liu, H., Sun, Y.-h., & Su, X.-N. (2009, 25-26 July 2009). *Application of Genetic Algorithm in Document Clustering.* Paper presented at the International Conference on Information Technology and Computer Science (ITCS).

Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., . . . Steinberg, D. (2007). Top 10 algorithms in data mining. *Knowl. Inf. Syst., 14*(1), 1-37. doi: 10.1007/s10115-007-0114-2

Xu, D., & Tian, Y. (2015). A comprehensive survey of clustering algorithms. *Annals of Data Science, 2*(2), 165-193.

Xu, R., & Wunsch, D. (2009). *Clustering*: Wiley-IEEE Press.

Xu, R., & Wunsch, D. C. (2010). Clustering Algorithms in Biomedical Research: A Review. *IEEE Reviews in Biomedical Engineering, 3*, 120-154. doi: 10.1109/rbme.2010.2083647

Zezula, P., Amato, G., Dohnal, V., & Batko, M. (2010). *Similarity Search: The Metric Space Approach*: Springer Publishing Company, Incorporated.

Zhao, Y., & Karypis, G. (2004). Empirical and Theoretical Comparisons of Selected Criterion Functions for Document Clustering. *Machine Learning, 55*(3), 311-331. doi: 10.1023/B:MACH.0000027785.44527.d6

Zhengyu, Z., Ping, H., Chunlei, Y., & Lipei, L. (2010). *A Dynamic Genetic Algorithm for Clustering Web Pages.* Paper presented at the 2nd International Conference on Software Engineering and Data Mining (SEDM).