A Multimodal Adaptive Genetic Clustering Algorithm

Sawsan Al-Malak
Computer Science Department, College of Computer & Information Sciences, King Saud University
Riyadh, Saudi Arabia
sawsan.almalak@gmail.com

Manar Hosny
Computer Science Department, College of Computer & Information Sciences, King Saud University
Riyadh, Saudi Arabia
mifawzi@ksu.edu.sa

ABSTRACT
Clustering is widely used in a variety of fields to find structures among data and extract useful knowledge. Recently, there has been an emergent need for robust and efficient techniques that can manage the exploding volume of data available in the World Wide Web or gathered from devices and sensors. However, clustering such data is challenging, due to the multimodal nature of this information. In this work, we introduce a novel Multimodal Adaptive Genetic Clustering (MAGC) algorithm that clusters information based on multiple features. Our approach adds feature weights as an extension to the chromosome, which represents a clustering solution, such that feature weights are also evolved and optimized alongside the original clustering solution. The number of clusters is also adaptive and is optimized during the search.

Keywords
Clustering; Genetic Algorithms; Multimodal Data.

1. INTRODUCTION
Due to the exploding amount of data currently available on the World Wide Web or gathered from devices and sensors, there has been an emergent need for robust and efficient techniques that can manage such data and extract useful information from it. Data clustering is defined as finding a natural grouping among data based on their features and/or properties. Clustering is applied in a variety of fields [1]. It also has many applications (e.g. indexing and data retrieval). However, due to the continuous increase in the complexity of data, attention is currently shifting towards multi-multimodal clustering [8]. For example, social media data is complex in structure and can be represented through multiple features (e.g. textual tags, geolocation, visual characteristics, etc.); it is *multimodal* in nature. Exploiting the multimodal nature of data is rational in this context, though. Specifically, since some features may be more important than others we can assign different weights to different features associated with the data, during the clustering process. A common approach to cluster data is to use each modality independently to create different clusters. This approach may overlook the relationships between modalities and may reduce the clustering performance. Multimodal clustering has been proven to perform better, since each modality contributes to forming a global meaning when combined with other modalities [9].

The clustering problem is considered an NP-hard grouping problem [5]. Metaheuristic algorithms in general are efficient in solving such complex problems, where they usually provide near-optimal solutions. The process of assigning weights to features using metaheuristic optimization techniques has been approached in some previous work. For example, the authors in [8,10] adopt an Ant Colony Optimization (ACO) technique in clustering a set of social media images. ACO is used to optimize the combination of features and their weights that result in the best objective value and handle large datasets in high-dimensional feature space. In this work, we will introduce a Multimodal Adaptive Genetic Clustering (MAGC) algorithm that clusters information based on multiple features. Our approach adds feature weights as an extension to the clustering solution, such that feature weights are also evolved and optimized alongside the original clustering solution. Optimizing feature weights aims to develop a clustering solution in which the most important features are targeted by adaptively adjusting their weights during the evolutionary process [6]. To the best of our knowledge, no work has used an adaptive Genetic Clustering Algorithm (GCA) with assigned feature weights that automatically optimize the discovery of semantically-related data clusters.

GECCO'16 Companion, July 20-24, 2016, Denver, CO, USA
ACM 978-1-4503-4323-7/16/07.
http://dx.doi.org/10.1145/2908961.2931633

2. METHODOLOGY
Our approach is an adaptive GCA where the number of clusters and feature weights are not previously determined. Rather, they are optimized together with the clustering solution.

2.1 Solution Representation
To represent a clustering solution, we use the label-based representation that represents the partitioning solution with integer encoding, and clusters' centroids are actual data objects (medoids) [6]. For the initial population, a number of solutions are generated with k randomly selected objects from N data objects. The value k (the number of clusters) is randomly chosen within a range of $[k\_min\ , k\_max]$, where $k\_min \geq 2$ and $k\_max \approx \sqrt{(N/2)}$ as recommended in [7]; k is stored in the first gene of the clustering solution. Gene values of medoids are -1, and gene values of other data objects are the indices of their nearest medoid. Feature weights will be an extension to the end of the chromosome, where its length is equal to the features considered in the clustering process. Weights are real numbers that fall in the range [0,1]. In a single chromosome, weights should sum up to 1.

2.2 Objective Function
The objective function of our algorithm depends on the quality of the clustering solution. Our chosen cluster validity measure is the Davies-Bouldin (DB) index [3], since it is suitable for a clustering algorithm with a variable number of clusters. It has also been proven to be computationally efficient [9].

2.3 Genetic Operators
For the crossover operator, parents pass down some properties to the children. We propose a Join and Split (J&S) crossover. In this crossover, parents randomly pass down the number of medoids they carry to the children. For example, if two parents contain partitions of k_1 and k_2 clusters, then one child will randomly have

k_1 clusters, and the other will have k_2 clusters. The $(k_1 + k_2)$ cluster medoids are then randomly distributed between the two children, taking into consideration that duplicate medoids are not allowed in the same child. If the same medoid appears in both parents, it should appear in both children, which makes the J&S crossover heritable [6]. After that, both parents will be maintained by having their objects reallocated to the nearest clusters medoids. The weights part of the chromosome is passed randomly to the children, each part as a whole. We apply mutation in a cluster-oriented way, where we add or remove a cluster to the solution and redistribute data objects. For the feature weights part, we apply mutation through subtracting a small value ε from one weight and adding it to another weight. We use Roulette Wheel Selection [4] as a selection strategy. We have set the population size to 100, and the crossover and mutation probabilities to p_c = 0.8 and p_m = 0.2, respectively. The termination condition halts the evolution if no improvement in fitness is observed within 10 consecutive generations.

3. EXPERIMENT AND RESULTS

Our algorithm is tested on the largest Flickr metadata collection that is available for studies on scalable similarity search techniques, the CoPhIR dataset [2]. For the current experiment, a subset of the images' features in the dataset was selected manually. The experiment aims to evaluate the algorithm's ability to cluster data semantically, as well as to compare its performance to a non-adaptive genetic clustering algorithm, which is an identical version of our algorithm excluding the features weights adaptation. First, a number of non-overlapping collections were collected from the dataset with variable sizes. The algorithm uses each subset to generate clustering solutions that are to be evaluated using the DB-index. Table 1 summarizes the average results of 10 runs for each data subset. A smaller DB-Index indicates a better clustering solution.

Table 1. Experimental Results

	With Adaptive Weights				Without Adaptive Weights			
Dataset Size	**Best DB-Index**	**Avg. DB-Index**	**Avg. No. of Generations**	**Computation Time (sec.)**	**Best DB-Index**	**Avg. DB-Index**	**Avg. No. of Generations**	**Computation Time (sec.)**
100	1.04	1.50	35.0	1.32	1.22	1.79	18	1.04
300	1.46	1.85	24.9	2.40	1.53	1.94	20.4	2.06
500	1.40	1.76	40.1	5.04	1.56	1.97	21.4	3.70
700	1.50	1.77	47.8	13.78	1.61	1.95	20.5	9.35
1000	1.51	1.72	62.4	27.07	1.64	1.91	24	14.96
1500	1.44	1.64	73.6	56.65	1.63	1.92	29.4	25.65
Avg.	1.39	1.71	47.3	17.71	1.53	1.91	22.3	9.46

From the reported results, we can observe that the quality of the fittest individual is slightly degrading as the dataset size increases, which indicates that the clustering becomes more difficult when the collection size gets larger. However, the algorithm is still able to converge with larger number of generations and without compromising the quality of the best solution, when the collection size increases. Table 1 also shows that the version of the algorithm without weights adaptation has larger overall DB-index, which in our case is undesirable. On average, weights adaptation provides approximately 10% improvement in the best solution quality and 12% improvement on the average solution quality. Moreover, the adaptation of weights was observed and recorded in order to detect any possible patterns throughout the experiment. It was observed that the resulting optimized weights were not biased towards specific features in all subsets of the data, where each run produced different weights. Nevertheless, for the fittest individuals for a single subset, patterns were detected to identify the strongest features that were able to cluster the subset semantically. For example, the fittest solutions obtained for the dataset of size 500 had relatively high weights for both location and title features.

4. CONCLUSION

In this work, we propose a genetic clustering algorithm that exploits multimodalities in clustering in order to produce semantically-related clusters. Our algorithm optimizes both the number of clusters and feature weights in order to accomplish this objective. Our future work will include enhancing the cohesiveness of the resulting clusters, optimizing the computation time of the algorithm, and testing MAGC on larger data subsets to put its scalability under emphasis.

5. REFERENCES

[1] Berkhin, P. A survey of clustering data mining techniques. In *Grouping multidimensional data*. Springer Berlin Heidelberg, 2006, 25–71.

[2] Bolettieri, P., Esuli, A., Falchi, F., et al. CoPhIR: a test collection for content-based image retrieval. *abs/0905.4627*, (2009).

[3] Davies, D.L. and Bouldin, D.W. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence 1*, (1979), 224–227.

[4] De Jong, K.A. Analysis of the behavior of a class of genetic adaptive systems. *Doctoral dissertation, University of Michigan, Dissertation Abstracts International 36*, 10 (1975).

[5] Falkenauer, E. *Genetic algorithms and grouping problems*. John Wiley & Sons, Inc., New York, NY, USA, 1998.

[6] Hruschka, E.R., Campello, R.J., Freitas, A.A., and De Carvalho, A.C. A survey of evolutionary algorithms for clustering. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews 39*, 2 (2009), 133–155.

[7] Mardia, K.V., Kent, J.T., and Bibby, J.M. Multivariate analysis. *Academic Press*, (1979).

[8] Nikolopoulos, S., Giannakidou, E., Kompatsiaris, I., Patras, I., and Vakali, A. Combining Multi-modal Features for Social Media Analysis. In S.M. Modeling, S.C.H.H. Computing, J. Luo, et al., eds., *Social Media Modeling and Computing*. Springer London, 2011, 71–96.

[9] Petrovic, S. A comparison between the silhouette index and the davies-bouldin index in labelling ids clusters. *Proceedings of the 11th Nordic Workshop of Secure IT Systems*, (2006), 53–64.

[10] Piatrik, T. and Izquierdo, E. Subspace clustering of images using Ant colony Optimisation. *16th IEEE International Conference on Image Processing (ICIP), Cairo*, (2009), 229–232.