CSCI 447 Project 4: Swarm-based Clustering Fall 2015

Brandon Fenton
Department of Mathematical Sciences
Montana State University
Bozeman, MT 59717-2400
Email: brandon.fenton@gmail.com

John Sherrill
Department of Mathematical Sciences
Montana State University
Bozeman, MT 59717-2400
Email: prof.sherrill@gmail.com

***Abstract*—Five seperate clustering algorithms were implemented in the interest of comparing convergence and clustering performance across 10 seperate datasets. Four of the five algorithms may be considered "swarm-based" methods as they use a population of candidate solutions that communally search for optimal solutions. Clustering performance was evaluated in terms of "quantization error" explain this choice later (found in literature). The algorithms were evaluated by comparing the distributions of quantization errors produced from multiple simulations across the 10 datasets. Since there were several algorithm, each with distinct parameters and multiple datasets tuning was performed manually on a per-example basis with initial reference values provided by the literature. The five algorithms implemented were: k-means, DB-scan, competitive learning, particle swarm optimization (PSO), and ant colony optimization (ACO). Of these five, it was hypothesized that DB-scan and ant colony optimization would yield the lowest quantization errors across all data sets as 1) neither makes broad assumptions about the problem and 2) the literature provided results seemed to suggest these were the most generally powerful classifiers. RESULTS IN FACT SHOW THAT THIS WAS EITHER WRONG OR RIGHT AND MAY BE DUE TO THE FACT THAT SOME CHARACTERISTIC ABOUT ALGORITHM.**

I. Introduction

A common data mining task is organizing data into k seperate classes, i.e., clustering the data into seperate clusters. While a general definition of a cluster does not exist, for the purposes of this project a cluster is defined as a group of data points that are close to one another in terms of Euclidean distance.

Given a set of points $D = \{\mathbf{x_1}, \ldots, \mathbf{x_n}\}$ and a number of clusters, n_c, let a partition P of D be defined as a set of disjoints subsets of D such that the union of all said subsets is D. That is, a partition P of D is a set $\{P_i\}$ such that $\cap_i P_i = \emptyset$ and $\cup_i P_i = D$. Let each P_i have centroids

$$\mathbf{c_i} = \frac{1}{|P_i|} \sum_{\mathbf{x} \in P_i} \mathbf{x}$$

where $|P_i|$ denotes the number of elements in P_i. Define the "quantization error" of such a partioning as

$$QE(P) = \frac{\sum_{i=1}^{n_c} \frac{1}{|P_i|} \sum_{\mathbf{x} \in P_i} |\mathbf{x} - \mathbf{c_i}|}{n_c}$$

This is the average Euclidean distance between each point and it's respective partition, averaged across all partitions. For this project we seek a partitioning (henceforth referred to as a clustering) $P = \{P_i\}$ with $|P| = n_c$ minimizing the quantization error. This measure of performance was chosen as it appeared in the literature [2] and was thus a metric that enabled comparisons of results.

Five clustering algorithms were implemented: k-means clustering, ant colony optimization (ACO), particle swarm optimization (PSO), DB-Scan, and a competitive learning method (CL). The results from k-means, DB-Scan, and CL were considered baselines as k-means was previously implemented and is not a swarm-based method **and some other stuff about DB-Scan and CL**. The ACO and PSO results were then compared with the others for both individual comparisons and swarm-based/non-swarm-based comparisons.

II. Algorithm Descriptions

A. k-Means Clustering

Several variations of the k-means clustering algorithm have existed since at least 1957 [1]. The version used for this project is the most common and simplest except that the stopping criterion is based on the number of iteration of the algorithm and not a check for convergence.

Given that there are n_c clusters to be formed, n_c points were randomly chosen from the data set D to be cluster centroids $\mathbf{c_i} i \in \{1, \ldots, n_c\}$. Each point $\mathbf{x} \in D$ was then assigned to the cluster C_i containing the centroid closest to $\mathbf{x}$, i.e. $\mathbf{x} \in C_i$. Each centroid, $\mathbf{c_i}$ was then updated to be

$$\mathbf{c_i} = \frac{\sum_{\mathbf{x} \in C_i} \mathbf{x}}{|C_i|}.$$

Data points were then reassigned to new clusters based upon the updated cluster centroids. What distinguishes this version from the most common implementations is that this reassignment is repeated t_{max} times, regardless of whether the cluster centroids have converged or not. This choice was made so that a more fair comparison could be drawn with the other swarm based clustering algorithms which are limited by a predefined number of iterations.

B. Ant Colony Optimization (ACO)

This guy.[2]

C. Particle Swarm Optimization (PSO)

This guy.[3]

D. DB-Scan

E. Competitive Learning

III. Experimental Approach

A. Datasets

A total of 10 seperate datasets were obtained from the UCI Machine Learning Repository, all of which represented classification problems [4]. However, the class feature was removed from each data set as the problem at hand is an *unsupervised learning* task. The data sets chosen were

1) Banknote Authentication
2) Wine
3) Iris
4) Seed
5) Wilt
6) Bran1
7) Bran2
8) Bran3
9) Bran4
10) Bran5

B. Tuning

For each dataset, the number of clusters was chosen, naturally, to be the number of classes from the classification feature that was removed. Also, in the particular implementation of k-means clustering the authors chose, there were no true parameters to tune. For the other algorithms, final tuning parameters are provided in tabular form:

TABLE I
ACO Tuning Results

Dataset	ρ	ϵ	α
Bank	0.005	0.3	1
Wine	0.005	0.3	1
Iris	0.005	0.3	1
Seed	0.005	0.3	1
Wilt	0.005	0.3	1
Bran1	0.005	0.3	1
Bran2	0.005	0.3	1
Bran3	0.005	0.3	1
Bran4	0.005	0.3	1
Bran5	0.005	0.3	1

IV. Results

V. Conclusion

Stuff.

TABLE II
PSO Tuning Results

Dataset	w	c_1	c_2
Bank	0.72	1.49	1.49
Wine	0.72	1.49	1.49
Iris	0.72	1.49	1.49
Seed	0.72	1.49	1.49
Wilt	0.72	1.49	1.49
Bran1	0.72	1.49	1.49
Bran2	0.72	1.49	1.49
Bran3	0.72	1.49	1.49
Bran4	0.72	1.49	1.49
Bran5	0.72	1.49	1.49

TABLE III
DB-Scan Tuning Results

Dataset	w	c_1	c_2
Bank	0.72	1.49	1.49
Wine	0.72	1.49	1.49
Iris	0.72	1.49	1.49
Seed	0.72	1.49	1.49
Wilt	0.72	1.49	1.49
Bran1	0.72	1.49	1.49
Bran2	0.72	1.49	1.49
Bran3	0.72	1.49	1.49
Bran4	0.72	1.49	1.49
Bran5	0.72	1.49	1.49

References

[1] Steinhaus, H. (1957). "Sur la Division des Corps Matriels en Parties". Bull. Acad. Polon. Sci. (in French) 4 (12): 801804.

[2] Runkler, Thomas A. (2005). Ant Colony Optimization of Clustering Models. International Journal of Intelligent Systems, 20(12), 1233-1251.

[3] Merwe V. D. and Engelbrecht, A. P. (2003). Data Clustering Using Particle Swarm Optimization. Proceedings of IEEE Congress on Evolutionary Computation 2003 (CEC 2003), Canbella, Australia, pp. 215-220.

[4] Lichman, M. (2013). *UCI Machine Learning Repository* [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[5] Fenton, B., Sherrill, J. (2015). Git repository: https://github.com/joncheryl/csci447/tree/master/project4.

TABLE IV
COMPETITIVE LEARNING TUNING RESULTS

Dataset	w	c_1	c_2
Bank	0.72	1.49	1.49
Wine	0.72	1.49	1.49
Iris	0.72	1.49	1.49
Seed	0.72	1.49	1.49
Wilt	0.72	1.49	1.49
Bran1	0.72	1.49	1.49
Bran2	0.72	1.49	1.49
Bran3	0.72	1.49	1.49
Bran4	0.72	1.49	1.49
Bran5	0.72	1.49	1.49

TABLE V
k-MEANS PERFORMANCE

Dataset	Mean QE	SD
Bank	0.72	1.49
Wine	0.72	1.49
Iris	0.72	1.49
Seeds	0.72	1.49
Wilt	0.72	1.49
Bran1	0.72	1.49
Bran2	0.72	1.49
Bran3	0.72	1.49
Bran4	0.72	1.49
Bran5	0.72	1.49

TABLE VI
PSO PERFORMANCE

Dataset	Mean QE	SD
Bank	0.72	1.49
Wine	0.72	1.49
Iris	0.72	1.49
Seeds	0.72	1.49
Wilt	0.72	1.49
Bran1	0.72	1.49
Bran2	0.72	1.49
Bran3	0.72	1.49
Bran4	0.72	1.49
Bran5	0.72	1.49

TABLE VII
ACO PERFORMANCE

Dataset	Mean QE	SD
Bank	0.72	1.49
Wine	0.72	1.49
Iris	0.72	1.49
Seeds	0.72	1.49
Wilt	0.72	1.49
Bran1	0.72	1.49
Bran2	0.72	1.49
Bran3	0.72	1.49
Bran4	0.72	1.49
Bran5	0.72	1.49

TABLE VIII
DB-SCAN PERFORMANCE

Dataset	Mean QE	SD
Bank	0.72	1.49
Wine	0.72	1.49
Iris	0.72	1.49
Seeds	0.72	1.49
Wilt	0.72	1.49
Bran1	0.72	1.49
Bran2	0.72	1.49
Bran3	0.72	1.49
Bran4	0.72	1.49
Bran5	0.72	1.49

TABLE IX
COMPETITIVE LEARNING PERFORMANCE

Dataset	Mean QE	SD
Bank	0.72	1.49
Wine	0.72	1.49
Iris	0.72	1.49
Seeds	0.72	1.49
Wilt	0.72	1.49
Bran1	0.72	1.49
Bran2	0.72	1.49
Bran3	0.72	1.49
Bran4	0.72	1.49
Bran5	0.72	1.49