Genetic Weighted K-means for Large-Scale Clustering Problems

Fang-Xiang Wu[1*], Anthony J. Kusalik[1,2], and W. J. Zhang[1]

[1]Division of Biomedical Engineering, [2]Department of Computer Science
University of Saskatchewan, Saskatoon, SK S7N 5A9, CANADA
*Email: faw341@mail.usask.ca

Abstract

This paper proposes a genetic weighted K-means algorithm called GWKMA, which is a hybridization of a genetic algorithm (GA) and a weighted K-means algorithm (WKMA). GWKMA encodes each individual by a partitioning table which uniquely determines a clustering, and employs three genetic operators (selection, crossover, mutation) and a WKMA operator. The superiority of the GWKMA over the WKMA and other GA-clustering algorithms without the WKMA operator is demonstrated.

1. Introduction

Weighted k-means attempts to decompose a set of objects into a set of disjoint clusters, taking into consideration the fact that the numerical attributes of objects in the set often do not come from independent identical normal distribution. Weighted k-means algorithms are iterative and use hill-climbing to find an optimal solution (clustering), and thus usually converge to a local minimum. Genetic algorithms (GAs) (Obitko 2002) offer heuristic solutions to avoiding local minima in optimization problems. Several GA-clustering algorithms have been previously reported, some of which are not hybridized with clustering algorithms (e.g., Maulik and Bandyopadhyay 2000), and thus their rates of convergence were very slow. Others are hybridized with k-means algorithms (e.g. Wu, et. al, 2003). The resultant algorithms inherit some drawbacks of unweighted k-means algorithms, for example, that the resultant clusters are spherical-shape. This paper proposes a genetic weighted k-means algorithm (GWKMA) which encodes solutions as partitioning strings and employs three genetic operators (natural selection, crossover and mutation) and one WKMA operator.

2. GWKMA

Denote a set of n objects by $D=\{x_1,x_2\cdots,x_n\}$, and a partition of D by Δ. Abusing notation, let x_i also stand for the feature vector of object x_i ($i=1,\cdots,n$). For the preset number of clusters, K, the cost function for a weighted k-means clustering technique may be defined by

$$J_G(\Delta)=\sum_{k=1}^{K}\sum_{x_i\in D_k}(x_i-\overline{m}_k)G(x_i-\overline{m}_k)' \qquad (1)$$

where

$$\overline{m}_k=\sum_{x_i\in D_k}x_i/n_k \qquad (2)$$

m_k and n_k are the mean and the number of objects in D_k, respectively, and G is an arbitrary symmetrical positive matrix with $\det(G)=1$. The cost function (1) of a weighted k-means algorithm can be reduced to

$$J(\Delta)=(\det(W(\Delta)))^{1/d} \qquad (3)$$

The objective of a weighted k-means algorithm is to find an optimal partition Δ_o which minimizes

$$J(\Delta_o)=\min_{\Delta}(\det(W(\Delta)))^{1/d} \qquad (4)$$

where d is the dimension of the object's feature vectors.

1. **Initialize** the population Δ^*.
2. $[\Delta^*, J(\Delta^*)]=WKM(\Delta^*, X, K, N)$, $g=1$.
3. **While** ($g\le GEN$)
4. $\quad\tilde{\Delta}^*=Selection(\Delta^*, X, K, N)$;
5. $\quad\Delta^*=Crossover(\tilde{\Delta}^*, n, N)$;
6. $\quad\Delta^*=Mutation(\Delta^*, Pm, n, d, K, N)$;
7. $\quad[\Delta^*, J(\Delta^*)]=WKM(\Delta^*, X, K, N)$;
8. $\quad g=g+1$;
9. **End while,** and set $\Delta_o=\Delta_1$
10. **Return** $J(\Delta_o)$ and the resultant partition Δ_o.

Figure1. Genetic weighted K-means Algorithm (GWKMA)

Our GWKMA is shown in Figure 1. In the following the encoding and selection, crossover and mutation, and WKMA operators are specified in detail.

Encoding: A partitioning string is used to express a clustering. A partitioning string is an integer string over the set $\{1,\cdots,K\}$ on which each position corresponds to an object and the number in a position represents the cluster to which the corresponding object is assigned. The search space consists of all integer strings s_Δ with length n over the set $\{1,\cdots,K\}$. A population is expressed by a set of partitioning strings representing its individuals (solutions), denoted by $\tilde{\Delta}^*$ or Δ^*.

Selection operator $\tilde{\Delta}^* = Selection(\Delta^*, X, K, N)$: For convenience of the manipulation, GWKMA always assigns the best individual found over time in a population to individual 1 and copies it to the next population. Operator $\tilde{\Delta}^* = Selection(\Delta^*, X, K, N)$ selects $(N-1)/2$ individuals from the previous population according to the probability distribution given by

$$P_s(s_{\Delta i}) = F(s_{\Delta i}) \Big/ \sum_{i=1}^{N} F(s_{\Delta i}) \quad (5)$$

where N (odd positive integer) stands for the number of individuals in a population, $s_{\Delta i}$ is the partitioning string of individual i, and $F(s_{\Delta i})$ represents the fitness value of individual i in the current population, and is defined as

$$F(s_{\Delta}) = TJ - J(s_{\Delta}) \quad (6)$$

where $J(s_{\Delta})$ is calculated by (3), and TJ is calculated by the following formula

$$TJ = \det(S) = \det(\sum_{x \in D} (x - \overline{m})'(x - \overline{m})) \quad (7)$$

Crossover operator $\Delta^* = Crossover(\tilde{\Delta}^*, n, N)$: This operator creates new (and hopefully better) individuals from selected parent individuals. In GWKMA, of two parent individuals, one is always the first individual, and the other is one of the $(N-1)/2$ individuals selected from the parent population other than the first individual by the selection operator. Here the crossover operator adopts the single-point crossover method (Obitko 2002).

Mutation operator $\Delta^* = Mutation(\Delta^*, Pm, n, d, K, N)$: Each position in an encoding string is randomly selected with a mutation probability Pm set by the user, and the number in the selected position is uniformly randomly replaced by another integer from the set $\{1, \cdots, K\}$. To avoid any singular partition (containing an empty cluster), after the previous operation, the mutation operator also randomly assigns K different objects to K different clusters, respectively.

WKMA operator $[\Delta^*, J(\Delta^*)] = WKMA(\Delta^*, X, K, N)$: The WKMA operator employs a relocation-iteration algorithms (Wu et al. 2003) with each individual s_{Δ} in population Δ^* as an initial partition, returning N new partitions Δ^* and their cost function values $J(\Delta^*)$. This operator also arranges N new individuals such that such that the first individual is the best one in population Δ^*.

3. Computational Experiments and Results

The proposed GWKMA was run on three datasets to illustrate its performance. Dataset 1 (Spath 1980) contains 89 towns (objects) in Bavaria (Germany) with each having four features. Dataset 2 (Iyer et al. 1999) contains expression profiles of 517 genes (objects), each having 12 expression values (features). Dataset 3 (Chu et al. 1998) contains expression profiles of 6118 genes with each having 7 expression values.

In the experiment, the number of clusters is fixed at 7 for Datasets 1 and 3, and 10 for Dataset 2. Other parameters are set as follows: the number of generations $GEN = 50$, the population size $N = 21$, and mutation probability $Pm = 0.05$. The Matlab™ software package was used to conduct most of our experiments. The performance of GWKMA was first compared with the GA without weighted k-means. The result showed that GWKMA reaches the best solutions in ten generations for all three datasets while the GA without the weighted k-means yields inferior solutions even if taken to 50 generations. Then the performance of GWKMA was compared with that of the weighted k-means algorithm without GA. To do this, the weighted k-means trials are run 500 times; the GWKMA trials are run 5 times. The results showed that the best solutions of GWKMA in 5 test runs are better than the best solutions of the weighted k-means in 500 test runs for all three datasets. Furthermore, the standard deviations in the 5 runs of GWKMA are less than one-thirtieth of those in 500 runs of the weighted k-means algorithm for all three datasets. This result shows GWKMA is more insensitive to the initial partitions than the weighted k-means algorithm.

In conclusion, this study proposed a genetic weighted K-means algorithm (GWKMA) which is a hybrid algorithm of the weighted K-means and a genetic algorithm. GWKMA was run on three real-life datasets. The results of the computational experiments showed that GWKMA can fulfil the clustering tasks not only on some small-scale datasets such as SS2 but also on large-scale datasets such as gene expression datasets. Furthermore, the results also showed that the GWKMA outperformed both the WKMA without GA and other GA-clustering algorithms without the WKMA operator.

References

Chu, J., et al. 1998. The transcriptional program of sporulation in budding yeast, *Science* 282, 699-705.

Iyer, V.R., et al. 1999. The transcriptional program in the response of human fibroblasts to serum, *Science*, 283, 83-87.

Maulik, U. and Bandyopadhyay, S. 2000. Genetic algorithm-based clustering technique, *Pattern Recognition*, 33: 1455-1456.

Obitko, M. 2002. Introduction to Genetic Algorithms. http:// cs.felk. cvut.cz /~xobitko/ga/.

Spath, H. 1980. Cluster Analysis Algorithms for Data Reduction and Classification of Objects. Ellis Horwood Limited, West Sussex, UK.

Wu F.X.; Zhang W.J.; and Kusalik A.J. 2003. A genetic k-means clustering algorithm applied to gene expression data, *Canadian Conference on AI 2003:* 520-526.