International Journal of Information Technology and Knowledge Management (ISSN: 0973-4414) July 2012, Volume 5, No. Issues 2, pp. 518-521

A Genetic Algorithm for Clustering in Data Mining

Kalpana Gupta
Research Scholar, Deptt. of Computer Science, CMJ University, Shillong, INDIA
kalpana_gupta106@yahoo.com

Abstract
Clustering techniques have obtained adequate results when are applied to data mining problems. Clustering is the process of subdividing an input data set into a desired number of subgroups so that members of the same subgroup are similar and members of different subgroups have diverse properties. Many heuristic algorithms have been applied to the clustering problem, which is known to be NP Hard. Genetic algorithms have been used in a wide variety of fields to perform clustering, however, the technique normally has a long running time in terms of input set size. In this paper we investigate the use of Genetic Algorithms to determine the best initialization of clusters, as well as the optimization of the initial parameters. The genetic algorithm uses the most time efficient techniques along with preprocessing of the input data set. The experimental results show the great potential of the Genetic Algorithms for the improvement of the clusters. The techniques of clustering are most used in the analysis of information or Data Mining, this method was applied to Data Set at mining

1. Introduction

Data Mining is to nugget out potential, hidden useful knowledge and information from abundant, incomplete, noisy, fuzzy and random practical data. Clustering is an important method in data mining. It attempts to partition a dataset into a meaningful set of mutually exclusive clusters according to similarity of data in order to make the data more similar within group and more diverse between groups. Clustering technique can be divided into 7 categories: hierarchical clustering, partitioning clustering, density-based clustering, grid-based clustering, character attribute joint clustering, multi-dimensional data clustering and NN clustering. K-means algorithm in partitioning clustering is the most widely used. Clustering algorithm could enumerate all the possible partitions in theory to obtain the best solution. But this is a representative NP-Hard problem. K-means was proposed by MacQueen in 1967. It uses the Heuristic information to make the search more objective in order that the searching efficiency is improved. Its basic idea is that the clustering number K is assigned, firstly creating an initial partition stochastically, then using iteration method to improve the partition through moving the clustering centroid continually until the best partition is obtained. Actually, the best solution is unnecessarily obtained using that searching method. But through the Heuristic information, using the mean value to denote the centroid of each cluster reduced the computing complexity and increased the searching efficiency. That makes it possible to obtain the best solution of massive data under certain efficiency constraint. The traditional k-means cluster algorithm has its inherent limitations: ?Random initialization could lead to different clustering results, even no result. ? The algorithm is based on objective function, and usually take the Gradient method to solve problem. As the Gradient method searched along the direction of energy decreasing, that makes the algorithm get into local optimum, and sensitive to isolated points. So the improvement on K-means aims at two aspects: optimization of initialization and improvement on global searching capability. In traditional K-means algorithm, the clustering centroid is moving through the mean value of each cluster. But that direction is not always consistent with the best centroid. So during the process, the solution may get worse and the searching in certain stage is blind. The improved K-means algorithm added another Heuristic information-the obtained best centroid to improve the searching probability around the best clustering centroid. That improved the stability of the algorithm.

Clustering

Clustering is a kind of unsupervised learning. Clustering is a method of grouping data that share similar trend and patterns. Clustering of data is a method by which large sets of data are grouped into clusters of smaller sets of similar data.

K-means Algorithm

The process of K-means includes determine the clustering number K, select K clustering centroids stochastically and partition the objects to the nearest clustering centroid to form a cluster according to the Nearest-Neighbor rule, then compute the mean value of each cluster and make it the new clustering centroid The classical K-means algorithm is described as follows :

1. Choose a value for K, the total number of clusters to be determined.
2. Choose K instances (data points) within the dataset at random. These are the initial cluster centers.
3. Use simple Euclidean distance to assign the remaining instances to their closest cluster center

4. Use the instances in each cluster to calculate a new mean for each cluster.
5. If the new mean values are identical to the mean values of the previous iteration the process terminates. Otherwise, use the new means as cluster centers and repeat steps 3-5.

drawbacks of K-means clustering

The final clusters do not represent a global optimization result but only the local one, and complete different final clusters can arise from difference in the initial randomly chosen cluster centers. We have to know how many clusters we will have at the first step

Introduction of GAs

* Inspired by biological evolution.
* Many operators mimic the process of the biological evolution including
* Natural selection
* Crossover
* Mutation

Elements consisting GAs

* Individual (chromosome):
* feasible solution in an optimization problem
* Population
* Set of individuals
* Should be maintained in each generation

Elements consisting GAs

* Genetic operators. (crossover, mutation...)
* Define the fitness function.
* The fitness function takes a single chromosome as input and returns a measure of the goodness of the solution represented by the chromosome.

Genetic Representation

* The most important starting point to develop a genetic algorithm
* Each gene has its special meaning
* Based on this representation, we can define
* fitness evaluation function,
* crossover operator,
* mutation operator.

The overall procedure is summarized as follows:

1. Choose a value for K, the total number of clusters to be determined.
2. Get K instances (data points) within the dataset by GA

```
Begin  initialize parameters
    establish the initial population for clustering
        while (not termination condition) do
            fitness evaluation
            Selection
            mutation
        end do
end
```

3. Use simple Euclidean distance to assign the remaining instances to their closest cluster center
4. Use the instances in each cluster to calculate a new mean for each cluster.
5. If the new mean values are identical to the mean values of the previous iteration the process terminates. Otherwise, use the new means as cluster centers and repeat steps 3-5.

1) Individual representation: The chromosomes are made up of real numbers to represent the coordinates of the cluster centers. The length of the chromosome is K_i m, where K_i denotes the number of clusters of the i^{th} individual and m denotes the number of object attributes. The first m genes denote the m di- mensions of the first cluster center, the next m genes rep- resent those of the second cluster center, and so on. For instance, let m = 2 and K_i = 3, then the individual {25.2 18.6 5.3 10.8 65.3 7.0} represents the coordinates of three cluster centers {(25.2 18.6)(5.3 10.8)(65.3 7.0)}.

2) Population initialization: For individual i, its number of clusters Ki is randomly generated in the range [Kmin,Kmax]. Here, Kmin is chosen to be 2 unless specified otherwise and
Kmax is chosen to be root of(N), where N denotes the number of objects. For initializing individual i, Ki distinct objects are chosen randomly from the data set and viewed as the initial cluster centers.
3) Fitness evaluation: The aim of clustering analysis is to divide a given data set into clusters. A resulting partition should possess the following properties: (1) homogeneity within the clusters, i.e. data that belong to the same cluster should be as similar as possible, and (2) heterogeneity between the clusters i.e. data that belong to different clusters should be as different as possible.
4) selection: The selection operation is implemented as follows:
Step 1: Given population *Qt*, where *t* denotes the number of generations, set *i* = 1 and choose the *ith* individual *Xti* .
Step 2: If *t* = 1, then individual *Xt i* is selected and proceed to Step 4.
Step 3: Individual *Xti* is compared with the *ith* individual *Xt*-1*i* in population *Qt*-1,
if *Fti* - *Ft*-1*i* > 0, then individual *Xti* is selected;
otherwise individual *Xt*-1*I* is selected. Here, *Fti* and *Ft*-1*i* denote the fitness values of
individuals *Xti* and *Xt*-1*i* , respectively.
Step 4: View the selected individual as the *ith* individual and let *i* = *i* + 1. If *i* <= *P*, then return to Step 2; otherwise output the selected population. Here, *P* denotes the population
size.
5) mutation: In this approach, there are two kinds of individuals, the best individuals and the others. The best individuals have the highest fitness. Here, we view the best individuals as the solutions with the "correct" number of clusters.
6) Termination criterion*:* In general, two stopping criteria are used in genetic algorithms. In the first, the evolution process is executed for a fixed number of generations and the best individual obtained is taken to be the optimal one. In the other, the algorithm is terminated if no further improvement in the fitness value of the best individual is observed for a fixed number of generations, and the best individual obtained is taken to be the optimal one.

Conclusions

As a fundamental problem and technique for data analysis, clustering has become increasingly important. Many clustering methods usually require the designer to provide the number of clusters as input. K-means algorithm is one of the most widely used clustering algorithms in spatial clustering analysis. It is easy and efficient But is also has limitations: It is sensitive to the initialization. It doesn't perform well in global searching and is easy to get into local optimization. In this paper, we propose a genetic algorithm based clustering method.

References

[1] 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.
[2]W. Pedrycz, Knowledge-based clustering, Wiley, 2005.
[3]S. Z. Selim and M. A. Ismail, "K-means-type algorithm: generalized convergence theorem and characterization of local optimality," IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 6, no. 1, pp. 81-87, 1984.
[4] I. Charon and O. Hudry, "The noising method: a new method for combinatorial optimization," Operations Research Letters, vol. 14, no. 3, pp. 133-137, 1993.
[5] S. Bandyopadhyay and U. Maulik, "Nonparametric genetic clustering: comparison of validity indices," IEEE Trans. on Systems, Man, and Cybernetics, Part C: Applications and Reviews, vol. 31, no. 1, pp. 120- 125, 2001.
[6] L. Y. Tseng and S. B. Yang, "A genetic approach to the automatic clustering algorithm," Pattern Recognition, vol. 34, no. 2, pp. 415-424, 2001.
[7] S. Bandyopadhyay and U. Maulik, "Genetic clustering for automatic evolution of clusters and application to image classification," Pattern Recognition, vol. 35, no. 6, pp. 1197-1208, 2002.
[8] H. J. Lin, F. W. Yang and Y. T. Kao, "An efficient GA-based clustering technique," Tamkang Journal of Science and Engineering, vol. 8, no. 2,
113-122, 2005.
[9] C. C. Lai, "A novel clustering approach using hierarchical genetic algorithms," Intelligent Automation and Soft Computing, vol. 11, no. 3,

143-153, 2005.
[10] D. Bhandari, C. A. Murthy and S. K. Pal, "Genetic algorithm with elitist model and its convergence," International Journal of Pattern Recognition and Artificial Intelligence vol. 10, no. 6, pp. 731-747, 1996.