Mixed clustering algorithm with artificial fish swarm and improved K-means

Hao Yang
yanghao1986com@qq.com
College of New Media, Zhejiang University Of Media and Communications, China

Keywords: clustering algorithm, artificial fish swarm, K-means

Abstract. In allusion to increase the speed and accuracy rate of clustering algorithm, the paper proposes mixed clustering algorithm with artificial fish swarm and improved K-means. Firstly, it leads artificial fish swarm algorithm to clustering algorithm and proposes artificial fish swarm clustering algorithm. Secondly, it improves traditional K-means algorithm and gives improved K-means algorithm. Finally, it gets mixed clustering algorithm with artificial fish swarm and K-means.

Introduction

At present, many clustering algorithms are used for clustering analysis. Along with clustering algorithms which are based on many ideas and theory are unceasingly appear, practical application of clustering analysis is also widespread day by day. In 1967, MacQueen proposed K-means algorithm which is under the condition of given clustering number and initial clustering centre[1]. K-means algorithm has the property of sample respective ownership category centre distance square sum is smallest. This clustering algorithm has become the most classic clustering algorithm in clustering analysis. In 1990, Kaufinan and Rousseeuwf proposed Partitioning Around Medoids algorithm which partitioned around centre point[2]. This clustering algorithm is only widespread applied in biology domain. In 1996, Zhang proposed clustering BIRCH algorithm which used layered method to balance iteration restriction[3]. This algorithm first time proposed carry on pre-treat database via local clustering. In 2007, Freg proposed Affinity Propagation clustering algorithm[4]. This algorithm is an algorithm of high quality clustering centre by message transmission between data points. It can quickly process large scale data. In recent years, with data mining technology and computation intelligent technology unceasingly development, it proposes some new clustering algorithms based on bionics idea. For example, clustering algorithm based on ant group algorithm. This paper mixed artificial fish swarm algorithm and improved K-means algorithm. It proposes mixed clustering algorithm with artificial fish swarm and improved K-means.

This paper is organized as follows. Firstly, it introduces artificial fish swarm algorithm to clustering algorithm and proposes artificial fish swarm clustering algorithm. Secondly, it improves traditional K-means algorithm and gives the steps of improved K-means algorithm. Finally, we get mixed clustering algorithm with artificial fish swarm and K-means.

Artificial fish swarm algorithm

In the midst of waters, most places of fish survival number are richly contained nutrient most places in the waters. It imitates fish swarm look for food behaviour based on this characteristic. Then it achieves global search optimization. This is basic idea of fish swarm algorithm. In the fish's activity, preying behaviour, swarming behaviour, following behaviour and random behaviour have close relationship with optimization proposition solution. How to use simple and effective way to structure these behaviours is the main question of algorithm implementation.

The individual state of artificial fish is expressed by vector $X=(x_1,x_2,\cdots,x_n)^{\mathrm{T}}$, where $x_i(i=1,2,\cdots,n)$ is the pre-optimization variable. Food density of artificial fish in the current location position is expressed by $Y=f(X)$, where Y is the objective function value. The distance between two artificial fish is expressed by $d_{i,j}=\left\|X_i-X_j\right\|$. *Visual* represents sensation distance of artificial fish. *Step* represents maximum step of artificial fish movement. δ is overcrowding scale factor.

Behaviour description of artificial fish swarm is as follows:

(1) Preying Behaviour.Fish in the water are swim freely. This is generally considered as a random swim. When fish find food, they quickly swim to food increasing gradually direction. Let X_0 be artificial fish current state, we random selected a state X_j. If $X_i < X_j$ in the maximum value question or $X_i > X_j$ in the minimum value question, then it goes a step further advance to this direction. Otherwise it random selects state X_j again. It judges whether satisfies advance condition. After repeated many times, if it does not satisfy advance condition, then it random moves one step.

(2) Swarming BehaviourArtificial fish in the roving process are naturally meet gather in swarm. This is one kind of life habit in order to guarantee swarm's survival and avoid hazard. The formation of artificial fish swarm is also one kind appear suddenly vivid demonstration.

(3) Following BehaviourWhen an artificial fish find food, nearby fish will follow its approach partners quickly arrive food position. Let X_i be artificial fish current state, explore current domain $d_{i,j}$<*Visable* partner number $n_f$and current partner Y_j is the maximum partner X_j.

(4) Random Behaviour.In the spare time, artificial fish in the water random, free is looking for food. Let X_i be artificial fish current state, artificial fish random selects a state X_j to movement in the context of perception. The random behaviour is non-deterministic.

(5) Restraint BehaviourIn the clustering process, for operation effect of swarming behaviour and random behaviour, it is easy to cause artificial fish's state become not feasible. Now it is need join corresponding restraint condition to carry on regularize. It makes artificial fish is from invalid state or unworkable state transition into feasible state or workable state.

The main idea of artificial fish swarm algorithm applied to clustering algorithm is as follows: It random puts artificial fishes into a three dimensional space. Every artificial fish has a random initial location and moves in the three dimensional space. It measures community similarity of artificial fish in partial environment. Artificial fishes achieve self organization clustering process by community similarity. It forms artificial fishes clustering in the three dimensional space.

Community similarity of artificial fish swarm is the comprehensive similarity among an artificial fish and all other artificial fishes in the partial environment. Basic measure formula of community similarity is $Y = \frac{1}{n}\sum_{i=1}^{n} d_{i,j} = \frac{1}{n}\sum_{i=1}^{n}\left\|X_i - X_j\right\|$,where $d_{i,j} = \left\|X_i - X_j\right\|$ represents the distance between two artificial fish.

Food density function of artificial fish in the current location position is the movement standard of artificial fish. Formulation main principle of food density function is community similarity. Community similarity is bigger, food density is smaller. Food density is bigger, community similarity is smaller. Artificial fish swarm clustering algorithm is described as follows:

1) Initialization number of artificial fish, sensation distance of each artificial fish, movement distance of step, three dimensional space region context, number of cycles.

2) It puts artificial fish random into three dimensional spaces. Each artificial fish has three dimensional coordinates $(x_1, x_2, x_3)^{\mathrm{T}}$. Number of cycles automatic add 1 begin execute a new cycle.

3) A group of artificial fish begin clustering cycle. It uses formula $Y = \frac{1}{n}\sum_{i=1}^{n} d_{i,j} = \frac{1}{n}\sum_{i=1}^{n}\left\|X_i - X_j\right\|$ to compute community similarity of every artificial fish. Then it explores clustering centre position X_c of current artificial fish and measures community similarity of artificial fish. It executes swarming behaviour and following behaviour by community similarity of clustering centre position and current artificial fish. Otherwise it executes preying behaviour.

4) If all artificial fishes in the group finish movement, then it executes next step. Otherwise it begin artificial fish clustering cycle.

5) If number of cycles less than maximum number of cycles, then Number of cycles automatic add 1. Otherwise output result, end artificial fish swarm clustering program.

Improved K-means algorithm

The main idea of K-means algorithm is as follows:

1) It random selects *k* points from data set as initial clustering centre.

2) Compute distances which are from each sample to clustering. The sample turns over to clustering which leaves closest clustering centre.

3) We get new clustering centre by computing new formation every clustering data object average value. If adjacent two time clustering centre has not any changed, then sample adjustment is over and clustering criteria function is converged.

K-means algorithm is using error square sum criterion function as clustering criteria function. Error square sum criterion function is defined as $E=\sum_{i=1}^{k}\sum_{p\in c_i}^{j}\left|p-m_i\right|^2$, where E is square error sum of all data objects, p is data object, m_i is average mean value of cluster c_i.

K-means algorithm random selects k objects as initial clustering centre from n data objects. Clustering result by K-means algorithm has very big uncertainty. Process of K-means algorithm is described as follows:

1) K-means algorithm based on objects average value in the cluster.

2) Input number k of cluster and database which contains n objects.

3) Output k clusters and minimum square error criteria.

Step of K-means algorithm is described as follows:

1) Initialization. It random selects k objects as initial cluster centre.

2) Number of cycles automatic add 1 begin execute a new cycle.

3) Each object assigns most similar cluster by object average value in the cluster.

4) Updating cluster average value until object average value of each cluster no longer change so far.

Improved K-means algorithm constructs a distance cost function and solves optimal cluster number k by distance minimum cost criterion. For n object set $X=(x_1,x_2,\cdots,x_n)^{\mathrm{T}}$ in the n space objects and these n space objects is clustered k cluster.Inter clusters distance L is distance sum from all clustering centres to global centre. $L=\sum_{i=1}^{k}\left|m_i-m\right|$, where m is average value of all samples, m_i is average value of samples in cluster c_i, k is number of clustering.Distance D within clusters is internal distance sum of clustering cluster. Internal distance of each cluster is distance sum from all samples in the cluster to centre. $D=\sum_{i=1}^{k}\sum_{p\in c_i}^{j}\left|p-m_i\right|$, where p is a sample.Distance cost function $F(s,k)$ is sum of inter clusters distance L and distance D within clusters. $F(s,k)=L+D=\sum_{i=1}^{k}\left|m_i-m\right|+\sum_{i=1}^{k}\sum_{p\in c_i}^{j}\left|p-m_i\right|$. It uses distance cost function $F(s,k)$ as space clustering validity test function to determine distance cost minimum criterion. When $F(s,k)$ achieves minimum, space clustering result is optimal. The optimal selection of k is $\min_k\left\{F(s,k)\right\}(k=1,2,3,\cdots,n)$.

Theorem 1 When $L=D$, space clustering number k reaches optimization, $k\le n^{1/2}$.

$F(s,k)$ has feature of simple structure and small computational complexity. $F(s,k)$ has a better test result.

Process of improved K-means algorithm is described as follows:

1) Based on K-means algorithm, it uses $F(s,k)$ to optimize value k.

2) Input database which contains n objects.

3) Output k^* clusters under distance cost function minimum condition.

Step of improved K-means algorithm is described as follows:

1) Compute space clustering number k reaches optimization, $k\le n^{1/2}$.

2) It uses K-means algorithm to achieve space clustering with the condition $k\le n^{1/2}$.

3) It uses distance cost function to respectively compute $F(s,k)$ with the condition different clustering number k.

4) It searches distance cost function minimum $F(s,k)^*$ and mark down k^*.

Mixed clustering algorithm

Artificial fish swarm algorithm has a fast convergence speed. It can be used to solve real-time requirement problem. For some occasion of accuracy is not high, you can quickly get a feasible solution with artificial fish swarm algorithm. Algorithm does not need strict problem mechanism model and problem precise description. Improved K-means algorithm is a simple and quick algorithm. It can solve clustering problem. Improved K-means algorithm has relatively scalable and high efficient for deal with large data set. Algorithm tries to find *k* partition which causes minimum square error value. Improved K-means algorithm only can be used in the situation of cluster average value is defined. Users must be given number of clusters. For different initial value, it may lead to different clustering results.

We can be mixed artificial fish swarm algorithm and improved K-means algorithm. Artificial fish swarm algorithm uses random ergodic principle to clustering analysis. Improved K-means algorithm uses determine/heuristic principle to clustering analysis. Artificial fish swarm algorithm can avoid local optimal occurrence, but it takes a long time. Improved K-means algorithm can quickly and efficiently assign data objects to corresponding clusters. Mixed clustering algorithm with artificial fish swarm and improved K-means no longer requires input initial segmentation and avoids initial error information lead to error clustering result. It can improve efficiency of artificial fish swarm algorithm and improved K-means algorithm.

Process of mixed clustering algorithm with artificial fish swarm and improved K-means is described as follows:

1) It uses artificial fish swarm algorithm to clustering analysis single data object.

2) Investigating the clustering result and selecting input points for clustering analysis by improved K-means algorithm.

3) It uses artificial fish swarm algorithm provides segmentation as input point and improved K-means algorithm to clustering analysis.

Summary

In this paper, mixed clustering algorithm with artificial fish swarm and improved K-means is proposed. It can increase the speed and accuracy rate of clustering algorithm. In the future work, we will study more methods to improve clustering algorithm.

Acknowledgments

This work was supported in part by the Introduction of Zhejiang University Of Media and Communications Scientific Research Grants Project(Z301B15521).

References

[1] MacQueen J,Some Methods for Classification and Analysis of Multivariate Observations, Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability,(1967)pp 81-297.

[2] Kaufinan L, Rousseeuw P,Finding Groups in Data: an Introduction to Cluster Analysis,NewYork: John Wiley and Sons,1990.

[3] Zhang T, Ramakrishnan R, Livny M ,An efficient data clustering method for very large databases, Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data,(1996)pp 103-144.

[4] Frey B, Dueck D ,Clustering by passing messages between data points, Science, 315(5814),(2007) pp 972-976

[5] Atsuyoshi N,An efficient query learning algorithm for ordered binary decision diagrams. Information and Computation. 201(2), (2005)pp 178-198.

[6] Maxim T, Andres M, Elena D,Bound-set preserving ROBDD variable orderings may not be optimum. IEEE Transactions on Computers. 54(2),(2005) pp 236-237.