Optimizing the Ant Clustering Model Based on K-Means Algorithm

Qin Chen, Jinping Mo
School of Computer, Electronics and Information, Guangxi University, Nanning,*P. R.* China
qchen218@163.com

Abstract

Ant clustering is one of effective clustering methods. Compares to other clustering methods, ant clustering algorithm has one outstanding advantage and one disadvantage. The advantage is that the total numbers of cluster is generated automatically ,and the disadvantage is that its cluster result is random and its result is influenced by the input data and the parameters, which leads low quality of its cluster result. In this paper, we propose an improved ant clustering algorithm based on K-Means, which optimizes the rules of ant clustering algorithm . In our system, we also decide the proper values of parameters Pdel and Iter by training the training datasets before we cluster. Experimental results demonstrate that the proposed method has a good performance.

Key words: K-Means , Rules,Ant Clustering Algorithm, Parameters

1.Introduction

Clustering is a process of grouping a set of data objects into clusters based on the information found in the data objects[1]. After completing the clustering process, data objects in a cluster are similar to each other and are different from data objects in other clusters. Because the grouping phenomenon for data objects can be captured through the clustering process, clustering plays an important role in various data analysis fields including statistics (McLachlan and Krishnan, 1997), pattern recognition (Webb, 2002), machine learning (Alpaydin, 2004), data mining (Tan et al., 2005),information retrieval (Wu et al., 2003), and bioinformatics (He et al., 2006). Among various clustering algorithms, K-means(Forgy, 1965; McQueen, 1967) is one popular and widespread partitioning algorithm because of its superior feasibility and efficiency in dealing with a large amount of data (Hand and Krzanowski, 2005). Ant clustering algorithm is another useful clustering algorithm because it is able to find utomatically a good partition over artificial and real data sets.Furthermore,it does not need the number of expected clusters to converge(Nicolas Labroche, 2002). Ant clustering can treat small to big sets of data with a great success but also demonstrate,that Ant clustering does not manage to find a good partition when an important number of clusters is expected.This may be due to the fact that there is only one rule that can create a new nest[2].

In this paper, we focus on how to assign an ant to the proper nest. For attaining this purpose, we use K-Means to optimize the rules of the Ant clustering algorithm.We also decide the proper values of parameters Pdel and Iter by training the training datasets before we cluster.

2. Related work

2.1 The Ant Clustering

The originality of Ant clustering is to model the chemical recognition system of ants to solve the unsupervised clustering problem[3]. It finds groups of similar objects as close as possible to the natural partition of the given data set. No assumptions are made about the representation of the objects. They may be described with numerical or symbolic values or with first order logic. All we need is the definition of a similarity measure which takes a couple of objects i and j as input and outputs a value Sim(i,j) between 0 and 1 . Value 0 means that the two objects are totally different, 1 means that they are identical.For one ant i , we define the following parameters[4]: (1)The Labeli,which is determined by the belonging nest of the ant i and is simply coded by a number, representative of the nest.(2) The Template,which is defined half by the genetic odor Genetic_i of the ant and half by an acceptance threshold Template_i.(3) An estimator M_i that reflects if the ant i is successful during its meetings with all encountered ants or not. (4) An estimator M_i^+which measures how well accepted is ant i in its nest.(5) An age Ai which, at the beginning, equals 0 and is used when updating acceptance threshold.(6) Estimates of the maximal similarity Max(Sim(i,·)) and mean similarity $\overline{Sim}(i,\cdot)$ observed during its meeting with other ants.

Considered thereafter two ants i and j ,there is

978-0-7695-3507-4/08 $25.00 © 2008 IEEE
DOI 10.1109/CSIE.2009.813

acceptance (or recognition) between i and j which is defined as :

Acceptance(i,j) $\Leftrightarrow (Sim(i,j) > Template_i) \wedge$ $(Sim\ (i,\ j) > Template\ _j)$

The Ant clustering use the follow behavioral rules associated with meetings[4]:

(1)The 1st rule : when two ants with no nest meet and accept each other,a new nest is created.

(2)The 2nd rule : when an ant with no nest an ant that already belongs to a nest,the ant that is alone joins the other in its nest.

(3)The 3rd rule:when two ants that belong to the same nest meet and accept, increments the estimators M and M_i^+.

(4)The 4th rule: when two nest mates meet and do not accept each other, decreases the estimators M and M_i^+ and the worst interpreted ant is ejected from the nest.

(5)The 5th rule : when two ants that belong to a distinct nest meet and accept each other, merger the two nests.

(6)The 6th rule: when no other rule applies,nothing happens.

2.2 K-Means clustering algorithm

K-means (MacQueen, 1967) is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed a priori. The main idea is to define k centroids, one for each cluster. These centroids shoud be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centroid. When no point is pending, the first step is completed and an early group age is done. At this point we need to re-calculate k new centroids as barycenters of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new centroid. A loop has been generated. As a result of this loop we may notice that the k centroids change their location step by step until no more changes are done. In other words centroids do not move any more. Finally, this algorithm aims at minimizing an *objective function*, in this case a squared error function. The objective function is defined as:

$$f = \sum_{j=1}^{k} \sum_{i=1}^{n} \| x_i^{(j)} - c_j \|^2$$

where $\| x_i^{(j)} - c_j \|^2$ is a chosen distance measure between a data point $x_i^{(j)}$ and the cluster center c_j, is an indicator of the distance of the *n* data points from their respective cluster centres.

3. Proposed algorithm

3.1 Modified clustering rules using K-Means

As we state before, the 3rd rule of the Ant clustering simply increments the estimators M_i and M_i^+ in case of acceptance and the 4th decreases the estimators M_i and M_i^+ in case of not accepting between the two ants and the worst interpreted ant is ejected from the nest.the estimator M_i^+ measures how well accepted is ant i in its nest ,but the value of M_i^+ do not represent that nest is the best for ant i. Ejecting an ant i from the nest according these rules can create accumulation error which leads to low clustering quality.In this paper ,we focus on how to assign an ant to the proper nest and use K-Means to modify ant clustering rules.

Let di be the distance between ant i and the centroid of the nest. We modify Ant clustering rules as follows:

(1) When two ants with no nest meet and accept each other, create a new nest and calculate the centroid of the nest.

(2)When an ant with no nest meets and accepts an ant that already belongs to a nest,the ant that is alone joins the other in its nest. And we renewal the centroid of the nest.

(3) In case of acceptance between two ants that belong to the same nest, calculate di for each ant and the ant which has larger value of di is ejected from the nest. And we renewal the centroid of the nest.

(4) When two ants that belong to a distinct nest meet and accept each other, merger the two distinct nest and renewal the centroid of the nest.

(5) When no other rule applies,nothing happens.

3.2 Parameters Settings

It has been shown in [2] that the quality of the convergence of Ant clustering mainly depends on two major parameters,namely the number of iterations of the meeting step Nb_{Iter}= $Iter * N$, and the probability of deleting the nests P_{del}. Too small or too large value of the Nb_{Iter} can influent the clustering result. Too small value can lead low rate of the data overlay and too large value can lead exceed-learn. Too small or too large value of the P_{del} can influent the clustering quality. Too small value can lead less number of the clusters and too large value can lead contrary case.So the values of the parameters Nb_{Iter} and P_{del} is very important.We describe hereafter how we can fix the value of these parameters .First,we present our measure of the performance of the algorithm and the data sets used for evaluation.

The performance of a clustering algorithm can be evaluated using various cluster validity measures (Maulik and Bandyopadhyay,2002;Pakhira et al.,2004). F-measure is one of the most popular measures for evaluating clustering result by combing the precision and recall ideas from information retrieval[5], and thus we use F-measure as an illustration example in this study. Assume we have p classes and q clusters, where q is usually equal to p. We calculate the recall and precision of that cluster for each given class. More specifically, for class i (1≤i≤p) and cluster j (1≤j≤q)

$$P(i,j)=Precision(i,j) = n_{ij}/n_j$$

$$R(i,j)=Recall(i,j) = n_{ij}/n_i$$

where n_{ij} is the number of members of class i in cluster j, n_j is the number of members of cluster j and n_i is the number of members of class i.The F-measure of class i and cluster j is then given by

$$F(i,j)=(2*R\ (i,j)*P\ (i,j))/(P\ (i,j)+R\ (i,j))$$

For an entire clustering the F-measure of any class is the maximum value it attains at any cluster and an overall value for the F-measure is computed by taking the weighted average of all values for the F-measure as given by the following.

$$F=\sum_{i=1}^{p}\frac{n_i}{n}\max{}_j\{F(i,j)\}$$

where the max is taken over all clusters and n is the number of all objects in the set. The larger the F-Measure is, the better the cluster quality is.

To evaluate performance of the proposed algorithm, the data sets (see http://kdd.Ics.Uci.edu for detail)have been used.The main characteristics of data are summarized in table 1.The field for each data set are:the number of objects(N),their associated number of attributes(M),and the number of clusters(K)

Table 1. Main characteristics of data set

Data set	N	M	K
Iris	150	4	3
Breast-cancer	178	10	2
Wine	683	13	3
Soybean	47	35	4
Glass	214	9	7

3.2.1 Parameter Nb_{Iter} Setting

As stated before,the value of the parameter Nb_{Iter} can influent the clustering result,that is, appropriate value can lead good clustering result.In our system,we test several values of Iter expressed as half of the minimal number of iterations.The figure 1 presents the results that we obtain for each data set in term of mean F-Measure according to the value Iter.

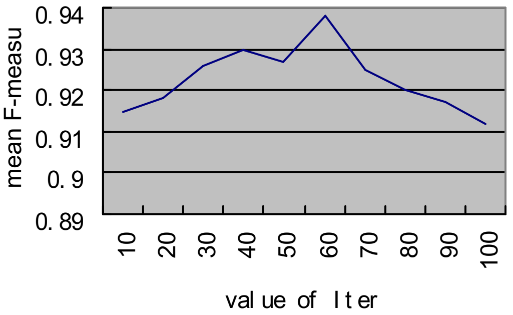

Figure 1 mean F-Measure for each Iter and each data set

According to this figure,we can see that the value of the mean F-measure achieves the largest when the value of Iter is 60.

3.2.2 Parameter P_{del} Setting

As stated before,the value of the parameter P_{del} can influent the clustering quality. that is, too small value can lead less clusters and too large value can lead contrary case. For the time being,the nest deletion criterion is only based on the number of ants that belong to this nest and a threshold fixed by the user. In our system,we test several values P_{del} of expressed as the nest deletion threshold.The figure 2 presents the results that we obtain for training data set(Iris) in term of mean number of the clusters according to the value P_{del} ranging from 0 to 0.2.

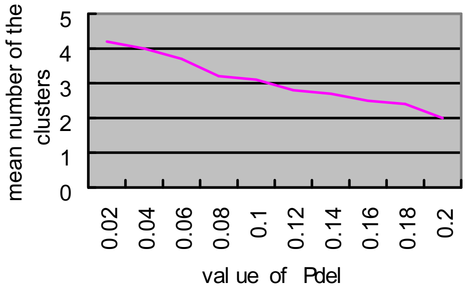

Figure 2 Number of the clusters for each Pdel over 10 runs

According to this figure,we can see that the value of the number of the clusters is most close to the actual number of the clusters(3) when the value Pd_{el} is 0.08.

4. Experiments and Results

In this section,we compare our method to the Ant clustering algorithm to evaluate the performance of our method.For each data set,we run 30 times each method and computer the mean F-measure(see in section 3.2). The table 2 shows the mean number of clusters effectively found (#clusters) and mean F-measure for each data set by both methods.

Table 2 Results obtained after 30 runs of each method applied over each data set

	#clusters		mean F-measure	
Data set	Ant	our method	Ant	our method
Iris	2.3	2.8	0.9518	0.9887
Breast-cancer	2.5	1.9	0.9074	0.9432
Wine	4.4	3.3	0.6396	0.7185
Soybean	3.6	4.2	0.9354	0.9457
Glass	5.7	7.2	0.8458	0.9214

The results from table 2 demonstrate that our algorithm performs better than the Ant clustering method. It seems to be mainly because Ant clustering manages to have,in general, a better appreciation of the number of clusters in the data.

5. Conclusion

In this paper we describe an improved ant clustering algorithm based-on K-Means,and we show how the parameters of the method can be set.We evaluate its performance against Ant clustering with real data sets and the results show that our proposed algorithm can do better than Ant clustering algorithm.

Acknowledgement

This work was supported by
(1)The Education Department of Guangxi Province under Grant NO. 200708LX106
(2) Project supported by Program to Sponsor Teams for Innovation in the Construction of Talent Highlands in Guangxi Institutions of Higher Learning

References

[1]Chieh-Yuan Tsai,Chuang-Cheng Chiu, “Developing a feature weight self-adjustment mechanism for K-means clustering algorithm”,Computational Statistics and Data Analysis,(2008,pp.4658-4672.

[2] N Labroche , N Monmarche , G Venturini, “A new clustering algorithm based on the chemical recognition system of ants ” 15th European Conference on Artificial Intelligence (ECAI 2002),Lyon FRANCE, 2002 , pp.345-349 .

[3] Nicolas Labroche , Nicolas Monmarché , Gilles Venturini .“AntClust :Ant Clustering and Web Usage Mining”, GECCO, 2003,pp.25-36 .

[4] Nicolas Labroche , Nicolas Monmarché , Gilles Venturini, “Web sessions Clustering with Artificial Ants Colonies” . EB/OL,2002 .

[5] Baeza-Yates,R.,and Ribeiro,B. “Modern Information Retrival”,ACM Press and Addison Wesley,1999.