An Improved Heuristic Ant-Clustering Algorithm

Yunfei Chen Yushu Liu Jihai Zhao

(Department of Computer Science and Engineering, School of Information Science and Technology, Beijing Institute of Technology, Beijing 100081)

E-mail: cyf_990@hotmail.com

Abstract: An improved heuristic ant-clustering algorithm (HAC) is presented in this paper. A device of 'memory bank' is proposed, which can bring forth heuristic knowledge guiding ant to move in the bi-dimension grid space. The device lowers the randomness of ants' moving and avoids the producing of "un-assigned data object". We have made some experiments on real data sets and synthetic data sets. The results demonstrate that HAC has superiority in misclassification error rate and runtime over the classical algorithm.

Keywords: ant-clustering, memory bank

1 Introduction

Ant-clustering algorithm was originally introduced for tasks in robotics by Deneubourg in 1991[1]. Lumer and Faieta modified the algorithm to extend to numerical data analysis by introducing a measure of dissimilarity between data objects[2]. Kuntz et al. applied it to the: graph-partitioning [3], text-mining [4], et al.. Wu and shi modified Deneubourg's basic model to derive CSI [5]. Note that Monmarché has introduced an interesting AntClass algorithm, a hybridization of an ant colony with the k-means algorithm, and using the misclassification error for evaluation purposes[6].

Contrasting with those classical clustering methods (like K-means and ISODATA), ant clustering boasts a number of advantages as Autonomy (Not any prior knowledge like initial partition or number of classes is required. Clusters are formed naturally through ant's collective actions.), Flexibility (Rather than deterministic search, a stochastic one is used to avoid the locally optimal.), Parallelism (Agent operations are inherently parallel).

While many of those advantages looked prefect, two important defects remain. Firstly, due to the fact that each time when ant is assigned a new data object the data object is selected at random, there may be some data objects that have never be assigned. We call them "un-assigned objects". Un-assigned objects lead to a high misclassification error rate of the algorithm, for they have never taken part in the clustering loop. The second defect is, by the random moving of ants, ant-clustering algorithms have a slow convergence rate. Result in the long time consume of clustering algorithm.

2 An Improved Heuristic Ant Clustering Algorithm

We present in this paper an improved Heuristic Ant Clustering (HAC) algorithm, which follows the basic ideas of LF and CSI [2,5]: Ants randomly move in their environment, a square grid with periodic boundary conditions. Data items that are scattered within this environment can be picked up, transported and dropped. Ants are likely to pick up data object with low similarity to others in the local neighborhood; they tend to drop them in the vicinity of similar ones. Finally, clusters are visually formed on the grid through ants' collective actions.

The main differences of the improved algorithm lie in:

(1) The device of 'memory bank' is proposed when realizing the simple agent---ant.

(2) Ant searches the global data object memory, rather than the whole data set, to find a new data object. Global data object memory stores the serial number of never-carried data object.

(3) Ant referentially searches his local environment memory for a 'best matching' position to drop a new data object.

The whole algorithm is presented in pseudo code in figure 1.

2.1 Similarity and Probability Conversion Function

Let us assume that an ant is located at site r and finds an object O_i at that site. The similarity of object O_i to the other type O_j in a local region is given by

$$f(i)=\begin{cases}\dfrac{1}{\sigma^2}\displaystyle\sum_{O_j\in Neigh(r)}\left[1-\dfrac{d(o_i,o_j)}{\alpha}\right], f(i)>0\\ 0 \qquad \text{otherwise}\end{cases} \tag{1}$$

Here, Neigh(r) denotes the local region. It is usually a square of $\sigma\times\sigma$ sites surrounding site r. The parameter α is a key coefficient that directly affects the number of cluster and con-

Briefe Introduction to Authors: Yunfei Chen was born in 1977. He is studying for doctorate. His main research areas include artificial intelligence, and data mining. Yushu Liu was born in 1941. He is a professor and doctoral supervisor. His main research areas include aided strategy and artificial intelligence.

```
1.Initialize all parameters;Randomly scatter data items on a plane;
2.for each it_ctr in 1 to # iterations Niterate do
3.  for each ant in 1to #ants Nant do
4.    while is_empty(global_data_object_list)=FALSE do
5.      if ant_is_load=FALSE then
6.        Similarity:=Compute_similarity(data_object(i),σ-neighborhood)
7.        P_pick:=Compute_pickup_probability(Similarity)
8.        If P_pick<Pr, :=random_generate_probability()then
9.            i:=random_select(never_carried_object)
10.       else ant_is_loaded=true
11.           ant_moving(random_coordinates_x,random_coordinates_y)
12.           remove_from_global_data_object_list(data_object(i))
13.     end if
14.   else
15.     while(is_empty(local_environment_stack)=FALSE)do
16.       Similarity:=Compute_similarity(data_object(i),pop_first
                    (local_environment_stack))
17.       P_drop:=Compute_drop_probability(Similarity)
18.       If P_drop>Pr: :=random_generate_probability()then
19.         ant_drop(data_object(i))
20.         ant_is_loaded=FALSE
21.         j:=random_select_data_obect(never_carried_object)
22.         break;
23.       end if
24.     end while
25.     If ant_is_loaded=true
26.       ant_moving(random_coordinates_x,random_ coordinates_y)
27.       follow the method like step 14~21;
28.     end if
29.   end if
30.  end while
31. end for
32.end for
33.Output the clustering result,Stop.
```

Fig.1 The main algorithm of HAC

vergence of the algorithm.

Similarity is converted to the probability of picking or dropping operations by probability conversion function.

$$p_{pick}(i)=\left(\frac{k_1}{k_1+f(i)}\right)^2 \tag{2}$$

$$p_{drop}(i)=\begin{cases}2f(i) & f(i)<k_2\\ 1 & f(i)\geq k_2\end{cases} \tag{3}$$

2.2 Memory bank

Real ants may perceive their surrounding environments and then memorize several important sites and objects.Inspired from this,we add a device of 'memory bank'to each ant.It is composed of'local environment memory'and'global data object memory'.

2.2.1 Local environment memory

Local environment memory stores the ever-carried objects' dropping positions.Ant referentially searches his local environment memory for a'best matching'position to drop a new data object.So local environment memory brings forth heuristic knowledge guiding ants' moving in the bi-dimension grid space.The'best matching'position i,defined by a pair of coordinate (x,y),is a cell with a σ-neighborhood where the data object has the biggest value of similarity.

2.2.2 Global data objects memory

Besides the data objects' dropping positions,ant has to remember the set of never-carried data objects.When ant is assigned a new data object,it searches the global data object memory to find a new one that has never been carried.The adding of global data objects memory has assured that not only the whole set of data objects can be surveyed at least once by a single ant,but also each data object can be surveyed many times through the collective actions of ant colony.Therefore the producing of 'un-assigned data object' is avoided.

3 Experiment results and analysis

We have applied the new HAC algorithm to several numerical databases including synthetic databases:ANT1(75,2,4),ANT2(500,2,4),ANT3(800,2,9)and real ones:Iris(150,4,3) and Soybean (307,35,19)from the MIC Machine Learning repository.

Tab.1 Results of the Clusters Number,Classification error rate and runtime

ANT1	k-means	CSI	HAC
#Clusters	4	4	4
Cl.Err	2.15%	3.02%	1.05%
Runtime	6.0	10.2	5.8
ANT2	k-means	CSI	HAC
#Clusters	4	4	4
Cl.Err	11.25%	11.35%	5.48%
Runtime	19.5	21.3	15.9
ANT3	k-means	CSI	HAC
#Clusters	9	9	9
Cl.Err	14%	13%	7.02%
Runtime	30.6	31.2	27.4
IRIS	k-means	CSI	HAC
#Clusters	3	3	3
Cl.Err	5.28%	5.66%	2.13%
Runtime	10.4	13.4	9.3
SOYBEAN	k-means	CSI	HAC
#Clusters	19	19	19
Cl.Err	8.21%	8.48%	3.98%
Runtime	15.5	15.7	13.1

3 evaluation measures were used to evaluate the resulting partition.They are the number of identified cluster(#Clusters), the misclassification error rate (Cl.Err)and the overall running time(Runtime).Table 1 gives the means value of the 3 evaluation measures(over 100 trials)of the 3 algorithms on 5 data sets.The results demonstrate that,if clear cluster structures exist within the data,the ant-clustering algorithm including:CSI and HAC,is quite reliable at identifying the correct number of clusters.In contrast with the k-means,the ant-clustering algorithm shows its strength in the ability to automatically determine the number of clusters.Comparing the runtimes of the 3 algorithms,we can see HAC is the fastest algorithm and its time consumer changes little with the scale of data set.So it is a fast clustering algorithm with prefect scalability.The CSI

(下转 158 页)

据包长度为 1k bits,平均 PSNR 增益如表 2:

表 1 测试序列参数

序列名称	格式	帧率	编码帧数
Container	CIF	15 fps	300
Coastguard	CIF	15 fps	300
Foreman	CIF	30 fps	300

表 2 MPEG-4 抗误码模式与非抗误码模式比较

序列名称	格式	DP 模式下平均 PSNR	非 DP 模式下平均 PSNR
Container	CIF	28.3	25.58
Coastguard	CIF	21.04	20.73
Foreman	CIF	22.05	19.09

在随机误码环境下(BER 10^{-3}),采用数据分类抗误码方法比不采用数据分类的编码模式平均 PSNR 高 2 dB。

5.1 不同打包长度的数据分类抗误码模式增益比较

数据打包长度采用表 3:

表 3 不同打包长度中冗余信息与有效荷载的比例

数据包长度(Kbits)	0.1	1	10
冗余信息/有效荷载(%)	48.3	5.71	0.34

在图 8 中,可以清晰地看到在编码 300 帧的序列中,采用数据分类抗误码保护的序列质量明显好于非抗误码模式的序列质量。而在不同的数据包尺寸的情况中,数据包尺寸越小,冗余信息在数据包中所占的比重越大,抗误码效果也越好。当数据包大小为 10 kbits,几乎包括整个一帧的所有宏块,因此其序列质量基本接近非抗误码模式下的效果。

(收稿日期:2004 年 2 月)

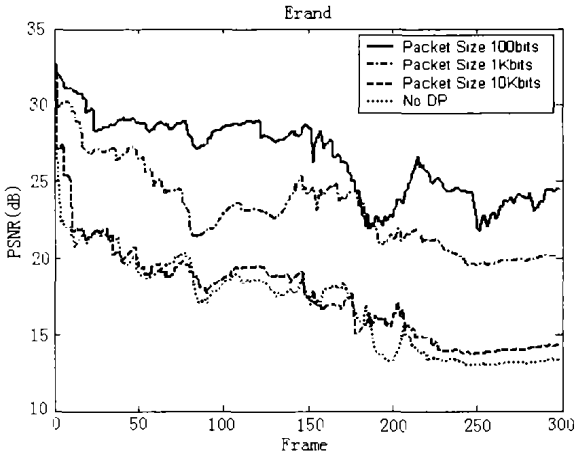

图 8 数据包尺寸对抗误码性能的影响

参考文献

1.Rai Talluri,Texas Instruments.Error-Resilient Video Coding in the ISO MPEG-4 Standard[J].IEEE Communications Magazine,1998

2.ITU-T Rec H 263.VeRMion 2,Video Coding for Low bitrate Communication[S].1998

3.R Talluri et al.Error Concealment by Data Partitioning[J].to appear in Sig Processing:Image Commun,1998

4.Takashima Wada,Murakami.ReveRMible Variable Length Codes[J].IEEE Trans on Commun,1995;43(2/3/4):158~162

5.ISO/IEC 14496-2.Information technology - Coding of audio-visual objects-Part 2 Visual[S].Second edition,2001-12

(上接 134 页)

algorithm is the slowest one,but compared with the k-means, the increasing gradient of its time consumption decreases with the growth of data set.We have further noted that the Cl.Err of HAC is apparently smaller than the other two.And it changes little with the growth of data sets.The HAC is a high accuracy clustering method.The Cl.Err of CSI is approximately equivalent to that of k-means.It is a little higher than k-means's on small data set,but a little lower on big data set, which indicates CSI has a slower increasing gradient of Cl.Err than k-means.

4 Conclusion

The research presents here an improved heuristic ant-clustering algorithm (HAC).A device of "memory bank"is proposed,which can bring forth heuristic knowledge guiding ant moving in the bi-dimension grid space.Hence the randomness of ants' moves decreases and the algorithm's convergence speeds up.In addition,the memory bank makes it possible for every object to be inspected before termination, which avoids the producing of"un-assigned data object".So the algorithm's misclassification error rate drops subsequently.

Experimental results demonstrate that HAC is a viable and effective clustering algorithm.Presently we have applied the HAC to the special data-mining model of artillery battlefield selecting in the field operation assistant decision system and got satisfied outcome.

References

1.Deneubourg J L,Goss S,Frank N et al.The dynamics of collective sorting:Robot-like ants and ant-like robots[C].In:Proc of the 1st International Conference on Simulation of Adaptive Behavior:From Animals to Animats,Cambridge,MA:MIT Press/Bradford Books,1991:356~363

2.Lumer E,Faieta B.Diversity and adaption in populations of clustering ants[C].In:Proc of the Third International Conference on Simulation of Adaptive Behavior:From Animals to Animats 3,Cambridge,MA:MIT Press,1994:501~508

3.P Kuntz,D Snyers,P Layzell.A stochastic heuristic for visualizing graph clusters in a bi-dimensional space prior to partitioning[J].Journal of Heuristics,1998;5(3):327~351

4.K Hoe,W Lai,T Tai.Homogenous ants for web document similarity modeling and categorization[C].In:Proc of the 3rd International Workshop on Ant Algorithms(ANTS 2002),Springer-Verlag,Berlin,Germany, vol 2463 of LNCS,2002:256~261

5.B Wu,Y Zheng,S Liu et al.SIM:A Document Clustering Algorithm Based on Swarm Intelligence[C].In:Proc of the IEEE World Congress on Computational Intelligence,Hawaiian,2002:477~482

6.Nicolas Monmarché,Mohamed Slimane,Gilles Venturini.AntClass:discovery of clusters in numeric data by an hybridization of an ant colony with the Kmeans algorithm[R].Internal Report,1999:(213,E3i)