广西科学院学报
Journal of Guangxi Academy of Sciences

2008,24(4):284~286
Vol. 24, No. 4 November 2008

一种新的 K-Means 蚁群聚类算法
A New AntClust Algorithm Based on K-Means Algorithm

莫锦萍，陈　琴，马　琳，苏一丹
MO Jin-ping, CHEN Qin, MA Lin, SU Yi-dan

(广西大学计算机与电子信息学院，广西南宁　530004)
(School of Computer, Electronics and Information, Guangxi University, Nanning, Guangxi, 530004, China)

摘要：针对蚁群聚类算法聚类质量不高的原因，使用 K-Means 算法改进蚁群聚类规则，提出一种新的 K-Means 蚁群聚类算法(KM-AntClust)，并通过实验验证新算法的聚类效果。实验结果表明，新的算法可以明显提高聚类质量。

关键词：聚类　蚁群算法　K-平均算法
中图法分类号：TP301　**文献标识码**：A　**文章编号**：1002-7378(2008)04-0284-03

Abstract: Due to the low clustering quality of AntClust algorithm, an improved AntClust algorithm is proposed, which optimize the rules of AntClust model with K-Means mind, is called KM-AntClust. Then the clustering effection is verified by experiments. Experimental results demonstrate that the new algorithm can significantly improve the quality of clustering.

Key words: clustering, ant colony algorithm, K-Means

聚类是数据在算法的指导下进行无人监督的分类。以 K-Means[1]和 K-Medoid[2]为代表的划分法是常用聚类算法中的一种。常用聚类算法多面向数值属性，而蚁群聚类算法(AntClust)[3,4]能处理任意类型的数据，具有强鲁棒性和适应性；但是其聚类结果随机，受数据和参数影响较大，聚类质量不高。本文使用 K-Means 算法思想改进蚁群聚类算法，提出一种新的 K-Means 蚁群聚类算法(KM-AntClust)，并在 UCI 数据集上对新算法的聚类效果进行测试。

1　传统蚁群聚类算法聚类质量分析

蚁群聚类方法是利用化学识别系统原理来聚类，它不需假设对象的表示，仅用相似度 $sim(i,j)$ 表示对象 i 和 j 的关系[3,4]。每只人工蚁均有标签 $Label_i$、基因 $Genetic_i$、和模板 $Template_i$[5]以及两个判断参数 M_i、M_i^+。算法规则如下：(1)两只无巢蚂蚁相遇时创建一个新巢；(2)无巢蚂蚁与一有巢蚂蚁相遇则将无巢蚂蚁归到对方所属巢中；(3)两只同巢蚂蚁 i、j 相遇，若相互接受，则增大 M_i、M_j 和 M_i^+、M_j^+ 值；(4)两只同巢蚂蚁 i、j 相遇，若互不接受，则减小 M_i、M_j 和 M_i^+、M_j^+ 值并将 M^+ 值小的蚂蚁移出巢；(5)两只不同巢蚂蚁相遇并相互接受则将两巢合并；(6)若不出现以上各情况，则不做任何操作。

蚁群聚类算法具有较好的鲁棒性和适应性，但是其聚类结果不稳定，主要原因是：第一，规则(3)依据 M_i^+ 判断踢出蚂蚁，但是根据算法思想，M_i^+ 为随机量，其值不仅与蚂蚁所属巢规模有关，还与循环次数相关。M_i^+ 是蚂蚁 i 被巢成员接受的程度，而不是反应蚂蚁与巢的依存关系，M_i^+ 大并不能说明此巢是蚂蚁 i 的最优归属，故依此踢出蚂蚁产生累积误差导致聚类质量降低。第二，循环迭代参数 $Iter$ 和删除概率 $Pdel$ 的设置不尽合理。如果循环次数 $NB_{Iter}=Iter * N$[5]不足，数据覆盖率低；太大则导致过学习。参数 $Pdel$ 太大聚出的类数目较少，相反则类过多，影响聚类质量。因此，参数 $Iter$ 和 $Pdel$ 的确定是保证聚类质量的重要环节。

收稿日期：2008-10-12
作者简介：莫锦萍(1984-)，女，硕士研究生，主要从事电子商务研究。

2 K-Means 蚁群聚类算法

2.1 使用 K-Means 改进蚁群聚类规则

K-Means 聚类算法基于误差平方和最小准则,聚类结果通常不受初始中心的影响,较为稳定。对于大数据集,其强伸缩性和高效性常使聚类结果以局部最优结束。因此我们引入 K-Means 算法改进聚类规则。

设 d_i 为蚂蚁 i 到其所属巢中心的距离,规则改进如下:(1)两只无巢蚂蚁 i,j 相遇时创建一新巢并计算巢中心;(2)无巢蚂蚁 i 与有巢蚂蚁 j 相遇则将蚂蚁 i 归到蚂蚁 j 所属巢中并更新该巢中心;(3)同巢互不接受的两只蚂蚁 i,j 相遇时,计算 d_i、d_j,将 d 值大的蚂蚁踢出巢并更新巢中心;(4)两只不同巢的蚂蚁相遇且相互接受时,将两巢合并并更新巢中心;(5)若不出现以上各情况,则不做任何操作。

2.2 K-Means 蚁群聚类算法

模拟蚂蚁相遇过程完成后,蚁群聚类算法通过将无巢蚂蚁分配到有巢且与其最相似的蚂蚁所属巢中来完成重新分配操作。有巢这一前提限制了所找到的蚂蚁实际上不一定最相似,这将导致聚类质量下降。因此用 K-Means 算法将蚂蚁分配到巢中心与其距离最小的巢来优化此操作。

设相遇过程完成后无巢蚂蚁个数为 q,存留的大类个数为 ClassNum,Class[ClassNum],蚂蚁总数为 N。K-Means 蚁群聚类算法描述为:

```
MinDisClassl(i , Class[]) /* 求巢中心距蚂蚁 i 最近的巢 */
  Clust()
    {
  for(r=1; r<=Iter * N; r++)
        {    i=rand()%n;
             j=rand()%n;
        Cluster according to the rules in Sect 2.2
    }
  Reassign()
  {
  for(i=1; i<q;i++)
  Label_i=MinDisLabel(i , Class[]);
  }
```

3 聚类实验

3.1 实验平台、数据集及度量标准

实验平台:PC (配置 Pentium 4,CPU 2.4GHz,内存 512M),操作系统是 Windows XP。算法使用 VC 编写。

数据集采用 UCI 公共数据库(http://kdd.Ics.Uci.edu)提供的数据集(见表 1)。

表 1 UCI 数据集

名称	分类	属性总数	样本总数
Irise	3	4	150
Breast-cancer	2	10	178
Wine	3	13	683

聚类性能评价采用文献[6]中介绍的 F-measure 方法。F-measure 组合了信息检索中查准率(precision)和查全率(recall)的思想。一个聚类 j 相对于分类 i 的 precision 和 recall 定义为:$P=\text{precision}(i,j)=N_{ij}/N_i$,$R=\text{recall}(i,j)=N_{ij}/N_j$,其中 N_{ij} 是在聚类 j 中分类 i 的数目,N_j 是聚类 j 所有对象的数目,N_i 是分类 i 所有对象的数目。分类 i 的 F-measure 定义为:$F(i)=2PR/(P+R)$。

对分类 i 而言,哪个聚类的 F-measure 值高,就认为该聚类代表分类 i 的映射,即 F-measure 表示分类 i 的评判分值。对聚类结果 λ 来说,其总 F-measure 可由每个分类 i 的 F-measure 加权平均得到:$F_\lambda=\dfrac{\sum_i(|i|*F(i))}{\sum_i|i|}$,其中 $|i|$ 为分类 i 中所有对象的数目。

3.2 实验步骤

步骤 1 使用 Breast-cancer 数据集进行参数训练,训练结果如图 1 和图 2 所示。从图 1 和图 2 结果可知,当 $Pdel=0.06$,$Iter=60$ 时聚类取得最佳效果,故实验中我们取 $Pdel=0.06$,$Iter=60$。

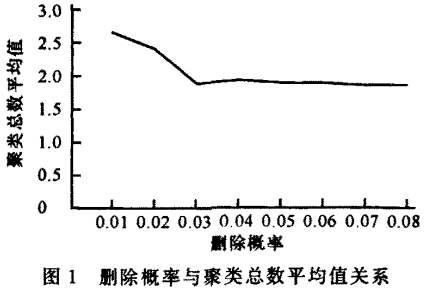

图 1 删除概率与聚类总数平均值关系

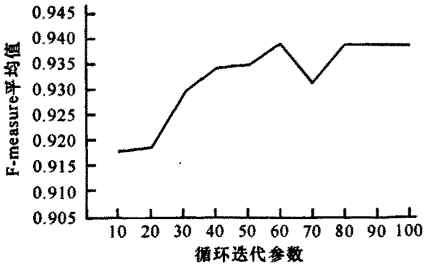

图2 循环迭代参数与F-measure平均值关系

步骤2 用AntClust算法对UCI数据集进行聚类。

步骤3 用改进的蚁群聚类算法(K-Means蚁群聚类算法)对UCI数据集进行聚类。

步骤4 结果比较与分析。

3.3 实验结果及分析

两个算法均运行10次后取F-measure平均值作为实验结果,结果如表2所示。

表2 UCI数据集聚类结果

数据集	聚类总数平均值		F-measure平均值	
	AntClust算法	K-Means蚁群聚类算法	AntClust算法	K-Means蚁群聚类算法
Irise	2.3	2	0.951805	0.988722
Breast-cancer	2.5	1.9	0.909741	0.935299
Wine	4.4	3.5	0.639769	0.715268

文献[7]中说明数据集Iris用于聚类时可以作两类处理。从表2可知,K-Means蚁群聚类算法获得更好的聚类效果,其F-measure平均值均比蚁群聚类算法的高,最高达到了0.988722;另外,改进算法聚类总数平均值也比蚁群聚类算法的更接近实际分类数。在实验中虽然聚类质量均有提高,但是有的效果并不明显,如表2中的Wine数据集,这主要是数据集中类与类间的差别不大,数据交叉重合比较多导致聚类效果不明显,聚类质量不高。

4 结束语

本文将K-Means聚类思想引入蚁群聚类算法中,并在UCI数据集上进行实验。实验结果表明,K-Means优化的蚁群聚类算法从距离角度反应蚂蚁与巢成员的接受程度,使得聚类有更加合理的判断依据,聚类质量得到了进一步的提高。

参考文献:

[1] MacQueen J. Some methods for classification and analysis of multivariate observations: proc of the 5th Berkeley Symp on Math Statist[C]. 1967:281-297.

[2] Kaufman J, Rousseeuw P J. Finding groups in data:an introduction to cluster analysis[M]. New York: John Wiley & Sons, 1990.

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

[4] Nicolas Labroche, Nicolas Monmarché, Gilles Venturini. Web sessions clustering with artificial ants colonies. [EB/OL]. [2006-01-12]. http://www. hant. i. univtours fr/webhant/pub/LabMonVen03a. www. pdf.

[5] Nicolas Labroche, Nicolas Monmarché, Gilles Venturini. AntClust: ant clustering and web usage ming [C]. Genetic and Evolutionary Computation, 2003:25-36.

[6] Yang Y, Kamei M. Clustering ensemble using swarm intelligence: IEEE Swarm Intelligence Symposium [M]. Piscataway, NJ: IEEE Service Center, 2003: 65-71.

[7] Parag M Kanade, Lawrence O Hall. Fuzzy ants as a clustering concept: proc of the 22nd International Conference of the North American Fuzzy Information Processing Society[C]. 2003: 227-232.

(责任编辑:韦廷宗)

瑞典研究显示高热量快餐食品易诱发老年痴呆症

老年痴呆症是发生在老年期及老年前期的一种原发性退行性脑病,指的是在没有意识障碍的状态下,记忆、思维、分析判断、情绪等方面出现障碍。瑞典卡罗林斯卡医学院的研究人员研究显示,高热量的快餐食品容易诱发老年痴呆症。研究人员在用脂肪、糖和胆固醇含量丰富的快餐食品喂食老鼠9个月后发现,老鼠脑中出现了类似老年痴呆症患者脑部的Tau蛋白纤维状缠结情况。Tau蛋白纤维状缠结是老年痴呆症患者典型的病理特征之一,这种蛋白纤维状缠结会影响患者的认知能力。研究人员指出,虽然上述成果对老年痴呆症的防治研究有启发,但是仍需进一步研究高热量快餐食品易诱发老年痴呆症的具体原因。

(据科学网)