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

2009,25(4):261~263
Vol. 25,No. 4 November 2009

一种基于序列相似性的蚁群聚类算法 SeqAntCluster
An Ant Colony Clustering Algorithm Based on Sequence Similarity: SeqAntCluster

马 凯,苏一丹,谭荣丽,梁胜勇
MA Kai,SU Yi-dan,TAN Rong-li,LIANG Sheng-yong

(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:将序列数据的相似度度量方法 S^3M 引入蚁群聚类算法中,提出一种基于序列相似性的蚁群聚类算法。该算法既继承了蚁群聚类算法原有的优点,又能有效地对序列数据聚类,更适合处理序列数据。

关键词:数据挖掘 蚁群聚类 序列 相似性
中图法分类号:TP311.13 **文献标识码**:A **文章编号**:1002-7378(2009)04-0261-03

Abstract: An ant colony clustering algorithm based on sequence similarity: SeqAntCluster was presented. A sequence similarity measure (S^3M) is introduced in Ant Colony Clustering Algorithm as similarity measure. The experimental results show SeqAntCluster can cluster sequence data effectively and of practical application value.

Key words: data minining, ant colony clustering, sequence, similarity

聚类就是将数据对象分组成为多个类,在同一个类中的对象之间具有较高的相似度,而不同类中的对象差别较大[1]。聚类在数据挖掘、统计学、机器学习、计算生物学以及商业智能等领域具有广泛的应用。

目前主要的聚类算法可以分为:划分的方法、层次的方法、基于密度的方法、基于网格的方法和基于模型的方法。随着技术的不断发展,一些新的生物启发聚类方法被相继提出,蚁群聚类算法便是其中的之一。蚁群聚类算法是1991年 Deneubourg 等[2]人根据某些种类的蚂蚁有规律垒堆尸体和将蚁卵按大小堆放的现象提出一种基于蚁群的聚类模型。Lumer 等[3]改进了 Deneubourg 的模型,提出 LF 算法并首次将其运用于数据分析。接着 Ramos[4]在 LF 的基础上又提出了新的蚁群聚类算法 Acluster,Acluster 蚁群聚类算法是对基本蚁群算法的改进并且在聚类质量和效率上都有很大改善。蚁群聚类算法具有灵活性、健壮性、分布性和自组织性等许多特性,这些特性使其非常适合本质上是分布、动态及需要交错的问题求解,如解决无监督的聚类问题。

许多领域涉及序列数据,例如 Web 使用挖掘的点击流数据,生物领域中的基因数据等。序列数据的一个共同特征就是数据本身无大小之分,只有序列内容和相对位置不同。传统方法使用欧几里德距离或向量夹角 cosine 作为相似性的度量,这些方法在相似度转换计算过程中丢失了许多序列信息,例如元素的先后顺序、相对位置等。张斌等[5]针对方向性数据的特点提出了一种基于方向相似性的蚁群聚类算法 ACCADS,该算法将方向性度量引入蚁群聚类算法作为相似性度量,有效地对高维方向数据聚类。本文将序列数据的相似度度量方法 S^3M[6]引入蚁群聚类算法中,提出一种基于序列相似性的蚁群聚类算法(SeqAntCluster)。此算法不但继承了蚁群聚类算法原有的优点,而且更适合处理序列数据。

1 SeqAntCluster 算法描述

SeqAntCluster 算法同其它各种聚类算法一样,按照不同策略划分簇,将具有某种相似程度的数据

收稿日期:2009-10-10
作者简介:马 凯(1978-),男,硕士研究生,主要从事数据库、电子商务研究。

划分到相应的簇里面，最终的实现都转化为数据之间的相似度计算问题。

1.1 序列数据的相似度度量

文献[5]的序列相似性度量方法 S^3M 同时考虑了序列与内容的相似性，由集合相似性和序列相似性两部分组成。

一个序列由有先后次序的项来构成，即有序的项集。

序列 S 定义为：$S=<a_1,a_2,\cdots,a_n>$，$a_1,a_2,\cdots,a_n$ 是序列中的有序的项集，序列的长度定义为：$|S|$；集合相似性定义为：

$$SetSim(A,B)=\frac{|A\cap B|}{|A\cup B|};$$

序列相似性定义为：

$$SeqSim(A,B)=\frac{LLCS(A,B)}{\max(|A|,|B|)},$$

$LLCS(A,B)$ 为 A、B 两个序列的最长公共子序列长度。序列 A 和 B 的相似度定义为：

$$S^3M(A,B)=p*\frac{LLCS(A,B)}{\max(|A|,|B|)}+q*\frac{|A\cap B|}{|A\cup B|},$$

其中 $p+q=1$ 而且 $p,q\geqslant 0$；p 和 q 决定了相关集合相似与序列相似的权重，可以根据不同的应用调整比例。

1.2 SeqAntCluster 算法描述

步骤1 初始化算法中蚂蚁个数 n，最大迭代次数 T，网格边长 M，邻域大小 s，参数 k1，k2等。

步骤2 根据公式计算数据对象的相似矩阵。

步骤3 将数据对象和蚂蚁随机分布到二维网格中。

步骤4 For t = 1 to T do

For i = 1 to n do

(1)Sum = 0; /* 计数器 */;

(2)计算蚂蚁 i 周围 s×s 的邻域内数据对象的数目；

(3)如果蚂蚁未负载而且其位置上有数据时，对其邻域内每个数据对象，在相似矩阵查找数据对的相似度值 sim，并计算 Pp。若 Pp 大于一个随机概率，计数器 Sum 加1，如果最终 Sum>=n/2或 n=0，则蚂蚁拾起该对象并标记自己已负载；

(4)如果蚂蚁为负载状态而且其位置上没有数据时，对其邻域内每个数据对象，在相似矩阵查找数据对的相似度值 sim，并计算 Pd。若 Pp 大于一个随机概率，计数器 Sum 加1，如果最终 Sum>=n/2，则蚂蚁放下该数据并标记自己未负载；

(5)蚂蚁根据概率转换函数的选择移动到一个新位置，释放信息素；

(6)每个格子的信息素以一定的速率蒸发。

步骤5 输出聚类结果。

2 SeqAntCluster 算法的实验及结果分析

2.1 实验平台、数据集及度量标准

实验平台为 PC 机(配置 Pentium 4，CPU 3.0 GHz，内存1GB)，操作系统是 Windows Server 2003，算法使用 Vb. net 2005实现。数据集采用 UCI 公共数据库(http://kdd. ics. uci. edu/) 提供的 MSNBC 数据集，该数据集收集自 msn. com 的服务器的 log 文件，包含989818条用户 session，平均长度为5. 7，session 序列中的每项代表用户请求的一个页面。我们抽取200条长度为6的 session 作为实验数据。

聚类性能的评价采用文献[7]介绍的差值平方和(SSE：sum of squared error)和文献[8]介绍的 Levensthein 距离(LD：Levensthein distance)。LD 衡量了从一个序列转换成另外一个序列的所需基本操作的最小代价，这些基本操作包括：插入、删除、替换3种。LD 越大表示序列间相互转换需要的操作越多，它们之间的相似度越低。平均 Levensthein 距离通常用来反映序列聚类的优劣程度。平均 Levensthein 距离 (ALD) 定义如下：

$$ALD=\frac{1}{k}\sum_{j=1}^{k}\frac{\sum_{j_s}^{C_j}LD(\hat{t}_j,t_{j_s})}{|C_j|},$$

其中 k 是总的聚类个数；$|C_j|$ 是第 j 个聚类的项的个数；$LD(\hat{t}_j,t_{j_s})$ 是第 j 个簇中的第 s 个元素和簇中心 $\hat{t}$ 之间的 levensthein 距离。

2.2 实验步骤

实验的第1步用 cosine 作为相似度的蚁群聚类算法(AntCluster 算法)对 UCI 数据集进行聚类；第2步用改进的使用 S^3M 作为相似度的蚁群聚类算法(SeqAntCluster 算法)对 UCI 数据集进行聚类；两步结果进行比较与分析。

2.3 实验结果及分析

从表1和表2对比得知使用 S^3M 相似度的 SeqAntCluster 算法聚类时的簇内 SSE 值小于使用 cosine 的 AntCluster 算法，这表明使用 S^3M 相似度的 SeqAntCluster 算法形成聚类的簇具有更高的相似度，聚类效果优于使用 cosine 相似度的 AntCluster 算法。从表3中可知使用 S^3M 相似度的 SeqAntCluster 算法聚类的 ALD 值小于使用 cosine

的 AntCluster 算法的值。因此用户 session 序列使用 S^3M 相似度聚类形成的簇比使用 cosine 能保留更多的序列信息,聚类的效果更好。

表1 AntCluster 算法的簇内 *SSE* 值

	C1	C2	C3	C3
C1	0	0.83	0.14	0.87
C2	0.83	0	0.72	0.21
C3	0.14	0.72	0	0.85
C4	0.87	0.21	0.85	0

表2 SeqAntCluster 算法的簇内 *SSE* 值

	C1	C2	C3	C3
C1	0	0.73	0.15	0.68
C2	0.73	0	0.52	0.21
C3	0.15	0.52	0	0.25
C4	0.68	0.21	0.25	0

表3 平均 levensthein 距离(*ALD*)

	AntCluster 算法	SeqAntCluster 算法
C1	4.52	4.43
C2	4.93	4.68
C3	5.56	3.41
C4	4.87	3.82
ALD	4.97	4.09

3 结束语

基于序列相似性的蚁群聚类算法将序列相似性度量引入蚁群聚类算法,除了继承原蚁群聚类算法的优点之外,更适合处理序列数据。实验结果表明,该算法能有效地对序列数据进行聚类,具有一定的实际应用价值。

参考文献:

[1] 韩家炜. 数据挖掘:概念与技术[M]. 北京:机械工业出版社,2006.

[2] Deneubourg J L,Goss S,Franks N. The dynamics of collective sorting: robot-like ant and ant-like robot: proceedings First Conference on Simulation of Adaptive Behavior:from Animals to Animats[C]. 1991:356-363.

[3] Lumer E D,Faieta B. Diversity and adaptation in populations of clustering ants:proceedings of the Third International Conference on Simulation of Adaptive Behaviour:From Animals to Animats 3[C]. 1994:501-508.

[4] Ramos V,Almeida F. Artificial ant colonies in digital image habitats-a mass behavior effect study on pattern recognition[C]. 2000:113-116.

[5] 张斌,苏一丹. 一种基于方向相似性的蚁群聚类算法 ACCADS[J]. 计算机与现代化,2008,151(3):86-89.

[6] Kumar P. Intrusion detection system using sequence and set preserving metric:proceedings of IEEE International Conference on Intelligence and Security Informatics [C]. 2005:498-504.

[7] Duda R O,Hart P E,Stork D G. Pattern classification [M]. New York:John Wiley and Sons,2001.

[8] Levenshtein L I. Binary codes capable of correcting deletions, insertions and reversals [J]. Soviet Physics-Doklady,1966,10:707-710.

(责任编辑:邓大玉)

(上接第258页)

索引进行 Huffman 编码。仿真实验表明,这种方法可以获得较高的数字图像压缩比。在实验过程中,我们发现 LBG 算法有计算量大和对初始码书敏感等缺点,这有待作进一步改进。

参考文献:

[1] 孙圣和,陆哲明. 矢量量化技术及应用[M]. 北京:科学出版社,2002.

[2] 马平. 数字图像处理和压缩[M]. 北京:电子工业出版社,2007.

[3] 张春田,苏育挺,张静. 数字图像压缩编码[M]. 北京:清华大学出版社,2006.

[4] 赵春晖,陈万海,张凌雁. 一种基于矢量量化的高光谱遥感图像压缩方法[J]. 哈尔滨工程大学学报,2006,27(3):447-449.

(责任编辑:韦廷宗)