

一种基于特征加权的蚁群聚类新算法

李玲娟, 李 冰

(南京邮电大学 计算机学院, 江苏 南京 210003)

摘 要:蚁群聚类算法作为一种群体智能的算法已经被证实可用于高维数据的聚类,能够快速有效地处理 Web 的海量、高维数据,但是传统的蚁群聚类算法并未考虑各维特征的贡献率,聚类的准确度有限。文中以优化聚类效果为目标,提出了一种基于特征加权的蚁群聚类新算法 FWACCA,在新算法中考虑了各维特征对分类贡献的多少,合理地使用了 Sigmoid 概率转换函数和主客观结合的赋权法。实验结果表明此新算法可以有效减少聚类出错率,提高聚类的准确性。

关键词:蚁群聚类;特征加权;概率转换

中图分类号:TP391

文献标识码:A

文章编号:1673-629X(2010)08-0067-04

A New Ant Colony Clustering Algorithm Based on Feature Weight

LI Ling-juan, LI Bing

(College of Computer, Nanjing University of Posts and Telecommunications, Nanjing 210003, China)

Abstract: It is approved that the ant colony clustering algorithm, as a kind of swarm intelligence algorithm, can be used in clustering the high-dimension data. It can quickly and efficiently deal with mass and high-dimensional Web data. But the contribution of each dimension does not take into account in the traditional ant colony clustering algorithm, and its clustering accuracy is limited. In order to get better clustering effect, presents a new ant colony clustering algorithm based on feature weight, which is named FWACCA. The algorithm considers the contribution degrees from different features, and properly applies the probability transition function named Sigmoid and both the subjective feature weighting method and the object feature weighting method. The experiment results show that the algorithm can reduce mistakes and enhance the clustering veracity.

Key words: ant colony clustering; feature weighting; probability transition

0 引言

聚类研究已经有很长的历史,几十年来,其重要性及与其他研究方向的交叉特性得到了人们的肯定。聚类是数据挖掘研究方向的重要研究内容之一^[1],事实上,它是一个无监督的分类,它没有任何先验知识可用。传统的聚类算法包括划分方法、层次方法、基于密度的方法、基于网格的方法和基于模型的方法等^[2],这些聚类方法在处理高维的 Web 数据时聚类效果并不理想。

蚁群聚类算法是近年来诞生的一种基于群体智能的算法^[3],在某些方面更加接近实际聚类问题,跟其他传统聚类方法相比,它能够快速有效地处理 Web 的海量、高维数据,并且不需要用先验知识决定输入参数

(簇的数目),在处理形式上也比较直观,便于理解,因而正在被广泛地研究与应用。

但是,现有的蚁群聚类算法并没有考虑各维特征对分类贡献的多少,而是假定了其贡献的均匀性。鉴于此,文中提出了一种基于特征加权的蚁群聚类新算法,将特征权重考虑进去,以期进一步完善已有的蚁群聚类算法。

1 蚁群聚类算法

蚁群聚类算法(Ant Colony Clustering Algorithm, ACCA)方法从原理上可以分为四种:

- (1)运用蚂蚁觅食的原理,用信息素来实现聚类^[4,5];
- (2)利用蚂蚁自我聚集行为聚类;
- (3)基于蚂蚁堆的形成原理实现数据聚类;
- (4)运用蚁巢分类模型,用蚂蚁化学识别系统进行聚类。

收稿日期:2009-12-07;修回日期:2010-03-08

基金项目:国家自然科学基金(60863001)

作者简介:李玲娟(1963-),女,辽宁辽阳人,教授,研究方向为数据挖掘、网络安全等。

文中的研究主要基于上述第三种方法。与此方法有关的研究和原理如下:

昆虫学家通过观察发现,很多种类的蚂蚁在打扫它们的巢穴时都会把尸体堆积在一起形成坟墓。Deneubourg 等曾用 *Pheidole pallidula* 蚂蚁做过蚁群构造墓地的试验^[6],随后他提出了一种解释上述蚁群聚类现象的基本模型 BM^[7]。这种模型主要是对蚂蚁拾起、移动、放下的行为方式进行建模。当一只任意移动的无负载的蚂蚁遇到一个物体时,如果该物体附近与之相同的物体越少,则拾起这个物体的概率越大;反之,放下这个物体的概率越大。一只随机运动的未负载的蚂蚁拾起一个物体的概率定义为:

$$P_p = \left(\frac{k_1}{k_1 + f} \right)^2 \quad (1)$$

其中, f 表示蚂蚁周围物体的感知因子(在机器人实现中定义为 $f = N/T$, 即 T 时间内遇到的物体个数 N), k_1 是阈值常数。当 $f \ll k_1$, 则 P_p 接近于 1, 即当周围没有多少物体时, 拾起一个物体的概率很大; 当 $f \gg k_1$, 则 P_p 接近于 0, 即在一个密集聚类中, 一个物体被移走的概率很小。一只随机运动的有负载的蚂蚁放下一个物体的概率定义为:

$$P_d = \left(\frac{f}{k_2 + f} \right)^2 \quad (2)$$

其中, k_2 是另一个阈值常数。当 $f \ll k_2$, 则 P_d 接近于 0; 而当 $f \gg k_2$, 则 P_d 接近于 1。放下行为大致遵守与拾起行为相反的规则。

Lumer 和 Faieta 将 Deneubourg 的基本模型推广应用到数据分析中, 提出了著名的 LF 算法^[8]。首先将 BM 中的两物体之间的二进制距离扩展到更多属性的物体和更复杂的距离, 其次将待聚类数据对象随机地投影到低维空间, 通常是一个平面, 每个单元只含有一个对象。然后将蚂蚁分布到这个空间内, 并以随机方式移动。假设在时刻 t , 一只蚂蚁在地点 r 发现一个数据对象 o_i , 则对象 o_i 与其邻域内所有对象的平均相似性定义为:

$$f(o_i) = \max \left\{ 0, \frac{1}{s^2} \sum_{o_j \in \text{Neigh}_{s \times s}(r)} \left[1 - \frac{d(o_i, o_j)}{\alpha(1 + ((v-1)/v_{\max}))} \right] \right\} \quad (3)$$

其中 α 为衡量相异度的参数, $d(o_i, o_j)$ 是两个对象 o_i 和 o_j 的欧式距离^[9], 表示两个对象的相似程度, $d(o_i, o_j)$ 的值在 $[0, 1]$ 之间, 对象间越相似, $d(o_i, o_j)$ 值越小, $f(o_i)$ 值越大; 反之, 对象越相异, $d(o_i, o_j)$ 值越大, $f(o_i)$ 值越小。 v 表示蚂蚁运动的速度(这里指的是一个时间单元内沿给定的网格轴线行走的网格单元数),

v_{\max} 为最大速度, $\text{Neigh}_{s \times s}(r)$ 表示地点 r 周围的以 s 为边长的正方形局部邻域。

拾起和放下概率定义为:

$$P_{p(o_i)} = \left(\frac{k_1}{k_1 + f(o_i)} \right)^2 \quad (4)$$

$$P_{d(o_i)} = \begin{cases} 2f(o_i), & \text{当 } f(o_i) < k_2 \text{ 时} \\ 1, & \text{当 } f(o_i) \geq k_2 \text{ 时} \end{cases} \quad (5)$$

其中的 k_1 和 k_2 为常数, 其作用与 BM 中的类似。

该算法实际上是一种基于网格和密度的聚类方法^[10]。为了便于处理高维数据空间, 将其映射到某一低维网格空间, 映射要确保簇内距离小于簇间距离, 同时网格的精细度将会影响聚类质量。

2 基于特征加权的蚁群聚类新算法(FWA-CCA)

在利用蚁群聚类算法进行聚类分析时, 总是假设特征提取是相当完善的, 构成模式矢量的特征是独立且无冗余的, 并且认为各维特征对分类的贡献是均匀的。而事实上, 构成样本特征矢量的各维特征来自不同的传感器, 存在量纲差异和精度以及可靠性的不同, 而且所选择的特征集也未必适合于模式的分类。鉴于此, 文中设计了基于加权的蚁群聚类新算法, 在该算法中考虑了各维特征对模式分类的不同贡献, 可以获得更有效的聚类分析结果。

目前关于权系数的确定方法有数十种之多, 根据计算权系数时原始数据的来源不同, 这些方法大致可分为两大类: 一类为主观赋权法, 其原始数据主要由专家根据经验主观判断得到, 如古林法、层次分析法、专家法等; 另一类为客观赋权法, 其原始数据由各指标在被评价单位中的实际数据形成, 客观赋权方法有很多种, 如 tf 算法、 idf 算法、 $\text{tf} * \text{idf}$ 算法和 $\text{mutual information}$ 算法等。较为常用的是 $\text{tf} * \text{idf}$ ^[11] 加权法, 也就是特征项频率 * 倒排文档频率加权法。

文中的 FWACCA 算法中采用的是主观方法与客观方法相结合的赋权法, 对于一些常识性的以及比较特殊的特征选项, 采用专家法直接赋值; 而对于大部分的其他特征选项则用客观权值赋值方法来进行计算, 借鉴了以上提到的文献^[11]的 $\text{tf} * \text{idf}$ 加权法。其权值计算公式为:

$$w_i = \text{tf}_i * \log_2(N/N_i + 0.01) \quad (6)$$

其中, tf_i 为第 i 个特征项出现的频次, N 为记录集中的数据总记录数, N_i 为出现第 i 个特征的数据记录数。

为了减少每个特征权值取值的不同给聚类造成的影响, 通常要将每个权值向量归一化到单位向量, 最后得到的权值计算公式如下:

$$w_i = \frac{tf_i * \log_2(N/N_i + 0.01)}{\sqrt{\sum_{k=1}^n (tf_{ki})^2 * [\log_2(N/N_i + 0.01)]^2}} \quad (7)$$

这样一来,原来对象属性间的距离 $d(o_i, o_j)$ 就要因权值因素的影响而有所变化,我们定义变化后的距离为 $d^*(o_i, o_j)$ 。在进行数据集聚类划分之前,必须对数据集进行预处理。假定一个网格内的数据集为 $O = \{o_1, o_2, \dots, o_n\}$, 包含 n 个样本,每一个样本为一个 s 维的矢量,即 $o_k = (o_{k1}, o_{k2}, \dots, o_{ks})$, 权值 $w = \{w_1, w_2, \dots, w_s\}$ 。那么预处理即为特征加权:

$$O'_k = w \cdot o_k \quad (8)$$

预处理后任意两个特征矢量的欧式距离可表示为

$$d^*(o_i, o_j) =$$

$$\sqrt{w_1 |o_{i1} - o_{j1}|^2 + w_2 |o_{i2} - o_{j2}|^2 + \dots + w_s |o_{is} - o_{js}|^2} \quad (9)$$

在 LF 算法中,概率转换函数运算需要两个参数 (k_1 和 $f(o_i)$), 其中 k_1 的选择对于概率的计算有着至关重要的作用,稍有不慎就会对聚类的效果产生很大影响。为此,选取对称式 Sigmoid 函数作为概率转换函数^[12],它在运算中只需调整一个参数,比 LF 算法中的公式(4)和(5)更简洁。

由此给出蚂蚁的拾起概率:

$$P_p = 1 - \text{sigmoid}(f(o_i)) \quad (10)$$

放下概率:

$$P_d = \text{sigmoid}(f(o_i)) \quad (11)$$

其中:

$$\text{sigmoid}(x) = \frac{1 - e^{bx}}{1 + e^{-bx}} \quad (12)$$

Sigmoid(x) 为自然指数形式,参数 b 越大,曲线饱和和越快,算法收敛速度也越快。

基于以上分析,可将文中设计的基于特征加权的蚁群聚类算法步骤描述如下:

1) 初始化算法中的蚂蚁个数 m , 对象个数 n , 最大迭代次数 T , 网格边长 Z , 局域边长 s 及其他参数 k_1 、 v 、 v_{\max} 、 b 等;

2) 将数据集集中的数据记录进行数据预处理,即按公式(7)进行特征向量权值运算;

3) 将数据对象和蚂蚁随机投影到一个二维网格中,一个网格中只允许放一个对象;

4) 重复迭代以下过程:

For $t = 1$ to T

{

For $i = 1$ to m

{

计算蚂蚁 i 周围边长为 s 的邻域内数据对象的数目以

及 $d^*(o_i, o_j)$

If 蚂蚁未负载且其位置上有数据

Then 按公式(10) 计算拾起概率 P_p

{ if P_p 大于一个随机概率,而同时该对象未被其他蚂蚁拾起

then 蚂蚁拾起该对象,随机移往别处,并标记自己已负载

else 蚂蚁拒绝拾起该对象,而随机选择其他对象

}

Else if 蚂蚁为负载状态

Then 按公式(11) 计算放下概率 P_d

{

if P_d 大于一个随机概率

then 蚂蚁放下该对象,并标记自己未负载,再重新选择一个新对象

else 蚂蚁拾起该对象继续移动到一个新位置

}

//for i

//for t

5) for $i = 1$ to n //对所有对象进行标记

{

if 一个对象是孤立的或它的邻域对象个数小于某一常数

then 标记该对象为孤立点

else 给该对象分配一个聚类序列号,并递归地将其邻域对象标记为同样的序列号

}

3 实验及结果分析

为了检验文中提出的基于特征加权蚁群聚类新算法的聚类准确度,基于人工采集的有关动物的分类数据进行了聚类实验。实验用的数据集共有 100 个记录,用 a_i 表示(i 为 1 到 100 的整数)。为了使结果更加明显,选取了三类动物:哺乳类 A、鸟类 B 和鱼类 C,包括老虎、狮子、斑马、天鹅、喜鹊、鲤鱼、鲸鱼等;每个记录由 10 个属性特征组成,如是否有鳍(isfin)、是否有翅膀(iswing)、是否用肺呼吸(breath)、水生与否(iswater)、有无脊椎(ischine)、恒温与否(temperature)、是否卵生(isegg)等等,而这其中除“是否用肺呼吸”、“是否卵生”和“是否有翅膀”的权值利用主观赋值法赋为 0.8 外,其他权值根据公式(7)计算出来。

分别利用 LF 算法和 FWACCA 算法进行聚类,结果分别如图 1 和图 2 所示。

由图中给出的实验结果可以看出,如果利用 LF 算法,鲸鱼就被聚类到了鱼类,而用 FWACCA 算法,鲸

鱼就可以准确地分到哺乳一类,而且新算法聚类效果更加明显,结果簇内的相似度较高,而簇间的相似度较低。

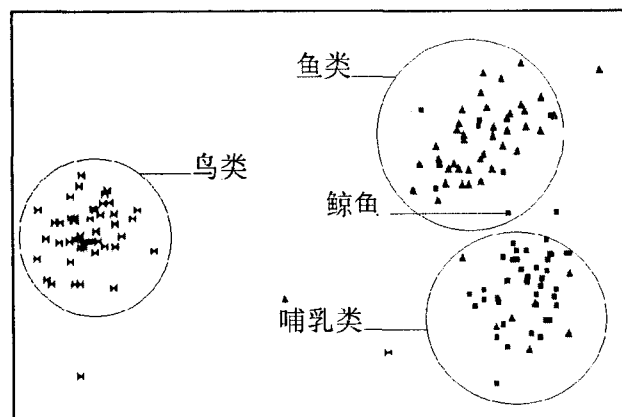


图 1 LF 算法聚类结果

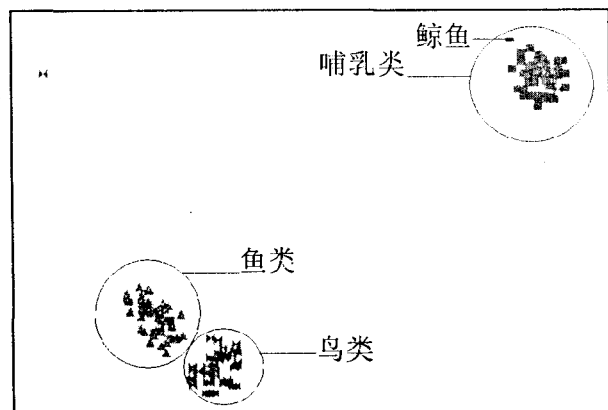


图 2 FWACCA 算法聚类结果

4 结束语

文中设计的蚁群聚类新算法,通过合理地使用概率转换函数 Sigmoid 函数和主客观结合的权值赋值法,弥补了已有的蚁群聚类算法的不足,能优化高维数据的聚类效果,这将对 Web 高维数据的聚类分析起到

有益作用。

参考文献:

- [1] 孙吉贵,刘洁,赵连宇. 聚类算法研究[J]. 软件学报, 2008,19(1):48-61.
- [2] 韩家伟,堪博. 数据挖掘:概念与技术[M]. 第2版. 范明,孟小峰译. 北京:机械工业出版社,2007.
- [3] Dorigo M, Bonabeau E, Theraulaz G. Ant algorithms and stigmergy[J]. Future Generation Computer Systems, 2000, 16: 851-871.
- [4] 杨新斌,孙京浩,黄道. 一种进化聚类学习新方法[J]. 计算机工程与应用, 2003,39(15):60-62.
- [5] 张惟皎,刘春煌,尹晓峰. 蚁群算法在数据挖掘中的应用研究[J]. 计算机工程与应用, 2004,40(28):197-193.
- [6] Bonabeau E, Dorigo M, Theraulaz G. Swarm intelligence - from natural to artificial system[M]. New York: Oxford University Press, 1999.
- [7] Deneubourg J L, Goss S, Franks N, et al. The dynamics of collective sorting: robot-like ant and ant-like robot[C]// Proceedings of the first conference on simulation of adaptive behavior: from animals to animats. Cambridge, MA: MIT Press, 1991:356-365.
- [8] Lumer E, Faieta B. Diversity and adaptation in populations of clustering ants[C]// Proceedings of the third international conference on simulation of adaptive behavior: from animals to animats. Cambridge, MA: MIT Press, 1994:499-508.
- [9] 张建华,江贺,张宪超. 蚁群聚类算法综述[J]. 计算机工程与应用, 2006(16):171-174.
- [10] 杨欣斌,孙京浩,黄道. 基于蚁群聚类算法的离群挖掘方法[J]. 计算机工程与应用, 2003,39(9):12-14.
- [11] Salton G, Buckley B. Term-Weighting approaches in automatic text retrieval[J]. Information Processing and Management, 1998,24(5):513-523.
- [12] Platt J C. Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods[M]. [s.l.]: MIT Press, 1999: 61-73.

(上接第 66 页)

- soaprdf/.

- [7] Li Yingjie, Yu Xueli, Geng Lili, et al. Research on Reasoning of the Dynamic Semantic Web Services Composition[C]// IEEE/WIC/ACM WI2006. [s.l.]:[s.n.], 2006.
- [8] Mallia A U, Singh M P. A Semantic Approach for Designing Business Protocols[C]// Proceedings of the 13th International World Wide Web Conference on Alternate Track Papers & Posters Table of Contents. New York, NY, USA: [s.n.], 2004:308-309.
- [9] Chen W, Mizoguchi R. Communication Content Ontology for Learner Model Agent in Multi-Agent Architecture[C]// Proceedings of AIED99 Workshop on Ontologies for Intelli-

gent Educational Systems. [s.l.]:[s.n.], 1999.

- [10] Tamma V, Wooldridge M, Dickinson I. An ontology-based approach to automated negotiation[C]// Proceedings of the IV Workshop on Agent Mediated Electronic commerce (AMEC IV). [s.l.]: Springer-Verlag, 2002:219-237.
- [11] 陈静,朱巧明,贡正仙. 基于 Ontology 的信息抽取研究综述[J]. 计算机技术与发展, 2007,17(10):84-86.
- [12] 孙志强. 多 Agent 系统通信机制的研究与相关运行支持环境的实现[D]. 北京:北京航空航天大学, 2004.
- [13] 耿丽丽,余雪丽. 基于 SOAP 的语义 WEB 服务通信协议协议的研究[D]. 太原:太原理工大学, 2003.