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

2010,26(4):406~408,411
Vol. 26, No. 4 November 2010

# 基于巢模板的核空间蚁群聚类算法
# Nest Template-Based Ant Clustering Algorithm in Kernel Space

覃 华,徐燕子,张 敏
QIN Hua, XU Yan-zi, ZHANG Min

(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**:为了改进蚁群算法因大量引入随机机制所引发的不稳定性,引入巢模板来改进聚类规则,提出一种基于巢模板的核空间蚁群聚类算法,并与原空间上的聚类算法进行比对。该算法用支持向量机的非线性映射函数把数据样本映射到核空间,再用巢模板记忆蚁群群体特征。核空间上的巢模板蚁群聚类算法能较好地处理特征复杂、类别多的数据集,其聚类结果比较接近真实情况,并且效果明显优于原空间上的聚类算法。

**关键词**:蚁群聚类 支持向量机 非线性映射函数 核函数 巢模板
**中图法分类号**:TP311 **文献标识码**:A **文章编号**:1002-7378(2010)04-0406-03

**Abstract**: If the features of data samples' are complex and with more categories, the ant clustering results are not satisfied. After the analysis of the main reasons, an idea that maps the data samples to kernel space by SVM' nonlinear mapping function is proposed. The features of data samples are recombined and highlighted in kernel space. The ant clustering algorithm is designed in kernel space and the nest template is been used to improve the stability and accuracy of algorithm. Experimental results on UCI datasets show that the clustering results of nest template ant clustering algorithm in kernel space are closer to the reality. The algorithm can proceed datasets which are complex and with more categories and the result is better than that in original space.

**Key words**: ant clustering, SVM, nonlinear mapping function, kernel function, nest template

蚁群聚类算法和传统的 K-means,C-means 等聚类算法相比,其主要特色在于它能利用群体智能来对任意的数据进行聚类,不需要事先设定簇的个数以及初始簇的中心,给实际应用带来了很大的方便[1,2]。但是如果数据的内在特征复杂,数据样本分布较为混乱,样本类别多、维数高时,蚁群聚类算法的聚类效果就会比较差[3~5]。其中关键的原因有:第一,复杂、高维、分布混乱的数据会直接导致空间距离不能很好地代表数据对象的本质特征,进而影响数据对象相似性的判断。第二,蚁群聚类过程中大量地采用随机机制,例如蚂蚁接受阈值的确定、蚂蚁的相遇等,再加上算法迭代参数 Iter 与删除概率 Pdel 的不确定性,导致算法不稳定。

由于支持向量机的非线性映射函数能把数据映射到一个高维的 Hilbert 空间(核空间),而在核空间中数据的特征会发生重组并凸显[1],所以核空间中数据的分布状态更接近于数据的客观事实,因此在核空间中对数据进行聚类,其效果要好于原空间[6~8]。

为了改进蚁群算法因大量引入随机机制所引发的不稳定性,本文引入巢模板来改进聚类规则,用巢模板来记忆蚁群群体特征,以群体特征弥补个体在特征识别时因随机性导致的不稳定性,提出一种基于巢模板的核空间蚁群聚类算法。

**收稿日期**:2010-09-02
**作者简介**:覃 华(1972-),男,副教授,主要从事最优化理论与数据挖掘研究。

## 1 核距离及基于巢模板的蚁群聚类规则

### 1.1 核空间中数据样本的欧氏距离公式

对于原空间上的两个 $p$ 维数据样本 $x=(x_1,x_2,\cdots,x_p)$ 和 $y=(y_1,y_2,\cdots,y_p)$,它们在原空间上的欧氏距离计算公式为

$$d(x,y)=\|x-y\|=\sqrt{\sum_{i=1}^{p}(x_i-y_i)^2}。\tag{1}$$

用支持向量机的非线性映射函数 $\varphi$ 把 $x$、$y$ 映射到核空间后,欧氏距离公式可写成向量内积形式

$$d(\varphi(x),\varphi(y))=\|\varphi(x)-\varphi(y)\|=\sqrt{(\varphi(x)-\varphi(y))\cdot(\varphi(x)-\varphi(y))}=\sqrt{\varphi(x)\cdot\varphi(x)-2\varphi(x)\cdot\varphi(y)+\varphi(y)\cdot\varphi(y)}。\tag{2}$$

其中的算符"·"表示向量的内积计算。

用支持向量机的核函数(SVM)[9]

$$K(x,y)=\varphi(x)\cdot\varphi(y)\tag{3}$$

改写(2)式得

$$d(\varphi(x),\varphi(y))=\sqrt{K(x,x)-2K(x,y)+K(y,y)}。\tag{4}$$

(4)式即为核函数表示的核空间中两样本的距离公式。

### 1.2 基于巢模板的蚁群聚类规则

巢模板的具体定义为

$$\|\text{nestTemp}_k\|^2=\begin{cases}0, & n_k=0,\\ \dfrac{1}{n_k^2}\displaystyle\sum_{i,j=1}^{n_k}d(X_i,X_j), & n_k>0,\end{cases}\tag{5}$$

其中 $n_k$ 代表第 $k$ 个巢中的蚂蚁数,$X_i(0<i\leqslant n_k)$ 代表巢中的某只蚂蚁。当巢中加入新的蚂蚁和剔除不合群的蚂蚁时,巢模板就按照以上定义进行更新。

基于巢模板的蚁群核聚类规则具体操作如下:

(1)巢中新加入蚂蚁:加强入巢的控制条件,除了比较相遇蚂蚁的模板外,还要比较蚂蚁与巢模板的相似性,以确定蚂蚁与巢中所有成员整体的关系,只有当两者同时满足接受阈值,蚂蚁才能加入。

(2)从巢中剔除不合群蚂蚁:以群体特性为依据,剔除与巢模板距离远的个体。

(3)分配无巢蚂蚁:利用蚂蚁和已知巢模板的对比,确定无巢蚂蚁的归属巢,将无巢蚂蚁分配到与它自身最接近的巢中。

## 2 算法描述

设每只蚂蚁都有相同的属性:标签、基因、模版,两个评价参数 $M_i$,$M_i^+$。对每一只蚂蚁 $a_i$ 定义参数:蚂蚁巢穴属性决定的标签 $\text{Label}_i$,用来代表巢。起初蚂蚁不受任何巢穴的影响,所以 $\text{Label}_i=0$,随后标签不断变换直到蚂蚁找到最好的巢为止。模板由蚂蚁的基因 $\text{Genetic}_i$ 和接受阈值 $\text{Template}_i$ 组成,其中 $\text{Genetic}_i$ 是数据集的对象且在算法过程中不断变化。$\text{Template}_i$ 在初始化节点获得,是蚂蚁与其他蚂蚁相遇期间观察到的最大相似度 $\text{Max}(\text{Sim}(i,\cdot))$ 和平均相似度 $\overline{\text{Max}(\text{Sim}(i,\cdot))}$ 的函数,是动态变化的。蚂蚁每次和其他蚂蚁相遇后按(6)式进行修改:

$$\text{Template}_i\leftarrow\frac{\overline{\text{sim}(i,\cdot)}+\text{Max}(\text{Sim}(i,\cdot))}{2}。\tag{6}$$

评价参数 $M_i$ 反映的是蚂蚁之间的相遇情况。相同巢的蚂蚁相遇时 $M_i$ 增加,反之则减少。开始时 $M_i=0$,在相遇过程中,则表示遇到同巢蚂蚁的个数,即蚂蚁 $a_i$ 所在巢的规模。$M_i^+$ 代表蚂蚁被巢成员接受的程度,当同巢蚂蚁相遇并且相互接受对方时,$M_i^+$ 值增大,否则 $M_i^+$ 值减小。$\text{Sim}(i,j)$ 表示 $i$ 对象和 $j$ 对象之间的相似度,$\text{Sim}(i,j)$ 的值为 0 和 1 之间,当 $\text{Sim}(i,j)=0$ 表示 $i$ 对象和 $j$ 对象完全不同,$\text{Sim}(i,j)=1$ 表示 $i$ 对象和 $j$ 对象完全相同。当蚂蚁学习到其接受阈值后,就可以进行模拟蚂蚁间的相遇操作了。两只蚂蚁相遇并且相互接受需满足:

$$\text{Acceptance}(i,j)\Leftrightarrow\begin{pmatrix}\text{Sim}(i,j)>\text{Template}_i\wedge(\text{Sim}(i,j)>\text{Template}_j)\\ \wedge(\text{Sim}(\text{nestTemp},j)>\text{Template}_j)\end{pmatrix}。\tag{7}$$

带巢模板的蚁群聚类算法:

步骤 1 蚂蚁的初始化过程。

步骤 2 对于任意的一只蚂蚁 $i\in[1,n]$,首先将数据集中第 $i$ 个对象的值赋予蚂蚁 $i$ 作为其遗传属性值,即 $\text{Genetic}_i\leftarrow i^{\text{th}}$。

步骤 3 蚂蚁所属簇的初始化。因各蚂蚁在算法开始时是打散的,故初始时其类簇标记为 0,$\text{Lable}_i\leftarrow 0$。

步骤 4 对于任意的一只蚂蚁 $i\in[1,n]$,随机选取 Iter 只蚂蚁并根据它们的遗传属性值 $\text{Genetic}_i$ 计算其与蚂蚁 $i$ 的相似度 $\text{Sim}(i,\cdot)$,根据

公式(6)初始化蚂蚁 $i$ 的模板阈值 $Template_i$。

步骤5 同时将各蚂蚁的 $M_i$ 和 $M_i^+$ 置0，$M_i \leftarrow 0$，$M_i^+ \leftarrow 0$，$i \in [1,n]$。利用核函数将蚂蚁映射到核空间，用(7)式计算接受阈值。

步骤6 模拟蚂蚁相遇过程，每次只随机选择两只蚂蚁根据算法规则进行模拟相遇操作。用带巢模板的蚁群核聚类规则对蚂蚁进行聚类。

步骤7 统计算巢中蚂蚁的总数及其删除概率(Pdel)。

步骤8 把小于 Pdel×$n$ 值的小巢删除。根据蚂蚁与巢模板的关系，将无巢的蚂蚁分配到与它最相似的巢中。

步骤9 如果迭代次数未结束，返回步骤4；迭代到算法结束。

## 3 算法比对

将原空间上的蚁群聚类算法(AC)，核空间蚁群聚类算法(ACK)和带巢模版的核空间蚁群核聚类算法(ACKT)进行对比。核函数统一选取非线性的径向基函数(RBF)，核函数工作参数统一选取 $1/k$，$k$ 为样本的属性个数，所有实验结果均取10次实验的平均值。

实验的硬件环境是 P4 CPU，3.00GHz，512M 内存的 PC 机，Windows XP 操作系统，用 Java (J2SE 6)编程实现算法。

采用 UCI 公共数据库提供的 Wine，Iris，Breast Cancer(BC)以及 Kdd Cup(99) 4个数据集。从中随机抽取4类网络攻击来测试，实验结果如表1所示。其中 N 代表聚类算法获得的类别数(簇的数目)，$\sigma$ 代表聚类类别的相对误差，它的定义为

$$\sigma = |a - A| / A \times 100\%。 \quad (8)$$

(8)式中，用 $a$ 表示近似数，$A$ 表示准确数，$\sigma$ 为 $a$ 相对于 $A$ 的相对误差，$\sigma$ 值越小，表示聚类得到的类别与真实类别越接近。

**表1 3种聚类算法的结果**

| Data | AC | | ACK | | ACTK | |
|---|---|---|---|---|---|---|
| | $N$ | $\sigma$ | $N$ | $\sigma$ | $N$ | $\sigma$ |
| Wine | 3.4 | 0.13 | 3.1 | 0.03 | 3 | 0 |
| Iris | 2.4 | 0.2 | 2.8 | 0.06 | 2.9 | 0.03 |
| KC'99 | 2.5 | 0.375 | 4 | 0 | 4 | 0 |
| BC | 2.5 | 0.25 | 2.1 | 0.05 | 2 | 0 |

从表1的数据可看出，对于相对容易的 Wine，Breast Cancer 数据集，蚁群聚类算法在原空间上所得到簇的数目明显偏离真实的类别数，而核空间蚁群聚类算法较接近真实类别数。说明在高维核空间上数据的分布状态与真实情况比较接近，在核空间上聚类效果好于原空间。

对于较复杂的 Iris 数据集，由于它的第3类与其它两类有交叉，在原空间上很难把第3类准确地区分出来，但是在核空间上聚类结果已经接近真实的3类。说明经过支持向量机的非线性函数映射后，数据的特征在核空间上发生重组和凸显，Iris 数据集中复杂的第3类就能够较容易地被识别出来。

对于类别较多的入侵检测数据集 KddCup'99，核空间蚁群聚类算法能准确识别出4类，原空间上的蚁群聚类效果明显比核空间上的差。说明类别多的数据样本被映射到核空间后，不同类别的数据样本因特征不同，在核空间上的分布有较大的差异，这有利于聚类算法把它们正确识别出来。

从表1还可以看出，引入巢模板后，核空间蚁群聚类算法的精度略有改进，巢模板的优化作用是有效的，同时也说明影响聚类效果的关键在于核空间的非线性映射，而对蚁群聚类算法的改进措施对聚类效果的提高贡献较小。

## 4 结束语

当数据样本特征复杂、类别多时，在原空间上对数据进行聚类其效果会比较差，主要原因是原空间上数据样本的欧氏距离不能正确反映数据的特征。而利用支持向量机的非线性映射函数，把数据样本映射到核空间后，数据样本的特征在核空间上会发生重组和凸显，数据样本按特征在核空间上重新分布，所以在核空间上进行聚类其效果更接近于事实。实验结果也表明，本文提出的基于巢模板的核空间蚁群聚类算法是有效的，聚类效果与真实情况比较接近。

**参考文献：**

[1] 裴振奎，李华，宋建伟，等. 蚁群聚类算法研究及应用[J]. 计算工程与设计，2008，19(29)：5009-5013.

[2] Li Shanfei，Yang Kewei，Huang Wei，et al. An improved ant-colony clustering algorithm based on the innovational distance calculation formula[C]. IEEE Third International Conference of Knowledge Discovery and Data Mining，2010：342-346.

[3] Elkamel A，Gzara M，Jamoussi，et al. An ant-based algorithm for clustering[C]. IEEE International Conference of Computer Systems and Applications，2009：76-82.

(下转第411页)

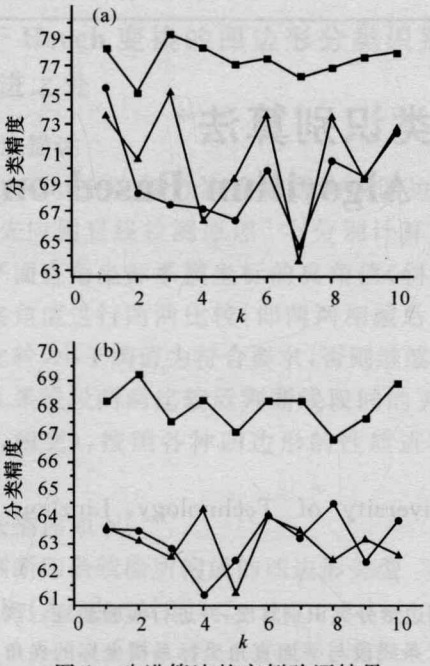


图 1 改进算法的实例验证结果

(a)两类分类问题;(b) 多类分类问题。

■:NNH,●:NNO,▲:NNM。

## 3 结束语

本文对实际生活中广泛应用的最邻近算法进行改进,并应用实例进行验证。改进的算法除了仍然具有容易理解,操作简单,效果明显的特点外,还能解决原算法经常出现的两事例距离计算问题,时序列中计算出现错误问题,无法适用到属性混合情形的问题,以及实现两事例有缺失数据时的距离计算。

**参考文献:**

[1] Cover T M, Hart P E. Nearest neighbor pattern classification[J]. IEEE Transactions on Information Theory, 1967,13(1):21-27.

[2] Vassilis Athitsos, Michalis Potamias, Panagiotis Papapetrou, Nearest Neighbor Retrieval Using Distance-Based Hashing[C]. ICDE, 2008:327-336.

[3] Han J, Kamber M. Data Mining: concepts and techniques: 2nd edition[M]. Morgan Kaufmann Publications, 2006.

[4] Little R, Rubin D. Statistical analysis with missing Data[M]. Wiley, 2002.

[5] 刘星毅. GBNN-填充缺失属性值算法[J]. 微计算机信息,2007,23(15):246-248.

[6] Yang Tao, Cao Longbing, Zhang Chengqi. A novel prototype reduction method for the K-Nearest neighbor algorithm with K>=1[M]. PAKDD, 2010: 89-100.

[7] 杨涛,骆嘉伟,王艳,等. 基于马氏距离的缺失值填充算法[J]. 计算机应用,2005,25(12):2868-2871.

[8] 何晓群. 多元统计分析[M]. 北京:中国人民大学出版社,2004.

(责任编辑:邓大玉)

(上接第 408 页)

[4] Chen Liang. IEEE an ant colony algorithm for text clustering[C]. International Conference of Computing, Control and Industrial Engineering, 2010:249-252.

[5] Mao Xinyan, Sun Binjie, Zhang Ying, et al. Color image segmentation method based on region growing and ant colony clustering[C]. IEEE WRI Global Congress of Intelligent Systems, 2009:173-177.

[6] Pal M, Foody G M. Feature selection for classification of hyperspectral data by SVM[J]. Geoscience and Remote Sensing, IEEE Transactions on, 2010, 48(5): 2297-2307.

[7] Liao Liang, Wang Dongyun, Wang Fengge, et al. A fast kernel-based clustering algorithm with application in MRI image segmentation[C]. IEEE International Conference of Intelligent Computing and Intelligent Systems, 2009:405-410.

[8] Jiang Quansheng, Jia Minping. Novel hybrid clustering algorithm incorporating artificial immunity into fuzzy kernel clustering for pattern recognition[C]. Control Conference, 2007:592-596.

[9] 王国胜. 核函数的性质及其构造方法[J]. 计算机科学,2006,33(6):172-178.

(责任编辑:尹 闯)