一种基于离群点信息的新型无监督聚类方法

吕天阳 王钲旋 左万利
(吉林大学计算机科学与技术学院,长春 130012)

摘 要 在图像检索领域,聚类分析技术有着广泛应用。因为在对图像进行聚类分析时,通常缺少可资利用的先验知识,所以需要采用无监督的聚类算法。为了适应图像检索的需要,提出了一种新型的无监督聚类方法,即根据离群点信息来自动确定聚类算法的终止时机。此方法还弥补了现有聚类算法在离群点识别、使用上的缺欠。为验证其可行性,用其改进了 CURE 和 ROCK 两个经典算法。实验表明,改进后的两个算法都能自动终止,并能取得优于以往的聚类效果。

关键词 图像检索 无监督聚类 离群点
中图法分类号: TP391.3 **文献标识码**: A **文章编号**: 1006-8961(2004)09-1095-06

A New Unsupervised Clustering Method Based on Ontlier Information

Lü Tian-yang, WANG Zheng-xuan, ZUO Wan-li
(*College of Computer Science and Technology, Jilin University, Changchun* 130012)

Abstract There are various applications of clustering analysis techniques in the field of image retrieval. For the lack of valuable prior knowledge in the image retrieval process, unsupervised clustering algorithms should be applied. This paper proposes a new unsupervised clustering method: clustering algorithms will automatically stop according to the outlier information. This method also complements the shortages of current clustering algorithms in outlier detection and using. To show its feasibility, the paper proposes several improvements on two classical clustering algorithms, CURE and ROCK. The empirical results show that by using new method, these two algorithms can stop automatically and also achieve better performance.

Keywords image retrieval, unsupervised clustering, outlier

1 引 言

图像检索技术,尤其是基于内容的图像检索技术,已成为当前的研究重点。在此领域中,聚类分析技术有着广阔的应用前景,它可用于以下几个方面:(1)用于图像分割,提取主要对象[1];(2)用于为大量的图像数据建立一致而有效的组织结构[2],例如在使用 Hierarchical 类型的聚类算法时,聚类过程的树型结构可作为索引结构;(3)图像的检索过程可视为查找与其最相近的数据簇[3]。

但是,在对图像进行聚类分析时,通常缺少可资利用的先验知识,即不清楚某单一图像应分为多少区域,也不了解由大量图像构成数据库中有多少数据簇,而且即使了解情况,先验知识也可能与数据的实际聚集方式存在矛盾,可是许多聚类算法却需要借助先验知识来指定算法,才能最终得到数据簇个数 k,如 K-means、Birch、CURE[4]和 ROCK[5]等算法。

为解决此问题,目前提出的一些无监督聚类算法[1,3]通常使用全局估价函数,当函数取得极值时,算法就自动终止,但其缺点是:(1)易陷入局部最优解;(2)需要引入新参数;(3)估价函数的选择较困难。

本文提出一种基于离群点的无监督聚类方法,即根据离群点信息,自动确定算法的终止时机。它不仅融合了离群点识别和聚类过程,还弥补了现有算法在离群点识别和使用上的缺欠。本文用到的主要

基金项目:国家自然科学基金项目(60373099)
收稿日期:2004-04-15;改回日期:2004-06-24

符号定义见表 1。

表 1 主要符号的约定

N	数据总数
M	数据维数
k	算法得到的数据簇个数
$\boldsymbol{C}_i$	第 i 个数据簇
n_i	属于 $\boldsymbol{C}_i$ 的数据个数
$d(\boldsymbol{C}_i,\boldsymbol{C}_j)$	簇 $\boldsymbol{C}_i$,$\boldsymbol{C}_j$ 的相似度(距离)

2 基于离群点信息的无监督聚类方法

在聚类时,如果缺少先验知识,就需要从所处理的数据中得到使算法自动终止的信息。

所谓聚类是使同类数据的相似度高,而离群点识别则是寻找与其他数据相似度极低的数据。可见,数据间的距离既能表示数据的相似程度,也能表示它们彼此"相离"的程度。

因此,新的无监督聚类方法将根据离群点信息来自动判断聚类算法的终止时机,其基本过程是:随着聚类的进行,当前最相似的两个数据簇 $\boldsymbol{C}_i$、$\boldsymbol{C}_j$ 间的差异增大,直到某一时刻,它们的差异明显,以至于可以认为它们互为离群点时,则算法自动终止。

新方法还弥补了现有聚类算法在离群点识别和使用上的缺欠。有些聚类算法,如 CURE[4]、ROCK[5]等算法,只是将较小的或聚类过程中增长缓慢的数据簇判定为离群点,这一做法只是将离群点识别作为聚类过程的副产品,并不能很好地识别离群点,更不能解释造成数据异常的原因,而且在识别出离群点后,这些算法只是简单地删除或搁置离群点,而没有利用离群点隐含的信息。

为测试新方法的可行性,用该方法改进了以下两个典型的聚类算法,以检查其效果。

2.1 CURE 算法

CURE 算法[4]用 r 个代表点刻画数据簇,由于其能够识别复杂形状的数据簇,且代表点可以保存在索引结构中,从而提高检索质量,并可降低索引结构占用的存储空间,因此,它适合图像检索的复杂情况。算法简介如下:

其基本流程为:(1)设定参数 k、r 和收缩率 α,并将每个输入数据作为单独的数据簇;(2)求每个簇的最近邻点,即代表点按收缩率 α 向数据簇中心收缩,则 $d(\boldsymbol{C}_i,\boldsymbol{C}_j)$ 等于收缩后两簇所有代表点间距离的最小值;(3)合并当前最相似的簇,确定合并后新产生的数据簇 $\boldsymbol{C}_{\text{new}}$ 的代表点及其最近邻点,若当前簇的数目大于 k,则转步骤(3),否则算法结束。

代表点的确定方法为:若 $r=n_i$,则 $\boldsymbol{C}_i$ 的所有数据都是代表点;否则,距簇中心最远的簇内数据点 a 为第 1 个代表点,距点 a 最远的点 b 为第 2 个代表点,以此类推,直到求得 r 个代表点。

为了使 CURE 算法能根据离群点信息自动终止,对它做如下 3 点改进:(1)聚类前,识别离群点;(2)聚类过程中,动态的计算数据簇间的相似性;(3)最后,依据离群点信息自动终止。

(1) 离群点识别

要使用离群点信息,则聚类算法需首先识别离群点。由于 CURE 算法只是将较小的数据簇或聚类过程中增长缓慢的数据簇判定为离群点,这时离群点识别只是聚类过程的副产品,如果聚类错误,则离群点识别也会出错,而且当原始类的数据量很小时,此方法失效。

因此,应采用专门的算法识别离群点。本文以基于距离的离群点概念[6]为基础,给出如下定义:

若 $d_{\text{NN}}(a)\times\xi>d_{\text{outlier}}$,则数据点 a 为离群点,其中,$d_{\text{NN}}(a)$为 a 点与其最近邻(nearest neighbor,NN)点间的距离,ξ 为考虑数据分布局部特征的因子,不妨设点 a 的最近邻点为点 b;如果点 b 的最近邻点不是点 a,则 $\xi=d_{\text{NN}}(a)/d_{\text{NN}}(b)$;否则 $\xi=d_{\text{sec}}(b)/d_{\text{NN}}(a)$,$d_{\text{sec}}(b)$为 b 与其第 2 近邻点间的距离。d_{outlier}为距离阈值,可参考数据均匀分布时的情况确定。此时,任意数据的 d_{NN}相同,可用 $\hat{d}$ 近似表示,$\hat{d}=\sqrt{\sum_{i=1}^{M}((a_{\max}^{(i)}-a_{\min}^{(i)})/\sqrt[M]{N})^2}$,$a_{\max}^{(i)}$、$a_{\min}^{(i)}$为所有数据第 i 维的最大值、最小值,则 $d_{\text{outlier}}=\hat{d}/\beta$,$\beta$ 为参数。

上述离群点识别方法能够很好的融入聚类算法中,如在读数据的过程中,可通过确定向量 $\boldsymbol{a}_{\max}$、$\boldsymbol{a}_{\min}$ 来求得 d_{outlier};对数据间距离的计算,可用于完成聚类和离群点识别两项工作。与其他的离群点识别算法相比[6],本文的方法使用的参数也更少。

(2)簇间距离

聚类过程中,在计算数据簇间的距离时,应考虑各数据簇的密度。

任意两数据簇间的距离 $d(\boldsymbol{C}_i,\boldsymbol{C}_j)$取决于:①合并前后,用 δ 刻画的数据簇密度 D 的变化;②$\boldsymbol{C}_i$、$\boldsymbol{C}_j$ 所有代表点间距离的最小值 $d_{\min}$。

合并后新产生的数据簇 C_{new} 的密度 $D(C_{new})$ 用 d_{min} 近代表示，C_i、C_j 的密度等于该数据簇代表点收缩后的平均距离。如果 $D(C_i)>D(C_{new})$，则 $\delta_i=D(C_i)/D(C_{new})$；否则 $\delta_i=D(C_{new})/D(C_i)$。同理可求 δ_j。

如果 n_i、n_j 都大于 1，则有 $d(C_i,C_j)=d_{min}(C_i,C_j)\times(\delta_i+\delta_j)/2$，否则 $d(C_i,C_j)=d_{min}(C_i,C_j)$。

很明显 $(\delta_i+\delta_j)/2\geqslant 1$，因此，$C_i$、$C_j$ 与 C_{new} 密度相差越大，$d(C_i,C_j))$ 的值越大，合并 C_i、C_j 的可能性就越小。

(3)自动终止

$d_{outlier}$ 用于识别离群点，由于其含有离群点信息，因此可用于判断算法的终止时机。

若当前最相似的两个数据簇 C_i、C_j 间的距离 $d(C_i,C_j)>d_{outlier}$，表明 C_i、C_j 的差异明显如同离群点，则算法自动终止。

至此，改进后的算法根据离群点信息和当前数据簇的情况，即可自动确定终止时机，不再需要参数 k。因其具有动态聚类的特点，故命名为 CURED(clustering using representatives and dynamically)。

2.2　ROCK 算法

传统的聚类算法适合处理数值型(real number)数据，如 K-Means、CURE 等，但在处理类别属性(categorical)的数据时，则存在较多缺欠[5]。ROCK 算法则针对类别属性的数据，与以往不同，它用公共近邻点数来衡量两数据间的相似性，这样就考虑了数据分布的局部特征。

算法简介：

ROCK 算法[5]用参数 θ 来判断两数据是否相邻，即判断任意两数据 a_1、a_2 间的相似性 $S(a_1,a_2)=|a_1\cap a_2|/|a_1\cup a_2|$；如果 $S(a_1,a_2)\geqslant\theta$，则 a_1、a_2 互为邻居；如果 a_3 是 a_1、a_2 的邻居，则 a_3 是 a_1、a_2 的公共近邻；a_1、a_2 的公共近邻总数记为 $N_{link}(a_1,a_2)$。

ROCK 算法认为，由于属于一类的数据间应有较多的公共近邻，因此，首先合并公共近邻数较多的数据簇。若数据簇 C_i、C_j 中所有数据的公共近邻数总和记为 $N_{link}[C_i,C_j]$，则 $d(C_i,C_j)=N_{link}[C_i,C_j]/((n_i+n_j)^{1+2f(\theta)}-n_i^{1+2f(\theta)}-n_j^{1+2f(\theta)})$，其中 $f(\theta)=(1-\theta)/(1+\theta)$。

算法基本流程为：(1)设定参数 k、θ，将每个输入数据作为单独的数据簇；(2)求每个簇的邻居，进而计算两数据簇的公共近邻数；(3)合并 $d(C_i,C_j)$ 最大的两个簇，并确定新产生的簇 C_{new} 与其他簇的公共近邻数和距离，若当前数据簇的数目大于 k，则转步骤(3)，否则算法结束。

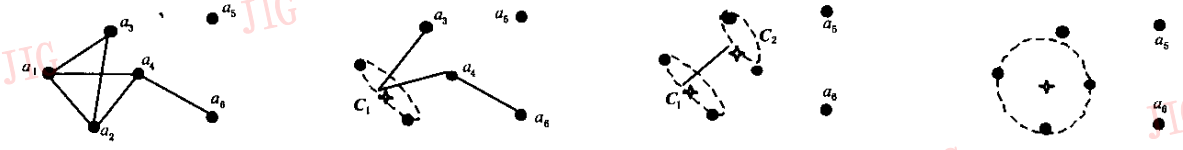
其中，C_{new} 与任意其他数据簇 C_p 的公共近邻数 $N_{link}[C_{new},C_p]=N_{link}[C_i,C_p]+N_{link}[C_j,C_p]$。

以下通过改进 ROCK 算法，使其在聚类过程中可动态评估数据簇的公共近邻，当算法得到若干不连通的子图时，则自动停止。

(1)离群点识别

参数 θ 可用于离群点识别，这一特性却被以往工作忽视[5,7]。

ROCK 算法是基于图的，即数据对应图中节点，而对应节点间有边相连，如果两数据 a_1、a_2 为邻居，则 a_1、a_2 两点间边的权值等于 $S(a_1,a_2)$。删除所有权值小于 θ 的边，将得到形如图 1(a)的图。由此可见，恰当的 θ 值，将使某些数据不是任何其他数据的邻居，例如图 1(a)中节点 a_5，很明显，这样的数据应被识别为离群点。



(a)　(b)　(c)　(d)

图 1　As-ROCK 算法的聚类过程示意图

(2)簇间距离

在聚类过程中，进一步使用公共近邻思想，即只有与 C_i、C_j 都相邻的数据簇 C_p 才被视为 C_{new} 的近邻，$N_{link}[C_{new},C_p]=N_{link}[C_i,C_p]+N_{link}[C_j,C_p]$；否则 $N_{link}[C_{new},C_p]=0$，即合并后仅保留 C_i、C_j 的公共近邻。

聚类过程的示意参见图 1，其中椭圆包含一数据簇中的节点。图 1(b)中节点 a_1、a_2 合并为簇 C_1，此时 $N_{link}[C_1,a_3]=N_{link}[a_1,a_3]+N_{link}[a_2,a_3]$；在图 1(c)中，节点 a_3、a_4 合并为簇 C_2 后，因为 a_6 只是 a_3 的邻居，所以簇 C_2 与 a_6 不相邻，即无边相连。

(3)自动终止

若按上述方法计算簇间距离，则随着聚类的进行，图中不断有边被删除，最终即得到若干不连通的子图(参见图 1(d))。由该图可见，子图间互不连通，说明彼此差异明显如同离群点。此时，算法自动终止，每个子图对应着一个求得的数据簇。

至此，改进后的算法能够自动终止，并取消参数 k。命名为 As-ROCK(auto-stopped ROCK)。

3 讨 论

CURE 算法和 ROCK 算法的区别明显，如，它们是针对不同类型的数据以及它们计算数据(簇)间距离的方法不同。尽管如此，两算法改进后都能借助离群点信息，自动终止。

3.1 离群点识别

聚类算法大致有以下两类方法获取离群点信息：(1)使用内嵌的机制；(2)使用专门的离群点识别方法。

如前所述，由于 CURE 识别离群点的方法存在诸多缺欠，因此本文使用专门的离群点识别算法。离群点识别有众多的算法可供选择，可参考如下原则来选择：离群点识别算法应与聚类算法有同样的或近似的机制，因为两者都要计算数据间的距离(或密度)，如果两者机制相似，就能用较小的计算量完成两项工作。

As-ROCK 算法则使用 ROCK 算法内嵌的机制来识别离群点，即不与其他数据相邻的数据被判定为离群点。

3.2 自动终止

ROCK 算法自动终止的情况相对简单，即如果当前的数据簇间没有公共近邻，则算法终止。

以下着重分析 CURED 算法自动终止的情况。

易证：滤除离群点后，任意数据点 a 与其最近邻间的距离 $d_{\mathrm{NN}}(a)\leqslant d_{\mathrm{outlier}}$。如果数据或数据簇间的距离在聚类过程中不发生改变，那么在聚类的任意阶段，由于当前相距最近的数据簇间的距离都将小于 d_{outlier}，此时算法将不会自动终止，因此在聚类过程中，CURED 算法在动态计算数据簇 $\boldsymbol{C}_i$、$\boldsymbol{C}_j$ 的相似性时，即考虑了两者的密度。一旦 $\boldsymbol{C}_i$、$\boldsymbol{C}_j$ 相距较远，且合并前后密度差异较大，则 $d(\boldsymbol{C}_i,\boldsymbol{C}_j)$ 将大于 d_{outlier}，此时算法终止。

3.3 复杂度分析

CURE 算法的复杂度为 $O(N^2)$。CURED 算法建立在 CURE 算法的基础上，全部改进带来的复杂度变化为 $O(r^2\times(N-k)+N)$。通常情况下 $r^2<N$，例如按文献[4]建议 r 取[10，15]，如果 $N>225$，则 $r^2<N$。可见 CURED 算法的复杂度与 CURE 算法相同。

在合并 $\boldsymbol{C}_i$、$\boldsymbol{C}_j$ 后，由于 ROCK 算法和 As-ROCK 算法都需要遍历两数据簇的全部邻居，两者区别仅在于遍历过程中，As-ROCK 尚需判断一数据簇是否是 $\boldsymbol{C}_i$、$\boldsymbol{C}_j$ 的公共近邻，因此改进前后复杂度并未发生变化。

3.4 大型图像数据库

对于大型数据库，文献[3，4]提出了先抽样，再聚类，最后扫描数据库的方法。改进后的算法可以沿用此方法处理大型图像数据库。值得注意的是，在对抽样的聚类结果中，离群点也有特殊意义，由于它可能是真正的离群点，也可能是某原始类的唯一样本，因此在扫描数据库的过程中，需注意离群点的作用。

4 实 验

实验采用信息熵(Entropy)H 和纯度(Purity)P[8]来评价聚类结果，信息熵和纯度定义如下：

$$H=\sum_{i=1}^{k}\frac{n_i}{N}\left(-\frac{1}{\log q}\sum_{j=1}^{q}\frac{n_i^{(j)}}{n_i}\log\frac{n_i^{(j)}}{n_i}\right)$$

$$P=\sum_{i=1}^{k}\frac{1}{N}\max_j(n_i^{(j)})$$

其中，$n_i^{(j)}$ 为原属于第 j 类的数据在求得的第 i 类中的个数。H 越小，P 越大，说明聚类效果越好。理想的情况下，$H=0.0$，$P=1.0$。

4.1 实验数据

由于文献[4]没有使用真实的高维数据做测试，且每个原始类的数据量很大(>100)，因此数据集 1 由 ORL 人脸图像库及以往获取的分属 42 人的 193 张样本照片[9]构成，同时用 PCA 方法得到 $M=100$ 的特征数据。为增强实验的实用性和说服力，每人的照片量在 1～5 张之间，以便使不同原始类的密度、大小存在差异。数据集 1 用于对比 CURE 和 CURED 算法。

数据集 2 为来自 UCI machine learning repository 的 Zoo 数据集，共 101 条记录，16 个属性值，用于对比 ROCK 和 As-ROCK 算法。

4.2 实验结果

(1)数据集 1 代表点数 r 的取值范围为[1，

5],收缩率 α 的取值范围为[0.00,1.00]。

表 2 给出 CURE 算法($k=42$)和 CURED 算法的最优聚类结果及相应参数值。表 3 进一步给出 CURED 算法得到的聚类个数,表中,0.157/0.765/51 表示 $H=0.157$,$P=0.765$,得到 51 个簇。结果表明,无论 α/r 取何值,CURED 算法都能自动停止,而且结果优于 CURE 算法的最佳聚类结果,其得到的数据簇个数大于 42,这表明先验知识和数据实际聚集方式间存在矛盾。

表 2　CURE 和 CURED 算法聚类效果评价对比

算　法	$H_{\min}$	$P_{\max}$	α/r
CURE	0.377	0.466	0.97/3
CURED	0.060	0.877	0.91/3

表 3　不同参数下 CURED 算法的聚类结果($\beta=5.8$)

r	α										
	0.01	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	0.99
1	0.157/0.765/51	0.144/0.777/49	0.140/0.765/50	0.119/0.788/51	0.167/0.760/49	0.165/0.771/50	0.105/0.821/55	0.135/0.793/56	0.128/0.793/57	0.092/0.838/61	0.160/0.777/55
2	0.179/0.704/47	0.224/0.670/45	0.188/0.721/49	0.195/0.710/48	0.152/0.760/52	0.138/0.765/55	0.111/0.788/59	0.108/0.810/63	0.102/0.827/66	0.073/0.860/70	0.092/0.838/70
3	0.179/0.721/50	0.201/0.709/48	0.176/0.732/51	0.147/0.777/55	0.139/0.782/56	0.128/0.771/55	0.095/0.827/62	0.103/0.832/67	0.082/0.855/68	0.064/0.872/72	0.118/0.816/68
4	0.148/0.777/54	0.160/0.760/52	0.148/0.777/55	0.149/0.777/55	0.163/0.754/55	0.107/0.810/57	0.115/0.805/63	0.108/0.816/66	0.064/0.872/63	0.064/0.872/72	0.118/0.816/68
5	0.148/0.777/54	0.160/0.760/52	0.148/0.777/55	0.149/0.777/55	0.150/0.771/55	0.104/0.816/58	0.097/0.827/62	0.108/0.816/66	0.138/0.788/62	0.0640/0.872/72	0.144/0.788/66

(2)数据集 2　文献[5]给出 $k=9$ 时,ROCK 算法的聚类结果。为做对比,表 4 中给出 As-ROCK 算法自动终止得到 9 个数据簇时的聚类结果。表 5 为各数据簇的详细情况。

表 4　ROCK 和 As-ROCK 算法聚类效果评价对比($k=9$)

算　法	$H_{\min}$	$P_{\max}$	θ
ROCK	0.070	0.881	0.74
As-ROCK	0.061	0.951	0.82

表 5　As-ROCK 算法的聚类结果($\theta=0.82$)

数据簇编号	原始类 1	原始类 2	原始类 3	原始类 4	原始类 5	原始类 6	原始类 7
Cluster 1*	0	0	1	0	0	0	0
Cluster 2*	0	0	0	0	0	0	1
Cluster 3	0	0	0	0	0	0	8
Cluster 4	38	0	0	0	0	0	0
Cluster 5	0	0	0	0	0	8	1
Cluster 6	0	0	3	0	4	0	0
Cluster 7	3	0	0	0	0	0	0
Cluster 8	0	20	1	0	0	0	0
Cluster 9	0	0	0	13	0	0	0

5　结　论

本文介绍了一种新型的基于离群点信息的无监督聚类方法,并用此方法改进了两个典型的聚类算法,进而分析它的一些性质。实验结果表明,改进后的两个算法能够自动终止,并取得优于以往的结果。下一步工作是将此方法用于图像检索领域中,尤其是探索 CURED 算法用于图像分割和组织图像数据库的潜力。

参 考 文 献

1　Rosenberger C, Chehdi K. Unsupervised clustering method with optimal estimation of the number of clusters: application to image segmentation[A]. In: International Conference on Pattern Recognition[C]. Barcelona, Spain, 2000, **1**:1656～1659.

2　Xiong Xuejian, Chan Kap Luk. Towards an unsupervised optimal fuzzy clustering algorithm for image database organization [A]. In: International Conference on Pattern Recognition[C], Barcelona, Spain. 2000, **3**:3909～3913.

3　Krishnamachari S, Abdel-Mottaleb M. Image browsing using hierarchical clustering[A]. In: Proceedings of the Fourth IEEE Symposium on Computers and Communications, ISCC'99[C], Red Sea, Egypt, 1999:301～307.

4　Guha S, Rastogi R, Shim K. CURE: an efficient clustering algorithm for large database[A]. In: Proceedings of the ACM SIGMOD Conference on Management of data [C]. Seattle, Washington, USA: ACM Press, 1998:73～84.

5　Guha S, Rastogi R, Shin K. ROCK: a robust clustering algorithm for categoical attributes[A]. In: Proceedings of the 15th International Conference on Data Engineering[C], Sydney, Australia, 1999: 512～521.

6　Knorr Edwin M, Ng Raymond T. A unified notion of outliers:

properties and computation [A]. In: Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining [C], Newport Beach, CA, USA, AAAI Press, 1997: 219～222.

7 Gupta C K, Ghosh J. Value balanced agglomerative connectivity clustering [A]. In: SPIE Conference on Data Mining and Knowledge Discovery Ⅲ [EB/OL]. Orlando, Florida, USA, 2001, http://citeseer. ist. psu. edu/gupta01value. html.

8 Zhao Ying, Karypis George. Criterion functions for document clustering: experiment and analysis[R]. Technical Report #01-40, University of Minnesota, Twin Cities. Minnesota, USA, 2001.

9 张悦.周春光.基于径向基函数网络人脸识别的研究[J].系统仿真学报, 2001, **13**(S2): 104～107.

吕天阳 1979 年生,2003 年在吉林大学计算机科学与技术学院毕业,获得计算机应用技术硕士学位,主要研究方向为聚类算法、信息可视化等。
E-mail:raynor1979@sina.com

王钲旋 1945 年生,教授,博士生导师。主要研究方向为计算机图研学、计算几何、图像处理等。

左万利 1957 年生,教授,博士生导师。主要研究方向为数据挖掘、数据库、操作系统等。

第一届中国生物特征识别竞赛 BVC2004

(Biometrics Verification Competition 2004)

随着国内生物特征识别技术的蓬勃发展,通过定期对现有算法的性能作一个客观、公正的评价来了解最新的研究现状,以便为未来的研究计划提供重要的参考。为此,中国科学院自动化研究所联合中国信息安全评测中心身份认证产品与技术测评中心,将共同组织举办第一届中国生物特征识别竞赛,包括指纹、人脸、虹膜三种生物特征的竞赛,自 2004 年 8 月启动,竞赛结果将在 2004 年 12 月 13～14 日在广州召开的第五届"中国生物识别学术会议"上公布。

为体现公平公正的原则,竞赛主办方不参加本次竞赛。在结果公布前,参赛人员可以决定是否匿名,主办方保证不泄露参赛人员的个人资料和程序信息,所提交的算法可以加密、加时间限制、加硬件狗或者做其它的技术处理,竞赛主办方承诺除了竞赛评测外不会将提交的算法用于其它用途。为了方便大家集中精力设计核心算法,主办方近期将公布源程序外壳用于下载,其中包括输入输出接口和读取数字图像数据部分。对于参赛人员,主办方免费提供大量的训练库。其中指纹包括 5 种设备获取的 1 000 幅图像,人脸包括用两种设备采集的 10 种不同表情和光照条件下的 2 000 幅图像,虹膜包括两种设备采集的两种不同光照条件下的 2 400 幅图像。详细的竞赛内容可参见以下网址:http://www.sinobiometrics.com/sinobiometrics'04.htm

欢迎从事生物特征识别技术研究的机构、产业界同行、广大科研工作者踊跃参加本次竞赛。

重要日期:

2004 年 8 月 1 日起公布竞赛通知

2004 年 8 月 10 日～9 月 30 日 接受参赛单位和参赛者的注册信息(网上注册,网址:http://www.sinobiometrics.com/sinobiometrics'04.htm)

2004 年 8 月 10 日～9 月 30 日 公布训练集

2004 年 9 月 1 日～10 月 20 日 开始接受提交的算法

2004 年 10 月 20 日～12 月 5 日 算法评测

2004 年 12 月 13 日 在第五届"中国生物识别学术会议"上公布算法竞赛结果,并颁发证书。

联系人:洪淼 电话:010-62659350

BVC2004 竞赛组委会

中国科学院自动化研究所生物特征认证与测评中心