

文章编号:1001-9081(2010)11-2949-03

## 基于距离的数据流离群点挖掘算法

杨显飞,张健沛,杨静,初妍

(哈尔滨工程大学 计算机科学与技术学院,哈尔滨 150001)

(yangxianfei@eyou.com)

**摘要:**传统的离群点挖掘算法无法有效挖掘数据流中的离群点。针对数据流的无限输入和动态变化等特点,提出一种新的基于距离的数据流离群点挖掘算法。通过 Hoeffding 定理及独立同分布中心极限定理,对数据流概率分布变化进行动态检测,利用检测结果自适应调整滑动窗口大小对数据流离群点进行挖掘。实验结果表明,该算法在人工数据集和真实数据集 KDD-CUP99 中可以对数据流中的离群点进行有效挖掘。

**关键词:**数据流;离群点;Hoeffding 定理;滑动窗口

**中图分类号:** TP301.6;TP311.13 **文献标志码:** A

## Algorithm for mining data stream outliers based on distance

YANG Xian-fei, ZHANG Jian-pei, YANG Jing, CHU Yan

(College of Computer Science and Technology, Harbin Engineering University, Harbin Heilongjiang 150001, China)

**Abstract:** The traditional algorithm of mining outliers cannot mine outliers in data stream effectively. Concerning the infinite input and dynamic change in data stream environment, a new algorithm for detecting data stream outliers based on distance was proposed. Change of data stream probability distribution was dynamically detected by Hoeffding theorem and independent identical distribution central limit theorem. Making use of detection outcome to self adaptation, sliding window size was adjusted to mine outliers in data stream. The experimental results show this algorithm can effectively mine data stream outliers in artificial data set and KDD-CUP99 date set.

**Key words:** data stream; outlier; Hoeffding theorem; sliding window

### 0 引言

随着信息产业的迅速发展,许多应用领域出现了大量连续达到的、潜在无限输入的数据有序序列,人们称之为数据流。离群点挖掘是数据流挖掘的重要组成部分之一,其目的是发现数据集中少量行为异常的数据对象,目前被广泛应用于信用卡欺诈检测、通信盗用以及网络入侵检测等。

迄今为止,静态数据集离群点挖掘算法的研究已经取得了广泛的研究成果<sup>[1-5]</sup>,但将其应用在数据流环境里,挖掘效果却难以让人满意。原因在于:静态数据集需要满足 i. i. d (independent identically distributed) 假设,即所有数据服从同一概率分布且彼此相互独立,因此学者们倾向于将 Hawkins 定义的离群点本质作为研究的出发点,Hawkins<sup>[1]</sup>认为“一个离群点是一个观察点,它偏离其他观察点如此之大以至于使人怀疑是由不同机制生成的”。然而在数据流环境里,产生数据的概率分布可随时间发生动态变化,因此静态数据集离群点挖掘算法无法有效挖掘数据流中的离群点。

另外,也有一些学者对数据流离群点挖掘研究进行了探讨<sup>[6-9]</sup>。由于数据流具有无限输入和动态变化等特点,这些算法大多采用滑动窗口的形式组织数据。一方面可以减少需要考虑的数据集规模,提高挖掘速度;另一方面,由于最近的数据更能体现当前的概率分布情况,当概率分布发生变化时,滑动窗口能够较好地适应新的数据分布情况。然而滑动窗口的引入使得窗口大小的合理设定非常困难,当概率分布不发

生变化时,较大的滑动窗口可以获得更加准确的挖掘结果,但当概率分布频繁发生变化时,较小的滑动窗口能够更快适应新的数据分布。为此,本文提出一种基于距离的数据流离群点挖掘算法(Data Stream Outliers Based on Distance, DSOBD)算法,通过概率分布变化检测自适应调整滑动窗口大小,从而更加准确地挖掘数据流中的离群点。

### 1 基于距离的数据流离群点挖掘算法

设  $o_1, o_2, \dots, o_i$  是特征空间中的数据流数据,将其装入数据块  $w$ , 数据块大小为  $n$ , 则  $s_i \in w_j$ , 当  $i = (j-1) \times n + 1, \dots, j \times n$ 。由于存储限制,仅保留最近到达的  $m$  个数据块。

#### 1.1 相关定义及推论

**定义 1** 基于距离的离群点。如果  $p_j < p$ , 称  $o_j$  为离群点, 其中  $p_j = |\{o_i \mid d(o_i, o_j) < r\}| / n$ ,  $d(o_i, o_j)$  是两个数据之间的距离,  $r$  表示给定的距离阈值,  $n$  是所有样本个数,  $p_j$  为数据  $o_j$  的离群度,  $p$  为离群度阈值。

设一个随机变量  $y$ ,  $y$  的最大取值范围是  $Y$  (概率的最大取值范围是 1, 信息收益的最大取值范围是  $\log c$ ,  $c$  是类别数), 对于  $y$  的  $n$  个独立观测值, 平均值为  $\bar{y}$ , Hoeffding 定理表明以  $1 - \sigma$  的概率下,  $y$  的真实平均值至少是  $\bar{y} - \varepsilon$ <sup>[10]</sup>。其中:

$$\varepsilon = \sqrt{\frac{Y^2 \ln(1/\sigma)}{2n}} \quad (1)$$

**推论 1** 根据 Hoeffding 定理, 设  $x_i \in \{0, 1\}$ ,  $i = 1, 2, \dots$ ,

收稿日期:2010-05-04;修回日期:2010-07-19。 基金项目:国家自然科学基金资助项目(60873037)。

**作者简介:**杨显飞(1979-),男,黑龙江哈尔滨人,博士研究生,主要研究方向:人工智能、数据挖掘;张健沛(1956-),男,黑龙江哈尔滨人,教授,博士生导师,主要研究方向:数据库、数据挖掘、软件理论;杨静(1962-),女,黑龙江哈尔滨人,教授,博士生导师,主要研究方向:数据库、数据挖掘、软件理论;初妍(1979-),女,黑龙江哈尔滨人,讲师,博士研究生,主要研究方向:数据挖掘、模式识别。

$n$  是二项分布随机变量  $x$  的  $n$  个独立观测值,  $x_i = 1$  的概率为  $p$ ,  $x_i = 0$  的概率为  $1 - p$ ,  $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ ,  $E(x) = p$ 。那么, 对于任意的小数  $\varepsilon > 0$  满足:

$$p(|\bar{x} - E(x)| \geq \varepsilon) \leq \sigma \quad (2)$$

其中:  $\sigma$  满足式(1),  $x$  的最大取值为 1。因此式(1)也可写成:

$$n = \frac{1}{2\varepsilon^2} \ln(1/\sigma) \quad (3)$$

**推论 2** 当数据流的概率分布稳定时, 设  $p_A$  是该概率分布产生的数据落入特征空间某一区域  $A$  内的真实概率,  $o = \{o_i\}$ ,  $i = 1, 2, \dots, n$  是该概率分布产生的一个数据集,  $o_A$  是该数据集中落入特征区域  $A$  的数据集合, 根据推论 1 可得:

$$p\left(\left|\frac{n_A}{n} - p_A\right| \geq \varepsilon\right) \leq \sigma \quad (4)$$

其中:  $n_A$  表示数据集  $o_A$  的势,  $\varepsilon$  和  $\sigma$  满足式(3)。

## 1.2 基于距离的数据流离群点挖掘方法

根据定义 1, 当推论 2 中数据集  $o$  由滑动窗口数据组成,  $A$  取以数据  $o_i$  为中心,  $r$  为半径的特征空间区域时, 则  $p_A$  是数据  $o_i$  在数据流中真实离群度,  $n_A/n$  是数据  $o_i$  在滑动窗口内的离群度。根据式(3) ~ (4) 可知, 当  $\sigma$  足够小,  $n$  足够大时, 可以利用  $o_i$  在滑动窗口内的离群度对其真实离群度进行有效估计, 并且随着  $n$  值的不断增加, 估计值无限靠近真实值。

如图 1 所示, 设  $B_i$  为最近到达的数据块,  $o_i$  为该数据块中的任意一个数据, 对其在数据流中的真实离群度进行估计, 根据前文分析, 当概率分布没有发生变化时, 较大的滑动窗口 ( $W_1$ ) 可以获得更好的估计值, 但当概率分布发生变化时, 滑动窗口仅设置为  $B_i$  ( $W_2$ ) 才能正确估计出  $o_i$  的真实离群度, 因此在设定滑动窗口前, 需要进行概率分布变化检测。

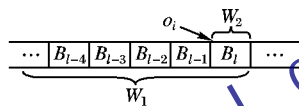


图1 滑动窗口大小示意图

## 1.3 数据流概率分布变化检测

设已存储的数据块为  $B_{i-m+1}, \dots, B_{i-1}, B_i$  (其中  $B_i$  为最近到达的数据块), 如果数据块  $B_i$  和  $B_{i-1}$  的概率分布相同, 则认为数据流的概率分布未发生变化。设  $o_s$  是数据块  $B_{i-1}$  中的任意一个数据, 计算其在数据块  $B_{i-1}$  和  $B_i$  中的离群度, 分别记为  $\bar{z}_{s,i-1}$  和  $\bar{z}_{s,i}$ , 根据式(4), 当概率分布未发生变化时有:

$$p(|\bar{z}_{s,i-1} - p_A| \geq \varepsilon) \leq \sigma$$

$$p(|\bar{z}_{s,i} - p_A| \geq \varepsilon) \leq \sigma$$

因此  $p(|\bar{z}_{s,i} - \bar{z}_{s,i-1}| \leq 2\varepsilon) \geq (1 - \sigma)^2$ , 为了下文叙述方便, 记下式为式(5)。

$$|\bar{z}_{s,i} - \bar{z}_{s,i-1}| \leq 2\varepsilon \quad (5)$$

设二项分布随机变量  $t, t_i = 1$  的概率是  $(1 - \sigma)^2$ ,  $t_i = 0$  的概率是  $1 - (1 - \sigma)^2$ 。根据独立同分布中心极限定理, 对于一个较大的  $k$  值,  $f(k)$  满足正态分布:

$$f(k) = \frac{\sum_{i=1}^k t_i - k(1 - \sigma)^2}{\sqrt{k(1 - \sigma)^2[1 - (1 - \sigma)^2]}}$$

因此在置信度为 0.95 的条件下:

$$\sum_{i=1}^k t_i \geq k(1 - \sigma)^2 - z_{0.025} \sqrt{k(1 - \sigma)^2[1 - (1 - \sigma)^2]} \quad (6)$$

其中:  $z_{0.025}$  表示正态分布的上 0.05 分位数, 在数据块  $B_{i-1}$  中随机抽取  $k$  个数据组成数据集, 记抽样数据集中满足式(5)的数据点个数为  $k_0$ , 根据式(6), 当  $k_0$  满足式(7) 时, 可以判定数据块  $B_{i-1}$  和  $B_i$  在 0.95 的概率条件下, 具有相同的概率分布。

$$k_0 \geq k(1 - \sigma)^2 - z_{0.025} \sqrt{k(1 - \sigma)^2[1 - (1 - \sigma)^2]} \quad (7)$$

## 1.4 基于距离的数据流离群点挖掘算法描述

**算法 DSODB**

**输入** 最近到达的数据块  $B_i$ , 能够保留的数据块个数  $n_b$ , 离群度阈值  $p$ , 抽取数据个数  $k$ 。

**输出** 数据块  $B_i$  中的离群点集合  $o_i$ 。

1)  $\omega = \{B_{i-m+1}, \dots, B_{i-2}, B_{i-1}\}$

// 滑动窗口已保留的数据块集合

2) 从数据块  $B_{i-1}$  中随机抽取  $k$  个数据, 计算其在数据块  $B_{i-1}$  和  $B_i$  的离群度  $\bar{z}_{s,i-1}$  和  $\bar{z}_{s,i}$ 。

3) 根据式(7) 判断数据块  $B_{i-1}$  和  $B_i$  是否具有相同的概率分布。如果相同,  $\omega = \{B_{i-m+1}, \dots, B_{i-1}, B_i\}$ , 否则  $\omega = \{B_i\}$ 。

4) 如果  $\omega$  包含  $n_b + 1$  数据块数据, 删除第一个数据块。

5) 对于数据块  $B_i$  中的每个数据点  $o_i$ , 通过计算其与  $\omega$  集合中所有数据的距离计算其离群度, 当离群度低于离群度阈值  $p$  时,  $o_i$  加入集合  $o_i$ 。

## 2 实验及分析

为了验证 DSORD 算法的有效性, 分别在人工数据集 Simulant-set1 和真实数据集 KDD-CUP99 上进行实验验证及分析。实验环境为 Windows XP 操作系统, CPU 3.0 GHz, 内存 512 MB, 算法实现工具 Matlab 7。

**实验 1** 2 维人工数据集 Simulant-set1 的生成方法如图 2 所示, 在 Matlab 环境中, 边长为 10 个单位的方形区域内等概率生成符合正态分布  $N(6.0, 1.0)$  和均匀分布  $(x - 2)^2 + (y - 2)^2 = 4$  的数据点, 随机将 1% 的数据点替换成空白区域均匀分布的噪声点, 共 40000 个数据。实验设计如下: 对同一数据集进行两次独立挖掘, 第一次采用数据流的形式以数据块为单位读入数据, 利用 DSORD 算法进行挖掘; 第二次使用完整数据集, 使用 Knorr 等人提出的基于距离的离群点算法 FAOD 算法进行挖掘<sup>[3]</sup>。DSORD 算法实验参数设置为:  $n = 5000, \sigma = 0.05, r = 1, n_b = 5, k = 1000$ 。FAOD 算法实验参数  $r = 1$ , 以数据块为单位两种算法挖掘出来的离群点个数如图 3 ~ 4 所示。

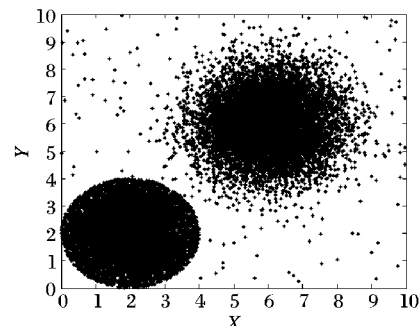


图2 在数据窗口中生成的数据分布

图 3 ~ 4 显示离群度阈值分别为 0.01 和 0.03 两种情况下两种算法挖掘出的离群点个数。从图 3 ~ 4 可以看出, 先前几个数据块中 DSORD 算法与 FAOD 算法挖掘的结果不一致程度较大, 但后面数据块中的挖掘结果趋于相同, 原因在于挖

掘之初滑动窗口较小,利用滑动窗口对数据的离群度进行估计其误差较大,随着滑动窗口的增大,其估计值越来越精确,因此后续数据块中两种算法的挖掘结果大致相同。同时也验证了当数据概率分布未发生变化时,本文提出的数据流概率分布变化检测算法能够不发假检测。

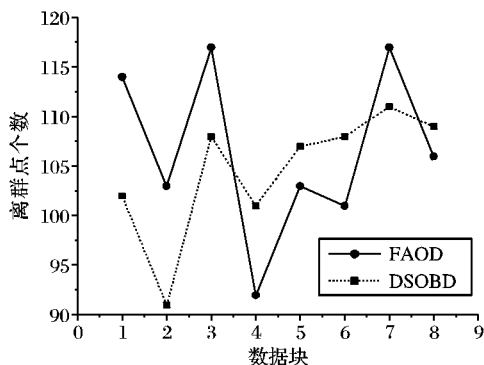


图 3  $p = 0.01$  时挖掘的离群点个数比较

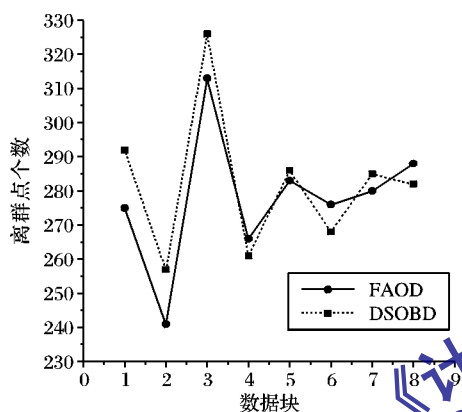


图 4  $p = 0.03$  时挖掘的离群点个数比较

**实验 2** 模拟动态数据流,检验 DSOBD 算法在分布函数发生变化时的挖掘性能。设计实验如下:模拟生成服从正态分布  $N(4.0, 1.0)$  的数据集 Simuland-set2,当概率分布函数发生变化时,其分布函数变为  $N(6.0, 1.0)$ ,共生成 50 000 个数据,随机生成分布函数变化点,噪声加入方式和实验参数同上。进行两次单独实验,第一次使用数据流的形式,利用 DSOBD 算法进行挖掘,第二次以概率分布函数变化为界,将数据集合成两部分,每一部分单独使用 FAOD 算法进行挖掘。可见第二次挖掘是理想情况下的挖掘。

图 5 为离群度阈值  $p = 0.03$  的条件下两次实验在每个数据块中的离群点挖掘结果的对比图,其他  $p$  值具有相似结果,从图 5 可知,在第 6 个数据块两种算法的挖掘结果相差比例远远高于其他数据块,其原因是随机生成的概率分布变化点在第 6 个数据块内,该块内数据包含了新旧两种概率分布产生的数据,因此 DSOBD 算法挖掘结果不理想,但第 7 个数据块内的离群点挖掘比较正常,说明 DSOBD 算法识别出了概率分布变化的情况,从而在滑动窗口中舍弃了前几个数据块的数据,随着概率分布持续稳定,其挖掘结果也越来越精确,从而进一步验证了实验 1 结果的可靠性。

真实数据集 KDD-CUP99 来源于网络入侵检测数据集 kddcup.data\_10\_percent.gz,共 494 021 条记录,每条记录有 42 个属性,包括 7 个离散属性,34 个连续属性和 1 个类别属性,对应正常模式或某种入侵模式,在进行实验前对数据进行预处理,去掉离散属性和类别属性,对剩余的连续属性按文献

[11]的方法进行归一化处理。由于离群点的挖掘任务是发现异于常规数据的少量异常数据,因此本实验仅对除攻击为 neptune 和 smurf 其余 20 种入侵攻击进行检测,原因是数据集中包含攻击为 neptune 的记录共 107 201 条,攻击为 smurf 的记录共 280 790 条,都高于正常模式属于 normal 记录的 97 279 条,因此离群点挖掘无法对其进行有效识别。取前 40 000 条数据进行实验,DSOBD 算法实验参数设置为  $n = 5\,000$ ,  $\sigma = 0.05$ ,  $r = 3$ ,  $n_b = 5$ ,  $k = 1\,000$ ,实验结果如图 6 所示。

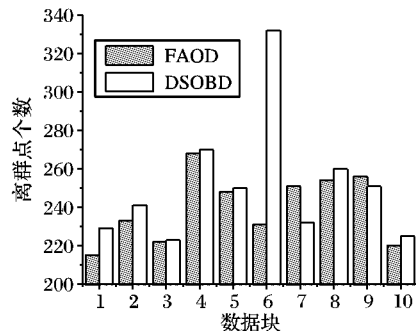


图 5 两种算法在进化数据流中挖掘效果对比

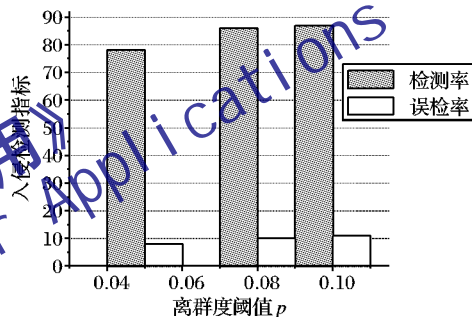


图 6 DSOBD 算法在 KDD-CUP99 数据集上的挖掘性能图

图 6 中检测率是被算法检测到的真实离群点数量与数据集的离群点数量的比值;误检率是被错误判定为离群点的数量与数据集中非离群点数量的比值。我们通过变化不同的离群度阈值来比较系统性能。从图 6 中可以看出,DSOBD 算法可以有效检测出网络入侵的攻击记录,随着离群度阈值的增大,其检测率也有所提高,但误检率也随之增大,原因在于随着离群度阈值的增大,处于边缘的正常数据也被误判为离群点。同时 DSOBD 算法仍然与 FAOD 算法相同,其挖掘结果受参数  $r$  和  $p$  的影响较大。

### 3 结语

针对数据流的无限输入和动态变化等特点,结合传统离群点的定义,本文提出一种基于距离的数据流离群点挖掘算法——DSOBD 算法。通过 Hoeffding 定理及独立同分布中心极限定理,对数据流概率分布变化进行检测,并根据其结果动态调整滑动窗口的大小,使 DSOBD 算法可以有效挖掘出数据流中的离群点。

#### 参考文献:

- [1] BREUNIG M M, KRIEGER H P, NG R T, et al. LOF: Identifying density-based local outlier [C]// Proceeding ACM SIGMOD'00 International Conference on Management of Data. Dallas, TEXAS: ACM, 2000: 93-104.
- [2] ZHANG T, RAMAKRISHNAN R, LIVNY M. BIRCH: An efficient data clustering method for very large databases [J]. ACM SIGMOD Record, 1996, 25(2): 103-114. (下转第 2973 页)



值相比较,Log-Gabor 小波能更好地表征人脸特征,但其产生的高维样本空间对特征提取造成了一定的难度,通过采用多通道策略,将这种高维样本空间分解成多个较低维数的样本空间,这将大大降低算法复杂度。(2D)<sup>2</sup>PCALDA 结合了

2DPCA 与 2DLDA 算法的优点,提取的特征既包含了样本的描述信息又包含了判别信息。在 ORL、Yale 与 CAS-PEAL-R1 数据库上的实验表明,结合本文方法是可行的、有效的与鲁棒的。

表 1 8 种方法在 ORL 数据库上的最佳平均识别率

%

算法	训练样本数					
	2	3	4	5	6	7
本文方法	87.50( $\sigma=1.13$ )	94.07( $\sigma=1.56$ )	96.00( $\sigma=0.97$ )	97.80( $\sigma=1.66$ )	98.50( $\sigma=1.32$ )	99.00( $\sigma=0.89$ )
(2D) <sup>2</sup> PCA	82.41( $\sigma=2.01$ )	89.50( $\sigma=2.16$ )	92.17( $\sigma=2.37$ )	94.95( $\sigma=1.74$ )	96.69( $\sigma=2.97$ )	96.83( $\sigma=2.13$ )
(2D) <sup>2</sup> LDA	83.43( $\sigma=1.71$ )	92.03( $\sigma=2.03$ )	93.55( $\sigma=1.56$ )	96.20( $\sigma=2.17$ )	97.28( $\sigma=1.71$ )	98.42( $\sigma=1.27$ )
(2D) <sup>2</sup> PCALDA	84.36( $\sigma=1.21$ )	92.66( $\sigma=1.45$ )	93.98( $\sigma=1.08$ )	96.02( $\sigma=1.34$ )	97.56( $\sigma=1.67$ )	98.56( $\sigma=1.31$ )
2DPCA	83.06( $\sigma=1.51$ )	88.69( $\sigma=1.20$ )	92.35( $\sigma=1.34$ )	94.10( $\sigma=1.55$ )	96.06( $\sigma=1.34$ )	97.25( $\sigma=1.22$ )
2DLDA	86.75( $\sigma=1.08$ )	90.54( $\sigma=1.17$ )	93.92( $\sigma=1.77$ )	95.40( $\sigma=1.91$ )	96.19( $\sigma=1.28$ )	97.08( $\sigma=0.97$ )
Gabor + LDA	79.56( $\sigma=2.17$ )	91.12( $\sigma=1.49$ )	95.75( $\sigma=2.38$ )	97.36( $\sigma=2.51$ )	98.29( $\sigma=2.03$ )	98.92( $\sigma=1.72$ )
Gabor + 2DLDA	84.33( $\sigma=1.61$ )	93.29( $\sigma=1.29$ )	95.87( $\sigma=1.48$ )	97.38( $\sigma=2.03$ )	98.30( $\sigma=1.77$ )	99.00( $\sigma=1.02$ )

表 2 8 种方法在 Yale 数据库上的最佳平均识别率

%

算法	训练样本数					
	2	3	4	5	6	7
本文方法	84.44( $\sigma=2.78$ )	89.33( $\sigma=3.02$ )	92.67( $\sigma=3.27$ )	93.56( $\sigma=3.81$ )	95.83( $\sigma=3.77$ )	96.22( $\sigma=3.19$ )
(2D) <sup>2</sup> PCA	71.41( $\sigma=3.79$ )	74.42( $\sigma=3.01$ )	77.05( $\sigma=4.11$ )	79.67( $\sigma=4.30$ )	80.33( $\sigma=3.16$ )	80.67( $\sigma=4.27$ )
(2D) <sup>2</sup> LDA	72.45( $\sigma=3.14$ )	83.50( $\sigma=4.05$ )	87.33( $\sigma=3.28$ )	91.22( $\sigma=4.27$ )	92.33( $\sigma=3.30$ )	93.56( $\sigma=3.97$ )
(2D) <sup>2</sup> PCALDA	72.87( $\sigma=2.37$ )	80.06( $\sigma=3.15$ )	84.27( $\sigma=4.06$ )	88.22( $\sigma=2.75$ )	89.97( $\sigma=3.31$ )	91.58( $\sigma=3.07$ )
2DPCA	70.82( $\sigma=4.12$ )	73.01( $\sigma=4.51$ )	76.57( $\sigma=3.78$ )	79.22( $\sigma=3.19$ )	80.69( $\sigma=4.37$ )	80.66( $\sigma=3.41$ )
2DLDA	70.15( $\sigma=4.17$ )	79.75( $\sigma=3.13$ )	84.10( $\sigma=3.76$ )	85.84( $\sigma=3.69$ )	90.13( $\sigma=4.02$ )	90.33( $\sigma=4.15$ )
Gabor + LDA	75.59( $\sigma=4.15$ )	84.29( $\sigma=3.78$ )	89.81( $\sigma=3.24$ )	90.90( $\sigma=4.11$ )	95.80( $\sigma=4.09$ )	96.22( $\sigma=3.21$ )
Gabor + 2DLDA	80.07( $\sigma=4.84$ )	86.04( $\sigma=4.09$ )	89.12( $\sigma=3.67$ )	91.33( $\sigma=3.39$ )	95.67( $\sigma=4.32$ )	96.17( $\sigma=3.68$ )

## 参考文献:

- [1] ASHOK R, NOUSHATH S. Subspace methods for face recognition [J]. Computer Science Review, 2010, 4(1): 1-14.
- [2] 程万里, 李伟生. 基于 Gabor-2DLDA 方法的人脸识别研究[J]. 计算机工程与应用, 2008, 44(35): 179-181.
- [3] WANG LIN, LI YONGPING, WANG CHENGBO, *et al.* 2D Gabor-face representation method for face recognition with ensemble and multichannel model [J]. Image and Vision Computing, 2008, 26(6): 820-828.
- [4] LIU CHENGJUN, WECHSLER H. Gabor feature based classification using the enhanced Fisher linear discriminant model for face recognition [J]. IEEE Transactions on Image Processing, 2002, 11(4): 467-476.
- [5] QI YONGFENG, ZHANG JIASHU. (2D)<sup>2</sup>PCALDA: An efficient approach for face recognition [J]. Applied Mathematics and Computation, 2009, 213(1): 1-7.
- [6] WANG JIAN, ZHANG D, FRANGI A F, *et al.* Two-dimensional PCA: A new approach to appearance based face representation and recognition [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2004, 26(1): 131-137.
- [7] LI MING, YUAN BAOZONG. 2D-LDA: A statistical linear discriminant analysis for image matrix [J]. Pattern Recognition Letters, 2005, 26(5): 527-532.
- [8] ZHANG DAOQIANG, ZHOU ZHI-HUA. (2D)<sup>2</sup>PCA: Two-directional two-dimensional PCA for efficient face representation and recognition [J]. Neurocomputing, 2005, 69(1/3): 224-231.
- [9] NOUSHATH S, H KUMAR G, SHIVAKUMAR P. (2D)<sup>2</sup>LDA: An efficient approach for face recognition [J]. Pattern Recognition, 2006, 39(7): 1396-1400.
- [10] FIELD D J. Relations between the statistics of natural images and the response properties of cortical cells [J]. Journal of Optical Society of America, 1987, 4(12): 2374-2394.

(上接第 2951 页)

- [3] JOHNSON T, KWOK I, NG R. Fast computation of 2-dimensional depth contours [C]// Proceedings of the fourth International Conference on Discovery and Data Mining. New York: AAAI, 1998: 224-228.
- [4] KNORR E M, NG R T. Algorithms for mining distance-based outliers in large datasets [C]// Proceedings of the 24th VLDB Conference. San Francisco: Morgan Kaufmann, 1998: 392-403.
- [5] HAWKINS D. Identification of outliers [M]. London: Chapman and Hall, 1980.
- [6] YAMANISHI K, TAKEUCHI J. A unifying framework for detecting outliers and change points from non-stationary time series data [C]// Proceedings of the eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2002: 676-681.
- [7] MUTHUKRISHNAN S, SHAH R, VITTER J S. Mining deviants in time series data streams [C]// Proceedings of the 16th International Conference on Scientific and Statistical Database Management. Washington, DC: IEEE Computer Society, 2004: 41-51.
- [8] HAN F, WANG Y M, WANG H P. Odabk: An effective approach to detecting outlier in data stream [C]// Proceedings of the Fifth International Conference on Machine Learning and Cybernetics. Washington, DC: IEEE, 2006: 1036-1041.
- [9] 周晓云, 孙志挥, 张柏礼, 等. 高维类别属性数据流离群点快速检测[J]. 软件学报, 2007, 18(4): 933-942.
- [10] HULTEN G, SPENCER L, DOMIGOS P. Mining time-changing data streams [C]// Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2001: 97-106.
- [11] 李洋, 方滨兴, 郭莉, 等. 基于只推式方法的网络异常检测方法[J]. 软件学报, 2007, 18(10): 2595-2604.