*Hans Journal of Data Mining* 数据挖掘, 2012, 2, 1-5
http://dx.doi.org/10.12677/hjdm.2012.21001 Published Online January 2012 (http://www.hanspub.org/journal/hjdm)

# Recognization of Blood Odor Based on Stepwise Discrimination Analysis*

**Chengsheng Long, Xin Wang, Dehua Wu#, Huidong Zhang, Jingning Qiang**

Nanjing Police Dog Research Institute of Public Security Ministry, Nanjing
Email: {#jqswdh, longchengsheng}@163.com

Received: Nov. 4th, 2011; revised: Dec. 8th, 2011; accepted: Dec. 15th, 2011.

**Abstract:** A model for recognization of blood odor based on stepwise discrimination analysis was established and the features come from the chromatographs of blood samples. The model was described in detail and its code was compiled by means of Matlab. The human blood and animal blood samples were used to train and test the model, which detailed the usage of the model. The results demonstrated that samples from different species could be distinguished.

**Keywords:** Pattern Recognization; Blood Odor; Stepwise Discrimination Analysis; Matlab

# 基于逐步判别分析的血液气味识别*

**龙成生，王　辛，吴德华#，张汇东，强京宁**

公安部南京警犬研究所，南京
Email: {#jqswdh, longchengsheng}@163.com

收稿日期：2011 年 11 月 4 日；修回日期：2011 年 12 月 8 号；录用日期：2011 年 12 月 15 日

**摘　要：**本文以血液气味色谱为基础，利用逐步判别分析法建立了血液气味识别模型，并对血液气味识别模型的建立进行了详细描述。以 Matlab 为计算工具，编写了血液气味识别模型的代码。以人体血液与犬、鸡的血液为例，讨论了血液气味识别模型的使用方法。血液气味识别模型能够正确区分人体血液与动物血液。

**关键词：**模式识别；血液气味；逐步判别分析；Matlab

## 1. 引言

血液气味是血迹搜索犬作业的物质基础。在命案快速侦破中，血迹搜索犬发挥了快速、便捷、准确定位等特殊作用[1,2]。在犯罪现场进行血迹搜索和血迹气味追踪时，血迹搜索犬能迅速找出现场血迹走向和附有血液气味的物证如凶器、血衣等。在血迹搜索犬的训练和使用中，血迹气味的质量是影响血迹搜索犬作业结果的关键[3,4]。有文献报道，人体血液气味与个体的健康状况[5]和所处环境[6]有密切关系。因此，血液气味的化学组成分析，可以揭示血迹气味的本质特征，有利于进一步提高血迹搜索犬的训练使用水平，提供血迹搜索犬进行气味作业的科学依据。

血迹搜索犬区分人体血液与其它血液的具体化学成分仍不清楚。但是可以肯定，这些化学成分在不同类别血液气味中存在差异性。研究这种差异性有利于扩大血迹搜索犬的应用范围，也为犬的气味识别机理的研究提供参考。模式识别(Pattern Recognition)是通过对表征事物或现象的各种形式数值、文字和逻辑关系信息进行处理和分析，达到对事物或现象进行描述、辨认、分类和解释的一个过程[7]。除了图像处理[8]、语音系统、文字识别等领域外，模式识别技术也被广泛应用于医学[9]、化学[10]、生物学[11]、食品[12]等领域。

*资助信息：公安部应用创新项目(2011YYCXNJJQ164)。
#通讯作者。

 

本文以血液气味样品色谱图为基础，利用逐步判别分析方法建立了血液气味识别模型，并以 Matlab 软件为计算工具编写了相关代码。利用建立的血液气味识别模型对不同来源的血液样品进行了识别，得到了较好结果。

## 2. 基于逐步判别分析的模式识别

逐步判别的基本思想：每一步选一个判别能力最显著的自变量进入判别函数，而且在每次选变量之前都对已经进入判别函数的诸变量逐个检验其显著性，如果发现有某个变量由于新变量的引入而变得不重要，即在判别函数中判别能力不显著时，就剔除这个变量，直到差别函数中包含的所有变量判别能力都显著时为止，其实施过程如图 1 所示。

## 3. 血液气味识别模型

### 3.1. 变量集的建立

在血液气味色谱图中，每一个色谱峰代表一个化合物，每个化合物都有特定的保留时间。血液气味识别模型的变量来自色谱图的保留时间和峰面积。

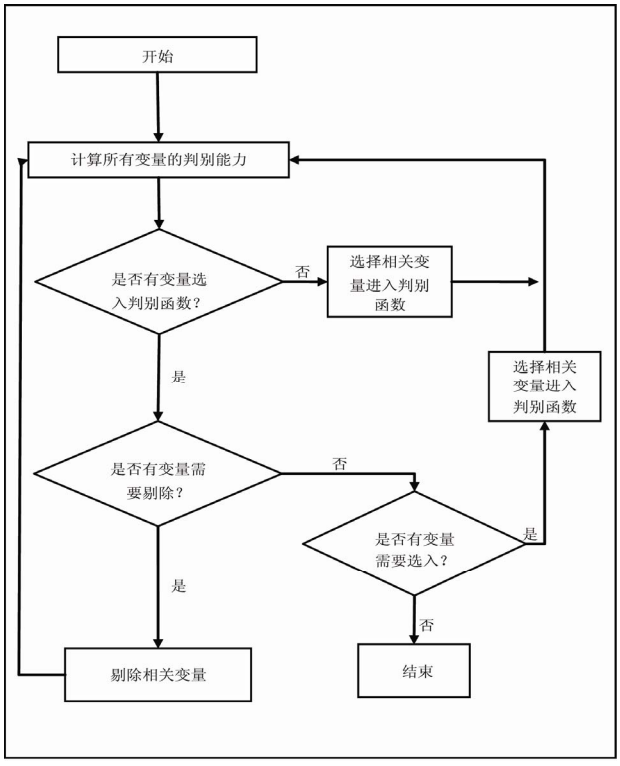


**Figure 1. Stepwise discriminant analysis process**
**图 1. 逐步判别分析流程图**

### 3.2. 特征向量选择

逐步判别分析过程已经包括特征向量优化过程。当然，我们可以根据经验选择待选变量作为自变量集，也可以去除明显的非特征变量。

### 3.3. 逐步判别分析过程(计算过程[13])

设变量数据为 $x_{igk}$，它表示第 $g$ 类第 $k$ 个样品第 $i$ 个变量的数值。

1) 计算分类均值 $\overline{x_{ig}}$ 和总均值 $\overline{x_i}$：

$$\overline{x_{ig}}=\frac{1}{n_g}\sum_{k=1}^{n_g}x_{igk}\ ,\ \overline{x_i}=\frac{1}{N}\sum_{g=1}^{G}\sum_{k=1}^{n_g}x_{igk}\ ,$$

式中 $i=1,2,\cdots,m$，表示第 $i$ 个变量；$g=1,2,\cdots,G$，表示第 $g$ 类；$n_g$ 表示第 $g$ 类的样品数量；$N=n_1+n_2+n_3+\ \cdots\ +n_G$，表示样品总数量。

2) 计算组内协方差矩阵 $W$ 和总协方差矩阵 $T$：

$W=\left(w_{ij}\right)_{m\cdot m}$，式中 $m$ 表示变量个数，

$$w_{ij}=\sum_{g=1}^{G}\sum_{k=1}^{n_g}\left(x_{igk}-\overline{x_{ig}}\right)\left(x_{jgk}-\overline{x_{jg}}\right)$$

$T=\left(t_{ij}\right)_{m\cdot m}$，式中 $m$ 表示变量个数，

$$t_{ij}=\sum_{g=1}^{G}\sum_{k=1}^{n_g}\left(x_{igk}-\overline{x_i}\right)\left(x_{jgk}-\overline{x_j}\right)$$

3) 逐步计算

假设已计算 $l$ 步(包括 $l=0$)，判别函数中引入了 $L$ 个变量，则第 $l+1$ 步的计算内容如下：

a) 计算全部变量的判别能力。若 $x_i$ 是未选变量，则 $U_{i|(L)}=\dfrac{w_{ii}^{(l)}}{t_{ii}^{(l)}}$；若 $x_i$ 是已选变量，则 $U_{i|(L-1)}=\dfrac{t_{ii}^{(l)}}{w_{ii}^{(l)}}$，其中 $w_{ii}^{(l)}$ 和 $t_{ii}^{(l)}$ 表示第 $l$ 步计算的结果。

b) 在已选变量中考虑剔除可能存在的最不显著的变量从已选变量中寻找最大的 $U_{i|(L-1)}$，即最小的 $F$ 值，将最大的 $U_{i|(L-1)}$ 记为 $U_{r|(L-1)}$ 并作 $F$ 检验：

$$F=\left(\frac{1-U_{r|(L-1)}}{U_{r|(L-1)}}\right)\frac{N-G-(L-1)}{G-1}$$，若 $F\le F_\alpha$（$F_\alpha$ 为给定值)，则把 $x_r$ 从判别函数中剔除出去，其后的计算见第 c)步；若 $F>F_\alpha$，则改为考虑从未选变量中选出最显著的变量，这时从未选变量中寻找最小的 $U_{i|(L)}$，即最大的 $F$ 值，将最小的 $U_{i|(L)}$ 记为 $U_{r|(L)}$ 并作 $F$ 检验：

$$F=\left(\frac{1-U_{r|(L)}}{U_{r|(L)}}\right)\frac{N-G-L}{G-1}$$，若 $F>F_\alpha$，则把 $x_r$ 引入判

别函数，其后计算见第 c)步。

c) 不论 $x_r$ 是选入还是剔除，都有相同的计算公式：

$$w_{ji}^{(l+1)}=\begin{cases} w_{rj}^{(l)}\big/w_{rr}^{(l)} & i=r, j\neq r \\ w_{ij}^{(l)}-w_{ir}^{(l)}w_{rj}^{(l)}\big/w_{rr}^{(l)} & i\neq r, j\neq r \\ 1\big/w_{rr}^{(l)} & i=r, j=r \\ -w_{ir}^{(l)}\big/w_{rr}^{(l)} & i\neq r, j=r \end{cases}$$

$$t_{ji}^{(l+1)}=\begin{cases} t_{rj}^{(l)}\big/t_{rr}^{(l)} & i=r, j\neq r \\ t_{ij}^{(l)}-t_{ir}^{(l)}t_{rj}^{(l)}\big/t_{rr}^{(l)} & i\neq r, j\neq r \\ 1\big/t_{rr}^{(l)} & i=r, j=r \\ -t_{ir}^{(l)}\big/t_{rr}^{(l)} & i\neq r, j=r \end{cases}$$

至此，第 $l+1$ 步计算结束，其后重复(1) ~ (3)进行下一步计算。在既不能剔除已选变量也无法引入新变量的情况下，逐步计算结束。

4) 计算判别系数

假设引入了 $L$ 个变量，并且得到 $w_{ji}^{(l)}$，则判别系数的计算为：

$$c_{ig}=(N-G)\sum_{j\in L}w_{ij}^{(l)}\overline{x_{jg}},\quad c_{0g}=-\frac{1}{2}\sum_{j\in L}c_{ig}\overline{x_{jg}},$$

$$i\in L; g=1,2,\cdots,G$$

判别函数为：

$$y_g(x)=\ln p_g+c_{0g}+c_{1g}x_1+c_{2g}x_2+\cdots+c_{mg}x_m$$，$p_g$ 为第 $g$ 类的先验概率。

5) 判别分类

设试样 $\boldsymbol{x}=(x_1, x_2, \cdots x_m)$并将之代入判别函数可得若 $y_g^*(x)=\max_{1\le g\le G}\{y_g(x)\}$，则把 $\boldsymbol{x}$ 归为第 $g$*类。

上述过程利用 Matlab (Version 7.0.0.19920 (R14))进行计算，算法中仅使用了 Matlab 的基本功能函数。

## 4. 实例分析

利用上述血液气味识别模型，对我们的实验数据进行了分析。样品分为两大类，人体血液气味和动物血液气味。以 10 个人体血液气味样品(类别 0)和 10 个动物血液气味样品(类别 1)作为训练集，对该模型进行了训练，再用训练好的模型对其它 8 个样品(类别 1)进行识别。具体过程如下：

### 4.1. 数据预处理

依据我们实验室的研究结果，我们选择了 9 个化合物作为变量，将各化合物的峰面积除以这些化合物的面积之和，得到各化合物的相对峰面积，数据如表 1 所示。

**Table 1. Data of training and test samples**
**表 1. 训练集与测试集数据**

| 编号 | 类别 | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ | $x_6$ | $x_7$ | $x_8$ | $x_9$ |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 0.0099 | 0.1594 | 0.1310 | 0.0053 | 0.0108 | 0.1705 | 0.0055 | 0.0090 | 0.4986 |
| 2 | 0 | 0.0073 | 0.1794 | 0.0848 | 0.0028 | 0.0113 | 0.1656 | 0.0048 | 0.0069 | 0.5373 |
| 3 | 0 | 0.0227 | 0.0000 | 0.0809 | 0.0000 | 0.0115 | 0.2678 | 0.0079 | 0.0132 | 0.5961 |
| 4 | 0 | 0.0122 | 0.0000 | 0.0661 | 0.0000 | 0.0075 | 0.1298 | 0.0030 | 0.0046 | 0.7769 |
| 5 | 0 | 0.0019 | 0.4968 | 0.0654 | 0.0000 | 0.0059 | 0.0539 | 0.0023 | 0.0034 | 0.3704 |
| 6 | 0 | 0.0188 | 0.0000 | 0.1907 | 0.0000 | 0.0124 | 0.1210 | 0.0105 | 0.0202 | 0.6263 |
| 7 | 0 | 0.0055 | 0.0000 | 0.0139 | 0.0008 | 0.0013 | 0.0382 | 0.0017 | 0.0030 | 0.9356 |
| 8 | 0 | 0.0349 | 0.0000 | 0.1089 | 0.0000 | 0.0150 | 0.4278 | 0.0089 | 0.0138 | 0.3906 |
| 9 | 0 | 0.0464 | 0.0000 | 0.1277 | 0.0000 | 0.0101 | 0.2900 | 0.0115 | 0.0205 | 0.4937 |
| 10 | 0 | 0.0370 | 0.0031 | 0.1100 | 0.0000 | 0.0086 | 0.2633 | 0.0095 | 0.0176 | 0.5508 |
| 11 | 1 | 0.0000 | 0.0000 | 0.0076 | 0.0000 | 0.0045 | 0.0000 | 0.0000 | 0.9879 | 0.0000 |
| 12 | 1 | 0.0100 | 0.0000 | 0.0017 | 0.0000 | 0.0054 | 0.0039 | 0.0040 | 0.9740 | 0.0011 |
| 13 | 1 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 1.0000 | 0.0000 |
| 14 | 1 | 0.6510 | 0.0000 | 0.0041 | 0.0000 | 0.1150 | 0.0514 | 0.0696 | 0.1089 | 0.0000 |
| 15 | 1 | 0.0000 | 0.0000 | 0.0003 | 0.0000 | 0.0041 | 0.0032 | 0.0040 | 0.9885 | 0.0000 |
| 16 | 1 | 0.9035 | 0.0000 | 0.0699 | 0.0000 | 0.0265 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |

**Continued**

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 17 | 1 | 0.0000 | 0.0000 | 1.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| 18 | 1 | 0.0832 | 0.0000 | 0.0006 | 0.0000 | 0.0034 | 0.0012 | 0.0000 | 0.9116 | 0.0000 |
| 19 | 1 | 0.0000 | 0.0000 | 0.0008 | 0.0000 | 0.0010 | 0.0000 | 0.0000 | 0.9982 | 0.0000 |
| 20 | 1 | 1.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| 21 | 1 | 0.7620 | 0.0000 | 0.0000 | 0.0000 | 0.2380 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| 22 | 1 | 1.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| 23 | 1 | 1.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| 24 | 1 | 0.8499 | 0.0000 | 0.0000 | 0.0000 | 0.1501 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| 25 | 1 | 0.5659 | 0.0000 | 0.2745 | 0.0000 | 0.0000 | 0.0000 | 0.1122 | 0.0000 | 0.0475 |
| 26 | 1 | 0.6907 | 0.0000 | 0.3094 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| 27 | 1 | 0.0045 | 0.0000 | 0.0105 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.9850 |
| 28 | 1 | 0.9207 | 0.0000 | 0.0000 | 0.0000 | 0.0793 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |

注：类别 0 和 1 分别代表人体血液气味样品和动物血液气味样品；$x_i$ 表示第 $i$ 个变量.

**Table 2. Result of training and test**
**表 2. 训练与测试结果**

| 编号 | 原归类 | *归类* | 编号 | 原归类 | *归类* |
|---|---|---|---|---|---|
| 1 | 0 | 0 | 15 | 1 | 1 |
| 2 | 0 | 0 | 16 | 1 | 1 |
| 3 | 0 | 0 | 17 | 1 | 1 |
| 4 | 0 | 0 | 18 | 1 | 1 |
| 5 | 0 | 0 | 19 | 1 | 1 |
| 6 | 0 | 0 | 20 | 1 | 1 |
| 7 | 0 | 0 | 21 | 1 | 1 |
| 8 | 0 | 0 | 22 | 1 | 1 |
| 9 | 0 | 0 | 23 | 1 | 1 |
| 10 | 0 | 0 | 24 | 1 | 1 |
| 11 | 1 | 1 | 25 | 1 | 1 |
| 12 | 1 | 1 | 26 | 1 | 1 |
| 13 | 1 | 1 | 27 | 1 | 0 |
| 14 | 1 | 1 | 28 | 1 | 1 |

### 4.2. 模型训练

训练基本参数为：先验概率 $p_1 = p_2 = 0.5$，变量数 $m = 9$，类别数 $G = 2$，样品总数 $N = 20$，$n_1 = n_2 = 10$。无论是选入新变量还是剔除已选变量，$F_\alpha$ 均设为 4.0。训练完成后，三个变量 $x_2$、$x_6$ 和 $x_9$ 构成特征向量。类别 1 和类别 2 的判别函数分别为：

$$y_1 = ln(10/20) - 228.2864 + 486.9481x_2 + 664.0131x_6 + 498.0996x_9,$$

$$y_2 = ln(10/20) - 0.0191 + 4.3028x_2 + 6.3263x_6 + 4.2214x_9。$$

### 4.3. 样品识别

用训练好的模型对测试样品进行识别，其结果如表 2 所示。编号 1 至 10 为人血样品(类别 1)，编号 11 至 20 为动物血液样品(类别 2)，编号 21 至 28 为测试样品。所有训练集(编号 1 至 20)样品都被正确归类；在测试集中，27 号样品被错误归类。

结果表明，在 28 个样品中(20 个训练样品，8 个测试样品)，有一个样品(编号 27)被归类错误，错误率为 3.6%。导致 27 号样品错误分类的原因可能是训练样品数量不足。训练样品数量的增加可以优化判别函数，从而减小错误率。在实际应用过程中，或进行深入研究时，可以增加样品数量获得更精确的结果。

## 5. 结论

利用逐步判别分析方法能很好地识别来源不同的血液样品的色谱图，从而达到识别血液样品的目的。在实际应用当中，增加训练集样本量可以提高血液气味识别模型的识别能力；也可以利用血液气味识别模型选择优化特征向量，对血液气味进行更深一步的研究。此外，模型筛选的特征向量对应一组特定的化合物。这些化合物对后续研究很有价值。

## 参考文献 (References)

[1] 温贤章, 范晓杰, 牛焕民. 血迹犬在现场工作中的应用与研究[A]. 第十三次全国养犬学术研讨会论文集[C], 北京: 中国畜牧兽医学会, 2009: 593-596.

[2] 李维福. 搜索微量血迹气味训练可行性探讨[J]. 中国工作犬

 

业, 2011, 2: 24-25.
[3] 刘凤义. 犬在血迹搜索训练中的要点[J]. 中国工作犬业, 2010, 4: 21.
[4] 董继霖. 浅谈血迹搜索犬的训练及使用[J]. 中国工作犬业, 2011, 27(2): 26-28.
[5] G. Horvath, H. Andersson and G. Paulsson. Characteristic odour in the blood reveals ovarian carcinoma. BMC Cancer, 2010, 10(1): 643.
[6] Y. S. Lin, P. P. Egeghy and S. M. Rappaport. Relationships between levels of volatile organic compounds in air and blood from the general population. Journal of Exposure Science & Environmental Epidemiology, 2008, 18(4): 421-429.
[7] 张学工. 模式识别[M]. 北京: 清大学出版社, 2010.
[8] Y. Wen, L. H. He and P. F. Shi. Face recognition using difference vector plus KPCA. Digital Signal Processing, 2012, 22(1): 140-146.
[9] A. Daemen, D. Timmerman, T. Van Den Bosch, *et al*. Improved modeling of clinical data with kernel methods. Artificial Intelligence in Medicine[URL], 2011. http://www.sciencedirect.com/science/article/pii/S0933365711001448#FCANote
[10] D. S. Cao, Y. Z. Liang, Q. S. Xu, *et al*. Exploring nonlinear relationships in chemical data using kernel-based methods. Chemometrics and Intelligent Laboratory Systems, 2011, 107(1): 106-115.
[11] W. Ye and R. T. Robbins. Stepwise and canonical discriminant analysis of longidorus species (nematoda: Longidoridae) from arkansas. Journal of Nematology, 2004, 36(4): 449-456.
[12] C. Simo, P. J. Martin-Alvarez, C. Barbas, *et al*. Application of stepwise discriminant analysis to classify commercial orange juices using chiral micellar electrokinetic chromatography-laser induced fluorescence data of amino acids. Electrophoresis, 2004, 16(25): 2885-2891.
[13] 许禄, 邵学广. 化学计量学方法(第二版)[M]. 北京: 科学出版社, 2004.

 