*Hans Journal of Data Mining* 数据挖掘, 2011, 1, 1-6
http://dx.doi.org/10.12677/hjdm.2011.11001 Published Online July 2011 (http://www.hanspub.org/journal/hjdm/)

# Research and Application of PRSD Studio Pattern Classification*

**Juan Liu, Hongyan Xu, Xiangou Zhu, Wenbin Liu**
[1]College of Physics & Electronic Information Engineering, Wenzhou University, Wenzhou
[2]The Key Laboratory of Low-Voltage Apparatus Intellectual Technology of Zhejiang, Wenzhou
Email: liujuan555ok@163.com; zhuxo@wz.zj.cn
Received: Jun. 24th, 2011; revised: Jul. 20th, 2011; accepted: Jul. 22nd, 2011.

**Abstract:** PRSD Studio is a MATLAB toolbox developed for studying and solving pattern recognition problems. And compared with the traditional pattern recognition PRTools, PRSD Studio has a few special features, such as bedienbarkeit, high speed, high visualization as well as embeding classifiers in your applications outside of the matlab environment. Therefore, after introducing the methods of pattern recognition and several classifiers, it elaborates the function and basic operations of the PRSD Studio pattern recognition toolbox by taking the iris.data as example, including dataset structure, feature selection, classifier design and performance evaluation.

**Keywords:** Pattern Recognition; Classifier; Matlab; PRSD Studio

# PRSD Studio 模式分类器研究与应用*

**刘　娟，徐红燕，朱翔鸥，刘文斌**
[1]温州大学物理与电子信息工程学院，温州
[2]浙江省低压电器智能技术重点实验室，温州
Email: liujuan555ok@163.com; zhuxo@wz.zj.cn
收稿日期：2011 年 6 月 24 日；修回日期：2011 年 7 月 20 日；录用日期：2011 年 7 月 22 日

**摘　要：**PRSD Studio 是一款专为研究解决模式识别问题而开发的 MATLAB 工具箱软件，与传统的模式识别工具箱 PRTools 相比，它操作方便，运行速度快，可视化程度高，且能调度到 matlab 环境外部并嵌入到自定义程序中使用。鉴于此，在介绍模式识别方法及几种分类器后，以鸢尾花数据文件 iris.data 为例，详细阐述了 PRSD Studio 模式识别工具箱的功能及使用方法，其中包括数据集构造、特征选择、分类器的设计及性能评价等。

**关键词：**模式识别；分类器；MATLAB；PRSD Studio

## 1. 引言

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。模式识别是信息科学和人工智能的重要组成部分，其主要应用领域包括图像分析与处理、语音识别、声音分类、通信、计算机辅助诊断、数据挖掘等学科。

目前，研究和解决模式识别问题的主要工具是 MATLAB 模式识别工具箱，用户能直接使用工具箱学习、应用和评估不同的方法，设计不同类型的分类器，解决模式识别领域内的特定问题。本文在介绍几种模式识别分类器的基础上，利用鸢尾花 iris.data 数据详

*基金项目：浙江省重大科技专项 2010C11G2250006；温州市科技计划项目 2008G0372；国家自然科学基金 60970065；浙江省杰出青年基金 ZJNSF：R1110261。

细阐述了 PRSD Studio 模式识别工具箱的主要功能及基本操作。

## 2. 模式分类器

解决模式识别问题的方法可以归纳为基于知识的方法和基于数据的方法。基于知识的方法一般归在人工智能的范畴，而基于数据的方法就是常说的模式识别，下面介绍几个基本术语：

样本集：若干样本的集合。

类或类别：定义在所有样本上的一个子集，处于同一类的样本在我们所关心的某种性质上是不可区分的，即具有相同的模式。

特征：指用于表征样本的观测，通常是数值表示的某些量化特征，有时也被称作属性。如果存在多个特征，则它们就组成了特征向量。

已知样本：指事先知道类别标号的样本。

未知样本：指类别标号未知但特征已知的样本。

基于数据的模式识别方法可以描述为：在类别标号 $y$ 与特征向量 $x$ 存在一定的未知依赖关系、但已知的信息只有一组训练数据对 $\{(x,y)\}$ 的情况下，求解定义在 $x$ 上的某一函数 $y'=f(x)$，对未知样本的类别进行预测，这个函数叫做分类器(classifier)[1]。根据不同的策略，可以构建不同的分类器。异于分类器的另一种模式识别叫做聚类，这种方法是根据样本特征将样本聚成几个类，使属于同一类的样本在一定意义上相似，而不同类之间的样本则有较大差异[2]，这种学习过程也称作非监督模式识别。

下面简要介绍 4 种模式识别分类器。

### 2.1. 贝叶斯分类器

贝叶斯分类器就是根据贝叶斯公式计算属于各个类的后验概率

$$p(\omega_i|x)=\frac{p(\omega_i,x)}{p(x)}=\frac{p(\omega_i)p(x|\omega_i)}{\sum_j(\omega_j)p(x|\omega_j)}$$

然后以后验概率最大化作为决策准则，即

若 $P(\omega_i|x)=\max_{j=1,\cdots,c}P(\omega_j|x)$，则 $x\in\omega_i$。

贝叶斯分类器的分类错误最小[3]，在模式识别研究中常常作为一个基准来评判一个分类器的性能。在有些场合下，不同决策造成的损失各有区别，例如，将一个没病的人误判为有病所带来的损失往往比把有病误判为没病大。这时，人们更关心的是一种风险最小的决策。

$$R(\alpha_i|x)=E[\lambda(\alpha_i,\omega_i)|x]=\sum_{j=1}^{c}\lambda(\alpha_i,\omega_i)P(\omega_j|x)$$

其中 $\lambda(\alpha_i,\omega_i)$ 是决策损失函数，表示对于状态为 $\omega_j$ 的向量 $x$，采取决策 $\alpha_i$ 所带来的损失。这即是最小风险贝叶斯决策[4]。

### 2.2. Fisher 分类器

Fisher 分类器是把 $d$ 维空间的样本投影到一条直线上，然后寻找最优投影方向，使样本在投影以后类间尽可能分开，类内尽可能聚集[1]。其判别准则是

$$\max J(\boldsymbol{w})=\frac{S\boldsymbol{S}_{\boldsymbol{B}}}{S\boldsymbol{S}_{\boldsymbol{W}}}=\frac{\boldsymbol{w}^T\boldsymbol{S}_{\boldsymbol{B}}\boldsymbol{w}}{\boldsymbol{w}^T\boldsymbol{S}_{\boldsymbol{W}}\boldsymbol{w}}$$

其中 $\boldsymbol{S}_{\boldsymbol{B}}$ 表示类间离散度矩阵，$\boldsymbol{S}_{\boldsymbol{W}}$ 表示类内离散度矩阵，$\boldsymbol{w}$ 表示投影方向。Fisher 分类器的最佳投影方向的解为 $w^*=S_w^{-1}(m_1-m_2)$。

### 2.3. $k$ 最近邻分类器

最近邻分类器的思想非常简单，就是一个新样本与已知样本中距离其最近的 $k$ 个样本逐一比较，并把其中大多数样本的类别作为新样本的类别[5]。最近邻决策表达式为

$$g_k(x)=\max_{i=1,\cdots,c}g_i(x)=\max_{i=1,\cdots,c}k_i$$

$k_i$ 表示近邻中属于某一类的个数。为了易于判断，通常 $k$ 取奇数如 1，3，5，7 等。研究表明，最近邻分类器的误差小于 2 倍的贝叶斯误差。

### 2.4. 支持向量机

二类支持向量机的分类目的是找到一个超平面使两类样本完全分开，如果训练样本可以被无误差地分开，并且每一类数据与超平面距离最近的向量与超平面之间的距离最大，则称这个超平面为最优超平面。从而，分类问题转化为求解最优超平面的问题[2]。其通过核函数来构造分类器，核函数的形式有多项式核函数，径向基核函数和 Sigmoid 函数。当核函数选为线性内积时就是线性支持向量机，非线性支持向量机

的问题通过引入特征变换转化成新空间的线性问题，其变换后的优化问题的解可描述为下式形式，

$$\max_{\alpha} \quad Q(\alpha)=\sum_{i=11}^{n}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j K\left(x_i\cdot x_j\right)$$

$$s.t.\ \ \sum_{i=1}^{n}y_i\alpha_i=0,\ 0\le\alpha_i\le C,\ \ i=1,\cdots,n$$

得到的决策函数函数下式所示[1]。

$$f(x)=\operatorname{sgn}\left(\sum_{i=1}^{n}\alpha_i^* y_i K(x_i,x)+b^*\right)$$

## 3. PRSD Stdio 操作实例

### 3.1. 软件介绍

PRSD Studio 是一款模式识别算法设计和调度的软件，由基于 MATLAB 的 PRTools 工具箱和基于 C 语言的执行库 libPRSD 组成，调度时不需要任何外部依赖关系。

与传统的模式识别工具箱 PRTools 相比，PRSD Studio 工具箱可以非常方便的描述数据集，选择指定标签的数据，在多维特征空间中查看对象分布的散点图。此外，它能将设计的分类器导出 MATLAB 环境外，快速嵌入到自定义应用程序中使用，大幅缩短了研究人员的开发时间。而且，它可在无需编译的情况下快速切换分类器，提高利用效率。http://prsdstudio.com/index.php/html/提供了大量的可用资源，其中有 PRSD Studio 模式识别工具箱的介绍。

### 3.2. 数据集描述

http://archive.ics.uci.edu/ml/datasets/Iris 是一个专为机器学习，智能系统等提供数据集的数据库，目前包含有 200 多个数据集。本文选择其中一个数据集 iris.data 为例说明 PRSD Studio 模式识别工具箱的使用。其中包含 150 个数据，它们主要描述鸢尾花的 4 个特征属性及其所属类别组成，数据格式如表 1 所示。

**Table 1. Data format of the iris.data**
**表 1. 鸢尾花的数据格式**

| 萼片长度 (cm) | 萼片宽度 (cm) | 花瓣长度 (cm) | 花瓣宽度 (cm) | 鸢尾花类别 |
|---|---|---|---|---|
| 5.1 | 3.5 | 1.4 | 0.2 | Iris-setosa |
| 6.4 | 3.2 | 4.5 | 1.5 | Iris-versicolor |
| 5.8 | 2.7 | 5.1 | 1.9 | Iris-virginica |

### 3.3. PRSD Studio 工具箱基本操作

#### 3.3.1. 构造数据集

iris.dat 文件中储存着鸢尾花的各个属性值及类别，将这些数据转成符合工具箱处理格式的命令如下：

```
load iris.dat；iris(:,5) = []；
lab = sdlab('Iris-setosa',50,'Iris-versicolor',50,'Iris-virginica',50)
A = sddata(Iris,lab)；
```

首先加载(load)数据文件到 MATLAB 环境中，然后去除类别列，再通过 sdlab 函数构造标签集，最后，由 sddata 函数构造出符合操作格式的数据集 A。

#### 3.3.2. 特征选择

在模式识别中，可能存在一些与分类问题无关的特征，从而影响分类器的设计及性能。因此，如何选取有效的特征是模式识别的一个基本问题。特征获取主要有两种方式：一种是特征选择，即从 $D$ 个特征中选出 $d(d<D)$个特征；另一种是特征提取，即通过适当的变换把 $D$ 个特征转换成 $d(d<D)$个新特征[6]。通过 sdfeatsel 函数和 sdpca 等函数进行特征选择和提取的操作，例如执行下述命令

```
pf = sdfeatsel(tr,'test','method','individual')
ts1 = ts*pf；    p=sdpca(A,3)
```

sdfeatsel 用单个(individual)特征选择的方法，随机地选择训练集 tr 中的 1 个特征，然后用测试集 ts 测试，这样得到的数据集 ts1 就只有一个特征了，’method’的可选参数还包含有’forward’(顺序前进法)、’backward’(顺序后退法)等；sdpca 用主成分分析的方法提取 3 个特征，即将这 4 个特征按重要性从大到小排列，然后提取出前 3 个主成分作为新特征。

数据集 A 的所有对象有 4 个特征，执行 sdscatter(A) 命令得到其在任意二个特征空间中的散点图，如图 1 所示，观察到在特征空间 1 与 2 中，对象的分布不利于由*和 O 所表示的对象的分类，而特征空间 1 与 4 则利于三类对象的分类。为了快速、有效的进行分类，故选择特征 1 和 4，先将数据集 A 随机的分裂成训练集 tr 和测试集 ts，选择出训练集 tr 中的特征 1 和 4，通过乘操作得到只有两个数据集 tr1 和 ts1，这样新得到的训练集和测试集都只有 2 个特征，执行命令如下：

```
[tr,ts] = randsubset(A,0.5)l
pf = sdfeatsel(tr,[1 4]);
tr1 = tr*pf;        ts1 = ts*pf;
```

### 3.3.3. 训练分类器

在 PRSD Studio 中，分类器由两部分组成，如图 2 所示，因此，训练分类器分为两步：选择训练模型和决策函数。数据样本通过训练模型后的输出叫软输出(估计样本属于每一类的概率密度)，决策函数对软输出进行转化得到一个确定的分类器。决策方法有两种：一种是基于阈值，另一种是基于权值。基于阈值的决策一般用于检测一个目标类，通过计算与目标类的相关性(如与目标类的距离)，设定一个阈值，判别小于等于阈值的样本为目标类，大于阈值的样本为非目标类，这种决策适应于两类分类或检测问题。基于权值的决策适应于多类分类操作，分类器提供一组可比较的权值，所有类的权值加和为 1，分类器的输出值与每类的权值相关[7]。

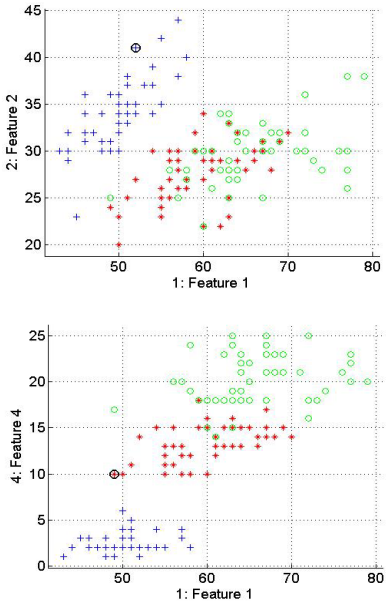



**Figure 1. Scatterplot distribution of the data set A**
**图 1. 数据集 A 的散点分布图**

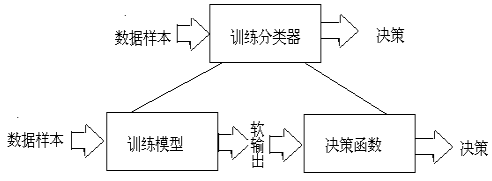



**Figure 2. Classifier component**
**图 2. 分类器的组成**

下面我们以 4 个分类器：朴素贝叶斯分类器，fisher 分类器，5-近邻分类器和支持向量机为例加以说明。在 Matlab 命令窗口中键入：

```
p1 = sdnbayes(tr1,'bins',3);   p2=sdfisher(tr1)
p3 = sdknn(tr1,'k',5);     p4 = sdsvc(tr1)
```

每个训练分类器都会得到一个软输出，nbayes 分类器采用直方图的形式进行分类，参数’bins’决定分类器对于每一类中直方图的条数。sdfisher 根据 fisher 准则投影到一个低维空间，并输出其线性判别。在默认情况下 svc 分类器选择用“RBF”径向基核函数，参数’type’可选择其它核函数，如’poly’多项式核函数或’linear’核函数等。

决策调用 sddecide 函数，例如执行 pd = sddecide(p) 命令就表示对分类器做决策，这样的决策过程使用默认操作点，所谓操作点就是阈值或权值。当然，操作点可以自行设置，如使用命令 sdops 设置两组基于权值的操作点，默认情况下调用第一组做决策，也可通过 setcurop 函数选择第二组权值，权值的不同使分类器的输出值也会不一样，下一节将给出分析结果，执行命令如下。

```
ops = sdops('w',[0.01 0.98 0.01;0.01 0.01 0.98;], tr1.lab.list)
ops1 = setcurop(ops,2)
pd5 = p2*ops;     pd6=p2*ops1
```

### 3.3.4. 评价分类器

分类器设计完成后，就需要分析评价其性能。目前，分类器性能评价标准有很多[8]，如训练错误率，测试错误率，交叉验证等。测试错误率是指从实际问题中的样本划分出一部分、或在目前已有样本之外有条件采集更多的样本来估计分类器性能[1]。当调用函数 sdconfmat 得到一个混合矩阵，就能看到分类器的测试错误率，其中’norm’参数是将结果标准化成小数的形式，执行如下命令：

```
dec1 = ts1*pd1
sdconfmat(ts1.lab,dec1,'norm')
```

表 2 列出了各个分类器的测试错误率，由表可知，朴素贝叶斯分类器对 Iris-setosa 类没有错误率，Iris-versicolor 有 0.12 的错误率，而 Iris-virginica 有 0.04 的错误率，分类器总的错误率为 0.16，依次计算可得，fisher 分类器的总错误率是 0.08，5-最近邻分类器的总

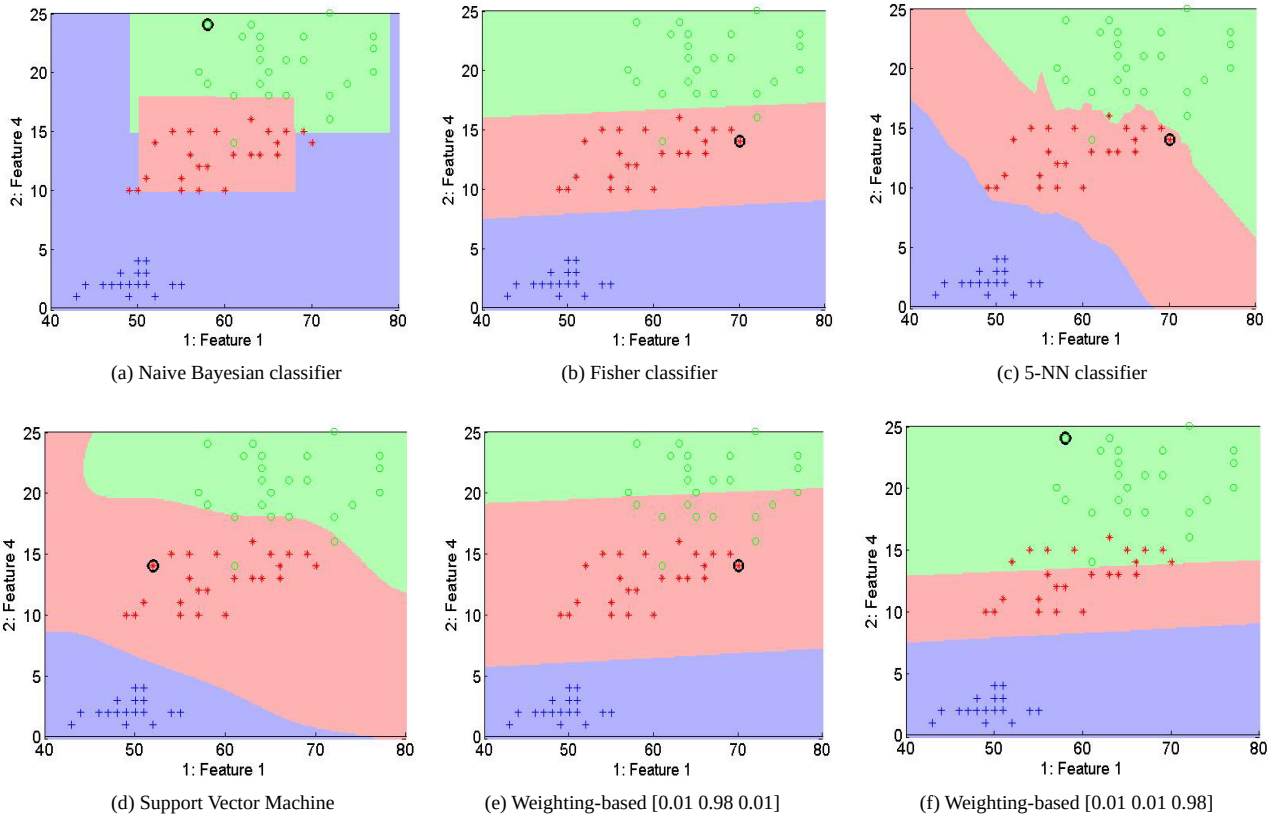
 

错误率为 0.08，支持向量机的总错误率为 0.2。因此，所设计的四个分类器对于测试集 ts1 来说，fisher 分类器和 5-最近邻分类器的性能最佳，而支持向量机的性能最差。分析原因可知，分类器的性能与样本的分布和分类器本身有关系，由于 Iris 数据集的对象较少，用最近邻法计算量小，且准确率高，因此测试结果与理论符合，故在实际应用中，需要综合考虑后选择合适的分类器。

sdscatter 的优点之一是能清晰的观察分类器的分类效果，下图 3 列出了各个分类器的分类图，由图可看出，fisher 分类器和 5-最近邻分类器的性能较好。对于图 e 设置第二类的权值较大，这样分类后第二类的错误率低，几乎为零；对于图 f 设置第三类的权值较大，分类后第三类的错误极低，也为零。这两个图表明，当某些类对于分类来说要使得其风险较小，可通过设置权值来满足这样的分类要求，即决策能自定义分类器的性能。

**Table 2. The test error of the four classifiers**
**表 2. 4 种分类器的测试错误率**

| classifier | True Labels | Decisions | | |
|---|---|---|---|---|
| | | Iris-setosa | Iris-versicolor | Iris-virginica |
| bayes | Iris-setosa | 1.000 | 0.000 | 0.000 |
| | Iris-versicolor | 0.080 | 0.880 | 0.040 |
| | Iris-virginica | 0.000 | 0.040 | 0.960 |
| fisher | Iris-setosa | 1.000 | 0.000 | 0.000 |
| | Iris-versicolor | 0.000 | 1.000 | 0.000 |
| | Iris-virginica | 0.000 | 0.080 | 0.920 |
| knn | Iris-setosa | 1.000 | 0.000 | 0.000 |
| | Iris-versicolor | 0.000 | 0.960 | 0.040 |
| | Iris-virginica | 0.000 | 0.040 | 0.960 |
| svc | Iris-setosa | 1.000 | 0.000 | 0.000 |
| | Iris-versicolor | 0.000 | 1.000 | 0.000 |
| | Iris-virginica | 0.000 | 0.200 | 0.800 |




(a) Naive Bayesian classifier (b) Fisher classifier (c) 5-NN classifier

(d) Support Vector Machine (e) Weighting-based [0.01 0.98 0.01] fisher classifier (f) Weighting-based [0.01 0.01 0.98] fisher classifier

**Figure 3. Classification map of classifiers**
**图 3. 分类器分类图**

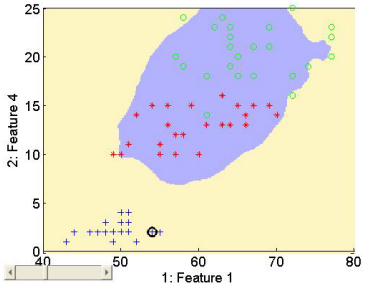
 

### 3.3.5. 检测机与分类器级联

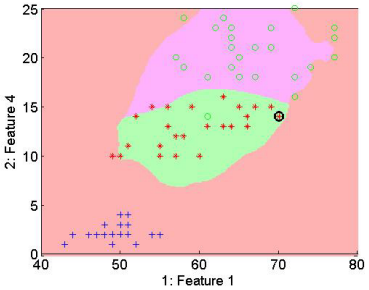
检测机也是一个分类器，其分类规则是只聚集于感兴趣的那一类，常称之为目标类，在很多情况下，可先用检测机分类出感兴趣的目标类，然后再对目标类中的所有类进行分类，这样可节省分类时间，提高分类效率[7]。例如，我们先从数据集中选出标签为’Iris-versicolor’和标签为’Iris-virginica’的类，并重新设置标签为’Iris-color’，构造一个检测机，检测目标类’Iris-color’，接着设计一个分类器与检测机级联，分类出目标中的另外两类对象，MATLAB 命令如下：

```
lab1 = sdrelab(lab,{[2,3] ‘Iri-color’});
tr3 = tr1(:,:,{'Iris-versicolor','Iris-virginica'});
p5 = sdmixture(tr3);        pd5=sddecide(p5)
pd6 = sddetector(tr3,’Iris-color’,sdparzen’, ’reject’, 0.01);      pc = sdcascade(pd6,'Iris-color',pd5)
```

从下图 4 可知，目标类‘Iris-color’是由第二类‘Iris-versicolor’和第三类’Iris-virginica’组成，因此检测机的作用就是检测目标类，将目标类从所有类中划分出来，由于目标类是由两类构成，因此，可通过级联一个分类器将目标类的两类进行分类，这就构成了检测机与分类器级联的形式，这种分类器的应用比较广泛。




(a) Classification map of the detector




(b) Classification map of detector/classifier cascade

**Figure 4. Detector/classifier cascade**
**图 4. 检测机与分类器级联**

### 3.3.6. 分类器的调度

用 PRSD Studio 设计的分类器能应用于任何使用 libPRSD API 自定义的外部应用程序中，使得程序运行速度快，无需编译而更换分类器，从而整体节省研究人员的开发时间。其具体步骤可描述为：(1) 初始化 libPRSD；(2) 加载决策后的输出；(3) 将决策后的输出添加到包含有输入的应用数据缓冲区；(4) 将决策后的输出添加到应用程序结果缓冲区[7]。

分类器的调度是通过 MEX 接口路由选择执行，sdexport 函数将用户设计的分类器导出 MATLAB 外部，然后再嵌入到用户自定义的应用程序中使用，如输入下述命令就将之前设计的朴素贝叶斯分类器调度到 MATLAB 外部，以.ppl 文件格式存储。

```
sdexport(pd1,’myclassifier.ppl’)
```

## 4. 总结

本文主要介绍了几种分类器的分类原理，详细叙述了 PRSD Studio 模式识别工具软件的基本操作，包括构造数据集，特征选择与提取，训练分类器，评价分类器，构造检测机与分类器级联，以及将分类器调度到 MATLAB 外部使用。对于模式识别分类器的设计与研究，PRSD Studio 工具箱提供了学习与研究的环境，http://prsdstudio.com/网站有大量的可供学习研究的资源，有兴趣的可继续进行深入的学习与研究。

## 参考文献 (References)

[1] 张学工. 模式识别[M]. 北京: 清大学出版社, 2010.
[2] 程丽丽. 支持向量机集成学习算法研究[D]. 哈尔宾工程大学, 2009.
[3] M. S. Wong, W. Y. Yan. Investigation of diversity and accuracy in ensemble of classifiers using bayesian decision rules. Beijing: International Workshop on Earth Observation and Remote Sensing Applications, 2008.
[4] 石洪波, 柳亚琴, 李爱军. 贝叶斯分类器的判别式参数学习[J]. 计算机应用, 2011, 31(4): 1074-1076.
[5] 王海芸, 李霞, 郭政等. 四种模式分类方法应用于基因表达谱分析的比较研究[J]. 生物医学工程学杂志, 2005, 22(3): 505-509.
[6] 曹苏群. 基于模糊 Fisher 准则的聚类与特征降维研究[D]. 江南大学, 2009.
[7] PRSD Studio 学习网站. Perclass User’s Guide [URL]. http://perclass.com/doc/guide/index.html.
[8] 宋枫溪, 高林. 文本分类器性能评估指标[J]. 计算机工程, 2004, 30(13): 107-109.