一种新颖混合贝叶斯分类模型研究*)

李旭升　郭耀煌
(西南交通大学经济管理学院　成都 610031)

摘　要　朴素贝叶斯分类器(Naive Bayesian classifier,NB)是一种简单而有效的分类模型,但这种分类器缺乏对训练集信息的充分利用,影响了它的分类性能。通过分析 NB 的分类原理,并结合线性判别分析(Linear Discriminant Analysis, LDA)与核判别分析(Kernel Discriminant Analysis, KDA)的优点,提出了一种混合贝叶斯分类模型 DANB (Discriminant Analysis Naive Bayesian classifier, DANB)。将该分类方法与 NB 和 TAN(Tree Augmented Naive Bayesian classifier, TAN)进行实验比较,结果表明,在大多数数据集上,DANB 分类器具有较高的分类正确率。

关键词　朴素贝叶斯分类器,线性判别分析,核判别分析,TAN 分类器

A Novel Hybrid Bayesian Classification Model

LI Xu-Sheng　GUO Yao-Huang
(School of Economics and Management, Southwest Jiaotong University, Chengdu 610031)

Abstract　Naive Bayesian classifier (NB) is a simple and effective classification model, but it is unable to make the best of the information of the training dataset, thus affecting its classification performance. On the basis of analyzing the classification principle of NB and integrating strongpoint of Linear Discriminant Analysis (LDA) and Kernel Discriminant Analysis (KDA), a new hybrid Bayesian classification model, DANB (Discriminant Analysis Naive Bayesian classifier), is proposed. DANB classifier is compared with NB and TAN (Tree Augmented Naive Bayesian classifier) by an experiment. Experiment results show that this model has higher classification accuracy in most datasets.

Keywords　Naive Bayesian classifier, Linear discriminant analysis, Kernel discriminant analysis, TAN classification

1　引言

朴素贝叶斯分类器(Naive Bayesian classifier,NB)是贝叶斯网分类器的一种,是目前公认的一种简单而有效的概率分类方法,其性能可与决策树、神经网络等算法相比,在某些领域表现性能优异[1,2]。然而,从 NB 方法中的"独立性假设"来看,与现实世界大多不相符合,于是人们从不同的角度思考[1,3~7],对模型进行改进,以调整 NB 中不现实的独立性假设,提高分类器的性能。这方面改进的主要方法是:放弃类变量已知条件下属性变量相互独立的假设以后,进一步表示属性变量之间可能存在的相依关系。Kononenko 的 semi-naive[3]贝叶斯分类器将属性集分割成若干个不相交的属性组,假设在不同组中的属性之间是相互独立的,而同一属性组内的各属性相互关联。Friedman 和 Goldszmidt[1]研究了具有树结构的 TAN 分类器(Tree Augmented Naive Bayesian classifier),允许每个属性节点最多可以拥有一个非类父节点。TAN 具有较好的综合性能,体现了学习效率与分类精度之间的一种适当的折衷[1,8]。BAN(Bayesian Network Augmented Naive Bayesian classifier)[1,9]进一步扩展了 TAN 的结构,允许属性之间可以形成任意的有向图,以表达更复杂的变量间的相依关系。然而,由于其结构学习的复杂性,与一般贝叶斯网络一样,BAN 结构的学习是不容易的(文[10]已证明贝叶斯网的学习是一个 NP-Complete 问题)。

从朴素贝叶斯分类器的学习过程来看,分类器本身也缺乏对训练样本集数据信息的充分利用。在分类器模型中,模型分别模拟每一个类的类条件联合概率分布,然后应用贝叶斯定理构建后验分类器。在分类器构建过程中抛弃了类与类之间信息,而这种信息正是分类所需要的。为了弥补这种缺陷,在认真分析了朴素贝叶斯模型结构特点以及构造分类器方法的基础上,本文提出了一种混合贝叶斯分类器 DANB (Discriminant Analysis Naive Bayesian classifier),给出了构造 DANB 分类器的算法,并实验比较了 DANB、NB 和 TAN,最后总结了本文的工作,并给出了下一步的研究方向。

2　朴素贝叶斯分类器

设 $U=\{X,C\}$是随机变量有限集,其中 $X=\{X_1,\cdots,X_d\}$是属性变量集;C 是类变量,取值范围为$\{c_1,\cdots,c_N\}$,x_i 是属性 X_i 的取值。样本 $x_i=(x_1,\cdots,x_d)$属于 c_i 的概率,由贝叶斯定理可表示为:

$$P(C=c_j|X=x_i)=P(C=c_j),P(X=x_i|C=c_j)/P(X=x_i)=P(c_j)\cdot P(x_1,\cdots,x_d|c_j)/P(x_1,\cdots,x_d) \quad (1)$$

其中 α 是正则化因子,$p(c_j)$是类 c_j 的先验概率,$p(x_1,\cdots,x_d|c_j)$是类 c_j 关于 x_i 的似然。

由概率的链式法则,式(1)可以表示为:

$$P(c_j|x_1,\cdots,x_d)=\alpha\cdot P(c_j)\cdot\prod_{i=1}^{d}P(x_i|x_1,\cdots,x_{i-1},x_{i+1},\cdots,x_d,c_j) \quad (2)$$

给定训练样本集 $D=\{x_1,\cdots,x_n\}$,分类任务的目标是对

*)国家自然科学基金资助课题(70371026)。**李旭升**　博士生,主要研究领域为机器学习、数据挖掘、模式识别、决策支持系统;**郭耀煌**　教授,博士生导师,主要研究领域为决策理论,运筹学。

训练样本集 D 进行分析，确定一个映射函数 $f:(x_1,\cdots,x_d)\to C$，使得对任意的未知类别的实例 $x_i=(x_1,\cdots,x_d)$ 可以标定类标签。

根据贝叶斯最大后验准则，给定某一实例 $x_i=(x_i,\cdots,x_d)$，贝叶斯分类器选择后验概率 $P(c_j\mid x_1,\cdots,x_d)$ 最大的类作为该实例的类标签。在实际应用中，NB 首先按照类标签把训练样本集分成几个子集，然后根据最大似然估计准则(Maximum Likelihood Criterion)，在每个由 c_j 标定的子集，对类条件属性的概率进行估计。对每一个离散属性 $P(X_j=x_{jk}\mid C=c_i)=n_{ijk}/n_i$，$n_{ijk}$ 为事件 $\{X_j=x_{jk}\}$ 在子数据集 $C=c_i$ 发生的频数，n_i 是在子数据集 $C=c_i$ 的样本数。对于连续属性，通常假定服从正态分布，可以表示为：

$$P(x_j\leqslant X_j<x_j+\Delta\mid C=c_i)=\int_{x_j}^{x_j+\Delta}g(x_j:\mu_{c_i},\sigma_{c_i})\mathrm{d}x_j \tag{3}$$

由导数的定义：

$$\lim_{\Delta\to 0}P(x_j\leqslant X_j<x_j+\Delta\mid C=c_i)/\Delta=g(x_j:\mu_{c_i},\sigma_{c_i}) \tag{4}$$

因此，

$$P(x_j\leqslant X_j<x_j+\Delta\mid C=c_i)\approx g(x_j:\mu_{c_i},\sigma_{c_i})\cdot\Delta \tag{5}$$

因子 Δ 出现在如式(1)的每一个类，当进行正则化后，它们将被抵消。其中：

$$g(x_j:\mu_{c_i},\sigma_{c_i})=\frac{1}{\sqrt{2\pi}\sigma_{c_i}}\mathrm{e}^{\frac{(x_j-\mu_{c_i})}{2\sigma_{c_i}^2}} \tag{6}$$

$$\mu_{c_i}=\frac{1}{n_i}\sum_{x\in D_i}x_j \tag{7}$$

$$\sigma_{c_i}=\frac{1}{n_i-1}\sum_{x\in D_i}(x_j-\mu_{c_i})^2 \tag{8}$$

从 NB 学习的分类过程来看，由于它没有使用类间信息，获得仅仅是对各个类别训练样本集分布的一种参数化的近似表达，这正是这类贝叶斯分类器本身的缺陷所决定的。即使在满足"类条件属性相互独立"的假定下，也不可能利用类间信息，这就引出了下面所探讨的 DANB 分类器。

3 判别分析和 DANB 分类器

3.1 线性判别分析

线性判别分析(Linear Discriminant Analysis，LDA)的目标是把 d_0 维空间的数据点投影到 d_1 维($d_1\leqslant d_0$)空间去，以实现不同类的样本点在 d_1 维空间能够形成相互分离的、各自内部紧凑的集合。当然，向任意的空间做投影，也有可能把这些不同类的数据混在一起，反而降低了分类的效果。通过选择适当的投影的空间，找到最大限度区分各类数据点的投影方向，这就是经典的可分性分析目标。

设有一组 d_0 维的训练样本 $x_1,\cdots,x_n$(x_i 为 d_0 维行向量，$n>d_0$)，它们分别属于 N 个不同的类别，即其中大小为 n_i 的样本子集 D_i 属于类别 c_i。对于 N 类问题，把 Fisher 线性判别准则做推广，就需要 $N-1$ 个判别函数。也就是说，投影问题实际上是从 d_0 维空间向 $N-1$ 维空间做投影，并且假设 $d_0\geqslant N$。为获得最好的分类效果，确定最佳的投影方向，需要定义下面的矩阵和向量：

类均值向量：$\boldsymbol{m}_i=\frac{1}{n_i}\sum_{x\in D_i}\boldsymbol{x}$ (9)

总体均值向量：$\boldsymbol{m}=\frac{1}{n}\sum_{x\in D}\boldsymbol{x}=\frac{1}{n}\sum_{i=1}^{N}n_i\boldsymbol{m}_i$ (10)

总体散布矩阵：$\boldsymbol{S}_t=\sum_{x\in D}(\boldsymbol{x}-\boldsymbol{m})(\boldsymbol{x}-\boldsymbol{m})'$ (11)

$$\boldsymbol{S}_t=\sum_{i=1}^{N}\sum_{x\in D_i}(\boldsymbol{x}-\boldsymbol{m}_i+\boldsymbol{m}_i-\boldsymbol{m})(\boldsymbol{x}-\boldsymbol{m}_i+\boldsymbol{m}_i-\boldsymbol{m})'=\sum_{i=1}^{N}\sum_{x\in D_i}(\boldsymbol{x}-\boldsymbol{m}_i)(\boldsymbol{x}-\boldsymbol{m}_i)'+\sum_{i=1}^{N}\sum_{x\in D_i}(\boldsymbol{m}_i-\boldsymbol{m})(\boldsymbol{m}_i-\boldsymbol{m})'$$
$$=\boldsymbol{S}_w+\boldsymbol{S}_b \tag{12}$$

类内散布矩阵：

$$\boldsymbol{S}_w=\sum_{i=1}^{N}\sum_{x\in D_i}(\boldsymbol{x}-\boldsymbol{m}_i)(\boldsymbol{x}-\boldsymbol{m}_i)' \tag{13}$$

类间散布矩阵：

$$\boldsymbol{S}_b=\sum_{i=1}^{N}n_i(\boldsymbol{m}_i-\boldsymbol{m})(\boldsymbol{m}_i-\boldsymbol{m})' \tag{14}$$

显然，类内散布矩阵表达样本到类内中心的距离，其越小，说明同类样本相对较集中；类间散布矩阵是类与类中心距离的度量，其越大，说明不同类样本可分性好。如果能够使得在投影后的空间，类内样本集中，类间样本分离，即可达到所需的目的。从维空间向维空间的投影是通过下列的个分类方程来进行的：

$$y_i=\boldsymbol{w}_i'\boldsymbol{x}\quad(i=1,\cdots,N-1) \tag{15}$$

如果把 y_i 看作是一个 $N-1$ 维的向量 y 的分量，把 w_i 看作是一个 $d_0\times(N-1)$ 矩阵的列向量，上式的投影方程可以表达为简单的矩阵方程：

$$\boldsymbol{y}=\boldsymbol{W}'\boldsymbol{x} \tag{16}$$

对原始训练样本 $D=\{x_1,\cdots,x_n\}$ 进行投影后，得到了投影后的新训练样本 $D'=\{y_1,\cdots,y_n\}$。这些新得到的样本本身具有它们自己的均值向量和散布矩阵，同前面定义相仿：

$$\tilde{\boldsymbol{m}}_i=\frac{1}{n_i}\sum_{y\in D'_i}\boldsymbol{y} \tag{17}$$

$$\tilde{\boldsymbol{m}}=\frac{1}{n}\sum_{y\in D'}\boldsymbol{y}=\frac{1}{n}\sum_{i=1}^{N}n_i\tilde{\boldsymbol{m}}_i \tag{18}$$

$$\tilde{\boldsymbol{S}}_w=\sum_{i=1}^{N}\sum_{y\in D'_i}\boldsymbol{y}(\boldsymbol{y}-\tilde{\boldsymbol{m}}_i)(\boldsymbol{y}-\tilde{\boldsymbol{m}}_i)' \tag{19}$$

$$\tilde{\boldsymbol{S}}_b=\sum_{i=1}^{N}\boldsymbol{n}_i(\tilde{\boldsymbol{m}}_i-\tilde{\boldsymbol{m}})(\tilde{\boldsymbol{m}}_i-\tilde{\boldsymbol{m}})' \tag{20}$$

容易证明：

$$\tilde{\boldsymbol{S}}_w=\boldsymbol{W}'\boldsymbol{S}_w\boldsymbol{W} \tag{21}$$

$$\tilde{\boldsymbol{S}}_b=\boldsymbol{W}'\boldsymbol{S}_b\boldsymbol{W} \tag{22}$$

上述各个方程说明了从高维空间向低维空间的投影过程中，类内散布矩阵和类间散布矩阵经历了怎样的变换。目标是寻找一个投影方向变换矩阵 W，能够在某种意义上，使得投影后的类间散布矩阵和类内散布矩阵的比值最大。离散度的一种简单的标量度量是散布矩阵的行列式的值。使用这样的度量方法，准则函数如下：

$$\max_{\boldsymbol{w}}J(\boldsymbol{W})=\frac{|\tilde{\boldsymbol{S}}_b|}{|\tilde{\boldsymbol{S}}_w|}=\frac{|\boldsymbol{W}'\boldsymbol{S}_b\boldsymbol{W}|}{|\boldsymbol{W}'\boldsymbol{S}_w\boldsymbol{W}|} \tag{23}$$

上式的一个重要性质是对于 W 的每一列 w_i($1\leqslant i\leqslant N-1$)，$w_i\to\gamma_i w_i$，$J(w)$ 的值并不因 γ_i 改变。所以总可以选择恰当的 w_i，使得分母 $w_i^TS_bw_i=1$，由此可以将式(23)的最大化问题转为下面等价的约束优化问题：

$$\min_{\boldsymbol{w}}:-\frac{1}{2}\boldsymbol{w}'_i\boldsymbol{S}_b\boldsymbol{w}_i \tag{24}$$

$$\text{s. t. }\boldsymbol{w}'_i\boldsymbol{S}_w\boldsymbol{w}_i=1 \tag{25}$$

由 Lagrange 定理，

$$l=-\frac{1}{2}\boldsymbol{w}'_i\boldsymbol{S}_b\boldsymbol{w}+\frac{1}{2}\lambda_i(\boldsymbol{w}'_i\boldsymbol{S}_w\boldsymbol{w}-1) \tag{26}$$

由极值条件，有：

$$\boldsymbol{S}_b\boldsymbol{w}_i=\lambda_i\boldsymbol{S}_w\boldsymbol{w}_i \tag{27}$$

由于 $\{w_1,\cdots,w_s\}=W(s\leqslant N-1)$，令 D 为特征向量 $\{w_1,\cdots,w_s\}$ 对应的特征值对角矩阵，可将式(27)写为：

$$S_bW=S_wWD \tag{28}$$

因为 S_b 是 N 个秩为 1 或为 0 的矩阵的和，其中只有 $N-1$ 个矩阵是相互独立的，所以 S_b 的秩为 $N-1$ 或更低。这样，非零的特征值至多只有 $N-1$ 个，所求的特征向量就对应这些非零的特征值。可以证明，在满足式(25)条件下，$W'S_wW$ 是单位矩阵。对于来自 N 个正态总体 $N_p(\mu_1,\Sigma),\cdots,N_p(\mu_N,\Sigma)$，容量分别为 $n_1,\cdots,n_N$ 的随机样本，由于

$$\begin{aligned}&S_t=S_w+S_b\Rightarrow W'S_tW=W'(S_w+S_b)W\\&\quad\Rightarrow W'S_tW=W'S_wW(I+D)\\&\Rightarrow W'S_tW=I+D\\&\Rightarrow\frac{1}{n-1}\tilde{S}_t=\frac{1}{n-1}(I+D)\end{aligned} \tag{29}$$

故投影后的样本协方差矩阵 $\tilde{S}_t/n-1$ 为对角矩阵。当样本容量足够大时，$\tilde{S}_t/n-1$ 收敛于 $W'\Sigma W$。因而，对于这种正态分布样本而言，投影后新样本的属性相互独立。实际上，对原样本的投影变换实现了双重目标：一是通过 *LDA* 获得了类与类之间的分离信息，同时实现了维度缩减，从 d_0 维降到了$\leqslant N-1$ 维；二是对 N 个正态总体，同协方差矩阵的样本而言，可以保在新的属性空间，类条件下属性变量相互独立，这正是 *NB* 的假定所要求的。

3.2 核判别分析

LDA 是比较有效的特征抽取与判别分析方法之一，在处理线性可分性较好的问题时，其性能与效率均是较优的，在应用中取得了较好结果[11]。然而，当模式识别问题的复杂度较高，线性判别分析手段就难以取得较好的分类效果，这时采用基于核函数的判别分析[12]往往能取得较好的分类精度。其基本思想是将原特征空间通过某种形式的非线性映射变换到一个高维空间，实现样本在新空间线性分离。由于在新空间中的线性方向对应于原特征空间的非线性方向，因此基于核的判别分析得出的判别方向对应原特征空间的非线性方向。相对于其他非线性方法，这种方法的独特和关键之处在于它巧妙地借助了“核函数”，而不需要对原特征空间进行任何直接的非线性映射，从而使计算判别矢量的工作变得相对容易。

设原样本空间通过非线性映射函数 ϕ 映射变换到希尔伯特空间(Hilbert Space)F，得到了映射后的样本 $D^\phi=\{\phi(x_1),\cdots,\phi(x_n)\}$。同 LDA 相似，在特征空间 F 定义下面的矩阵和向量：

类均值向量：$$m_l^\phi=\frac{1}{n_l}\sum_{i=1}^{n_l}\phi(x_l) \tag{30}$$

总体均值向量：

$$m^\phi=\frac{1}{n}\sum_{i=1}^{n}\phi(x_i)=\frac{1}{n}\sum_{i=1}^{N}n_im_l^\phi \tag{31}$$

总体散布矩阵：$$S_t^\phi=S_w^\phi+S_b^\phi \tag{32}$$

类内散布矩阵：

$$S_w^\phi=\sum_{i=1}^{N}\sum_{\phi(x_j)\in D_i^\phi}(\phi(x_j)-m_l^\phi)(\phi(x_j)-m_l^\phi)' \tag{33}$$

类间散布矩阵：

$$S_b^\phi=\sum_{i=1}^{N}n_i(m_l^\phi-m^\phi)(m_l^\phi-m^\phi)' \tag{34}$$

采用与 LDA 相同的准则，在空间寻找最佳的投影方向等同于求解下面的广义特征方程：

$$S_b^\phi W^\phi=S_w^\phi W^\phi D^\phi \tag{35}$$

由于

$$\begin{aligned}&S_b^\phi W^\phi=S_w^\phi W^\phi D^\phi\Leftrightarrow S_b^\phi W_\phi=(S_t^\phi-S_b^\phi)W^\phi D^\phi\Leftrightarrow S_b^\phi W^\phi(I+D^\phi)\\&=S_t^\phi W^\phi D^\phi\end{aligned} \tag{36}$$

因此，求式(35)所对应的 W^ϕ 等同于求

$$S_t^\phi W^\phi=S_t^\phi W^\phi\Lambda^\phi \tag{37}$$

这里 $\Lambda^\phi=D^\phi(I+D^\phi)^{-1}$ 为特征值对角矩阵。为了把 LDA 推广到非线性的情况，采用 Mercer 核来实现映射，定义为：

$$k(x_i,x_j)=k_{ij}=\phi'(x_i)\phi(x_j) \tag{38}$$

依据再生核理论[12](Theory of Reproducing Kernels)，特征向量 w^ϕ 必定在于 F 中所有训练样本张成的空间，因此存在系数 $\alpha_i,i=1,\cdots,n$ 使得：

$$w^\phi=\sum_{i=1}^{n}\alpha_i\phi(x_i) \tag{39}$$

将式(39)代入(37)并考虑(30)～(34)，可以证明[13]，对于 Λ^ϕ 中的每一个特征值 λ^ϕ，有：

$$\lambda^\phi\tilde{K}\tilde{K}\alpha=\tilde{K}J\tilde{K}\alpha \tag{40}$$

其中：$$\tilde{K}=K-\frac{1}{n}1K-\frac{1}{n}K1+\frac{1}{n^2}1K1 \tag{41}$$

α 为 λ^ϕ 所对应特征向量$(\alpha_1,\cdots,a_n)'$，记 Λ^ϕ 的特征值所对应的特征向量矩阵为 A；1 是元素均为 1 的 $n\times n$ 矩阵；J 是由 J_l 构成的 $N\times N$ 块对角矩阵，$l=1,\cdots,N$，J_l，是元素为 $1/n_l$ 的 $n_l\times n_l$ 矩阵，K 表示由 k_{ij} 构成的 $n\times n$ 核矩阵，$i=1,\cdots,n$；$j=1,\cdots,n$。样本或实例 z 经非线性映射在核空间 w^ϕ 判别方向的投影为：

$$(w^\phi)'\phi(z)=\sum_{i=1}^{n}\alpha_ik(x_i,z) \tag{42}$$

这里，α_i 由式(40)确定。注意到函数 $\phi(x)$ 的具体形式并不必要显式地表达，它暗含于选择的核函数中。通过选择不同的核函数，如多项式核、高斯核等，就可以实现广泛的非线性映射。因此，如果原样本经过非线性核映射后，能够在空间形成同协方差的正态分布样本，并通过向判别方向投影，则同 LDA 一样可以保证投影后生成的属性变量在类条件下相互独立，并实现提取类间非线性分离信息的目的。

3.3 DANB 分类器

定义(DANB 分类器) 样本原属性集 $X(X_1,\cdots,X_d)$，类变量为 C(取值$\{1,\cdots,N\}$)，将 X 向最大线性可分空间投影(或由核函数实现到空间非线性映射后，向最大线性可分方向投影)，投影后的样本属性集为 $Y=\{Y_1,\cdots,Y_l\}$，$l\leqslant N-1$，Y 向量的每一分量均是向量 X(或 k_{ij})的线性组合。在给定类变量的条件下，假设 Y 每一分量是相互独立的，满足这一条件的贝叶斯分类模型称为 DANB。

构造 DANB 分类器的关键是寻找 W(或 A)。一旦确定了 W(或 A)，即可将原样本投影到新的样本空间。这时候分类器在新的样本空间进行学习，获得模型参数。原样本属性集 $X=\{X_1,\cdots,X_d\}$ 中任意两个属性间可能存在一定的依赖关系，投影后在新样本空间，构造属性假定为相互独立(如原样本是来自同协方差多元正态分布或经过非线性核映射后为同协方差多元正态分布，则为真实)。在样本完成变换后，分类器执行 NB 的功能，在新样本空间进行参数学习，获得 DANB 分类器。对于来自未知类的实例，首先要进行投影变换，然后用 DANB 进行分类。

DANB 分类模型算法伪代码描述：

```
(1) begin: select LDA or KDA
(2)    if LDA then
(3)       compute ,mi,m,Sw,SB
(4)       solve (28)式得
(5)       project: 将原样本集向 W 投影,得投影后新样本集
(6)    endif
(7)    if KDA then
(8)       select Kernel unction(如:
(9)       多项式核,高斯核等)
(10)      compute,K,K
```

(11) solve (40)式得 A
(12) project:由(42)式计算出经核非线性映射后的新样本集
(13) endif
(14)train: 用新样本集训练 NB分类器的参数
(15)end

表 1 试验中所用数据集描述

	Data Set	#Attributes	#Class	#Instances	
				Train	Test
1	diabetes	8	2	786	CV-10
2	german-numeric	24	2	1000	CV-10
3	glass	9	7	214	CV- 5
4	ionshere	34	2	351	CV-10
5	iris	4	3	150	CV-10
6	liver-disorder	6	2	345	CV-10
7	segment	19	7	2310	CV-10
8	sonar	60	2	208	CV-10
9	vehicle	18	4	846	CV-10
10	waveform	40	3	5000	CV-10

4 实验评估

在表 1 所列 10 个数据集上进行试验，数据来源于 UCI 库[14]。试验的目的主要是对 DANB 与 NB 和 TAN 分类器在每个数据集上的分类正确率进行比较，其中 TAN 分类器采用的是基于条件互信息熵的方法[1]。分类器的精度测试采用分层交叉验证(stratification cross-validation, CV)。目前本文没有考虑处理缺失数据和性质变量(如有或没有某种特性)，故均选取无数据缺失和性质变量的数据集。当采用 KDA 提取判别信息时，由于 K 为 $n\times n$ 矩阵，因此对于较大的数据集，考虑到计算量问题，在每层仅随机抽取部分训练样本的数据来计算判别信息。虽然，这样做不能充分利用数据信息，但可以加快计算速度，减少对计算机资源的要求，是计算精度与计算效率的一种折衷。本实验采用多项式核与高斯核做非线性映射，当然满足 Mercer 定理[12]，其它形式的核函数也可以使用。

多项式核(Polynomial Kernel)：$k(x,y)=(x\cdot y+1)^d$，d 为多项式的阶数。

高斯核(Gaussian Kernel)：$k(x,y)=\exp(-\|x-y\|^2/2\sigma^2)$，$\sigma$ 为可调参数。

从表 2 试验结果可以看出，DANB 在大部分实验数据集上取得了最好的分类性能。对 1－2－5－9－10 数据集分类正确率比 NB 和 TAN 分类器的分类正确率都要高，且在所有数据集上明显超过 NB，仅在 3－7－8 数据集上表现低于 TAN，这些可以从图 1 中 NB 与 DANB 对比和 TAN 与 DANB 对比看出。另外，在图 1 中还给出了当采用 LDA 时，DANB 同 NB 和 TAN 的对比。可以发现，在多数情况下，仅采用 LDA 计算判别方向，就能获得较好的分类精度。同时，由于 DANB 具有维度缩减的作用，比如 4 数据集，属性从 34 维降到 1 维，10 数据集从 40 维降到 2 维，且分类性能上升，这无疑是非常有用的。一旦获得分类器后，对未标类样本的分类是极为简单的。通常属性维度远高于类别数，因而 DANB 所带来的好处就显而易见了。

表 2 DANB 和 NB、TAN 的试验结果比较

	DataSet	NB	TAN	DANB					
				LDA	Polynomial Kernel		Gaussian Kernel		
					d=2	d=3	sigma=1	sigma=2	sigma=2.5
1	diabetes	0.76	0.75	**0.77**	0.76	0.76	0.67	0.66	0.66
2	german-numeric	0.73	0.74	**0.77**	0.74	0.71	0.70	0.70	0.69
3	glass	0.44	0.72	0.58	0.53	0.53	0.64	0.65	0.64
4	ionshere	0.75	0.92	0.85	0.89	0.92	0.94	0.94	**0.95**
5	iris	0.95	0.94	**0.98**	**0.98**	**0.98**	0.97	0.97	**0.98**
6	liver-disorder	0.52	0.56	0.63	0.68	**0.69**	0.60	0.63	0.65
7	Segment*	0.75	**0.95**	0.91	0.80	0.67	0.16	0.17	0.21
8	sonar	0.61	**0.78**	0.65	0.62	0.67	0.67	0.67	0.71
9	vehicle	0.45	0.72	**0.79**	0.45	0.42	0.29	0.27	0.29
10	Waveform*	0.80	0.81	0.83	0.84	0.82	0.33	0.37	0.75

注：右角上方带 * 的数据集，提取核判别方向信息时，按每层随机抽取 20%的训练数据计算判别方向。

结论 朴素贝叶斯分类器是一种简单而有效的分类算法，但它的独立性假定使其无法表达实际数据中属性间存在的相依关系。目前有许多方法和技术采用进一步表达相依关系或通过变量选择以满足独立性假定，来改进朴素贝叶斯分类器的性能。本文提出了一种新型分类模型 DANB，它从另一角度出发，致力于弥补 NB 不能提取类间信息的缺陷，通过使用线性判别分析与核判别分析的方法寻找类间最大可分离的投影空间，然后再将原样本向最大可分离空间投影，获得新样本，以判别量为新属性，用 NB 算法在新样本中进行学习，从而将 NB 与判别分析方法有机地结合起来。在满足各类(或经非线性映射后各类)同协方差矩阵、正态分布的假定下，不仅新样本属性即判别量是严格满足类条件独立的，而且实现了维度缩减，因此 DANB 是对原样本信息的精确表达。当然，各类同协方差矩阵、多元正态分布是较强的假定，现实中往往难以满足。在这种情况下，DANB 对原始信息的表达是粗略的，判别量之间可能不再满足类条件独立假定。正如 NB 在不满足类条件独立性假定时，却同样表现出良好的分类性能，因此预计 DANB 也会如此。当然，进一步肯定需要大量的实证研究。实际上，表 1 中的实验数据集也并非都能满足这样强的条件，表现效果也都很出色。但毕竟试验量小了些，在此不作结论，留着以后进一步研究。在不满足多元正态性的场合，线性判别量可以看成是对总样本信息提供了一种近似。由于线性判别量是大量属性变量的线性组合，它们往往接近正态变量[15]，这也解释了为什么试验中仅采用 LDA 的 DANB 就表现良好分类性能的部分原因；当样本经过恰当的非线性映射后，在新的特征空间获得类间最佳可分离后，可以作出同上结论，也可解释采用 KDA 与 NB 结合提高分类性能的原因。然而，当采用 KDA 来提取分离信息时，

同其它核分析方法一样，选择恰当的核函数将是一个困难的问题。这往往需要对特定的分类问题的深入理解，在此基础上才能选择出合理的核函数。从表 2 中可以看到不同的核函数对分类性能产生了极大的影响，不恰当的核函数将造成分类性能下降而不是提高，如何选择恰当的核函数是需要进一步研究的问题。

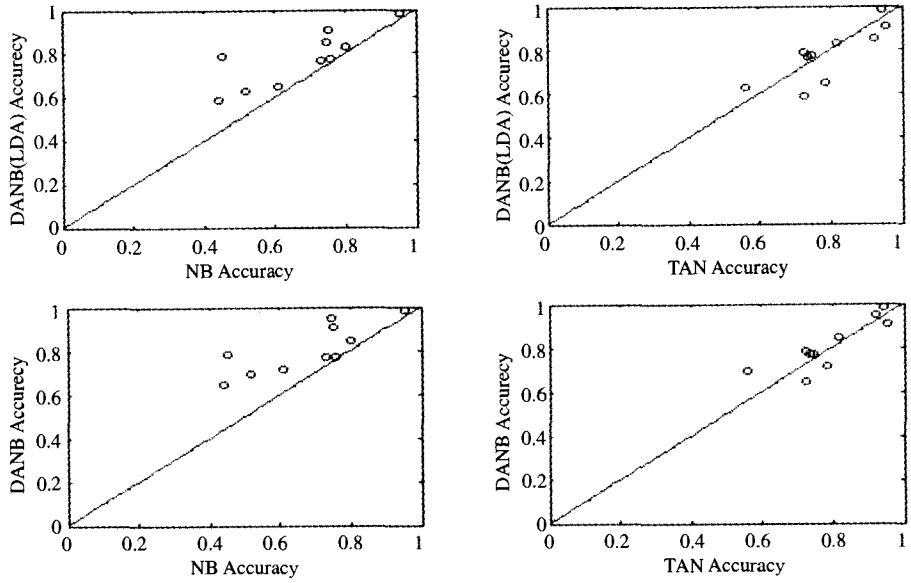

图 1　DANB、NB 和 TAN 分类精度对比散点图

注：对角线以上的点表示 DANB(DANB(LDA))的分类精度大于对比项的分类精度。

参 考 文 献

1 Fried N, Geiger D, Goldszmidt M. Bayesian network classifiers. Machine Learning, 1997, 29(2-3):131～163
2 Langley P, Iba W, Thompson K. An analysis of Bayesian classifiers. In: Proc. of the 10th National Conf on Artificial Intelligence. Menlo Park: AAAI Press, 1992. 223～228
3 Kononenko I. Semi Bayesian classifier. In: Proc. of the 6th European Working Session on Learning. New York: Springer-Verlag, 1991. 206～219
4 Pazzani M J. Searching for dependencies in Bayesian classifiers. In: Learning from Data: Artificial Intelligence and Statistics V. New York: Springer-Verlag, 1996. 239～248
5 Langley P, Sage S. Induction of selective Bayesian classifiers. In: Proc. of the 10th Conf. on Uncertainty in Artificial Intelligence. San Francisco: Morgan Kaufmann Publishers, 1994. 399～406
6 Webb G I, Pazzani M J. Adjusted probability naive Bayesian induction. In: Proc. of the 11th Australian Joint Conf. on Artificial Intelligence. Berlin: Springer-Verlag, 1998. 285～295
7 Kohavi R. Scaling up the accuracy of Naive-Bayes classifiers: A decision-tree hybrid. In: Proc. of the 2nd Int'l Conf. on Knowledge Discovery and Data Mining. Menlo Park: AAAI Press, 1996. 202～207
8 Keogh E J, Pazzani M J. A comparison of Distribution-based and Classification-based Approaches. In: Proc. of the Uncertainty' 99: The 7th Int'l Workshop on Artificial Intelligence and Statistics. San Francisco: Morgan Kaufmann Publishers, 1999. 225～230
9 Cheng J, Greiner R. Comparing Bayesian network classifiers. In: Proc of the 15th Conf on Uncertainty in Artificial Intelligence. San Francisco: Morgan Kaufmann Publishers, 1999. 101～108
10 Chickering D M, Geiger D, Heckerman D. Learning Bayesian networks is NP-complete. In: Learning from Data: Artificial Intelligence and Statistics V. New York: Springer-Verlag, 1996. 121～130
11 Xu Y, Yang J, Jin Z. Theory analysis on FSLDA and ULDA. Pattern Recognition, 2003, 36(12): 3031～3033
12 Shawe-Taylor J, Cristianini N. Kernel Methods for Pattern Analysis. Cambridge, UK, New York : Cambridge University Press, 2004
13 Baudat G, Anouar F. Generalized Discriminant Analysis Using a Kernel Approach. Neural Computation, 2000, 12: 2385～2404
14 http://www. ics. uci. edu/～ mlearn/MLRepository. html. Irvine, CA: University of California, Department of Information and Computer Science
15 Johnson R A, Wichern D W. 实用多元统计分析. 陆璇 译. 北京：清华大学出版社，2001

(上接第 101 页)

(6)根据公式(3-12)求出第 i 次的 ϕ_i；

(7)$i=i+1$，转入第 4 步；

(8)将节点上"已经处理"的标识清除；

(9)转入第 2 步。

$\theta(i)$是单调递减的，保证距离 N_l，N_m 越远的近邻对 N_l，N_m 相似性的影响越小。

通过精确语义相似性分析，最终得到的结果保存在精确语义相似性字典中(Exact Semantic Similarity Dictionary，简称 ESSD)。

$$ESSD=\eta^{\infty}=\{\langle N_l,N_m,\phi\rangle \mid N_l\in N(IS_1),N_m\in N(IS_2)\} \quad (3\text{-}12)$$

结论　在从半结构化的信息源建立本体的过程中，为了统一地分析和处理不同的信息源，本文提出了一个统一的概念模型，将各信息源转换为 SDS-G。该模型不仅完整地表现了各信息源的内涵，在计算语义相关性和语义距离时，还考虑了外延的影响。基于转换后得到的 SDS-G 模型，以节点名、节点属性、节点近邻为比较特征，使用节点名相似性分析方法、0 近邻相似性分析方法计算节点的基本语义相似性分析，然后使用近邻相似性分析方法修正已得到的相似性，提炼模式间更精确的语义相似性。本文提出的方法还解决了相似性分析中类型冲突的问题，方法是半自动化的，仅在早期需要少量的人类专家的参与。本文提出的方法在实验中也取得了较好的效果。

参 考 文 献

1 Gruber T. What is an Ontology? http://www-ksl. stanford. edu/kst/whst-is-an-ontology. html, 28/01/2002
2 Bergamaschi S, Castano S, Vincini M. Semantic integration of semistructured and structured data sources. SIGMOD Rec, 1999, 28(1):54～59
3 Buneman P, Davidson S, Fernandez M, et al. Adding structure to unstructured data. In: Proc. of International Conference on Database Theory (ICDT'97), 1997. 336～50