一种基于属性分割的产生式/判别式混合分类器*

石洪波，柳亚琴
（山西财经大学 信息管理学院，太原 030031）

摘 要：为了利用产生式和判别式方法各自的优势，研究了基于属性分割的产生式/判别式混合分类模型框架，提出了一种基于属性分割的产生式/判别式混合分类器学习算法 GDGA。其利用遗传算法，将属性集 X 划分为两个子集 X_G 和 X_D，并相应地将训练集 D 垂直分割为两个子集 D_G 和 D_D，在两个训练子集上分别学习产生式分类器和判别式分类器；最后将两个分类器合并形成一个混合分类器。实验结果表明，在大多数数据集上，混合分类器的分类正确率优于其成员分类器。在训练数据不足或数据属性分布不清楚的情况下，该混合分类器具有特别的优势。

关键词：分类；产生式；判别式；属性分割；遗传算法
中图分类号：TP181 **文献标志码**：A **文章编号**：1001-3695(2012)05-1654-05
doi:10.3969/j.issn.1001-3695.2012.05.014

Generative/discriminative hybrid classifier based on attributes partitioning

SHI Hong-bo, LIU Ya-qin
(*School of Information Management, Shanxi University of Finance & Economics, Taiyuan* 030031, *China*)

Abstract: In order to exploit the best of generative and discriminative approaches, on the basis of investigating a framework of generative/discriminative hybrid model based on attributes partitioning, this paper proposed a learning algorithm of generative/discriminative hybrid classifier based on attributes partitioning, GDGA. This algorithm divided the attributes X into two subsets, X_G and X_D, by applying genetic algorithms, and vertically partitioned the training set into two subsets D_G and D_D accordingly. Then it trained a generative classifier and a discriminative classifier on D_G and D_D respectively. In the final, it constructed a generative/discriminative hybrid classifier by combining the generative classifier and the discriminative classifier. Experimental results show that the generative/discriminative hybrid classifier performs better than its generative component and discriminative component on most data sets. This hybrid classifier has particular advantage in the case of unclear attribute distribution or not enough training data.

Key words: classification; generative; discriminative; attributes partitioning; genetic algorithms

0 引言

产生式方法(generative approaches)和判别式方法(discriminative approaches)是两种不同的分类方法[1~4]，具有各自的思路和框架。从概率的角度看，两种方法的建模对象不同。在产生式方法中，对实例 x 和相应类 y 的联合概率分布 $p(x, y) = p(x|y)p(y)$ 建模，采用联合似然函数最大化的方法估计条件概率 $p(x|y)$ 和先验概率 $p(y)$。对于新实例 x，利用 $p(x|y)$ 和 $p(y)$ 估计后验概率 $p(y|x)$，以确定 x 的所属类。为了估计条件概率 $p(x|y)$ 的值，需要对其服从的分布进行假设。一般地，当训练数据量较小或者假设的分布符合数据的真实分布时，产生式分类器优于与其对应的判别式分类器[2,5]。

与产生式方法不同，判别式方法并不关心每个类中数据的概率分布 $p(x|y)$，而是直接对类后验概率分布 $p(y|x)$ 建模，利用条件似然函数最大化或与不同损失函数相关联的目标函数的最优化来估计参数，由此产生的后验概率分布 $p(y|x)$ 用于新实例 x 类值 y 的估计。由于判别式方法直接关注于分类任务，因此，当给定足够训练数据时，判别式方法常常显示出很好的分类正确率[2,6]。

产生式和判别式方法具有不同的特点和优势，如何充分利用两种分类模型的优势，构造出具有更优分类性能的模型和学习算法值得研究。在现有的产生式/判别式混合方法中，主要有三种研究思路：

a)基本模型与辅助模型相结合的混合方法。在这类方法中，以产生式和判别式模型之一作为基本分类模型，另一种模型为辅助模型，将产生式模型中的部分结果用于判别式模型中或者利用判别式方法学习产生式模型。Jaakkola 等人[7]研究了从产生式模型中抽取用于判别式分类方法的核函数的模式。Tu[8]提出了一种混合分类器学习框架，它利用辅助样本获得伪负类数据，加上原有的正类数据递归地学习判别式模型，最后该模型收敛于产生式模型。

b)基于包含两种分类模型成分的目标函数的混合方法。McCallum 等人[9]提出了一个多条件学习框架 MCL，其目标函数是产生式概率和判别式概率的加权积。Lasserre 等人[10]提出一个产生式与判别式的混合框架，从一个共同的联合模型导出的判别式和产生式成员分类器具有各自的参数集。

收稿日期：2011-10-11；**修回日期**：2011-11-24 **基金项目**：国家自然科学基金资助项目(60873100)；山西省自然科学基金资助项目(2009011017-4)

作者简介：石洪波(1965-)，女，山西太原人，教授，博士，主要研究方向为机器学习、数据挖掘(shb710@163.com)；柳亚琴(1981-)，女，山西柳林人，讲师，博士研究生，主要研究方向为智能信息管理.

c)基于属性分割的混合方法。Raina 等人[6,11]提出了基于属性分割的产生式/判别式混合分类器,将属性矢量分割成具有不同权值的多个子矢量。在 Kang 等人[12,13]提出的产生式/判别式混合方法中,分别采用启发式搜索方法和统计测试方法将属性集 X 分割为两个子矢量。

在基于属性分割的产生式/判别式混合方法中,关键问题是如何分割属性集 X。文献[6,11]根据文本数据领域的特征,将属性集分割为文本正文和附加信息(包括标题、作者、链接等)两个子集,在两个子集上分别采用产生式和判别式方法。这种属性分割方法需要结合应用领域的数据特征,当应用领域数据特征不明确时,则很难采用这种分割属性集的方法。文献[12]采用一个迭代的、启发式的方法来分割属性集,该方法从属性分割($X_D = \{\}$, $X_G = X$)开始,从 X_G 中逐步移动单个属性到 X_D,使得属性的移动产生最大的性能改进,直到找不出这样的属性为止。这种方法与属性选择采用的包装法的思路类似,它将属性集的分割与分类算法相结合,可以得到较好的分类准确率。但文献[12]采用的单向搜索的启发式策略,可能得到属性分割空间中的局部最优解。文献[13]利用条件概率分布的统计测试(单变量 Shapiro-Wilk 测试)来分割属性。这种方法与属性选择采用的过滤法思路类似,在该方法中,属性集的分割过程独立于具体的分类算法,可看做是分类算法的预处理。这种方法的分类正确率依赖于属性分割所采用的评价函数的合理性。文献[13]采用的是单变量统计测试,而不是对属性集上的完全统计测试,因此有可能无法得到最佳属性分割。

遗传算法[14]是一种随机、自适应搜索算法,具有内在的隐并行性和很好的全局寻优能力。本文给出了基于属性分割的产生式/判别式混合分类模型框架,提出了一种基于属性分割的产生式/判别式混合分类器学习算法 GDGA(a learning algorithm of generative/discriminative hybrid classifier by applying genetic algorithms)。

1　产生式与判别式模型

产生式和判别式模型是两种不同的分类模型。产生式模型对包括训练实例 x 和相应类标签 y 的完整统计模型 $p(x,y)$ 建模。由于 $p(x,y)$ 可以拆分为先验概率 $p(y)$ 和条件概率 $p(x|y)$ 两个概率分布项,因此可以分别对 $p(y)$ 和 $p(x|y)$ 建模。条件概率分布 $p(x|y)$ 反映了每个类中数据的概率分布或生成过程,因此也称 $p(x|y)$ 为数据生成过程(data generating process, DGP)。从训练数据中学习数据生成过程 $p(x|y)$ 是产生式模型学习的关键。

与产生式模型不同,判别式模型则是关注于分类问题本身,将计算资源聚焦于与分类任务紧密相关的后验概率分布 $p(y|x)$,直接对 $p(y|x)$ 建模。判别式模型的这种直接对分类问题建模的思想与奥卡姆剃刀(Occam's razor)的原则一致。

为了清楚地解释产生式和判别式模型的差别,文献[8]用两类问题解释了产生式和判别式模型,如图 1 所示。图中的小圆圈表示正类实例,小方块表示负类实例,分别用“+1”和“-1”表示正类和负类。给定数据实例 x,判别式模型估计 x 是正类或负类的概率 $p(y|x)$,而产生式模型则通过对 $p(x|y=+1)$ 和 $p(x|y=-1)$ 建模来得到每个类中实例 x 的条件概率,即 x 的生成过程。

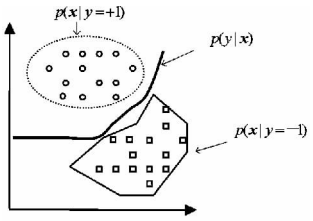

图1　产生式与判别式模型的说明

从图 1 中可以很明显地看出,判别式模型主要关注于如何将正类和负类数据分开,而产生式模型则试图理解每个类的基本信息,如各类的分布情况。如果一个数据实例属于正类,但是离分类边界很远,那么在这种情况下,利用判别式模型则可能会计算出一个很高的概率值预测这个实例属于正类,但利用产生式模型计算出的概率值却可能预测这个实例不属于正类。

此外,还有许多关于产生式和判别式模型理论和实验比较的研究结果。线性判别分析 LDA (linear discriminant analysis)和线性逻辑斯蒂回归 LLR(linear logistic regression)分别是典型的产生式模型和判别式模型。Efron[5]对两者进行了比较,得出下列观察结果:如果条件概率 $p(x|y)$ 和先验概率 $p(y)$ 的建模正确,则 LDA 比 LLR 更有效;如果训练数据不符合假设的分布,LLR 的分类正确率则优于 LDA 的正确率。Rubinstein 等人[1]也提出建议,如果分类模型正确性的可信度较高,则最好采用产生式方法。Ng 等人[2]以 LLR 为判别式分类器、以朴素贝叶斯为产生式分类器,对产生式和判别式方法进行了理论和实验比较,结果发现:判别式分类器具有较低的渐近误差。而产生式分类器以较快的速度接近其渐近误差,也就是说,对于较大的训练数据集,判别式分类器的正确率更高;而对于较小的训练数据集,产生式分类器的正确率则更高。

2　产生式/判别式混合模型学习算法

2.1　基于属性分割的产生式/判别式混合模型

令大写字母表示变量名,如 X_i、Y,黑体大写字母表示变量集,如 $\boldsymbol{X}_D$,小写字母用来表示相应大写字母表示的变量或变量集的具体取值,如 x_D 是变量集 $\boldsymbol{X}_D$ 的一个取值,y 是变量 Y 的一个取值。考虑一个离散或连续属性变量的有限集 $\boldsymbol{X} = \{X_1, X_2, \cdots, X_n\}$ 和类变量 Y,独立同分布训练集 $D = \{(x_1, y_1), \cdots, (x_i, y_i), \cdots, (x_m, y_m)\}$,其中 $x_i = (x_{i,1}, \cdots, x_{i,n})$。

以文献[1,2,5]等的研究结果为基础,本文给出基于属性分割的产生式/判别式混合分类模型。分类的目标是正确预测给定实例 x 的类值 y。如果分类性能度量准确,则 x 的最佳预测应该是使类后验概率 $p(y|x)$ 取最大值的类[15]。在基于属性分割的产生式/判别式混合模型中,属性集 $\boldsymbol{X}$ 被分割为两个子集 $\boldsymbol{X}_G$ 和 $\boldsymbol{X}_D$,其中 $\boldsymbol{X}_G$ 将应用于产生式模型,$\boldsymbol{X}_D$ 将应用于判别式模型。因此,混合分类模型的目标是估计给定实例 $x = (x_G, x_D)$ 类值为 y 的后验概率 $p(y|x_G\ x_D)$。

依据贝叶斯定理和链规则,$p(y|x_G, x_D)$ 可以表示为

$$p(y|x_G,x_D) = \frac{p(x_G)p(y|x_G)p(x_D|x_G,y)}{p(x_G,x_D)} \tag{1}$$

为了使得产生式分类器和判别式分类器能够独立训练,本文假设给定类 y 时,属性集 $\boldsymbol{X}_G$ 和 $\boldsymbol{X}_D$ 是块条件独立,即 $p(x_G, x_D|y) = p(x_G|y)\ p(x_D|x_G,y) = p(x_G|y)p(x_D|y)$。这一假设比朴素贝叶斯分类器的属性条件独立假设弱得多。给定这一假设,则式(1)右边可以表示为

$$\frac{p(x_G)}{p(x_G,x_D)}\times p(y|x_G)p(x_D|y)= \tag{2}$$

$$\frac{p(x_G)p(x_D)}{p(x_G,x_D)}\cdot\frac{p(y|x_G)p(y|x_D)}{p(y)} \tag{3}$$

对于新实例 $x=(x_G,x_D)$,项 $(p(x_G)p(x_D))/p(x_G,x_D)$ 与类值无关,其值不会影响分类的结果。因此,根据最大后验准则,基于属性分割的产生式/判别式混合分类模型框架可表示为

$$\arg\max_y \frac{p(y|x_G)p(y|x_D)}{p(y)} \tag{4}$$

式(4)表示的混合模型框架与文献[13]采用的混合分类模型框架类似,但是文献[13]假设每个类的先验概率 $p(y)$ 都是相等的,其目标是使 $p(y|x_G)p(y|x_D)$ 取最大值,这种先验概率相等的假设对不平衡数据集的分类有一定影响。

式(4)表示的混合分类模型由两个成员分类器构成,一个是产生式分类器,另一个是判别式分类器。每个成员分类器各自拥有一组参数。依据属性分割 (X_G,X_D),相应地产生两个训练子集 D_G 和 D_D。产生式分类器通过训练子集 D_G 上的似然函数最大化或对数似然函数最大化来估计产生式分类器的最佳参数 $\hat{\theta}$。产生式分类器的对数似然函数可表示为

$$LL_{\text{gen}}(\theta)=\sum_{i=1}^{m}\log\{p(y_i)p(x_i|y_i)\} \tag{5}$$

判别式分类器通过训练子集 D_D 上的条件似然函数最大化或条件对数似然函数最大化来估计判别式分类器的最佳参数 $\hat{\alpha}$。判别式分类器的条件对数似然函数可表示为

$$CLL_{\text{dis}}(\alpha)=\sum_{i=1}^{m}\log p(y_i|x_i) \tag{6}$$

2.2 基于属性分割的产生式/判别式混合分类器学习算法

为了学习上述的产生式/判别式混合分类器,一个关键问题是如何分割属性集 X。作为一种归纳学习策略,遗传算法具有内在的隐并行性和全局寻优能力,已经显示出比许多随机搜索方法和局部搜索方法更好的改进。遗传算法对最优属性分割问题的可应用性是很明显的。本文利用遗传算法,采用类似于属性选择包装法的思路实现属性集分割。

应用遗传算法主要考虑个体的表示方法、适应度函数的选取、个体选择策略等问题。

1)个体编码和表示 群体中的每个个体代表一个候选的属性分割方案。对于给定的属性集 $X=\{X_1,X_2,\cdots,X_n\}$,最简单的表示形式是二元表示,即每个属性分割由一个 n 维二元矢量来表示(其中 n 为属性个数),矢量的每一位对应于一个属性,如果矢量的某位取值为"1",表明该位对应的属性被选入 X_D 属性集;否则,对应的属性被选入属性集 X_G。

2)适应度函数 遵循多准则最优化函数的思路,本文设计的适应度函数 $\text{fitness}(X_G,X_D)$ 包括两个准则:$|X_D|$ 和 $\text{accuracy}(X_G,X_D)$,表示为

$$\text{fitness}(X_G,X_D)=(\text{accuracy}(X_G,X_D),1/|X_D|) \tag{7}$$

(1)$\text{accuracy}(X_G,X_D)$ 具有属性分割 (X_G,X_D) 的产生式/判别式混合分类器的分类正确率。分类正确率表示分类器的分类性能,有助于不同分类器之间的比较,所以本文选择它作为适应度函数的主要部分。产生式/判别式混合分类器的正确率采用测试集上混合分类器正确分类的实例占全部测试实例的百分比来估计。

(2)$|X_D|$ X_D 中的属性个数。在评估属性分割的每一轮中,不同属性分割的分类正确率可能相同。为了区分具有相同分类正确率的不同属性分割的重要性,本文在适应度函数中引入 X_D 中的属性个数。一般地,判别式分类器训练的时间复杂度较高,为了提高产生式/判别式混合分类器的学习效率,希望选择 X_D 中属性个数较少的属性分割,即 $1/|X_D|$ 值较大的属性分割。

(3)个体选择策略 个体选择采用转盘赌的策略。该策略按照一个个体的适应度值占当前属性分割种群中所有个体的适应度值之和的比例来决定该个体被选择的概率。这种方法不能保证最好的属性分割进入下一代,仅仅能够保证最好的属性分割有很大机会进入下一代。设当前种群为 $\{(X_G,X_D)^{(1)},\cdots,(X_G,X_D)^{(i)},\cdots,(X_G,X_D)^{(K)}\}$,其中的个体数为 K,则个体 $(X_D,X_G)^{(i)}$ 被选择的概率为

$$P(X_G,X_D)^{(i)}=\frac{\text{fitness}(X_G,X_D)^{(i)}}{\sum_{j=1}^{K}\text{fitness}(X_G,X_D)^{(j)}} \tag{8}$$

当进化代数达到固定值或达到一个极值使得后续的迭代不再能够产生更好的结果时,算法就终止。以下为完整的 GDGA 算法。

输入:具有属性集 X 的训练数据集 D,产生式分类器 Gen,判别式分类器 Dis。

输出:产生式/判别式混合分类器。

a)随机选取属性集 X 的 K 个不同属性分割 $(X_G,X_D)^{(i)}$ $(i=1,2,\cdots,K)$ 作为初始群体 $P^{(0)}$;

b)调用过程 $\text{compute\_fitness}(X_G,X_D,D)$ 计算群体 $P^{(0)}$ 中的每个属性分割 $(X_G,X_D)^{(i)}$ 的适应度值;

c)$j=1$;

d)Repeat{

(a)按照式(8)计算群体 $P^{(j-1)}$ 中每个属性分割 $(X_G,X_D)^{(i)}$ 被选择的概率 $P(X_D,X_G)^{(i)}$;

(b)选用转盘赌策略,从群体 $P^{(j-1)}$ 中选择部分属性分割,组成群体 $P^{(j)}$;

(c)应用交叉和变异算子形成新个体,对 $P^{(j)}$ 进行重组;

(d) 调用过程 $\text{compute\_fitness}(X_G,X_D,D)$ 计算群体 $P^{(j)}$ 中的每个属性分割 $(X_G,X_D)^{(i)}$ 的适应度值;

(e)$j=j+1$;

}Until 满足终止条件;

e)返回具有最佳适应度值的混合分类器。

过程 $\text{compute\_fitness}(X_G,X_D,D)$

(a)根据属性分割 (X_G,X_D),将训练集 D 分割为子集 D_G 和 D_D;

(b)在 D_G 和 D_D 上分别训练分类器 Gen 和 Dis;

(c)估计混合分类器的适应度值 $\text{fitness}(X_G,X_D)$;

(d)返回适应度值 $\text{fitness}(X_G,X_D)$。

2.3 算法复杂度

以朴素贝叶斯和线性逻辑斯蒂回归作为 GDGA 的两个成员分类器来分析 GDGA 算法的复杂度。朴素贝叶斯的理论时间复杂度为 $O(mn_G)$,而线性逻辑斯蒂回归的时间复杂度为 $O(mn_D^2t)$。其中:m 是训练实例数;n_G 是 X_G 中的属性个数;n_D 是 X_D 中的属性个数;t 是线性逻辑斯蒂回归中的迭代次数。假设遗传算法的最大迭代次数为 q,在遗传算法的每一次迭代中,朴素贝叶斯和线性逻辑斯蒂回归最多各执行 s 次,则 GDGA 算法的时间复杂度为 $O(qsmn_G+qsmn_D^2t)$。从以上分析可知,该算法的时间复杂度与属性分割的结果有关,当 X_D 中的属性个数较多时,其时间复杂度较高;反之,时间复杂度则较低。由于算法中两个成员分类器是在两个训练子集上分别训练的,训练过程和结果相互不影响,因此可以考虑同时学习两个成员分类器。在分布式环境下,遗传算法每次迭代中的若干

对产生式和判别式可以在分布式环境的每个节点上分别运行,最后将各节点的结果合并,从而提高算法的效率。

3　实验

3.1　实验数据和方法

为了检验产生式/判别式混合分类算法 GDGA 的有效性,从 UCI 机器学习数据库[16] 中选取了 17 个实验数据集。在选取的实验数据集中,既有实例个数较少的数据集,如 Balloon、Promoter gene sequences,也有实例个数较多的数据集,如 Annealing、Tic-tac-toe end game;既有两个成员分类器的分类正确率差异较大的数据集,如 Annealing、Tic-tac-toe end game、House-vote-84,又有两个成员分类器的分类正确率差异较小的数据集,如 Echocardiogram、Wine recognition、Promoter gene sequences。表 1 列出了 17 个数据集的实例个数、属性个数、类个数等信息。

表 1　训练数据集描述

数据集	实例个数	属性个数	类个数
Annealing	898	38	5
Balance scale	625	4	3
Balloon	76	4	2
Bridges2	108	11	6
Cleveland	303	13	2
Echocardiogram	132	8	2
Horse-colic	368	22	2
House-vote-84	435	16	2
Ionosphere	351	34	2
Lymphography	148	18	4
Pima indians diabetes	768	8	2
Promoter gene sequences	106	57	2
Sonar	208	60	2
Tic-tac-toe end game	958	9	2
Transfusion	748	4	2
Wdbc	569	30	2
Wine recognition	178	13	3

实验的目标是对产生式/判别式混合分类器 GDGA 与它的两个成员分类器进行比较。文献[2]对线性逻辑斯蒂回归和基于正态分布的朴素贝叶斯分类器进行了理论和实验的比较,得出一些重要的结论。因此,本文采用基于正态分布的朴素贝叶斯分类器作为产生式分类器,采用线性逻辑斯蒂回归作为判别式分类器。

所有的实验在 weka[17] 系统上运行,它提供了机器学习许多流行的学习算法。本文利用 weka 系统中的 weak. Classifiers. NaiveBayes(简记为 NB)和 weak. Classifiers. SimpleLogitic(简记为 LLR)作为两个成员分类器,并在 weka 上实现了基于属性分割的产生式/判别式混合分类器 GDGA。对于包含丢失值的数据集,实验中统一将其看做一个单独值来处理。

3.2　实验结果和分析

实验对 GDGA 分类器与它的两个成员分类器——朴素贝叶斯分类器 NB 和线性逻辑斯蒂回归分类器 LLR——在每个数据集上的分类正确率进行比较。每个分类器的分类正确率是测试集上正确预测的实例数占总实例数的百分比,采用 10 重交叉验证估计分类器的正确率,10 重交叉验证在每个数据集上分别测试 10 次,每次采用不同的 10 重划分方法。

表 2 列出了 NB、LLR 和 GDGA 三个分类器在 17 个数据集上的 10 次 10 重交叉验证测试的平均正确率及标准离差。最后一行列出了三个分类器在 17 个数据集上分类正确率的平均值,GDGA 分类器比 NB 分类器提高约 4.6%,比 LLR 分类器提高 1.2%。图 2 显示了 GDGA 分类器与朴素贝叶斯分类器和线性逻辑斯蒂回归分类器比较的散布图。在图 2 中,对角线之上的点对应于 GDGA 分类正确率较高的数据集;对角线之下的点对应于 GDGA 分类正确率较低的数据集。从图 2 中可以直观地看出,GDGA 分类器在大多数数据集上比单个朴素贝叶斯或线性逻辑斯蒂回归分类器具有更好的分类正确率。

表 2　三个分类器的分类正确率

数据集	NB	LLR	GDGA
Annealing	81.9376 ± 0.30	88.5412 ± 0.29	91.3252 ± 0.56
Balance scale	90.576 ± 0.18	87.68 ± 0.33	90.576 ± 0.18
Balloon	79.0790 ± 2.19	76.7105 ± 3.78	80.2632 ± 1.75
Bridges2	67.6190 ± 1.27	68.8571 ± 1.27	69.0476 ± 2.39
Cleveland	82.869 ± 0.63	82.3763 ± 0.77	83.4323 ± 1.14
Echocardiogram	75 ± 2.49	74.4595 ± 2.16	75.4054 ± 3.42
Horse-colic	77.3370 ± 0.56	80.1902 ± 0.61	81.7663 ± 1.08
House-vote-84	90.1379 ± 0.07	96.0229 ± 0.55	95.8851 ± 0.48
Ionosphere	82.5071 ± 0.47	87.8063 ± 0.60	90.9117 ± 0.94
Lymphography	82.8378 ± 0.97	83.3108 ± 2.40	84.5270 ± 1.37
Pima indians diabetes	75.6641 ± 0.38	77.0833 ± 0.52	76.7578 ± 0.60
Promoter gene sequences	91.5094 ± 1.18	91.3207 ± 1.77	91.4151 ± 2.45
Sonar	68.0769 ± 1.27	74.2308 ± 1.74	75.9615 ± 1.28
Tic-tac-toe end game	69.6451 ± 0.09	98.2568 ± 0.37	98.2568 ± 0.09
Transfusion	75.0669 ± 0.25	76.8316 ± 0.25	79.0374 ± 0.22
Wdbc	93.3392 ± 0.24	97.5219 ± 0.40	97.4692 ± 0.43
Wine recognition	97.2472 ± 0.41	97.5281 ± 0.54	97.5281 ± 0.76
Average	81.2029	84.6311	85.8568

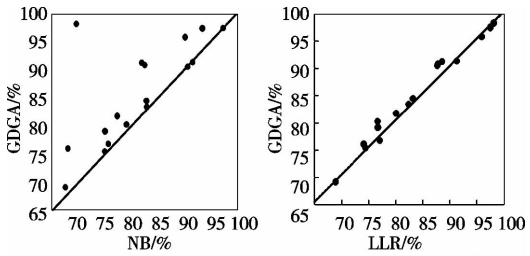

图2　混合分类器与朴素贝叶斯和线性逻辑斯蒂回归分类准确率比较的散布图

从表 2 中可以观察到,对于某些数据集,如 Echocardiogram、Wine recognition 和 Promoter gene sequences,两个成员分类器的分类正确率不存在明显差异,GDGA 在这些数据集上的分类正确率接近于其两个成员的正确率。分析其原因,可能是由于在这些数据集上,朴素贝叶斯模型中定义的 DGP 既不是很好也不是很差。与这些数据集不同,对于 House-vote-84 和 Tic-tac-toe end game 等数据集,两个成员分类器的分类正确率存在明显差异,GDGA 的分类正确率接近于它的具有较高正确率的成员分类器的正确率。总体来看,几乎对于所有的数据集,GDGA 的分类正确率总是优于和接近于它的两个成员分类器中的优者。

图 3 给出了在训练数据集上建模时属性分割的结果。图中的横坐标表示 17 个数据集,纵坐标表示属性组中的属性个数。每个数据集对应两组属性,第一组属性用实心柱表示,第二组属性用斜线柱表示。每一对柱子上面的数值代表两组属性中的属性个数。例如,数据集 Annealing 对应于数值对(5,33),表示数据集 Annealing 的 38 个属性分为两个子集,分别包含 5 个属性和 33 个属性。依据这个属性分割,数据集 Annealing 被分割为两个子集,朴素贝叶斯和线性逻辑斯蒂回归分别在两个数据子集上训练。

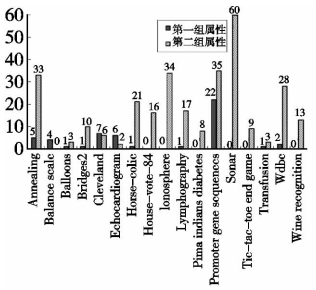

图3 在训练集上的属性分割结果

从图 3 中可以观察到,部分数据集的分割结果出现了属性分割子集之一为空的情况,如 Balance scale、House-vote-84、Pima indians diabetes 和 Tic-tac-toe end game。对于这些数据集,GDGA 认为这些数据集的全部属性的 DGP 都是误定义或者都不存在误定义,因此,GDGA 混合分类器只使用了其中一个成员分类器,其分类正确率与该成员分类器的分类正确率相同或相近,如 GDGA 和朴素贝叶斯在 Balance scale 上的分类正确率相同,而 GDGA 和线性逻辑斯蒂回归在 Tic-tac-toe end game 上的分类正确率相同。对于属性分割的两个子集均不为空的数据集,GDGA 在大部分数据集上均取得了较好的分类性能。例如数据集 Annealing,算法将其分割为包含 5 个和 33 个属性的两个子集,混合分类器比两个成员分类器的分类性能明显提高。从图 3 中还可以看到,对于绝大多数数据集,第二组属性中的属性个数均多于第一组属性中的属性个数。这种现象与本文选择的两个成员分类器有关,如果选择其他成员分类器或改变产生式分类器的模型假设,属性分割的结果就可能发生较大的变化。

在实验过程中还发现以下现象:在原始数据集和它的不同数据子集上建模,得到的最佳属性分割可能不同。例如,对于数据集 Ionosphere,在全部训练数据上建模时属性分割出现了子集为空的情况(图 3),但在 10 重交叉验证评估时,参与建模的数据是全部训练数据的 90%,10 次建模得到的两个属性子集均不为空。分析其原因,可能是属性分割的结果与训练数据集紧密相关,当数据足够多时,抽取原始训练数据 90% 的数据形成的数据子集不会改变原始训练数据集的分布,从而不会影响属性分割的结果;但当原始训练数据集的数据量不足以描述其分布时,数据量的减少将可能会改变数据的分布,从而影响属性分割的结果。

4 结束语

本文研究了基于属性分割的产生式/判别式混合分类模型框架,提出了一种基于属性分割的产生式/判别式混合分类模型学习算法。该算法利用遗传算法将属性集 X 分割为两个子集 X_G 和 X_D,相应地,数据集 D 被划分为两个子集 D_G 和 D_D,产生式分类器和判别式分类器分别在 D_G 和 D_D 上进行训练;最后,采用某种方法将产生式分类器和判别式分类器合并形成产生式/判别式混合分类器。实验结果表明,在大多数数据集上,混合分类器的分类正确率优于或接近于它的两个成员分类器中的优者。

根据一般的观察,如果 DGP 的分布假设定义准确,产生式分类器比相应的判别式分类器具有更好的性能;如果 DGP 的分布假设与数据的实际分布相差较远,它将比相应的判别式分类器差。但是在实际应用中,对每个数据集常常难以判断模型的 DGP 定义是否很好地描述了数据的真实生成过程,在这种情况下,混合分类器 GDGA 具有特别的优势。在 GDGA 算法中,适应度的选择非常重要,如何选择更合适的适应度函数,以提高学习算法的准确率和效率值得进一步研究。

参考文献:

[1] RUBINSTEIN Y D, HASTIE T. Discriminative vs informative learning [C]//Proc of the 3rd International Conference on Knowledge Discovery and Data Mining. Menlo Park, CA: AAAI Press, 1997: 49-53.

[2] NG A Y, JORDAN M I. On discriminative vs generative classifiers: a comparison of logistic regression and naive Bayes[C]//Proc of Conference on Advances in Neural Information Processing Systems. Cambridge: MIT Press, 2002: 841-848.

[3] BISHOP C M, LASSERRE J. Generative or discriminative? Getting the best of both worlds (with discussion) [C]//Proc of the 8th World meeting on Bayesian Statistics. UK: Oxford Univer Sity Press, 2007: 3-24.

[4] XUE J H, TITTERINGTON D M. On the generative-discriminative tradeoff approach: interpretation, asymptotic efficiency and classification performance[J]. Computational Statistic and Data Analysis, 2010, 54(2): 438-451.

[5] EFRON B. The efficiency of logistic regression compared to normal discriminant analysis[J]. Journal of American Statistical Association, 1975, 70(352): 892-898.

[6] RAINA R, SHEN Y, NG A Y, *et al*. Classification with hybrid generative/discriminative models [C]//Advances in Neural Information Processing Systems. 2003: 545-552.

[7] JAAKKOLA T S, HAUSSLER D. Exploiting generative models in discriminative classifiers [C]//Proc of Conference on Advances in Neural Information Processing Systems Ⅱ. Cambridge: MIT Press, 1999: 487-493.

[8] TU Zhou-wen. Learning generative models via discriminative approaches[C]//Proc of IEEE Conference on Computer Vision and Pattern Recognition. Washington DC: IEEE Computer Society, 2007: 1-8.

[9] McCALLUM A, PAL C, DRUCK G, *et al*. Multi-conditional learning: generative/discriminative training for clustering and classification [C]//Proc of the 20th National Conference on Artificial Intelligence. Boston: AAAI Press, 2006: 433-439.

[10] LASSERRE J A, BISHOP C M, MINKA T P. Principled hybrid of generative and discriminative models [C]//Proc of IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Washington DC: IEEE Computer Society, 2006: 87-94.

[11] FUJINO A, UEDA N, SAITO K. A hybrid generative /discriminative approach to text classification with additional information[J]. Information Processing and Management, 2007, 43(2): 379-392.

[12] KANG C, TIAN Jin. A hybrid generative /discriminative Bayesian classifier[C]//Proc of the 19th International Florida Artificial Intelligence Research Society Conference. Melbourne Beach, Florida: AAAI Press, 2006: 562-567.

[13] XUE Jing-hao, TITTERINGTON D M. Joint discriminative-generative modelling based on statistical tests for classification [J]. Pattern Recognition Letters, 2010, 31(9): 1048-1055.

[14] DEJONG K. Learning with genetic algorithm: an overview[J]. Machine Learning, 1988, 3(2-3): 121-138.

[15] DUDA R, HART P. Pattern classification and scene analysis[M]. New York: Wiley, 1973.

[16] FRANK A, ASUNCION A. UCI machine learning repository [EB/OL]. (2010). http://archive. ics. uci. edu/ml.

[17] WITTEN I H, FRANK E. Data mining: practical machine learning tools and techniques with Java implementations[M]. Seattle: Morgan Kaufmann Publishers, 2000.