中国科学: 信息科学 2022 年 第 52 卷 第 10 期: 1792–1807
SCIENTIA SINICA Informationis
论文

基于特征增广的生成 – 判别混合模型构建方法

张文钧[1], 蒋良孝[1,2*], 张欢[1]

1. 中国地质大学计算机学院, 武汉 430074
2. 智能地学信息处理湖北省重点实验室, 武汉 430074
* 通信作者. E-mail: ljiang@cug.edu.cn

收稿日期: 2021–06–06; 修回日期: 2021–08–27; 接受日期: 2021–10–07; 网络出版日期: 2022–10–10

国家自然科学基金联合基金重点项目 (批准号: U1711267) 和中央高校基本科研业务费专项资金项目 (批准号: CUGGC03) 资助

摘要 从概率框架的角度来看, 生成模型首先由数据学习联合概率分布, 然后再求出条件概率分布, 通常具有更快的收敛速度; 而判别模型由数据直接学习条件概率分布, 往往具有更高的准确率. 生成 – 判别混合模型作为二者的有效结合, 同时集成了它们的优点. 然而, 现有方法在构建混合模型时, 需要将原始特征划分为两个独立的特征空间, 分别用于训练生成模型和判别模型. 特征划分不仅提升了模型的时间复杂度, 还削弱了原始特征空间的表达能力. 为了解决这一问题, 本文提出了一种基于特征增广的生成 – 判别混合模型构建方法. 该方法首先利用生成模型学习条件概率分布, 然后将学到的条件概率分布作为新特征增广到原始特征空间中, 最后在增广后的特征空间中训练判别模型并预测最终的分类结果. 该方法利用特征增广的思想做模型混合, 无需对原始特征进行划分, 具有较低的时间复杂度, 同时还增强了原始特征空间的表达能力. 在 36 个经典 UCI 标准数据集上的实验结果表明, 所提方法不仅具有有效性和通用性, 还遵循了偏差 – 方差权衡原则.

关键词 生成模型, 判别模型, 特征增广, 条件概率分布, 偏差 – 方差权衡

1 引言

从概率框架的角度来看, 机器学习所要实现的是基于有限的训练实例尽可能准确地估计出测试实例的条件概率分布. 概率统计模型依据条件概率分布估计策略的差异, 主要可分为生成模型和判别模型 [1]. 生成模型首先由数据学习联合概率分布, 然后依据联合概率分布求得条件概率分布. 生成模型表示了由给定实例产生预测类标记的生成关系, 典型的生成模型包括朴素贝叶斯 (naive Bayes, NB)、隐马尔科夫模型 (hidden Markov model, HMM) 和线性判别分析 (linear discriminant analysis, LDA) 等. 判别模型由数据直接学习条件概率分布, 更加关心给定实例应当被预测为什么样的类标记, 典型的判别模型包括逻辑回归 (logistic regression, LR)、支持向量机 (support vector machine, SVM)、k 近

引用格式: 张文钧, 蒋良孝, 张欢. 基于特征增广的生成 – 判别混合模型构建方法. 中国科学: 信息科学, 2022, 52: 1792–1807, doi: 10.1360/SSI-2021-0199
Zhang W J, Jiang L X, Zhang H. A feature augmentation-based method for constructing generative-discriminative hybrid models (in Chinese). Sci Sin Inform, 2022, 52: 1792–1807, doi: 10.1360/SSI-2021-0199

邻 (k-nearest neighbor, KNN)、决策树 (decision tree, DT) 和广义加性模型 (generalized additive model, GAM) 等 [2]. 生成模型和判别模型各有优缺点, 具体来说: 生成模型的收敛速度更快, 当样本容量增加时可以更快地收敛于真实模型, 此外生成模型可以还原出联合概率分布, 可被用于含有隐变量的学习问题; 判别模型直接对条件概率分布进行学习, 往往学习的准确率更高 [2].

为了同时集成生成模型和判别模型的优点, 学者们提出了多种生成 – 判别混合模型构建方法. 其中 Rubinstein 和 Hastie [3] 最早提出将生成模型和判别模型结合在一起, 他们试图将满足生成模型假设的特征用于学习生成模型, 其余特征用于学习判别模型, 最后组合生成模型和判别模型的输出作为最终输出. 按照这种基于特征划分的构建思路, 逐渐衍生出了许多不同的生成 – 判别混合模型构建方法 [4~7]. 然而, 特征划分不仅提升了模型的时间复杂度, 还削弱了原始特征空间的表达能力.

为了解决这一问题, 本文提出了一种基于特征增广的生成 – 判别混合模型构建方法. 该方法的主要步骤如下: 首先利用生成模型学习条件概率分布, 然后将学到的条件概率分布作为新特征增广到原始特征空间中, 最后在增广后的特征空间中训练判别模型并预测最终的分类结果. 本文的主要贡献包括以下 3 个方面:

(1) 概述了生成模型和判别模型的比较研究以及现有的生成 – 判别混合模型构建方法;

(2) 提出一种基于特征增广的生成 – 判别混合模型构建方法, 该方法利用特征增广的思想做模型混合, 无需对原始特征进行划分, 具有较低的时间复杂度, 同时还增强了原始特征空间的表达能力;

(3) 依据该方法, 选择 NB 和隐朴素贝叶斯 (hidden naive Bayes, HNB) 作为生成模型, LR 作为判别模型, 分别构建了 NB-LR 和 HNB-LR 混合模型, 并通过实验验证了新方法不仅有效、通用, 还遵循偏差 – 方差权衡原则.

2 相关工作

对于生成模型和判别模型的比较研究, 已经有很长的历史. Eforn [8] 通过仿真实验和理论分析, 对 LDA 和 LR 进行比较, 结论证明生成模型 LDA 的收敛速度快于判别模型 LR. Rubinstein 和 Hastie [3] 在 3 种模拟数据集上对生成模型和判别模型进行比较, 结果发现: 当模拟数据集的类密度满足高斯 (Gaussian) 分布时, 生成模型 LDA 的分类性能优于判别模型 LR; 当模拟数据集采用对数样条密度时, 判别模型 GAM 的分类性能优于生成模型 NB; 当模拟数据集中实例较少时, NB 的平均分类性能优于 GAM. 除此之外, Ng 和 Jordan [9] 最早对生成模型 NB 和判别模型 LR 进行了实验比较和理论研究, 结果发现 NB 和 LR 的分类性能与训练集大小有关. 具体而言, 若训练集足够大, LR 的分类准确度往往更高, 而当训练集较小时, NB 的分类准确度可能更高. 由于生成模型和判别模型各有优点, 如何构建生成 – 判别混合模型来集成两者的优点, 逐渐成为相关研究的重点.

在过去 20 年里, 有大量关于构建生成 – 判别混合模型的研究工作, 试图将生成模型和判别模型有效地结合在一起. Rubinstein 和 Hastie [3] 最早提出将生成模型和判别模型结合在一起, 他们试图将满足生成模型假设的特征用于学习生成模型, 其余特征用于学习判别模型, 最后组合生成模型和判别模型的输出作为最终输出. 不过遗憾的是, 虽然他们提出将生成模型和判别模型结合在一起, 但是并没有给出详细实验结果来验证他们的想法.

Raina 等 [10] 构建了一种以 NB 为生成模型, LR 为判别模型的混合模型, 来解决文本分类问题. 他们选择 USENET 的成对新闻组作为实验数据, 每条数据中包含两个区域, 分别是主题区域和信息区域. 主题区域的单词往往少于信息区域, 因此应当分别设置不同的权重. Raina 等利用 LR 从数据集中学习单词权重, 并在加权后的数据集上学习 NB, 最终得到了优于 NB 和 LR 的分类结果. Fujino

等 [11] 构建了和 Raina 等相似的混合模型, 来处理包含多种信息 (标题、超链接等) 的文本分类问题. 他们在每种信息中都分别训练了一个生成模型 NB, 然后利用最大熵原理学习到的权值对多个 NB 的输出进行加权组合. 他们在 4 个文本数据集上进行了实验, 实验结果表明混合模型的分类性能可同时优于 NB 和 LR. Bishop 和 Lasserre [12] 提出了一种启发式方法, 利用生成模型和判别模型的对数似然函数的凸组合在两者之间进行插值, 并将得到的混合模型应用于静态图像中的物体识别. 实验结果表明, 混合模型在生成模型和判别模型间得到了平衡, 同时集成了两者的优点, 取得了更好的表现.

上述混合模型都只针对于解决固定应用场景下的分类问题. 与这些方法不同, Kang 和 Tian [4] 提出了一种不限定于固定应用场景的构建生成 – 判别混合模型的新方法, 在该方法中, 生成模型为 NB, 判别模型为 LR. 该方法将原始特征空间划分为两部分, 分别用于训练 NB 和 LR. 具体来讲, 他们依据分类精度贪婪地为 LR 添加训练特征, 直到分类精度不再提高, 剩余未添加的特征用于训练 NB, 由此构建的混合模型被称为 HBayes-NB. 他们在 20 个不同的数据集上测试了 HBayes-NB 的性能, 并使用交叉验证来评估 HBayes-NB 的分类误差, 实验结果表明 HBayes-NB 对所有 20 个数据集的平均分类误差最小. 然而, 由于采用贪婪的特征划分策略, HBayes-NB 容易陷入局部最优, 且训练所需计算量也很大.

不同于上述混合模型构建思路, Xue 和 Titterington [5] 设计了一种生成 – 判别混合模型, 其中生成模型为 LDA, 判别模型为 LR. 该混合模型通过对所有原始特征进行单变量正态性测试, 将那些满足测试的特征用于训练生成模型 LDA, 其余特征用于训练判别模型 LR, 通过这种过滤的特征划分策略极大地降低了构建混合模型的计算量. Xue 和 Titterington [5] 在 6 个只有数字特征的数据集上测试了该混合模型的性能, 他们发现, 对于较小规模的数据集, 混合模型的分类错误率比 LR 和 LDA 更低.

Tan 等 [6] 通过削弱混合模型中生成模型 NB 的特征条件独立假设, 提出了一种混合模型的构建方法. 他们首先假设所有原始特征都用于训练 NB, 然后每次在用于训练 NB 的特征中计算所有特征对的条件互信息, 找到条件互信息最高的一组特征对作为待划分特征对. 然后对于这组特征对中的两个待划分特征, 从去除这两个特征之外剩余的训练 NB 的特征中各找一个特征, 使得各自组成的特征对的条件互信息最高. 最后比较这两组特征对的条件互信息, 将值更高的那一个特征对中的待划分特征用于训练 LR. 重复上述过程直到用于训练 NB 的特征不足两个, 或所有特征对的条件互信息均不大于 0.05. 该方法通过削弱 NB 的特征条件独立假设, 降低了混合模型的偏差, 但由于忽略了方差的影响, 所得混合模型的分类性能优于 NB, 但不如 LR.

在最新的工作中, Tan 等 [7] 对这一工作进行了完善, 提出了一种基于偏差 – 方差权衡的构建策略. 其中混合模型的偏差, 用平均的条件互信息度量, 而混合模型的方差, 则通过一个基于训练集实例个数的指数函数度量. 给定一个原始特征, 他们通过这种度量方法, 分别求出该特征用于训练 NB 产生的偏差增量和用于训练 LR 产生的方差增量, 并通过比较两者的大小决定该特征用于训练 NB 还是用于训练 LR. 用划分后的特征分别训练 NB 和 LR, 并组合两者的输出最为混合模型的最终输出. 实验结果证明, 他们构建的混合模型的分类性能同时优于 NB 和 LR.

3 方法

根据相关工作中的概述, 目前大多数的生成 – 判别混合模型在构建时, 都需要将原始特征划分到两个独立的特征空间, 分别用于训练生成模型和判别模型. 然而, 特征划分不仅提升了模型的时间复杂度, 还削弱了原始特征空间的表达能力. 为了解决这一问题, 本文提出了一种基于特征增广的生成 – 判别混合模型构建方法. 本节将阐述如何使用本文提出的新方法来构建生成 – 判别混合模型. 首先,

本节解释了使用生成模型做特征增广的原因, 并给出本文所选生成模型的详细定义; 之后, 本节介绍如何使用判别模型完成分类任务, 并给出本文所选判别模型的详细定义; 最后, 本节对混合模型的结构做详细说明, 并分别给出混合模型在训练阶段和测试阶段的详细算法流程和复杂度分析.

3.1 基于生成模型做特征增广

生成模型首先由数据学习联合概率分布, 然后再求出条件概率分布, 因此必然考虑

$$P(C|A_1,\ldots,A_j,\ldots,A_m)=\frac{P(A_1,\ldots,A_j,\ldots,A_m,C)}{P(A_1,\ldots,A_j,\ldots,A_m)},\tag{1}$$

其中 $A_1,\ldots,A_j,\ldots,A_m$ 表示 m 个特征变量, C 表示类变量. 依据式 (1), 生成模型需要先学习联合概率分布 $P(A_1,\ldots,A_j,\ldots,A_m,C)$, 再在此基础上求出条件概率分布 $P(C|A_1,\ldots,A_j,\ldots,A_m)$. 由于生成模型在学习条件概率分布时考虑到联合概率分布的影响, 因此若将生成模型学习的条件概率分布作为新特征增广到原始特征空间中, 再供判别模型学习, 有助于判别模型在分类过程中考虑到联合概率分布的影响.

在构建混合模型时, 本文分别选择 NB 和 HNB 作为生成模型. 其中, NB 在计算式 (1) 中联合概率分布 $P(A_1,\ldots,A_j,\ldots,A_m,C)$ 时引入特征条件独立假设, 特征条件独立假设要求给定类变量 C 时所有特征变量完全相互独立, 这使得 NB 的结构如图 1(a) 所示, 此时联合概率分布可通过以下形式计算:

$$P(A_1,\ldots,A_j,\ldots,A_m,C)=P(C)\prod_{j=1}^{m}P(A_j|C).\tag{2}$$

根据全概率法则, 式 (1) 中分母 $P(A_1,\ldots,A_j,\ldots,A_m)$ 等于特征变量 $A_1,\ldots,A_j,\ldots,A_m$ 与类变量 C 的联合概率分布对 C 中所有类标记的累和, 因此当使用 NB 为一个具体的测试实例 $\boldsymbol{x}$ 分类时, 若 $\boldsymbol{x}$ 可由特征值向量 $\langle a_1,\ldots,a_j,\ldots,a_m\rangle$ 表示, 其中 a_j 是第 j 个特征 A_j 的特征值, 用 c_k 表示类变量 C 的第 k 个类标记, NB 使用式 (3) 来计算给定测试实例 $\boldsymbol{x}$ 预测为类标记 c_k 的条件概率:

$$P(c_k|\boldsymbol{x})=\frac{P(c_k)\prod_{j=1}^{m}P(a_j|c_k)}{\sum_{k=1}^{q}P(c_k)\prod_{j=1}^{m}P(a_j|c_k)},\tag{3}$$

其中 $P(c_k|\boldsymbol{x})$ 为 NB 将测试实例 $\boldsymbol{x}$ 预测为类标记 c_k 的条件概率, $P(c_k)$ 是类标记 c_k 的先验概率, $P(a_j|c_k)$ 是给定类标记 c_k 时特征变量 A_j 取值为 a_j 的概率, $P(c_k)$ 和 $P(a_j|c_k)$ 分别用式 (4) 和 (5) 来估计:

$$P(c_k)=\frac{\sum_{i=1}^{n}\delta(c_i,c_k)+1}{n+q},\tag{4}$$

$$P(a_j|c_k)=\frac{\sum_{i=1}^{n}\delta(a_{ij},a_j)\,\delta(c_i,c_k)+1}{\sum_{i=1}^{n}\delta(c_i,c_k)+n_j},\tag{5}$$

其中 n 是训练实例个数, q 是类标记个数, c_i 是第 i 个训练实例的类标记, a_{ij} 是第 i 个训练实例的第 j 个特征值, n_j 是特征 A_j 的特征值个数, $\delta(\cdot)$ 是二元函数, 两个参数相同时为 1, 否则为 0.

特征条件独立假设的引入极大地降低了 NB 的复杂度, 使其成为鲁棒模型. 不过, 在实践中 NB 要求的特征条件独立假设很难成立, 这也给 NB 造成了较大的偏差. 目前已有大量文献探讨了如何通过削弱特征条件独立假设来提高 NB 的分类性能, 这些方法主要分为两类, 分别是结构扩展 [13,14] 和特征加权 [15,16]. HNB [14] 是一个经典的基于结构扩展的 NB 改进模型, 其网络结构如图 1(b) 所示. 图中 A_{hpj} 表示特征 A_j 的隐藏父特征节点, 包含了其他所有特征对 A_j 的影响. 作为 NB 的改进模型,

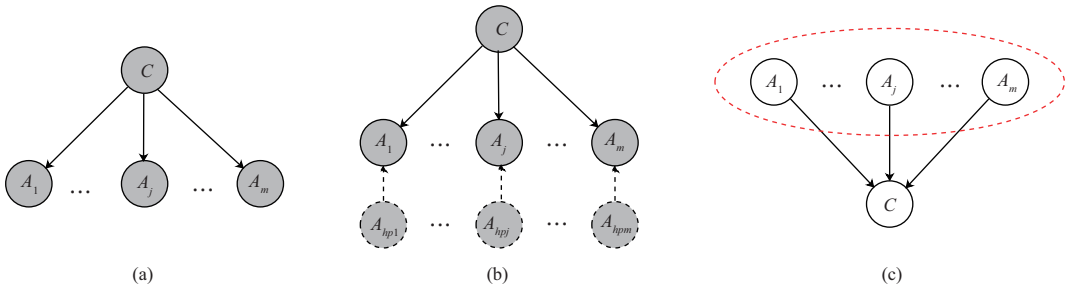

图 1 (网络版彩图) 模型的贝叶斯网络结构

Figure 1 (Color online) Bayesian network structure of models. (a) NB; (b) HNB; (c) LR

HNB 拥有更好的分类性能, 本文分别使用 NB 和 HNB 作为生成模型来构建混合模型, 以探究生成模型对新方法的重要性和新方法的通用性.

3.2 基于判别模型完成分类任务

依据前文分析, 判别模型直接由数据学习条件概率分布 $P(C|A_1,\ldots,A_j,\ldots,A_m)$, 往往学习的准确率更高 [2], 因此在本文所提的构建方法中选择由判别模型完成最终的分类任务. 在构建混合模型时, 由于考虑到特征增广会造成特征冗余, 因此本文选择基于 L2 正则化的 LR 作为判别模型. L2 正则化为 LR 的参数添加约束, 降低了 LR 因特征冗余而陷入过拟合的风险.

给定测试实例 $\boldsymbol{x}$, LR 假设对数几率 (log odds) 满足线性分布, 得到

$$\ln\frac{P\left(c_k|\boldsymbol{x}\right)}{P\left(c_q|\boldsymbol{x}\right)}=\boldsymbol{x}\boldsymbol{w}_k^{\mathrm{T}}+b_k,\quad k=1,2,\ldots,q-1, \tag{6}$$

其中 c_k 表示类变量 C 的第 k 个类标记, c_q 表示类变量 C 的最后一个类标记, $\boldsymbol{w}_k$ 和 b_k 便是 LR 需要学习的参数. 为了便于讨论, 本文用 $\hat{\boldsymbol{x}}$ 代替 $(\boldsymbol{x},1)$, 用 $\boldsymbol{\beta}_k$ 代替 $(\boldsymbol{w}_k,b_k)$, 使得 $\boldsymbol{x}\boldsymbol{w}_k^{\mathrm{T}}+b_k$ 简化为 $\hat{\boldsymbol{x}}\boldsymbol{\beta}_k^{\mathrm{T}}$. 针对多分类问题, 本文统一采用 $P\left(c_q|\boldsymbol{x}\right)$ 作为对数几率中的分母 [17]. 为了使得 $\sum_{k=1}^{q}P(c_k|\boldsymbol{x})=1$, 可从式 (6) 得出

$$P\left(c_q|\boldsymbol{x},\boldsymbol{\beta}\right)=\frac{1}{1+\sum_{k=1}^{q-1}\exp\left(\hat{\boldsymbol{x}}\boldsymbol{\beta}_k^{\mathrm{T}}\right)}, \tag{7}$$

$$P\left(c_k|\boldsymbol{x},\boldsymbol{\beta}\right)=\frac{\exp\left(\hat{\boldsymbol{x}}\boldsymbol{\beta}_k^{\mathrm{T}}\right)}{1+\sum_{k=1}^{q-1}\exp\left(\hat{\boldsymbol{x}}\boldsymbol{\beta}_k^{\mathrm{T}}\right)},\quad k=1,2,\ldots,q-1, \tag{8}$$

其中 $\boldsymbol{\beta}=\{\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_k,\ldots,\boldsymbol{\beta}_{q-1}\}$. LR 通过最小化负对数似然来优化参数 $\boldsymbol{\beta}$, 本文采用的目标函数为

$$\boldsymbol{\beta}^*=\arg\min_{\boldsymbol{\beta}}\ell(\boldsymbol{\beta})+r(\boldsymbol{\beta}), \tag{9}$$

其中 $\ell(\boldsymbol{\beta})$ 表示负对数似然项, $r(\boldsymbol{\beta})$ 表示正则化项 [18], 分别依据式 (10) 和 (11) 计算:

$$\ell(\boldsymbol{\beta})=-\sum_{k=1}^{q}\sum_{i=1}^{n_k}\ln P(c_k|\boldsymbol{x}_i^{c_k}), \tag{10}$$

$$r(\boldsymbol{\beta})=\sum_{k=1}^{q-1}\frac{1}{2}\boldsymbol{\beta}_k\boldsymbol{\beta}_k^{\mathrm{T}}, \tag{11}$$

其中 n_k 表示训练集中类标记为 c_k 的实例个数, $\boldsymbol{x}_i^{c_k}$ 表示第 i 个类标记为 c_k 的训练实例.

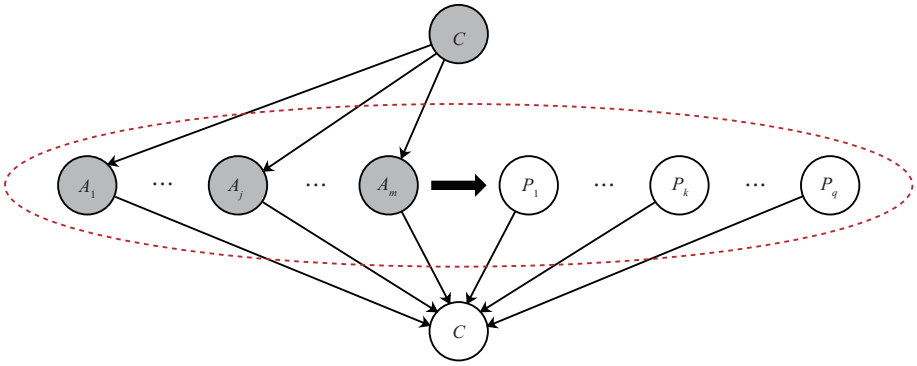

图 2 (网络版彩图) NB-LR 的贝叶斯网络结构

Figure 2 (Color online) Bayesian network structure of NB-LR

NB 要求特征条件独立假设成立, 而 LR 在训练和测试过程中无需对特征变量之间的结构关系做任何假设. Rijmen [19] 认为 LR 也可建模为贝叶斯网络, 其贝叶斯网络结构如图 1(c) 所示, 图中特征变量周围的虚线表示 LR 无需假设特征变量的之间的结构关系.

3.3 基于特征增广构建生成 – 判别混合模型

本小节以 NB-LR 混合模型为例介绍基于特征增广构建混合模型的详细过程. NB-LR 混合模型的贝叶斯网络结构如图 2 所示, 图中 $A_1,\ldots,A_j,\ldots,A_m$ 表示 m 个原始特征变量, $P_1,\ldots,P_k,\ldots,P_q$ 表示使用条件概率分布增广的新特征变量, 新特征变量的数目等于类标记数目 q. 基于特征增广的构建方法先在包含 m 个原始特征变量的原始训练集上学习一个 NB, 然后回代原始训练集中的每个训练实例, 通过 NB 计算每个训练实例的条件概率分布, 并用所学条件概率分布对该训练实例做特征增广, 最后在由增广后的训练实例组成的新训练集上学习 LR 并由 LR 给出最终分类结果.

给定原始训练集 D, NB-LR 混合模型的训练过程描述为算法 1. 首先依据式 (4) 和 (5) 在 D 上训练一个 NB, 然后将 D 中的每一个训练实例 $\boldsymbol{x}$ 分别回代到训练好的 NB 中计算其条件概率分布, 并将此分布增广到 $\boldsymbol{x}$ 中, 增广后的训练实例 $\boldsymbol{x}_a$ 可以由特征值向量 $\langle a_1,\ldots,a_j,\ldots,a_m,p_1,\ldots,p_k,\ldots,p_q\rangle$ 表示, 其中 a_j 是 $\boldsymbol{x}_a$ 在原始特征 A_j 的特征值, 仍等于 $\boldsymbol{x}$ 在原始特征 A_j 上的特征值; p_k 是 $\boldsymbol{x}_a$ 在增广的新特征 P_k 上的特征值, 等于 NB 将实例 $\boldsymbol{x}$ 预测为类标记 c_k 的条件概率 $P(c_k|\boldsymbol{x})$, 可由式 (3) 计算得到. 增广后的训练实例构成了新的训练集 D_a, 最后依据式 (9)∼(11) 在 D_a 上学习一个 LR.

给定测试实例 $\boldsymbol{x}$, NB-LR 混合模型的测试过程描述为算法 2. 与训练过程类似, 首先需要将测试实例 $\boldsymbol{x}$ 输入到 NB 中, 依据式 (3) 估计 $\boldsymbol{x}$ 的条件概率分布, 将所得分布增广到 $\boldsymbol{x}$ 得到增广后的实例 $\boldsymbol{x}_a$. 然后遍历类标记 c_k, 依据式 (7) 和 (8) 求条件概率分布 $\{P(c_1|\boldsymbol{x}_a,\boldsymbol{\beta}),\ldots,P(c_k|\boldsymbol{x}_a,\boldsymbol{\beta}),\ldots,P(c_q|\boldsymbol{x}_a,\boldsymbol{\beta})\}$. 最后找到分布中最大的概率值, 返回与之对应的类标记作为预测结果.

根据算法 1, 新方法在训练阶段进行特征增广时, 需要将 n 个训练实例回代到 NB 中计算条件概率分布, 仍用 q 表示类标记个数, m 表示原始特征个数, 则 NB 的分类时间复杂度为 $O(qm)$, 因此特征增广在训练阶段的时间复杂度为 $O(nqm)$; 由于特征增广为每个训练实例增加了 q 个新特征, 因此特征增广在训练阶段的空间复杂度为 $O(nq)$. 根据算法 2, 特征增广在分类阶段需要将测试实例输入 NB 得到其条件概率分布, 因此特征增广在分类阶段的时间复杂度为 $O(qm)$, 空间复杂度为 $O(q)$. 综上可知, 特征增广的时间复杂度与原始特征个数 m 满足线性关系, 且由于特征增广在训练阶段的时间复杂度受到训练实例个数 n 的影响, 因此本文所提构建方法的效率在小数据集上更有优势.

Algorithm 1 基于特征增广构建 NB-LR 混合模型 (训练过程)

Input: Training set D.
Output: Generative model NB, discriminative model LR.
for k from 1 to q **do**
  Estimate $P(c_k)$ and $P(a_j|c_k)$ of generation model NB by Eqs. (4) and (5) based on D;
end for
Create a new training set $D_a = \emptyset$;
for each training instance $\boldsymbol{x}$ from D **do**
  Create a new training instance $\boldsymbol{x}_a = \boldsymbol{x}$;
  for k from 1 to q **do**
    Calculate p_k of generation model NB by Eq. (3);
    Augment p_k into $\boldsymbol{x}_a$;
  end for
  Put $\boldsymbol{x}_a$ into D_a;
end for
Learn parameter $\boldsymbol{\beta}$ of discriminative model LR until convergence by Eqs. (9)∼(11) based on D_a;
return Generative model NB, discriminative model LR.

Algorithm 2 基于特征增广构建 NB-LR 混合模型 (测试过程)

Input: Test instance $\boldsymbol{x}$, generative model NB, discriminative model LR.
Output: The predicted class label of $\boldsymbol{x}$.
Create a new test instance $\boldsymbol{x}_a = \boldsymbol{x}$;
for k from 1 to q **do**
  Calculate p_k of generation model NB by Eq. (3);
  Augment p_k into $\boldsymbol{x}_a$;
end for
Create a new conditional probability distribution $P = \emptyset$;
for k from 1 to q **do**
  if $c_k = c_q$ **then**
    Calculate $P(c_k|\boldsymbol{x}_a, \boldsymbol{\beta})$ of discriminative model LR by Eq. (7);
  else
    Calculate $P(c_k|\boldsymbol{x}_a, \boldsymbol{\beta})$ of discriminative model LR by Eq. (8);
  end if
  Put $P(c_k|\boldsymbol{x}_a, \boldsymbol{\beta})$ into P;
end for
return The class label corresponding to the maximum probability in P.

4 实验与结果

本节设置了两组比较实验来验证本文所提混合模型构建方法的有效性和通用性. 在第 4.1 小节中, 首先介绍了详细的实验设置和实验数据; 在第 4.2 小节中, 围绕用本文所提方法构建的 NB-LR 混合模型展开实验, 验证了本文所提构建方法的有效性; 在第 4.3 小节中, 将 NB-LR 混合模型中的 NB 替换为 HNB, 构建了 HNB-LR 混合模型, 探讨了本文所提构建方法的通用性.

4.1 实验设置和实验数据

实验使用到的模型包括: NB, HNB, LR, SVM, KNN, C4.5, HBayes-NB, 以及使用本文所提构建方法所得的混合模型 NB-LR, NB-NB, LR-NB, LR-LR 和 HNB-LR. 其中, NB, HNB, LR 代表了组成混合

表 1 实验使用数据集的标号和名称

Table 1 The indexs and names of datasets used in the experiments

ID	Dataset	ID	Dataset	ID	Dataset	ID	Dataset
1	anneal.ORIG	10	credit-a	19	ionosphere	28	sick
2	anneal	11	credit-g	20	iris	29	sonar
3	audiology	12	diabetes	21	kr-vs-kp	30	soybean
4	autos	13	glass	22	labor	31	splice
5	balance-scale	14	heart-c	23	letter	32	vehicle
6	breast-cancer	15	heart-h	24	lymph	33	vote
7	breast-w	16	heart-statlog	25	mushroom	34	vowel
8	colic.ORIG	17	hepatitis	26	primary-tumor	35	waveform-5000
9	colic	18	hypothyroid	27	segment	36	zoo

模型的单个模型; SVM, KNN, C4.5 代表了其他较为典型的判别模型; HBayes-NB 是相关工作中介绍的一类经典的生成 – 判别混合模型; NB-LR 和 HNB-LR 是基于特征增广构建的生成 – 判别混合模型; NB-NB, LR-NB, LR-LR 是基于特征增广构建的衍生混合模型, 分别代表了生成 – 生成混合模型、判别 – 生成混合模型, 以及判别 – 判别混合模型. 上述所有模型的实现均在国际机器学习与数据挖掘实验平台 (Waikato environment for knowledge analysis, WEKA) [20] 上完成. 除了基于特征增广的构建方法, 其余模型均可直接使用 WEKA 平台官方集成的代码, 其中 LR 和 SVM 需额外添加 LIBLINEAR [18] 库. 实验中模型的具体参数设置如下: LR 和 SVM 统一采用 L2 正则化, 统一对增广的数值特征进行归一化处理, KNN 中的近邻数依据经验设置为 3, 其余参数均取 WEKA 平台规定的默认值.

本文的实验使用了 WEKA 平台推介使用的 36 个经典 UCI [21] 标准数据集, 这些数据集代表了广泛的应用领域和数据特征, 表 1 列出了这些数据集在本文中的使用标号和对应名称. 由于实验选择的生成模型不适用于处理具有缺失值和连续特征的数据集, 因此本文进行了如下预处理工作: 首先, 使用离散特征的众数 (mode) 替换了离散特征的缺失值, 使用连续特征的均值 (mean) 替换了连续特征的缺失值; 然后, 采用基于信息熵的最小描述长度法 (minimum description length, MDL) 将连续特征离散化为离散特征.

4.2 基于 NB-LR 混合模型的实验

为了验证本文所提方法的有效性, 本文针对 NB-LR 混合模型设计了一组比较实验, 其中混合模型 NB-LR 的比较对象包括 NB, LR 和 NB-NB, LR-NB, LR-LR, HBayes-NB 混合模型. NB 和 LR 是组成 NB-LR 混合模型的单个模型, NB-NB, LR-NB, LR-LR 是采用本文所提方法构建的衍生混合模型, HBayes-NB 是一类现有经典的混合模型. 依据前文分析, 新方法中先依靠生成模型为判别模型做特征增广, 弥补判别模型忽略联合概率分布对分类的影响的不足; 再利用判别模型准确率高的优点完成分类任务, 使得混合模型能获得更好的分类性能. 三类衍生混合模型与生成 – 判别混合模型相比, 虽然也采用了基于特征增广的构建方法, 然而均破坏了本文构想的生成模型和判别模型在混合模型中的作用: NB-NB 混合模型未采用判别模型给出最终分类结果; LR-NB 混合模型采用判别模型做特征增广, 采用生成模型给出最终分类结果; LR-LR 混合模型未采用生成模型做特征增广. 实验中选择单个模型和衍生混合模型作为 NB-LR 的比较对象, 试图验证本文所提新方法的有效性以及新方法在结构上的合理性; 选择现有混合模型作为 NB-LR 的比较对象, 试图验证本文所提新方法在分类性能和时间复杂度上的优势.

本文采用分类精度和训练时间评估模型的性能, 表 2 和 3 分别给出了分类精度和训练时间的详细比较结果, 表中所有的观测值都是通过十次十折交叉验证获得的平均值. 根据显著性水平为 $\alpha = 0.05$ 的成对双尾 t 检验 [22], 表 2 和 3 中 $\bullet$ 表示 NB-LR 在相应数据集上的评估指标显著优于其比较模型, $\circ$

表 2 NB-LR 在分类精度上的详细实验结果 (%)

Table 2 The detailed experimental results of NB-LR on accuracy (%)

Dataset	NB-LR	NB	LR	NB-NB	LR-NB	LR-LR	HBayes-NB
anneal.ORIG	93.47±2.73	92.66±2.72	82.23±2.15 •	91.74±2.70 •	84.99±3.90 •	84.61±2.84 •	92.74±2.66
anneal	97.88±1.50	96.13±2.16 •	96.64±1.70 •	31.89±4.20 •	94.90±2.05 •	97.14±1.44	97.55±1.91
audiology	77.58±6.31	71.40±6.37 •	77.14±6.65	75.12±7.70	75.36±7.19	77.24±6.77	72.50±6.55 •
autos	77.37±9.42	72.30±10.31 •	65.74±9.89 •	75.67±9.77	68.84±10.22 •	66.32±9.67 •	75.51±9.75
balance-scale	69.93±3.99	71.08±4.29	69.98±3.93	66.78±5.32 •	66.74±5.30 •	69.88±3.89	69.87±3.77
breast-cancer	71.91±7.08	72.94±7.71	71.82±7.03	72.74±7.95	71.48±7.39	71.81±7.05	72.97±8.08
breast-w	96.94±1.72	97.25±1.79	97.14±1.84	97.05±1.79	97.27±1.89	96.97±1.83	97.15±1.74
colic.ORIG	75.63±6.51	73.62±6.83	77.97±5.69	73.86±6.49	77.58±6.16	78.24±5.50	73.26±6.40
colic	83.67±5.21	81.39±5.74	83.83±5.64	81.12±5.75	84.05±5.53	83.74±5.69	81.60±5.97
credit-a	86.55±3.72	86.25±4.01	85.74±4.13	86.16±3.92	86.00±4.04	85.77±4.15	87.07±3.48
credit-g	76.59±3.77	75.43±3.84	76.09±3.41	74.16±4.13 •	75.37±3.63	75.95±3.44	75.07±3.81
diabetes	78.53±4.59	77.85±4.67	78.58±4.40	78.10±4.67	79.05±4.45	78.36±4.39	78.27±4.30
glass	74.03±9.10	74.39±7.95	65.47±8.42 •	70.59±9.31	71.31±8.75	68.02±8.46 •	74.86±8.03
heart-c	82.44±6.99	83.60±6.42	80.49±7.79	82.52±6.87	81.06±7.48	80.49±7.86	83.00±6.57
heart-h	83.89±6.31	84.46±5.92	83.24±6.32	83.86±6.53	83.59±6.45	83.31±6.22	83.51±6.30
heart-statlog	83.22±6.33	83.74±6.25	83.33±6.22	83.85±6.33	83.33±6.33	83.44±6.22	84.37±6.11
hepatitis	87.95±8.41	84.22±9.41	88.54±7.01	84.23±9.17	84.49±9.13	88.40±7.24	84.54±9.59
hypothyroid	99.04±0.50	98.48±0.59 •	95.91±0.74 •	98.42±0.58 •	98.49±0.53 •	98.51±0.69 •	98.50±0.61 •
ionosphere	91.51±3.73	90.77±4.76	85.73±4.58 •	90.77±4.75	93.56±3.73	86.53±4.55 •	91.08±4.67
iris	94.47±5.45	94.47±5.61	91.27±7.56	94.47±5.36	95.27±5.55	93.13±5.72	95.33±5.48
kr-vs-kp	96.52±1.08	87.79±1.91 •	95.92±1.18 •	87.93±1.88 •	95.93±1.23 •	96.92±0.91	94.47±1.62 •
labor	94.00±9.81	93.13±10.56	90.87±11.35	94.03±10.60	94.20±9.00	92.27±10.41	96.13±8.42
letter	79.03±0.90	74.00±0.88 •	69.42±0.90 •	73.18±1.00 •	73.77±0.80 •	75.16±0.83 •	74.80±0.88 •
lymph	85.82±9.07	84.97±8.30	83.42±9.96	83.62±9.07	86.04±8.68	82.94±9.97	85.96±7.50
mushroom	99.52±0.24	95.52±0.78 •	95.99±0.73 •	96.46±0.73 •	96.34±0.68 •	96.70±0.57 •	98.99±0.36 •
primary-tumor	48.02±5.80	47.20±6.02	46.26±6.50	27.94±5.31 •	38.56±7.26 •	45.72±6.13 •	46.78±6.55
segment	94.32±1.45	91.71±1.68 •	92.78±1.54 •	91.89±1.67 •	94.60±1.39	93.49±1.59	93.50±1.73
sick	97.36±0.75	97.10±0.84	97.28±0.78	96.28±0.96 •	96.82±0.93 •	97.26±0.72	97.36±0.78
sonar	86.55±7.22	85.16±7.52	85.40±7.85	85.21±7.64	85.96±7.68	85.40±8.19	84.92±7.58
soybean	93.97±2.47	92.20±3.23 •	94.29±2.13	92.56±2.56 •	92.64±2.95 •	94.10±2.20	92.43±3.03 •
splice	95.96±1.12	95.42±1.14 •	91.40±1.48 •	95.40±1.15 •	93.56±1.32 •	91.79±1.46 •	95.45±1.16 •
vehicle	73.09±3.96	62.52±3.81 •	71.63±3.75	62.59±3.99 •	71.68±3.89	71.84±3.73	65.71±4.47 •
vote	96.02±2.97	90.21±3.95 •	96.04±2.94	90.23±3.93 •	95.38±2.92	96.00±2.97	91.61±4.12 •
vowel	71.51±3.96	65.23±4.53 •	54.89±4.66 •	63.79±4.09 •	62.28±4.58 •	60.80±4.42 •	66.74±4.64 •
waveform-5000	86.63±1.43	80.72±1.50 •	86.95±1.20	83.42±1.32 •	86.49±1.41	86.88±1.19	84.73±1.48 •
zoo	96.05±5.46	93.98±7.14	93.17±6.59	97.05±5.13	96.75±4.84	95.25±5.55	96.05±5.60
$W/T/L$	–	14/22/0	12/24/0	17/19/0	13/23/0	10/26/0	11/25/0
Average	85.47	83.31	82.85	80.96	83.71	83.62	84.29

则相反. 表中任意一个比较模型的 $W/T/L$ 表明: 与比较模型相比, NB-LR 的评估指标在 W 个数据集上显著提升, 在 T 个数据集上基本持平, 在 L 个数据集上显著降低. 作为各模型在 36 个数据集上的整体性能指标, 平均分类精度和平均训练时间被分别汇总在表 2 和 3 的底部. 最后, 基于表 2 和 3, 本文利用 KEEL (knowledge extraction based on evolutionary learning) 数据挖掘软件做了系统的威尔克森 (Wilcoxon) 符号秩检验 [23], 来进一步比较每一对模型之间的整体性能差异, 详细的比较结果如表 4 和 5 所示, 表中 • 表示相应列中的模型显著优于相应行中的模型, 而 ∘ 表示相应行中的模型显著优于相应列中的模型. 主对角线下方的显著性水平为 $\alpha = 0.05$; 主对角线上方的显著性水平为 $\alpha = 0.1$. 从上述实验的比较结果来看, NB-LR 的整体性能显著优于其比较对象, 这充分验证了本文所提构建方法的有效性. 这些发现总结如下:

(1) 在 36 个数据集上, 依据本文所提方法构建的 NB-LR 混合模型的平均分类精度为 85.47%, 显著高于其所有比较对象: NB (83.31%), LR (82.85%), NB-NB (80.96%), LR-NB (83.71%), LR-LR (83.62%), HBayes-NB (84.29%). 混合模型 NB-LR 取得了最高精度, 表明依据本文所提方法构建的混合模型的

表 3 NB-LR 在训练时间上的详细实验结果 (s)

Table 3 The detailed experimental results of NB-LR on training time (s)

Dataset	NB-LR	NB	LR	NB-NB	LR-NB	LR-LR	HBayes-NB
anneal.ORIG	0.09±0.03	0.00±0.00 ○	0.07±0.01	0.02±0.00 ○	0.07±0.01	0.15±0.02 ●	206.53±3.83 ●
anneal	0.09±0.01	0.00±0.00 ○	0.07±0.01 ○	0.02±0.00 ○	0.07±0.01 ○	0.14±0.02 ●	212.68±4.00 ●
audiology	0.07±0.01	0.00±0.00 ○	0.03±0.01 ○	0.03±0.01 ○	0.03±0.00 ○	0.07±0.01	525.09±22.53 ●
autos	0.10±0.01	0.00±0.00 ○	0.11±0.01	0.00±0.00 ○	0.10±0.01	0.23±0.03 ●	11.79±0.28 ●
balance-scale	0.01±0.00	0.00±0.00 ○	0.00±0.00 ○	0.00±0.00 ○	0.00±0.00 ○	0.01±0.00 ●	0.06±0.01 ●
breast-cancer	0.01±0.00	0.00±0.00 ○	0.01±0.00	0.00±0.00 ○	0.01±0.00 ○	0.01±0.01 ●	0.32±0.03 ●
breast-w	0.01±0.00	0.00±0.00 ○	0.00±0.00 ○	0.00±0.00 ○	0.00±0.00 ○	0.01±0.00 ●	0.72±0.10 ●
colic.ORIG	0.04±0.01	0.00±0.00 ○	0.04±0.01	0.00±0.00 ○	0.04±0.01	0.09±0.01 ●	15.79±0.53 ●
colic	0.01±0.01	0.00±0.00 ○	0.01±0.00	0.00±0.00 ○	0.01±0.00	0.01±0.01	8.24±0.46 ●
credit-a	0.03±0.00	0.00±0.00 ○	0.03±0.00	0.00±0.00 ○	0.02±0.01	0.05±0.01 ●	4.05±0.14 ●
credit-g	0.04±0.01	0.00±0.00 ○	0.03±0.01	0.00±0.00 ○	0.03±0.00 ○	0.07±0.01 ●	15.97±0.61 ●
diabetes	0.00±0.00	0.00±0.00 ○	0.00±0.00 ○	0.00±0.00 ○	0.00±0.00 ○	0.01±0.00 ●	0.52±0.05 ●
glass	0.01±0.01	0.00±0.00 ○	0.00±0.00	0.00±0.00 ○	0.01±0.00	0.01±0.00	0.35±0.02 ●
heart-c	0.00±0.00	0.00±0.00 ○	0.00±0.00 ○	0.00±0.00 ○	0.00±0.00 ○	0.00±0.00	1.30±0.06 ●
heart-h	0.00±0.00	0.00±0.00 ○	0.00±0.00 ○	0.00±0.00 ○	0.00±0.00 ○	0.00±0.00 ●	1.24±0.14 ●
heart-statlog	0.00±0.00	0.00±0.00 ○	0.00±0.00	0.00±0.00 ○	0.00±0.00 ○	0.00±0.00 ●	0.93±0.07 ●
hepatitis	0.00±0.00	0.00±0.00 ○	0.00±0.00 ○	0.00±0.00○	0.00±0.00 ○	0.00±0.00	2.04±0.09 ●
hypothyroid	0.16±0.02	0.00±0.00 ○	0.13±0.01 ○	0.04±0.01 ○	0.13±0.01 ○	0.28±0.02 ●	247.14±8.88 ●
ionosphere	0.03±0.01	0.00±0.00 ○	0.03±0.00	0.00±0.00 ○	0.03±0.00	0.05±0.01 ●	41.31±0.81 ●
iris	0.00±0.00	0.00±0.00 ○	0.00±0.00	0.00±0.00 ○	0.00±0.00	0.00±0.00 ●	0.02±0.01 ●
kr-vs-kp	0.07±0.01	0.00±0.00 ○	0.05±0.01 ○	0.02±0.01 ○	0.05±0.01 ○	0.10±0.01 ●	486.27±35.47 ●
labor	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.47±0.09 ●
letter	161.23±25.32	0.00±0.00 ○	78.81±2.71 ○	1.59±0.05 ○	76.34±2.09 ○	199.54±15.72 ●	1165.76±82.97 ●
lymph	0.01±0.00	0.00±0.00 ○	0.01±0.00 ○	0.00±0.00 ○	0.01±0.00	0.01±0.00 ●	2.13±0.28 ●
mushroom	0.95±0.06	0.00±0.00 ○	0.94±0.04	0.04±0.00 ○	0.91±0.04	2.09±0.08 ●	243.59±13.35 ●
primary-tumor	0.05±0.01	0.00±0.00 ○	0.03±0.01 ○	0.01±0.00 ○	0.03±0.01 ○	0.08±0.02 ●	6.09±0.31 ●
segment	2.17±0.11	0.00±0.00 ○	1.82±0.07 ○	0.03±0.01 ○	1.74±0.05 ○	4.14±0.16 ●	100.00±5.38 ●
sick	0.04±0.00	0.00±0.00 ○	0.02±0.01 ○	0.02±0.01 ○	0.03±0.01 ○	0.05±0.01 ●	282.18±18.95 ●
sonar	0.00±0.00	0.00±0.00 ○	0.00±0.00 ○	0.00±0.00 ○	0.00±0.00 ○	0.00±0.00 ○	265.97±10.19 ●
soybean	0.19±0.02	0.00±0.00 ○	0.12±0.01 ○	0.04±0.01 ○	0.11±0.01 ○	0.28±0.04 ●	154.29±5.21 ●
splice	1.53±0.18	0.00±0.00 ○	1.23±0.06 ○	0.05±0.01 ○	1.13±0.05 ○	2.51±0.10 ●	3874.72±463.27 ●
vehicle	0.09±0.01	0.00±0.00 ○	0.08±0.01	0.01±0.00 ○	0.08±0.01 ○	0.18±0.02 ●	18.38±0.76 ●
vote	0.00±0.00	0.00±0.00 ○	0.00±0.00 ○	0.00±0.00 ○	0.00±0.00	0.00±0.00 ●	5.36±0.29 ●
vowel	0.67±0.05	0.00±0.00 ○	0.42±0.04 ○	0.01±0.01 ○	0.40±0.02 ○	1.11±0.05 ●	11.86±1.23 ●
waveform-5000	1.26±0.06	0.00±0.00 ○	1.11±0.04 ○	0.06±0.01 ○	1.08±0.05 ○	2.37±0.08 ●	1340.68±47.79 ●
zoo	0.00±0.00	0.00±0.00 ○	0.00±0.00	0.00±0.00 ○	0.00±0.00 ○	0.00±0.00 ●	1.13±0.09 ●
$W/T/L$	–	0/1/35	0/15/21	0/1/35	0/12/24	29/6/1	36/0/0
Average	4.69	0.00	2.37	0.06	2.29	5.94	257.08

表 4 威尔克森测试的精度比较结果 (NB-LR)

Table 4 Classification accuracy comparison results of the Wilcoxon tests (NB-LR)

Method	NB-LR	NB	LR	NB-NB	LR-NB	LR-LR	HBayes-NB
NB-LR	–	○	○	○	○	○	○
NB	●	–					●
LR	●		–		●	●	●
NB-NB	●			–	●		●
LR-NB	●			○	–		
LR-LR	●		○			–	
HBayes-NB	●	○		○			–

整体性能优于其组成模型 NB 和 LR, 依据这一方法构建的衍生混合模型, 以及现有经典的混合模型 HBayes-NB, 初步验证了本文所提构建方法的有效性.

(2) 依据显著性水平为 $\alpha = 0.05$ 的成对双尾 t 检验, NB-LR 的分类性能显著优于其所有比较对象, 在 36 个数据集上与各比较对象相比的结果为: NB (14 胜 0 负), LR (12 胜 0 负), NB-NB (17 胜 0

表 5　威尔克森测试的训练时间比较结果 (NB-LR)

Table 5　Training time comparison results of the Wilcoxon tests (NB-LR)

Method	NB-LR	NB	LR	NB-NB	LR-NB	LR-LR	HBayes-NB
NB-LR	–	●	●	●	●	○	○
NB	○	–	○	○	○	○	○
LR	○	●	–	●	●	○	○
NB-NB	○	●	○	–	○	○	○
LR-NB	○	●		●	–	○	○
LR-LR	●	●	●	●	●	–	○
HBayes-NB	●	●	●	●	●	●	–

负), LR-NB (13 胜 0 负), LR-LR (10 胜 0 负), HBayes-NB (11 胜 0 负). 本文构建的 NB-LR 的分类精度相较于各比较对象均在 10 个以上的数据集上显著提升, 无显著下降, 这组结果从单个数据集的比较中再次验证了 NB-LR 的有效性.

(3) 依据表 4 所示威尔克森符号秩检验的分类精度比较结果, 本文构建的 NB-LR 在 36 个数据集上的整体性能显著优于所有的比较对象, 包括 NB, LR, NB-NB, LR-NB, LR-LR, HBayes-NB, 再次验证了所提方法的有效性. 3 个衍生模型中, 生成 – 生成混合模型 NB-NB 的分类性能弱于判别 – 生成混合模型 LR-NB, 判别 – 判别混合模型 LR-LR 和判别 – 生成混合模型 LR-NB 的分类性能优于 LR, 但均不如生成 – 判别混合模型 NB-LR, 这验证了基于特征增广方法构建生成 – 判别混合模型在结构上的合理性.

(4) 在 36 个数据集上, 依据本文所提方法构建的 NB-LR 混合模型的平均训练时间为 4.69 s, 不如其组成模型 NB (0 s) 和 LR (2.37 s), 但显著优于依据特征划分构建的混合模型 HBayes-NB (257.08 s). 与 HBayes-NB 相比, NB-LR 在显著性水平为 $\alpha = 0.05$ 的成对双尾 t 检验的结果为 36 胜 0 负, 表 5 所示的威尔克森符号秩检验的训练时间比较结果同样证明 NB-LR 显著优于 HBayes-NB. 这表明和经典的基于特征划分构建的混合模型相比, 本文所提新方法具有更低的时间复杂度.

为了进一步观察本文构建的 NB-LR 混合模型的偏差 – 方差权衡效果, 本文统计了 NB-LR 混合模型相较于 NB 和 LR 的偏差与方差变化. 本文中偏差与方差的估计与文献 [24] 保持一致. 在上述使用的所有实验数据集上, 本文用 NB (LR) 的偏差和方差减去本文构建的 NB-LR 混合模型的偏差和方差, 并将相减的结果汇总于图 3 (图 4). 图 3 和 4 采用堆积柱形图展示偏差与方差的变化, 偏差和方差的变化通过不同底纹的柱形图表示. 图中纵轴度量偏差与方差变化的大小和正负, 其中柱形图位于 0 刻度线以上则表示 NB-LR 的结果较 NB (LR) 有降低, 位于 0 刻度线以下则相反, 柱状图的高度反映了变化的大小; 横轴对应数据集标号, 数据集标号表明了数据集的名称和顺序, 具体与表 1 保持一致.

从图 3 和 4 所示的实验结果可知, 依据本文所提方法构建的 NB-LR 混合模型在分类全部的 36 个数据集时: 相较于 NB 偏差在 31 个数据集上得到了降低, 方差在 12 个数据集上得到了降低; 相较于 LR 的偏差在 26 个数据集上得到了降低; 方差在 23 个数据集上得到了降低. 这表明了依据本文所提方法构建混合模型在完成分类任务中遵循了偏差 – 方差权衡的原则, 即混合模型相较于其组成模型可同时取得更低的偏差和方差, 因此混合模型具有更好的泛化性能.

4.3　基于 HNB-LR 混合模型的实验

在本文所提的构建方法中, 混合模型最终采用判别模型给出分类结果, 因此混合模型的分类性能和其中的判别模型直接相关; 而生成模型的作用在于通过联合概率分布学习条件概率分布, 并以此做特征增广供判别模型学习, 因此生成模型间接作用于混合模型的分类性能. 本组实验用 HNB 替换 NB 构建 HNB-LR 混合模型, 尝试分析生成模型对混合模型的分类性能的影响. HNB 通过结构扩展的方法

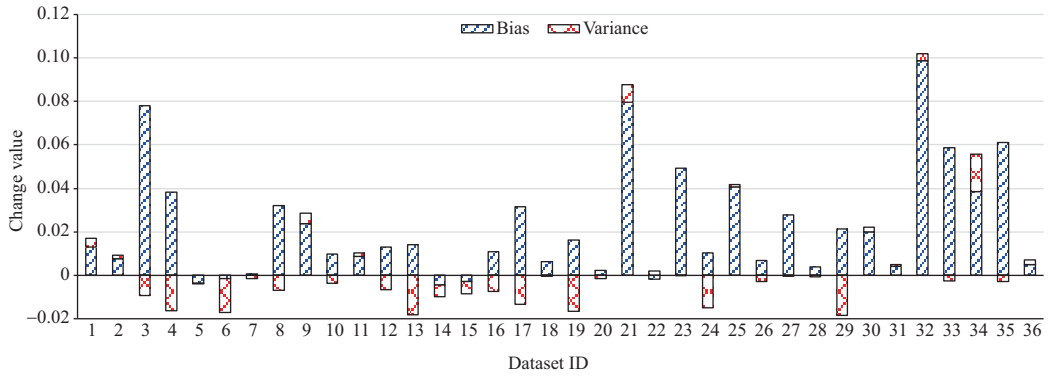

图 3 (网络版彩图) NB-LR 相较于 NB 的偏差 – 方差变化

Figure 3 (Color online) Changes in bias and variance (NB-LR vs. NB)

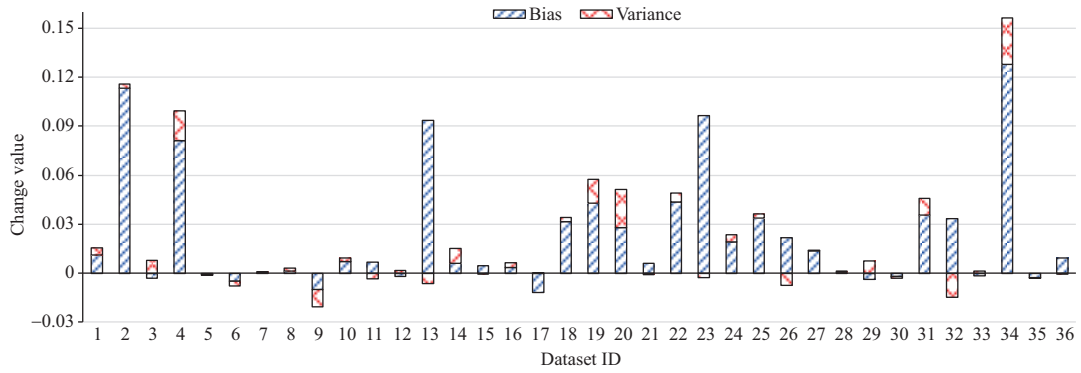

图 4 (网络版彩图) NB-LR 相较于 LR 的偏差 – 方差变化

Figure 4 (Color online) Changes in bias and variance (NB-LR vs. LR)

削弱了 NB 的特征条件独立假设, 因此可以得到更加准确的条件概率分布, 通过比较 HNB-LR 和 NB-LR 的分类表现可以观察混合模型的性能与增广的条件概率分布之间的关系. 由于混合模型最终由判别模型给出分类结果, 因此除了 HNB 和 LR 外, 本文还设置了 3 个判别模型 SVM, KNN, 以及 C4.5 作为 HNB-LR 的比较对象, 进一步测试了 HNB-LR 的分类性能. 本组实验的实验数据、实验设置、实验方法和显著性检验方法均与前文保持一致, 表 6 给出了分类精度的详细比较结果, 表 7 给出了基于分类精度的威尔克森符号秩检验的详细比较结果.

从表 6 和 7 所示的比较结果来看, HNB-LR 的分类精度显著优于其所有比较对象, 再次验证了新方法的有效性, 同时也验证了新方法的通用性, 即新方法同样适用于除了 NB 以外的其他模型. 同时, 实验结果同样表明, 依据新方法构建的混合模型虽然最终由判别模型完成分类任务, 但是选择合适的生成模型做特征增广对混合模型分类性能的提升同样重要. 这些发现总结如下:

(1) 在 36 个数据集上, 依据本文所提方法构建的 HNB-LR 混合模型的平均分类精度为 86.47%, 显著高于其所有比较对象: HNB (85.86%), LR (82.85%), SVM (83.16%), KNN (83.73%), C4.5 (83.95%). 本文构建的 HNB-LR 的分类精度取得了最高值, 这表明依据本文所提方法构建的混合模型不仅结构简单, 并且可以适用于不同的生成模型, 具有较强的通用性. 同时 HNB-LR 的平均分类精度较 NB-LR 再次提升 1% 也表明混合模型虽然由判别模型给出最终分类结果, 但是选择合适的生成模型做特征增

表 6 HNB-LR 在分类精度上的详细实验结果 (%)

Table 6 The detailed experimental results of HNB-LR on accuracy (%)

Dataset	HNB-LR	HNB	LR	SVM	KNN	C4.5
anneal.ORIG	95.22±2.19	95.29±2.04	82.23±2.15 •	84.98±2.33 •	93.11±2.25 •	92.47±2.37 •
anneal	98.66±1.08	98.33±1.22	96.64±1.70 •	97.24±1.45 •	97.84±1.34	98.78±0.91
audiology	76.73±6.21	69.04±5.83 •	77.14±6.65	81.70±6.84 ○	65.80±7.48 •	77.22±7.69
autos	82.04±8.39	82.17±8.60	65.74±9.89 •	64.65±10.31 •	78.54±7.98	76.39±9.55
balance-scale	69.60±3.95	69.05±3.75	69.98±3.93	70.14±3.97	68.94±3.74	69.32±3.89
breast-cancer	70.67±6.94	73.09±6.11	71.82±7.03	71.22±7.06	73.33±5.50	75.26±5.04 ○
breast-w	96.42±1.91	96.32±2.01	97.14±1.84	96.87±1.90	96.90±1.93	94.65±2.51 •
colic.ORIG	75.01±5.77	74.06±5.79	77.97±5.69	77.29±6.05	71.63±5.95	78.57±5.41
colic	82.39±5.72	82.09±5.86	83.83±5.64	83.63±5.49	80.22±6.61	84.72±5.95
credit-a	85.96±3.96	85.91±3.70	85.74±4.13	85.61±4.19	85.07±3.67	86.58±3.53
credit-g	75.89±3.82	76.12±3.72	76.09±3.41	75.84±3.34	71.30±3.45 •	72.17±3.49 •
diabetes	78.25±4.19	76.81±4.11	78.58±4.40	78.53±4.53	77.39±4.41	77.34±4.91
glass	79.57±8.82	77.80±8.40	65.47±8.42 •	68.49±8.64 •	76.91±8.16	75.23±9.46 •
heart-c	81.69±7.32	82.31±6.81	80.49±7.79	80.55±7.81	81.25±7.12	77.32±6.20 •
heart-h	83.51±5.87	84.87±6.03	83.24±6.32	84.06±5.99	83.62±6.27	80.96±6.90
heart-statlog	82.67±6.49	82.33±6.55	83.33±6.22	83.52±6.43	80.89±6.94	82.26±7.32
hepatitis	87.89±7.62	88.26±7.28	88.54±7.01	86.86±7.97	87.17±8.14	81.33±9.48 •
hypothyroid	99.17±0.46	98.95±0.48	95.91±0.74 •	98.67±0.51 •	97.10±0.76 •	99.28±0.42
ionosphere	92.08±4.02	91.82±4.33	85.73±4.58 •	88.98±4.26	90.71±4.73	89.49±5.12
iris	93.53±5.88	93.80±5.86	91.27±7.56	93.07±5.83	94.67±5.69	93.87±4.89
kr-vs-kp	97.21±0.80	92.36±1.30 •	95.92±1.18 •	96.85±0.99	96.56±1.00	99.44±0.37 ○
labor	95.97±8.51	94.87±9.82	90.87±11.35	91.63±10.81	91.70±11.13	87.13±15.32
letter	89.45±0.68	88.20±0.66 •	69.42±0.90 •	63.65±2.70 •	90.44±0.56 ○	78.75±0.78 •
lymph	86.01±8.46	85.84±8.86	83.42±9.96	83.05±10.27	83.47±8.41	76.51±10.14 •
mushroom	99.96±0.06	99.94±0.10	95.99±0.73 •	96.31±0.67 •	100.00±0.00	100.00±0.00
primary-tumor	47.61±5.70	47.66±6.21	46.26±6.50	46.16±6.53	39.44±5.37 •	41.01±6.59 •
segment	96.83±1.15	95.88±1.19 •	92.78±1.54 •	91.71±1.87 •	93.86±1.50 •	95.23±1.37 •
sick	97.48±0.73	97.56±0.74	97.28±0.78	97.11±0.90	97.10±0.81	97.82±0.75
sonar	86.11±7.44	84.63±7.34	85.40±7.85	85.93±7.67	81.65±8.56	80.60±8.85
soybean	94.03±2.36	93.88±2.47	94.29±2.13	94.74±2.36	91.54±2.90 •	92.63±2.72
splice	96.33±1.11	95.84±1.10 •	91.40±1.48 •	91.43±1.51 •	77.59±1.84 •	94.17±1.28 •
vehicle	73.26±3.80	72.37±3.35	71.63±3.75	71.10±3.85	72.02±3.95	70.77±3.86
vote	95.93±2.94	94.43±3.18 •	96.04±2.94	96.11±3.13	93.58±3.48 •	96.27±2.79
vowel	87.23±3.12	85.12±3.65 •	54.89±4.66 •	54.04±4.41 •	79.23±4.12 •	79.54±4.01 •
waveform-5000	86.59±1.44	86.21±1.44	86.95±1.20	86.89±1.28	78.65±1.78 •	76.36±1.77 •
zoo	96.15±5.44	97.73±4.64	93.17±6.59	95.35±6.07	95.15±6.55	92.61±7.33
W/T/L	–	7/29/0	12/24/0	10/25/1	11/24/1	13/21/2
Average	86.47	85.86	82.85	83.16	83.73	83.95

表 7 威尔克森测试的精度比较结果 (HNB-LR)

Table 7 Classification accuracy comparison results of the Wilcoxon tests (HNB-LR)

Method	HNB-LR	HNB	LR	SVM	KNN	C4.5
HNB-LR	–	○	○	○	○	○
HNB	•	–	○	○	○	○
LR	•	•	–			
SVM	•	•		–		
KNN	•	•			–	
C4.5	•	•				–

广对混合模型分类性能的提升同样重要.

(2) 依据显著性水平为 $\alpha = 0.05$ 的成对双尾 t 检验, HNB-LR 显著优于其所有比较对象, 在 36 个数据集上与各比较对象相比的结果为: HNB (7 胜 0 负), LR (12 胜 0 负), SVM (10 胜 1 负), KNN (11 胜 1 负), C4.5 (13 胜 2 负). 本文构建的 HNB-LR 显著优于 HNB, 表明混合模型仍可有效提升生成模型的分类性能, 证实了混合模型的通用性; LR 无法显著优于其他经典的判别模型, 而 HNB-LR 显著优

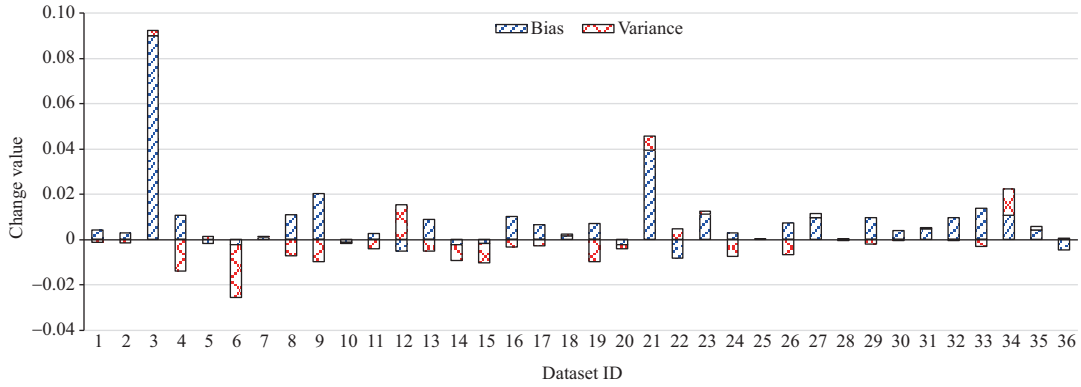

图 5 (网络版彩图) HNB-LR 相较于 HNB 的偏差 – 方差变化

Figure 5 (Color online) Changes in bias and variance (HNB-LR vs. HNB)

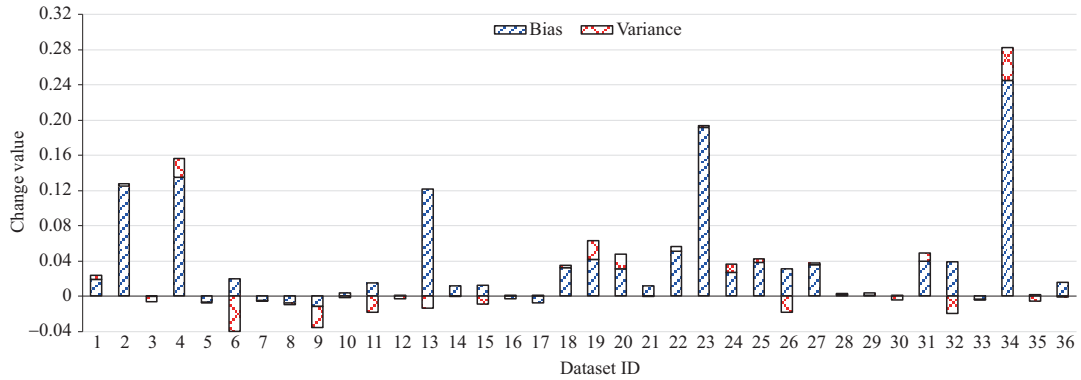

图 6 (网络版彩图) HNB-LR 相较于 LR 的偏差 – 方差变化

Figure 6 (Color online) Changes in bias and variance (HNB-LR vs. LR)

于其他经典的判别模型, 这表明选择合适的生成模型做特征增广对判别模型的性能提升是显著的.

(3) 依据表 7 所示的威尔克森符号秩检验的比较结果来看, 本文构建的 HNB-LR 显著优于所有的比较对象, 包括 HNB, LR, SVM, KNN, C4.5. 值得注意的是, HNB-LR 并不直接使用 HNB 的分类结果, 而是由 LR 给出分类结果, 但其分类性能仍然显著优于 HNB, 这再次表明使用本文所提方法构建混合模型时采用生成模型学习的条件概率分布做特征增广是正确的、有效的.

同样, 为了进一步观察 HNB-LR 混合模型的偏差 – 方差权衡效果, 本文统计了 HNB-LR 混合模型相较于 HNB 和 LR 的偏差与方差变化, 统计方法与前文一致, 并将结果汇总于图 5 和 6. 图 5 和 6 所示的实验结果与 NB-LR 混合模型的实验结果基本一致, HNB-LR 混合模型相较于其组成模型同时取得了更低的偏差和方差, 遵循偏差 – 方差权衡的原则, 具有更好的泛化性能.

5 总结与展望

本文以构建生成 – 判别混合模型为切入点, 概述了生成模型和判别模型的比较研究以及现有的生成 – 判别混合模型构建方法, 在此基础上提出了一种基于特征增广的生成 – 判别混合模型构建方法, 并基于此方法构建了 NB-LR 和 HNB-LR 混合模型, 通过多组实验和显著性检验对新方法的有效性与

通用性进行验证. 实验结果表明依据本文所提新方法构建的混合模型显著提升了单个模型的分类性能, 混合模型为判别模型引入了生成模型学习的条件概率分布, 有效提升了判别模型的分类性能, 同时混合模型也遵循偏差 – 方差权衡的原则, 具有更好的泛化性能.

不过, 目前本文采用的数据集都是较低维的数据集, 当数据集的特征维度很高, 如文本分类数据集时, 增广的条件概率分布的影响力不再明显, 因此混合模型的性能提升效果也就不再明显. 如何将本文所提出的方法适用于高维分类问题是我们接下来要关注的主要问题. 此外, 本文所提的混合模型最终采用判别模型给出分类结果, 因此仍然具备判别模型的优势, 如何将已有针对判别模型的改进工作引入混合模型的构建, 进一步提升生成 – 判别混合模型的分类性能, 同样是我们关注的问题.

参考文献

1 Zhou Z H. Machine Learning. Beijing: Tsinghua University Press, 2016 [周志华. 机器学习. 北京: 清华大学出版社, 2016]
2 Li H. Statistical Learning Method. 2nd ed. Beijing: Tsinghua University Press, 2019 [李航. 统计学习方法. 第二版. 北京: 清华大学出版社, 2019]
3 Rubinstein Y D, Hastie T. Discriminative vs. informative learning. In: Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining, 1997. 49–53
4 Kang C, Tian J. A hybrid generative/discriminative Bayesian classifier. In: Proceedings of the 19th International Florida Artificial Intelligence Research Society Conference, 2006. 562–567
5 Xue J H, Titterington D M. Joint discriminative-generative modelling based on statistical tests for classification. Pattern Recogn Lett, 2010, 31: 1048–1055
6 Tan Y, Shenoy P P, Chan M W, et al. On construction of hybrid logistic regression-naive Bayes model for classification. In: Proceedings of the 8th International Conference on Probabilistic Graphical Models, Lugano, 2016. 523–534
7 Tan Y, Shenoy P P. A bias-variance based heuristic for constructing a hybrid logistic regression-naïve Bayes model for classification. Int J Approx Reason, 2020, 117: 15–28
8 Efron B. The efficiency of logistic regression compared to normal discriminant analysis. J Am Stat Assoc, 1975, 70: 892–898
9 Ng A Y, Jordan M I. On discriminative vs. generative classifiers: a comparison of logistic regression and naive Bayes. In: Proceedings of Advances in Neural Information Processing Systems, Vancouver, 2001. 841–848
10 Raina R, Shen Y, Ng A Y, et al. Classification with hybrid generative/discriminative models. In: Proceedings of Advances in Neural Information Processing Systems, Vancouver, 2003. 545–552
11 Fujino A, Ueda N, Saito K. A hybrid generative/discriminative approach to text classification with additional information. Inf Process Manage, 2007, 43: 379–392
12 Bishop C M, Lasserre J. Generative or discriminative? Getting the best of both worlds. Bayes Stat, 2007, 8: 3–24
13 Friedman N, Geiger D, Goldszmidt M. Bayesian network classifiers. Mach Learn, 1997, 29: 131–163
14 Jiang L X, Zhang H, Cai Z H. A novel Bayes model: hidden naive Bayes. IEEE Trans Knowl Data Eng, 2009, 21: 1361–1371
15 Jiang L X, Zhang L G, Li C Q, et al. A correlation-based feature weighting filter for naive Bayes. IEEE Trans Knowl Data Eng, 2019, 31: 201–213
16 Zaidi N A, Cerquides J, Carman M J, et al. Alleviating naive Bayes attribute independence assumption by attribute weighting. J Mach Learn Res, 2013, 14: 1947–1988
17 Theodoridis S, Koutroumbas K. Pattern Recognition. 4th ed. Orlando: Academic Press, 2008
18 Fan R E, Chang K W, Hsieh C J, et al. LIBLINEAR: a library for large linear classification. J Mach Learn Res, 2008, 9: 1871–1874
19 Rijmen F. Bayesian networks with a logistic regression model for the conditional probabilities. Int J Approx Reason, 2008, 48: 659–666
20 Witten I H, Frank E, Hall M A. Data Mining: Practical Machine Learning Tools and Techniques. 3rd ed. San Francisco: Morgan Kaufmann, 2011

21 Dua D, Graff C. UCI Machine Learning Repository. Irvine: University of California, 2017
22 Nadeau C, Bengio Y. Inference for the generalization error. Mach Learn, 2003, 52: 239–281
23 Singh P K, Sarkar R, Nasipuri M. Significance of non-parametric statistical tests for comparison of classifiers over multiple datasets. Int J Comput Sci Math, 2016, 7: 410–442
24 Webb G I. MultiBoosting: a technique for combining boosting and wagging. Mach Learn, 2000, 40: 159–196

A feature augmentation-based method for constructing generative-discriminative hybrid models

Wenjun ZHANG[1], Liangxiao JIANG[1,2*] & Huan ZHANG[1]

1. *School of Computer Science, China University of Geosciences, Wuhan* 430074*, China*;
2. *Hubei Key Laboratory of Intelligent Geo-Information Processing, Wuhan* 430074*, China*
* Corresponding author. E-mail: ljiang@cug.edu.cn

Abstract From the perspective of probability framework, the generative model first learns the joint probability distribution from the data and then calculates the conditional probability distribution with a faster convergence speed. However, the discriminative model learns the conditional probability distribution directly from the data, thus often demonstrating higher accuracy. As an effective combination of the generative and discriminative models, the generative-discriminative hybrid model integrates their advantages. However, the existing methods must divide the original features into two independent feature spaces to train the two models. Feature division not only increases the time complexity of the model but also weakens the expression ability of the original feature space. To solve this problem, this paper proposes a feature augmentation-based method for constructing the generative-discriminative hybrid model. First, this novel method uses the generative model to learn the conditional probability distribution. Then, it augments the learned conditional probability distribution as new features into the original feature space. Finally, it trains the discriminative model in the augmented feature space and predicts the final classification result. The new method offers several advantages, including using feature augmentation to mix the models, not requiring feature division, exhibiting low time complexity, and enhancing the expression ability of the original feature space. The experimental results on 36 classical UCI benchmark datasets show that the new method is not only effective and universal but also follows bias-variance trade-off.

Keywords generative model, discriminative model, feature augmentation, conditional probability distribution, bias-variance trade-off