计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2020,29(3):155−160 [doi: 10.15888/j.cnki.csa.007161]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

利用协变量调整控制混杂因子的鲁棒文本分类①

董园园

(齐鲁师范学院, 济南 250013)

摘　要: 针对目前很多文本分类方法很少控制混杂变量, 且分类准确度对数据分布的鲁棒性较低的问题, 提出一种基于协变量调整的文本分类方法. 首先, 假设文本分类中的混杂因子 (变量) 可在训练阶段观察到, 但无法在测试阶段观察到; 然后, 以训练阶段的混杂因子为条件, 在预测阶段计算出混杂因子的总和; 最后, 基于 Pearl 的协变量调整, 通过控制混杂因子来观察文本特征和分类变量对分类器的精度影响. 通过微博数据集和 IMDB 数据集验证所提方法的性能, 实验结果表明, 与其他方法相比, 所提方法处理混杂关系时, 可以得到更高的分类准确度, 且对混杂变量具备鲁棒性.

关键词: 协变量调整; 混杂变量; 文本分类; 文本特征; 鲁棒性

引用格式: 董园园.利用协变量调整控制混杂因子的鲁棒文本分类.计算机系统应用,2020,29(3):155–160. http://www.c-s-a.org.cn/1003-3254/7161.html

Robust Text Categorization Using Covariates to Control Confounding Factors

DONG Yuan-Yuan

(Qilu Normal University, Jinan 250013, China)

Abstract: Aiming at the problem that many documents categorization methods seldom control hybrid variables and have low robustness to data distribution, a documents (text) categorization method based on covariate adjustment is proposed. Firstly, it is assumed that the confounding factors (variables) in text categorization can be observed in the training stage, but not in the testing stage. Then, the sum of confounding factors is calculated in the prediction stage under the condition of the confounding factors in the training stage. Finally, based on Pearl's covariate adjustment, the accuracy of text features and classification variables to the classifier is observed by controlling the confounding factors. The performance of the proposed method is verified by microblog data set and IMDB data set. The experimental results show that the proposed method can achieve higher classification accuracy and robustness against mixed variables than other methods.

Key words: covariate adjustment; confounding variables; text classification; text features; robustness

1　引言

文本分类[1]方法的研究已经超过了 50 年, 该类方法大多被应用于专题文献分类. 然而, 随着科技的发展和创新, 跨学科领域如计算社会科学[2]、公共卫生监测[3]和流行病学[4]等都对文本分类提出了新要求. 这些领域的待分类对象通常是在线文本[5], 预测标签则可能是健康状况、政治立场或人类表情等差异化的术语. 这些变化对文本分类 (或称为文档分类) 提出了新要求和新挑战.

目前, 已有很多研究者对文本分类进行了探讨和研究, 如文献[6]为了提取更多的可信反例和构造准确高效的分类器, 提出一种基于聚类的半监督主动分类

① 基金项目: 山东省社会科学规划研究项目 (17CTYJ03)
Foundation item: Social Science Planning Research Project of Shandong Province (17CTYJ03)
收稿时间: 2019-04-26; 修改时间: 2019-05-21; 采用时间: 2019-05-23; csa 在线出版时间: 2020-02-28

方法, 该方法利用聚类技术和正例文档共享尽可能少的特征, 从未标识数据集中尽可能多地移除正例, 但该方法仅适用于较少文本特征的情况. 文献[7]中提出一种基于聚类的改进 KNN 算法, 采用改进统计量方法进行文本特征提取, 依据聚类方法将文本聚类为几个簇, 最后利用改进的 KNN 方法对簇类进行分类, 但该方法难以提高文本分类效率. 还有一些学者开发出控制混杂因子[8]的方法, 包括匹配[9]、分层和回归分析[10]. 文献[11]开发出了用于因果图模型的测试方法, 用于确定哪种结构允许使用后门调整对混杂因子进行控制. 文献[12]提出了一种基于 LDA 模型的文本分类方法, 应用 LDA 概率增长模型对文本集进行主题建模, 在文本集的隐含主题-文本矩阵上训练 SVM, 构造文本分类器, 具有较好的分类效果.

上述方法各有特点, 但其主要缺点是: 混杂变量不能得到很好地控制, 从而造成分类器的错误输出. 本文的目的是对影响分类的因子进行控制和调整, 使得文本分类器具有良好的准确性和鲁棒性. 因此, 本文基于 Pearl 的后门调整方法[11], 提出了一种基于协变量调整的文本分类方法, 该方法以训练阶段的混杂变量为条件, 在预测阶段计算出混杂变量的总和. 另外, 本文还进一步探讨了该方法的参数影响, 以允许对预期调整的强度进行调整. 实验结果表明, 所提方法能够提高分类器的鲁棒性, 即使在混杂因子与目标变量之间的关联从训练集到测试集发生倒置的极端情况下, 也能保持较高的准确度.

2 用于文本分类器的协变量调整

2.1 文本分类中得分协变量调整

假设研究目的是估计变量 X 对变量 Y 的因果效应, 但无法进行随机对照实验. 则已知混杂因子变量 Z 的一个充分集, 可以使用式 (1) 估计对因果关系:

$$p(y|do(x)) = \sum_{z\in Z} p(y|x,z)p(z) \tag{1}$$

该公式称作协变量调整 (也称为后门调整). 协变量标准是一个图形化测试, 决定 Z 是否是估计因果效应变量的一个充分集, 并要求 Z 中不存在 X 的子节点, 且 Z 会阻止 X 和 Y 之间包含指向 X 的每一条路径. $p(y|x) \neq p(y|do(x))$, 其中的符号“do”表示假设 X=x.

协变量调整已经在因果推理问题中得到了充分研究, 但本文的研究是文本分类中的应用. 假设已知一个训练集 $D = \{(x_i, y_i, z_i)\}_{i=1}^{n}$, 集合中的每个实例均包含一个术语特征向量 x, 一个标签 y 和一个协变量 z. 本文的目的是对一些新实例 x_i 的标签 y_i 进行预测, 同时控制一个未观测到的混杂因子 z_i. 也即: 本文假设混杂因子可在训练阶段观察到, 但无法在测试阶段观察到.

所提方法的有向图模型如图 1 所示, 给出了对文本分类的一种省略混杂因子 Z 的判别式方法, 假设混杂因子对 $P(Y|Z)$ 中的向量和目标标签均有影响, 用已观察到的向量 x 为条件的 logistic 回归分类器对 $P(Y|X)$ 进行建模, 该模型的结构确保 Z 可以满足用于调整的协变量标准.

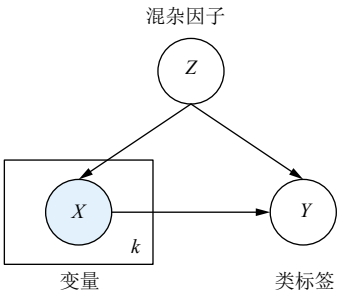


图 1 本文方法的有向图模型

虽然协变量调整方法通常用于识别 X 对 Y 的因果关系, 但并没有解释任何因果关系. 然而, 式 (1) 给出了一个框架, 在控制 Z 中 X 为已知时, 作出对 Y 的预测. 这样, 可以训练一个分类器, 对 $P(Y|Z)$ 从训练数据到测试数据发生变化的情况下具备鲁棒性.

本文使用式 (1) 对测试样本 x 进行分类. 假设对于训练样本, z 为已观测状态, 但没有在测试样本中观察到. 因此, 需要从已标记的训练数据中估计两个变量 $p(y|x,z)$ 和 $p(z)$. 假设 x_i 是一个二进制特征向量, y_i 和 z_i 则是二进制变量. 对于 $p(z)$, 可使用最大似然估计:

$$p(z = k) = \frac{\sum_{i\in D} 1[z_i = k]}{|D|} \tag{2}$$

式中, $1[\cdot]$ 是一个指示函数; D 表示训练集; $p(z)$ 表示为训练集中指示函数之和与训练集样本数的比例. 对于 $p(y|x,z)$, 使用 L2-正则化 logistic 回归, 计算过程可以参考文献[13].

2.2 对调整强度进行调节

从实施的角度来看, 上述方法可表述为: 使用上述

最大似然估计 (Maximum Likelihood Estimation, MLE) 计算出$p(z)$. 本文通过对每个实例x_i, 附加上两个分别表示 z=0 和 z=1 的额外特征$c_{i,0}$和$c_{i,1}$, 高效地计算出$p(y|x,z)$. 如果$z_i = 0$, 将第一个特征设为v_1, 将第二个特征设为 0; 若$z_i = 1$, 则将第二个特征设为v_1, 将第一个特征设为 0. 默认情况下设$v_1 = 1$, 但可因情况而定. 为了对一个新实例进行预测, 使用式 (1) 计算后验概率.

考虑到术语特征向量 x 通常包含数以千计的元素变量, 而协变量 z 添加两个额外特征就能够对分类产生较大影响. 为了理解这一点, 可以考虑正则化 logistic 回归中的权重训练不足的问题[13]. 由于在文本分类中使用了数以千计的相互联系和重叠的变量, 对一个 logistic 回归模型进行的优化涉及到相关变量系数间的权衡问题, 以及由 L2 正则化惩罚所决定的系数量级. 在这个设定中, 少数高预测性特征的存在会导致低预测性特征的系数低于期望数值, 因为高预测性特征在模型中占据主导地位, 会导致在低预测性特征设定中的模型性能较差. 因此, 本文通过引入 z 的特征 (一个潜在的高预测性特征), 故意对 x 中的术语系数进行不充足训练. 例如, 若 z 指的是性别, 则通过使用协变量调整, 使得与其他术语相比, 性别指示性术语具有相对较低量级的系数. 通过对协变量调整的强度进行调节, 改写了 L2 正则化 logistic 回归[13]对数似然函数, 对术语向量的系数和混杂因子的系数进行区分:

$$L(D,\theta) = \sum_{i\in D} \log p_\theta(y_i|x_i,z_i) - \lambda_x \sum_k (\theta_k^x)^2 - \lambda_z \sum_k (\theta_k^z)^2 \quad (3)$$

式中, θ^x为术语向量系数; θ^z为混杂因子系数; θ为θ^x和θ^z的串联参数; λ_x和λ_z分别为控制术语系数和混杂因子系数的正则化强度. 在默认情况下设$\lambda_z = \lambda_x = 1$. 但是, 通过设$\lambda_z < \lambda_x$, 能够降低混杂因子系数$\theta^z$的量级惩罚. 这使得系数$\theta^z$在分类决策中发挥比$\theta^x$更重要的作用, 并增加$\theta^x$中不充分训练的数量. 本文通过提高$v_1$的混杂因子特征数值, 同时将其他特征数值保持为 0, 达到这个效果. 由于本文没有将特征矩阵标准化, 增加v_1的数值同时保持 x 的数值不变, 能够促使θ^z的数值较小, 并有效地使得对θ^z的 L2 惩罚相对小于θ^x.

3 实验与分析

本文使用了 3 个公开数据集进行实验, 其中混杂因子 Z 和分类变量 Y 之间的关系在训练集和测试集中有所差异. 有以下两种情况: 直接控制训练和测试数据之间的差异; Z 和 Y 之间的关系发生了突然.

为对具有不同的$P(Y|Z)$分布的训练/测试集进行采样, 假设已有包含元素$\{(x_i,y_i,z_i)\}$标注后的数据集D_{train}和D_{test}, 其中y_i和z_i为二进制变量. 本文引入了一个偏差参数$P(y = 1|z = 1) = b$; 根据定义可知$P(y = 0|z = 1) = 1 - b$. 对于每个实验, 从每个集合进行不放回抽样$D'_{\text{train}} \subseteq D_{\text{train}}$, $D'_{\text{test}} \subseteq D_{\text{test}}$. 为了模拟$P(Y|Z)$中的变化, 对于训练和测试使用不同的偏差项$b_{\text{train}}$和$b_{\text{test}}$. 因此, 根据以下约束条件进行采样:

1) $P_{\text{train}}(y = 1|z = 1) = b_{\text{test}}$;

2) $P_{\text{test}}(y = 1|z = 1) = b_{\text{test}}$;

3) $P_{\text{train}}(Y) = P_{\text{test}}(Y)$;

4) $P_{\text{train}}(Z) = P_{\text{test}}(Z)$.

其中, 3) 和 4) 的两个约束条件是为了隔离对$P(Y|Z)$中的变化影响. 因此, 从训练数据到测试数据中, 保持$P(Y)$和$P(Z)$不变, $P(Y|Z)$则会发生变化. 本文对$P(Y,Z|X)$的联合分布进行建模, 使用一个 logistic 回归分类器, 其中标签在 Y 和 Z 的积空间中. 在测试阶段, 本文对 z 的可能分配进行求和, 以计算 y 的后验分布.

3.1 实验数据及设置

为构建微博数据集, 本文使用微博信息流应用编程接口采集包含上海和杭州地理坐标的博文. 该实验在 4 天时间中 (2016 年, 6 月 15 日至 6 月 18 日) 共收集了 246 930 条包含上海坐标的博文和 218 945 条包含杭州坐标的博文. 通过移除删减, 并对采集到的博文进行了二次采样, 保留 6000 个用户的博文, 使得所有用户的性别和地理位置均匀分布. 其中, 用户的性别作为预测其位置的混杂变量. 因此, 设$y_i = 1$表示上海, $z_i = 1$表示男性. 构建这个数据集的方式使得数据在 4 种可能的 y/z 配对中均匀地分布.

本文对电影评论中的情感进行预测, 并使用来自“豆瓣”等的 IMDB 电影数据将影片类型作为混杂因子. 该数据集中包括 50 000 条来自 IMDB 的影片评论, 这些评论带有正面或负面的情感标签. 移除了英语或中文停用词和出现次数不到 10 次的术语, 使用一个二进制向量来表示特征的存在与否. 电影是否是由 IMDB 分类所确定为“动作”类型影片作为一个混杂因子. 由此, 对于动作影片, 本文设$z_i = 1$, 对于其他类型影片则

设$z_i = 0$. 这个数据集在 4 种可能的标签/混杂因子配对中是不均匀分布的. 大约 18% 的影片为动作电影, 而对动作电影带有正面情感的评论约占 5%.

对于微博和 IMDB, 本文在训练/测试中进行了变化模拟, 将训练集和测试集的偏差值 b 设为 0.1–0.9, 并对一些分类模型的准确度进行了比较. 对于每对b_{train}和b_{test}, 抽样 5 段训练/测试的分割样本, 并计算平均准确度.

3.2 对比的模型

本文对以下模型进行了比较:

协变量调整 (BA): 即本文所提方法, 通过设置混杂特征的数值$v_1 = 10$, 进行强度更高的协变量调整的模型, 该模型表示为 BAZ10.

Logistic 回归 (LR): 本文研究的主线是一个标准 L2 正则化 logistic 回归分类器, 该分类器不会为混杂因子做任何调整, 仅简单地对$P(Y|X)$进行建模.

二次采样 (LRS): 在训练阶段, 一种移除偏差的简单方式是选择数据的子样本, 使得$P(Y|Z)$均匀分布. 当存在一个较强的混杂偏差时, 该方法会丢弃很多实例, 且实例数量会随着混杂因子数量的增加而进一步减少.

匹配 (M): 匹配通常被用于从观测研究中评估因果效应. 对于每个 $y=i, z=j$ 的训练实例, 采样另一个训练实例, 其中 $y\neq i, z=j$.

3.3 结果分析

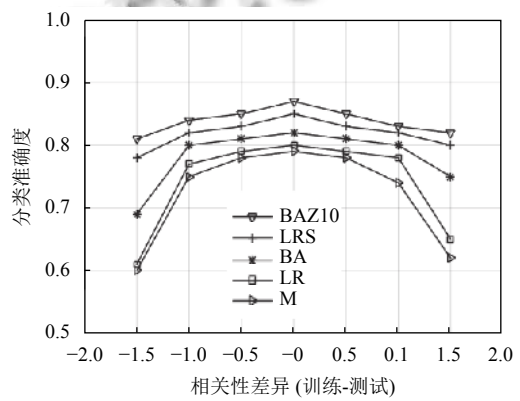
对于微博和 IMDB 电影数据, 本文分别建立了两组实验. 随着训练和测试偏差的差异变化, 研究测试准确度的变化情况. 另外, 计算 Z 和 Y 之间的皮尔森相关性[14], 并给出在测试阶段和训练阶段相关性的差异.

3.3.1 微博实验

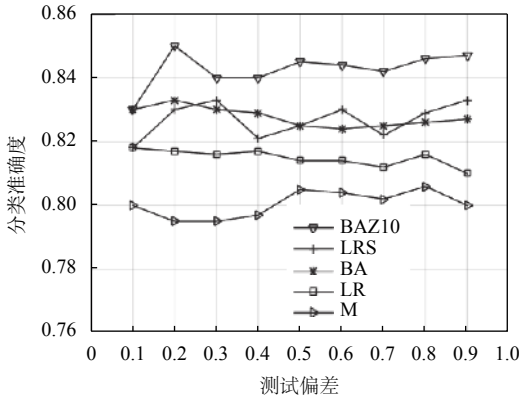
微博数据的实验结果如图 2 所示, 在极值区域表现最佳的方法是 BAZ10 和 LRS. 这两种方法在区间[–1.6, –0.6]和区间[0.6, 1.6]中的性能超过了其他分类器: 与 BA 方法相比, 超过了 15 分; 与 LR 和 M 方法相比超过了 20 分; 比 SO 方法超过了 30 分. 而在这个区间之外的中间区域, BAZ10 方法的性能仅次于 BA 和 LR 方法. 此外, 当相关性差异为 0 时, BAZ10 方法的最大准确度损失大约为 2 分. 这一结果表明, BAZ10 方法对混杂因子的鲁棒性明显高于 LR 方法, 前者在混杂因子影响较小的情况下, 仅会产生最低限度的准确度损失.

在所有训练偏差上, 每个测试偏差的平均准确度如图 2(b) 所示. BA 和 BAZ10 方法在总体上比其他方法的准确度更高. SO 的总体性能不佳, 与其他方法相比, 其准确度要低 4 到 8 分.

为了找到 BAZ10 方法比其他方法的准确度和鲁棒性更高的原因, 本文给出了当偏差值为 0.9 时, LR、BA 和 BAZ10 分类器的系数, 如图 3 所示. 由图 3 可知, 根据χ^2统计数据, 10 个最能预测类标签的特征和 10 个最能预测混杂变量的特征. 与位置相关的特征权重在协变量调整方法中有少许下降, 但依然保持相对较重要的地位. 与之相反, 与性别相关的特征权值在协变量调整方法中则非常接近 0.



(a) 训练集和测试集不同时微博预测准确度的变化



(b) 训练偏差上测试偏差的平均准确度

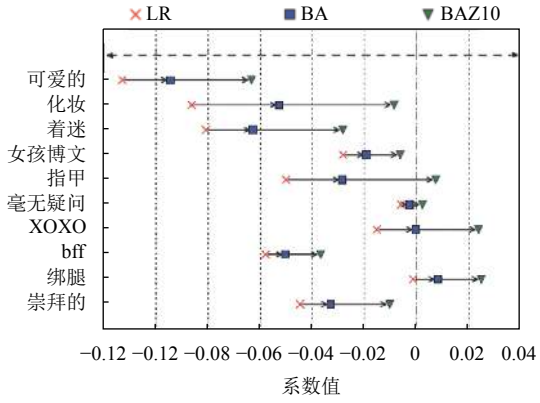
图 2 微博数据的实验结果

已知拟合数据偏差的强度时, Simpson 悖论[15]的特征百分比如图 4 所示, 其中, 微博数据中大概包含 22 K 的特征. 由图可知, BAZ10 的 Simpson 悖论特征数量相对保持不变; 而在其他方法中, 该特征数量则在偏差接近极值时迅速增长.

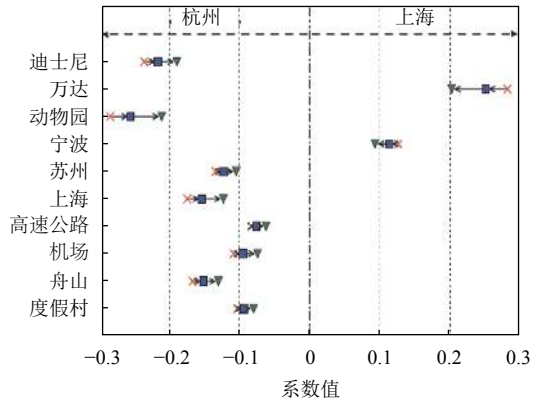
3.3.2 IMDB 实验

图 5 给出了 IMDB 数据的实验结果. 结果显示, BA 和 BAZ10 是对混杂偏差的鲁棒性最好的方法.

LRS 方法鲁棒性最低. LRS 方法的结果不理想原因可能是在 IMDB 数据中, y/z 变量的分布不均衡, 使得 LRS 方法每次仅能在很小比例的训练数据上拟合. 这也是在 IMDB 实验中, 整体准确度的变化幅度要比微博实验小得多的原因.



(a) 与混杂因子 (性别) 相关性最高的10个特征



(b) 与类标签 (位置) 相关性最高的10个特征

图 3 根据卡方统计的结果

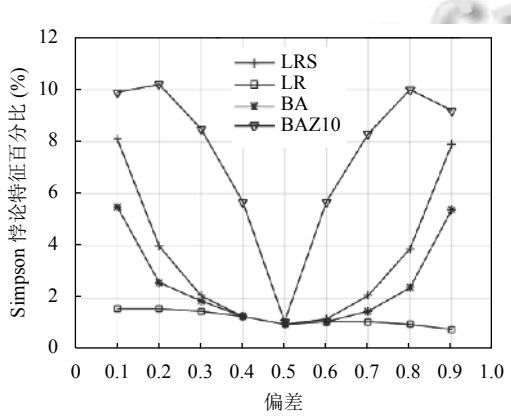
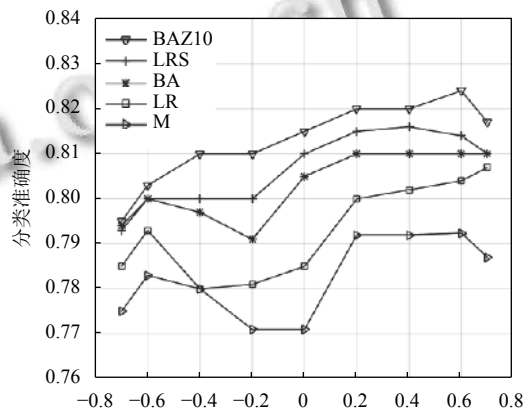


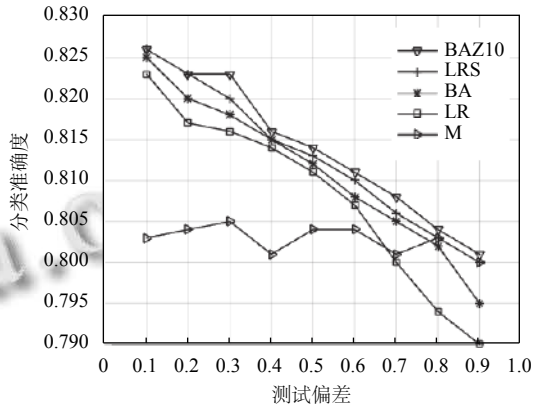
图 4 Simpson 悖论的特征百分比

3.4 参数分析

对于 IMDB 和微博实验, 本文还计算了一个成对的 t-测试, 以使用相关性差异的每个数值对 BAZ10 和 LR 方法进行比较. 实验结果发现, 在 19 个案例中, BAZ10 的性能优于 LR; 在 8 个案例中, LR 的性能优于 BAZ10; 而在 5 个案例中, 结果并没有明显差别. 从图中还可以观察到, 当测试数据与训练数据相对于混杂因子非常类似时, BAZ10 的性能大致相当或稍弱于 LR 方法; 然而, 当测试数据在混杂因子上与训练数据不同时, BAZ10 的性能要优于 LR.



(a) 训练集和测试集不同时准确度的变化



(b) 测量偏差与准确度关系

图 5 IMDB 数据的实验结果

总之, 当在混杂因子的影响中存在极端和突然转变时, 最好的方法是丢弃发生该转变之前的大部分数据. 然而, 一旦在该转变后可用的实例数量适中时, BAZ10 方法能够作出调整解决混杂偏差的问题.

关于参数分析, 本文以 BA 方法为例, 图 6 给出了控制着协变量调整强度的v_1参数影响. 该图给出了c_0和c_1成比例系数绝对值的变化, 及当v_1在微博中增加时准确度的变化. 这些结果是在训练数据集偏差差异较大

的情况下产生的, 从图 6 中可看到, 当v_1小于 10^{-1} 时, 准确度较低, 但较稳定. 然后随v_1的增加而增长, 并且在 $v_1 = 10$时, 开始大幅攀升. 这个数据集中, 准确度在两个峰值之间出现了 15 点增益. 对于所有实验中给定 $v_1 = 10$, 使用交叉验证可以选择出能够产生期望鲁棒性的v_1数值.

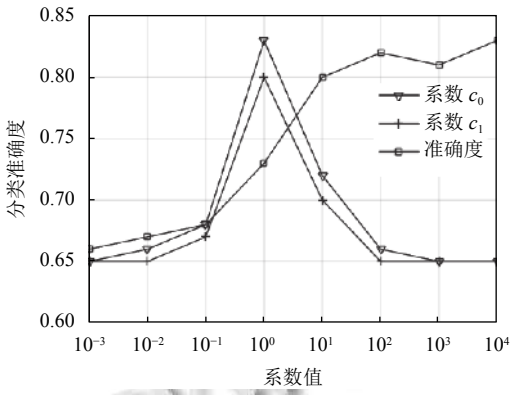


图 6 混杂因子特征系数和准确度

4 结论与展望

本文提出了一个快速有效的文本分类方法, 即使用协变量调整来控制混杂因子. 在 3 个不同的数据集上, 本文发现协变量调整能够在混杂关系从训练数据到测试数据发生变化时, 提高分类器的鲁棒性, 并且在混杂偏差很大的情况下, 可以使用一个额外的参数对调整的强度进行调节. 协变量调整不但能够降低与混杂因子相关的系数量级, 而且可以纠正与目标类标签相关联的系数标注.

未来本文将研究在训练阶段仅有 Z 的带噪估计, 以及 Z 是一个变量向量的情况.

参考文献

1 苏金树, 张博锋, 徐昕. 基于机器学习的文本分类技术研究进展. 软件学报, 2006, 17(9): 1848–1859.

2 宋胜利, 王少龙, 陈平. 面向文本分类的中文文本语义表示方法. 西安电子科技大学学报 (自然科学版), 2013, 40(2): 89–97, 129.

3 王啸宇, 郭代红, 徐元杰. 基于文本分类技术的住院患者药源性变态反应自动监测模块研究. 中国药物应用与监测, 2016, 13(2): 117–120.

4 Fukuchi K, Sakuma J, Kamishima T. Prediction with model-based neutrality. Proceedings of European Conference on Machine Learning and Knowledge Discovery in Databases. Prague, Czech Republic. 2013. 499–514.

5 黄章树, 叶志龙. 基于改进的 CHI 统计方法在文本分类中的应用. 计算机系统应用, 2016, 25(11): 136–140. [doi: 10.15888/j.cnki.csa.005393]

6 刘露, 彭涛, 左万利, 等. 一种基于聚类的 PU 主动文本分类方法. 软件学报, 2013, 24(11): 2571–2583. [doi: 10.3724/SP.J.1001.2013.04467]

7 周庆平, 谭长庚, 王宏君, 等. 基于聚类改进的 KNN 文本分类算法. 计算机应用研究, 2016, 33(11): 3374–3377, 3382.

8 王宇达. 因果效应和统计推断[硕士学位论文]. 北京: 北京邮电大学, 2015.

9 Mariani J, Antonietti L, Tajer C, *et al*. Gender differences in the treatment of acute coronary syndromes: Results from the Epi-cardio registry. Revista Argentina de Cardiología, 2013, 81(4): 287–295. [doi: 10.7775/rac.v81.i4.2330]

10 Breitenstein MK, Pathak J, Simon G. Studying the confounding effects of socio-ecological conditions in retrospective clinical research: A use case of social stress. AMIA Joint Summits on Translational Science Proceedings, 2015, 2015: 41–45.

11 Pearl J. On measurement bias in causal inference. Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence. Catalina Island, CA, USA. 2010. 425–432.

12 吴江, 侯绍新, 靳萌萌, 等. 基于 LDA 模型特征选择的在线医疗社区文本分类及用户聚类研究. 情报学报, 2017, 36(11): 1183–1191. [doi: 10.3772/j.issn.1000-0135.2017.11.010]

13 赵谦, 孟德宇, 徐宗本. $L_{1/2}$ 正则化 Logistic 回归. 模式识别与人工智能, 2012, 25(5): 721–728. [doi: 10.3969/j.issn.1003-6059.2012.05.001]

14 介科伟. 基于 Pearson 相关性分析的高校学生恋爱模型. 首都师范大学学报 (自然科学版), 2017, 38(6): 8–13.

15 吴小安. 辛普森悖论——逻辑进路和因果进路之争. 自然辩证法通讯, 2018, 40(5): 53–59.