计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2021,30(2):188−193 [doi: 10.15888/j.cnki.csa.007782]

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

基于 BIG-WFCHI 的微博信息关键特征选择方法①

殷仕刚[1], 安　洋[1], 蔡欣华[2], 屈小娥[2]

[1](西安理工大学 信息化管理处, 西安 710048)
[2](西安理工大学 计算机科学与工程学院, 西安 710048)
通讯作者: 殷仕刚, E-mail: yinsg@xaut.edu.cn

摘　要: 特征选择是用机器学习方法提高转发预测精度和效率的关键步骤, 其前提是特征提取. 目前, 特征选择中常用的方法有信息增益 (Information Gain, IG)、互信息和卡方检验 (CHI-square test, CHI) 等, 传统特征选择方法中出现低频词引起的信息增益和卡方检验的负相关、干扰计算等问题, 导致分类准确率不高. 本文首先针对低频词引起的信息增益和卡方检验的负相关、干扰计算等问题进行研究, 分别引入平衡因子和词频因子来提高算法的准确率; 其次, 根据微博信息传播的特点, 结合改进的 IG 算法和 CHI 算法, 提出了一种基于 BIG-WFCHI (Balance Information Gain-Word Frequency CHI-square test) 的特征选择方法. 实验分析中, 本文采用基于最大熵模型、支持向量机、朴素贝叶斯分类器、KNN 和多层感知器 5 种分类器对两个异构数据集进行了测试. 实验结果表明, 本文提出的方法能有效消除无关特征和冗余特征, 提高分类精度, 并减少运算时间.

关键词: 微博信息; 特征选择; 机器学习; 信息增益; 卡方检验

引用格式: 殷仕刚,安洋,蔡欣华,屈小娥.基于 BIG-WFCHI 的微博信息关键特征选择方法.计算机系统应用,2021,30(2):188–193. http://www.c-s-a.org.cn/1003-3254/7782.html

Key Feature Selection Method for Weibo Information Based on BIG-WFCHI

YIN Shi-Gang[1], AN Yang[1], CAI Xin-Hua[2], QU Xiao-E[2]

[1](Department of Information Management, Xi'an University of Technology, Xi'an 710048, China)
[2](School of Computer Science and Engineering, Xi'an University of Technology, Xi'an 710048, China)

Abstract: Feature selection, whose premise is feature extraction, is a key step to improve the accuracy and efficiency in retweeting prediction through achine learning methods. Currently, the approaches commonly adopted in feature selection include Information Gain (IG), mutual information, and CHI-square test (CHI). In the traditional feature selection methods, such problems of IG and CHI as negative correlation and interference calculation elicited by low-frequency words lead to low classification accuracy. In view of these problems, we introduce a balance factor and a word frequency factor in this study to increase the algorithm accuracy. Then, according to the spread characteristics of Weibo information, combined with the improved IG and CHI algorithms, we propose the feature selection method based on Balance Information Gain-Word Frequency CHI-square test (BIG-WFCHI). Furthermore, we experimentally test the proposed method with five classifiers including maximum entropy model, support vector machine, naive Bayes classifier, K-nearest neighbor, and multi-layer perceptron on two heterogeneous data sets. The results show that our method can effectively eliminate both irrelevant and redundant features, increase the classification accuracy, and reduce the running time.

Key words: Weibo information; feature selection; machine learning; Information Gain (IG); CHI-square test (CHI)

① 基金项目: 国家自然科学基金 (61672027)
Foundation item: National Natural Science Foundation of China (61672027)
收稿时间: 2020-06-15; 修改时间: 2020-07-14; 采用时间: 2020-07-27; csa 在线出版时间: 2021-01-27

目前, 作为现实社会网络的延伸, 微博平台已经成为网民表达意见、交流信息的热门网站平台. 据中国互联网络信息中心 (CNNIC) 第 45 次《中国互联网络发展状况统计报告》显示[1], 微博是我国三大社交应用之一. 在抗击新冠肺炎疫情过程中, 上亿用户通过微博关注最新疫情、获取防治服务、参与公益捐助. 截至 2020 年 2 月 4 日, 微博热搜榜上疫情相关话题的占比超过 60%. 显然, 新兴媒体已经渗透到我们的生活中, 给我们的信息获取和社会互动带来了巨大的变化. 然而, 由于缺乏对内容的即时审查, 虚假信息极易产生和迅速传播, 给社会带来负面影响. 因此, 准确、有效地预测微博的传播范围, 对于防止虚假信息传播具有重要意义.

利用机器学习方法预测微博传播范围的前提是提取微博转发特征. 因此, 选择有效的特征是提高预测精度和效率的关键步骤, 通过选择有效的特征可以在不损失处理速度和性能的前提下消除不相关和冗余的特征.

通过对文献 [2,3] 实验结果的分析, 发现: 1) IG 和 CHI 方法表现良好, 表明高频词有利于分类; 2) 相反, MI 有效性较差的原因在于其固有的低频词优势, 这一缺陷导致了预测能力差和学习能力差[4]. 代六玲等[5] 在研究中也发现将单一的方法进行组合应用可以提高特征选择的准确率, 并大幅度缩短分类训练时间. 李玉鑑等[6] 将 DF 和 CHI 相结合不仅保留了 CHI 方法能够考虑特征词项与类别相关的优点, 而且利用文档频率 DF 值来去除掉低频词, 降低了 CHI 对低频特征词的权重, 增强了对关键特征的识别能力. Qian 等[7] 将信息理论与集合论理论相结合, 解决了特征选择中的不完全数据问题, 但分类数据与数值数据的共存却悬而未决问题. Wang 等[8] 进一步发现, 为了克服 CHI 的缺陷, CHI 常常与词频等其他因素相结合. Guyon 等[9] 也发现 IG 受到冗余相关特征的影响.

通过以上分析, 发现直接简单的将 DF 和 CHI 进行结合很难去除冗余特征. 相反, 它甚至可以忽略低频词的关键特征. 本文对传统 IG 和 CHI 特征选择方法进行了研究分析, 针对 IG 算法低频特征词对运算结果产生干扰的问题, 引入平衡因子进行调节; 针对 CHI 算法存在的负相关问题, 引入词频因子来提高算法准确率. 在此基础上, 根据微博信息传播特点, 结合改进的 IG 和 CHI 算法, 提出了一种基于 BIG-WFCHI (Balance Information Gain-Word Frequency CHI-square test) 的特征选择方法. 最后, 以 2017 年微博数据和 Reddit 社区数据, 测试 BIG-WFCHI 的性能. 实验结果表明 BIG-WFCHI 特征选择方法能够提高信息分类准确率, 且降低了运算时间和成本.

1 BIG-WFCHI 微博信息关键特征选择方法

1.1 信息增益

在信息论中, 熵表示信息中包含的平均信息量. 对于特征, 熵度量它们对分类的有用程度. 假设特征 t 有 m 个可能值, $v=\{v_1|\ v_2\ |\cdots|\ v_m\}$, $p_i(i=1,2,\cdots,m)$ 是 v_i 的概率, 那么 t 的信息熵可以定义为:

$$H(t)=-\sum_{i=1}^{m}p_i\log_2 p_i \tag{1}$$

其中, 较低的熵表示更简单的分布. 注意, 熵为 0 意味着所有的样本都有相同的值. 相比之下, 熵越大表明无序分布越多. 当特征分布均匀时, 在 $\log_2 m$ 处达到最大熵.

信息增益是根据系统的原始熵与系统具有固定特征的条件熵之差定义的, 它描述了特征的信息量. 一般来说, 一个特征越不确定, 它包含的信息就越多. 特征 t 的 IG 定义为式 (2)[3,10].

$$\begin{aligned}IG(t)=&H(c)-H(c|t)\\=&-\sum_{i=1}^{n}p(c_i)\log_2 p(c_i)+p(t)\sum_{i=1}^{n}P(c_i|t)\log_2 p(c_i|t)\\&+p(\bar{t})\sum_{i=1}^{n}p(c_i|\bar{t})\log_2 p(c_i|\bar{t})\end{aligned} \tag{2}$$

其中, $p(c_i)$ 表示类别 c_i 的出现概率, 对于特征 t 和类别集 $C=\{c_i,\ i=1,2,\cdots,n\}$, IG 利用类别 c_i 中 t 出现 ($p(t)$) 和不出现 ($p(\bar{t})$) 的概率来度量其在 C 上的信息增益, 因此, 较大的信息增益表示 t 对 C 的贡献较大, 这使得 IG 方法更有可能选择信息增益较大的特征到一个类别.

1.2 卡方检验

卡方检验 (CHI-square test)[6] 又称为 χ^2 检验, 是检验特征是否服从某一理论分布或假设分布的假设检验之一, 属于自由分布的非参数检验.

其基本思想是, 首先假设 H_0 是真的, 然后基于 H_0 计算 χ^2 来描述观测值与期望值之间的差距. 利用 χ^2 分布和自由度, 可以得到当前统计量在 H_0 下的概率 p.

卡方检验可以用来衡量特征 t 和类别 c_i 之间的相关性. 假设 t 和 c_i 服从单自由度的 χ^2 分布. 其中, N 表示数据集的大小; B 表示 c_i 中具有特征 t 的子集的大小; D 表示 c_i 中不具有特征 t 的子集的大小; L 表示 c_i 中不具有特征 t 的子集的大小, M 表示 c_i 中不具有特征 t 的子集的大小. c_i 中特征 t 的 χ^2 值为:

$$\chi^2(t,c_i)=\frac{N(BM-DL)^2}{(B+L)(D+M)(B+D)(L+M)} \tag{3}$$

当 $\chi^2(t, c_i)$=0 时, 特征 t 和 c_i 是独立的, χ^2 的值越大它们的相关性越强.

对于多类问题, 首先计算 t 和 c_i 的 χ^2 值, 然后分别在整个数据集上测试特征 t 的 χ^2 值.

$$\chi^2_{\text{avg}}(t)=\sum_{i=1}^{n}P(c_i)\chi^2(t,c_i) \tag{4}$$

$$\chi^2_{\max}(t)=\max_{1\le i\le n}\{\chi^2(t,c_i)\} \tag{5}$$

其中, n 表示类别数. 式 (4) 是分类特征的平均 χ^2 值, 式 (5) 是最大值. 根据 χ^2 值得到排序后的特征列表, 然后根据排序后的列表选择特征.

1.3 基于 BIG-WFCHI 的微博信息关键特征选择算法

信息增益和 χ^2 方法只计算整个数据集中每个特征的频率, 而不考虑特定类别的特征 (转发/不转发). 这两种方法只关注具有一定特征的微博数量, 而不关注特定类别微博的频率. 这夸大了低频特征的作用, 导致分类精度下降[11].

因此, 除了使用基于微博数量的统计方法外, 还需要考虑所有类别特征的概率分布, 本文引入词频因子 E 作为标准度量, 它表示出现在一个类别中的特征的总频率.

设在微博数据集中, 属于类别 C_i 的微博是 $d_1, d_2, \cdots, d_n$, 特征 t 微博 $d_k(1\le k\le n)$ 中出现的次数为 $f_{ik}(t)$, 特征 t 在 C_i 中出现的次数为 $f_i(t)$. 词频因子 E 为特征 t 在某类 C_i 中出现的总词频, 如式 (6) 所示.

$$f_i(t)=\sum_{k=1}^{n}f_{ik}(t) \tag{6}$$

除了上述导致结果不理想的原因外, 传统的信息增益方法更有可能选择在一个特定类别中出现较少而在其他类别中出现较多的特征, 而不是在一个特定类别中出现较多而在其他类别中出现较少的有价值特征. 为了解决这个问题, 需要设置一个平衡因子, 以确保当一个特定类别的无关特征 (或受影响较小的特征) 发生时, 该参数变为负值或非常小的正值, 表明该特征具有负相关性或贡献较小. 平均值可以是一个简单有效的标准来衡量特征对类别的影响. 因此, 本文引入平衡因子 F 为:

$$F=df_i(t)-\overline{df_i(t)} \tag{7}$$

平衡因子 F 为分类 C_i 中包含特征 t 的微博数与各分类出现特征 t 的微博平均数的差值, 如式 (7) 所示. 其中, $df_i(t)$为在分类 C_i 中包含特征 t 的微博数; $\overline{df_i(t)}$ 为数据中各分类出现特征 t 的微博平均数, $\overline{df_i(t)}=\frac{1}{n}\sum_{i=1}^{n}df_i(t)$, n 为数据集的分类个数.

通过式 (2)、式 (6) 和式 (7) 得出:

$$\begin{aligned}BIG(t)&=IG\times E\times F\\&=E\times F\times\left[-\sum_{i=1}^{n}p(c_i)\log_2p(c_i)+p(t)\sum_{i=1}^{n}p(c_i|t)\log_2p(c_i|t)\right.\\&\quad\left.+p(\bar{t})\sum_{i=1}^{n}p(c_i|\bar{t})\log_2p(c_i|\bar{t})\right]\end{aligned} \tag{8}$$

因此, IG 避免忽略特定类别中的特征频率, 并选择在特定类别中出现较少但在其他类别中出现较多的特征.

从式 (3) 可以看出, D 和 L 变大, 而 B 和 M 变小. 即 $DL>BM$, 这意味着由于特定类别的频率较低, 特征的统计值被夸大. 因此, 这些非最优特征更有可能被选择. 这就是所谓的负相关[12]. 为了克服这个问题, 如式 (9) 所示, 对式 (3) 进行限定.

$$\begin{aligned}&WFCHI(t,c)\\&=\chi^2(t,c_i)\\&=\begin{cases}\dfrac{N(BM-DL)^2}{(B+L)(D+M)(B+D)(L+M)}\times E, & BM-DL>0\\0, & BM-DL\le 0\end{cases}\end{aligned} \tag{9}$$

基于上述对 IG 和 CHI 特征选择方法优缺点的分析, 结合两个引入的词频因子 E 和平衡因子 F, 提出一种基于 BIG-WFCHI 特征选择算法. 其计算方法如式 (10) 所示.

$$BIC-WFCHI(t)=BIG(t)\times WFCHI(t,c) \tag{10}$$

为了更加准确的描述 *BIG-WFCHI* 算法, 引入以下两个定义: *BIG-WFCHI* 离散度和 *BIG-WFCHI* 特征类间差值.

定义 1. BIG-WFCHI 离散度, 记为 D_p*BIG-WFCHI*,

表示每个类别中特征 *BIG-WFCHI* (以下简称 *IC*) 值的分散程度, 用式 (11) 中的 D_p 表示.

$$D_p(IC)=\frac{\sum_{i=1}^{m}\sum_{j=1}^{n}\left(IC_{ij}-\overline{IC_i}\right)^2}{n} \tag{11}$$

其中, m 表示特征总数, n 表示类别数量, IC_{ij} 表示第 i 个特征在第 j 个分类的 *BIG-WFCHI* 值, $\overline{IC_i}$ 为第 i 个特征在所有类中 IC_{ij} 的平均值.

BIG-WFCHI 离散度可以用来测量特征的冗余度. 具有较大 *BIG-WFCHI* 离散度的特征具有较强的识别能力, 即它们对分类更具价值.

定义 2. BIG-WFCHI 特征类间差值, 记为 D_f 表示在类间最大 *IC* 值与第二 *IC* 值的差值, 如式 (12) 所示.

$$D_fIG-CHI=\max(IC_i)-\max'(IC_i) \tag{12}$$

其中, $\max(IC_i)$表示第 i 个特征在指定类中最大的 *BIG-WFCHI* 值, $\max'(IC_i)$表示第 i 个特征在指定类中第二大值. D_f 值越大说明特征越特征在特定类别中的分布越密集. 也就是说, 这个特征对分类更为关键.

利用 D_p 和 D_f 进一步分析特征的冗余度, 可以减少特征的维数, 去除冗余特征, 缩短运行时间.

BIG-WFCHI 算法的主要步骤如算法 1.

算法 1. BIG-WFCHI 算法

输入: 原始数据集 $S(t_1, t_2,\cdots, t_n)$, 阈值$\mathcal{E}_p$ 和$\mathcal{E}_f$;
输出: 最优特征子集 S_{best};

(1) 将 S 中各个特征数据进行规范化. 在处理多个特征数据时容易出现运算结果偏向数值较大的特征项, 导致计算结果出现偏置问题. 在本文各个特征值被规范在 1–10 之间以此来规避偏置问题;
(2) 初始化每个特征的 $IC(t)$=0;
(3) 利用式 (6) 和式 (7) 计算 E 和 F 的值;
(4) 根据式 (8) 计算 S 中各个特征 $(t_1, t_2,\cdots, t_n)$ 的$BIG(t)$值;
(5) 根据式 (9) 计算出 S 中特征项的$WFCHI(t,C_i)$值;
(6) 根据$BIG(t)$值和$WFCHI(t,C_i)$计算出式 (10) 中$BIG-WFCHI$值, 对应每个特征项按照降序进行排列放入特征集 S_0;
(7) 根据定义 1, 对 S_0 特征 t_i $(i=1,2,\cdots,n)$ 计算离散度 $D_p(IC)$, 将 S_0 中 $D_p(IC)<\mathcal{E}_p$ 的特征项存入 S_1 特征子集中直到特征集 t_i 中为空;
(8) 根据定义 2, 计算 S_1 特征子集中特征项的 $D_f(IC)$ 值, 将 S_1 特征子集中 $D_f(IC)>\mathcal{E}_f$ 的特征存入 S_{best} 特征子集中.

在这里, 本文利用 E 和 F 来减少低频特征和负相关引起的干扰, 然后根据 D_p 和 D_f 选择特征. 不同的数据集需要不同的阈值$\mathcal{E}_p$ 和$\mathcal{E}_f$, 其中极小的数据集不利于选择, 而较大的数据集去除了一些关键的分类特征. 本文分别以 D_p 和 D_f 的平均值作为阈值$\mathcal{E}_p$ 和$\mathcal{E}_f$.

2 实验分析

2.1 数据集与实验环境

本文采用 2017 年新浪微博数据为实验数据集, 并以 Reddit 社区的“披萨随机行为”为样本 1, 测试 BIG-WFCHI 的通用性. 这两个数据集分别命名为 WBdataset 和 PZdataset. 它们的属性如表 1 所示.

表 1 实验数据集

数据子集	样本数量	类别数	转发成功率(%)
PZdataset	5600	2	11.76
WBdataset	40 000	2	16.234

由于 WBdataset 和 PZdataset 中只有两种状态: retweeted 和 not retweeted, successful 和 not successful, 因此本文将预测视为二值分类. 通过分析 WBdataset 和 PZdataset 的数据记录, 分别提取了 20 个原始特征 [13]. 采用 IG, CHI, BIG-WFCHI, TF-IDF[14,15] 分别从这两个原始特征集中选择在每个方法中贡献大多数值的前 10 个特征作为主要特征.

为了验证 BIG-WFCHI 方法的有效性, 本文选取了 LIBSVM(SVM)[16]、MaxEnt(ME)[17]、Naive-Bayes 分类器 (NBC)、K 近邻 (KNN) 和多层感知器 (MLP) 5 种分类器. 这些分类器通常用于机器学习, 它们的分类结果在效果上有所不同[17–20]. 在此, 本文简要说明了这些方法的主要参数选择. 考虑到数据集是稀疏矩阵, 本文选择 SMO[19] 作为优化算法, 在 LIBSVM 中选择 RBF 作为核函数, 使得数据集有更好的性能. 在 KNN 中, k 的值是通过交叉验证确定的, 得到的最佳结果介于 100 到 150 之间. 由于一个分类模型的精度不是本文研究的重点, 所以在 MLP 中只设置了一个隐藏层.

2.2 实验结果与分析

实验中采用 10 倍交叉验证. 对于每个特征选择方法、分类器和数据集, 我们执行 10 次运行, 然后报告结果的平均值、标准差和弗里德曼检验.

表 2 显示了 4 种特征选择方法的精度. 最高的分类精度用粗体加下划线和突出显示. 从表 2 可以看出, 本文提出的方法在支持向量机、KNN、NBC 和 MLP 上达到了最佳的精度. 该方法在基于 ME 的 PZdataset 和 WBdataset 上分别取得了最佳精度和次优精度.

在 10 倍交叉验证中, 由于每次运行时都会更改训练数据集和测试数据集, 因此分类精度会有所不同. 为了显示精度之间的差异, 表 3 中执行了 10 次运行的标准差. 结果表明, 在两个数据集上, 基于 BIG-WFCHI 的

分类精度标准差在 KNN 中是最小的. 在其他分类器中, 基于 BIG-WFCHI 的标准差也是小的有理数.

表 2 IG、CHI、BIG-WFCHI 和 TF-IDF 的分类精度 (%)

分类器	数据集	IG	CHI	BIG-WFCHI	TF-IDF
ME	PZdataset	80.86	81.01	**84. 30**	81.09
	WBdataset	**93.99**	92.59	93.34	91.80
SVM	PZdataset	79.94	80.16	**81.75**	78.87
	WBdataset	90.15	91.32	**93.55**	91.25
KNN	PZdataset	75.51	75.10	**75.56**	75.18
	WBdataset	86.88	86.99	**88.65**	83.26
NBC	PZdataset	75.06	74.52	**75.27**	75.14
	WBdataset	85.38	86.38	**87.92**	83.98
MLP	PZdataset	65.08	62.89	**69.91**	63.70
	WBdataset	70.32	72.05	**81.28**	69.13

表 3 IG、CHI、BIG-WFCHI 和 TF-IDF 的标准偏差

分类器	数据集	IG	CHI	BIG-WFCHI	TF-IDF
ME	PZdataset	0.0051	0.0046	**0.0029**	0.1054
	WBdataset	**0.0407**	0.0462	0.0459	0.0412
SVM	PZdataset	0.1288	**0.1011**	0.1015	0.1120
	WBdataset	0.0569	0.0575	**0.0527**	0.0547
KNN	PZdataset	0.0032	0.0029	**0.0004**	0.0010
	WBdataset	0.1428	0.1204	**0.1199**	0.1378
NBC	PZdataset	**0.0003**	0.0007	0.0027	0.0027
	WBdataset	0.1395	0.1332	**0.1224**	0.1396
MLP	PZdataset	0.0068	0.2493	**0.0399**	0.0081
	WBdataset	**0.0483**	0.0508	0.0514	0.0509

进一步探讨 10 次运行结果之间是否存在显著差异, 本文对这些分类结果进行了 Friedman 检验. 在所有的测试中, 选择变量无显著性差异作为零假设, 0.05 作为置信水平. 由于篇幅的限制, 本文在只表 4 中显示 WBdataset 上的测试结果. 所有的精确都大于 0.05, 这意味着我们接受了零假设. 10 次 10 倍交叉验证没有显著性差异. 因此, 预测结果的均值是可靠的. 在 PZdataset 上的测试结果显示了相同的结论.

表 5 显示了基于 IG、CHI、BIG-WFCHI 和 TF-IDF 选择的特征的不同分类器分类结果的 AUC 值. 从这些 AUC 值可以很容易地看出 BIG-WFCHI 优于其他 3 种选择方法.

图 1 和图 2 显示基于 IG、CHI、BIG-WFCHI 和 TF-IDF 选择的特征的不同分类器分类结果的 ROC 曲线. 可以看出, 在 4 种分类器中, BIG-WFCHI 选择的特征具有最好的分类效果.

实验结果表明, 在不同的数据集或分类器下, 基于 BIG-WFCHI 选择的特征子集, 分类精度可以提高或至少保持在同一个数量级. 通过以上讨论, BIG-WFCHI 方法可以更有效地选择信息量更大的特征, 实现特征选择具有实际意义.

表 4 WBdataset 中 IG、CHI、BIG-WFCHI 和 TF-IDF 的 Friedman 检验

特征选择方法	检验统计量	ME	SVM	KNN	NBC	MLP
IG	χ^2	8.287	5.368	12.331	8.914	2.953
	p	0.506	0.801	0.195	0.515	0.906
CHI	χ^2	12.226	8.172	17.580	12.096	5.562
	p	0.201	0.518	0.132	0.217	0.786
BIG-WFCHI	χ^2	3.941	2.217	6.083	3.742	1.104
	p	0.915	0.925	0.671	0.921	0.949
TF-IDF	χ^2	9.027	5.838	13.946	8.591	2.237
	p	0.435	0.778	0.173	0.484	0.917

表 5 IG、CHI、BIG-WFCHI 和 TF-IDF 的 AUC 值

分类器	数据集	IG	CHI	BIG-WFCHI	TF-IDF
ME	PZdataset	0.8583	0.8373	**0.8899**	0.7361
	WBdataset	0.8593	0.8511	**0.8766**	0.8458
SVM	PZdataset	0.8463	0.8245	**0.8726**	0.7439
	WBdataset	0.7118	0.7003	**0.8274**	0.6959
KNN	PZdataset	0.8224	0.8050	**0.8532**	0.7593
	WBdataset	0.8187	0.8108	**0.8349**	0.805
NBC	PZdataset	0.8314	0.8109	**0.8587**	0.7494
	WBdataset	0.6925	0.7108	**0.8014**	0.6601
MLP	PZdataset	0.8583	0.8373	**0.8899**	0.7361
	WBdataset	0.8593	0.8511	**0.8766**	0.8458

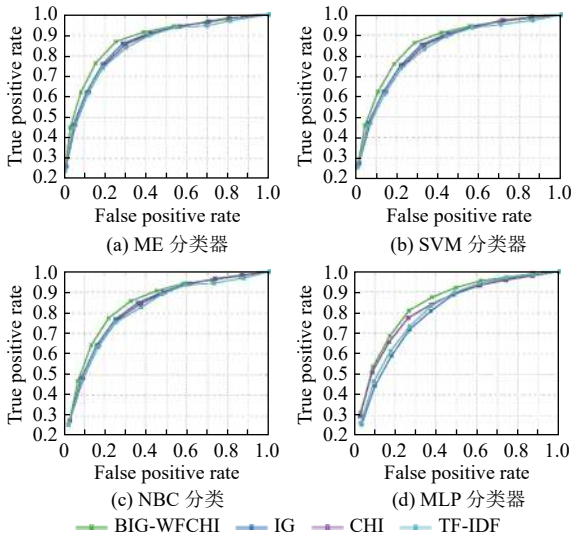

图 1 基于 WBdataset 相关选择特征的不同分类器的 ROC 曲线

3 结论

本文以转发预测为例, 讨论了信息增益、互信息和卡方检验等方法在特征选择中的应用, 但这些方法存在负相关和可能对计算结果产生干扰等缺陷. 本文

引入平衡因子和词频因子来提高算法准确率; 其次, 提出了一种 BIG-WFCHI 特征选择方法. 实验结果表明, 该方法克服了上述缺陷, 消除了冗余贡献, 提高了 ME、支持向量机、NBC、KNN 和 MLP 等分类器的效率.

随着网络数据复杂度和规模的迅速增加, 特征选择变得越来越重要. BIG-WFCHI 特征选择方法能去除冗余特征, 有助于减少计算时间, 节省存储空间, 提高机器学习效率. 因此, 为特征选择提供了一种有效的方法.

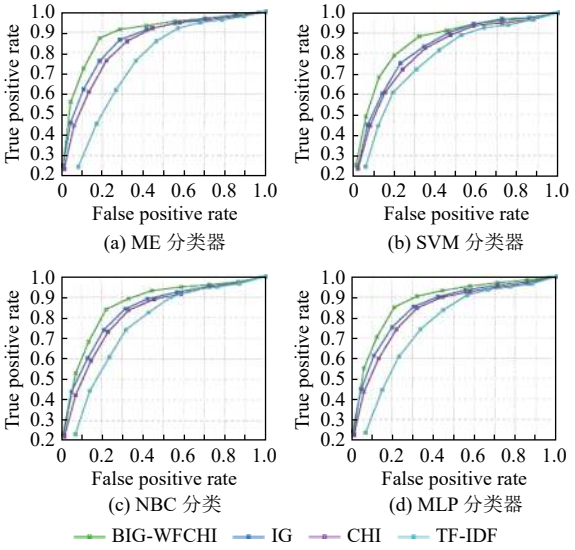

图 2 基于 PZdataset 相关选择特征的不同分类器的 ROC 曲线

参考文献

1 于朝晖. CNNIC 发布《第 45 次中国互联网络发展状况统计报告》. 网信军民融合, 2020, (5): 26–27.

2 Yang YM, Pedersen JO. A comparative study on feature selection in text categorization. Proceedings of the 14th International Conference on Machine Learning. Nashville, TN, USA. 1997. 412–420.

3 Yin CX, Zhang HJ, Zhang R, *et al*. Feature selection by computing mutual information based on partitions. IEICE Transactions on Information and Systems, 2018, E101-D(2): 437–446. [doi: 10.1587/transinf.2017EDP7250]

4 Doquire G, Verleysen M. Mutual information-based feature selection for multilabel classification. Neurocomputing, 2013, 122: 148–155. [doi: 10.1016/j.neucom.2013.06.035]

5 代六玲, 黄河燕, 陈肇雄. 中文文本分类中特征抽取方法的比较研究. 中文信息学报, 2004, 18(1): 26–32. [doi: 10.3969/j.issn.1003-0077.2004.01.005]

6 李玉鑑, 周兰珍, 操卫平. 基于 DF 和 CHI 的联合特征提取方法及其应用. 北京工业大学学报, 2008, 34(9): 995–1000.

7 Qian WB, Shu WH. Mutual information criterion for feature selection from incomplete data. Neurocomputing, 2015, 168: 210–220. [doi: 10.1016/j.neucom.2015.05.105]

8 王皓, 孙宏斌, 张伯明. PG-HMI: 一种基于互信息的特征选择方法. 模式识别与人工智能, 2007, 20(1): 55–63. [doi: 10.3969/j.issn.1003-6059.2007.01.009]

9 Guyon I, Elisseeff A. An introduction to variable and feature selection. The Journal of Machine Learning Research, 2002, 3: 1157–1182.

10 Huang NT, Li RQ, Lin L, *et al*. Low redundancy feature selection of short term solar irradiance prediction using conditional mutual information and Gauss process regression. Sustainability, 2018, 10(8): 2889. [doi: 10.3390/su10082889]

11 Yu L, Liu H. Efficient feature selection via analysis of relevance and redundancy. Journal of Machine Learning Research, 2004, 5: 1205–1224.

12 王宏威, 李国和. 基于属性相似度的连续型特征选择方法. 渤海大学学报 (自然科学版), 2014, 35(4): 350–355.

13 李勇军, 尹超, 于会, 等. 基于最大熵模型的微博传播网络中的链路预测. 物理学报, 2016, 65(2): 020501. [doi: 10.7498/aps.65.020501]

14 Sharma N, Kaur G, Verma A. Survey on text classification (Spam) using machine learning. International Journal of Computer Science and Information Technologies, 2014, 5(4): 5098–5102.

15 Forman G. BNS feature scaling: An improved representation over TF-IDF for SVM text classification. Proceedings of the 17th ACM Conference on Information and Knowledge Management. Napa Valley, CA, USA. 2008. 263–270.

16 Chang CC, Lin CJ. LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2011, 2(3): 27.

17 Tan SB, Cheng XQ, Wang YF, *et al*. Adapting naive Bayes to domain adaptation for sentiment analysis. Proceedings of the 31th European Conference on Information Retrieval. Toulouse, France. 2009. 337–349.

18 路永和, 何新宇. 基于维度索引表的改进 KNN 分类算法. 情报理论与实践, 2014, 37(5): 102–106.

19 李飞, 李红莲. 支持向量机大规模样本快速训练算法. 北京信息科技大学学报 (自然科学版), 2012, 27(2): 83–87.

20 Malouf R. A comparison of algorithms for maximum entropy parameter estimation. Proceedings of the 6th Conference on Natural Language Learning. Taipei, China. 2002. 1–7.