doi: 10.3969/j.issn.0490-6756.2019.01.013

最大判别特征选择算法在文本分类的优化研究

刘 云, 黄荣乘
(昆明理工大学信息工程与自动化学院, 昆明 650500)

摘 要: 采用朴素贝叶斯分类器进行文本分类时,特征选择方法的好坏直接影响到分类器的性能. 本文提出一种最大判别(MD)特征选择算法,由训练得到 N 个类的概率分布后,通过对样本进行测试并得到其特征向量 d 中每个特征词区分类别的能力,并构造出了一个新的特征向量 ε 用于分类,使得从中选取的部分特征词具有最大的类别区分能力. 仿真结果表明,与 cMFD,CSFS 和 CMFS 三种特征选择算法相比,MD 特征选择算法能在选取较少特征词情况下,获得更高的分类精度.

关键词: 相对熵; 杰弗里斯散度; 多项式朴素贝叶斯分类器; 特征选择
中图分类号: TN929.5 **文献标识码**: A **文章编号**: 0490-6756(2019)01-0065-06

Bayesian classifier-based maximum discriminant feature selection algorithm for text classification

LIU Yun, *HUANG Rong-Cheng*
(School of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650500, China)

Abstract: When using Naive Bayes classifier to classify texts, the feature selection method has a direct impact on the performance of the classifier. In this paper, a maximum discrimination (MD) feature selection algorithm is proposed. After N types of probability distributions are obtained through training, the ability to distinguish the categories of each feature in its feature vector d is acquiredby testing the sample, and a new feature vector ε is constructed for classification, the selected features from the feature selection have the maximum discrimination capacity for text categorization. Simulation results show that compared with cMFD, CSFS and CMFS feature selection algorithms, MD feature selection algorithm can obtain higher classification accuracy when fewer features are selected.

Keywords: KL divergence; Jeffrey divergence; Multinomial naive bayes classifier; Feature selection

1 引 言

随着电子文本文档越来越多的使用,将未知类别的文档自动判定为事先定义好的类别是非常有必要的[1-3]. 文档中会存在一些与分类无关的冗余特征,当用含有这些冗余特征的特征向量进行分类时会产生过拟合[4,5]现象,导致分类器的分类性能降低. 所以需要进行特征选择,并用选出的特征子集来对文本进行分类,使得分类器的分类性能得到优化.

Rogerio 等人提出基于类别的最大特征选择算法(Category-dependent Maximum f Features-

收稿日期: 2018-04-10
基金项目: 国家自然科学基金(61262040)
作者简介: 刘云(1973-)男, 云南昆明人, 副教授, 研究方向为无线通信研究. E-mail: liuyunkm@126.com
通讯作者: 黄荣乘. E-mail: 1908616172@qq.com

per Document,cMFDR)[6]. 该算法对训练集中的所有文档进行特征选择,它会排除所有未达到阈值的文档,且阈值依赖于类别,以确保每个类别都具有不同数量的特征. Zhou 等人提出了类子空间特征选择算法(Class Subspace Feature Selection,CSFS)[7]. 利用主成分特征提取方法获得较低维的类子空间,然后为子空间选择出最相关的特征. Feng 等人提出了综合度量特征选择算法(Comprehensive Measurement Feature Selection,CMFS)[8]. 该算法考虑了在所有类下特征词的分布信息,综合测量了特征词在类间和类内的重要性以对特征进行排名选取.

基于朴素贝叶斯分类器本文提出了最大判别特征选择算法(Maximum Discrimination,MD),通过训练集得到 N 个类的条件概率分布后,为测试样本构造了一个二分类假设检验器来对样本的类别进行两次假设检验. 用相对熵(KL 散度)[9,10]分别计算出在两种假设检验下测试样本中的特征词区分两个类别分布的能力,通过分析 KL 散度与贝叶斯分类器分类误差的性质[10]后将两个假设检验下的 KL 散度求和得到了杰弗里斯散度(J 散度)[11,12].

在面对 N 类分类问题时,对于 N 个类别使用了 N 个二分类假设检验器,给定一个特征向量规定为 d 的测试样本,根据类条件概率分布,首先计算出该测试样本假设为第一类时其特征向量 d 中的第一个特征词的 KL 散度,然后按此方法依次算出该样本假设为第 N 类时第一个特征词的 KL 散度,并将这 N 个 KL 散度相加即得到了第一个特征词的 JMH 散度. 对于 M 个特征词,就得到了 M 个 JMH 散度的分数,并降序排列构成一个新的特征向量 ε. 在对样本进行分类时,从特征向量 ε 中选出 r 个具有最大区分类别能力的特征词构成特征向量 S_r,并将样本用特征向量 S_r 表示,基于朴素贝叶斯分类器,计算出样本属于各类时的后验概率后,取后验概率最大的类别作为该样本的类别.

2 相关工作

2.1 样本的表示

在文本分类中,从给定的数据集中,生成一个有着 M 个不同特征词的字典. 当给定一个样本,计算每个词出现的次数,根据词袋生成一个特征向量 $d=[x_1,x_2,\ldots,x_M]^T$,x_i 中的第 i 项对应于字典中第 i 个特征词出现的次数. 它也被称为词语频率(Term Frequency,TF),其中每个类的 TF 分布通常可以用特定多项分布模型来建模.

2.2 朴素贝叶斯

多项式朴素贝叶斯分类器(Multinomial naive Bayes,MNB)是用术语频率来表示样本的最著名分类方法之一. 通过训练,得到在 c 类和文档长度 l 的条件下样本 d 的类条件概率分布为

$$p(d|c,l;\theta_c^m)=\frac{l}{x_1!\ x_2!\ldots x_M!}\prod_{i=1}^{M}p_{ic}^{x_i} \tag{1}$$

其中,$l=\sum_{m=1}^{M}x_m$,在 c 类下样本 M 个特征词的概率构成向量 $\theta_c^m=[p_{1c},p_{2c},\ldots,p_{Mc}]^T$,$i=1,2,\ldots,M$,$p_{ic}\geq 0$ 且 $p_{1c}+p_{2c}+\ldots+p_{Mc}=1$. 为了避免零概率问题,应用“拉普拉斯平滑”技术,使用 MLE 方法,在 c 类下每个特征的概率 $\hat{p}_{ic}$ 的估计值为

$$\hat{p}_{ic}=\frac{l_{ic}+\lambda_1}{l_c+\lambda_2} \tag{2}$$

其中,l_{ic} 是 c 类中第 i 个特征词出现的次数,l_c 是 c 类中特征词的总数. $\lambda_1=1$ 和 $\lambda_2=M$ 是恒定的的平滑参数.

为了简化朴素贝叶斯分类规则,假定所有类的文档长度都为 l,即类的概率密度函数 $p(d|c,l)=p(d|c)$,$p(c|d,l)=p(c|d)$. 当给定要分类的新样本 D_t 时,用词袋生成其特征向量 d_t 并应用以下规则进行分类如下.

$$\begin{aligned} c^* &= \arg\max_{c\in\{1,2,\ldots,N\}} p(c|d_t;\theta_c) \\ &\propto \arg\max_{c\in\{1,2,\ldots,N\}} \log p(d_t|c;\theta_c)+\log p(c) \end{aligned} \tag{3}$$

其中,$p(d_t|c;\theta_c)$,$i=1,2,\ldots,N$ 是基于 MNB 分类器的类条件概率分布由(1)给出.

2.3 二元假设检验

首先考虑一个二分类问题,其中通过训练,c_1 类的分布为 $P_1=p(x|c_1;\theta_1)$,c_2 类的分布为 $P_2=p(x|c_2;\theta_2)$. 对一个测试样本进行二分类假设检验,如果一个样本分为 c_1 类,则接受假设 H_1,如果一个样本分为 c_2 类,则接受 H_2,即有 $p(x|c_1)=p(x|H_1)$和 $p(x|c_2)=p(x|H_2)$,并且 $p(x|H_i)$代表在其它类别下的类条件概率分布.

根据信息论[8],给定一个特征向量为 d 的测试样本 x 时 P_1 和 P_2 两个概率分布之间的 KL 散度 $KL(P_1,P_2)$定义为

$$KL(P_1,P_2)=\int_x p(x|H_1)\log\frac{p(x|H_1)}{p(x|H_2)}dx=E_{p_1}\left[\log\frac{p(x|H_1)}{p(x|H_2)}\right] \tag{4}$$

其中,$E_{P_1}[x]$表示在概率分布 P_1 下 x 的期望. 根据式(1)和方程(4),x 中的特征词在两个多项式分布 P_1 和 P_2 之间的 KL 散度为

$$KL(P_1,P_2)=\sum_{i=1}^{M}p_{i1}\log\frac{p_{i1}}{p_{i2}} \tag{5}$$

其中,$0\leq p_{ic}\leq 1$ 且 $\sum_{i=1}^{M}p_{ic}=1,i=1,2,\ldots,M,c=1,2$.

在方程(3)的 MAP 规则下,如果

$$\begin{aligned}&\log p(x|H_1)+\log p(H_1)>\log p(x|H_2)+\log p(H_2)\\&\Rightarrow\log\frac{p(x|H_1)}{p(x|H_2)}>-\log\frac{p(H_1)}{p(H_2)}=\gamma\end{aligned} \tag{6}$$

则将样本 x 分为类 c_1,即接受 H_1. 其中对数似然比 $\log\left(\frac{p(x|H_1)}{p(x|H_2)}\right)$鉴别样本 x 属于 c_1 类还是 c_2 类 h.

由式(6)知,在 P_1 分布下对数似然比的期望值有

$$KL(P_1,P_2)>\gamma \tag{7}$$

在贝叶斯分类器中,从(4)的 KL 散度定义知,$KL(P_1,P_2)$代表假设样本属于 c_1 类时样本中的特征词区分 P_1 分布和 P_2 分布的能力大小,KL 越大则样本中的特征词越容易区分类别. 随着中心极限定理的扩展,由切尔诺夫[13,14]证明,当有大量样本时,类型Ⅰ误差 α,即错误地接受 H_1 的概率,渐近地有

$$\lim_{n\to\infty}\log\frac{1}{\alpha^*}=\lim_{n\to\infty}KL(P_1,P_2;O_n) \tag{8}$$

其中,O_n 表示 n 个独立的样本,且当有无数样本时,KL 散度值越大Ⅰ型误差越小.

KL 散度不是对称的. 同样,在方程(3)中的 MAP 规则下,$KL(P_2,P_1)$代表假设样本属于 c_2 类时样本中的特征词区分 P_2 分布和 P_1 分布的能力大小,如下式.

$$KL(P_2,P_1)>-\gamma \tag{9}$$

当有大量样本时,Ⅱ型误差 β^*,即错误地接受 H_2 的概率为

$$\lim_{n\to\infty}\log\frac{1}{\beta^*}=\lim_{n\to\infty}KL(P_1,P_2;O_n) \tag{10}$$

为了使Ⅰ型误差和Ⅱ型误差以渐近的方式最小化,可以使用 J 散度[11,12].

$$J(P_1,P_2)=KL(P_1,P_2)+KL(P_2,P_1) \tag{11}$$

通过结合等式(7)和方程(9),有

$$KL(P_1,P_2)>\gamma>-KL(P_2,P_1) \tag{12}$$

因为 $KL(P_1,P_2)\geq 0,KL(P_2,P_1)\geq 0$,一个较大的 $J(P_1,P_2)$会使Ⅰ型误差和Ⅱ型误差渐近地变小. 因此,为使分类误差最小化(也就是使 KL 或 J 散度最大化),就要选择出具有最大判别类别能力的特征词,然而,J 散度只适用于二分类问题,接下来扩展多重假设检验的 J 散度(即多分类).

3　MD 特征选择算法

3.1　多分布假设检验

将二元假设检验进行扩展,其中通过训练,N 个类别的分布为 $P=\{P_1,P_2,\ldots,P_N\}$,根据 J 散度,给定一个测试样本,对于 N 个分布得到了 $JMH(P_1,P_2,\ldots,P_N)$散度,定义如下.

$$JMH(P_1,P_2,\ldots,P_N)=\sum_{i=1}^{N}KL(P_i,\bar{P}_i) \tag{13}$$

其中,$\bar{P}_i$ 是除了第 i 个分布外剩下的 $N-1$ 个分布的联合,定义为

$$\bar{P}_i=\sum_{k=1,k\neq i}^{N}\omega_{ki}P_k \tag{14}$$

由于每个二分类检验器的 KL 散度之和 $KL(P_i,\bar{P}_i)+KL(\bar{P}_i,P_i)$,都能衡量样本中的特征词区分 P_i 分布与 $\bar{P}_i$ 分布的能力,针对 N 分类问题,构造了 N 个二元假设检验器,得到了测试样本中每个特征区分 N 个类别能力(JMH 散度).

在每个二元假设检验器中,$\bar{P}_i$ 由所有其余 $N-1$个类别的混合分布表示,先验系数 φ 由下式给出.

$$\varphi_{ki}=\frac{p_{ck}}{\sum_{m=1,m\neq i}^{N}p_{cm}} \tag{15}$$

其中,p_{cm} 是 c_m 类的先验概率. 当 $N=2$ 时,$JMH(P_1,P_2)=J(P_1,P_2)$.

在 N 类分类问题中,给出了基于最大 JMH 散度的特征选择算法. 根据类条件概率分布,用式(5)和式(13)计算出测试样本中每个特征词的 JMH 散度分数,并降序排列,得到一个特征索引集

$\varepsilon=(e_1,e_2,\ldots,e_M)$,其中,这个算法的计算复杂度是 $O(MN)$. 对于 N 类分类问题的最大判别特征选择算法如算法 1.

算法 1 最大判别特征选择算法

输入:

(1) M 个特征的概率参数向量:$\theta_c=[p_{1c},p_{2c},\ldots,p_{Mc}]$,$c=1,2,\ldots,N$;

(2) N 个类别的先验概率:$p_1,p_2,\ldots,p_N$;

算法:

for $i=1:M$ do

1) 形成两个变量 x_i 和 $\bar{x}_i$,x_i 代表第 i 个特征词,$\bar{x}_i$ 代表其余的 $M-1$ 个特征词,P_{ic} 表示类的分布,$c=1,2,\ldots,N$;

for $c=1:N$ do

2) 形成两个分布 P_{ic} 和 $\bar{P}_{ic}$,$\bar{P}_{ic}$ 代表变量 $\bar{x}_i$ 在其余 $N-1$ 类下的联合分布;

3) 用式(5)计算出第 i 个特征词在 P_{ic} 分布和 $\bar{P}_{ic}$ 分布之间的 KL 散度 $KL(P_{ic},\bar{P}_{ic})$;

end

4) 计算第 i 个特征词的 JMH 散度:$JMH_i=\sum_{c=1}^{N}KL(P_{ic},\bar{P}_{ic})$;

end

5) 根据 JMH 得分将 M 个特征降序排列:$JMH_{e1}>JMH_{e2}>\cdots>JMH_{eM}$;

输出:

排名后的特征索引 $\varepsilon=\{e_1,e_2,\ldots,e_M\}$

从 $\varepsilon=(e_1,e_2,\ldots,e_M)$ 中所选的 r 个特征词有最大区分类别的能力:

$$B_r^*=\arg\max_{S_r\subset B,|S_r|=r}JMH(P_1,P_2,\ldots,P_N|S_r)=\arg\max_{S_r\subset B,|S_r|=r}\sum_{i=1}^{N}KL(P_i,\bar{P}_i|S_r)\tag{16}$$

其中,$JMH(P_1,P_2,\ldots,P_N|S_r)$ 是式(13)在特征子集 S_r 下的 JMH 散度.

3.2 评价指标

在对文本进行分类时,通常用准确率[15],精确率和召回率来作为评估系统的指标,其定义如下.

$$\begin{cases}\text{准确率}(Accuracy)=\dfrac{TP+TN}{TP+TN+FN+FP}\\ \text{精确率}(\text{Pr}ecision)=\dfrac{TP}{TP+FP}\\ \text{召回率}(\text{Re}call)=\dfrac{TP}{TP+FN}\end{cases}\tag{17}$$

其中,TP 表示正确判定属于此类的文档数,FP 表示错误的判属此类的文档数,FN 表示错误判定不属于此类的文档数,TN 表示正确的判定不属于此类的文档数. 将精确率和召回率相结合构成了 F1 指标[16],更全面的对分类器的性能进行评价,定义如下.

$$F_1=\frac{2\times\text{Pr}ecision\times\text{Re}call}{\text{Pr}ecision+\text{Re}call}\tag{18}$$

4 仿真分析

4.1 数据集说明

本文采用了 REUTERS 中的 ModApte 数据集[17]. REUTERS 是世界前三大的多媒体新闻通讯社,提供各类新闻和金融数据. REUTERS 中的 ModApte 数据集有 65 个类由 8293 个文档构成,由于训练数据集较大,随机选择了 4994 个文档作为训练集,3299 个文档作为测试集,并选取数据集 REUTERS-10 中的前 10 个类.

4.2 仿真分析

为了验证本文所提算法的性能:基于朴素贝叶斯分类器,本文将 MD 算法与 cMFDR,CMFS 和 CSFS 三种算法进行对比. 用分类准确率和 F_1 指标来评估三种特征选择算法的性能,其中特征词数量从 10 个到 1000 个.

图 1 和图 2 分别代表了 MD 算法和 cMFDR、CMFS 和 CSFS 算法进行实验时其准确率和 F1 指标的曲线图. 由图 1 看出当特征词数量为 100 时,提出的 MD 算法的准确率接近最高为 93.5%,而 cMFDR 算法和 CSFS 算法当特征词数量为 1000 时准确率才接近 MD 算法. 由图 2 看出当特征词数量为 80 时,提出的 MD 算法的 F1 指标接近最大值为 88.35%,在特征词个数较少时 F1 指标大幅领先于其余三种算法.

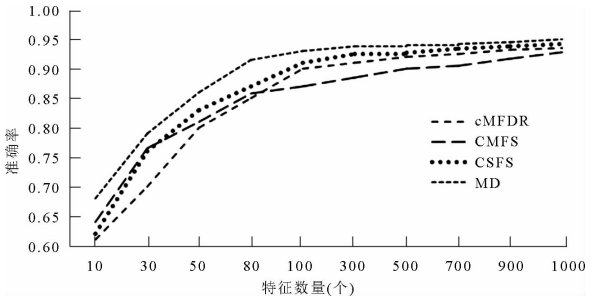

图 1 四种不同算法的准确率对比图

Fig. 1 Comparison of the accuracy of four different algorithms

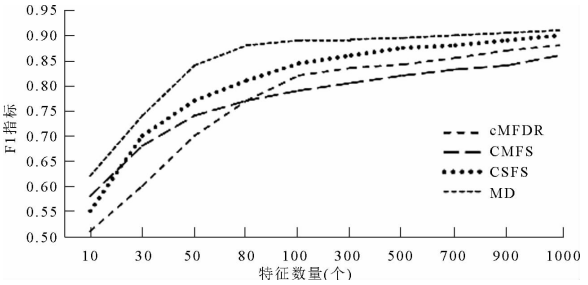

图 2　四种不同算法的 F_1 指标对比图

Fig. 2　Comparison of the F_1 of four different algorithms

从数据比较得知，MD 特征选择算法与其他三种特征选择算法相比，MD 特征选择算法在特征词数量相对少的时候，就可以获得较高的准确率和 F_1 指标，其他对比算法需要更多的特征词数量参考，才能达到相近的准确率和 F_1 指标，也就是说 MD 特征算法只需要选取少量特征词就能获得同其它算法一样的分类精度.

其中，CSFS 算法运用了主成份分析法，会把所有样本当作一个类别处理，寻找一个平方误差最小下的最优线性映射，而忽略了类别属性，忽略的投影方向可能恰好包含了有区分度的特征，而本文所提出的 MD 算法运用 KL 散度直接针对类别间的差异性，选择出使类别间差异最大的特征，所以 MD 特征选择算法的分类性能高于 CSFS 特征选择算法.

5　结　论

在对文本分类时，针对特征选择影响分类精度的问题，本文提出了一种对原始特征词进行排序的 MD 特征选择方法. 由训练得到 N 个类的概率分布后，通过分析测试样本中每个特征词区分类别的能力后构造出了一个新的特征向量 ε 用于分类. 仿真结果表明，对比 cMFDR，CMFS 和 CSFS 算法，由于该算法没选用对于类来说重要的特征词进行分类而是选用有最大类别区分能力的特征词进行分类，通过选取少量排名后的主要特征词就几乎达到最高的分类准确率和 F1 指标，使得朴素贝叶斯分类器对文本分类精度大幅提高. 下一步将所提出的特征选择方法与其他先进的机器学习算法相结合，以加强对罕见类别的学习.

参考文献：

[1]　陈波. 基于循环结构的卷积神经网络文本分类方法[J]. 重庆邮电大学学报：自然科学版，2018，30：705.

[2]　王岩，张波，薛博. 基于 FOA-SVM 的中文文本分类方法研究[J]. 四川大学学报：自然科学版，2016，53：759.

[3]　高云龙，左万利，王英，等. 基于集成神经网络的短文本分类模型[J]. 吉林大学学报：理学版，2018，56：933.

[4]　刘丹枫，刘建霞. 面向深度学习过拟合问题的神经网络模型[J]. 湘潭大学自然科学学报，2018，40：96.

[5]　Li J，Ozog P，Abernethy J，*et al*. Utilizing high-dimensional features for real-time robotic applications：Reducing the curse of dimensionality for recursive Bayesian estimation[C]//Proceedings of the 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Daejeon：IEEE，2016.

[6]　Fragoso R C P，Pinheiro R H W，Cavalcanti G D C. Class-dependent feature selection algorithm for text categorization[C]//Proceedings of the 2016 International Joint Conference on Neural Networks (IJCNN). Vancouver，BC：IEEE，2016.

[7]　Zhou X F，Guo L，Wang T Y. Text feature selection based on class subspace[C]//Proceedings of the 2014 IEEE International Conference on Data Mining Workshop. Shenzhen：IEEE，2014.

[8]　Feng L Z，Zuo W L，Wang Y W. Improved comprehensive measurement feature selection method for text categorization[C]//Proceedings of the 2015 International Conference on Network and Information Systems for Computers. Wuhan：IEEE，2015.

[9]　Kullback S. Information theory and statistics[M]. North Chelmsford，MA，USA：Courier Corporation，1997.

[10]　Cover T M，Thomas J A. Elements of information theory[M]. Hoboken，NJ，USA：Wiley，2012.

[11]　Legrand L，Grivel E，Giremus A. Jeffrey′s divergence between autoregressive moving-average processes[C]//Proceedings of the 2017 25th European Signal Processing Conference (EUSIPCO). Kos island，Greece：IEEE，2017.

[12]　Grivel E，Saleh M，Omar S. Comparing a complex-valued sinusoidal process with an autoregressive process using Jeffrey′s divergence[C]//Proceedings of the 2017 25th European Signal Processing Conference (EUSIPCO). Kos island，Greece：IEEE，2017.

[13]　Dong J X，Niu D B，Song E. An approach based on

Chernoff distance to sparse sensing for distributed detection [C] //Proceedings of the 2017 20th International Conference on Information Fusion (Fusion), Xi′an: IEEE, 2017.

[14] 陈齐根. 基于切尔诺夫界的泊松试验和的尾部概率估计及其应用 [J]. 重庆科技学院学报, 2013, 15: 156.

[15] 范大鹏, 张凤斌. 一种基于并行免疫网络的大数据分类算法 [J]. 江苏大学学报: 自然科学版, 2018, 39: 581.

[16] 阳馨, 蒋伟, 刘晓玲. 基于多类特征池化的文本分类算法 [J]. 四川大学学报: 自然科学版, 2017, 54: 287.

[17] Kull M, Flach P A. Machine learning and knowledge discovery in databases [J]. Lect Notes Comput Sci, 2014, 8725: 18.

引用本文格式:

中 文: 刘云, 黄荣乘. 最大判别特征选择算法在文本分类的优化研究[J]. 四川大学学报: 自然科学版, 2019, 56: 65.

英 文: Liu Y, Huang R C. Bayesian classifier-based maximum discriminant feature selection algorithm for text classification [J]. J Sichuan Univ: Nat Sci Ed, 2019, 56: 65.