

一种基于类别的混合型文本特征降维

刘海峰, 刘守生, 姚泽清

(解放军理工大学 理学院, 江苏 南京 210007)

摘 要: 提出了一种基于特征选择和特征抽取的混合型文本特征降维方法. 通过一种改进的优势率方法进行初次特征选择, 将文本表示为以类别属性为行向量的矩阵形式; 再使用一种改进的最大散度差特征抽取方法进行二次特征抽取. 在最大限度减少信息损失的前提下实现了文本特征的二次降维. 对中文文本的分类实验结果表明, 提出的特征降维方法具有良好的分类效果.

关键词: 文本分类; 特征选择; 特征抽取; 特征降维

中图分类号: TP391

文献标识码: A

文章编号: 1000-7180(2010)10-0013-05

A Mixed Text Reducing Feature Method Based on Sort

LIU Hai-feng, LIU Shou-sheng, YAO Ze-qing

(Institute of Sciences, PLA University of Science and Technology, Nanjing 210007, China)

Abstract: This paper presents a mixed method of reducing the text features which based on the feature selection and the feature extraction. Firstly, we carry out the first selection through improving on the odds ratio. The text is expressed with the matrix vector which composes of sort attributes. Then, we use an improved scatter difference to extract the features again. In the condition of the least information lost, we have completed the text feature reduction twice. The result which based on the Chinese text categorization shows that this method has a better precision in the text categorization.

Key words: text categorization; feature selection; feature extraction; feature reduction

1 引言

文本自动分类(Text Categorization, TC)是指在预先给定的类别标记集合下, 根据测试文本内容对其类属进行自动判定. 文本自动分类技术广泛应用于信息分类、信息过滤和信息检索等领域. 基于Web的信息检索的紧迫现实需求有力推动着文本自动分类技术的发展, 作为数据挖掘和机器学习领域的重要研究热点, 文本自动分类技术研究得到了学者的持续关注.

文本信息是以自然语言形式存在的非结构化数据. 使用计算机对其进行自动分类处理主要涉及文本的结构化表示、文本的特征降维、文本分类器设计等方面. 其中, 文本特征的高维性及文本向量的数据稀疏性问题是影响文本分类效率的主要瓶颈. 数百

篇的训练文本其特征维数也有上万维甚至数万维, 而文本分类的训练集通常是由数千篇文本组成, 高维的文本特征使得普通的计算平台难以承受. 因此, 文本特征降维的效率对文本自动分类的效果影响甚大, 文本特征降维技术成为文本自动分类的关键性技术之一^[1].

2 基于向量空间模型的文本标引模式的问题

将文本表示为由其特征项(通常以文本里的词或词组表示)组成的向量形式, 这种文本表示模型称为向量空间模型. 向量空间模型以其语义完整, 结构简单, 易于数学处理而得到了广泛的应用. 但是该模式也有其固有的缺陷, 影响了其对文本标引的效果.

收稿日期: 2009-08-18; 修回日期: 2009-10-18

基金项目: 国家自然科学基金项目(70571087)

2.1 文本的向量表示模式下存在的问题及解决途径

目前,向量空间模型是文本结构化表示的主流模型.另外还有基于概率分布、二维视图等模型.在向量空间模型下的相关技术研究主要集中在特征项的赋权方法上.此时训练集被表示成所谓的文本—特征项矩阵 $A = (w_{ij})_{m \times n}$ 形式.其中元素 w_{ij} 表示特征项 t_i 在文本 d_j 中的权重, $i = 1, 2, \dots, m; j = 1, 2, \dots, n$.而权重 w_{ij} 的赋权方式通常使用许多方法.但是无论使用何种方法,文本最终被表现为由诸特征项相应的权重构成的文本向量 $d_j = (w_{1j}, w_{2j}, \dots, w_{mj})$, 向量的各个分量 w_{ij} 的取值是基于整个训练集而得到.因此这种向量形式的文本表示没有利用(比如基于 TF-IDF 的特征赋权)或没有充分利用(比如基于 IG、 χ^2 统计等方法)文本的类别信息.这一不足之处影响了向量空间模型对文本的标引能力.

我们将文本 d 的向量表示 $d = (w_1, w_2, \dots, w_m)$ 改变为用矩阵形式进行表示:

$$d = \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1s} \\ w_{21} & w_{22} & \cdots & w_{2s} \\ \vdots & \vdots & & \vdots \\ w_{r1} & w_{r2} & \cdots & w_{rs} \end{bmatrix}$$

式中, w_{ij} 表示特征项集合在第 i 个类别里权重按照由大到小顺序的第 j 个特征项 t_j 在文本 d 里的权重,而 d 中的行向量 $(w_{k1}, w_{k2}, \dots, w_{ks})$ 则表示第 k 个类别 c_k 对于文本 d 在标引方面的重要程度. s 表示每个类别所选取的用于表示文本类别向量的特征项个数, r 表示文本的类别数, $i = 1, 2, \dots, r; j = 1, 2, \dots, s$.

这种以文本在各个类别下的类别属性度量为行向量的表示形式,不仅考虑了文本特征在训练集上的重要程度,同时考虑了文本在各个类别上的类别属性,体现了文本在不同类别之间标引上的差异.因此,这种以矩阵形式表示文本提高了特征项对文本的表示能力.

下面考虑在各个类别内使用合适的特征选择方法,从特征集 $T = \{t_1, t_1, \dots, t_m\}$ 里选择出相应的诸类别特征用于文本的矩阵表示,同时达到文本特征的初次降维目的.

2.2 一种基于改进优势率的文本矩阵表示

优势率(Odds Ratio, OR)方法度量的是某一特征项 t_k 作为类别 c_i 的类内特征相比较作为其他类别的特征时性能上的差异.研究结果表明,在样本分布

相对均匀条件下在几种常用的特征选择方法中,优势率方法的性能略为占有^[2].常见的基于优势率的计算公式如下:

$$\begin{aligned} OR(t_k, c_i) &= \log \frac{p(t_k | c_i)[1 - p(t_k | \bar{c}_i)]}{p(t_k | \bar{c}_i)[1 - p(t_k | c_i)]} \\ &= \log \frac{N(t_k, c_i) \times N(\bar{t}_k, \bar{c}_i)}{N(\bar{t}_k, c_i) \times N(t_k, \bar{c}_i)} \quad (1) \end{aligned}$$

式中, c_i 表示第 i 类训练文本, $p(t_k | c_i)$ 表示类 c_i 内特征 t_k 出现的概率, $p(t_k | \bar{c}_i)$ 表示在其他类别里特征 t_k 出现的概率, $N(t_k, c_i)$ 为含有特征 t_k 的 c_i 类的文本数, $N(\bar{t}_k, \bar{c}_i)$ 为非 c_i 类且不含有特征 t_k 的文本数, $N(\bar{t}_k, c_i)$ 为非 c_i 类且含有特征 t_k 的文本数, $N(t_k, \bar{c}_i)$ 为 c_i 类不含有特征 t_k 的文本数.

式(1)体现了含有特征项 t_k 的 c_i 类文本对于所有非 c_i 类的文本来说在类别判定上的优势.两者的比值越大,越能体现 t_k 相对于非 c_i 类别的特征来说更适合于作为 c_i 类别的特征.

式(1)的赋权模式当不同类别的特征项的交集接近空集时效率较高.即不同类别的文本的特征也互不同时,效果较好.进一步分析式(1)可以看出:特征 t_k 相对于确定类别 c_i 的优势率不仅与其在 c_i 类里出现的次数有关(越多越好),还与其在非 c_i 类文本里出现的次数相关(越少越好).这就造成了那些属于多类别的特征项其优势率不高.而式(1)中分母 $N(\bar{t}_k, \bar{c}_i) \times N(\bar{t}_k, c_i)$ 既要求在非 c_i 类内含有特征 t_k 的文本数越少越好,同时要求在 c_i 类内不含有特征 t_k 的文本数越少越好.这两个要求的“合力”放大了“非 c_i 类内含有特征 t_k 的文本数越少越好”这一要求对特征选择的影响力度,过于强调了“非 c_i 类文本不含有特征 t_k ”对优势率的作用,从而抑制了 t_k 对 c_i 类文本的分类作用.基于此因素考量,对式(1)进行如下的改进:

$$OR(t_k, c_i) = \log \frac{N(t_k, c_i) \times N(\bar{t}_k, \bar{c}_i)}{N(\bar{t}_k, \bar{c}_i)} \quad (2)$$

即减少非 c_i 类文本不含有特征 t_k 这一要求对特征选择的干扰.

使用式(2)文本类别特征的选择,用以构造类别特征向量.

综上所述,提出一种基于改进优势率的文本矩阵表示方法如下:

(1) 首先对训练集进行预处理,得到特征项集合 $T = (t_1, t_1, \dots, t_m)$;

(2) 通过式(2)计算特征项在各个类别里的优势率 $OR(t_k, c_i)$;

(3) 对于 c_i 类文本,选择前 s 个(这里 $s \ll m$) 优势率值最大的特征项用于表示文本 d 在 c_i 类里的权重,这里 $i = 1, 2, \dots, r$;

(4) 将文本 d 表示成矩阵形式:

$$d = (w_{ij})_{r \times s}$$

式中, w_{ij} 表示在第 i 个类别里权重值由大到小顺序的第 j 个特征项 t_j 在文本 d 里的权重,其赋权方式采用常用的 TF-IDF 方式计算, $i = 1, 2, \dots, r; j = 1, 2, \dots, s$.

(5) 两文本 d_k, d_l 之间的相似度使用矩阵的 2-范数进行度量:

$$\text{sim}(d_k, d_l) = \frac{1}{(\sum_{i,j} |w_{kij} - w_{lij}|^2)^{\frac{1}{2}}}, \quad (3)$$

此时,训练集被表示为 $R^{r \times s}$ 矩阵空间,每个文本均被表示为一个相应的矩阵.这种基于选择的特征降维方法将文本特征维数由 m 维降低为 $r \times s$ 维.由于在实际分类时 r 很小,而 $s \ll m$,所以使用矩阵表示文本不仅提高了文本的标引能力,特征维数也得到有效的降低.

3 基于改进的最大散度差鉴别准则的二次特征降维

借助矩阵进行文本表示,使用改进的优势率赋权方法进行了文本类别特征选择,实现了文本特征的有效降维.但是,由于文本矩阵的行数即类别向量的维数是使用式(2)对 m 个特征进行选择得到的前 s 个特征项,一般来说训练集的特征项个数 m 通常很大(常常高达上万维),而选择出来的特征项个数 s 相比较 m 来说要小得多.因此对于每个类别向量来说,均是以丢失其余 $m - s$ 个特征项含有的信息为代价达到维数压缩的目的.为了减少信息丢失过多,这种形式的特征选择不应该使得 s 过于小(比如 $s < 20$).所以我们在进行特征选择时,实验中取 s 约为 m 的二十分之一,然后使用特征抽取方法进一步对特征维数进行压缩,在尽量减少信息缺失原则下达到文本特征的有效降维.

3.1 基于最大散度差准则的特征抽取及其特点

作为文本特征降维的另一种主要模式,特征抽取方法是通过映射把测量空间的数据投影到特征空间,获得原始特征集相应的像集,进而对像集进行评估.特征抽取本质上是完成测量空间到特征空间的变换.作为经典的线性特征抽取方法,Fisher 鉴别准则^[3]被公认为是最好的线性特征抽取方法.也使得

基于 Fisher 准则所发展起来的各种特征降维方法研究得到逐步深入.由于 Fisher 线性鉴别方法在使用时经常会遇到小样本问题,即类内散布矩阵 S_w 的奇异性问题,此时模型无法使用.人们常常采用的是“近似最优”的处理方法:或者对 S_w 进行一个小的扰动使其成为非奇异阵,以扰动后的矩阵代替原来的类内散布矩阵进行投影向量的求解;或者使用“零空间”方法^[4]在 S_w 的零空间内寻找极大化类间散布矩阵 S_b 的一组标准正交投影向量.但是这种方法是丢失 S_w 零空间以外的所有分类信息为代价.上述处理方法本质上讲都属于近似解法.针对这一问题,Song^[5]提出了最大散度差线性鉴别准则:

$$J(W) = w^T(S_b - cS_w)w, \quad |w| = 1 \quad (4)$$

式中, c 为参数,用以调节类内散度相对于类间散度在决策方面的重要程度.其思想是寻找一个最佳鉴别矢量,使得高维模式样本沿着该方向投影后类间散度和类内散度之差最大.这种方法解决了 S_w 奇异性问题对 Fisher 准则的困扰,同时将矩阵求逆计算转化为计算散度差,计算复杂度得以降低.

在原始投影数据为文本矩阵的条件下,最大散度差鉴别准则简述如下:

设 $c_i (i = 1, 2, \dots, r)$ 表示训练文本类别, c_i 内的文本数为 $n_i, i = 1, 2, \dots, r; N = \sum_{k=1}^r n_i$ 为训练文本数,记 $d_{k_i} = (w_{k_{ij}})_{r \times s}$ 表示第 i 类的第 k 个文本, $l = 1, 2, \dots, r; k = 1, 2, \dots, n_l; \bar{d}_{k_i} = \frac{1}{n_i} \sum_{d_{k_i} \in C_i} d_{k_i}$ 为类 c_i 内样本均值, $i = 1, 2, \dots, r; \bar{d}$ 为所有训练样本均值;训练样本的类内散布矩阵 S_w 、类间散布矩阵 S_b 分别记为

$$S_w = \frac{1}{N} \sum_{i=1}^r \sum_{d_{k_i} \in C_i} (d_{k_i} - \bar{d}_{k_i})^T (d_{k_i} - \bar{d}_{k_i}),$$

$$S_b = \sum_{i=1}^r \frac{n_i}{N} (\bar{d}_{k_i} - \bar{d})^T (\bar{d}_{k_i} - \bar{d}) \quad (5)$$

基于最大散度差的鉴别准则为

$$J(W) = \text{tr}(W^T(S_b - cS_w)W),$$

$$(w_i, w_j) = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases} \quad (6)$$

其思想是寻找一组正交的单位投影向量 W 使得式(6)达到最大.而 $W = (w_1, w_2, \dots, w_t)_{s \times t}$ 由矩阵 $S_b - cS_w$ 最大的前 t 个特征值相应的单位正交特征向量组成^[6],此时准则式(6)可变形为

$$J(W) = \sum_{i=1}^t w_i^T (S_b - cS_w) w_i \quad (7)$$

而文本矩阵 $\mathbf{d} = (w_{ij})_{r \times s}$ 通过投影变换:

$$\tilde{\mathbf{d}}_{r \times t} = \mathbf{d}_{r \times s} \times \mathbf{W}_{s \times t} \quad (8)$$

转化为 $r \times t$ 阶矩阵, 由于 $t \ll s$, 因此在基于式(2)的特征选择的基础上, 文本矩阵维数进一步由 $r \times s$ 降为 $r \times t$, 达到了二次特征降维目的.

最大散度差准则函数的物理意义是投影变换后特征空间的样本具有最大类间散度和最小类内散度. 它解决了 Fisher 鉴别准则面临的类内散布矩阵的奇异性问题, 同时在最大散度差模式下不再需要求解逆矩阵, 从而算法的速度有了很大的提高.

最大散度差方法可以很好分开类间距离较大的类别, 但是在类间距离较小的类别上分类上效果欠佳. 针对这一问题, 可以考虑对类间矩阵进行加权, 目的是将类间距离较小的类别拉开, 使得投影后的相集具有更好的可分性.

3.2 一种加权的最大散度差特征抽取方法

分析式(7)可以看出, 参数 c 能够平衡类内散度相对于类间散度的比重, 现在问题主要是类间散度

$\mathbf{S}_b = \sum_{i=1}^r \frac{n_i}{N} (\bar{\mathbf{d}}_{k_i} - \bar{\mathbf{d}})^T (\bar{\mathbf{d}}_{k_i} - \bar{\mathbf{d}})$ 对间距很小的类别之间的区分能力较差. 为此, 考虑使用径向基函数^[7]为加权手段调节诸类别训练样本相应的协方差值, 以达到将间距较小的类别拉开的目的. 即将式(5)改进为

$$\tilde{\mathbf{S}}_b = \sum_{i=1}^r \frac{n_i}{N} \times \exp\{-\|(\bar{\mathbf{d}}_{k_i} - \bar{\mathbf{d}})\|_2^2 / \sigma^2\} \times (\bar{\mathbf{d}}_{k_i} - \bar{\mathbf{d}})^T (\bar{\mathbf{d}}_{k_i} - \bar{\mathbf{d}}) \quad (9)$$

而最大散度差准则相应改进为

$$\mathbf{J}(\mathbf{W}) = \sum_{i=1}^t w_i^T (\tilde{\mathbf{S}}_b - c\mathbf{S}_w) w_i \quad (10)$$

这里 $\mathbf{W} = (w_1, w_2, \dots, w_t)_{s \times t}$ 为矩阵 $\tilde{\mathbf{S}}_b - c\mathbf{S}_w$ 最大的前 t 个特征值相应的单位正交特征向量组成.

此时, 通过式(9)中的参数 σ 的不同取值以调节类间距离, 以便找到最优的投影矩阵. 其物理意义是将各个类别样本的中心均向着总体样本中心的方向拉到以 $\bar{\mathbf{d}}$ 为球心的高维单位球的球面上. 诸类别中心向量在单位球面上的分布不均匀, 由于径向基函数随着矩阵 2-范数的增大而递减, 所以通过式(9)可以将间距较小的类别赋以较大的权重以拉大间距, 而间距较大的类别之间的距离缩小, 从而达到调节训练样本的类间距的目的.

4 一种混合型的文本特征二次降维方法

综上所述, 提出的一种混合型文本特征二次降

维算法表述如下:

(1) 对训练文本集进行预处理, 得到原始特征集 $T = \{t_1, t_1, \dots, t_m\}$;

(2) 通过式(2)计算特征集 $T = \{t_1, t_1, \dots, t_m\}$ 里各特征项在诸类别里的优势率 $OR(t_k, c_i), i = 1, 2, \dots, r; k = 1, 2, \dots, m$;

(3) 对于 c_i 类文本, 选择前 s 个优势率值最大的特征项构成 c_i 类的类别特征集, 相应的类别向量为 $(w_{1i}, w_{2i}, \dots, w_{si})$, 表示文本 $\mathbf{d}$ 在 c_i 类里相应的属性, $i = 1, 2, \dots, r; w_{ki}$ 赋权方式采用 TF-IDF 公式计算, $k = 1, 2, \dots, s$, 本文实验中取 $s = 270$;

(4) 文本 $\mathbf{d}$ 表示成矩阵形式: $\mathbf{d} = (w_{ij})_{r \times s}$

(5) 使用式(10)进行特征抽取, 取矩阵 $\tilde{\mathbf{S}}_b - c\mathbf{S}_w$ 最大的前 t 个特征值相应的单位正交特征向量组成投影矩阵 $\mathbf{W} = (w_1, w_2, \dots, w_t)_{s \times t}$, 本文实验中取 $s = 270, c = 0.7, \sigma^2 = 5$;

(6) 使用式(8)对由步骤(4)得到的文本矩阵 $\mathbf{d} = (w_{ij})_{r \times s}$ 进行投影变换, 将其维数进一步压缩为 $r \times t$ 维, 本文试验中取 $t = 75$;

(7) 两文本 d_k, d_l 之间的相似度使用式(3)进行度量;

(8) 分类器使用常用的 KNN 分类器^[8], 其中取参数 $k = 17$.

5 实验及其结果分析

对上述特征降维方法的分类效果进行了试验. 试验数据为从新浪、百度下载的 4 010 篇文本, 其中分为经济(420 篇)、环境(426 篇)、体育(729 篇)、计算机(585 篇)、文艺(610 篇)、农业(458 篇)以及军事(782 篇)共 7 类. 试验时采用 3 折交叉试验法, 将 4 010 篇文本分为 3 份, 2 份为训练集, 1 份为测试集; 每份轮流作为测试集循环测试共 3 次, 取平均值为测试结果. 使用东北大学自然语言处理实验室提供的分词软件进行分词, 使用禁用词表过滤停用词、人工删除冷僻的特低频词、并剔除了虚词、助词、人称代词、特高频词等预处理后建立候选特征项集得到特征个数 $m = 5\,462$ 个. 分类效果评估指标使用常用的查准率、查全率以及 F_1 测试值:

查准率 = 分类的正确文本数 / 实际分类文本数;

查全率 = 分类的正确文本数 / 应有文本数;

$F_1 = (\text{查准率} \times \text{查全率} \times 2) / (\text{查准率} + \text{查全率})$

试验数据结果统计如表 1 所示.

表 1 文本分类试验结果统计								
类 别	经济	军事	体育	计算机	文艺	农业	环境	平均
查准率	0.827 3	0.858 2	0.882 1	0.921 7	0.893 6	0.928 5	0.837 9	0.878 5
查全率	0.867 6	0.907 8	0.913 6	0.897 6	0.837 1	0.883 7	0.879 2	0.883 8
F_1 值	0.847 0	0.882 3	0.897 6	0.909 5	0.864 4	0.905 5	0.858 1	0.880 6

从表 1 可以看出,本文提出的混合型特征降维算法在体育、军事、计算机与农业类别的文本分类效果较好,经济与环境两类准确率略低一点,由于这两类文本内容相对接近一点,其共用词组相对多一些,文本数也略少一些.平均查准率、查全率与 F_1 测度均在 0.88 左右,特征降维的效果还是令人满意的^[9].

6 结束语

特征选择和特征抽取作为文本特征降维的两种主要模式在文本分类中已经得到了广泛的应用.如何有效地利用两类模型的优点,使之相互补充而提高特征降维的效率,考虑各种方法的组合型、混合型等方式的使用,同时对具体模型进行必要的改进使其更好地表示文本,本文的研究对于文本挖掘领域的研究是有益的.这些工作的研究还有待深入,还需继续进行.

参考文献:

[1] Yang Y, Pedersen J O. A comparative study on feature selection in text categorization[C]//Proceedings of ICML-97, 14th International Conference on Machine Learning. US, Nashville, 1997:412-420.

[2] 刘海峰,姚泽清,汪泽焱,等.基于位置的文本特征加权方法研究[J].微电子学与计算机,2009,26(2):188-192.

[3] 杨建,杨静宇,叶辉. Fisher 线性鉴别分析的理论研究及其应用[J].自动化学报,2003,36(4):563-566.

[4] 刘海峰,姚泽清,刘守生,等.文本分类中基于核的非线性判别[J].应用科学学报,2008,26(6):627-631.

[5] Hong Z Q, Yang J Y. Optimal discriminant plane for a small number of samples and design method of classifier on the plane[J]. Pattern Recognition, 1991, 24(4): 317-324.

[6] Chen Li Fen, Liao H Y Mark, Ko M T, et al. A new LDA-based face recognition system which can solve the small sample size problem[J]. Pattern Recognition, 2000, 33(10):1713-1726.

[7] 宋枫溪,程科,杨静宇,等.最大散度差和大间距线性投影与支持向量机[J].自动化学报,2004,30(6):890-896.

[8] 刘海峰,姚泽清,王元元,等.一种基于散度差组合型文本特征降维方法[J].微电子学与计算机,2008,25(8):69-75.

[9] 王建国,林宇生,刘锁兰,等.改进的广义主分量分析及在人脸识别中的应用[J].计算机工程与设计,2008,29(11):2901-2905.

作者简介:

刘海峰 男,(1962-),博士,副教授,CCF 会员.研究方向数据挖掘、文本分类.

刘守生 男,(1965-),博士,副教授.研究方向神经网络.

姚泽清 男,(1960-),教授.研究方向统计分析、数据挖掘.