

基于灰关系与信息增益融合的文本分类模型

刘海峰¹, 李凯齐², 汪泽焱¹

(1 解放军理工大学 理学院, 江苏 南京 210007;

2 解放军理工大学 指挥自动化学院, 江苏 南京 210007)

摘 要: 针对信息增益模型在文本分类中的不足之处, 提出了一种基于灰关系与信息增益的文本分类算法. 首先基于改进的 χ^2 统计进行类别特征选择用于类内文本表示, 提高类别中心向量的表示能力; 其次针对 IG 模型对低频词赋权过大问题, 提出了基于频数和位置的改进加权方法; 最后提出了基于灰关系的文本相似度计算途径, 改善了基于距离的相似度计算模式的不足. 试验表明, 此算法提高了文本分类效率.

关键词: 信息增益; 灰色理论; χ^2 统计; 位置加权; 文本分类

中图分类号: TP391

文献标识码: A

文章编号: 1000-7180(2012)10-0094-05

Text Categorization Model Based on Fusing of Gray Theory and Information Gain

LIU Hai-feng¹, LI Kai-qi², WANG Ze-yan¹

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

2 Institute of Command Automation, PLA University of Science and Technology, Nanjing 210007, China)

Abstract: In view of the information gain model defects in the text classification, this article puts forward a text classification algorithm based on the grey relation and information gain. Firstly, we improved a method of χ^2 statistics in sort feature selection in order to express text. In this way, we improve the precision of the class center vector. Secondly, according to the IG model weights the low frequency words too bigger, we put forward an improved weighted method basing on frequency and position. Lastly, we put forward a new way in text similarity calculation in order to improve the shortcomings of the similarity calculation model that based on distance. Subsequent text categorization test shows that this paper puts forward an improved IG method and enhances the text classification efficiency.

Key words: information gain; gray theory; χ^2 statistics; position weighting; text categorization

1 引言

文本分类(text categorization, TC)是指在给定的分类体系下,根据文本内容自动确定文本所属类别的过程.作为文本挖掘领域的重要研究内容,文本分类技术在提高信息利用的实时性、有效性和准确性上具有重要的现实意义.基于统计的机器学习方法在文本分类中的应用研究形成了文本自动分类技术研究的主流方向.文本信息是以自然语言形式

为主体存在的一种半结构化或非结构化信息载体,因此对文本进行合理的结构化表示是基于统计的文本挖掘方法的前提和基础.目前,文本结构化表示方法主要有向量空间模型及其变种、基于特征项的概率分布模型、二维视图模型等.其中向量空间模型以其概念简单、应用方便等优点,成为最有效的文本表示模型之一,是目前文本结构化表示的主流模式^[1].因此,在以向量形式表示文本的模式下,提高文本自动分类效率的关键在于两个方面^[2]:

收稿日期: 2012-01-18; 修回日期: 2012-04-05

基金项目: 国家自然科学基金资助项目(71071161)

(1)有效的特征降维以改善文本向量的稀疏性问题,避免维数灾难现象;

(2)合理的文本特征赋权方法以提高特征项对文本的表示能力.

2 一种改进的 IG 特征选择方法

2.1 IG 特征选择方法及其特点

文本特征降维主要分为特征抽取和特征选择两种模式.特征选择是借助某个评估函数对原始特征项逐一评分,根据分值大小选取若干个分值最高的特征项作为原始特征集的一个子集用于文本表示,其本质上是对特征集的一种简约.基于选择的特征降维方法以其语义完整、易于理解、算法复杂度较低而得到了广泛的使用.常用的特征选择方法有比较经典的如 TF-IDF、 χ^2 统计、互信息、信息增益等,其中信息增益和 χ^2 统计的效果最好^[3].

信息增益(Information Gain,IG)是一种基于熵的评估方法.与其他常用的特征选择方法比如 TF-IDF 方法相比,它不仅考虑了类内出现的特征项对文本标引的力度,同时也考虑了那些在类内未出现的特征项对文本分类的影响,度量了特征项在文本里出现前后信息熵的改变量,体现的是某一特征项的存在与否对类别预测的影响,是目前效果比较好的特征选择方法.IG 计算的是特征项 t_k 出现前训练集 T 的信息熵 $H(T)$ 与 t_k 出现后的条件信息熵 $H(T|t_k)$ 之间信息的增量:

$$\begin{aligned} IG(t_k) &= H(T) - H(T|t_k) = \\ &= - \sum_{i=1}^r P(c_i) \log_2 P(c_i) + \\ &= P(t_k) \sum_{i=1}^r P(c_i|t_k) \log_2 P(c_i|t_k) + \\ &= P(\bar{t}_k) \sum_{i=1}^r P(c_i|\bar{t}_k) \log_2 P(c_i|\bar{t}_k) \end{aligned} \quad (1)$$

这里 t_k 表示第 k 个特征项, r 为文本类别数, $p(c_i)$ 表示 c_i 类文本出现的概率, $P(t_k)$ 表示 t_k 在训练集里出现的概率, $P(\bar{t}_k)$ 表示 t_k 在训练集里没有出现的概率.从 n 个特征项里选择信息增益最大的前 p 个特征项构成文本特征向量.

作为信息论在文本信息处理方面应用效率较高的特征选择方法之一,IG 模型不仅考虑了类内出现的特征项对文本表示能力,同时也考虑了那些在类内未出现的特征对文本分类的影响力度.IG 的特点是便于理解,应用方便,效率较高.但是该方法也存在一些问题.

2.2 IG 方法的不足与改进思路

IG 模型有其特点,但是其缺陷也是明显的.第一,待分类文本 d 的类别判定是以 d 与各个类别 c_i 的类别中心向量

$$d_{c_i} = \frac{1}{|c_i|} \sum_{d_j \in c_i} d_j \quad (2)$$

之间相似度大小为标准.基于式(1)的特征项评估是在整个训练集上进行的,特征选择后所有类别文本共用一个特征集表示,因此这种文本表示方式就不能充分体现出不同类别 c_i 与 c_j 之间的类别特征需求方面的差异,因为式(1)的特征选择方式不是针对各个类别文本自身需求选择出适用于各自类别表示的特征集.这一问题造成 c_i 类的类别中心向量 $d_{c_i} = \frac{1}{|c_i|} \sum_{d_j \in c_i} d_j (j=1,2,\dots,r)$ 不能很好地体现 c_i 的类别中心位置.第二,IG 模型考虑了类内未出现的特征项对分类影响.事实上,类内未出现的特征可能对文本类别判定方面具有一定贡献,但这种贡献的力度往往小于其对类别判定上所带来的干扰力度^[4].特别是在训练集的分类分布和特征分布高度不平衡的情况下,绝大多数类都是负类而且绝大多数特征都不出现,此时式(1)的函数值主要是由不出现的特征所决定,此时信息增益的效果就会大大降低.第三,低频词的作用被放大.对于多数低频词特别是在训练集里分布相对均匀的低频词,式(1)中 $P(\bar{t}_k)$ 很大,相比较于不考虑类内未出现的特征项的赋权模式,IG 模式下低频词所对应的权值变大,其结果是将一些对分类作用意义不大的低频词选为文本特征.IG 方法造成了低频词的作用被放大的同时,相应的大多数中频词的信息量却利用不足.而在使用停用词表删除特高频词和特低频词,对文本进行预处理后的特征项集合里具有最好分辨力的词恰好是中频词^[5],因此解决 IG 模型对低频词权重放大、对中频词权重的抑制现象可以改善其特征选择效率.

针对上述第一个问题,考虑到 χ^2 统计模型的特点,本文分两步进行特征选择:先使用改进的 χ^2 统计方法对各个类别的文本分别选择相应的最优特征集用于类内文本表示、进而用于类别中心的表示,再使用 IG 模型进行特征选择用于训练集的文本表示.但是考虑 IG 模型存在的第二和第三个问题,本文通过引入修正因子对 IG 模型进行了改进,实现对低频词权重的适度抑制,同时增大高、中频词的文本表示的能力,提高 IG 模型对文本的整体表示能力.

2.3 基于改进 χ^2 统计的类别特征选择

2.3.1 基于 χ^2 统计的类别特征选择特点

χ^2 统计是一种效果较好的特征选择方法^[3],能较好地体现特征项与类别之间的相关信息,反映出不同特征对类别判定方面信息含量的大小. 特征项 t_k 和 c_i 类文本之间的相关性由下式度量:

$$\chi^2(t_k, c_i) = \frac{m(AD - CB)^2}{(A + C)(B + D)(A + B)(C + D)}, \quad i = 1, 2, \dots, r \quad (3)$$

这里 A 表示属于 c_i 且包含 t_k 的文本数, B 表示不属于 c_i 且包含 t_k 的文本数, C 表示属于 c_i 不包含 t_k 的文本数, D 表示不属于 c_i 且不包含 t_k 的文本数, m 为训练集文本数.

用 χ^2 统计方法来进行特征抽取是基于如下假设:在指定类别文本中出现频率高的词条与在其他类别文本中出现频率比较高的词条,对判定文本是否属于该类别都是很有帮助的. χ^2 统计量体现了特征项 t_k 与类别 c_i 之间的相关程度. 该模式认为 t_k 与 c_i 之间符合自由度为 1 的 χ^2 分布 $\chi^2(1)$. χ^2 统计量综合考虑了特征项 t_k 和类别 c_i 内的文本同现情况与不同现况下 t_k 在类别 c_i 的文本中的重要程度. χ^2 值越大,说明 t_k 对属于 c_i 中的文本进行定类的作用越大. 该方法在同类别文本的类内特征选择方面优点突出:那些是在 c_i 内频数较大(以文本数记),但是在整个训练集内出现的频数相对较低的特征项对于该类别文本的表现能力强,这些特征项多数可以作为 c_i 的类别特征. 而分析式(2)可以发现,这些特征项在式(2)中对应的 A 、 D 值较大而 B 、 C 值较小, $AD \gg BC$, 从而式(2)值较大,在 χ^2 统计模式下这些特征项多数会被选择为类 c_i 的文本特征用于 c_i 类的文本表示. 这也是我们为什么使用 χ^2 统计进行类别特征选择的原因.

2.3.2 基于改进的 χ^2 统计的类别特征选择

但是 χ^2 统计模型也有其不足之处. 主要的问题之一就是降低了低频词的权重^[6]. 其实一些特定类别特别是稀有类别的低频词对类别标引能力强,但是由于式(2)的对应值过小而被删除. 针对这一问题,考虑对式(2)引入权重因子进行修正.

记 $tf_i(t_k)$ 表示在类 c_i 内特征项 t_k 出现的频数(以文本数记),则

$$\lambda_{ik} = \frac{\max_{1 \leq j \leq n} \{tf_i(t_j)\} - \min_{1 \leq j \leq n} \{tf_i(t_j)\} + 1}{tf_i(t_k) - \min_{1 \leq j \leq n} \{tf_i(t_j)\} + 1}$$

为 $tf_i(t_k)$ 的递减函数,这里 n 为特征数. 进而

$$\chi_{\text{new}}^2(t_k, c_i) = \lambda_{ik} \frac{n(AD - CB)^2}{(A + C)(B + D)(A + B)(C + D)} \quad (4)$$

随着 $tf_i(t_k)$ 的减小而增大. 这种加权可以改善 χ^2 统计对低频词的过抑制现象. 借助式(4)进行类别 c_i 的类别特征选择,是基于特征与类别相关性角度的度量,而 χ^2 统计模型的特点是以特征项与具体类别相关性度量为目的,对同类别的文本的表示能力更强.

在解决了文本类别特征选择问题后,我们仍然使用信息增益方法进行文本特征选择,但是针对常用的信息增益方法的缺点也进行了改进,以提高其文本标引能力.

3 一种改进的信息增益特征选择模型

针对信息增益模型的第二、第三问题,考虑从两个方面入手引入修正因子对 IG 模型进行改进,实现对低频词权重的适度抑制,改善中频词对文本的表示能力.

3.1 基于特征频数的信息增益模型改进

χ^2 统计是基于含有(或不含有)特征项的文本数量即文本的频数为依据对特征项在文本类别判定上的重要程度的分析,而信息增益的统计对象是特征项的频数即词频. 因此在以词频为统计对象的前提下,对低频词的作用的量化方式与 χ^2 统计的思路不同. 一般说来,在类别分布相对平均的条件下如果 t_k 频数越大且分布的越平均,说明其对文本的标引作用越大,应该调高其权重;反之如果特征频数较小,应该调低其权重. 但是 IG 模型却放大了低频词的作用. 基于这方面考虑,记 $tf_i(t_k)$ 表示特征项 t_k 在类 c_i 内频数, $i = 1, 2, \dots, r$.

引入权重修正因子

$$\mu(t_k) = \frac{\sum_{i=1}^r tf_i(t_k)}{\sqrt{\sum_{k=1}^n \left(\sum_{i=1}^r tf_i(t_k) \right)^2}}, k = 1, 2, \dots, n \quad (5)$$

将式(1)改进为

$$IG_{\text{improve}}(t_k) = \mu(t_k) \times IG(t_k), \quad k = 1, 2, \dots, n \quad (6)$$

显然,在训练集里频数高的特征项的修正因子值大于频数低的特征项的修正因子值,因此以式(5)作为特征项 t_k 的权重可以提高 IG 模型下高、中频词的权重,达到降低低频词权重的目的.

3.2 基于特征位置的 IG 模型进一步改进

同一个词由于其在文本里的位置不同,它对于文本的表示能力也会存在着较大的差异.在文本的不同位置的特征项对文本表达能力从大到小顺序为:标题>摘要>关键词>副标题>段首句>其他^[7],基于此考虑,我们根据文本内容所在的位置将训练集分割为三个“伪训练集” $S_l, l=1,2,3$,其中 S_1 由训练集的标题、摘要、关键词组成, S_2 由训练集的副标题和段首句组成, S_3 由训练集的除去前两个集合外的其他部分组成.在进行特征项文本赋权时,考虑分两步计算特征项 t_k 的权重 $MI(t_k)$:

首先根据式(1)在伪训练集 $S_l, l=1,2,3$ 内分别计算 t_k 相应的 IG 值,并将其记为 $IG(t_k)_l, l=1,2,3$;而基于位置因素度量的 t_k 对应的 IG 值为

$$IG_{\text{improve}}(t_k) = \sum_{l=1}^3 \theta_l IG(t_k)_l \quad (7)$$

考虑到词频因素,结合式(6),得到 t_k 的 IG 最终权重为

$$\overline{IG}_{\text{improve}}(t_k) = \mu(t_k) \times IG_{\text{improve}}(t_k) = \mu(t_k) \times \sum_{l=1}^3 \theta_l IG(t_k)_l \quad (8)$$

其中权重修正因子 $\mu(t_k)$ 由式(5)计算,位置加权系数 θ_l 根据文献[7]的试验结果,本文中取为: $\theta_1=0.5, \theta_2=0.3, \theta_3=0.2$.

这样利用式(5)实现了文本特征集的权重计算,选择权重值最大的前 p 个特征构成文本特征集,实现基于选择的文本特征降维.

4 基于灰关系的文本相似度计算

在向量模型下,待分类文本 d 与类别中心向量 $d_{c_i} = \frac{1}{|c_i|} \sum_{d_i \in c_i} d_i$ 之间的相似度计算大多基于欧氏距离或者向量的夹角余弦.但是不同特征项在文本表示方面的作用比较复杂,不是简单的线性关系就能描述,因而通常文本之间的基于欧式距离的相似度计算模式不能很好地解决这个问题.另外,同一个特征项在不同类别之间的相异性信息也没有得到利用.而灰色系统理论中的灰关联系数的计算方式却能够改善这方面的不足.

灰色系统研究的是部分信息明确、部分信息未知的小样本、贫信息的不确定性系统,它通过对已知的部分信息的处理去开发、实现对现实世界的确切描述和认识.灰色系统理论对多因素、非线性问题的处理效果较好^[8].灰色关联系统的关联度分析方法

能够根据问题中各个因素之间的相似或相异程度来衡量诸因素之间的关联程度,这种方式可以从一个侧面体现同一个特征项在不同文本之间的相关性信息.

完成特征降维后,假定 $d=(t_1, t_2, \dots, t_p)^T$ 为待分类文本, $d_{c_i}=(t_{c_i,1}, t_{c_i,2}, \dots, t_{c_i,p})^T, i=1,2, \dots, r$ 为 c_i 的类别中心向量,利用灰色理论中的灰关系分析^[8]方法, d 和 d_{c_i} 之间的第 k 维灰关系系数为

$$g_k(t_k, t_{c_i,k}) = \frac{\Delta_{\min}(k) + \theta \cdot \Delta_{\max}(k)}{\Delta_j(k) + \theta \cdot \Delta_{\max}},$$

$$k=1,2, \dots, p.$$

式中, $\Delta_{c_i}(k) = |t_k - t_{c_i,k}|, \Delta_{\min}(k) = \min_j \min_k \{\Delta_{c_i}(k)\}, \Delta_{\max}(k) = \max_j \max_k \{\Delta_{c_i}(k)\}, \theta \in (0,1)$ 是分辨率控制参数,文中试验取 $\theta=0.8$.每一维的灰关系系数计算完成后, d 和 d_{c_i} 之间的相似度可采用每一维加权平均的方法计算:

$$\text{sim}(d, d_{c_i}) = \sum_{k=1}^p \alpha_k \cdot g(t(k), t_{c_i}(k)) \quad (9)$$

式中,参数 α_k 为第 k 维灰关系系数相应的权重.本文试验中为简便记,取 $\alpha_k = \frac{1}{p}$.

5 一种灰关系与信息增益融合的文本分类模型

综上所述,提出一种基于信息增益的改进型文本分类算法如下:

(1)对训练集进行分词、删除停用词等预处理,得到特征项集合 G ;

(2)在各个类别 $c_i, i=1,2, \dots, r$,内分别使用式(4)对集合 G 进行类别特征选择,取最大的 p 个值对应的特征项组成 c_i 类别相应的类别特征集 $H_i = \{t_{i1}, t_{i2}, \dots, t_{ip}\}, i=1,2, \dots, r$ 用于 c_i 的类内文本表示,并由式(2)计算 c_i 的类别中心向量;

(3)使用式(8)计算集合 G 中特征 $t_k(k=1,2, \dots, n)$ 在训练集里的权重,取前 p 个最大值对应的特征项构成训练集的特征集用于训练集的文本表示;

(4)使用式(9)计算待分类文本 d 和类别文本向量 d_{c_i} 之间的相似度,将分类文本 d 划归到与其相似度最大的类别文本相应的类别内.

6 实验结果及其分析

本文对文本分类模型的分类效率进行了试验.试验数据来自 TanCorpV1.0 中科院文本分类语料

库^[9],选取其中的 7 个类 2 820 篇文本进行实验,每类约 400 篇文本. 试验时采用 3 折交叉试验法,将 2 820 篇文本平均分为 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. 取向量维数 $p=120$,效果评估函数使用常用的查准率、查全率和 F_1 测试值,其中:

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 / 类内应有文本数,

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

为了比较改进 IG 模型的分类效果,我们将以经典 IG 进行的特征选择模型(以式(1)进行文本特征选择,以式(2)计算 c_i 的类别中心向量 d_{c_i} ,以向量夹角余弦计算 d 与 d_{c_i} 之间相似度)所相应的 F_1 测试值以 F_{old} 列出,本文提出的改进方法相应的数据分别加以下标 new 以示区别. 分类结果统计如表 1 所示.

表 1 文本分类测试结果

类别	体育	卫生	房产	教育	汽车	财经	娱乐	平均
查准率 _{new}	0.896 2	0.885 9	0.904 1	0.854 9	0.910 2	0.887 9	0.909 2	0.892 6
查全率 _{new}	0.912 7	0.872 3	0.899 2	0.887 1	0.898 3	0.893 7	0.887 3	0.892 9
$F_{1\text{new}}$	0.900 7	0.879 0	0.901 6	0.870 7	0.904 2	0.890 8	0.898 1	0.892 2
查准率 _{old}	0.821 5	0.801 9	0.812 7	0.851 6	0.836 5	0.841 2	0.834 6	0.828 6
查全率 _{old}	0.832 6	0.825 3	0.831 5	0.814 8	0.832 4	0.802 6	0.810 1	0.821 3
$F_{1\text{old}}$	0.827 0	0.813 4	0.822 0	0.832 8	0.833 9	0.821 4	0.822 1	0.824 7

从表 1 可以看出,本文提出改进 IG 文本分类方法取得了较为理想的分类效果. 相比较于常用的基于 IG 的特征选择文本特征降维方法,文本分类效果上平均查准率提高约 7.2%,平均查全率提高约 8.0%,平均 F_1 测试值提高了约 7.6%,新方法的 F_1 测试值达到 89.2%,因此本文提出基于灰关系与信息增益融合的文本分类方法效果还是令人满意的.

7 结束语

本文针对信息增益模型在文本分类中的不足之处进行了研究. 提出了基于改进的 χ^2 统计方法进行类别特征选择,用于文本类别中心向量的表示,使得类别中心向量更能真正体现类别的中心位置;针对 IG 模型在对低频词在类别表示作用上赋权过大问题,提出了基于频数和位置的改进加权方法,改善了 IG 模型在特征选择方面对文本的表示能力. 其实特征项与文本之间、特征项与文本类别之间、特征项相互之间的关系十分复杂,在现有的各种降维方法基础上,探索更加合理的特征降维方式,进一步改善各种模型在理论上的缺欠,是提高文本分类精度的一种可行的途径. 将不同的学科知识进行有效的融合,探讨多学科知识在文本分类上的可行的应用途径,对文本信息处理有着现实性需求,可以丰富文本信息处理的手段和方法,这也是我们今后进一步的研究工作.

参考文献:

- [1] 焦远锋,李万龙,郑山红,等. 一种新的两阶段 FCM 聚类算法[J]. 微电子学与计算机,2009,26(4):40-42.
- [2] 牛强,王志晓,陈岱,等. 基于支持向量机的 Web 文本分类算法[J]. 微电子学与计算机,2006,23(9):102-104.
- [3] Yang Yiming, Pedersen J O. A comparative study on feature selection in text categorization [C]// Proceedings of the 14th International Conference on Machine Learning (ICML297),1997:412-420.
- [4] 刘庆和,梁正友. 一种基于信息增益的特征优化选择方法[J]. 计算机工程与应用,2011,47(12):130-136.
- [5] 方清华. 信息检索加权理论与技术:基于 VSM 模型的分析[J]. 情报杂志,2008,27(6):73-76.
- [6] 肖婷,唐雁. 改进的 χ^2 统计文本特征选择方法[J]. 计算机工程与应用,2009,45(14):136-140.
- [7] 侯汉清,张成志,郑红. Web 概念挖掘中标引源加权方案初探[J]. 情报学报,2005,24(1):87-92.
- [8] 董尤心,张杰,唐宏,等. 效能评估方法研究[M]. 北京:国防工业出版社,2009.
- [9] 谭松波,王月粉. 中文文本分类语料库-TanCorpV1.0 [DB/OL]. (2010-04-10) [2012-01-18]. <http://www.searchforum.org.cn/tansongbo/corpus.htm>.

作者简介:

刘海峰 男,(1962—),博士,副教授. 研究方向模式识别、文本挖掘.

李凯齐 男,(1982—),博士生. 研究方向文本挖掘.

汪泽焱 男,(1972—),副教授. 研究方向统计分析.