结合同义向量聚合和特征多类别的KNN分类算法

林啟锋 蒙祖强 陈秋莲
(广西大学计算机与电子信息学院 南宁 530004)

摘 要 特征选择是文本分类的关键阶段,其选择过程将影响文本分类速度与精度。χ^2 统计量能很好地体现词和类别之间的关系,是文本分类领域特征提取阶段的重要方法之一。分析了 χ^2 统计量在文本分类中的应用,发现CHI向量所表达的与各类别关系的特征词无法全面表达出此类的概念含义,依赖于训练集中出现的特征情况,且该向量仅用于特征选择阶段;针对 χ^2 统计量特征词的表达局限及其向量没有得到充分利用的问题,提出结合同义向量聚合和特征多类别的改进KNN分类算法,该方法能够综合考虑特征所表达的含义,且通过特征集多类别矩阵使CHI向量也能在分类阶段起到提高整个算法效率的作用。实验结果与分析表明,该改进算法明显提高了文本分类效率,并且提高了分类的精度。

关键词 文本分类,χ^2 统计量,特征集多类别矩阵,KNN
中图法分类号 TP18 **文献标识码** A

KNN Text Categorization Algorithm Based on Semantic-Vector-Combination and Multiclass of Feature

LIN Qi-feng MENG Zu-qiang CHEN Qiu-lian
(College of Computer, Electronics and Information, Guangxi University, Nanning 530004, China)

Abstract Feature selection is the key stage in the text categorization, and the processing of it will affect the speed and accuracy of text classification. χ^2 statistic is a important methods in feature selection of text categorization since it measures the dependence between a term and a class effectively. Nevertheless, we found the feature in the vectors of CHI can not fully express the means of concept and it depends the training text set, and the vectors of CHI are used only for the phase of feature selection after the analysis of the application of χ^2 statistic in the text categorization. So this paper proposed an improved kNN text categorization algorithm based on Semantic-Vector-Combination and Multi-class of feature, in which the feature considers the means of concept, and the matrix of multiclass of features will improve the efficiency of algorithm in the stage of categorization. The results and analysis of experiments show that the efficiency of categorization is improved and its accuracy is also enhanced.

Keywords Text categorization, χ^2 statistic, Feature-MultiClass-Matrix, K-Nearest neighbor

1 引言

现有的比较著名的分类方法有贝叶斯(Bayes)、K最近邻法(KNN)、支持向量机(SVM)、神经元网络(Nnet)等[1,2]。其中许多分类方法采用向量空间模型(VSM),通过特征提取,使每个文本都表示为一个向量,方便进行快速的计算;因而特征提取是建立向量空间模型的基础,特征提取及特征在各类别间分步的准确性将影响分类的最终结果。本文通过分析特征提取阶段常用的评估函数 χ^2 统计量,发现该特征提取方法虽然能够反映出特征词和类别之间的关系,但它的统计过程并没有完全反映出这类语义特征与类别的关系,只是反映出了训练文档中出现的某些特征词与类别的关系;且正如文献[3,4]指出,在使用 χ^2 分步的统计量时,对CHI的取值,往往只考虑该特征词与所有类别之间的 χ^2 平均值或最大值,丢弃了许多有用的信息。为了充分利用统计量的相关性,目前存在一些统计量的改进方法,如向量聚合技术[3],但该技术聚合的是分布相同的特征词,对于这样的特征词无法保证其语义相关,且这样的聚合依赖于特征的分布情况,而特征分布情况依赖于训练集中存在的文本,因此这种方法较适用于话题的检测与追踪,而对于文本自动分类适用性较低。文献[4]的文本特征统计量叠加也是一种较常用的充分利用统计量的改进方法,该方法虽然能充分利用各特征在各类上的分布情况,但叠加的过程更使得分类从宏观上体现文本分类特性,忽略了各特征词的语义特性。因此本文通过对 χ^2 统计量特点的分析,提出结合同义向量聚合和特征多类别的KNN分类算法;实验及理论证明,该算法极大减少了KNN文本分类的计算量,提高了文本分类的精度。

2 文本分类中的 χ^2 统计量

2.1 特征选择过程中的 χ^2 统计量

文本特征选择也就是通过各种评估函数从文本中选出能

到稿日期:2013-05-21 返修日期:2013-07-09 本文受国家自然科学基金项目(61063032),广西自然科学基金项目(2012GXNSFAA053225)资助。

林啟锋(1987-),男,硕士生,主要研究方向为人工智能、数据挖掘,E-mail:ssynkqtd@163.com;蒙祖强(1974-),男,博士,教授,主要研究方向为人工智能、数据挖掘;陈秋莲(1974-),女,硕士,副教授,主要研究方向为智能优化、CAD。

够代表文本且能反映出文本与各类别之间相关程度的特征词选择过程。文献[5,6]对常见的特征词权重评估函数做了比较，常用的权重计算方法有：特征频率/反文档频率 TFIDF、信息增益 IG、互信息 MI、χ^2 统计量 CHI、文本证据权等。在中文文本分类中使用较频繁的且在某些数据集(如 Reuters-21578 和 OHSUMED)[7]中被认为分类效果较好的特征提取方式是统计量 CHI 方法。

统计量 CHI 方法认为词和类别之间的关系符合 χ^2 的分步规律。通过计算每个特征词的统计量，可以得到该特征词与各类别之间的联系，若 χ^2 值越小，该词和某个类别之间的相性越小，则该词独立于该类之外；反之，χ^2 值越大，相关性越大，则该词与该类关系越紧密，由此我们有理由认为包含该词的文本属于该类别的概率越大。

设类别集合为 $\boldsymbol{C}=\{c_1,c_2,\cdots,c_n\}$，文本总数为 N，词为 t，A 表示 t 和 c_i 同时发生的次数，B 表示 t 发生而 c_i 不发生的次数，C 表示 c_i 发生而 t 不发生的次数，D 表示 t 和 c_i 都不发生的次数，而 CHI 公式如下[7]：

$$\chi^2(t,c_i)=\frac{N(A*D-C*B)^2}{(A+B)*(A+C)*(B+D)*(C+D)} \quad (1)$$

由于 χ^2 统计量在计算过程中能较好地体现出该特征词与各类别之间的相关性，使得 CHI 方法在文本的特征选择中得到广泛的应用，通常取词条在所有类别中的 χ^2 平均值或最大值为其 CHI 值。

2.2 同义特征词 CHI 向量聚合

在特征空间中，由于特征维数较大，难免会出现词义相同的情况，若对于不同的表达方式，降低计算机识别能力，必然会因此减小文本间的相似度，造成不必要的误差。由此，我们在计算特征词的统计量时，若只是计算每个特征词的统计量，显然得到的特征词 χ^2 分步只能反映出这个特征词在给定的训练文档下与各类别之间的相关性，若该训练集或训练文本包含若干个意思相同的特征词，存在若干个 CHI 向量，则不仅无法综合全面地体现出该类特征词与类别之间的关系，而且特征向量的维数会出现虚高的现象，由此得到的特征词与类别的相关性会过于依赖所给定的训练文本集；若由此导致测试文本无法正确分类，则必然降低整个算法的准确率和召回率。

因此希望得到的特征词 χ^2 分布不仅仅只是表示该特征词与各类别之间的关系，而且要能表达语义相同的这类词与各类别之间的相关性。为了得到这类词与各类别的相关性，我们把表达相同语义的各词的 CHI 向量进行聚合，从而得到能更加综合、全面地反映出这类词与各个类别的相关性的 CHI 向量。

通过聚合规则，我们可以把语义相似或相同一类词的 CHI 向量聚合在一起，使其更加全面、综合地反映出这类词与各个类别之间的相关性，从而极大地减少特征词与类别的相关性对训练文本的依赖。

聚合规则：对于同义特征词 a 和 b 的两个 CHI 向量，设两个向量为：

$\chi^2(a,C)=(\chi^2(a,C_1),\chi^2(a,C_2),\cdots,\chi^2(a,C_n))$，$\chi^2(b,C)=(\chi^2(b,C_1),\chi^2(b,C_2),\cdots,\chi^2(b,C_n))$，则把特征词 a 和 b 的 CHI 向量进行聚合，过程为：

$\chi^2(a+b,C)=(\chi^2(a,C_1)+\chi^2(b,C_1),\chi^2(a,C_2)+\chi^2(b,C_2),\cdots,\chi^2(a,C_n)+\chi^2(b,C_n))$ (2)

2.3 χ^2 统计量特征的多类别性

CHI 方法在特征选择中表现较好，但目前的分类算法把花费大量时间计算得到的 CHI 向量仅用于特征选择；且在特征选择的过程中并没有使得该向量得到充分利用，没有极大地体现出特征与各个类别之间的相关性，通常只是利用了其最大相关的值，或者平均值，无法利用该特征词与各类别之间相关性的差异，例如，对于取最大值的情况，假如若干个类中有两个类较相似，那必然会引起有些特征与相似的两个类别的相关性都较大，若只是取相关性最大的值，则会忽略第二大相关的类别，必然会引起不必要的误差，甚至因若干个特征误差导致某些文本被错分到另一个类；同理，对于取均值的情况也必然会引起一些误差，由于均值相同，其分布有可能存在许多较大的差异。

基于 CHI 统计量没有得到充分利用的情况，本文在使用 CHI 向量时，不采用取其最大值或均值的方法，而将保留整个 CHI 向量，不论其相关性大小，因为相关性大的在分类过程中对分类结果的影响也大，相关性小或者不相关的对分类结果的影响较小或者不影响。由此我们可以理解为一个特征是属于多个类别的，根据每个类别相关性大小来决定该特征属于该类别的程度。为方便理解与计算，针对提到的包含多个特征的特征集多类别矩阵，我们把 CHI 向量理解为特征多类别性向量。

定义 1(特征多类别性向量) 设 $\boldsymbol{mcv\_a}=(x_1,x_2,\cdots,x_n)$ 表示特征 a 的多类别向量，其中 x_i 值的大小表示该特征属于该类的程度，其中 $x_i=\chi^2(a,C_i)$。

利用定义 1，我们可以很容易得到每一个特征对于各类别的归属程度，但在得到每一个文本各个特征的 CHI 向量时，正如 2.2 节中提到的特征存在义同形异的情况，使得每一个特征词的 CHI 向量无法综合全面地表达出这类特征与每个类的相关程度，因此我们将进行同义特征 CHI 向量的聚合。为方便计算，我们引入特征集多类别性矩阵，即反映出多个特征的多类别性数学表示，以便进行文本向量的特征多类别表示和包含训练文档集所有特征的特征多类别表示及计算。

定义 2(特征集多类别矩阵) 设特征集向量表示为 $\boldsymbol{V}=(t_1,t_2,\cdots,t_n)$，第 i 个特征在 m 个类上所对应的多类别性向量为 $\boldsymbol{mvc\_t_i}=(x_{i1},x_{i2},\cdots,x_{im})$，则特征集多类别矩阵为 $\boldsymbol{T}$：

$$T=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix} \quad (3)$$

式中，x_{ij} 表示第 i 个特征属于第 j 个类别的程度，$x_{ij}=\chi^2(t_i,C_j)$。

基于特征词的 χ^2 分布特性以及 KNN 的高计算量，本文将把训练文本集中的所有特征词进行同义特征 CHI 向量聚合；利用 2.2 节的聚合规则，把训练文本集中所有文本的所有特征进行同义向量聚合，最终将得到一个与这些类别相关的包含训练集中所用特征词的特征集多类别矩阵；该矩阵将应用于第 3 节中改进 KNN 分类算法的初次类别判定。

3 基于两次类别判定的改进 KNN 算法

KNN 算法的计算量较大，往往限制了它在即时领域的应

用，目前在降低其计算量方面做了一些研究，衍生出了各种各样的改进算法[8,9]，如类的中心向量算法，或选出一些类的代表文本进行类似的计算，但这些方法的使用，往往也引入了分类的误差，降低了分类的准确率。本文将通过二次判别方法进行类别判定，通过初次判别选出可能性最大的若干类，从而使用原始的 KNN 算法进行第二次详细判定。这样可以在不引入误差的情况下，减少不必要的计算量。

3.1 类别初判定

对于类别的初次判定，我们基于 KNN 算法思想，选出最相似的 w 个类，从而把这 w 个类中的文本当成训练文本集，进行 KNN 分类。为选出这 w 个最相似类(最近邻)，我们基于包含训练集中各类特征词的特征集多类别矩阵，把待分类文本与该矩阵进行计算，得到待分类文本与各类别的相似程度，从中选出相似度最大的 w 个类作为训练子集进行分类。具体方法如下：

设某个测试文档向量 $\boldsymbol{a}=(w_1,w_2,\cdots,w_n)$；包含训练集中各类特征词的特征集多类别矩阵为 $\boldsymbol{T}$；$\mathbf{simV}(a,\boldsymbol{C})$表示测试文本与各个类相似的向量，$sim(a,C_i)$表示测试文本与第 i 个类的相似值；

$$\begin{aligned}\mathbf{simV}(a,\boldsymbol{C})&=a*\boldsymbol{T}\\&=(w_1,w_2,\cdots,w_n)\begin{pmatrix}x_{11}&x_{12}&\cdots&x_{1m}\\x_{21}&x_{22}&\cdots&x_{2m}\\\vdots&\vdots&\vdots&\vdots\\x_{n1}&x_{n2}&\cdots&x_{nm}\end{pmatrix}\\&=(\mathrm{sim}(a,\boldsymbol{C}_1),\mathrm{sim}(a,\boldsymbol{C}_2),\cdots,\mathrm{sim}(a,\boldsymbol{C}_m))\end{aligned}\tag{4}$$

式中，$\mathrm{sim}(a,\boldsymbol{C}_i)=\sum_{k=1}^{n}w_k*x_{ki}$。

通过式(4)我们能得到待测试文本与 m 个类之间的相似程度，进而选出其中相似度最大的 w 个类(本文的 w 我们取总类别数的 1/4)，从而在由 w 个类所包含的文本组成的训练集上进行 KNN 算法，这样我们在对文本进行分类时就能极大地减少算法的计算量，使其分类效率提高至原来的 m/w 倍。

3.2 改进 KNN 算法

基于特征集多类别矩阵初次判定，我们把原来包含 m 个类别的训练集缩小到只包含 w 个类别的训练集，从而计算测试文本与 w 个类训练文本之间的相似度，找出与测试文本最相似的 k 个文本，并根据这 k 个文本判定待测试文本的类别。此过程不仅能减少不必要的计算量，还能避免与大量无关类的文本进行相似度计算时所带来的噪声。在给出具体改进算法之前，先给出特征集多类别矩阵的获取算法，从而结合矩阵进行类别的初判定及后续的二次判定，特征集多类别矩阵获取算法如下：

输入：训练集 Tr，类别集 C，类别总数 m

输出：特征集多类别矩阵 T

1)依次求出训练集每个文本中每个特征多类别性向量(即 CHI 向量)；

2)把特征集多类别矩阵 T 初始化为 0 行 0 列空矩阵；

3)依次判断每个特征 t_i，在矩阵 T 中是否存在与 t_i 同义的特征的多类别性向量；

4)若矩阵 T 中已存在与特征 t_i 同义的特征的多类别性向量，则通过 2.2 节的聚合规则把同义特征的多类别性的向量进生聚合，转 6)；

5)若矩阵 T 中没有与特征 t_i 同义的特征的多类别性向量，则把该特征的多类别性向量加放矩阵的最后一行；

6)若所有的特征已处理完毕，则算法结束，否则转 3)。

对于步骤 3)特征词间是否同义的判断，我们将基于知网中词的各种定位关系来进行；通过特征集多类别矩阵获取算法得到特征集多类别矩阵 $\boldsymbol{T}$ 后，便可通过式(4)计算每一篇测试文本与每个类别的相似度，从中选出最相似的 w 个类别，从而达到降低计算量的目的。改进 *KNN* 具体算法的过程如下：

输入：类别集 C，测试集 Te，训练集 Tr，最相似文本数 k，初判别最相似的类别数 w。

输出：各测试文本的类别

1)根据特征集多类别矩阵获取算法，先得到特征集多类别矩阵 T；

2)取新的测试文本 d，利用式(4)，计算 d 与特征集多类别矩阵 T 的乘积，获得该文本与各个类别之间的相似度向量；

3)从该相似度向量中提取相似度最高的 w 个类别，作为初判定的结果；

4)依据初判定的结果，在 w 个类中选出与文本 d 最相似的 k 个最近邻；

5)根据这 k 个最近邻在 w 个类中的分布，来判定文本 d 最终属于哪一个类；

6)若所有的测试文本均已分类，算法结束，否则转 2)。

由上述算法可以知道，对于改进 KNN 算法中的初次类别判定，能缩小该测试文本所属类别的范围，从而没必要与所有类别的训练文档进行相似度的计算。对于训练集在类别分布较平均的情况下，改进 KNN 算法的计算量变成原始算法的 w/m。

4 性能测试与评估

为了对该算法的性能进行分析，我们用 VC++6.0 实现本文的算法。数据集方面，由于中文的文本分类没有一个公共的通用的语料库对各种分类算法进行评估，且本文不涉及分词方面的研究，因此本文选择了中科院计算机所的 TanCorp V 1.0 语料库[11]作为算法性能测试数据集。该数据集共包含 14150 篇文本，划分为 12 个类，具体分布情况如表 1 所列。由于这些文档在 12 个类中的分布不均匀，因此我们随机地选取每个类中的 100 篇文本作为测试文本，其余皆作为训练使用(由于地域类包含文档较少，我们只从 150 篇中选取 50 篇作为测试文档)。

表 1 TanCorp V1.0 数据集的样本分布

类别	财经	体育	教育	卫生	房产	科技	艺术	娱乐	人才	地域	汽车	电脑
文本数	819	2805	808	1406	935	1040	546	1500	608	150	590	2943

为了验证改进的算法是否在分类效率上得到提高，我们分别对从 12 个类中抽出的 1150 篇文档进行了分类时间的比较，KNN 算法中对于 k 的取值，本文置 $k=10$，且初次分类 $w=3$(为分类总数的 1/4)，对于特征集多类别矩阵 T，我们认为它是分类规则的获取阶段，因此本文所讨论的分类时间不包含该规则获取对时间的消耗。

从图1可以看出，基于特征集多类别矩阵的类别初判定缩小了类别范围，有效地提高了文本的分类效率，尽管有些类（如体育类和电脑类）并没有像3.1节中所说的提高 m/w 倍，那是因为分类过程中类别初判定也需要一定量的时间；且训练样本分布得极度不平均，由于体育类和电脑类所占的训练样本的总数较大，尽管待判定类别数变成总数的 w/m，但其训练样本子集 w 个类所包含的训练文本数却不是训练样本总数的 w/m。

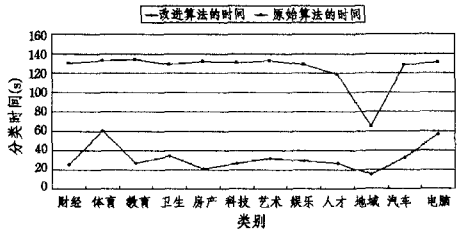

图1 改进算法与原始算法分类时间对比

通过图1，我们知道训练文本集在类别间的分布不均使得改进算法的分类时间曲线较原始的算法波动较大（不考虑地域类，此处产生较大的波动是因为测试文本数较小的原因），因为原始算法对于任何一个待分类文本，它的训练集文本总数总是恒定不变的，而对于改进的算法，在样本分布不均的训练集中，每一个待分类文本所属的可能类别不同，可能类别所包含的文本总数也必然不同，该文本的分类时间必然受此影响，由此可知改进算法分类效果的提高容易受训练集在类别间分布的影响。

对于一个文本分类算法的改进，若以牺牲其准确率、召回率来换取其效率的提升，显然这样的改进是毫无意义的；为此我们对该改进算法的准确率进行验证，并与传统的KNN算法进行对比；在给出实验数据与分析之前，我们先给出准确率微平均和宏平均的计算公式：

$$\text{微平均准确率}(Micro\_precision)=\frac{\text{所有类别正确分类文本数之和}}{\text{训练文本总数}}$$

$$\text{宏平均准确率}(Macro\_precision)=\frac{\text{所有类别准确率之和}}{\text{类别总数}}$$

下面将对比改进算法与传统KNN算法的准确率，并对其进行分析；为了了解 w 取值对分类性能的影响，我们将对 w 取不同值：

通过表2可知，改进的算法不仅没有牺牲其准确率，反而得到提升，那是因为在分类过程中不像传统方式那样把某个特征简单地归到某一类中，而是保留了它与各类的相关程度，使其没有因省略除了最大相关类外的其它类的相关性而引起误差，也使得这些误差没有因为每个文本包含大量特征而得到叠加；且在初次判别后，避免了该测试文本与其它不相关类的文本进行相似度计算带来的噪声。通过对 w 取值的不同我们发现，该算法中 w 的取值与KNN中对 k 的取值情况相似，若取值过大，则会引入误差，若取值过小，则会由于考虑得不够全面而同样会引起分类性能的下降。

在准确率上，表2中的宏平均相对于微平均较低，那是因为宏平均比较容易受小类的影响，如艺术与娱乐这种界限较为模糊的类的存在；且类别中文本数量越大，能代表其类别的特征越多，特征集多类别矩阵能反映出更多该类别的信息，使得在分类过程中更加不容易分错。

表2 两种方法的微平均和宏平均准确率对比

	传统KNN	改进KNN (w=4)	改进KNN (w=3)	改进KNN (w=2)
微平均	0.9033	0.9287	0.9312	0.9243
宏平均	0.8471	0.9152	0.9173	0.9079

综上所述，结合同义向量聚合和特征多类别的改进KNN分类算法不仅在效率上提升得较为明显，且其分类的准确率也得到了提高。

结束语 本文分析了 χ^2 在特征选择过程中由于CHI取值方式（如最大值、均值等）产生的分布信息没有得到充分应用的情况，提出保留其有用的信息，衍生出特征集多类别矩阵，使这些信息在分类过类过程中也得到充分利用，从而有效提高KNN算法的分类效率；但该方法在处理界限较模糊的类较多的情况时，为了不引起误差和降低准确率，w 不得不取较大的值，使得算法的效率不是那么明显，甚至因为要多计算文本与特征集类别矩阵而降低效率，因此如何锐化类别之间的差异将成为下一步研究的重点。

参考文献

[1] Yang Yi-ming, Liu Xin. A re-exam ination of text categorization methods[C]//Proceedings, 22nd Annual Internaitonal ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR). 1999: 42-49

[2] 陈雅芳，徐从富. 中文文本分类方法研究[D]. 杭州：浙江大学，2012

[3] 李莹，张晓辉，王华勇，等. 一种应用向量聚合技术的KNN中文文本分类方法[J]. 小型微型计算机系统，2004，25(6)：993-996

[4] 印鉴，谭焕去. 基于统计量的KNN文本分类算法[J]. 小型微型计算机系统，2007，28(6)：1094-1097

[5] 林少波，杨丹. 中文文本分类特征提取方法的研究与实现[D]. 重庆：重庆大学，2011

[6] 申红，吕宝粮，内山将夫，等. 文本分类的特征提取方法比较与改进[J]. 计算机仿真，2006，23(3)：222-224

[7] Yang Y, Pedersen J P. A comparative study on feature selection in text categorization[C]//Proceedings of the Fourteenth International Conference on Machine Learning (ICML'97). 1997: 412-420

[8] 王爱平，徐晓艳，国玮玮，等. 基于改进KNN算法的中文文本分类方法[J]. 微型机与应用，2011，30(18)：8-10

[9] Y Gao, P Jin-yan, F Gao. Improved Boosting Algorithm through Weighted K-Nearest Neighbors Classifier[C]//Proceedings, 3rd International Conference on Computer Science and Information Technology (ICCSIT). 2010: 36-40

[10] 董振东，董强. 知网简介[EB/OL]. http://www.Keenage.-com, 2012-7-23

[11] 谭松波，王月粉. 中文文本分类语料库 TanCorpV1.0[EB/OL]. http://lcc.software.ict.ac.cn/tansongbo/corpusl.php, 2010-10-23