文章编号：1004-4353(2021)02-0175-05

基于相对熵的 KNN 文本分类方法的研究

崔东虎， 赵亚慧， 崔荣一*
（延边大学 工学院，吉林 延吉 133002）

摘要：为提高处理文本相似度的效果，提出了一种基于相对熵度量文本差异的 KNN 算法. 该算法首先对文本进行预处理(分字与删去停用字)和构建特征字字典；然后计算训练集中所有文本特征字的概率，并组成训练集(特征字概率矩阵)；最后计算预测文本的特征字概率向量，并通过计算和统计 K 个预测文本与训练集文本间相对熵最小的文本类别个数后将数目最多的类别作为测试样本的类别. 实验结果表明，该算法的分类效果不仅显著优于传统 KNN、SVM、Decision Tree、朴素 Bayes 算法的分类效果，且在小样本数据情况下还明显优于 RNN 算法.

关键词：文本分类；KNN 算法；相对熵；欧氏距离
中图分类号：TP391.1 **文献标识码**：A

Research on text classification of *K*-nearest neighbor algorithm based on relative entropy

CUI Donghu， ZHAO Yahui， CUI Rongyi*
（*College of Engineering*, *Yanbian University*, *Yanji* 133002, *China*）

Abstract: To improve the effectiveness of processing text similarity, a KNN algorithm based on the relative entropy measure of text feature differences was proposed in this paper. Firstly, the algorithm preprocessed the text, including character separation and deletion of stop characters, and constructed a feature character dictionary. Then the probabilities of all the text feature characters in the training set were calculated, and the training set (probability matrix of feature character) was formed. Finally, we calculated the probability vector of feature characters of the predicted text, and counted the number of text categories with the lowest relative entropy between the K predicted texts and the training set, and used the category with the highest number as the category of the test sample. The experimental results show that the classification effect of this algorithm is not only significantly better than that of the traditional KNN, SVM, Decision Tree, and Naive Bayes, but also significantly better than that of RNN algorithm in the case of small sample data.

Keywords: text classification; KNN algorithm; relative entropy; Euclidean distance

0 引言

随着全球互联网络的普及以及网络信息的海量化，人们对文本自动分类技术愈加关注. 近年来，基于机器学习的文本分类方法因具有人工干预少、分类速度快和精度高等优点而受到学者们的广泛关注[1-2]. 目前，较为成熟的文本分类方法有 K 近邻(K-Nearest Neighbor，KNN)算法[3]、朴素贝叶斯(Naive Bayes，NB)算法、决策树(Decision Tree，DT)算法、支持向量机(support vector machines，SVM)、深度学习模型[4]等，这些模型的主要分类流程为文本信息获取、分词处理、特

收稿日期：2021-02-17 * **通信作者**：崔荣一(1962—)，男，博士，教授，研究方向为模式识别、智能计算.
基金项目：国家语委科研项目(YB135-76)；延边大学外国语言文学一流学科建设项目(18YLPY13)

征提取、文本向量表示、算法处理及性能评价[5-6]. 其中KNN算法虽然具有计算简单、准确率高的优点[7-8],但其只适合于描述低维度特征向量间的差异,难以满足实际需要;而基于深度学习的分类模型虽然分类效果较好,但是需要大量的数据和训练时间[9-10]. 为了提高KNN算法在高维特征空间中的效果,本文以单字概率作为文本特征,提出了一种基于相对熵度量文本特征差异的KNN算法,并通过实验验证了本文分类方法的有效性.

1 相关理论

在本文中,文本特征用文字在文本中出现的概率来表示,用相对熵计算两个样本之间的概率差异. 当计算所得的相对熵差值较大时,说明两个样本的差异较大;当计算所得的相对熵差值较小时,说明两个样本间的差异较小[11].

1.1 相对熵

相对熵(亦称KL散度)虽然不是严格意义上的距离,但是它可以有效描述两个概率分布之间的差异. 设$P(x)$和$Q(x)$是随机量X的两个概率分布,则P对Q的相对熵的计算公式[12]为:

$$D_{\mathrm{KL}}(P\|Q)=\sum_{x}P(x)\log\frac{P(x)}{Q(x)}. \tag{1}$$

在具体的文本分类问题中,P为训练集文本中每个特征的概率分布,Q为测试集文本中每个特征的概率分布. 本文采用式(1)度量样本间的概率分布差异.

1.2 KNN算法

KNN算法[13]是一种基于实例的学习方法,该算法分为训练和分类两个阶段. 在训练阶段,算法对文本特征进行抽取,并将文本以特征向量的形式定义到实数域中,即将文本内容形式化为向量空间中的点. 在分类阶段,首先按与训练阶段同样的方法将待分类文本表示为特征向量,然后计算该待分类文本与训练样本集中每个文本的距离,最后找出与待分类文本最近的K个邻居,并由这K个邻居中的多数类别来决定待分类文本的类别.

KNN算法中的关键步骤是测量样本间距离. 欧氏距离[14]是测量样本间距离的常用函数,该函数虽然具有计算简单、方便的优点,但是随着特征维度的增加其区分不同特征的能力逐渐变弱,同时因值域大的变量在计算中占据主导作用,因此此时计算出的样本间距离会出现较大误差,进而影响分类的准确率. 为了克服欧氏距离的缺点,本文采用相对熵度量样本之间的差异.

2 分类器设计

2.1 分类处理流程

本文采用数据预处理、模型训练、分类结果预测3个阶段进行文本分类. 分类的核心思想为:采用相对熵计算测试样本与训练样本间的差异,以此找出相对熵最小的K个值,并统计这K个样本中的多数类别. 主要处理步骤描述如下:

步骤1 构造数据集. ①对收集到的语料进行分字处理,并删去停用字,以防止文本维度过大而导致计算消耗过大;②将文本向量化,向量的值是各特征字出现的概率;③将数据分为训练集和测试集.

步骤2 训练KNN分类器. 将所有由特征字概率组成的训练样本向量组合成矩阵.

步骤3 利用KNN算法对测试集进行分类. ①计算测试样本中各特征字出现的概率,并将测试样本表示为特征字的概率向量;②计算测试样本和每个训练集样本的向量所对应的相对熵,并统计相对熵中K个最小的相对熵所对应的训练集样本的类别个数;③将上述结果中的多数类别作为测试样本的类别.

2.2 文本表示与KNN分类器训练

文本表示方法通常采用向量空间模型. 向量空间模型采用TF-IDF方法来计算词频矩阵,但当文本特征维度较大时,采用TF-IDF方法会导致计算消耗过大,进而会降低分类效率. 由于以特征字作为文本特征可以有效减少特征维数,因此本文选取特征字作为特征,以此计算相对熵. 训练样本集可表示为:

$$D=\{(\boldsymbol{x}_i,c_i)\mid i=1,2,\cdots,n\}. \tag{2}$$

其中:$\boldsymbol{x}_i=(x_i^1,x_i^2,\cdots,x_i^T)^{\mathrm{T}}$是一个$T$维向量,即特征维数为$T$(字典中总字数);$x_i^j$表示第$i$个训练样本的第$j$个特征分量值,即第$i$个文本中第$j$个字的出现概率;$c_i$表示第$i$个样本相应的类别,$c_i$属于类别标签集$C$($C=\{1,2,\cdots,m\}$,$m$为类别

数）；n 为训练样本总数.

测试样本集可表示为：

$$Y=\{(\boldsymbol{y}_j)\mid j=1,2,\cdots,n\}. \tag{3}$$

其中：$\boldsymbol{y}_j=(y_j^1,y_j^2,\cdots,y_j^T)^{\mathrm{T}}$ 是一个 T 维向量；y_j^i 表示第 j 个训练样本的第 i 个特征分量值，特征分量值为当前文本中该字出现的概率.

由于在一个文本数据中可能不会出现字典中的所有字，因此概率矩阵中就有可能出现 0. 但由于计算相对熵时 P 或 Q 的概率不能为 0，因此本文采用平滑的方法处理文本，以此避免零概率情况的发生. 文本中的字概率可表示为：

$$P_{ij}=\frac{N_{ij}+1}{N_i+T}. \tag{4}$$

其中：P_{ij} 为第 i 个文本中第 j 个特征字出现的概率，N_{ij} 为第 i 个文本中第 j 个特征字出现的次数，N_i 为第 i 个文本包含的总字数，T 为字典总字数. 第 i 个文本的特征字概率向量可表示为：

$$\boldsymbol{P}_i=(P_{i1},P_{i2},\cdots,P_{iT})^{\mathrm{T}}. \tag{5}$$

KNN 分类器的训练数据可表示为矩阵：

$$\boldsymbol{A}=\begin{pmatrix}P_{11} & \cdots & P_{1T}\\ \vdots & \ddots & \vdots\\ P_{n1} & \cdots & P_{nT}\end{pmatrix}. \tag{6}$$

其中 T 为特征维数，n 为样本数量. 式(6) 中的每一行表示的是特定文档的特征字概率行向量.

2.3　分类算法

本文利用相对熵判断两个样本之间的概率差异，并据此选出特征字概率分布最近的 K 个文本，具体方法为：

1) 在需要分类的文本中，利用式(4) 计算出每个字的出现概率，并按式(5) 把测试文本 $\boldsymbol{d}$ 表示为测试向量 $\boldsymbol{P}_d$；

2) 计算测试向量 $\boldsymbol{P}_d$ 与式(6) 训练集中每一行之间的相对熵，并将相对熵进行升序排序；

3) 根据给定的 K 值取 K 个与测试文本相对熵最小的训练集文本，并统计其中的各类别数目.

测试样本类别的表达式为：

$$c(\boldsymbol{d})\leftarrow \underset{c\in C}{\operatorname{argmax}}\sum_{i=1}^{k}\delta(c,c_i). \tag{7}$$

其中：$c(\boldsymbol{d})$ 为测试文本 $\boldsymbol{d}$ 的类别，c_i 为训练样本 x_i 的类别，δ 定义为 $\delta(a,b)=\begin{cases}1, & a=b;\\ 0, & a\neq b.\end{cases}$

3　实验结果与分析

实验所用的新闻数据集来源于搜狗数据实验室的共 3 000 篇新闻文本，新闻类别分为体育、计算机、经济 3 类.

3.1　计算文本相似度

由于本文的分类实验需要使用相对熵来度量文本间的差异，因此在分类前需要计算出测试文本与所有训练集文本之间的相对熵. 表 1 为计算所得的某一测试文本(均包含 5 个文本)与 3 个新闻类别间的相对熵. 由表 1 可以看出，测试文本与体育类新闻的相对熵相对最低，由此可判断出该测试文本应为体育类新闻.

表 1　测试文本与不同新闻类别间的相对熵

新闻类别	测试文本				
	文本 1	文本 2	文本 3	文本 4	文本 5
计算机	0.969 7	1.052 0	1.053 0	1.086 0	1.167 0
经济	0.650 0	0.739 9	0.740 2	0.741 7	0.744 8
体育	0.401 5	0.405 4	0.480 4	0.504 8	0.540 3

图 1 为表 1 中的测试文本与 100 个训练集文本间的相对熵的分布情况. 从图 1 中可以看出，体育类新闻所对应的相对熵均低于经济类新闻、计算机类新闻所对应的相对熵，由此可进一步判断出上述测试文本应为体育类新闻.

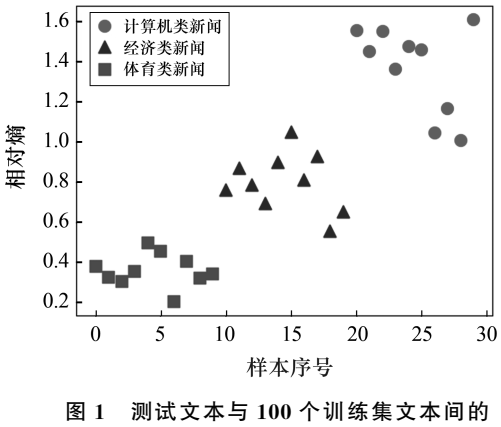

图 1　测试文本与 100 个训练集文本间的相对熵的分布情况

3.2　分类实验

分类实验的步骤为：① 计算出测试样本与所有训练样本间的相对熵，然后对求出的相对熵进行升序排序；② 按照排序结果找出相对熵最小的 K 个训练样本；③ 统计 K 个样本的类别，并根据公式(7) 判断测试文本的类别. 取不同 K 值时基

于相对熵的KNN算法的分类效果如表2所示.由表2可以看出,随着K值的增大,分类的准确度、精度、召回率、F1值均呈下降趋势,所以本文选取K值为1.

表2 K取不同值时的分类效果

K值	准确度	精度	召回率	F1值
1	0.9732	0.9732	0.9734	0.9730
2	0.9732	0.9732	0.9734	0.9730
3	0.9554	0.9554	0.9567	0.9546
4	0.9554	0.9554	0.9570	0.9544
5	0.9495	0.9495	0.9491	0.9495

3.3 算法验证

为了进一步验证基于相对熵的KNN算法的分类效果,本文将该算法与传统KNN方法(基于特征字概率欧氏距离的KNN算法和基于特征字字频欧氏距离的KNN算法)、支持向量机(SVM)、决策树(Decision Tree)、贝叶斯(Bayes)以及循环神经网络(RNN)等算法的分类效果进行对比.

1)与基于特征字概率欧氏距离的KNN算法进行对比.训练集为特征字概率矩阵,待分类文本为特征字概率向量.当K取不同值时基于特征字概率欧氏距离的KNN算法的分类效果如表3所示.对比表2和表3可知,基于相对熵的KNN算法的分类效果在各指标上均显著优于基于特征字概率欧氏距离的KNN算法的分类效果.

表3 K取不同值时基于特征字概率欧氏距离的KNN算法的分类效果

K值	准确度	精度	召回率	F1值
1	0.8823	0.8823	0.8837	0.8823
2	0.8823	0.8823	0.8837	0.8823
3	0.8823	0.8823	0.8856	0.8820
4	0.8823	0.8823	0.8850	0.8824
5	0.8529	0.8529	0.8553	0.8530

2)与基于特征字字频欧氏距离的KNN算法进行对比.实验中将特征字字频矩阵作为训练集,待分类文本为特征字字频向量.表4为K取不同值时基于特征字字频欧式距离的KNN算法的分类效果.对比表4和表2可知,基于相对熵的KNN算法的分类准确率在各指标上依然高于基于特征字字频欧氏距离的KNN算法的分类效果.

表4 K取不同值时基于特征字字频欧氏距离的KNN算法的分类效果

K值	准确度	精度	召回率	F1值
1	0.9287	0.9287	0.9298	0.9285
2	0.9287	0.9287	0.9298	0.9285
3	0.9287	0.9287	0.9279	0.9285
4	0.9376	0.9376	0.9372	0.9370
5	0.9287	0.9287	0.9275	0.9278

3)与SVM、Decision Tree、朴素Bayes算法进行对比.实验使用特征字字频作为文本的特征表示.3种算法的分类效果见如表5.对比表5和表2可知,基于相对熵的KNN算法的分类效果最优.

表5 3种算法的分类效果

分类算法	准确度	精度	召回率	F1值
Decision Tree	0.8307	0.8307	0.8280	0.8277
SVM	0.9724	0.9724	0.9717	0.9719
Bayes	0.8550	0.8550	0.8544	0.8536

4)与RNN算法进行对比.图2为基于相对熵的KNN算法与RNN算法在不同文本数据量时的分类效果.由图2可以看出:当文本数据小于2700个时,基于相对熵的KNN算法的分类效果优于RNN算法,并且数据量越少,基于相对熵的KNN算法的分类效果就越明显;当文本数据大于2700个时,RNN算法的分类效果优于KNN算法的分类效果,并且随着文本数据量的增加,RNN算法分类效果的优势越为明显.

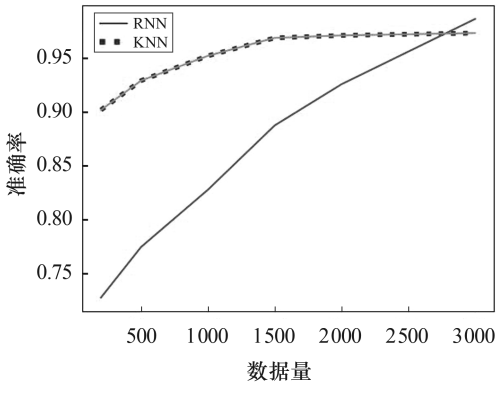

图2 基于相对熵的KNN算法与RNN算法在不同文本数据量时的分类效果

4 结论

研究表明,本文提出的基于相对熵的KNN

算法的分类效果显著优于基于欧氏距离的 KNN 算法和 SVM、Decision Tree、朴素 Bayes 算法的分类效果，并且在小样本的情况下还显著优于 RNN 算法的分类效果，因此本文方法在文本分类中具有良好的应用价值. 本文在研究中使用的文本表示方法未能考虑特征间的重要性差异，因此在今后的研究中我们将对重要程度不同的特征进行加权处理，从而更好地进行文本表示，以提升本文方法的效果.

参考文献：

[1] 刘娇，崔荣一，赵亚慧，等. 跨语言文献相似度的分析方法[J]. 延边大学学报(自然科学版)，2016，42(2)：151-155.

[2] 张雷，崔荣一. 基于编辑距离的词序敏感相似度度量方法[J]. 延边大学学报(自然科学版)，2020，46(2)：140-144.

[3] 邵珊珊. 基于 KNN 的分类方法及其应用研究[D]. 秦皇岛：燕山大学，2019.

[4] 张冲. 基于 Attention-Based LSTM 模型的文本分类技术的研究[D]. 南京：南京大学，2016.

[5] 王亚林，陈忍忍. 不同机器学习算法在分类问题中的应用比较[J]. 黑龙江科学，2021，12(4)：16-18.

[6] ABDULATEEF S，KHAN N A. Machine learning based sentiment text classification for evaluating treatment quality of discharge summary[J]. Information (Switzerland)，2020，11(5)：17.

[7] KIBANOV M，BECKER M，MUELLER J，et al. Adaptive kNN using expected accuracy for classification of Geo-spatial data[J]. ACM Press，2018，18(33)：857-865.

[8] HUANG X，XIONG L，LIU Y，et al. An improved KNN short text classification algorithm based on category features[J]. Computer Engineering and Science，2018，40(3)：148-154.

[9] NOUSHAHRH G，AHMADI S. Multitask learning for text classification with deep neural networks[C]//International Conference on Innovative Techniques and Applications of Artificial Intelligence. Springer-Verlag：Springer International Publishing，2016：119-133.

[10] 万家山，吴云志. 基于深度学习的文本分类方法研究综述[J]. 天津理工大学学报，2021，37(2)：41-47.

[11] 马燕. 基于相对熵的作品作者判定方法[J]. 文教资料，2014(31)：131-133.

[12] HUIBIN L，WEI C，AGUS S. Relative entropy based method for probabilistic sensitivity analysis in engineering design[J]. Journal of Mechanical Design，2006，128(2)：326-336.

[13] 胡春月. 基于 KNN 算法的佚名诗词作者概率研究[J]. 技术与市场，2020，27(11)：69-70.

[14] 丁义，杨建. 欧式距离与标准化欧式距离在 k 近邻算法中的比较[J]. 软件，2020，41(10)：135-136.