一种改进的KNN文本分类算法

樊存佳 汪友生 边 航
(北京工业大学电子信息与控制工程学院 北京 100124)

摘 要：当今大数据时代，文本数据占相当大的比重，作为有效管理和组织文本数据的方法，分类逐渐成为关注的热点。KNN是一种经典的分类算法，针对其分类速度和分类精度无法同时兼顾的不足，采用改进的K-Medoids聚类算法裁剪对KNN分类贡献小的训练样本，从而减少KNN相似度的计算量，并定义代表度函数有差别地处理测试文本的K个最近邻文本，以提高KNN的分类精度。实验结果表明，改进后的方法在分类速度上和分类精度上均有明显地提高。

关键词：文本分类；KNN；裁剪训练样本；代表度函数
中图分类号：TP391 **文献标识码**：A **国家标准学科分类代码**：510

Improved KNN text classification algorithm

Fan Cunjia Wang Yousheng Bian Hang
(College of Electronic and Control Engineering, Beijing University of Technology, Beijing 100124, China)

Abstract: Text data accounts for a large proportion in the era of big data nowadays, text classification, as an effective method of managing and organizing text data, has attracted much attention. KNN is a classic classification algorithm, but its classification speed and accuracy cannot be considered synchronously. Aiming at this shortage, the improved K-Medoids clustering algorithm is adopted to cut the training samples which make little contribution to the classification, to reduce the KNN similarity computation. The representativeness function is defined in order to treat K nearest neighbor samples of testing text differently, to enhance the accuracy of KNN. The results show that the improved method performs better than the traditional method both in speed and accuracy of classification.

Keywords: text classification; KNN; cut the training samples; the representativeness function

1 引 言

当今大数据时代，挖掘数据潜在的价值至关重要。数据挖掘作为发现数据潜在价值的技术，引起极大关注。大数据中文本数据占相当大的比例，而文本分类作为有效组织和管理文本数据的数据挖掘方法，成为普遍关注热点之一。它在信息过滤、信息组织和管理、信息检索、数字图书馆以及垃圾邮件过滤等方面得到广泛应用。常用的分类方法，如K最近邻(K-nearest- neighbor, KNN)[1]，贝叶斯(naive bayes, NB)[1]以及支持向量机(support vector machine, SVM)[2]等。KNN作为经典的分类方法之一，有实现简单、鲁棒性高等优点；但也存在很多缺点，以至于不能适用于很多实际应用中。KNN的不足主要包括以下两个方面：第一，分类过程中因相似度计算量巨大而耗费大量时间，导致分类效率低；第二，分类性能容易受训练样本的影响，当数据出现严重不均匀分布时，分类器性能可能受到严重影响，甚至变得极差[3]。针对KNN分类过程计算量大的问题，将很多研究者的改进总结为3个方面：第一，改进特征选择方法，将那些对分类贡献小的特征词舍弃，实现对VSM(vector space model)模型的有效降维[4]；第二，通过选取原始训练文本集中的一些代表文本作为新的训练文本集或者删除原来训练文本集中的某些对分类贡献小的文本，将删除后剩余的文本作为新的训练文本集[5-8]；第三，设计快速搜索算法，以加快待分类文本的K个最近邻文本的搜索速度[9]。

针对目前各种KNN改进型算法，在速度和精度上难以兼顾的情况，一方面采用改进的K-Medoids聚类算法以裁剪对KNN分类贡献小的训练样本；另一方面为提高KNN算法分类精度，定义代表度函数$u(d_i,C_j)$，实现有差别地处理测试文本的K个最近邻文本。实验结果表明，本文改进后的KNN算法在分类速度和分类精度上均有较明显的提高。

收稿日期：2015-08

2 KNN 算法

计算待分类样本与已知训练样本的欧氏距离或余弦相似度，找到与待分类文本距离最近或者相似度最大的 K 个最近邻文本，再根据 K 个最近邻文本的类别来判断待分类文本的类别[10]，简单地说，K 个最近邻文本中大多数属于某个类别，则样本也属于这个类别[11]。具体步骤如下：

假设训练文本集为 S，其中有 N 个类别 $C_1,C_2,\cdots,C_N$，S 的总文本数为 M，特征向量维度阈值为 n。$d_i=\{x_{i1},x_{i2},\cdots,x_{ij},\cdots,x_{in}\}$ 表示 S 中的一个文本的特征向量形式 $(0<i\leqslant M)$，x_{ij} 表示 d_i 的第 j 维的权重 $(0<j\leqslant n)$。待分类文本的特征向量形式为 $d=\{X_1,X_2,\cdots,X_j,\cdots,X_n\}$，$X_j$ 表示 d 的第 j 维的权重 $(0<j\leqslant n)$。余弦相似度计算如式(1)所示。

$$Sim(d,d_i)=\frac{\sum_{j=1}^{n}(X_j x_{ij})}{\sqrt{\sum_{j=1}^{n}(X_j{}^2)}\sqrt{\sum_{j=1}^{n}(x_{ij}{}^2)}} \tag{1}$$

通过式(1)找到待分类文本的 K 个最近邻文本后，最后通过式(2)计算待分类文本 d 属于每个类别的权重，将待分类文本归到权重最大的类别。

$$W(d,C_j)=\sum_{i=1}^{K}Sim(d,d_i)y(d_i,C_j) \tag{2}$$

式中：$y(d_i,C_j)$ 为类别属性函数，如式(3)所示。

$$y(d_i,C_j)=\begin{cases}1,d_i\in C_j\\0,d_i\notin C_j\end{cases} \tag{3}$$

3 改进的 KNN 算法

KNN 在相似度计算时训练文本集的全部文本都要参与运算，分类过程的时间复杂度与训练文本的规模成正比[12]，从而导致 KNN 的分类效率很低。当训练集为海量文本集时，相似度计算消耗时间急剧增加，算法将失去实用性。另一方面，在 KNN 算法中的类别属性函数 $y(d_i,C_j)$，对于测试文本的 K 个最近邻文本“一视同仁”，不考虑它们对所在类别的重要程度，即类别中的每个文本对于该类别的代表性不同[13]，这在一定程度上影响分类精确度。针对 KNN 的以上两个不足，本文对其进行了改进。

3.1 基于改进的 K-Medoids 算法的样本裁剪

针对 KNN 相似度计算涉及全部训练文本导致计算效率低的问题，很多研究者都从删减原始训练集合中的对分类贡献小的文本入手，采用聚类算法将训练文本分为 m 个簇，利用待分类文本衡量训练文本的价值，删减对分类贡献小的训练文本后得到新的训练文本集[6,14-15]。这些方法虽能一定程度提高分类速度，但各自都有局限性，如文献[14]的 DBSCAN 聚类算法需要依靠个人经验来确定参数；文献[15]的 K-means 算法存在对噪声和孤立点数据敏感、只有在训练文本集的每个簇的平均值被定义的情况下才能使用；K-Medoids 是一种经典的且应用较为广泛的聚类算法，有着计算简单、快速的特点[16]，且它具有较高的鲁棒性，对于训练样本的裁剪可以弥补 K-means 和 DBSCAN 的不足，但它存在初始化中心点敏感的问题，该算法在进行中心点替换时采用全局替换方法，效率相当低[6]。本文根据文献[17]对 K-Medoids 算法进行初始中心点选择和替换中心点搜索策略的改进，将改进后的 K-Medoids 算法应用到对训练文本的裁剪。

假设训练文本集为 S，S 包含 $C_1,C_2,\cdots,C_N$ 这 N 个类别，共包括文本数为 M。基于改进的 K-Medoids 算法的训练样本裁剪算法步骤如下。

1)初始中心点选择的优化

①对于训练文本集 S，指定其需要划分为 m 个簇，$m=3\times N$；

②为每个簇随机选取一个中心点 O_i（$0<i\leqslant m$）；

③计算训练文本集 S 中剩余非中心点文本与这 m 个中心点的余弦相似度，将它们分配到相似度最大的簇中；

④在每个簇内，以簇内每个点作为中心点，计算它与簇内其他文本的距离和，选择距离和最小的点为新的中心点 O_i；

2)替换中心点搜索策略的优化

①选择一个未选择的中心点 O_i，这是第 j 次迭代（j 从 0 到 m 取值），共进行 m 次迭代。替换中心点集 U 不再是全局非中心点集，而是 O_i 的邻近范围，这个范围是指距中心点 O_i 最近的 j 个簇包含的所有非中心点文本构成的区域；

②在中心点候选集 U 中选择一个未被选择过的非中心点 Q，计算 Q 和 O_i 的平方误差之差，记录在集合 E 中，直到 U 中的所有非中心点都被选择过。

③如果 $\min(E)<0$（集合 E 中最小值小于 0），用集合 E 中最小值对应的非中心点替换原中心点，替换后得到新的 m 个中心点的集合。把剩余的对象分配给相似度最大的中心点所代表的簇，重新从步骤①开始执行；

④如果 $\min(E)>0$ 或 $\min(E)=0$，算法结束，最终得到 m 个聚类中心点；

3)训练样本裁剪

计算待分类文本与 m 个聚类中心的相似度，如果 $Sim(D,O_i)<T_i$（T_i 为第 i 个簇的簇内阈值，即簇内文本与该簇中心点的最小相似度），说明待分类文本与该簇内的文本相似度相当低，所以可以把该簇包含的文本裁剪掉，否则把该簇内包含的文本加入到新的训练文本集。

3.2 代表度函数的定义

在 KNN 算法中的类别属性函数 $y(d_i,C_j)$，对于待分类文本的 K 个最近邻文本“一视同仁”，不考虑它们对所在类别的重要程度，这在一定程度上会影响分类精度。文献[18]定义的隶属度为训练文本到所属类别中心的距离的倒数，用隶属度代替 KNN 中的类别属性函数，相比传统 KNN 算法，分类精度得到了提高。但是隶属度的定义仅仅从两向量的距离因素考虑，忽视了两向量的角度因素。文献[19]定义的类间影响因子来代替 KNN 中的类别属性函数，

它仅仅考虑训练文本到所属类别中心的余弦相似度，忽视了两向量的距离因素。所以，综合考虑两向量之间的距离和夹角的因素从而定义了代表度函数。将定义的代表度函数代替类别属性函数应用到KNN权重计算中。

设训练文本 d_i 的已知类别为 C_j，则将 d_i 对于类别 C_j 的重要程度定义为代表度函数 $u(d_i,C_j)$，如式(4)所示。

$$u(d_i,C_j)=\frac{1}{Dist(d_i,\overline{C}_j)}\times Sim(d_i,\overline{C}_j) \tag{4}$$

式中：$\overline{C}_j$ 表示类别 C_j 中心向量，是将类别 C_j 的所有文本向量相加再求平均。$Dist(d_i,\overline{C}_j)$ 表示训练文本 d_i 到所属类别 C_j 的类别中心的欧式距离，$Sim(d_i,\overline{C}_j)$ 为训练文本 d_i 与所属类别 C_j 的类别中心的余弦相似度，如式(1)所示。由式(4)知，当 d_i 与类别 C_j 中心距离越小，d_i 与类别 C_j 中心的余弦相似度越大，则 d_i 对于 C_j 的代表性越强，即 d_i 对于 C_j 的代表度越大；相反，当 d_i 与类别 C_j 中心距离越大，d_i 与类别 C_j 中心的余弦相似度越小，则 d_i 对于 C_j 的代表性越弱，即 d_i 对于 C_j 的代表度越小。

KNN算法中类别属性函数 $y(d_i,C_j)$ 将用式(5)替换。

$$y(d_i,C_j)=\begin{cases}u(d_i,C_j), & d_i\in C_j\\ 0, & d_i\notin C_j\end{cases} \tag{5}$$

3.3 改进的KNN算法流程

在KNN文本分类算法中，将改进的K-Medoids算法应用到对训练文本裁剪。在KNN算法分类过程中用代表度 $u(d_i,C_j)$ 来代替类别属性函数 $y(d_i,C_j)$。本文改进算法的流程如图1所示。

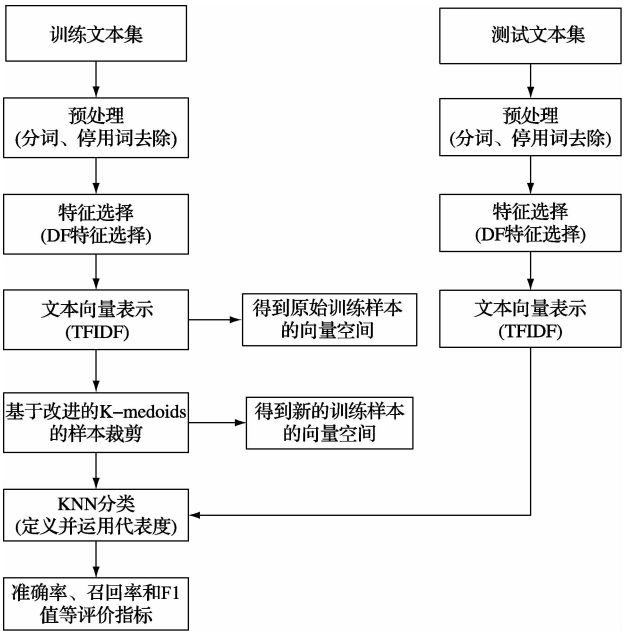

图1 本文改进算法流程

本研究采用中科院开源分词软件ICTCLAS2015对训练文本集和测试文本集进行分词、停用词去除等预处理。采用文档频率DF(document frequency)方法对分词后的训练文本和测试文本进行特征选择，得到每个文本的特征词。将训练文本和测试文本分别根据自己的特征词表示为向量形式，每一维的权重用 $TFIDF=TF\times IDF$ 来计算[20]。按照基于改进的K-Medoids算法的样本裁剪步骤对训练文本集进行裁剪，得到新的训练文本集 S_{new}。对待分类文本进行KNN分类，其中用式(5)替换类别属性函数。最后通过比较KNN算法改进前后的查准率、查全率和 F_1 值，以及时间指标，来验证改进算法的性能。

$$查准率=\frac{分类器在\ C_j\ 上分类正确的文本数}{分类器判别为\ C_j\ 的文本数} \tag{6}$$

$$查全率=\frac{分类器在\ C_j\ 上分类正确的文本数}{C_j\ 真正包含的文本数} \tag{7}$$

$$F_1=\frac{2\times 查准率\times 查全率}{查准率+查全率} \tag{8}$$

4 实验结果与分析

实验环境：Windows 7 32位操作系统，处理器为Intel E7400，内存3 GB和Microsoft Visual Studio 2008开发工具。实验采用的语料库是复旦大学发布的中文语料库，训练集与测试集不重叠。选取的文本类别包括：艺术、教育、历史、法律、交通、政治等20个类别。训练文本集包含文本总数为9 804，测试文本集包含文本总数为9 833。各类训练文本集和测试集包含文本数量如表1所示。本实验的分类效果是通过查准率、查全率以及 F_1 值来评价，速度是通过消耗时间来衡量。

表1 各类训练集和测试集包含文本数

类别	训练集	测试集
艺术	740	742
教育	59	61
历史	466	468
法律	51	52
交通	57	59
政治	1 024	1 026
文学	33	34
哲学	44	45
空间	640	642
能量	32	33
电子	27	28
通信	25	27
计算机	1 357	1 358
矿藏	33	34
环境	1 217	1 218
农业	1 021	1 022
经济	1 600	1 601
医疗	51	53
军事	74	76
运动	1 253	1 254
	9 804	9 833

根据文献[7]，传统的KNN算法在$K=5$时分类效果最好。在K取5,10,15,20,25,30时，运用本文方法分别进行实验，结果显示该算法在$K=10$时效果最佳，所以这里仅给出$K=10$时的结果。把本文算法分别与传统的KNN算法、基于K-Medoids的KNN算法进行比较，实验数据如表2和3所示，表2中给出本文算法与传统KNN算法分别得到最佳分类效果时的查准率、查全率以及F_1值，还给出本文算法与基于K-Medoids的KNN算法在$K=10$时的查准率、查全率以及F_1值，表3给出时间指标。

表2 3种算法实验结果

类别	传统KNN算法($K=5$)			基于K-Medoids的KNN($K=10$)			本文算法($K=10$)		
	查准率/%	查全率/%	F_1值/%	查准率/%	查全率/%	F_1值/%	查准率/%	查全率/%	F_1值/%
艺术	92.77	91.89	92.33	93.26	91.63	92.44	93.53	91.34	92.42
教育	10.45	21.67	14.10	38.89	41.64	40.22	40.12	44.51	42.20
历史	70.23	61.47	65.56	67.65	65.80	66.71	69.54	67.14	68.32
法律	86.56	91.31	88.87	84.96	88.20	86.55	85.23	87.66	86.43
交通	75.32	73.29	74.29	87.24	85.36	86.29	90.65	88.09	89.35
政治	50.23	62.66	55.76	49.32	52.07	50.66	53.21	52.20	52.70
文学	20.65	20.41	20.53	32.33	29.34	30.76	33.42	33.70	33.56
哲学	48.56	50.08	49.31	51.40	53.67	52.51	53.62	58.12	55.78
空间	91.23	87.43	89.29	91.65	88.20	89.89	92.12	89.23	90.65
能量	59.78	63.45	61.56	59.89	63.79	61.78	61.34	65.20	63.21
电子	58.64	54.38	56.43	46.54	42.14	44.23	48.59	44.25	46.32
通信	73.66	77.60	75.58	74.56	78.86	76.65	75.69	80.22	77.89
计算机	95.89	93.05	94.45	95.47	93.37	94.41	96.58	94.52	95.54
矿藏	62.87	64.18	63.52	72.98	75.98	74.45	73.89	77.90	75.84
环境	88.69	85.26	86.94	88.42	86.72	87.56	90.24	87.58	88.89
农业	91.45	93.02	92.23	91.65	95.05	93.32	92.98	96.17	94.55
经济	90.68	88.98	89.82	90.78	84.73	87.65	90.45	86.50	88.43
医疗	60.32	63.63	61.93	56.38	63.48	59.72	58.97	62.01	60.45
军事	65.14	63.46	64.29	71.89	73.02	72.45	77.85	74.47	76.12
运动	82.98	86.39	84.65	83.62	87.59	85.56	85.65	87.49	86.56
平均值	68.81	69.68	69.07	71.44	72.03	71.69	73.18	73.42	73.26

表3 时间性能

时间/min	传统KNN算法	本文算法
训练时间	0	70.65
分类时间	340.45	156.31
总时间	340.45	226.96

从表2可以看出，本文改进的KNN算法与传统KNN算法进行对比，在查准率、查全率以及F_1值上均有提高。在查准率上，平均提高了6.35%，只有在历史、政治、电子及医疗上略有下降；在查全率上，平均提高了5.37%，只有在艺术、法律、政治、电子、经济及医疗上略有下降；在F_1值上，平均提高了6.07%，只有法律、政治、电子、经济及医疗上略有下降。与基于K-Medoids的KNN算法进行比较，在查准率上，各个类别的查准率全面提高，平均提高了2.44%；在查全率上，平均提高了1.93%，除了在艺术、法律、医疗及运动这4个类别上略有下降；在F_1值上，平均提高了2.19%，只有在艺术和法律两个类别上略有下降。从表3可以看出，本文算法与传统KNN算法进行比较，分类时间上提高了1.18倍，总时间提高了50%。

综上所述，本文改进的KNN算法分别与传统的KNN算法、基于K-Medoids的KNN算法进行比较，在查准率、查全率以及F_1上均得到一定程度的提高，时间性能上较传统KNN算法得到显著提升。

5 结 论

KNN是一种实现简单、鲁棒性高的经典分类算法，但它并不是一种高效的算法。一方面，为了降低相似度计算量，本文引入基于改进的K-Medoids聚类算法删减训练文本集，提升了时间性能；另一方面，为了有差别处理待分类文本的K个最近邻文本，本文定义代表度函数$u(d_i,C_j)$代替类别属性函数$y(d_i,C_j)$，提高了分类精度。实验结果表明，本文这两个方面的改进是可行的和有效的。另外，在进行样本裁剪后减少了相似度计算量，但是不可避免地会裁剪掉有用的信息。所以今后的努力方向是研究

更有效的和更有针对性的样本裁剪方法。

参考文献

[1] SEBASTIANI F. Machine Learning in Automated Text Categorizaition[J]. ACM Computing Surveys, 2002,34(2): 1-47.

[2] 谭熊,余旭初,秦进春,等.高光谱影像的多核SVM分类[J].仪器仪表学报,2014,35(2):405-411.

[3] 周靖,刘晋胜.特征联合熵的一种改进K近邻分类算法[J].计算机应用,2011,31(7):1785-1788.

[4] 钱晓东,王正欧.基于改进KNN的文本分类方法[J].情报科学,2005,23(4): 550-554.

[5] 李荣陆,胡运发.基于密度的KNN文本分类器训练样本裁剪方法[J].计算机研究与发展,2004,41(4):539-545.

[6] 罗贤锋,祝胜林,陈泽健,等.基于K-Medoids聚类的改进KNN文本分类算法[J].计算机工程与设计,2014,35(11):3864-3867.

[7] LI B, YU S, LU Q. An improved K-nearest-neighbor algorithm for text categorization[J]. Expert Systems with Applications,2012,39(1):1503-1509.

[8] 王煜,张明,王正欧,等.用于文本分类的改进KNN算法[J].中文信息学报,2007,21(13):159-162.

[9] 刘海博,郗亚辉,王煜.用于文本分类的快速KNN算法[J].河北大学学报:自然科学版,2008,28(3):322-326.

[10] 王煜.基于决策树和K最近邻算法的文本分类研究[D].天津:天津大学,2014.

[11] 石欣,印爱民,张琦.基于K最近邻分类的无线传感器网络定位算法[J].仪器仪表学报,2014,35(10):2238-2247.

[12] 杨林波.快速文本分类研究[D].江苏:江南大学,2008.

[13] 陈小莉.基于信息增益的中文特征提取算法研究[D].重庆:重庆大学,2008.

[14] 苟和平,景永霞,冯百明,等.基于DBSCAN聚类的改进KNN文本分类算法[J].科学技术与工程,2013,13(1): 219-222.

[15] 刘海峰,姚泽清,苏展,等.文本分类中基于K-means的类偏斜KNN样本剪裁[J].微电子学与计算机,2013,29(5): 24-28.

[16] 赵烨,黄泽君.蚁群K-medoids融合的聚类算法[J].电子测量与仪器学报,2012,26(6):800-804.

[17] 夏宁霞,苏一丹,覃希.一种高效的K-Medoids聚类算法[J].计算机应用研究,2010,27(12):4517-4519.

[18] 江涛,陈小莉,张玉芳,等.基于聚类算法的KNN文本分类算法研究[J].计算机工程与应用,2009,45(7):153-155.

[19] 吴春颖,王士同.一种改进的KNN Web文本分类方法[J].计算机应用研究,2008,25(11):3275-3277.

[20] BRUNO T, SASA M, DZENANA D. KNN with TF-IDF based framework for text categorization[J]. Procedia Engineering, 2013,69(1):1356-1364.

作者简介

樊存佳,1990年出生,硕士研究生。主要研究方向为数据挖掘等。

E-mail:fancunjia@emails.bjut.edu.cn

汪友生(通讯作者),1965年出生,工学博士,副教授,主要研究方向为图像处理、数据挖掘等。

E-mail:wangyousheng@bjut.edu.cn