基于改进 KNN 算法的中文文本分类方法

王爱平，徐晓艳，国玮玮，李仿华
（安徽大学 计算智能与信号处理教育部重点实验室，安徽 合肥 230039）

摘 要：介绍了中心向量算法和 KNN 算法两种分类方法。针对 KNN 分类方法在计算文本相似度时存在的不足，提出了改进方案。新方案引入了中心向量分类法的思想。通过实验，对改进的 KNN 算法、中心向量算法和传统的 KNN 算法应用于文本分类效果进行了比较。实验结果表明，改进的 KNN 算法较中心向量法和传统的 KNN 算法在处理中文文本分类问题上有较好的分类效果，验证了对 KNN 算法改进的有效性和可行性。

关键词：文本分类；中心向量法；KNN；相似度
中图分类号：TP39 文献标识码：A 文章编号：1674-7720(2011)18-0008-03

Text categorization method based on improved KNN algorithm

Wang Aiping, Xu Xiaoyan, Guo Weiwei, Li Fanghua
(Ministry of Education Key Laboratory of Intelligent Computing & Signal Processing, Anhui University, Hefei 230039, China)

Abstract: This paper mainly introduces the central vector algorithms and KNN algorithms two classification method. According to KNN classification method in calculating text the shortcomings of the similarity, put out one improved scheme. The new scheme introduces the idea of central vector classification method. At last an empirical study of using the improved KNN algorithm, the central vector algorithm and the traditional KNN algorithm to categorize the Chinese text is conducted. The result of the experiment shows that, compared with central vector algorithm and traditional KNN algorithm, improved KNN algorithm has better categorization effect of the Chinese text, and verify the validity and feasibility of improvement KNN algorithm.

Key words: texts categorization; central vector algorithm; KNN; similarity

由于互联网上可用的文本信息的迅速增长，在信息搜集中，常会有急需查找和组织相关的信息来获得所需要的文本知识，因此文本自动分类技术就变得越来越重要，同时，提高文本自动分类的整体效果也成了一种新的挑战。目前常用的文本分类算法有朴素贝叶斯(Native Bayes)[1]、K 近邻算法 KNN(K Nearest Neighbor)[2]、支持向量机 SVM(Support Vector Machine)[3]等。其中 K 近邻分类算法是一种基于统计的分类方法，具有思路简单、易实现、无需训练过程等优点，因此得到了广泛应用。相关研究证明，K 近邻算法是向量空间模型下最好的分类算法之一。

尽管如此，K 近邻算法仍然存在很多不足，本文针对其中的不足之处提出了改进的方法。

1 基于近邻的分类方法

1.1 中心向量法

中心向量法[4]的基本思想是，根据属于某一类别的所有训练文本向量，计算该类别的中心向量，在进行分类时，计算待分类文本向量与每个类别中心向量的相似度，然后将其归入与之相似度最大的那个类别。该方法也可以看成是 K 近邻分类方法的一种特殊情况，其有效地降低了分类时的开销。类中心向量的求法通常有三种，本文采用如下的计算方法：

将某一类别中所有的文本向量求和得到类中心向量，表示成公式为：

$$C_i=\sum_{k=1}^{n} d_{ik} \tag{1}$$

其中，n 表示类 i 中的文本个数，d_{ik} 表示类 i 中的第 k 篇文本。

1.2 传统的 K 近邻算法

K 近邻[2]分类方法是一种懒惰的、有监督的、基于实例的机器学习方法。该算法的基本思路是，先将训练文本集中的所有文本表示成向量的形式，再将这些文本向

量组成文本向量集并储存起来。当待分类文本到达时，计算这篇文本与训练文本集中每一个文本的相似度，并且将计算得到的值按降序排列，找出排在最前面的 K 篇文本，然后根据这 K 篇文本所属的类别来判断待分类文本的类别。计算文本相似度的方法通常有欧氏距离、向量内积和夹角余弦三种。本文采用夹角余弦计算文本之间的相似度，公式如下：

$$Sim(d_1,d_2)=\frac{\sum_{i=1}^{n}W_{1i}W_{2i}}{\sqrt{\sum_{i=1}^{n}W_{1i}^2\sum_{i=1}^{n}W_{2i}^2}} \tag{2}$$

其中，W_{1i} 和 W_{2i} 分别表示文本 d_1 和 d_2 的文本向量中第 i 个特征项的权重。求出的余弦值越大说明两个文本的相似度越大，两个向量所代表的文本就越可能属于同一个类别，反之，两个向量所代表的文本属于同一个类别的可能性就越小。依据这 K 篇文本将待分类文本进行归类的方法是，对这 K 篇文本与待分类文本的相似度按公式(3)求和，将属于同一个类的文本的相似度求和，然后对每个类所求的和进行排序，将待分类文本分到相似度和比较大的那个类中。

$$T_j(d)=\sum_{i=1}^{k}T_j(d_i)Sim(d,d_i) \tag{3}$$

其中，k 表示选取的文本数，$T_j(d_i)$ 表示文本 d_i 是否属于 C_j 类，如果属于，则值为 1，否则值为 0；$Sim(d,d_i)$ 即为式(2)所求。

2 改进的 K 近邻算法

研究表明，在用相似度计算方法计算两个文本向量之间的相似度时，并没有考虑待分类文本与训练文本所属的类别之间是否存在相似性，因此将所求的结果运用到分类中时可能会导致分类结果的不准确。针对这一不足，本文将中心向量分类方法的思想引入到了相似度计算公式中，对夹角余弦相似度计算公式进行了改进。

2.1 改进的相似度计算公式

首先引入权向量 V_i，V_i 表示类别 i 的权向量，V_{ij} 表示权向量 V_i 的第 j 个特征值，初始化 V_{ij} 为 1。改进的夹角余弦相似度计算公式和相似度求和公式如下：

$$Sim(d_1,d_2,V_i)=\frac{\sum_{j=1}^{n}V_{ij}W_{1j}W_{2j}}{\sqrt{\sum_{j=1}^{n}W_{1j}^2\sum_{j=1}^{n}W_{2j}^2}} \tag{4}$$

$$T_j(d,V_i)=\sum_{i=1}^{k}T_j(d_i)Sim(d,d_i,V_i) \tag{5}$$

2.2 针对文本在每个类别中分布不均匀情况的处理

假如只有 A、B 两种类别，若标记为 A 类别的文本 d_0 被归到 B 类别中，即 $T_A(d_0,V_A)$ 小于 $T_B(d_0,V_B)$，则增大类别 A 的权向量 V_A，减小类别 B 的权向量 V_B，直到 $T_A(d_0,V_A)$ 小于 $T_B(d_0,V_B)$，经过几轮增减操作后，待分类文本 d_0 就会被正确归类。

上述方法克服了每个类别中文本篇数不平均的问题。如果类别 A 是一个包括较多篇数文本的类别，类别 B 是一个包括较少篇数文本的类别，根据传统的 K 近邻分类算法，类别 B 中的样本有可能被归到类别 A 中，则对类别 B 的权向量进行的增加操作的次数就会比类别 A 的权向量的多。经过几轮增减操作后，含有较少篇数文本的类别 B 就可能有较大的权向量 V_B。因此，引入权向量可以克服文本在每个类别中分布不均匀的问题。

2.3 相似度计算方法的改进策略

首先，引入两个向量，类别 i 不可变的中心向量 C_i^u 和可变的中心向量 C_i^a，C_i^u 的计算公式如式(1)，初始化 C_i^a 为 C_i^u，即 $C_{ij}^{a,0}=C_{ij}^u$，其中 C_{ij}^u 表示向量 C_i^u 的第 j 个特征值，$C_{ij}^{a,0}$ 表示向量 C_i^a 的第 j 个特征值，上标中的 0 表示此次正在进行的增减操作，由此可得 $V_{ij}^0=1$。在每次增减操作后，都必须对训练文本集中的所有文本进行分类，如果标记为类别 A 的文本 d_0 被归到 B 类别，就用如下公式来调整 $C_A^{a,0}$、$C_B^{a,0}$、V_A^0 和 V_B^0：

$$C_{A,j}^{a,0+1}=C_{A,j}^{a,0}+\text{increase\_weight}\times W_{0j}\quad W_{0j}>0 \tag{6}$$

$$V_{A,j}^{a,0+1}=\frac{C_{A,j}^{a,0+1}}{C_{A,j}^{u}}\quad W_{0j}>0 \tag{7}$$

$$C_{B,j}^{a,0+1}=\begin{cases}C_{B,j}^{a,0}-\text{reduce\_weight}\times W_{0j} & \text{reduce\_weight}\times W_{0j}<C_{B,j}^{a,0}\\ 0 & \text{reduce\_weight}\times W_{0j}\geq C_{B,j}^{a,0}\end{cases} \tag{8}$$

$$V_{B,j}^{a,0+1}=\frac{C_{B,j}^{a,0+1}}{C_{B,j}^{u}}\quad W_{0j}>0 \tag{9}$$

其中，式(6)和式(7)是进行增加操作的公式，式(8)和式(9)是进行减少操作的公式，increase_weight 和 reduce_weight 是每次进行增减操作的权值。

3 性能评价方法

分类的准确度和速度是评价一种文本分类算法的标准。其中，分类速度取决于分类规则的复杂程度，而分类的准确度主要是参照通过专家思考判断后对文本的分类结果与人工分类结果的相近程度，越相近其分类的准确程度就越高，这里包含了评价文本分类算法的两个指标：准确率(Precision)和召回率(Recall)[5]。由于准确率和召回率分别表示分类效果的两个不同方面，因此通常使用 F_1 测试值统筹评估分类结果[6]。另外有微平均和宏平均两种计算准确率、召回率和 F_1 值的方法[7]。在计算分类的各个评价指标时，先建立如表 1 所示的二值分类列联表。

可用如下的公式计算准确率(Precision)、召回率(Recall)、F_1 值、宏 F_1 值(MacroF_1)和微 F_1 值(MicroF_1)：

$$Precision=\frac{A}{A+B} \tag{10}$$

表 1 二值分类列联表

	属于该类的文本数	不属于该类的文本数
判定为属于该类的文本数	A	B
判定为不属于该类的文本数	C	D

$$Recall=\frac{A}{A+C} \tag{11}$$

在式(10)中，如果 $A+B=0$，则 $Precision=1$，在式(11)中，如果 $A+C=0$，则 $Recall=1$。

$$F_1=\frac{2Precision\times Recall}{Precision+Recall} \tag{12}$$

$$\text{Macro}F_1=\frac{\sum_{k=1}^{N}F_{1k}}{N}\times 100\% \tag{13}$$

$$\text{Micro}Recall=\frac{\sum_{k=1}^{N}A_k}{\sum_{k=1}^{N}A_k+\sum_{k=1}^{N}C_k}\times 100\% \tag{14}$$

$$\text{Micro}Precision=\frac{\sum_{k=1}^{N}A_k}{\sum_{k=1}^{N}A_k+\sum_{k=1}^{N}B_k}\times 100\% \tag{15}$$

$$\text{Micro}F_1=\frac{2\times\text{Micro}Precision\times\text{Micro}Recall}{\text{Micro}Precision\times\text{Micro}Recall}\times 100\% \tag{16}$$

4 实验与结果分析

在实验中，增减操作的权值用来控制每次增减操作的步长，它会影响实验的结果，当把增减操作的权值都设为 1.0 时，进行增减操作可以使基于 K 近邻算法的分类方法达到比较稳定的性能改进。进行增减操作的最大次数也是一个比较难确定的值，但是实验表明，当把增减操作最大次数设为 5 时，可以获得较好的分类效果。

实验数据选取中文语料库中的 4 个类别作为训练文本集，每类文本的篇数不等。改进的 K 近邻算法的分类结果如表 2、表 3 和图 1 所示。

从 2 表可以看出，对于各个类别，使用改进的 K 近邻分类算法后其准确率、召回率和 F_1 值都比使用中心向量法和传统的 K 近邻算法有明显的提高。从图 1 可以看出，如果从整体上评价测试结果，使用传统的 K 近

表 2 改进的 K 近邻算法在各个类上的分类结果

分类方法	指标/%	政治	教育	经济	计算机
中心向量法	准确率	77.5	71.4	79.2	79.6
	召回率	62.0	90.0	76.0	78.0
	F1 值	68.9	79.7	77.6	78.8
传统的 KNN	准确率	83.8	76.3	79.3	80.4
	召回率	62.0	90.0	84.0	82.0
	F1 值	71.3	82.6	81.6	81.2
改进的 KNN	准确率	89.0	90.9	93.9	100.0
	召回率	96.4	95.9	86.1	93.9
	F1 值	92.3	93.3	89.8	96.9

表 3 该技能的 K 近邻算法在整体上的测试结果

分类方法	微 F_1 值/%	宏 F_1 值/%
中心向量法	80.55	81.52
传统的 KNN	81.88	82.35
改进的 KNN	85.44	85.85

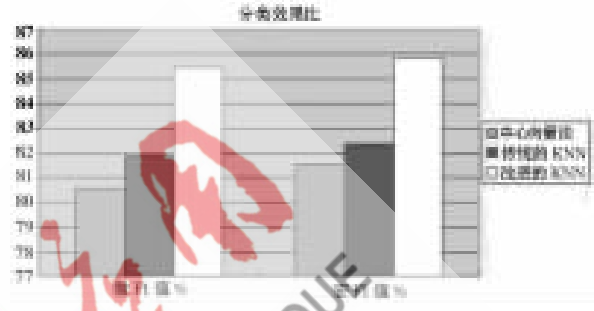

图 1 整体分类效果比较图

邻算法的分类效果在微 F_1 值和宏 F_1 值都比使用中心向量算法提高近 1 个百分点，使用改进的 K 近邻算法的分类效果在微 F_1 值和宏 F_1 值又都比传统的 K近邻算法提高近 3 个百分点。所以，改进的K近邻算法比中心向量算法和传统的K近邻算法有较好的分类效果。

本文提出的改进的 K 近邻算法，与传统的K近邻算法相比，引入了中心向量分类算法的思想，在相似度计算方面进行了改进。从实验结果可以得到，改进的 K 近邻分类算法的分类效果比传统的 K 近邻算法高出 3 个百分点，同时也验证了对算法改进的有效性和可行性。下一步的工作就是通过进一步学习其他的分类算法，尝试将其他的分类算法引入到 K 近邻分类算法中，以达到更高的分类效果。

参考文献

[1] 宫秀军，孙建平，史忠植.主动贝叶斯网络分类器[J]. 计算机研究与发展，2002，39(5)：74-79.

[2] 张 宁，贾自艳，史忠植.使用 KNN 算法的文本分类[J]. 计算机工程，2005，31(8)：171-173.

[3] JOACHIMS T. Text categorization with support vector machines: learning with many relevant features[C].In Proceeding of ECML-98, 10th European Conference on Machine Learning, Berlin：Springer-Ver-lag, 1998：137-142.

[4] 王新丽.中文文本分类系统的研究与实现[D].天津大学硕士研究生论文，2007.

[5] 曹勇，吴顺祥.KNN 文本分类算法中的特征选取方法研究[J].科技信息(科技·教研)，2006(12)：26-28.

[6] 柴春梅，李翔，林祥.基于改进 KNN 算法实现网络媒体信息智能分类[J].计算机技术与发展，2009，19(1)：1-4.

[7] 刘怀亮，张治国，马志辉，等.基于 SVM 与 KNN 的中文文本分类比较实证研究[J].信息系统，2008，31(6)：941-944.

(收稿日期：2011-05-27)

作者简介：

王爱平，女，1956 年生，教授，主要研究方向：数据挖掘、人工智能、编译技术、计算机仿真以及滤波算法收敛性等。

徐晓艳，女，1987 年生，硕士研究生，主要研究方向：数据挖掘。

国玮玮，女，1986 年生，硕士研究生，主要研究方向：数据挖掘。