情报工程 ISSN 2095-915X
第2卷 第6期 8-16 2016年12月

TECHNOLOGY INTELLIGENCE ENGINEERING
ISSN 2095-915X Vol.2 No.6 8-16 Dec 2016

doi:10.3772/j.issn.2095-915x.2016.05.002

基于训练集裁剪的加权K近邻文本分类算法

孙新，欧阳童，严西敏，尚煜茗，郭文浩

（北京理工大学计算机学院北京市海量语言信息处理与云计算应用工程技术研究中心 北京 100081）

摘要：文本分类是信息检索领域的重要应用之一，由于采用统一特征向量形式表示所有文档，导致针对每个文档的特征向量具有高维性和稀疏性，从而影响文档分类的性能和精度。为有效提升文本特征选择的准确度，本文首先提出基于信息增益的特征选择函数改进方法，提高特征选择的精度。KNN(K-Nearest Neighbor)算法是文本分类中广泛应用的算法，本文针对经典KNN计算量大、类别标定函数精度不高的问题，提出基于训练集裁剪的加权KNN算法。该算法通过对训练集进行裁剪提升了分类算法的计算效率，通过模糊集的隶属度函数提升分类算法的准确性。在公开数据上的实验结果及实验分析证明了算法的有效性。

关键词：文本分类，特征选择，信息增益，最近邻分类算法

中图分类号：TP391，TP181

The Weighted KNN Text Categorization Algorithm Based on Training Set Cutting

SUN Xin, OUYANG Tong, YAN XiMin, SHANG YuMing, GUO WenHao

(Beijing Engineering Research Center of Massive Language Information Processing and Cloud Computing Application, School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China)

Abstract: Text categorization is one of the key research fields in the information retrieval. Feature selection is an important part in the document processing, and imposes great influence on the document

基金项目：本文受国家973课题（2013CB329605）的资助。

作者简介：孙新（1975–），副教授，博士，研究方向：人工智能，email：sunxin@bit.edu.cn；欧阳童（1993–），硕士研究生，主要研究方向：机器学习，自然语言处理；严西敏（1993–），硕士研究生，研究方向：人工智能，智能医疗，大数据处理；尚煜茗（1992–），硕士研究生，主要研究方向：云计算，自然语言处理；郭文浩（1993–），硕士研究生，主要研究方向：云计算，自然语言处理。

classification. In this paper, an improved feature selection algorithm based on information gain was proposed to improve the accuracy of text feature selection effectively. Moreover, K-Nearest Neighbor (KNN) algorithm is used widely in text categorization, and the advantages of this method are high accuracy and stability.However, the number of training samples and their position may influence the classification performance of the KNN algorithm, thus we proposed the weighted KNN classification algorithm based on training set cutting, and the accuracy of the classification algorithm can be improved by the rough sets and the concept of membership function. Finally, this research tested the new algorithm based on the text categorization experiment, and the results indicated that the effectiveness of the proposed algorithm.

Keywords: Text categorization, feature selection, information gain, KNN algorithm

1. 引言

文本分类是处理和组织大规模文本数据的关键技术，其主要任务是在预先给定的类别标记集合下，根据文本内容判定它的类别，广泛应用于搜索引擎、快速资料分检、自动文摘、信息资料推送和信息过滤等领域。

目前关于文本分类算法的研究很多，主要有基于规则的决策树方法、基于连接的人工神经网络、基于统计的朴素贝叶斯、K- 最近邻分类算法 (K Nearest Neighbor，KNN) 和支持向量机 (SVM) 等 [1]。KNN 算法 [2] 即 K 近邻算法是一种经典的统计模式识别方法，由最近邻算法 [3] 演化而来的。KNN 算法原理简单易于实现，具有高稳定性和高准确性，并且鲁棒性好，是目前应用最广泛的分类算法之一。但是，训练样本的数量和分布位置很大程度影响了分类的准确程度。因此，如何保持 KNN 性能稳定、准确率高的优势，又能提高分类速度，是 KNN 研究的一个关键难题。近年来学者们提出许多针对 KNN 的改进算法，主要是通过一定的策略来直接减少需要比较的样本数，或通过降低特征维数以减少计算量，从而达到提高分类效率的目的。

此外，特征空间的维数过高也是文本分类面临的主要困难问题之一。特征选择是中文文本分类过程的重要预处理环节，特征选择的效果直接影响文本的分类准确率 [4]。如何降低特征空间的维数，提高分类的效率和精度，成为文本分类中需要首先解决的问题。

目前，文本分类研究中常用的特征选择方法主要有：文档频率 (Document Frequency)、X^2 统计 [5]、信息增益（Information Gain, IG）、互信息（Mutual Information, MI）等 [6]。

文档频率方法 [7] 认为文档集中文档频率低的特征项，其带有的类别信息也应该较少，根据这一原理将文档频率低于一定值的特征项进行剔除，但该方法可能会错误地过滤某个低频特征词、影响精度。X^2 统计算法旨在找到对分类作用相关度较大的特征项，但 X^2 统计算法处理低频特征有其局限性。互信息方法将两个事物之间的相关程度作为判断特征项与类别之间相关性的度量，如果特征项与某类别的互信息量最大，代表该特征项在该类别中出现的频率最多，说明该特征词对该类别具有较高的区分意义 [6]，但互信息方法没有考虑特征词的发生频率，在互信息评估时会倾向选择罕见的特征词。信息增益的方法使用信息增益代表特征项对文本分类的作用大小，信息增益越大说明特征项具有的信息量越大，对分类的意义越大。文献 [7] 通过实验从不同角度比较发现，DF、IG 的出色表现说明高频词汇确实对

文本分类有益，而 MI 性能差的原因是其特征选择倾向于罕见词，信息增益是最有效的特征选择方法之一。信息增益已成为文本分类研究中常用的特征选择算法。

本文首先针对现有的文本分类方法存在生成特征向量维数高、依赖训练集、忽略低频关键词等不足的问题，提出一种改进的信息增益的特征选择算法，抽取高维向量中对分类有重要意义的信息实现降维，减小下一步分类计算的计算量。同时改进的信息增益计算方法提高了特征选择的准确程度，能够在后续文本分类算法中提高分类的准确率。然后，为解决 KNN 方法计算量大、类别标定函数精度不高的问题，提出基于训练集裁剪的加权 KNN 算法，将训练集进行裁剪，提高算法的计算效率；引入类内距离、样本距离和皮尔逊相关系数改进隶属度参数，提高算法的准确度，并给出实验验证。

2 基于信息增益的文本特征选择改进算法

在文本文类中，经过分词、文本特征表示后，文本转化为由特征项构成的多维向量，但文本特征表示后向量维度通常过大。向量空间的高维性和文档表示向量的稀疏性不但增加了分类的时间复杂度和空间复杂度，而且还大大影响到分类的精度。

文本特征选取的方法一般是将特征项进行权重排序，首先需要对特征词进行权重判断，按照一定的策略将对文本分类起到重要作用的特征词赋予较高的权重，将分类作用较小的词赋予较低的权重，然后将特征词按照权重排序，根据一定的选取策略剔除掉某些对分类意义不大的特征项。这样可以降低文本向量的维度，提高分类算法的效率。

文本特征选取中，首先要确定如何对特征项进行权重赋值，然后根据权重制定特征项的剔除策略，保留对文本分类有重要意义的特征项。根据特征选取的选取方式分为特征选择和特征提取 [8]。简单来讲，特征选择方式是指使用一定的策略，从原有的具有大量特征项的特征集中去掉不相关或对分类意义不大、区分度较低的特征项，从而达到降维的目的，提高分类算法的精确度与运算速度。例如原始特征集具有 n 个特征项：$(W_1, W_2, W_3, \ldots, W_n)$，其中选取对分类最有意义的 k 个特征项进行降维处理（k<n），那么通过特征选取后得到 $(W_1, W_2, W_3, \ldots, W_k)$。特征选择是一种使用减少特征项的处理方式。而特征提取方式则是通过映射的方式，例如哈希映射，将一个高维的向量映射到低维的空间中，以达到特征项降维、提高分类计算效率的目的。

信息增益 [9] 是一种基于熵的评估方法，是指某特征项在文本中出现与否所产生的信息熵之差。信息增益代表特征项对文本分类的作用大小，信息增益越大说明特征项具有的信息量越大，对分类的意义越大。IG 考虑在特征项出现和不出现的情况下，特征项在文本所在类别中所占的信息量的多少。前后差别越大所携带的信息量越大，该特征项的分类能力越强，并以此信息量作为特征项的权值进行特征提取。信息增益的计算公式如公式（1）：

$$\mathrm{IG(t)} = -\sum_{i=1}^{N} P(c_i)\log_2 P(c_i) + P(t)\sum_{i=1}^{N} P(c_i|t)\log_2 P(c_i|t) + P(\bar{t})\sum_{i=1}^{N} P(c_i|\bar{t})\log_2 P(c_i|\bar{t}) \quad (1)$$

其中 $P(c_i)$ 为属于类别 c_i 的文档在整个文档集中出现的频率，$P(t)$ 为在文档集中含有特征项 t 的文档的概率，$P(c_i|t)$ 为包含特征项 t 的文档属于类别 c_i 的概率，$P(c_i|\bar{t})$ 为不包含特征项 t 的文档属于类别 c_i 的的概率。

传统的信息增益方法认为信息增益的大小与特征项的分类能力正相关，主要关注点都是基于文档数目，相对于特征项在文本中的信息则关注较少。假设有 A、B 两个特征词，含有特征词 A 的文档和含有特征词 B 的文档分别在某类别中出现的频率相比于其他分类都很大，那么根据传统的信息增益公式计算，这两个特征项的信息增益应该是近似的；但是假设其中一个特征词在文本中大量出现，而另一个特征词在文本中仅出现较少的次数，那么直观的来看，出现频率较大的特征词具有更强的类别代表性，然而根据传统的信息增益计算方法却无法体现出这一特性。

因此可以考虑将特征词在文本中出现的词频作为信息增益的参数。

定义 1 特征词的类内概率：特征词在类别 C 中出现的次数与该类别中词的总数的比值，定义为该特征词的类内概率，即特征词在类别 C 中出现的平均概率为：

$$P(t|C)=\frac{N(t)}{N(C)} \tag{2}$$

通过定义可以得出。一个词的类内概率越高，则这个词的在 C 类内各个文本中出现的比例越大，越具有分类代表性。

在考虑了特征词的类内概率的基础上，假设出现如下情况，两个特征词 t_1、t_2 的类内概率相同，其中 t_1 在少部分文档中出现频率极高，在其他文档中出现频率低；而 t_2 在每个文档中都以相近的概率出现。显然特征词 t_2 比特征词 t_1 更具有类别代表性，因此给出特征词的类内方差的定义，如公式（3）所示。

$$D(t|C)=1+\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[\frac{N(t_i)}{N(C_i)}-\frac{N(t)}{N(C)}\right]^2} \tag{3}$$

类内方差越大说明此特征词在类内分布越不平衡，特征词的类内方差与分类代表性成反比关系。

因此可以得到基于信息增益的特征选择函数的参数计算公式为：

$$\varphi(t)=\frac{P(t|C)}{D(t|C)}=\frac{\frac{N(t)}{N(C)}}{1+\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[\frac{N(t_i)}{N(C_i)}-\frac{N(t)}{N(C)}\right]^2}} \tag{4}$$

根据公式（1）和公式（4）可以得到基于信息增益的特征选择函数计算公式为：

$$\begin{aligned}NewIG(t)&=IG(t)*\varphi(t)\\&=\left[-\sum_{i=1}^{N}P(c_i)\log_2 P(c_i)+P(t)\sum_{i=1}^{N}P(c_i|t)\log_2 P(c_i|t)\right.\\&\quad\left.+P(\bar{t})\sum_{i=1}^{N}P(c_i|\bar{t})\log_2 P(c_i|\bar{t})\right]\\&\quad*\frac{\frac{N(t)}{N(C)}}{1+\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[\frac{N(t_i)}{N(C_i)}-\frac{N(t)}{N(C)}\right]^2}}\end{aligned} \tag{5}$$

利用公式（5）计算特征词的信息增益，并升序排列，根据选定的阈值进行剔除筛选，就可以抽取高维向量中对分类有重要意义的信息、进而实现降维，减小下一步分类计算的计算量。同时改进的信息增益计算方法提高了特征选择的准确程度，能够在一定程度上提高分类的准确率。

3 基于训练集裁剪的加权 K 近邻分类算法

3.1 K 近邻分类算法

KNN 算法基于空间向量模型来度量两个文本之间的相似水平，基本思想是：首先需要将被分类的文本进行空间向量特征表示，将其表示成为向量形式，再在向量空间中取出与被分类样本最相似（距离最近）的 K 个训练样本，已知 k 个训练集样本的类别，认为被分类样本类别一定存在于这 K 个训练样本的类别中，按照训练样本所属分类的多少来对被分类样本进行分类。

由 KNN 的基本原理可以知道，K 的取值将影响到分类的准确性，通常经验的取值策略为 $K \le \sqrt{N}$，N 为训练样本数。KNN 算法的一般步骤如下：

（1）计算测试样本向量与训练集样本向量的距离 sim(X，d_i)（例如余弦距离），其中 d_i 为第 i 个临近的训练样本，X 为测试样本，按照距离的大小进行排序运算。

（2）按照预先取定的 K 值，选取距离最近的 K 个训练样本。

（3）统计取出的 K 个训练样本所属的分类按照公式（6）计算类别从属度。

$$P(X,\ C_i) = \sum_{d_i \in KNN} \mathrm{sim}(X,\ d_i)\, y(d_i,\ C_i) \tag{6}$$

$$y(d_i,\ C_i) = \begin{cases} 1 & d_i \in C_i \\ 0 & d_i \notin C_i \end{cases}$$

（4）取最大的 P(X，C_i) 标示测试样本为类别 C_i。

在进行 KNN 分类算法时需计算所有训练样本与测试样本的距离，寻找与测试样本距离最近的 K 个训练样本。算法结束。

KNN 算法原理简单易于实现，具有高稳定性及较高的准确性，并具有高度的鲁棒性，应用范围广。但 KNN 算法同样具有缺陷，如下：

（1）KNN 算法在计算 K 个近邻集合时需要计算测试样本与全部训练样本的距离，因此计算量大。

（2）KNN 算法在选取近邻点的数量上很大程度影响了分类的准确程度。

（3）KNN 算法只考虑的训练样本在数量上的影响，没有考虑训练样本在测试样本周围的分布情况。

3.2 基于训练集裁剪的加权 KNN 算法

采用 KNN 算法进行分类时，经典的 KNN 算法需要将被测试文本的特征向量与全部训练集文本特征向量进行计算，得出被测试文本与训练集中文本的距离，然后选择出 K 个最近邻的训练样本，将训练样本所属分类个数进行统计，最后根据数量最大的类别对被测试样本进行标定。其中将被测试样本的特征向量与整个训练集中文本的特征向量进行计算是 KNN 算法效率低下的关键因素，因此对训练集样本进行裁剪，可以有效的降低计算量，提升算法效率；其次为了避免因受被测试文本特征向量所处空间周围训练样本特征向量密度影响，优化分类标定策略可以提升分类的准确程度。

通过以上问题的分析，本文提出基于训练集裁剪的加权 KNN 分类算法。首先给出训练集的裁剪方法，然后引入皮尔逊相关系数衡量文本之间相似度的大小进行权重标注，最后给出算法的

详细描述。

(1) 训练集的裁剪方法

首先给出以下定义：

定义 2 类别区域中心：通过将文本进行特征表示，可以将文本表示为多维空间内的一个点，在一个类别中的所有该类别特征向量都存在于一个高维区域中，因此定义高维区域的中心作为该类别的区域中心。计算公式为：

$$O_i = \frac{\sum_{x=1}^{n_i} d_x(w_1, w_2, \dots, w_k)}{n_i} \quad (7)$$

定义 3 区域最大半径：在确定类别的区域中心后，可以认为属于该类别的训练样本的特征向量都分布于以区域中心为球心的超球体中，因此可以定义区域最大半径为以该球心为中心距离球心最远的点的距离。计算公式为：

$$r_i = \max\{dis(d_{1\to x}, O_i)\} \quad (8)$$

其中$dis(d_{1\to x}, O_i)$为某类别中各个点到区域中心的距离。

被测试样本到类别之间的距离定义为被测试样本到区域中心的距离与区域最大半径之差：

$$dis(t, C_i) = dis(t, O_i) - r_i \quad (9)$$

进行分类时，如果某些点具有极大的相似性，在空间中将呈现出这些点汇聚在一起，可以认为它们形成一个超球体。根据以上原理，测试文本的空间分布一定会落在其应该被分类的超球体内部或附近，或者说距离特定类别超球体的距离比其他不相关类别的超球体更近。

依据这个思想，本文提出训练集的裁剪方法：在构建分类器时，保存各个类别的区域中心（超球体的球心）坐标和类别区域（超球体）最大半径，将其设定为分类器底层结构。在进行下一步分类时首先计算测试样本到类别的距离，将测试样本到类别之间距离升序排列，依次从距离测试样本最近的类别中找出距离测试样本空间特征向量坐标最近邻的 K 个点更新训练集，并将新的训练集根据距离测试样本的距离升序排序，逐次迭代，直到在新的类别中没有比训练集中的点距离测试样本更近的点。

在进行训练集裁剪之后，不需要计算被测试文本到所有训练集文本的空间特征向量距离，只需要计算个别按照空间超球体划分的类别中某些坐标的距离，因此能够有效的减少计算量。

(2) 引入皮尔逊相关系数进行权重标注

针对测试样本受空间内点密度影响的问题，本文将 KNN 算法中的分类计算方法进行改进，根据模糊集的概念对新的训练集中样本空间特征向量进行权重标注。

首先介绍模糊集的概念 [10]。模糊集用于表达模糊性，又被称为模糊子集。模糊集是区别于普通集合的概念，普通集合是具有共性的一类对象的集合体，元素之间的共性是明确的，因此判断元素是否属于某个集合也是确定的，既是或否。但是在实际应用中有许多的概念并不是确定的，因此出现了模糊集理论来描述这一不确定性。模糊集的定义为：在域 U 上，对于任意的 x ∈ X，假设 A 是集合 X 到区间 [0,1] 的一个映射，那么 x → A(x)，则称 X 是 A 上的模糊集，其中 A(x) 为 x 对模糊集 A 的隶属度 [11]。

在文本特征向量空间中，越靠近类别区域中心的点越具有该类别的特征，换句话说为越靠近区域中心的点的类别确定性就越高；相反越在区域边缘的点对于其所处类别的确定性或关联性越低。因此引入模糊集的隶属度来度量这个不确定性，将 KNN 算法中的类别属性函数加入隶属度参数作为权重。

由前文阐述的理论同样可以推知，如果测试样本的空间特征点越靠近训练集中样本的空间特征点，那么这两个文本之间就越相似，这两个文

本具有更加相似的属性特征，那么可以认为测试样本特征点与训练样本特征点之间的距离也一定程度上反映了文本的隶属度关系。

只考虑空间特征向量两个点之间的距离，既只考虑多特征文本在同一属性（单一维度）的差异而求得的距离并不能完全说明两个文本之间相似度的大小，同时还需要综合考虑不同属性值对类别判断的作用大小，因此引入皮尔逊相关系数 [12] 来评价这一参数。皮尔逊相关系数可以用来衡量变量之间的相关度，取值范围是 [−1,1]。相关系数的绝对值越大，那么两个量之间的相关度越高，如果这两个量线性相关，则相关系数变成 −1 或 1。两个变量之间的相关系数越高，从一个变量去预测另一个变量的精确度就越高，这是因为相关系数越高，就意味着这两个变量的共同部分越多。

皮尔逊相关系数计算公式为：

$$\rho_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2}\sqrt{\sum_i (y_i - \bar{y})^2}} \quad (10)$$

综上所述本文提出的隶属度函数为：

$$\text{membership}(t, d_i, C_i) = \frac{1}{dis(d_i, C_i)} * \frac{1}{dis(t, d_i)} * \rho_{td_i} \quad (11)$$

因此根据 KNN 算法得到改进后的类别属性函数为：

$$P(X, C_i) = \sum \text{membership}(t, d_i, C_i) \quad (12)$$

最终根据类别的类别属性函数值，选择最大的值的类别进行被测试文本的类别标注。

基于训练集裁剪的加权 KNN 算法的具体步骤如算法 1 所示：

算法 1 基于训练集裁剪的加权 KNN 算法

Step1：保存类别区域中心、类别区域最大半径，计算测试样本到类别的距离按照升序排序。

Step2：取出测试样本到类别边距最小的区域索引。

Step3：计算测试样本到所选区域中各个点的距离，按照升序排序。

(a) 取前 K 个作为被测试文本的当前 K 近邻集合。

(b) 如果该类别中点的个数小于 K 那么将所有点加入当前 K 近邻集合。

记录下来当前 K 近邻集合的最大距离 disMax，并将该类从被测试样本到所选区域距离集合中剔除。

Step4：(a) 若测试样本到所选区域距离集合的最小值小于 disMax，则代表在该类别中可能存在到被测试样本的距离比当前 K 近邻集合中的点更近的点，或当前 K 近邻集合中点的个数小于 K，那么计算被测试文本到所选区域中各个点的距离，按照升序排序，再与当前 K 近邻集合做并集，然后更新当前 K 近邻集合。

(b) 如果被测试样本到所选区域距离集合的最小值大于 disMax，那么认为其他区域中不会有更近的点，因此当前 K 近邻集合成为 K 最近邻集合。

Step5：根据公式（12）计算被测试数据的分类进行标定。

4 实验结果分析

对于分类算法的分类准确程度通常使用准确率和召回率作为评估标准。召回率是指分类后，

对于正确的分类结果数占分类结果总数的百分比。召回率越高则代表该算法在该类别上所遗漏的样本数越少。准确率是指经过分类算法分类后，分类正确的样本数与分类结果中属于该类的样本数的比值。准确率越高则代表此分类算法出错的可能越小。

本文采用准确率作为评价分类结果的指标，分成两种情况进行对比试验。选取 UCI 公共数据集中的 Heart Disease 数据集中的克利夫兰数据集进行测试。图 1 为数据集中部分数据，数据共 14 列，前 13 列为用于分类的属性，第 14 列为类别预测值。

样本分为 A、B 两组，A 组中训练集样本均匀分布，B 组中训练集样本由程序生成随机数随机选取。

A 组训练集样本中每种类别数据各选取 50 条数据作为训练样本，选取三类数据集中剩余数据中随机选取各 30 条数据作为测试样本。

B 组训练集样本中每种类别数据各随机选取 30−80 条数据作为训练样本，经过随机程序选取后类别 1 选取 42 条，类别 2 选取 67 条，类别 3 选取 60 条，类别 4 选取 58 条选取三类数据集中剩余数据中随机选取各 30 条数据作为测试样本。

age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	class
67	1	4	160	286	0	2	108	1	1.5	2	3	3	2
67	1	4	120	229	0	2	129	1	2.6	2	2	7	1
37	1	3	130	250	0	0	187	0	3.5	3	0	3	0
41	0	2	130	204	0	2	172	0	1.4	1	0	3	0
56	1	2	120	236	0	0	178	0	0.8	1	0	3	0
62	0	4	140	268	0	2	160	0	3.6	3	2	3	3
57	0	4	120	354	0	0	163	1	0.6	1	0	3	0
63	1	4	130	254	0	2	147	0	1.4	2	1	7	2
53	1	4	140	203	1	2	155	1	3.1	3	0	7	1
57	1	4	140	192	0	0	148	0	0.4	2	0	6	0
56	0	2	140	294	0	2	153	0	1.3	2	0	3	0
56	1	3	130	256	1	2	142	1	0.6	2	1	6	2
44	1	2	120	263	0	0	173	0	0	1	0	7	0
52	1	3	172	199	1	0	162	0	0.5	1	0	7	0
57	1	3	150	168	0	0	174	0	1.6	1	0	3	0
48	1	2	110	229	0	0	168	0	1	3	0	7	1
54	1	4	140	239	0	0	160	0	1.2	1	0	3	0
48	0	3	130	275	0	0	139	0	0.2	1	0	3	0
49	1	2	130	266	0	0	171	0	0.6	1	0	3	0
64	1	1	110	211	0	2	144	1	1.8	2	0	3	0
58	0	1	150	283	1	2	162	0	1	1	0	3	0
58	1	2	120	284	0	2	160	0	1.8	2	0	3	1
58	1	3	132	224	0	2	173	0	3.2	1	2	7	3
50	0	3	120	219	0	0	158	0	1.6	2	0	3	0
58	0	3	120	340	0	0	172	0	0	1	0	3	0
66	0	1	150	226	0	0	114	0	2.6	3	0	3	0

图 1 Uci 数据集中 Heart Disease 数据集部分数据

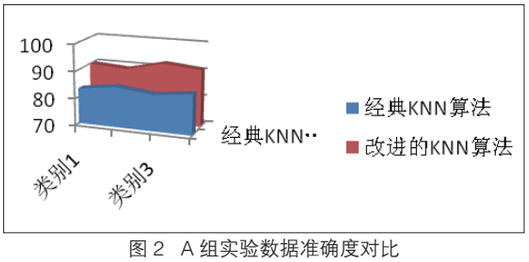

图 2　A 组实验数据准确度对比

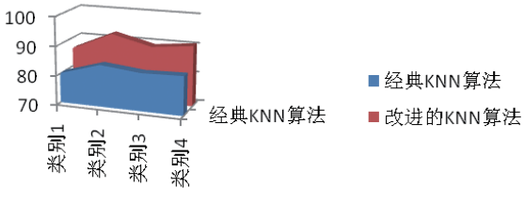

图 3　B 组实验数据准确度对比

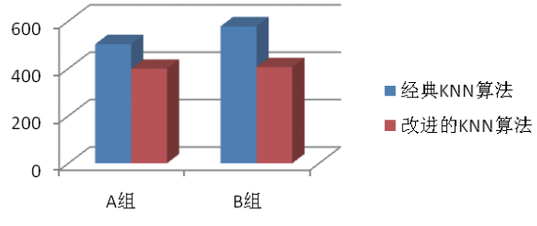

图 4　AB 组实验数据消耗时间对比

根据实验结果图 2、3、4 所示，基于训练集裁剪的加权KNN算法通过对训练样本进行裁剪，并且利用模糊集理论改进类别属性函数，其计算结果不论是从分类的准确度还是从分类的效率方面都优于经典的 KNN 算法。

5 结语

文本分类中需要将分词后的文本进行特征表示，将文本信息表示为空间特征向量。为了降低向量空间矩阵的维数，以减轻分类、聚类、信息挖掘等后续工作的计算压力，需要对特征提取结果进行降维。为此，本文首先提出了基于信息增益的文本特征选择改进算法，将对分类意义不大的、带有噪声的特征项去除，引入类内概率和类内方差参数，提高了特征选择的计算精度。

然后，针对 KNN 方法计算量大、类别标定函数精度不高的问题，提出基于训练集裁剪的加权 KNN 算法，根据模糊集的概念对新的训练集中样本空间特征向量进行权重标注，将训练集进行裁剪，显著提高了算法的计算效率，引入类内距离、样本距离和皮尔逊相关系数改进隶属度参数，有效提高了算法的准确度。对典型样本的分类实验结果验证了改进算法的可行性、有效性，分类效果得到较好的改善。

参考文献

[1] 路永，张宇楠 . 中文文本分类中基于和声搜索算法的特征选择方法 [J]. 情报学报 , 2015, 34(11): 1203-1213.

[2]Yu L, Liu H. Efficient Feature Selection via Analysis of Relevance and Redundancy[J]. Journal of Machine Learning Research, 2004, 5(12):1205-1224.

[3]Soucy P, Mineau G W. A Simple KNN Algorithm for Text Categorization[C]// IEEE International Conference on Data Mining. 2001:647.

[4] 徐燕，李锦涛，王斌，等 . 文本分类中特征选择的约束研究 [J]. 计算机研究与发展 , 2008, 45(4):596-602.

[5] 张爱华，荆继武，向继 . 中文文本分类中的文本表示因素比较 [J]. 中国科学院大学学报 , 2009, 26(3):400-407.

[6]Liu H, Sun J, Liu L, et al. Feature Selection with Dynamic Mutual Information[J]. Pattern Recognition, 2009, 42(7):1330-1339.

[7]Yang Y, Pedersen J O. A Comparative Study on Feature Selection in Text Categorization[C]// Fourteenth International Conference on Machine Learning. Morgan Kaufmann Publishers Inc. 1998:412-420.

[8] 林少波，杨丹，徐玲 . 基于类别相关的新文本特征提取方法 [J]. 计算机应用研究 , 2012, 29(5):1680-1683.

[9]Quinlan J R. C4.5：Programs for Machine Learning [M]. San Mateo, Morgan Kaufmann Publishers Inc, 1993.

[10]Pawlak Z. Rough Set[J]. International Journal of Computer & Information Sciences, 1982, 11(5):341-356.

[11]Szczuka M S, Kryszkiewicz M, Ramanna S, et al. Rough Sets and Current Trends in Computing[J]. Lecture Notes in Computer Science, 2010, 6086(5):xx,853.

[12] 盛骤，谢式千，潘承毅 . 概率论与数理统计 (第四版)[M]. 北京：高等教育出版社 , 2008.