基于 KNN 与自动检索的迭代近邻法在自动分类中的应用[1)]

杨建良 王永成

(上海交通大学计算机系,上海 200030)

摘要 本文研究了一种基于 KNN 与自动检索的自动分类算法——迭代近邻法(Iterative KNN, I-KNN),用以解决 KNN 算法在小样本库的环境下分类效果不佳的问题。在无法得到足够的定类样本时,通过检索的方法将待分样本的局部主题特征放大,进而得到足够定类的相似样本。实验证明,迭代近邻法既增加了获取相似样本的几率,同时也有效地控制了样本相似度条件限制放宽后可能引入的分类噪声,在实际应用中能较好地提升自动分类系统的查全率和查准率。

关键词 KNN 自动分类 自动检索 迭代近邻法

Application of Iterative-KNN Based on KNN and Automatic Retrieval in Automatic Categorization

Yang Jianliang and Wang Yongcheng

(*Shanghai Jiaotong University*, *Shanghai* 200030)

Abstract In this paper, an approach called I-KNN, based on KNN and automatic retrieval, is proposed in order to improve the performance of KNN algorithm with a small-scale document database. If we cannot get enough similar documents in such condition, we magnify part of the document subjects to find more similar documents by Internet retrieval. It is proved by experiment that I-KNN can improve the recall and precision of automatic categorization system.

Keywords KNN, automatic categorization, automatic retrieval, iterative KNN.

1 概 述

随着网络的普及,"信息过载"现象日益突出:人们能获取的信息资源总量非常大,但其中无用信息泛滥,从而难以提取出真正有效的信息资源[1]。作为一种有效的信息处理方法,自动分类技术将信息按照一定的体系进行分类整理,从而可大大提高用户搜集信息的效率。

目前,比较常用的自动分类算法有 Bayes 法[2],KNN 法,SVM 法[3],LLSF 法[4],VMS 法[5]等。其中,KNN 方法以其对分类体系独立性强的优势,在自动分类领域有着较多的应用。但是,KNN 方法最大的缺陷之一在于其对训练样本库的容量要求比较大,因此不适用于小样本情况下的自动分类。本文提出了一种基于 KNN 与自动检索的迭代近邻法,使得采用该方法的分类系统能较好地应用于小样本库的情况。另外,本文还通过实验证明了该改进的 KNN 算法在实际应用中能较好地提升自动分类系统的查全率和查准率。

收稿日期:2003 年 4 月 22 日

作者简介:**杨建良**,男,1979 年生,硕士研究生,主要研究方向为信息自动分类。**王永成**,男,1939 年生,博士生导师,主要研究方向为网络智能信息处理。

1) 国家自然科学基金资助项目(60082003)。

2 传统的KNN法

2.1 KNN法的基本思想

KNN法即K最近邻法,最初由Cover和Hart于1968年提出[6],是一个理论上比较成熟的方法。该方法的思路非常简单直观:根据传统的空间向量模型,文本的内容被形式化为特征空间中的加权特征向量,即 $D = D(T_1, W_1; T_2, W_2; \cdots; T_n, W_n)$。文档向量中的各个维度分别对应于用于表征文档的各个特征属性[7]。如果一个样本所在特征空间中的K个最相似(即特征空间中最邻近)的样本中,大多数样本属于某一个类别,则该样本也属于这个类别。该方法在定类决策上只依据最邻近的一个或者几个样本的类别来决定待分样本所属的类别。

2.2 KNN法的一般过程

(1) 主题特征抽取

对待分样本进行降维处理,通过主题特征抽取将待分样本表示为加权特征向量的形式。

(2) 相似样本选择

在特征向量空间中,两个空间向量的相似度可以用两个向量之间的点积来计算获得。对于待分样本,需要计算其与训练样本库中每一个训练样本之间的相似度。经过排序,可以获得前K篇训练样本作为所选取的相似样本。

(3) 相关类别的选取

在得到 $N(N \leqslant K)$ 个与待分样本最相似的训练样本后,将对这N个训练样本所涉及的所有M个类别分别进行类别相关度计算。经过对类别相关度排序,可以根据分类系统的实际需求选择 $m(m \leqslant M)$ 个类别作为最终的分类结果输出。

2.3 KNN的优点

KNN方法虽然从原理上也依赖于极限定理,但在类别决策时,只与极少量的相邻样本有关。因此,采用这种方法可以较好地避免样本的不平衡问题[1]。从分类过程来看,KNN最直接地利用了样本和样本之间的关系,减少了类别特征选择不当对分类结果造成的不利影响,可以最大程度地减少分类过程中的误差项。另外,对于一些类别特征不明显的类别而言,KNN法更能体现出其分类规则独立性的优势,使得方便快捷的分类自学习的实现成为可能。

3 与检索相结合的I-KNN法

3.1 传统KNN法的不足之处

传统的KNN方法有着两点明显的不足之处:

(1) 样本相似度计算量较大

每一个待分样本都需要计算其与训练样本库中所有样本的相似度,才能求得与其最近邻的K个样本。对于一个有上百万训练样本的样本库的系统而言,庞大的计算量将阻碍分类速度达到用户的实际需求。对于这个问题,一种方法是采用其他的快速分类方法对待分样本进行预分类,将训练样本限制在一定的类别空间中,从而减少无用的训练样本之间的相似度计算。另外还可以采用主题概念链倒排的方法,降低KNN算法的时间复杂度,提高实际分类效率。

(2) 样本库容量依赖性较强

样本库容量依赖性较强的问题对KNN法在实际应用中的限制更大:有不少类别无法提供足够的训练样本,使得KNN算法所需要的相对均匀的特征空间条件无法得到满足。当表示待分样本的特征向量落在较为稀疏的训练特征向量空间中时,就可能产生以下的问题:如果将样本的相似性阈值定得相对较高的话,那么能找到足够定类的相似样本的概率就相对减小;如果将样本的相似性阈值定得相对较低的话,那么最终的分类结果的误分率会大大增加。如图1和图2所示,其中黑色十字代表与待分样本实际类别所匹配的训练样本,白色星号表示与待分样本实际类别不匹配的相似性噪声样本。

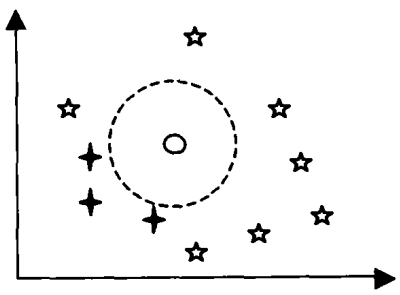

图1 阈值较高导致相似样本不足

3.2 迭代近邻法的基本思想

在实际应用中,特别是在自动分类系统的使用初期,样本库的容量不可能一下子达到KNN算法理论上所需要的"样本数量较大,分布相对均匀"的要

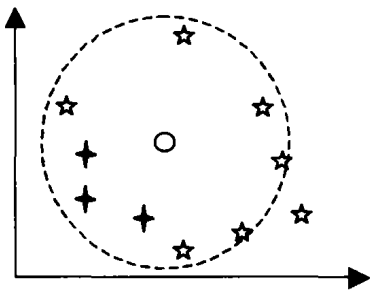

图 2　阈值较低导致相似性噪声过多

求。因此，我们采用一种扩展的 KNN 法——迭代近邻法(Iterative KNN, I-KNN)对上述的问题进行改进。

迭代近邻法的基本思想是：首先在相似样本的选择上使用较高的相似性阈值，以减少相似性噪声对分类结果的影响；如果在较高阈值范围内无法获得足够定类的训练样本的话，则把待分样本中抽取出的主题概念作为检索项，以自动检索的方式从 Internet 上获取多个相似样本；然后对这些相似样本进行 KNN 分类得到二次近邻样本，并以这些二次近邻样本的分类结果决策出原待分样本的最终分类结果。该扩展的 KNN 分类过程与传统的 KNN 法(即一次近邻法)过程可以分别用图 3 和图 4 来表示。图中白色圆点表示待分样本，黑色圆点表示通过 Internet 检索得到的相似样本，黑色十字代表与待分样本实际类别所匹配的训练样本，白色星号表示与待分样本实际类别不匹配的相似性噪声样本。

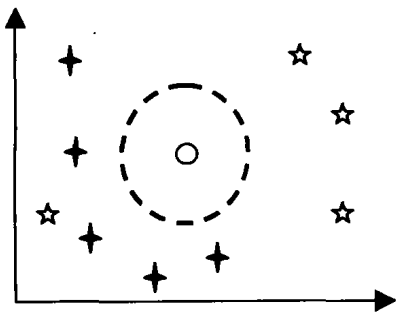

图 3　采用一次近邻法对相似样本进行选择

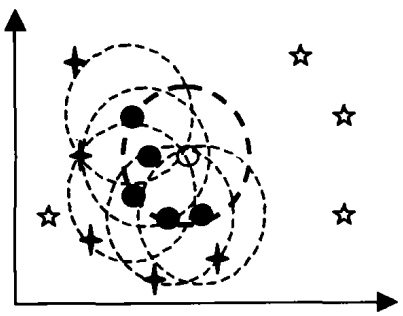

图 4　采用迭代近邻法对相似样本进行选择

从图中可以看到，与传统的 KNN 法相比，I-KNN 法通过对 Internet 的自动检索，补充了多个相似样本。考虑到 Internet 可以被视为一个开放性的大容量数据库，其样本容量远远大于本地的样本库，因此通过自动检索并过滤后所得到的相似样本与待分样本之间有着更高的相似度，所以可以采用这些相似样本来对待分样本进行定类决策。但这些相似样本与一次近邻样本(训练样本)不同，它们没有事先确定的类别，不能作为直接的定类依据。因此，在获得这些相似样本后，还需要对这些样本进行 KNN 分类得到二次近邻样本，然后根据二次近邻样本的 KNN 分类结果来决策出原有待分样本的最终类号。这里需要考虑的一个问题是，当对相似样本进行 KNN 分类的时候，也可能无法获得足够的定类样本，那么就需要对该相似样本进行迭代的二次近邻选取过程，以确定其类别。不过在实际应用中，考虑到分类速度和系统的稳定性，一般不考虑使用迭代的二次近邻选取，而是采用多个相似样本分类结果投票的方式来决定最终的分类结果。

从样本本身的角度来分析，迭代近邻法的实质就是：在样本库中难以找到与待分样本的主题特征完全匹配的相似样本时，采用检索的方法对待分样本的最重要的主题特征进行“局部放大”，然后再对“局部放大”后的主题特征重新在样本库中检索。这样既增加了获取相似样本的几率，同时也有效地控制了样本相似度条件限制放宽后可能引入的分类噪声。

3.3　迭代近邻法的基本过程

I-KNN 的基本过程如图 5 所示。

3.4　自动检索的实现

在二次近邻样本选取的过程中，我们采用 Google 的网页检索的方法来获取相似样本：

(1) 选取权值最高的 N 个主题概念作为检索项；

(2) 根据检索项和 Google 的检索 CGI 接口重构检索 URL；

(3) 分析检索结果页面，得到 M 个相似性页面的 URL；

(4) 根据 M 个 URL 获得对应的相似性页面的内容，经 HTML 预处理后得到相似样本。

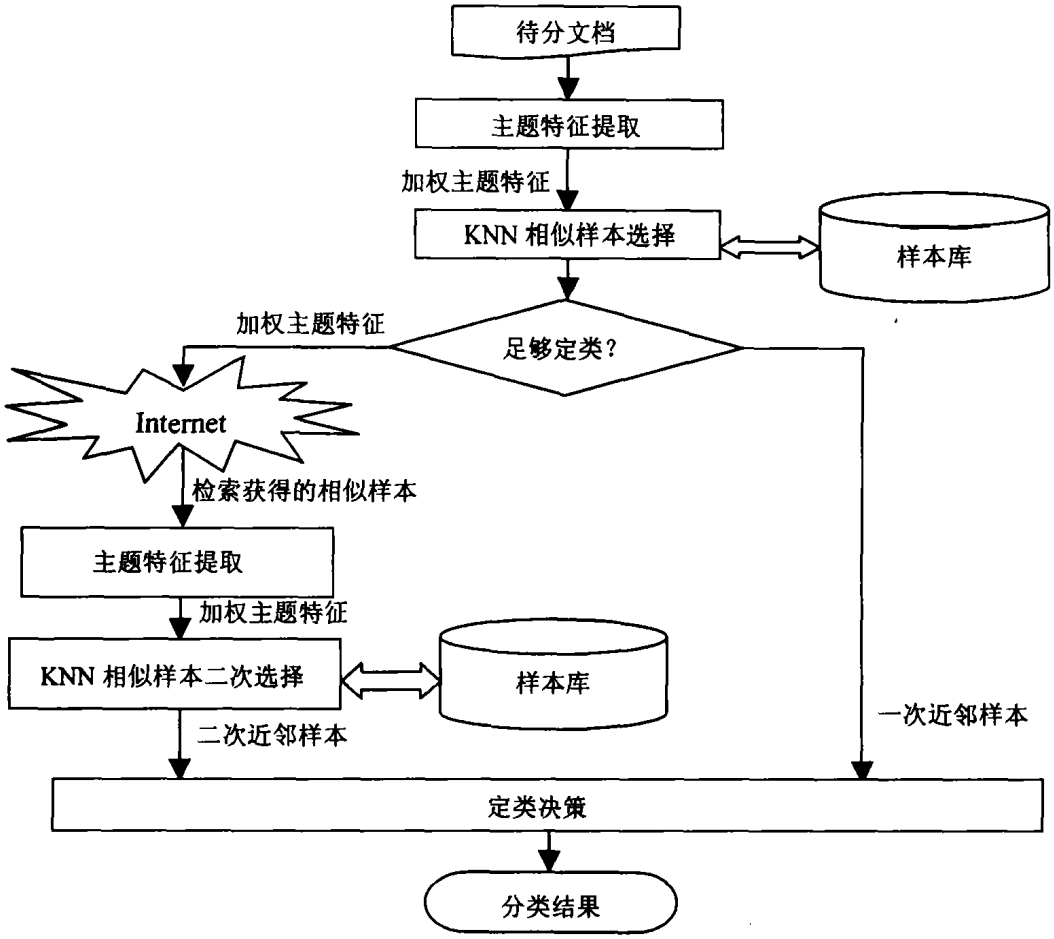

图 5 迭代近邻法基本过程

4 实验结果及分析

4.1 实验方法介绍

本次实验针对的是网上的新闻,根据预设的分类体系分为 10 个大类 30 个子类。从新浪、搜狐及中华网上获取了 3075 篇新闻,其中 75 篇作为开放测试样本,另外 3000 篇作为训练样本。为了测试小样本条件下迭代近邻法的分类效果,我们根据需要将训练样本库的容量预设为 600 篇,然后逐次增长至 1500 篇和 3000 篇。

4.2 实验结果及分析

表 1 KNN 算法和 I-KNN 算法在不同小样本容量环境下的比较

	600 篇训练样本	1500 篇训练样本	3000 篇训练样本
KNN 查全率	30.67	52.00	69.33
I-KNN 查全率	54.67	66.67	77.33
KNN 查准率	65.71	73.58	81.25
I-KNN 查准率	68.33	74.62	81.69

从表 1 可以看到,无论是在何种样本容量的小样本库环境下,采用迭代近邻法的分类系统对整体的查全率和查准率都有不同程度的改善。

从系统的查全率来看,在小样本库环境下采用迭代近邻法能有较大程度的提高:在 600、1500 和 3000 篇训练样本的环境下,系统的查全率分别有 78.25%,28.21%和 11.54%的提升。从系统的查准率来看,改进前后的整体查准率非常接近,采用迭代近邻法的系统略高于原有系统。这说明通过迭代近邻法从网上采集的样本在特征空间中的分布情况与待分样本非常相近,因此采用从网上获取相似样本的方法是可行的。

5 总 结

实验证明,本文所提出的基于 KNN 和自动检索的优化分类算法——迭代近邻法能使自动分类系统的整体查全率和查准率有一定的提高,特别是在训练样本数量不足的环境下优化效果更为明显。采用该优化算法的自动分类系统在小样本情况下的实际

应用能力得到了显著的增强,并在 3000 篇样本容量的环境下达到了可实际应用的程度。

参 考 文 献

1 尹中航. 网络新闻智能分类技术的研究与实现:[学位论文]. 上海:上海交通大学计算机系,2002

2 Mitchell T.. Machine Learning. McGraw: Hill, 1996.

3 Vapnik V.N.. The Nature of Statistical Learning Theory [M]. NY: Springer-Verlag, 1995

4 Yang Y.. Expert network: effective and efficient learning from human decision in text categorization and retrieval. Proceedings of the Fourth Annual Symposium on Document Analysis and Information Retrieval(SIGIR'94), 1994. 13 ~ 22

5 G. Salton, M.E. Lesk. Computer evaluation of indexing and text processing. Journal of the ACM, 1968: 15 (1), 8 ~ 36

6 T.M. Cover, P.E. Hart. Nearest neighbor pattern classification. IEEE Trans. on Inf. Theory, 1967, IT-13: 21 ~ 27

7 Kwok-Yin Lai, Wai Lam. Automatic Textual Document Categorization Using Multiple Similarity-Based Models. SDM' 2001, Nov. 2001.

(责任编辑　许增棋)