一种新的基于密度的 k－最近邻文本分类器训练样本约减方法

A New Density－Based Method for Reducing the Amount of Training Samples in k－NN Text Classification

徐义峰　（衢州学院信息与电子工程系　浙江衢州 324000）
陈春明　（桂林电子科技大学图书馆　广西桂林 541004）
徐云青　（衢州学院信息与电子工程系　浙江衢州 324000）

摘　要：本文针对 k－最近邻方法分类效率不高的问题，提出了一种基于密度的训练样本集约减算法。该方法通过计算训练样本集中各类别的类别密度及整个训练集的平均密度，去掉高密度类别中的部分样本，使训练样本集具有更好的代表性。实验表明，该方法不仅提高了 k－最近邻方法的分类效率，而且对其分类准确率也有一定程度的提高。

关键词：文本分类　k－最近邻方法　训练样本

1　引言

k－最近邻方法[1]（k－Nearest Neighbor，k－NN）作为一种基于统计的简单、有效、非参数的分类方法，在文本分类中得到广泛使用，并取得了很好的效果。其基本思想是在训练样本中找到测试样本的 k 个最近邻，然后根据这 k 个最近邻的类别来决定测试样本的类别。k－最近邻方法是一种基于要求的或懒惰的学习方法，它存放所有的训练样本，直到测试样本需要分类时才建立分类，这样与测试样本比较的可能近邻数量（即训练样本个数）较大时，会有很大的计算代价，训练文本分布的不均匀也会造成分类准确率的下降。

目前主要通过两种途径来减小 k－最近邻方法的计算量：一种途径是设计快速搜索算法，在尽量短的时间内找到测试样本的最近邻[2,3]；另一种途径是在原来的训练样本集中选取一些代表样本作为新的训练样本，或删除原来的训练样本集中的某些样本，将剩下的样本作为新的训练样本，从而达到减小训练样本集的目的[4,5,6,7,8]。但是这些方法在训练样本集中每增加或删除一个样本时，都要对样本进行一次测试，反复迭代直到样本集不再变化，这对于有成百上千甚至上万的训练文本来说，其工作量是非常大的。文献[9]提出了一种基于密度的 k－近邻文本分类器训练样本裁剪方法，针对整个训练样本集，根据训练样本的分布密度对其进行裁剪，使训练样本的分布尽量均匀。这种方法不仅减少了训练样本的数量，使 k－最近邻方法的计算量降低，而且削弱了训练样本分布的不均匀性对分类性能的影响。但这种方法需要计算整个训练样本集中任意两个样本间的距离（或相似度），其计算复杂度为 O(n2)，而且还需要确定三个参数（MinPts、LowPts、ε）的值。为此，本文提出一种新的基于密度的更为简单的样本约减方法。

本文所给方法的基本思路是：首先计算训练集中每类文本的密度及整个训练集的平均密度，对于类别密度大于平均密度的类别，去掉一些样本，从而减少训练集中样本的个数。去掉样本时，每次都是选择相似度最高的两个样本中的一个去掉，这样既减少了训练样本的个数，又不影响训练样本的代表性。既能提高分类效率，又能保证分类准确率不降低。

2　基本概念

为了对训练样本的分布密度进行衡量，从而实现

训练样本集的约减，本文给出如下一些概念。

给定一个训练样本集 $D=C_1\cup C_2\cup\cdots\cup C_k$，$C_i(i=1,2,\cdots,k)$ 代表一个样本类别，且有 $C_i\cap C_j=\Phi(i,j=1,2,\cdots,k,i\neq j)$。$C_i=\{X_1,X_2,\cdots,X_{ni}\}$，其中 $X_j(j=1,2,\cdots,ni)$ 是类别 C_i 中的一个样本。

定义 1：训练样本集 D 的某一类别 C_i 中，两个样本 X、Y 之间的近邻程度用相似度表示，计算公式为：

$$s(X,Y)=\frac{\sum x_i * y_i}{\sqrt{(\sum x_i^2)*(\sum y_i^2)}}$$

定义 2：训练样本集 D 的某一类别 $Ci=\{X_1,X_2,\cdots,X_{ni}\}$ 的类密度的计算公式为：

$$Den(C_i)=\frac{\sum_{i,j=1}^{ni} s(X_i,X_j)}{ni*ni}$$

定义 3：训练集样本 D 的密度计算公式为：$Den(D)=\frac{\sum_{i=1}^{k} Den(C_i)}{k}$

3 基于文本密度的训练样本集约减算法

基于上述定义，本文给出训练样本集约减算法，描述如下：

输入：训练样本集 D；

输出：约减后的训练样本集 D；

步骤：

(1) 对于 D 中的每个类别 C_i，计算其中任意两个样本间的相似度；

(2) 计算 D 中每个 C_i 的类密度 $Den(C_i)$；

(3) 计算训练集 D 的密度 Den(D)；(4) 对于 D 中的每个 C_i，如果其类别密度 $Den(C_i)$ 高于训练集密度 Den(D)，则从 C_i 中约减掉 T 个文本，否则转(6)；

① 找 Ci 中相似度最高的两个样本，如果两个样本都没有标记去掉其中的任意一个，对另一个加上标记；

② 如果两个样本都有标记，去掉其中任意一个；

③ 如果只有一个样本有标记，去掉没有标记的样本；

④ 如果已去掉 T 个样本，转(5)，否则转(4－1)；

(5) 重新计算类别密度 $Den(C_i)$，转(4)；

(6) 结束，输出约减后的训练样本集 D。

算法分析：对于训练集 D 的某一类别 $C_i=\{X_1,X_2,\cdots,X_{ni}\}$ 来说，计算其样本两两之间的相似度的时间复杂度是 $O(ni^2)$，所以整个算法的时间复杂度为 $O(ni^2)$，而且该算法不需要确定参数。

4 实验分析

实验环境：P4CPU，512MB 内存，80GB 硬盘，Windows2000 操作系统，VC ++ 6.0 编程语言，Access2000 数据库。

实验数据：本文所用的语料库来自网站"中文自然语言处理开放平台"，总计 18630 篇有效文本，其中训练文本 9586 篇，测试文本 9044 篇。全部已经分类，总共分为 20 个类别。

实验结果：根据本文设计的基于文本密度的训练样本集约减算法，分四个档次对训练文本集进行约减，分别约减 15.85%、18.74%、20.67% 和 25.46%，约减前后分别用 k－最近邻方法进行分类，分类准确率如表 1 所示。

表 1　训练文本约减与分类准确率(k－最近邻法)

训练文本篇数 \ 分类准确率			基于 k－最近邻方法的分类准确率(%)
0	原始数据	9586 篇	71.44
1	约减 15.85%	8067 篇	72.07
2	约减 18.74%	7790 篇	72.35
3	约减 20.67%	7605 篇	72.68
4	约减 25.46%	7145 篇	71.58

从表 1 可以看出训练文本个数的适当约减并没有降低分类准确率，相反还有一些提高。约减 15.85% 的训练文本数据意味着可以降低分类时 15.85% 的计算量，约减 25.46% 的训练文本数据意味着可以降低分类时 25.46% 的计算量，这对于提高 k－最近邻方法的分类效率是非常有效的。

4 结束语

k－最近邻方法是一种比较广泛使用的文本分类方法，但其分类性能也受到一些因素的制约。本文提

(下转第 64 页)

(上接第 128 页)

出并设计了一种基于文本密度的训练样本集约减算法,并通过实验,结果表明,该方法在保证分类准确率的前提下,能够有效约减训练样本集,从而降低算法分类时的计算量,这对于提高 k－最近邻方法的分类效率是非常重要的。

参考文献

1 Y Yang, X Lin. A re－examination of text categorization methods. In: The 22nd Annual Int'l ACM SIGIR Conf on Research and Development in Information Retrieval, New York: ACM Press, 1999.

2 S. Belkasim, M. Shridhar, M. Ahmadi. Pattern classification using an effective KNNR[J]. Pattern Recognition Letter, 1992, 25(10): 1269－1273.

3 V. E. Ruiz. An algorithm for finding nearest neighbors in (approximately) constant average time[J]. Pattern Recognition Letter, 1986, 4(3): 145－147.

4 P. Hart. The condensed nearest neighbor rule[J]. IEEE Trans on Information Theory, 1968, 14(3): 515－516.

5 D. Wilson. Asymptotic properties of nearest neighbor rules using edited data[J]. IEEE Trans on Systems, Man and Cybernetics, 1972, 2(3): 408－421.

6 P. Devijver, J. Kittler. Pattern Recognition: A Statistical Approach[M]. Englewood Cliffs: Prentice Hall, 1982.

7 L. Kuncheva. Editing for the k－nearest neighbors rule by a gentic algorithms[J]. Pattern Recognition Letters, 1995, 16(8): 809－814.

8 L. Kuncheva. Fitness functions in editing kNN reference set by genetic algorithms[J]. Pattern Reconition, 1997, 30(6): 1041－1049.

9 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, 41(4): 539－545.