文章编号: 1003-0077(2007)03-0076-07

用于文本分类的改进 KNN 算法

王 煜[1,2], 王正欧[2], 白 石[3]

(1. 河北大学 数学与计算机学院, 河北 保定 071002; 2. 天津大学 系统工程研究所, 天津 300072;
3. 沧州市城建档案馆, 河北 沧州 061000)

摘 要: 最近邻分类器是假定局部的类条件概率不变, 而这个假定在高维特征空间中无效。因此在高维特征空间中使用 k 最近邻分类器, 不对特征权重进行修正就会引起严重的偏差。本文采用灵敏度法, 利用前馈神经网络获得初始特征权重并进行二次降维。在初始权重下, 根据样本间相似度采用 SS 树方法将训练样本划分成若干小区域, 以此寻找待分类样本的近似 k_0 个最近邻, 并根据近似 k_0 个最近邻和 Chi-square 距离原理计算新权重, 搜索出新的 k 个最近邻。此方法在付出较小时间代价的情况下, 在文本分离中可获得较好的分类精度的提高。

关键词: 计算机应用; 中文信息处理; 文本分类; 神经网络; Chi-square 距离; KNN 算法
中图分类号: TP391 **文献标识码**: A

An Improved KNN Algorithm Applied to Text Categorization

WANG Yu[1,2], WANG Zheng-ou[2], BAI Shi[3]

(1. School of Computer and Mathematics, Hebei University, Baoding, Hebei 071002, China;
2. Institute of Systems Engineering, Tianjin University, Tianjin 300072, China;
3. Urban Construction Archives of Cangzhou, Cangzhou, Hebei 061000, China)

Abstract: Nearest neighbor classification assumes locally constant class conditional probabilities. The assumption becomes invalid in feature space with high dimension. When KNN classier is used in feature space high dimension, severe bias can be introduced if the weights of features are not amended. In this paper, initial weights of text features are acquired based on sensitivity method firstly, and the second dimension reduce is done. Then training samples are divided into many groups based on sample similarity and the initial weights by using SS tree, k_0 approximate nearest neighbors of unknown sample are acquired by using SS tree. Weights are computed again based on k_0 approximate nearest neighbors and chi-square distance theory. K nearest neighbors are acquired based on new weights. Little time is spent, but the better accuracy of text categorization is acquired.

Key words: computer application; Chinese information processing; text categorization; neural network; Chi-square distance; KNN algorithm

1 引言

随着文本信息量的快速增长, 文本分类已成为信息检索、知识挖掘和管理等领域的关键技术。文本分类的精确程度取决于特征提取的科学性和分类算法的科学性。现有的文本分类方法主要有支持向量机(SVM)、K 最近邻(KNN)、决策树、线性最小二乘法估计(LLSF)和贝叶斯分类算法(Bayes)等。KNN 是一种传统的模式识别方法, 被广泛的应用于文本自动分类研究, 在准确率上表现出众[1]。KNN 方法是一种非参数的分类技术, 在基于统计的模式识别中非常有效, 对于未知和非正态分布可以取得较高的分类准确率。

最近邻分类器假定局部类的条件概率不变, 而这个假定在高维特征空间中无效[2,3]。因而在高维

收稿日期: 2006-09-25 **定稿日期**: 2007-03-12
基金项目: 国家自然科学基金资助项目(60275020)
作者简介: 王煜(1971—), 女, 副教授, 博士, 主要研究方向为数据挖掘、文本挖掘。

特征空间中使用欧氏距离公式的 k 最近邻分类器,如不对特征权重进行修正就会引起严重的偏好。因此在高维特征空间中使用 k 最近邻分类器,权重调整对分类性能至关重要。

文本分类中,对于分类起主要作用的特征数量远远低于文本特征空间本身的维数,相当多维对于文本分类意义不大,甚至是噪声数据,因此降维处理在文本分类中是必须的。虽然对于文本的高维特征向量可以通过降维处理得到相对维数较低的特征空间,但这样的特征空间的维数仍然是高维的。因此单靠特征降维无法解决最近邻分类器的假定(局部的类条件概率是不变的)在高维空间不成立的问题。

人们研究了各种权重的学习调整方法,有些方法是根据训练样本确定各个特征在分类中的作用,提前设定权重[4],这些方法在低维特征空间效果明显,随着维数增高,其性能逐步下降。因此有些方法是根据待测试样本和其临近样本确定特征权重[2,3]。这类方法以局部类的条件概率是可变的为前提,因此可以较好减少高维空间中 k 最近邻分类方法中的偏差,提高 k 最近邻分类器的分类性能。例如文献[2]在分类的时候寻找待分类样本的 k_0 个最近邻样本,然后据此计算新特征权重,再根据新权重寻找 k 个最近邻,这个过程还可重复,可极大提高分类精度。但会极大增加分类时间。

本文方法采用灵敏度法获得初始权重,根据初始权重设定加权欧氏距离公式,并采用神经网络进行第二次降维处理;据此加权欧氏距离公式测量样本间的相似度,采用 SS 树方法将训练样本空间划分成若干个小区域。在给出一个待分类样本 x 时,首先查找中心距离距 x 最近的 sk 个区域,然后根据加权距离公式在这几个区域内查找出 x 的近似 k_0 个最近邻。根据这 k_0 个最近邻和 Chi-Square 距离原理重新计算权重,再根据新的权重寻找 k 个最近邻,判别出 x 的类别。本文的方法在付出较小时间代价的情况下,可较大提高分类精度。

2　基于神经网络的权重调整和特征选择

神经网络作为数据挖掘的工具有多种应用,其中之一就是用于特征选择,同时可以用于特征权重的计算。本节采用灵敏度法[5]来进行特征权重的计算,并根据灵敏度大小进行特征选择。

2.1　基于神经网络的权重调整

将训练样本库的样本作为前馈神经网络的输入,样本的类别作为输出,当达到一定的训练精度时训练结束。然后逐个删除某个特征输入结点,再训练一个神经网络分类器,则删除该特征前后的分类器精度之差即为灵敏度法的计算依据。

假定某个训练样本库 T 中具有 J 类样本 n 个,样本特征维数为 m,计算各特征权重的具体步骤为:

1) 将整个训练样本库作为前馈神经网络的训练样本,采用 BP 算法,对神经网络进行训练,直到收敛为止。这样得到了一个神经网络分类器。此时神经网络分类器对训练样本库的样本 h 分类的预测值为 p_h^0。

2) 计算每个特征的灵敏度: 对每一个特征 i,将训练样本中所有样本的第 i 个特征的值均改为 0,其他特征值不变,形成新的样本库 B_i,然后在样本库 B_i 的基础上,按照第 1) 步的方法重新训练神经网络分类器,此时神经网络分类器对训练样本库的样本 h 分类的预测值为 p_h^i。则可根据公式(1) 计算特征 i 的灵敏度:

$$S_i = \frac{1}{n}\sum_{h}^{n}\frac{|p_h^0 - p_h^i|}{p_h^0} \qquad (1)$$

S_i 大,说明特征 i 对分类贡献大,S_i 小,说明特征 i 对分类贡献小。$\frac{|p_h^0 - p_h^i|}{p_h^0}$ 表示 p_h^i 对于 p_h^0 的相对误差的绝对值。

3) 计算各个特征的权重: 将特征 i 的 S_i 进行标准化计算,即可得到各特征初始权重 w_i,方法如公式(2)。

$$w_i = S_i \Big/ \sum_{j}^{m} S_j \qquad (2)$$

2.2　基于神经网络的特征选择

将特征向量的所有特征按照灵敏度由高到低进行排序,然后删除部分排序靠后的特征,再用神经网络对剩余特征进行训练,得到的误差在允许范围内就继续删除。这样剩余的特征就是对分类重要的特征。

为了减少神经网络在特征删除工作中的运算量,本节采用一种新的神经网络特征选择方法,在本文中称为"二分选择法"。首先按照 2.1 节中的灵敏度将特征由高到低进行排序。然后以二分法方式查找某个特征 R,以此特征 R 为界,将排在其后的全部特征删除。此方法极大减少了神经网络进行特征选择的运算量。本节基于神经网络的特征选择算法的具体步骤为:

1) 设定允许误差为 e;

2) 将所有特征按照 2.1 节中计算出来的灵敏度由高到低进行排序,形成队列 Q,此时特征数量为 m;

3) $i \leftarrow 1; j \leftarrow m; R = m$;

4) mid $\leftarrow [(i + j)/2]$,取队列前 mid 的特征作为训练样本的新的特征向量空间,去掉其余的特征,形成新的样本库 C;//$[(i + j)/2]$表示$(i + j)/2$ 后取整数;

5) 按照新的样本库 C 建立新的神经网络分类器,分类器对样本库 C 中所有样本分类的误差之和为 ce;

6) 如果 ce $\leq e$,则 $j \leftarrow$mid - 1, R = mid;否则 $i \leftarrow$mid + 1;

7) 如果 $i < j$,转向第 4) 步执行;

8) 将队列 Q 中 R 后面的特征从样本的特征向量空间中删除,得到 m维的特征向量空间;

9) 整理训练样本库,将样本库中样本按照新的特征向量空间表示,形成新的样本库 F。

3 基于 SS-tree 方法的训练样本空间区域划分

SS-Tree[6]就是根据数据之间的相似性作为标准,为数据集合建立一棵 B^+-Tree。为了减少分类中增加的寻找初始 k_0 个最近邻的时间,本文根据样本间的相似度,采用 SS 树方法将训练样本空间划分成若干个小区域。

假定每个结点的孩子数量在[b,B]之间。建立 SS-Tree,首先将所有训练样本链成一个待插入样本队列,初始化一棵空树。然后将待插入样本一个个插入 SS 树中。本文中,SS 树的建立遵循的准则为:

1) 选择插入结点的标准为中心点距离待插入样本最近的结点;

2) 为保持每个结点孩子数量不超过 B,在结点的孩子数量超过 B 时,如果结点所对应样本集合的样本重新插入的数量少于给定比例,则将该结点删除,将所有样本插入待插入队列,否则要对结点进行分裂,以尽量减少分裂后两个区域的面积总和为准则进行结点分裂;

3) 在删除结点时,这个结点的父结点的孩子数量可能会小于 b,为保持每个结点孩子数量不少于 b,此时就要进行结点合并工作,合并时以合并后面积增加最小为标准。

为了记录区域划分,将 SS 树所有孩子为样本的那层结点的区域划分保存在一张表中。为了减少后期查找样本启动磁盘的时间,要按照此表中的顺序重新排列训练样本,然后进行存储。

4 基于局部自适应度量的权重获取

4.1 Chi-square 距离

假定训练样本库具有 n 个样本,这些样本可以分为 J 类。x 为待分类样本,y 是 x 的最近邻,计算样本相似度的距离公式 $D(x,y)$ 应该满足公式 $E[(r^*(x) - r(x,y)]^2$ 值最小[7],其中 $r(x,y) = \sum_{j=1}^{J} Pr(j|x)(1 - Pr(j|y))$, $r^*(x) = \sum_{j=1}^{J} Pr(j|x)(1 - Pr(j|x))$, $Pr(j|x)$ 是 x 的分类条件概率。把 $r(x,y)$ 和 $r^*(x)$ 代入 $E[(r^*(x) - r(x,y)]^2$,得到使 $E[(r^*(x) - r(x,y)]^2$ 最小的 Chi-square 距离如公式(3)。

$$D(x,y) = \left(\sum_{j=1}^{J} Pr(j|x)(Pr(j|y)(1 - Pr(j|x)) \right)^2 \quad (3)$$

根据待分类样本 x 和样本 y 之间的分类条件概率之差,文献[8]提出改进的 Chi-square 距离公式(4)。

$$D(x,y) = \sum_{j=1}^{J} (Pr(j|y) - Pr(j|x))^2 \quad (4)$$

为了进一步改进 Chi-square 距离公式,文献[2]加入权重 $1/Pr(j|x)$,得到加权 Chi-square 距离公式(5)。公式(5)比公式(4)优越之处在于权重 $1/Pr(j|x)$ 会在 x 与 y 可能不属于同一类时增加 x 与 y 的距离[2]。

$$D(x,y) = \sum_{j=1}^{J} \frac{(Pr(j|y) - Pr(j|x))^2}{Pr(j|x)} \quad (5)$$

公式(5)表达的是样本的真实分类和估计分类的差,而这里的研究目的是要找出每个特征在样本集合中局部的分类能力。Chi-square 距离公式(5)可以帮助研究每个特征预测分类条件概率 $Pr(j|x)$ 的能力,因此可以在此基础上确定每个特征分类时的权重。

4.2 权重计算

根据 Chi-square 距离的讨论,可知 $Pr(j|x)$ 是 x 的函数。可通过 $Prt(j|x_i = a)$ 来计算 $Pr(j|x)$ 的条件期望,x_i 表示 x 的第 i 个组成部分,假定 x_i 的

值为 a。这里 $Prt(j|\ x_i = a)$ 计算方法如公式(6)[2]:

$$Prt(j\mid x_i = a) = E[Pr(j\mid x)\mid x_i = a] = \int Pr(j\mid x)\, p(x\mid x_i = a)\, dx \tag{6}$$

其中 $p(x|\ x_i = a)$ 是其他输入变量的条件密度。

公式(7)中的 $r_i(z)$ 代表了特征 i 在 $x_i = z_i$ 时的预测 $Pr(j|\ z)$ 的能力。$Prt(j|\ x_i = z_i)$ 越接近 $Pr(j|\ z)$, 特征 i 在 z 处的预测类别的能力越强。$r_i(z)$ 的值越小,对应的特征 i 对分类的作用越大,特征 i 的权重也应该越大; $r_i(z)$ 的值越大,对应的特征 i 对分类的作用越小,特征 i 的权重也应该越小。

$$r_i(z) = \sum_{j=1}^{J} \frac{[Pr(j\mid z) - Prt(j\mid x_i = z_i)]^2}{Prt(j\mid x_i = z_i)} \tag{7}$$

可根据样本 x 的近邻确定特征 i 在 x 周围的局部范围内的分类能力,计算方法如公式(8):

$$\bar{r}_i(x) = \frac{1}{k_0}\sum_{z\in N(x)} r_i(z) \tag{8}$$

其中 $N(x)$ 表示 x 的最近邻的样本集合,样本个数假定为 k_0 个。

根据公式(9)可以确定在对待分类样本 x 求最近邻时的距离公式中各特征的权重。

$$w_i(x) = \exp\left(cR_i(x)\right) \Big/ \sum_{l=1}^{m}\exp(cR_l(x)) \tag{9}$$

其中, c 为调整 r_i 对 w_i 影响程度的系数, m 为样本特征空间的维数, $R_i(x) = \max_{j=1}^{m}\{\bar{r}_j(x)\} - \bar{r}_i(x)$。

实际应用中,需要根据训练样本计算未知的 $Pr(j|\ z)$ 和 $Prt(j|\ x_i = z_i)$ 的近似值。如果样本 z 的最近邻集合为 $N_1(z)$, 样本数量为 k_1, 则 $\Pr(j|\ z)$ 的近似估计 $Pra(j|\ z)$ 如公式(10):

$$Pra(j\mid z) = \frac{\sum_{h=1}^{n} 1(x_h \in N_1(z))\, 1(class_h = j)}{\sum_{h=1}^{n} 1(x_h \in N_1(z))} \tag{10}$$

其中, $1(\cdot)$ 函数表示如果 $(\ )$ 内条件成立,则函数值为 1,如果 $(\ )$ 内条件不成立,函数值为 0; $class_h$ 表示样本 x_h 所属的类别。$Prt(j|\ x_i = z_i)$ 的近似估计 $Prta(j|\ x_i = z_i)$ 的计算如公式(11):

$$Prta(j\mid x_i = z_i) = \frac{\sum_{x_h\in N_2(z)} 1(|\ x_{hi} - z_i\ | \le \Delta_i)\, 1(class_h = j)}{\sum_{x_h\in N_2(z)} 1(|\ x_{hi} - z_i\ | \le \Delta_i)} \tag{11}$$

其中, x_{hi} 是第 h 个样本的第 i 个特征, $N_2(z)$ 是 z 的最近邻集合,样本数量为 k_2, $N_2(z)$ 是比 $N_1(z)$ 范围大的最近邻集合。Δ_i 的选择方法是让满足 $|\ x_{hi} - z_i\ | \le \Delta_i$ 的样本个数为 L, 即 $\sum_{h}^{n} 1(|\ x_{hi} - z_i\ | \le \Delta_i)\, 1(x_h \in N_2(z)) = L$。

5　改进的 KNN 文本分类算法

文献[2]中提出了一种 Adaptive Metric nearest-neighbor 算法(简称为 ADAMENN 算法)。ADAMENN 算法首先根据权重为 1 的欧氏距离公式寻找待分类样本 x 的 k_0 个最近邻,然后为 k_0 个最近邻中的每个样本寻找 k_1 和 k_2 个最近邻,以计算 $Pra(j|\ z)$、$Prta(j|\ x_i = z_i)$; 然后根据公式(9)重新计算权重后,再次寻找 x 的 k 个最近邻; 如果效果不好,还要重复上述工作。但是重复一次就要在训练样本库中作一次寻找 k_0 个最近邻和 k_0 次寻找 k_2 个最近邻的工作。ADAMENN 算法根据局部样本获得的权重可以使 KNN 分类器的分类精度得到提高,但是分类速度会非常慢。使计算量大的 KNN 算法的计算量更大,在高维的文本特征空间中会因为其计算量过大而严重降低其实用性。

本文针对 ADAMENN 算法的缺陷提出了一种改进的 KNN 算法,本文中称其为 MKNN 算法。为了保证分类精度的提高,权重开始不采用 1,而是采用 2.1 节灵敏度方法获得的权重,并在分类前根据此权重计算出每个训练样本在训练样本库中的 k_1 和 k_2 个最近邻,计算出 $Pra(j|\ z)$、$Prta(j|\ x_i = z_i)$ 并记录。然后采用 SS 树将训练样本空间按照样本间相似度划分成许多小区域,仅在距离待分类样本近的区域内快速查找待分类样本的近似 k_0 个最近邻,以此来求新的权重,从而降低了时间代价。并且为了减少在分类时付出的时间代价,MKNN 算法在分类时候只重新计算权重一次。

在 MKNN 算法中,虽然会增加采用 SS 树进行样本区域划分的时间开销,但这是预处理的时间,没有占用分类的时间,并且只要训练样本库不改变,就不需要再次执行。与 ADAMENN 算法相比,它可以降低分类时所用时间,随着对样本分类次数的增加,总体时间消耗也会小于 ADAMENN 算法。并且寻找每个训练样本的 k_1 和 k_2 个最近邻也都在分

类之前处理完成(同样如果训练样本库不改变,也不需要再次执行)。这样分类时间与 ADAMENN 算法相比大大降低了。从而在分类时,为修改权重付出的时间代价比较小。针对传统 KNN 算法而言,MKNN 算法总体时间开销有所增大,分类时间开销的增大比较小,但分类精度却大大提高。

本节提出的改进 KNN 算法在本文中称为 MKNN 算法,其具体步骤如下:

1) 采用 VSM 表示文本特征,建立原始样本空间;

2) 采用文献[9]中的改进 χ^2 统计量和模式聚合方法对文本特征空间进行降维处理;

3) 设新的特征空间为 m 维,应用 TFIDF 公式计算特征值 TFIDF 计算如公式(12);

$$x_{ij} = \log(tf_{ij} + 1) \times \left(1 - \sum_{i=1}^{n} \frac{\frac{tf_{ij}}{df_j}\log\left(\frac{tf_{ij}}{df_j}\right)}{\log(n)}\right) \tag{12}$$

其中,x_{ij} 为第 j 个词在第 i 篇样本中的特征值,n 为训练样本库样本数量,tf_{ij} 为第 j 个特征在第 i 篇样本中出现的频率,df_j 为第 j 个特征在样本库中出现的样本数量。

文献[10]证明:此计算公式相对于其他常见公式可以对文本检索等操作取得最好的结果。

4) 采用本文 2.1 节的神经网络权重调整方法,获得每个特征的权重 w 作为初始权重,确定测量样本相似度的距离公式(13)中权重;并根据第2.2 节中的神经网络特征提取方法进行第二次特征提取;

$$D(x,y) = \sqrt{\sum_{i=1}^{m} w_i (x_i - y_i)^2} \tag{13}$$

5) 采用第 3 节的方法,根据距离衡量公式(13)建立训练样本库的 SS 树,将训练样本空间根据样本间的相似度划分成许多小的区域;

6) 确定参数 k_0、k_1、k_2、k、L 和 c:

k_0:待分类样本 x 最近邻集合 $N_0(x)$,样本数量 k_0。这个含有 k_0 个样本的集合用于获取待分类样本处的特征权重的局部训练样本,k_0 表示局部范围的大小。k_0 不宜太大,太大就不能获得待测试样本处的特征的局部分类能力,k_0 也不能太小,太小就可能成为某个数据点处的特征的分类能力,而不是测试样本处的特征的局部分类能力;另外,参数 k_0 要根据 k 值进行选择,k 值大 k_0 就要大,k 值小 k_0 就要小;

k_1:z 的最近邻样本集合 $N_1(z)$ 中样本的个数,z 是 $N_0(x)$ 中的样本,k_1 要小,小数值可降低估计偏差;

k_2:z 的最近邻样本集合 $N_2(z)$ 中样本的个数,z 是 $N_0(x)$ 中的样本,k_2 不能太大也不能太小,文献[2]的经验为 $k_2 = 0.15\ n$,n 为训练样本库的样本数量;

k:待测试样本周围的 k 个近邻,用于计算待测试样本的类别;

L:距 z 距离不超过 $\triangle$ 的样本数量,z 是 $N_0(x)$ 中的样本,文献[2]的经验为 $L = 0.5\ k_2$;

c:c 为公式(9)调整 r_i 对 w_i 影响程度的系数;

7) 查找每个训练样本 z 的 k_1 和 k_2 个最近邻,根据公式(10)(11)计算出 $Pra(j|z)$、$Prta(j|x_i = z_i)$;

8) 当给定一个待分类样本 x 时,分类的具体步骤为:

①查找中心点距离 x 最近的 sk 个区域(SS 树划分的)。根据公式(13)在 sk 个区域内查找出含有 k_0 个最近邻样本的集合 $N_0(x)$,sk 的值不能太大,太大会提高时间代价,也不能太小,太小会降低分类精度;

②以 x 的 k_0 个最近邻为基础,根据公式(8)计算出 $\bar{r_i}(x)$,然后根据公式(9)计算新权值 w;

③采用新的特征权重查找 x 的 k 个最近邻,根据 k 个最近邻将待分类样本 x 归属为权重最大的那个类别。每类权重 P 计算方法如公式(14):

$$P(x,C_j) = \sum_{i=1}^{k} Pa(a_i,C_j)\, wa(a_i,C_j)$$
$$j = 1,2,\cdots,J \tag{14}$$

其中 $wa(a_i,C_j)$ 是 x 的 k 个最近邻中的样本 a_i 将 x 分类到类别 C_j 的权重,本文采用 $1/D(x,a_i)$,$Pa(a_i,C_j)$ 如下:

$$Pa(a_i,C_j) = \begin{cases} 1 & a_i \text{ 是类别 } C_j \text{ 的样本} \\ 0 & a_i \text{ 不是类别 } C_j \text{ 的样本} \end{cases}$$

6 仿真实验

为了验证算法的有效性和正确性,本文从新浪网上下载了 1 665 篇新闻为仿真实验中的数据,其中 1 110 篇新闻作为训练样本,其余 555 篇新闻作

为测试样本,这些样本共分 6 类,训练样本、测试样本的分布情况如表 1。

表 1　样本分布

类名	环保	法治	体育	军事	娱乐	财经
训练样本篇数	120	313	246	143	124	164
测试样本篇数	60	157	124	72	61	81

在训练样本中,除去虚词后,统计出 8 649 个词,计算出每个词的 chi 值并由高到低进行排序,选择 chi 值高的前 4 000 个单词按照文献[10]方法进行降维处理,得到 179 个特征,经过神经网络进行一步特征提取后,得到 75 个特征。在相同的环境下,采用传统 KNN 算法,对既不采用灵敏度方法进行权重调整也不采用神经网络特征提取的方法(实验一)、只采用灵敏度方法调整权重而不采神经网络特征提取的方法(实验二)和既采用灵敏度方法调整权重又采神经网络特征提取的方法(实验三,本文称为 WKNN 算法)的比较结果如表 2 所示。

表 2　灵敏度权重修正和神经网络特征提取效果的实验比较

	特征维数	k = 10		K = 20	
		分类精度	错误分类篇数	分类精度	错误分类篇数
实验一	179	61.44 %	214	65.59 %	191
实验二	179	81.08 %	105	80.36 %	109
实验三	75	79.1 %	116	79.82 %	112

在相同的环境下,采用 WKNN 算法和 MKNN 算法建立文本分类器对测试样本进行分类。在本文 MKNN 算法中,建立 SS 树时,b 取值为 $\max\{K,20\}$,B 取值 b 值的 2 倍,sk 取值不能太大也不能太小,本文仿真试验中,b 取值 20,B 取值 40,sk 取值 7(sk 曾选择 3,5,7,9,11,最后 7 最为合适,选择 9,11 时间代价增高,分类精度提高非常小,选择 3,5 时分类精度偏低);选择 $k_1=1$,$k_2=160$(k_1 曾选择 1,2,3,4,5,6,当参数 $k_1>2$ 时,其分类精度低于 WKNN 算法,$k_1=1$ 分类精度最好;参数 k_2 选择方法根据文献[2],也采用过 $k_2=100,200$,分类精度均不理想);$k=10$ 时选择 $k_0=50$(k_0 也曾选择 30,40,60,70,选择 30,40,70 时分类精度均低于 WKNN 算法,选择 60 时分类精度提高很小),$k=20$ 时选择 $k_0=70$(k_0 也曾选择 60,80,90 分类精度均低于 WKNN 算法,选择 80 时分类精度接近 WKNN 算法)。在上述参数下,分类结果如表 3 所示。

表 3　WKNN 算法、MKNN 分类的实验比较

	k = 10			K = 20		
	分类精度	错误分类篇数	平均分类时间	分类精度	错误分类篇数	平均分类时间
WKNN	79.1 %	116	3.49 秒	79.69 %	112	3.50 秒
MKNN	89.55 %	58	4.33 秒	90.27 %	54	4.37 秒

从表 3 可以看出,本文的 MKNN 算法的分类精度比 WKNN 算法有大幅度提高。这样可以在付出较小时间代价的情况下,获得了较大分类精度的提高。

7　结论

采用灵敏度法获得初始特征权重,然后根据样本间相似度,采用 SS 树将训练样本库划分成许多的小区域,据此可快速查找待分类样本的近似 k_0 个最近邻。之后,根据 k_0 个近似最近邻和 Chi-Square 距离理论重新计算权重后,再根据新的权重寻找待分类样本的 k 个最近邻,从而在付出较小时间代价的情况下,可获得较大分类精度的提高。另外,本文采用神经网络方法进行第二次降维处理时,注意了神经网络降维运算量大的问题,采用二分法将运算量大大降低。

本文中 MKNN 算法只提高了文本分类准确率,为了使 KNN 算法更加实用,需要在分类速度上进行改进,进一步提高分类性能。

参考文献:

[1] 代六玲,黄河燕,陈肇雄. 中文文本分类中特征抽取方法的比较研究[J]. 中文信息学报,2004,18(1):26-32.

[2] Carlota Domeniconi, Jing Peng, Dimitrios Gunopulos. Locally Adaptive Metric Nearest-Neighbor Classification[J]. IEEE TRASACTIONS ON PATTERN ANAL YSIS AND MACHINE INTELLIENGCE. 2002, 24(9):1281-1285.

[3] Jing Peng, Douglas R. Heisterkamp, H. K. Dai. LDA/SVM Driven Nearest Neighbor Classification [J]. IEEE TRASACTIONS ON NEURAL NETWORKS. 2003,14 (4):940-942.

[4] 王晓晔,王正欧. K-最近邻分类技术的改进方法[J]. 电子信息学报, 2005,27(3):487-491.

[5] Setiono R, Liu H. Neural network feature selector [J]. IEEE TRANSCATIONS ON NEURAL NET-

WORKS, 1977, 8(3): 654-662.
[6] David A. White, Ramesh Jain. Similarity indexing with the SS-tree[A]. In: Proceedings of the 12th International Conference on Data Engineering[C]. 1996, 516-523.
[7] Weitschereck D, Aha D W, Mohri T. A review and empirical evaluation of feature weighting methods for a class of lazy learning algorithms [J]. AI Review. 1997, 11(2): 273-314.
[8] T. Hastie, R. Tibshirani. Discriminant Adaptive Nearest Neighbor Classification [J]. IEEE TRASACTIONS ON PATTERNANALYSIS and MACHINE INTELLIGENCE. 1996, 18(6): 607-615.
[9] 王煜, 王正欧. 基于模糊决策树的文本分类规则抽取[J]. 计算机应用, 2005 年, 25(7): 634-1637.
[10] 周水庚, 关佶红, 胡运发. 隐含语义索引在中文文本处理中的应用研究[J]. 小型微型计算机系统, 2001, 22(2): 239-243.

《综合型语言知识库》通过技术鉴定

北京大学计算语言学研究所完成的研究成果《综合型语言知识库》于 2007 年 2 月 13 日在京通过技术鉴定。鉴定会由教育部主持。鉴定委员会由来自清华大学、北京航空航天大学、中科院软件所、中科院计算所、教育部语言文字应用研究所、北京语言大学、中国科学技术信息研究所的 9 位专家组成。张钹院士任主任, 怀进鹏教授任副主任。

鉴定委员会听取了俞士汶教授的研制报告、孙斌博士的技术报告、测试组组长孙茂松教授的测试报告以及用户报告, 审查了相关资料, 并进行了认真的讨论。鉴定委员会对该项成果的鉴定结论如下:

1. 北京大学计算语言学研究所自 1986 年以来, 在 863、973、自然科学基金、社会科学基金等国家计划的支持下, 历时 20 多年, 建成了《综合型语言知识库》。

2. 其中, 《现代汉语语法信息词典》包括 34 个数据文件, 收录词语 8 万条, 描写的语法属性总项数超过 360 万项, 是目前国内外最有影响的汉语词汇知识库; 《汉语短语结构知识库》包含 600 余条汉语短语规则, 涵盖了汉语基本短语结构的各种合理组合; 《中英文概念词典》实现对词网中近 10 万个英文概念的汉语对应, 是全球多语词网建设中具有标志性的一项成果; 《现代汉语大规模基本标注语料库》切分标注的总量超过 5 000 万字, 《汉英双语对齐语料库》规模达 80 万句对, 规模大、质量高。

3. 《综合型语言知识库》是一个在逻辑上有机联系的整体。在语言基础资源方面, 提出并制定了一系列规范, 使得各成员之间的属性互相参照对应, 知识库中既包含词、短语、句子、篇章等不同语言单位, 又涉及汉语、英语等不同语言, 并从词法、句法和语义等不同角度进行信息描述, 而信息描述融合了词典中的显性知识和语料库中的隐性知识, 是语言信息处理的基础资源和重要保证, 在工程实践中又进一步发展了面向语言信息处理的汉语语法理论体系。

4. 在这些资源基础上, 开发的基于语料库的双语词典编纂平台实现了语料库处理技术和词典编纂技术的整合, 有利于辞书编纂手段的现代化; 通过对汉语词语切分、词性标注和命名实体识别等关键技术创新, 研制了文本信息提取系统。

5. 《综合型语言知识库》已得到广泛应用, 并向国内外大公司和研究机构转让许可使用权 150 余次, 取得了显著的经济效益和社会效益。

鉴定委员会认为: 《综合型语言知识库》开创性地实现了汉语词语的大规模归类与属性描述, 很好地处理了基础研究与应用研究的关系, 形成了基础资源建设与应用系统开发相互支撑、相互促进的良性模式, 其规模、深度、质量和应用效果在我国语言工程实践中是前所未有的。该成果是以汉语为核心的多语言知识库建设中最全面、最重要的研究成果, 总体上达到了国际领先水平, 一致同意通过鉴定。

建议继续推进语言知识库的研究与开发, 进一步推广应用。

张化瑞