K 特征线法在文本分类上的应用

The Application of K Nearest Feature Line Method to the Text Classification

杨 昂

(湖南大学计算机系 长沙410082)

Abstract A novel pattern classification method, called nearest feature line(NFL) has been proposed . In this paper we propose a novel classification method, called the k nearest feature line(kNFL), for text classification and approach it to application. The experiment result shows that kNFL presents better performance than NFL, k nearest neighbors (kNN) and nearest neighbor(NN),especially when the training sets are small.

Keywords kNFL,kNN,classification

1. 引言

随着 Internet 的迅速发展,网上信息成几何级数不断增长,如何从中找出人们需要的文献是信息检索要处理的重要问题。近期的一个研究热点是网络文本信息处理。文本信息处理包括信息检索、文本分类和信息过滤。自动文本分类是其中的重要环节。比较成熟的统计分类模型是向量空间法,它将文献表示为向量,将过滤器的设计转化为机器学习的问题。自动文本分类算法是一种有监督的机器学习算法。通过领域专家手工分类的文献训练集进行训练,得到统计模型,再用算法考察被测试的文献属于哪类的可能性最大。

在模式识别中广泛使用的分类方法有近邻法,就是说对未知样本 x,我们比较 x 与 $N=\sum_{i=1}^{c}N_i$ 个已知样本之间的距离,并决策 x 与离它最近的样本同类。kNN 是最近邻法的一个推广。这个方法就是取未知样本 x 的 k 个近邻,看这 k 个近邻中多数属于哪一类,就把 x 归为哪一类。最近特征线法(nearest feature line,NFL)是近两年由 Stan z. Li[1]在人脸识别中提出的新的分类方法,在人脸识别中取得了很好的效果。我们把 NFL 用于文本分类,并结合 kNN 和 NFL 的思想,提出了新的分类方法,k 最近特征线法(kNFL)。我们在文本分类中比较了这几种分类算法。实验说明 kNFL 具有最好的性能。它的平均精度是最高的。对于小样本集合(每类样本小于10)有明显优势。在分类算法的比较测试中,kNFL 表现出较强的适应性,在与 k 近邻法、贝叶斯法、NFL 和近邻法的比较中,它的精度是最高的。

2. K 最近特征线算法

2.1 相关工作

kNFL 算法是对 NFL 的一种扩充。NFL 最先是由 Stan Z. Li 在1998年提出的,在人脸识别和图像处理的分类中有很好的实验结果[1]。NFL 假设训练模型中每类中至少有两个样本点,将同一类中任意两点连线,认为线上的点也能代表此类的特征,称为特征线(feature line,FL),未知样本 x 到特征线的距离叫做特征线距离。NFL 就是求出所有的特征线距离,判别最短特征线距离所属的类就是未知样本点所属的类。一类中已知样本 x1,x2,特征线为$\overline{x_1x_2}$,点 x_p 是特征线上 x 所投影的点,定义特征线距离为:

$$d(x,\overline{x_1x_2})=\|x-x_p\|$$

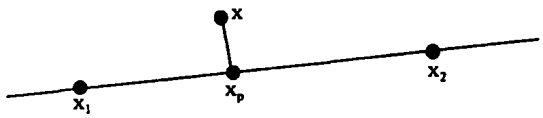

图1 由已知样本 x1,x2产生特征线$\overline{x_1x_2}$。未知样本 x 在$\overline{x_1x_2}$的投影为 x_p

点 x_p 由以下公式计算:$x_p=x_1+\mu(x_2-x_1),\mu\in\mathcal{R}$ (1)

因为$\overline{x_px}$与$\overline{x_1x_2}$垂直,我们有

$$(x_p-x)\cdot(x_2-x_1)=[x_1+\mu(x_2-x_1)-x]\cdot(x_2-x_1)=0 \quad (2)$$

这里是向量的点积运算,得到:

$$\mu=(x-x_1)\cdot(x_2-x_1)/(x_2-x_1)\cdot(x_2-x_1) \quad (3)$$

特征线 FL 是两个样本点的线性组合,扩大了已知样本的规模。设类 c 中有 $N_c>1$个已知样本,那么就有 $K_c=N_c(N_c-1)/2$条线来代表这个类。

NFL 算法实现过程是:

1. 设 $x^c=\{x_i^c\,|\,1\leqslant i\leqslant N_c\}$是类 c 的 N_c 个样本点集合,x_i^c,x_j^c 是类 c 的两个不同的样本点,共有 K_C 条特征线$\overline{x_i^cx_j^c}$。

2. 求这个类中每对特征线$\overline{x_i^cx_j^c}$到未知样本 x 的距离 $D(\overline{x_i^cx_j^c})$,若小于最小距离 $d_{\min}$,则未知样本 x 属于当前类;

3. $c=c+1$,若 $c\leqslant M$,转入2;

4. 判断未知样本属于最短特征线距离所属类。

NFL 扩长了样本点的数目,对于人脸识别有明显的优势,因为同一个人的图像会有不同的表情,侧面角度和亮度等一系列的变化,而训练库是有限的,不可能把一个人的所有情况下的图像都录入。NFL 把同类的不同样本点联系起来,使有限的样本空间变为无限的连续空间,增大了样本点容量,有很强的适应能力,能分辨出在不同角度,亮度下的人脸图像,它比普通算法具有明显的优势。

近邻法是 Cover 和 Hart 于1968年提出的[6],直至现在仍是模式识别非参数法中最重要的方法之一。kNN 算法思想很简单:给一个待识别的样本,系统在已标记类别的训练集中找到最近的 k 个近邻,看这 k 个近邻中多数属于哪一类,就把待识别的样本归为哪一类。当 k 取1时就是近邻法(nearest-neighbor,NN),取最近的样本点所属的类为未知样本的类别。NN 强调最近点的重要性,而 kNN 则降低了机会误差,从

整体考虑，是一种更为普遍的方法，错误率也更低。

我们在kNN算法的启发下，将NFL扩展为kNFL，降低了算法的错误率，提高了分类精度。但kNFL，NFL算法需要将所有特征线(FL)存入计算机中，每次决策都要计算待识别样本与全部特征线之间的距离并进行比较，因此存储量和计算量都很大。

2.2 kNFL

NFL算法认为样本空间是连续空间，有限的已知样本不足以代表类空间。它把同类已知样本看成是有联系的样本点，它们的连线上的点也能代表样本空间，使有限的样本变为无限样本。在人脸识别中，每个样本点代表一幅特征人脸，同一个人的表情，脸的转向，照片亮度是千差万别的，而训练样本只能是有限的，NFL把有限样本扩充到连续，无穷多的样本，增加了样本容量，提高了精确度和效率，使NFL的错误率低于其他算法。

在NFL中，未知样本的类别由最近的特征线(feature line，FL)决定，最近特征线NFL的类别就是 x 的类别。

在文本分类中，我们采用公式(4)的相似度距离，两篇文献相似度越高，距离值越大。kNFL算法如下：

1. 求所有FL的距离 $d_1, d_2, \cdots, d_N$，将它们排序；

2. 保留前 k 个 d_i 的距离值，类别和对应样本点；

3. 在前 k 个 d_i 中按类求和 $d^c = \sum_{c(d_i)=c, 1\leqslant i\leqslant k} d_i, 1\leqslant c\leqslant M$；

4. $c(x)=c(d^*), d^* = \max_{d^c, 1\leqslant c\leqslant M} d^c$，其中 $c(x), c(d^*)$ 表示 x, d^* 所属类别。

这样有一个最近距离决定未知样本类变为有k个最近距离决定，降低了偶然误差，提高了算法适应性。

3. 文本分类模型

向量空间模型是由Salton等人在六十年代末到七十年代初期提出并发展起来的[7]。这一模型将给定的文本(文献、查询、或文献的一段等)转换成一个维数很高的向量。它的最大特点是可以方便地计算出两个向量的近似度，即向量所对应的文本相似性。所有文献用向量表示，也就是将搜索到的文档材料进行特征项抽取，形成特征向量，而用户查询时，则针对特定的查询向量，比较它与所有文献的相似度，并依相似度大小将文献排序提交给用户。向量空间模型使用以下的一些知识：

·文献D(Document)：泛指各种机器可读的记录，通常指一篇文献。

·特征项t(Term)：也称为索引项，是指出现在文档d中且能够代表该文档性质的基本语言单位，主要由词或短语来构成，这些基本语言单位统称为项，于是文献和查询均可用项构成向量来表示 $D=(t_1, t_2, \cdots, t_n)$。

·特征项权值 w_{ik}(Term weight)：对于有n个不同的项的系统，文献 $D=(t_1, t_2, \cdots, t_n)$，项 $t_k(1\leqslant k\leqslant n)$ 常常被赋予一个数值 w_{ik}，表示它在文献中的重要程度，称为项 t_k 的权重。因此，我们一般用 $D=(w_{i1}, w_{i2}, \cdots, w_{in})$ 的形式表示文献。也就是指特征项 t_k 代表文档 d_i 的能力大小。w_{ik} 的计算采用特征项频率 tf_{ik} 和反比文献频率 idf_k 计算：

$$w_{ik}=tf_{ik}\cdot idf_k=tf_{ik}\cdot(\log_2(N/n_k)+1) \qquad (4)$$

TF·IDF：TF(Term Frequency)指某个特征项t在文档d中出现的次数，显然，TF越大，则特征项t越能够代表文档d，t的权值应当与之成正比。DF(Document Frequency)指整个文档集合中，包含特征项t的文档个数，直观上看，DF越大，即包含特征项t的文档越多，则特征项t越代表文档d的能力越小，t的权值应当与之成反比；所以人们就提出了IDF(Inverse Document Frequency)，既然DF与t的权值成反比，那么IDF应当与t的权值成正比，通常 $IDF=\log(N/DF)$，其中，N代表文档集合中的文档个数。TF·IDF能够有效地表达特征项t在文档d中的重要性。

在公式(4)中，tf_{ik} 表示特征项 t_k 在文档 d_i 中出现的频率，N代表文档集合中的文档数，n_k 代表在文档集合中出现特征项 t_k 的文档数目。从公式(4)可知，tf_{ik} 越大，w_{ik} 值越大；同样 n_k 越小，w_{ik} 值也越大，说明该特征项 t_k 更能够代表文档 d_i 的内容。

·向量空间模型：设文档集合中共有m个不同的特征项 $t_1, t_2, \cdots, t_m$，分别计算文档 $d_i(i=1, \cdots, N)$ 的特征项 $t_1, t_2, \cdots, t_m$ 的特征项权值，由这些特征项权值所构成的向量 $(w_{i1}, w_{i2}, \cdots, w_{im})$ 成为文档 d_i 的向量。

由于特征项 $t_1, t_2, \cdots, t_m$ 互不相同，我们可以将文档向量看作是m维欧氏空间的向量。这样，文档之间的相似程度通过向量的形式转化为向量之间的数学计算模式，使得在进行文档归类以及查询匹配过程中的计算过程比较简单、快速。

·相似度(Similarity)：两个文献 d_1 和 d_2 的内容之间的相关程度通常用相似度来表示。在向量空间模型中，我们借助于向量之间的某种距离来表示文献间的相似度。通常用向量之间的夹角的余弦来计算。设文档 d_1 和 d_2 向量表示是：$d_1=(w_{11}, w_{12}, \cdots, w_{1m})$，$d_2=(w_{21}, w_{22}, \cdots, w_{2m})$，则夹角余弦公式如公式(5)：

$$\mathrm{Sim}(d_i, d_j)=\cos\theta=\frac{\sum_{k=1}^{m} w_{ik}\cdot w_{jk}}{\sqrt{(\sum_{k=1}^{m} w_{ik}^2)(\sum_{k=1}^{m} w_{jk}^2)}} \qquad (5)$$

4. 实验及结果

我们在VC++6.0上开发了一个文本分类系统Archer。系统源程序使用C++，数据库从网上下载，数据库为2000篇新闻组短文，已被分为20类，每类100篇短文。按上章介绍的文本分类模型，将文献看成是一个词集，词之间没有关系。首先对所有文献做词法分析，找出所有不同的词。从2000篇英文短文中共找出36864个不同的词。在建模型时可以取权值最大的，最能代表文献特性的前N个词。库建好后，测试时系统从数据库中随机选取一定比例的文献作为测试集，即未知样本集，余下的作为训练集，即已知样本集，按模型对训练集建模。然后根据算法设定每篇文章中每个词的权值。于是一篇文献可以看成是多维特征向量空间中的一点。kNN，NN，kNFL，NFL均用公式(5)计算两点的距离，距离越大表明两点越相似。我们对数据库进行测试，实验结果表明平均精度kNFL最好。特别是在训练集较小时(每类篇数<10)有明显优势，见表1。对于不同词库规模也有很好的适应性，见图2。

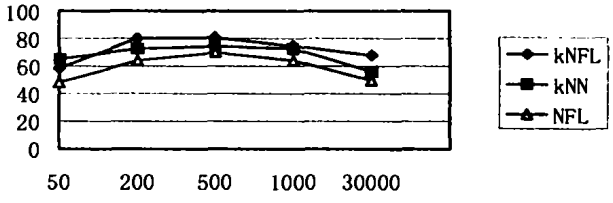

图2 kNFL，kNN，NFL在词库规模变化时的平均精度 p×100

(下转第56页)

```
y.y.y.y.0: udp 0 [ttl ox10] (ttl 49 , id 47349, bad cksum area!)[tos 0x10] (ttl 241, id 17965)
                4510 0038 462d 0000 f101 5da6 yyyy yyyy
                xxxx xxxx 0303 f470 0000 0000 4510 0030
                b8f5 0000 3111 aaea xxxx xxxx yyyy yyyy
                0870 0000 0008 0000
```

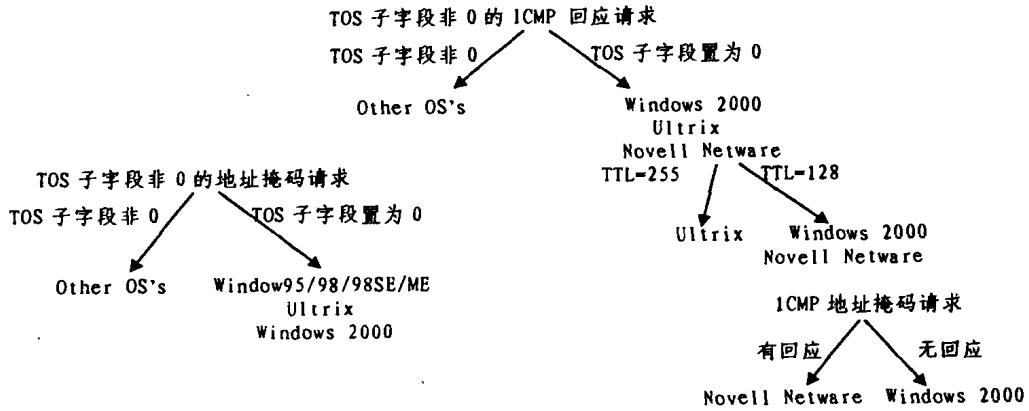

图4 使用 TOS 子字段非0的 ICMP 回应请求等进行指纹探测

从上面的数据包中可以看出原始 IP 报头和 ICMP 端口不可达出错报文数据部分中 IP 报头不同的地方：

1)IP 报头16位的总长度字段 原 UDP 数据报大小为28字节，但回应包引用的出错数据部分 IP 报头中的总长度字段值却为48，多了20个字节。

2)TTL 字段值 回应包引用的出错数据部分 IP 报头中的 TTL 字段值为49，比原 UDP 数据报中的 TTL 值小15，说明中间经过了15跳点。

3)部校验和字段 因为总长度字段和 TTL 字段都变化了，回应包引用的出错数据部分 IP 报头中的首部校验和字段值自然也发生了改变。

4)UDP 首部校验和字段 回应包引用的出错数据部分前64比特数据中的 UDP 首部校验和字段等于0。

在实验中还发现 AIX 4.1在返回的数据包中除了上述第四个不同点外，其余3个不同点都存在。根据这些特征就可以区分不同版本的 AIX 操作系统。

向运行 BSDI 4.x 的机器关闭的 UDP 端口发送 UDP 数据包，监听数据包后发现不同点：

1)IP 报头 16 位的总长度字段 原 UDP 数据报大小为28字节，但回应包引用的出错数据部分 IP 报头中的总长度字段值却为48，多了20个字节。

2)IP TTL 字段值 回应包引用的出错数据部分 IP 报头中的 TTL 字段值为53，中间经过了64-53＝11跳点。

3)IP 首部校验和字段 回应包引用的出错数据部分 IP 报头中的 IP 首部校验和字段等于0。其他操作系统也有各自的特征。

结束语 指纹探测技术是一项非常复杂的技术，它需要了解和掌握各种操作系统之间的细微特征差异，并能够远程捕获这些差异。遵循 ICMP 协议可以构造出种类众多的报文格式，恰好满足了指纹特征多样性的要求，而且各种操作系统的 ICMP 协议栈在实现上确实存在一定的差异。因此，基于 ICMP 协议的指纹探测技术一定会有广阔的发展空间。

参 考 文 献

1 Stevens W R. TCP/IP 详解 卷1：协议. 机械工业出版社，2000

2 Arkin O. Understanding some of the ICMP Protocol's Hazards. 2000

3 Comer D E. 用 TCP/IP 进行网际互联(第一卷). 电子工业出版社，1998

(上接第48页)

表1 算法在不同规模下的平均精度 p×100

训练集大小	平均精度 p×100				
(每类篇数)	KNFL	KNN	NFL	NN	NB
10	72.45	67.34	60.87	45.89	70.43
20	74.64	69.50	70.34	50.56	75.68
30	76.37	75.42	66.45	63.43	78.53
50	80.45	77.56	69.59	68.57	79.23
90	82.57	77.95	67.53	69.54	80.34

结论及今后工作 kNFL 结合了 kNN 与 NFL 的优点，扩大了训练集规模，改进了平均精度，实验表明对于高维样本和训练规模不大时，kNFL 有明显的优势。kNFL 是一种正确率较高的、简单的分类算法。kNFL 和 NFL 可以应用于更多的模式识别的领域，特别是向量空间具有明显连续性的模型中。算法的通用性和应用价值需要进一步在实验中检验。

参 考 文 献

1 边肇祺. 模式识别. 北京：清华大学出版社，第二版，1999. 136～161

2 Dasarathy B V. Nearest Neighbor(NN) Norms: NN Pattern Classification Techniques. McGraw-Hill Computer Science Series. IEEE Computer Society Press, Las Alamitos, California, 1991

3 Lam W, Ho C Y. Using a generalized instance set for automatic text categorization. In: Proc. of the 21th Ann Int ACM SIGIR Conf. on Research and Development in Information Retrieval, 1998. 81～89

4 Vapnic V. The Nature of Statistical Learning Theory. Springer, New York, 1995

5 Rennie J, McCallum A. Efficient Web Spidering with Reinforcement Learning. ICML-99, 1999

6 Li S Z, Lu J. Face Recognition Using the Nearest Feature Line Method. IEEE Trans. Neural Networks, 1999, 10(2): 439～443

7 Wong S G A, Yang C S. A vector space model for automatic indexing. Communications of the ACM, 1975, 18(11): 613～620