一种改进的快速 k-近邻分类算法

乔玉龙[1], 潘正祥[1,2], 孙圣和[1]
(1. 哈尔滨工业大学自动化测试与控制系, 黑龙江哈尔滨 150001; 2 国立高雄应用科技大学电子工程系, 台湾高雄)

摘 要: 本文提出了一种新的有效的 k-近邻分类快速算法. 该算法利用向量的方差和在小波域中的逼近系数得出两个重要的不等式. 在搜索 k-近邻的过程中, 首先判断每个训练向量是否满足这两个不等式, 由此排除大量不可能成为 k-近邻的向量, 从而可以快速的找到未知样本的 k 个近邻, 使得在保持 k-近邻法分类性能不变的情况下, 分类的效率得到很大地提高. 最后, 我们以纹理分类为例验证算法的有效性.

关键词: k-近邻; 小波变换; 纹理分类
中图分类号: TP391. 4 文献标识码: A 文章编号: 0372-2112 (2005) 06-1146-04

Improved K Nearest Neighbors Classification Algorithm

QIAO Yu-long[1], PAN Jeng-shyang[1,2], SUN Sheng-he[1]
(1. *Dept. of Automatic Test and Control, Harbin Institute of Technology, Harbin, Heilongjiang 150001, China;*
2. *Dept. of Electronic Engineering, National Kaoksiung University of Applied Science, Taiwan, China*)

Abstract: A novel and efficient algorithm is proposed to reduce the computational complexity for KNN classification. It uses two important features, the approximation coefficient of a fully decomposed feature vector with Haar wavelet and variance of the corresponding untransformed vector, to produce two efficient test conditions. Since those vectors that are impossible to be the k closest vectors in the design set are kicked out quickly by these conditions, this algorithm saves largely the classification time and has the same classification performance as that of the exhaust search classification algorithm. Experimental results based on texture image classification will verify our proposed algorithm.

Key words: *K* nearest neighbors; wavelet transform; texture classification

1 引言

k-近邻法是一种非常简单有效的分类方法, 广泛应用于模式识别的各个领域. 从字义上看, 该方法就是找出未知样本 x 的 k 个近邻, 看这 k 个近邻中多数属于哪一类, 就把 x 归为那一类. 具体的说, 假设有 L 个类 $c_1, c_2, \cdots, c_L$, 第 i 个类的训练样本集为 w_i, 整个训练样本集为 $\bigcup_{i=1}^{L} w_i$, 样本总数为 Ω, $y_i (i = 1, 2, \cdots, \Omega)$ 表示第 i 个训练样本. 给定未知样本 x 和距离测度, 首先从 Ω 个训练样本中找出 x 的 k 个近邻, $k_i (1 \leqslant i \leqslant L)$ 表示这 k 个近邻中属于第 i 类的样本数, 那么把 x 归为类 c_I, 其中 $I = \arg\max_i k_i$, 这就是所谓的 k-近邻规则(分类方法). 在这篇文章中, 我们用向量表示样本或者样本的特征向量, 分类中采用 Euclidean 距离.

虽然 k-近邻法应用很广, 但它一个最大的缺点就是计算复杂性高. 为了从训练样本中找到 x 的 k 个近邻, 不得不计算 x 与所有样本间的距离, 因为这个很大的计算负担, k-近邻法在实时性很强的场合应用很少. 为了提高 k-近邻法的效率, 已有很多文献[1~5]提出减少计算负担的算法. 这些算法大致可分为两类, 第一类是在保持或近似保持分类性能不变的条件下, 减少训练样本集的大小[1,2]; 第二类是在保持分类性能不变的情况下, 采用快速算法[3~5], Huang 和 Wen[4] 提出了一种基于小波域部分距离搜索的 k-近邻搜索算法(WKPDS), 对一个待分类的未知样本, 此算法在小波域中利用部分距离搜索方法快速地找到 k 个近邻, 从而适当地减少计算复杂性. 文献[5]提出快速算法(KWENNS), 在搜索 k 个近邻前, 先将特征向量变换到小波域, 再利用两个向量逼近系数和距离间的关系这一重要特征排除大部分不可能是 k 个近邻的向量, 进而大大提高分类速度. 显然, 也可以将第一类算法和第二类算法相结合, 进一步提高分类的效率. 因此对第二类算法的研究就显得更为重要.

本文利用向量的方差和在小波域中的逼近系数这两个重要特征, 提出了一种新的快速分类算法, 并以纹理分类为例, 从分类所需时间和找到 k 个近邻所需的乘法、加法、比

收稿日期: 2004-08-07; 修回日期: 2005-02-20

较、开方运算量这两个方面验证算法的有效性.

2 先前的一些算法

从对 k- 近邻方法的描述我们看到, 完成分类最重要的就是要找到 k 个近邻, 如果用穷尽搜索, 那就得计算未知样本与所有训练样本间的距离, 当训练样本数较多或者表示样本的特征向量的维数较高时, 计算复杂性就越高. 因此提出许多快速搜索的方法, 进而提高分类速度.

在着重介绍小波域中的快速算法之前, 首先简要介绍一下小波变换.

2.1 小波变换

在过去的十几年里, 小波变换一直是数学和工程的研究焦点, 并且已被成功地应用到各个领域, 特别是信号处理. 本小节介绍一些必要的知识, 有关小波变换的详细内容, 请参见文献[6, 7].

Harr 小波可以按下面的公式分解向量 $\boldsymbol{x}$,

$$X_{s_{j,n}} = \frac{1}{\sqrt{2}}[X_{s_{j-1,2n}} + X_{s_{j-1,2n+1}}]$$

$$X_{d_{j,n}} = \frac{1}{\sqrt{2}}[X_{s_{j-1,2n}} - X_{s_{j-1,2n+1}}]$$

其中 $X_{s_{j,n}}$ 和 $X_{d_{j,n}}$ 分别表示尺度为 j 时的逼近系数和细节系数, 原向量 $\boldsymbol{x}$ 就是尺度为零时的逼近系数, 当 $\boldsymbol{x}$ 的维数 N 满足 $2^{n-1} < N \leqslant 2^n$ 时, 经过 n 层小波变换, $\boldsymbol{x}$ 可以被完全分解, 此时唯一的逼近系数表示为 C_x, 逼近系数和细节系数组成小波系数向量. 当向量的维数不是 2 的幂次时, 进行小波变换时的边界处理可参见文献[5]. 由 Harr 小波的正交性, 对任意两个向量 $\boldsymbol{x}$ 和 $\boldsymbol{y}$ 有 $d(\boldsymbol{x}, \boldsymbol{y}) = d(\boldsymbol{X}, \boldsymbol{Y})$, 其中 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 分别是 $\boldsymbol{x}$ 和 $\boldsymbol{y}$ 小波系数向量, $d(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{\sum_{i=0}^{N-1}(x_i - y_i)^2}$ 是 $\boldsymbol{x}$ 和 $\boldsymbol{y}$ 间的 Euclidean 距离, 因此, 在时域中找的 k 个近邻与在小波域找到的相同, 这是小波域算法的根本出发点.

2.2 WKPDS 算法

Hwang 和 Wen[4] 提出了一种基于小波域部分距离搜索的 k- 近邻搜索算法(WKPDS). 因为小波变换后, 逼近系数包含重要的信息(文献[5]给出了详细地分析), 并且小波变换使得特征向量的能量集中到少数系数上, WKPDS 可以快速地找到未知样本的 k 个近邻, 所以在保持分类性能不变的情况下, 大大节约了分类时间. WKPDS 的基本思想是: 在计算某个训练向量与未知向量间的距离的过程中, 始终判断累加的部分距离是否已超过目前 k 个近邻中与未知向量相距最大的那个距离, 一旦超出则终止距离的计算. 具体地说, 对任意的输入向量 $\boldsymbol{x}$, 假设当前 k 个近邻为 $\boldsymbol{y}_{j_1}, \cdots, \boldsymbol{y}_{j_k}$, 相应的距离为 $d(\boldsymbol{X}, \boldsymbol{Y}_{j_1}) \cdots d(\boldsymbol{X}, \boldsymbol{Y}_{j_k})$, 并且 $d(\boldsymbol{X}, \boldsymbol{Y}_{j_1}) \leqslant d(\boldsymbol{X}, \boldsymbol{Y}_{j_2}) \leqslant \cdots \leqslant d(\boldsymbol{X}, \boldsymbol{Y}_{j_k})$. 如果训练样本 $\boldsymbol{y}_j$ 满足

$$\sum_{i=0}^{s}(X_i - Y_{ji})^2 \geqslant d^2(\boldsymbol{X}, \boldsymbol{Y}_{j_k}),\ s \leqslant l-1$$

$$\text{那么 } d(\boldsymbol{X}, \boldsymbol{Y}_j) \geqslant d(\boldsymbol{X}, \boldsymbol{Y}_{j_k}) \qquad (1)$$

其中, $\boldsymbol{X}, \boldsymbol{Y}_j, \boldsymbol{Y}_{j_1}, \cdots, \boldsymbol{Y}_{j_k}$ 分别是 $\boldsymbol{x}, \boldsymbol{y}_j, \boldsymbol{y}_{j_1}, \cdots, \boldsymbol{y}_{j_k}$ 的小波系数向量, l 是小波系数向量的维数, 说明 $\boldsymbol{y}_j$ 不可能是 k 个近邻中的一个. 虽然 WKPDS 不是很有效, 但它可以和其它方法结合来提高分类效率.

2.3 WKENNS 算法

令 $\boldsymbol{x} = \{x_0, x_1, \cdots, x_{N-1}\}$, $\boldsymbol{x}$ 的均值为 $M_x = \frac{1}{N}\sum_{i=0}^{N-1} x_i$. 对任意两个向量 $\boldsymbol{x}$ 和 $\boldsymbol{y}$, 文献[8]中指出 $d(\boldsymbol{x}, \boldsymbol{y}) \geqslant \sqrt{N}|M_x - M_y|$, 其中 $d(\boldsymbol{x}, \boldsymbol{y})$ 表示 $\boldsymbol{x}$ 和 $\boldsymbol{y}$ 间的 Euclidean 距离. Pan 和他的合作者[5]分析了向量被 Harr 小波完全分解后的逼近系数与均值间的关系, 并得出下面的结论: 对 N 维向量 $\boldsymbol{x}$, 如果 $2^{n-1} < N \leqslant 2^n$, 那么经过 n 级小波分解, 有 $C_x = N \cdot M_x \backslash \sqrt{2^n}$, 结合以上结果, 进一步得出: 如果训练样本 $\boldsymbol{y}_j$ 满足,

$$C_{y_j} \leqslant C_x - \sqrt{N/2^n}d(\boldsymbol{X}, \boldsymbol{Y}_{j_k}) \text{ 或者 } C_{y_j} \geqslant C_x + \sqrt{N/2^n}d(\boldsymbol{X}, \boldsymbol{Y}_{j_k})$$

$$\text{那么 } d(\boldsymbol{X}, \boldsymbol{Y}_j) \geqslant d(\boldsymbol{X}, \boldsymbol{Y}_{j_k}) \qquad (2)$$

其中 C_x, C_{y_j} 分别表示用 Harr 小波完全分解 $\boldsymbol{x}$ 和 $\boldsymbol{y}_j$ 后得到的逼近系数, $\boldsymbol{X}$, $\boldsymbol{Y}_j$, $\boldsymbol{Y}_{j_k}$ 分别为 $\boldsymbol{x}$, $\boldsymbol{y}_j$, $\boldsymbol{y}_{j_k}$ 的小波系数向量. 这就是说, 当判断 $d(\boldsymbol{X}, \boldsymbol{Y}_j)$ 与 $d(\boldsymbol{X}, \boldsymbol{Y}_{j_k})$ 的关系时, 我们可以先判断(2)是否成立, 如果成立, 那么 $\boldsymbol{y}_j$ 不可能是 k 个近邻中的一个, 通过检验不等式使得大部分不可能是 k- 近邻的向量被排除, 减少了距离的计算, 从而大大提高了分类的速度.

3 本文的算法

令 $\boldsymbol{x}$ 的方差为 $V_x = \sqrt{\sum_{i=0}^{N-1}(x_i - M_x)^2}$, 文献[9]证明 $d(\boldsymbol{x}, \boldsymbol{y}) \geqslant |V_x - V_y|$, 也就是说, 对某一向量 $\boldsymbol{z}$, 如果 $V_z \leqslant V_x - d(\boldsymbol{x}, \boldsymbol{y})$ 或者 $V_z \geqslant V_x + d(\boldsymbol{x}, \boldsymbol{y})$, 那么 $d(\boldsymbol{x}, \boldsymbol{z}) \geqslant d(\boldsymbol{x}, \boldsymbol{y})$, 这样, 在判断 $d(\boldsymbol{x}, \boldsymbol{z})$ 与 $d(\boldsymbol{x}, \boldsymbol{y})$ 的关系时, 如果上面的不等式成立, 就不必计算 $d(\boldsymbol{x}, \boldsymbol{z})$, 而快速的得到 $d(\boldsymbol{x}, \boldsymbol{z}) \geqslant d(\boldsymbol{x}, \boldsymbol{y})$, 所以此不等式可用来提高 k- 近邻法的效率.

对任意输入向量 $\boldsymbol{x}$, 假设在已搜索过的训练向量中 $\boldsymbol{y}_{j_1}, \cdots, \boldsymbol{y}_{j_k}$ 为当前的 k 个近邻, 相应的距离为 $d(\boldsymbol{X}, \boldsymbol{Y}_{j_1}) \cdots d(\boldsymbol{X}, \boldsymbol{Y}_{j_k})$, 并且 $d(\boldsymbol{X}, \boldsymbol{Y}_{j1}) \leqslant (\boldsymbol{X}, \boldsymbol{Y}_{j2}) \leqslant \cdots \leqslant d(\boldsymbol{X}, \boldsymbol{Y}_{j_k})$, 为了判断新的训练向量 $\boldsymbol{y}_j$ 与 $\boldsymbol{x}$ 的距离是否小于 $d(\boldsymbol{X}, \boldsymbol{Y}_{j_k})$, 本文的新算法主要包含三步: 首先判断是否 $C_{y_j} \leqslant C_x - \sqrt{N/2^n}d(\boldsymbol{X}, \boldsymbol{Y}_{j_k})$ 或者 $C_{y_j} \geqslant C_x + \sqrt{N/2^n}d(\boldsymbol{X}, \boldsymbol{Y}_{j_k})$ 成立, 如果成立, 那么 $d(\boldsymbol{X}, \boldsymbol{Y}_j) \geqslant d(\boldsymbol{X}, \boldsymbol{Y}_{j_k})$, 即 $\boldsymbol{y}_j$ 不可能是 $\boldsymbol{x}$ 的 k- 近邻, 否则, 进行第二步, 检验 $V_{y_j} \geqslant V_x + d(\boldsymbol{X}, \boldsymbol{Y}_{j_k})$ 或者 $V_{y_j} \leqslant V_x - d(\boldsymbol{X}, \boldsymbol{Y}_{j_k})$ 是否成立, 如果成立, 则 $d(\boldsymbol{X}, \boldsymbol{Y}_j) \geqslant d(\boldsymbol{X}, \boldsymbol{Y}_{j_k})$, 那么 $\boldsymbol{y}_j$ 被排除在 $\boldsymbol{x}$ 的 k- 近邻之外. 最后, 如果此训练向量不能用以上两步排除, 我们采用 WKPDS 判断向量 $\boldsymbol{y}_j$ 与 $\boldsymbol{x}$ 的距离与 $d(\boldsymbol{X}, \boldsymbol{Y}_{j_k})$ 间的大小关系.

有关算法涉及到的初始化以及详细步骤可参见文[5]. 我们注意到新算法利用两个重要的特征: 向量的方差和在小波域中的逼近系数, 并且算法是在小波域中进行的, 所以在搜索未知向量的 k- 近邻前, 首先要计算未知向量的方差, 并将其变换到小波域. 同时在利用新的算法前, 必须进行必要的预处理.

预处理步骤 1: 计算所有训练向量 $\boldsymbol{y}_i$ 的方差 V_{y_i};

预处理步骤 2: 用 Harr 小波完全分解所有的训练向量, 并且按照它们的逼近系数从小到大重新排列所有的训练向量 (文献[5]中表明这步的好处).

这些预处理步骤都可以离线完成, 因此, 可将它们看成分类器的训练.

本文的算法在保持分类性能不变的情况下, 能快速的完成分类, 和算法 WKENNS 相比, 新算法增加了一步判断, 排除许多不可能成为 k- 近邻的向量. 本文的算法看起来和文献[9]提出的算法相似, 但是新算法的目的是结合时域和小波域的消除条件以提高 k- 近邻分类方法的效率.

4 实验结果

本节以纹理分类为例验证新算法的效率, 实验数据包括文献[10]中的 30 幅 Brodatz 纹理图像 $D3$、$D6$、$D9$、$D12$、$D15$、$D16$、$D18$、$D19$、$D24$、$D29$、$D34$、$D38$、$D41$、$D47$、$D48$、$D49$、$D54$、$D60$、$D65$、$D68$、$D72$、$D84$、$D85$、$D88$、$D92$、$D94$、$D98$、$D102$、$D107$、$D112$, 这些图像可从互联网[11, 12]下载. 每个图像的大小为 512× 512, 从每幅图像中随机选择 400 个大小为 128 × 128 的子图像, 其中 300 幅作为训练样本图像, 其他的 100 幅用来测试算法. 文献[13]将小波系数极值的密度作为特征来分割纹理图像, 本文利用此特征进行纹理分类, 并用样条小波[14]提取特征. 我们分别从分类时间和所用内存这两个方面进行比较, 实验结果列于表 1 和 2 中.

表 **1** 特征向量维数为 **8** 时的比较结果

算法	内存	$k=5$		$k=3$	
		误分数	时间(s)	误分数	时间(s)
WKPDS	8	13	21.41	13	21.10
WKENNS	8	13	1.81	13	1.49
本文算法	9	13	0.99	13	0.76

表 **2** 特征向量维数为 **13** 时的比较结果

算法	内存	$k=5$		$k=3$	
		误分数	时间(s)	误分数	时间(s)
WKPDS	15	4	28.59	5	27.87
WKENNS	15	4	4.18	5	3.58
本文算法	16	4	2.3	5	1.82

当特征向量的维数是 8 时, 相对算法 WKPDS 和 WKENNS 而言, 本文的算法分别平均减少了 96% 和 47% 的分类时间; 当特征向量的维数为 13 时, 新算法的分类时间平均为 WKPDS 的 7%, WKENNS 的 53%. 虽然新算法需要少量的额外内存, 但它可以大大减少分类时间, 同时保持分类性能不变.

为了进一步验证算法的有效性, 表 3 列出了当特征向量维数分别为 8 和 13 时, 搜索最近邻所需乘法(Mul)、加法(Add)、比较(Com)和开方(Squ)算术运算量的平均结果.

表 **3** 搜索最近邻所需算术运算量的比较结果

维数	算法	Mul	Add	Com	Squ
8	WKPDS	12471.3	33947.8	21456.3	0
	WKENNS	1050.2	2094.4	1494.9	6.0
	本文算法	346.4	718.8	1656.5	13.0
13	WKPDS	17146.4	43296.4	26117.4	0
	WKENNS	2510.5	5008.0	3443.6	6.5
	本文算法	962.1	1959.1	3577.8	13.9

从表 3 可以看出, 和 WKPDS 相比, 本文的算法虽然增加了极少量的开方运算, 但却减少了大量的乘法和加法. 相对 WKENNS 而言, 新算法以少量的比较和开方运算为代价, 大大减少了乘法和加法运算.

无论是从分类所需的时间还是从算术运算量来看, 新算法都是非常有效的.

5 结论

为了减少 k- 近邻分类算法的计算负担, 本文提出了一种快速算法. 新算法利用两个消除条件排除许多不可能是 k 个近邻的训练样本, 同时利用 WKPDS 判断不能被消除条件排除的那些向量, 从而快速地完成分类. 同先前的算法相比, 虽然新算法需要少量的额外内存, 但它可以大大减少分类时间, 同时保持分类性能不变.

参考文献:

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

[2] Q B Xie, C A Laszlo, R K Ward. Vector quantization technique for nonparametric classifier design[J]. IEEE Trans Pattern Anal Machine Intell, 1993, 15(12): 1326- 1330.

[3] K Fukunaga, P M Narendra. A branch and bound algorithm for computing k- nearest neighbors[J]. IEEE Trans Computers, 1975, 24(7): 750 - 753.

[4] W J Hwang, K W Wen. Fast KNN classification algorithm based on partial distance search[J]. Electron Lett, 1998, 34(21): 2062- 2063.

[5] J S Pan, Y L Qiao, S H Sun. A fast K nearest neighbors classification algorithm[J]. IEICE Trans Fundamentals, 2004, E87-A(4): 961 - 963.

[6] S Mallat. A theory of multiresolution signal decomposition: the wavelet representation[J]. IEEE Trans Patt Anal and Mach Intell, 1989, 11(7): 674- 693.

[7] M Vetterli, J Kovacevic. Wavelet and Subband Coding[M]. NJ: Prerrtice- Hall, Englewood Cliffs, 1995.

[8] L Guan, M Kamel. Equal-average hyperplane partitioning method for vector quantization of image data[J]. Pattern Recognit Lett, 1992, 13(10): 693- 699.

[9] C H Lee, L H Chen. Fast closest codeword search algorithm for vector quantization[J]. IEE Proc - Vis Image Signal Process, 1994, 141(3): 143- 148.

[10] P Brodatz. Textures: A Photographic Album for Artists & Designers

[M]. Dover, New York, 1966.

[11] Brodatz Textures By Trygve Randen[DB/OL]. http://www.ux.his.no/~tranden/brodatz.html.

[12] USC-SIPI Image Database[DB/OL]. http://sipi.usc.edu/services/database/database.cgi?volume=textures.

[13] J S Pan, J W Wang. Texture segmentation using separable and non-separable wavelet frames[J]. IEICE Trans Fundamentals, 1999, E82-A(8): 1463-1474.

[14] S Mallat, S Zhong. Characterization of signals from multiscale edges [J]. IEEE Trans Patt Anal and Mach Intell, 1992, 14(7): 710-732.

作者简介:

乔玉龙　男, 2000 年 7 月获得哈尔滨工业大学信息与计算科学专业学士学位, 2002 年 7 月获得哈尔滨工业大学计算数学专业硕士学位, 现在为哈尔滨工业大学自动化测试与控制系博士研究生, 目前主要致力于小波变换、图像处理、纹理分析、模式识别和机器学习的研究. Email: qiaoyulong@dsp.hit.edu.cn

潘正祥　男, 1996 年获英国爱丁堡大学博士学位, 现任台湾国立高雄应用科学技术大学电子工程系教授, 并被哈尔滨工业大学自动化测试与控制系聘为特聘教授, 他在国际上发表了 40 多篇期刊文章和 90 多篇会议论文, 目前主要致力于数据挖掘、信息安全和图像处理研究.

孙圣和　男, 哈尔滨工业大学自动化测试与控制系教授, 博士生导师, 曾任国务院仪器科学与技术学科评议组成员, 现任测控所所长, 兼任中国计量学会常务理事、电子计量专业委员会主任、中国电子学会电子测量与仪器分会副主任、电子测量与仪器学报编委会主任委员等职务, 发表学术论文 240 篇, SCI 和 EI 收录 164 篇, 出版专著 6 部, 目前的研究领域包括自动化测试系统, 电子系统故障模拟、诊断、检测与排除, 图像处理与模式识别, 信息安全技术及应用.