第15卷 第2期 自 动 化 学 报 Vol. 15, No. 2
1989 年 3 月 ACTA AUTOMATICA SINICA Mar., 1989

加权 K-NN 分类器及其应用

周 伟
（杭州电子工业学院）
易 丹 波
（杭州通信广播电视技术研究所）

摘 要

本文介绍了加权 K-NN 分类器的基本原理，并给出了相应的算法，着重对权函数的形式进行了讨论，提出了两种新的权函数定义公式. 将加权 K-NN 分类器用在白血球自动分类系统中，分类精度比 K-NN 分类器有明显提高.

关键词：K-NN 分类器；权函数；加权 K-NN 分类器；白血球；二分树.

K-NN 分类器是较为常用的一种非参数分类器，在某些场合下应用时，能得到较为满意的结果[1]. 仔细考察 K-NN 分类器，可发现它有明显的缺陷：K-NN 分类器没有考虑待识样本与 K 个最近邻标准样本(训练集)的相似程度. 显然，一个标准样本与待识样本越相似，这个标准样本应该在分类时起的作用也越大，反之亦然. 为此，有人提出了加权 K-NN 分类器[2].

一、加权 K-NN 分类器

针对 K-NN 分类器存在的不足，人们提出了相应的改进措施，最常用的是距离加权 K-NN 分类器.

1. 加权 K-NN 分类器的基本原理及算法

假设：一训练样本空间 Ω，共分为 M 类，$\Omega = \Omega_1 \cup \Omega_2 \cdots \cup \Omega_M$，已知它们的类别属性，训练样本集可表示为 $X = \{x^1, x^2, \cdots x^N\}$，其中 N 为训练样本总个数，可写成 $x^n = x_m^i \in \Omega_m$. $m = 1, 2, \cdots M$; $i = 1, 2, \cdots N_m$; $n = 1, 2, \cdots N$，且有 $\sum_{m=1}^{M} N_m = N$，待识样本为 $y \in \Omega$.

进一步假设：待识样本 Y 在训练样本空间 Ω 中的 K 个最近邻按其属性可表示为 x_m^v，其中 $v = 1, 2, \cdots K_m$; $m = 1, 2, \cdots M$ (排除 $K_m = 0$ 的情况)，且有 $\Sigma K_m = K$，与 x_m^v 相应的权为 W_m^v.

本文于 1987 年 5 月 9 日收到.

加权 K-NN 分类器的基本思想为，赋予与待识样本 y 更接近的标准样本比较远的标准样本有更大的“权”，也就是说，在分类中起更大的作用．这个“权”函数随待识样本与标准样本之间距离的变化而变化，更确切地讲，定义权函数是距离度量 $d(y, x_m^i)$ 的减函数，即有：$W_m^i = W[d(y, x_m^i)]$．

加权 K-NN 分类器的具体算法如下：

第一步　输入待识样本 y，设置K值，$1 \leqslant K < N$；

第二步　设起始值 $n = 1$；

第三步　计算待识样本 y 与 x^n 之间的距离，

IF($n \leqslant K$)THEN 将 x^n 归入到 y 的K个最近邻中，

ELSE IF (x^n 比原先的K个最近邻更接近于 y)，THEN 用 x^n 替换K个最近邻中的最远者，置 $n = n + 1$；

第四步　IF($n \leqslant N$) THEN 转至第三步；

第五步　依据距离度量 $d(y, x_m^v)$ 确定权函数 W_m^v；

第六步　计算判决函数 $C(\Omega_m/y) = \sum_{v=1}^{K_m} W_m^v$；

第七步　IF[$C(\Omega_m/y) > C(\Omega_\lambda/y)$ 对于所有 $\lambda \neq m$，$\lambda = 1, 2, \cdots M$（排除 $K_\lambda = 0$ 的情况）成立]，

THEN 将 y 分类到 Ω_m 中去；

第八步　IF[出现 $C(\Omega_m/y)$ 的最大值多于一个的情况]，

THEN 将 y 分类到首先找到的具有最大 $C(\Omega_m/y)$ 值的类别 Ω_m 中去；

第九步　结束．

2. 几种权函数的定义式

上面给出了加权 K-NN 分类器的具体算法，其中一个很关键的问题是如何根据距离度量 $d(y, x_m^v)$ 来确定权函数值 W_m^v．文献[2]中介绍了三种权函数的定义式，设K个最近邻中的最远点和最近点的距离分别为 $d_{\max}$，$d_{\min}$．

第一种定义式为

$$W_m^v = \begin{cases} \dfrac{d_{\max} - d(y, x_m^v)}{d_{\max} - d_{\min}}, & 1 < K < N, \\ 1, & K = 1. \end{cases}$$

第二种定义式为

$$W_m^v = \frac{1}{d(y, x_m^v)}, \quad \text{要求 } d(y, x_m^v) \neq 0.$$

第三种定义式为

$$W_m^v = K - j + 1.$$

式中 $j(1 \leqslant j \leqslant K)$ 为K个最近邻按距离度量大小排队的序号，$j = 1$ 对应最近样本，$j = K$ 对应最远样本．

以上介绍的三种权函数定义式均可用在加权 K-NN 分类器中，但都有不足之处，三种定义式随距离增大而递减的速度是恒定的，然而，具有可调递减速度的权函数定义式使

用起来更方便.

下面给出两种更灵活合理的权函数定义式,假设同前.

第四种定义式为

$$W_m^v = \begin{cases} \exp\left\{\alpha\left[\dfrac{d_{\max} - d(y, x_m^v)}{d_{\max} - d_{\min}}\right]\right\}, & 1 < K < N, \\ 1, & K = 1. \end{cases}$$

式中 $\alpha(0 < \alpha < \infty)$ 用来调节权函数的递减速度.

第五种定义式为

$$W_m^v = \begin{cases} \exp\{-\alpha[d(y, x_m^v) - d_{\min}]\}, & 1 < K < N, \\ 1, & K = 1. \end{cases}$$

二、加权 K-NN 分类器的应用

文献 [1] 中介绍了一种在微机上实现白血球自动分类的方法,建立了细胞个数均为97的训练集和考试集各一个,自动分类系统采用 K-NN 分类器,定义距离度量 $d(y, x_m^i)$ 如下:

$$d(y, x_m^i) = (y - x_m^i)^T \Sigma^{-1} (y - x_m^i).$$

式中 Σ 为第 m 类标准样本的协方差矩阵. 适中选择 $K = 5$, 分类结果较为满意, 如果采

表1 不同分类器下的精度比较 ($K = 5$)

精度%/条件		特征数 3	4	5	6	7	8
训练集	K-NN	78.35	79.38	80.41	83.50	85.57	85.57
	加权 K-NN 权函数一	85.57	86.60	88.66	88.66	90.72	90.72
	加权 K-NN 权函数二	79.38	80.41	81.44	83.51	85.57	87.63
	加权 K-NN 权函数三	78.35	78.35	82.47	84.54	86.60	87.63
	加权 K-NN 权函数四	86.60	88.66	91.75	92.78	92.78	93.81
	加权 K-NN 权函数五	85.57	86.66	89.69	91.75	93.81	93.81
考试集	K-NN	75.26	77.32	77.32	78.35	78.35	78.35
	加权 K-NN 权函数四	81.44	80.41	83.51	83.51	85.57	86.60
	加权 K-NN 权函数五	82.47	84.54	84.54	86.60	87.63	88.66

表2 K 值对精度的影响(入选特征数为七)

精度/条件 \ K值	3	4	5	6	7	8	9	10	11	12
K-NN	79.38	79.38	80.41	80.41	81.44	81.44	80.41	78.35	77.32	77.32
加权 K-NN, 权函数四	89.69	89.69	91.75	91.75	92.78	92.78	93.81	93.81	93.81	93.81

用本文介绍的加权 K-NN 分类器，分类精度可望得到提高.

表 1 给出了利用加权 K-NN 分类器的分类结果. 为了便于比较，同时还给出了利用 K-NN 分类器的结果. 可见，加权 K-NN 分类器能改善分类精度. 本文提出的两种权函数定义式效果则更好，考试集上的精度比训练集略低一些. 对应于第四、第五种定义式，分别选择 $\alpha = 2$ 和 $\alpha = 0.2$. 表 2 给出了在训练集上 K-NN 分类器和加权 K-NN 分类器在不同K值下的分类精度. 可见，对于 K-NN 分类器，当 $K \geqslant 9$ 时，分类精度反而下降，而对于加权 K-NN 分类器，分类精度是K值的不减函数，故K值的选择较为方便.

本文介绍的加权 K-NN 分类器以极小量的速度下降换取了分类精度的可观提高，这显然是可取的.

参　考　文　献

[1] 周伟、王承训，白细胞自动分类问题的研究，杭州电子工业学院学报，2(1986)，93—100.
[2] Dudani, S. A., The Distance-Weighted K-Nearest-Neighbor Rule, *IEEE Trans.*, **SMC-15** (1976), 325—327.

A WEIGHTED K-NN CLASSIFIER AND ITS APPLICATION

ZHOU WEI

(*Hangzhou Institute of Electronic Engineering*)

YI DANBO

(*Hangzhou Institute of Communication, Broadcasting and Television Technology*)

ABSTRACT

This paper introduces the basic principle of a weighted-K-NN classifier and presents the corresponding algorithm. The form of the weight function are discussed emphatically, and two new definition formulas of weighted function are produced. The weighted-K-NN Classifier has been applied to a system for the automated classification of white blood cells, and the classification accuracy is much higher than the K-NN Classifier.

Key words ——K-NN classifier; weighted function; weighted-K-NN classifier; white blood cells; binary tree.