增刊 第1期
1985年8月

自 动 化 学 报
ACTA AUTOMATICA SINICA

11卷

关于K最近邻判决规则错误率的新上界

俞 庠
（北京工业大学）

摘 要

本文得出在二类情况下K最近邻判决规则错误率R_K的上界的闭合形式表达式

$$R_K \leqslant (1+C_K)R^*.$$

式中 $C_K = 1/(2.898\sqrt{K} - 1.898)$，$R^*$是Bayes判决规则错误率。这个表达式不但包含最近邻判决规则错误率R_1的上界表达式$R_1 \leqslant 2R^*$，而且所有的奇数K的数值与级数形式的上确界非常接近。

一、引 言

在统计模式识别中，最重要的分类规则是 Bayes 判决规则和K最近邻判决规则[1]。虽然 Bayes 判决规则有最小的错误率，但是它要求已知先验概率 $P(\omega_i)$ 和条件概率密度 $p(X|\omega_i)$。这里 X 表示待分类的未知模式，ω_i 表示第i个类别，$i = 1,2,\cdots,m$。同时，在 $p(X|\omega_i)$ 中，还包含若干未知的参数。因此，在许多实际问题中，通常使用非参数形式的K最近邻判决规则。

这个判决规则不需要任何先验知识，因而比较实用。可是，它的错误率R_K总是比Bayes判决规则的错误率R*大，R*是R_K的下界。确定R_K的上界是相当困难的。为此，人们致力于寻找R_K与R*之间的关系，以确定用R*表示的R_K的上界。

Cover和Hart早已证明[2]，在m类问题中，当样品数目n 很大时，由最近邻判决规则引起的错误率R_1小于Bayes判决规则错误率R*的两倍，即

$$R_1 \leqslant (2 - \frac{m}{m-1}R^*)R^* < 2R^*. \qquad (1)$$

他们还证明R_K的上确界具有如下的形式：

$$R_K \leqslant (1+c_k)R^*. \qquad (2)$$

式中 c_k 是 K 的一个常数序列。当 m = 2 和 K = 1 时，由不等式（1）可知，$c_1 = 1$；当 $K > 1$ 时，$c_k < 1$。后来，Devijver在二类情况下证明了一个不紧凑的上界[3]，即

$$R_K \leqslant \left(1 + 2\frac{K!!}{(K+1)!!}\right) R^*. \qquad (3)$$

式中 符号!!表示半阶乘，最近邻数目K为任何正整数。

本文于1983年10月 8 日收到。

最近，Devroye在K≥5且为奇数的二类情况下，求得R_K的比较紧凑的上界[4]，即

$$R_k \leqslant \left[1+\alpha\frac{\sqrt{K-1}}{K-3}\left(1+\frac{\beta}{\sqrt{K-3}}\right)\right]R^*. \tag{4}$$

式中常数$\alpha = 0.3399424150\cdots$，$\beta = 0.9749687445\cdots$。

二、结 果

在二类情况下，K为奇数时，用R^*表示的R_k的级数形式上确界为[5]

$$R_k \leqslant (1+c_k)R^*.$$

式中

$$C_k = \sup\left\{\frac{1-2R^*}{R^*}\sum_{i=0}^{\frac{K-1}{2}}\binom{K}{i}(1-R^*)^i R^{*k-i}\right\}. \tag{5}$$

当K为固定值时，先把上式右边大括弧里的函数对R^*求导，算出使C_k为极大值的R^*，再代入这个等式，就能得到C_k，进而得到R_k的上界。然而，在实际应用中，K通常很大，利用等式（5）求出C_k是很麻烦的。

为此，首先用等式（5）算出前十一个C_k，K = 1,3,…,21，见表1。

表1

K	1	3	5	7	9	11	13	15	17	19	21
C_k	1	0.316	0.218	0.173	0.147	0.130	0.117	0.107	0.100	0.093	0.088

然后运用最小二乘法，以双曲线型函数拟合上述数据点，最后求得反映C_k与K之间关系的闭合形式的近似公式。假设

$$C_k = 1/(a+b\sqrt{K}), \tag{6}$$

为了确定两个待定常数a和b，令

$$x = \sqrt{K},\quad y = 1/C_k, \tag{7}$$

这样，等式（6）变成

$$y = a + bx. \tag{8}$$

于是求解的问题被归结为用x的线性函数y(x)来拟合由代换（7）算出的新的数据点集（见表2）。

表2

x_i	1	1.732	2.236	2.646	3	3.317	3.606	3.873	4.123	4.359	4.583
y_i	1	3.165	4.587	5.780	6.803	7.692	8.547	9.346	10.000	10.753	11.364

由此得到最小二乘法的方程组为

$$\sum_{i=1}^{11} w_i[(a+bx_i)-y_i] = 0, \tag{9}$$

$$\sum_{i=1}^{11} w_i\{(a+bx_i)-y_i\}x_i=0\,. \tag{10}$$

为了确保拟合曲线通过数据点$x_1=1$，$y_1=1$，假定权w_i是未知的，而$w_i=1$，这里$i=2,3,\cdots,11$，并且增加一个关系式

$$a+b=1\,, \tag{11}$$

解（11）和（9）或（10）式的联立方程组，得

$$a=-1.898\cdots,\quad b=2.898\cdots$$

最后得到公式

$$C_k=\frac{1}{2.898\sqrt{K}-1.898}\,. \tag{12}$$

三、讨　　论

为了对各种方法所得到的K最近邻判决规则错误率 R_k 的上界的紧凑程度作一比较，可以认为等式（5）的结果是最紧凑的，因而可以看作上确界。同时，算出式（3），（4）和(12) 中的C_k，一并列于表3。另外，使等式（5）的C_k为极大值的R^*各值也列于表3。

表3　C_k 的各种上界对比

K	C_k				R^*
	上确界	式(12)	式(4)	式(3)	
1	1	1	—	1	0
3	0.316	0.320	—	0.750	0.226
5	0.218	0.218	0.574	0.625	0.295
7	0.173	0.173	0.310	0.547	0.331
9	0.147	0.147	0.224	0.492	0.354
11	0.130	0.130	0.181	0.451	0.369
13	0.117	0.117	0.154	0.419	0.381
15	0.107	0.107	0.136	0.393	0.390
17	0.100	0.099	0.122	0.371	0.397
19	0.093	0.093	0.112	0.352	0.404
21	0.088	0.088	0.104	0.336	0.409
23	0.083	0.083	0.097	0.322	0.413
25	0.079	0.079	0.091	0.310	0.417
27	0.076	0.076	0.087	0.299	0.420
29	0.073	0.073	0.082	0.289	0.423
31	0.070	0.070	0.079	0.280	0.426
33	0.068	0.068	0.076	0.272	0.429

从表3可以看出，不等式（3）中的上界，除了K = 1 时 $C_1 = 1$ 外，其余各值都与由等式（5）算出的上确界相差甚大；不等式（4）中的上界，只有当K值很大时，才能比较接近上确界，而 K 值越小，差距就越大，当 K = 1 和 3 时，则无法得到上界；对于等式（12），仅当K = 3 时有不大的误差，其余各个上界几乎与相应的上确界完全相等。

显而易见，用公式（12）计算c_k是十分简便的，尤其是当K非常大时更显出它的优点。值得指出的是，这里并没有利用K≥23的数据，如用这个公式计算后面的c_k，仍然得到令人满意的结果。可以推断，当 $K\to\infty$ 时，$c_k \to \alpha/\sqrt{K}$。这里 $\alpha = 1/2.898$，与正态分布的近似结果是一致的，因而K很大时这个公式也是适用的。

在许多实际分类事例中，样品数目 n 往往是很大的，一般要求选取较多的K最近邻。在使用K最近邻判决规则时，如何选取恰当的K值，使得分类过程不至于太复杂，而它的错误率尽量接近于最佳的Bayes判决规则错误率，是一个十分重要的问题。因此，能用一个简单而精确的表达式近似地算出K最近邻判决规则错误率的上确界是有现实意义的。

参 考 文 献

〔1〕 Chen C. H., A review of statistical pattern recognition, in Pattern Recognition and Signal Processing, ed. by C. H. Chen, 117～132 (Alphen aan den Rijn, Sijthoff and Noordhoff, 1978).

〔2〕 Cover T. M. and Hart P. E., Nearest Neighbor Pattern Classification, IEEE Trans. on Information Theory, IT-**13** (1967), 21～27.

〔3〕 Devijver P.A., New Error Bounds With the Nearest Neighbor Rule, IEEE Trans. on Information Theory, IT-**25** (1979) 749～753.

〔4〕 Devroye L., Some Properties of the K-nearest Neighbor Rule, Proceedings of the 5th International Joint Conference on Pattern Recognition (1980), 103～105.

〔5〕 福永圭之介著，陶笃纯译，统计图形识别导论，科学出版社（1978），197～199.

NEW UPPER BOUNDS ON THE ERROR RATE OF THE K-NEAREST NEIGHBOR DECISION RULE

Yu Xiang

(Beijing Poly—Technical Institute)

Abstract

In this paper, the following closed form expression for the upper bounds on the error rate R_K of the K-nearest neighbor decision rule for the two-class problem is obtained:

$$R_K \leqslant (1 + c_K)R^*$$

where $c_K = 1/(2.898\sqrt{K} - 1.898)$ and R^* is the error rate of the Bayes decision rule. This expression not only contains one on the error rate R_1 of the nearest neighbor decision rule $R_1 \leqslant 2R^*$, but also numerically gives a good approximation to the series form supremum for all K odd.