# 证据理论 k-NN 规则中确定相似度参数的新方法

刘 明[1,2], 袁保宗[1], 唐晓芳[1]
(1. 北京交通大学信息科学研究所, 北京 100044; 2. 河北大学电子信息工程学院, 河北保定 071002)

摘 要: 本文提出了一种确定证据理论 k-NN 分类规则中相似度参数的新方法. 对于一个模式识别问题, 我们首先为每一模式类求得一个参考最近邻距离, 使其在最小错误率意义下将训练样本集中属于该模式类的样本与其他样本分离, 然后根据所得参考最近邻距离计算相似度函数参数. 该方法在训练集比较小、样本非高斯分布条件下仍然能够计算出比较准确的参数, 使得相应的分类错误率较小, 而且时间复杂度比 L. M. Zouhal 的方法低约 4-8 倍.

关键词: 证据理论; 基本概率赋值函数; k-近邻分类
中图分类号: TN911.73 文献标识码: A 文章编号: 0372-2112 (2005) 04-0766-03

# A New Approach to Determine the Similarity Parameters in Evidence-Theoretic k-NN Rule

LIU Ming[1,2], YUAN Bao-zong[1], TANG Xiao-fang[1]
(1. *Institute of Information Science, Beijing Jiaotong Univ., Beijing* 100044, *China*;
2. *College of Electronic and Information Engineering, Hebei Univ., Baoding, Hebei* 071002, *China*)

**Abstract:** This paper presents a new approach to determine the similarity parameters in the Evidence-Theoretic k-NN Classification Rule. Given a pattern recognition problem, we first compute a reference nearest neighbor distance to separate samples of one class from other samples with least error rate, and then calculate the similarity parameters based on the obtained distance. Under the condition of small scale samples with non-gaussian distribution, the proposed method can get more suitable parameters and thus reduce classification error rate. Furthermore, its computation complexity is 4-8 times lower than that of L. M. Zouhal's method.

**Key words:** evidence theory; basic probability assignment function; k-nearest neighbor classification

## 1 引言

证据理论提供了一种很有潜力地表达、合成不确定信息的方法, 从而为信息融合提供了一种有效途径. 在模式识别领域, 关于证据理论的研究主要有两个方向: 其一是基于证据理论的分类器融合[1,2]; 其二是直接应用实际问题中的信息解决分类问题[3,4]. Denoeux 的研究属于第二个方向, 他首先在文献[3]中提出了基于证据理论的 k-NN 分类规则, 然后在文献[5]中设计了基于证据理论 k-NN 的神经网络分类器, 最近 Petit 与 Denoeux 在文献[6]中又提出了基于证据理论的非参数回归方法, 取得了很多很有价值的成果.

在 Denoeux 提出的方法中一个关键问题是如何根据样本之间距离判断它们属于同一类别的可能性即相似度, 这是模式识别研究领域中一个十分重要的问题[7,8]. Denoeu 假定相似度随距离的变化是一负指数函数, 对于不同的类别采取不同的参数, 确定不同的相似度函数. 然而关于如何确定这些参数, 他没有提出有效的方法. 在他的算法中采用了一种非常简单的处理方式, 只能适用于样本高斯分布的情况. 在文献[9]中 Zouhal 提出了基于统计学习的参数优化算法, 比较 Denoeux 的方法有很大的改进. 但是 Zouhal 的方法也存在一些问题: (1) 各个类别的多个参数要一起优化, 每个训练样本的分类结果的表达式是非线性的, 从而导致一个比较复杂的优化问题; 为了求解, 她假定训练样本的数量非常多, 从而导致她的方法不适用于小样本情况. (2) 最终结果依赖于初值, 而没有提供一个有效的求初值的方法. 基于上述原因我们重新研究了根据距离确定样本之间相似度问题, 提出了新的解决方法, 在该方法中首先采用统计方法估计每一模式类的参考最近邻距离, 然后基于该参考最近邻距离计算相似度函数参数. 我们分别在模拟数据集和实际数据集上对新方法进行了测试, 实验结果验证了这种方法在小样本、非高斯分布情况下的性能优于 Zouhal 的优化方法.

## 2 证据理论 k-NN 分类规则

对于一个模式识别问题, 假设模式类为: $T=\{\omega_1、\omega_2、\omega_3\cdots\omega_C\}$; 训练样本集为: $\{(x_1, L_1), (x_2, L_2)\cdots(x_N, L_N)\}$. 其中 $C$ 为类别数, $x_i$, $i=1, 2\cdots N$ 为训练样本, $L_i$ 是 $x_i$ 的标号.

对于一个输入待识别样本 $x_s$, 假定 $F_s$ 是它的 $k$ 近邻, $F_s$ 中任一样本 $x_i$ 的类别标号为 $L_i=\omega_q$. 那么 $(x_i, \omega_q)$ 可以看成是一个独立的支持对 $x_s$ 进行分类的证据, 它所包含的信息可以用一个基本概率赋值函数 *BPA* 来表示:

$$m^{s,i}(\{\omega_q\})=\alpha \quad (1)$$

$$m^{s,i}(T)=1-\alpha \quad (2)$$

这里 $\alpha\in[0, 1]$, 其大小取决于样本 $x_i$ 与 $x_s$ 的距离 $d$, 它随 $d$ 的增大而减小, 这一变化规律可以用一个函数来描述,

收稿日期: 2004-03-19; 修回日期: 2004-11-24
基金项目: 国家自然科学主任基金(No. 60441002); 大学重大项目基金(No. 2003SZ002)

我们称该函数为相似度函数. Denoeux 在文献[3]中假定相似度函数的形式为:

$$\alpha = \alpha_0 e^{-r_q d^{\beta}} \tag{3}$$

式中 $0 < \alpha_0 < 1$, $r_q > 0$, $\beta \in \{1, 2\}$.

将 $\Phi_s$ 中各个 $x_i$ 对应的 *BPA* $m^{s,i}$ 根据 Dempster 的合成规则合成起来, 可以得到:

$$m^s = \bigoplus_{x_i \in F_s} m^{s,i} \tag{4}$$

然后可以根据 $m^s$ 计算 $x_s$ 属于各个模式类的置信度并进行分类.

## 3　相似度函数参数的确定

估计相似度函数是证据理论 $k$ 近邻方法的一个关键问题, 关于参数 $\alpha_0$ 和参数 $r_q$ 的选取, Denoeux 在文献[3]中没有深入研究, 他只是根据经验设定 $\alpha_0 = 0.95$, 并令 $r_q$ 取训练样本中每一模式类的类内平均距离的倒数. 这种方法仅仅适用于样本高斯分布的情况.

### 3.1　Zouhal 参数优化算法

为了得到分类误差的近似表达式, Zouhal 假定: (1) $\alpha_0 = 0.95$. (2) 训练集样本数 $N$ 足够多, 而 $k$ 充分小, 从而训练集中样本与其 $k$ 近邻之间距离趋近于零.

训练集中任意样本 $x_i$ 的类别属性可以用一个矢量 $\boldsymbol{t}_i = (t_{i,1}, t_{i,2}, \cdots, t_{i,C})$ 表示[9]. 对 $x_i$ 用证据理论 $k$-NN 规则进行分类, 结果也可以用一个矢量 $\boldsymbol{P}_i = (P_{i,1}, P_{i,2}, P_{i,3}, \cdots, P_{i,C})$ 表示, $\boldsymbol{P}_i$ 的任一分量 $P_{i,j}$ 表示 $x_i$ 属于模式类 $\omega_j (j = 1, 2, \cdots, C)$ 的置信度. 然后我们可以按下面公式计算分类误差 $E$:

$$E = \sum_{i=1}^{N} \sum_{j=1}^{C} (P_{i,j} - t_{i,j})^2 \tag{5}$$

样本 $x_i$ 与其 $k$ 近邻之间的距离用一个矢量 $\boldsymbol{d}$ 表示, 把 $P_{i,j}$ 看成是关于 $\boldsymbol{d}$ 的函数. 如果 Zouhal 的假设式(2)成立, $\boldsymbol{d}$ 趋近于零向量, 那么可以将 $P_{i,j}$ 在零向量附近用泰勒级数展开:

$$P_{i,j}(d) \cong P_{ij}(0) + \nabla_d P_{i,j} \mid_{d=0} d \tag{6}$$

根据式(5)、式(6)可以得到平均分类误差 $E$ 的近似表达式, 然后采用优化算法可以求出相似度函数中的参数, 具体方法见文献[9].

如果 Zouhal 提出的假定条件得到满足, 那么她的方法可以得到局部的近似最优解, 但是在实际问题中这两个条件往往不能得到满足. 另外 Zouhal 在求解最优化问题时采用了梯度下降算法, 所以最终得到的结果依赖于初值, 而她没有给出一个有效地求初值的方法.

### 3.2　新的方法

在 Denoeux 与 Zouhal 的方法中都假定 $\alpha_0$ 为与 $q$ 无关的常数, 这一假定是没有根据的, 去掉这一假定, 用 $\alpha_q$ 代替 $\alpha_0$, 并且令 $\beta = 1$, $d_q = 1/r_q$, 我们得到:

$$\alpha = \alpha_q e^{-\frac{d}{d_q}} \tag{7}$$

将 $e^{-\frac{d}{d_q}}$ 看成是一个判断, 那么 $\alpha_q$ 可以理解为对这一判断所打的折扣, 或者说该判断的可靠性. 而 $d_q$ 可以理解为模式类 $\omega_q$ 的参考最近邻距离.

关于 $d_q$ 的信息应该不仅仅包含在 $\omega_q$ 对应的训练样本中, 不属于 $\omega_q$ 的训练样本也从另外一个方面提供了 $d_q$ 的信息. 对于每一个属于模式类 $\omega_q$ 的样本, 假设它与其在类 $\omega_q$ 内的最近邻之间距离为 $d_i$, 我们可以认为它提供了一个信息: $d_q \geqslant d_i$; 对于每一个不属于模式类 $\omega_q$ 的样本 $x_j$, 假设它与其在类 $\omega_q$ 内的最近邻之间距离为 $d_j$, 那么我们也可以认为它提供了一个信息: $d_q \leqslant d_j$. 将所有这样的信息综合起来考虑, 我们可以认为 $d_q$ 是一个阈值, 它在最小错误率意义上根据样本与其在模式类 $\omega_q$ 中最近邻的距离将训练集分成属于模式类 $\omega_q$ 的样本和不属于类 $\omega_q$ 的样本. 而对应的分类错误率 $c_q$ 反映了模式类 $\omega_q$ 与其他模式类的混淆的程度, 我们可以认为它也反映了根据 $d_q$ 计算相似度的误差, 从而有 $\alpha_q = 1 - c_q$.

关于本文介绍的有关 $d_q$ 与 $d_q$ 的确定方法, 可以通过以下的计算获得.

对于模式类 $\omega_q$ 我们可以将训练集中的样本分成两类: $T_q$ 与 $T_u$, $T_q$ 表示属于 $\omega_q$ 的样本集合, $T_u$ 表示不属于 $\omega_q$ 的样本集合.

对于 $T_q$ 中的每一个样本 $x_i$, 在 $T_q$ 中求其最近邻并计算距离 $d_i$, 然后根据 $d_i$ 定义一个函数 $Q(d)$:

$$Q(d) = \sum_{d_i \geqslant d} d_i \Big\backslash \sum_i d_i \tag{8}$$

$Q(d)$ 表示大于 $d$ 的 $d_i$ 在全部 $d_i$ 中所占的比例, 它是一个单调下降函数, 满足:

$$\begin{cases} Q(0) = 1 \\ \lim\limits_{d \to \infty} Q(d) = 0 \end{cases} \tag{9}$$

对于 $T_u$ 中的每一个样本, 在 $T_q$ 中求其最近邻并计算距离 $d'_i$, 然后我们也根据 $d'_i$ 定义一个函数 $U(d)$:

$$U(d) = \sum_{d_i \leqslant d} d_i \Big\backslash \sum_i d_i \tag{10}$$

$U(d)$ 表示小于 $d'_i$ 的 $d_i$ 在全部 $d'_i$ 中所占的比例. $U(d)$ 是一个单调上升函数, 它满足:

$$\begin{cases} U(0) = 0 \\ \lim\limits_{d \to \infty} U(d) = 1 \end{cases} \tag{11}$$

根据式(9)、式(11)可以推出, $Q(d)$ 与 $U(d)$ 对应的曲线相交于一点 $H$, 如图 1 所示. 点 $H$ 的横坐标即为所求的参考最近邻距离 $d_q$, 而其纵坐标 $Q(d_q)$ 为 $1 - \alpha_q$.

## 4　实验与分析

我们首先在模拟数据集上进行实验, 分别对 Denoeux 的证据理论 k-NN 分类规则(ET)、Zouhal 的带有参数优化的证据理论 $k$-NN 分类规则(ETO)、采用本文方法确定相似度函数的证据理论 $k$ 近邻方法(NET)、投票最近邻(KNN)、加权最近邻方法进行了测试. 对于加权最近邻方法我们又分别采用了三种相似度函数, W-1 中使用 Denoeux 定义的相似度函数, W-2 中使用根据 Zouhal 方法求出的相似度函数, W-3 中使用本文方法求出的相似度函数.

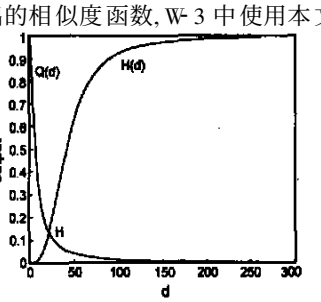

图 1 曲线 $Q(d)$ 与 $U(d)$

在实验中我们选择了 ELENA 项目中的 Clouds 数据集, 它由 5000 个二维样本组成, 我们选取 300 个样本作为训练集, 另外选取 1000 个样本作为测试集, 实验结果如表 1 所示, 可以看到证据理论方法优

于其他 $k$ 近邻方法,而本文方法优于 ETO 方法和 ET 方法.从三种加权 $k$ 近邻方法的比较可以看出根据本文方法求出的相似度函数要优于根据 Denoeux 和 Zouhal 的方法求出的相似度函数.

表 1 Clouds 数据中取 300 个样本训练时各分类器的错误率

| K | ETO | ET | KNN | NET | W- 1 | W- 2 | W- 3 |
|---|---|---|---|---|---|---|---|
| 4 | 0.1460 | 0.1290 | 0.1540 | 0.1310 | 0.1300 | 0.1480 | 0.1290 |
| 5 | 0.1320 | 0.1440 | 0.1440 | 0.1320 | 0.1440 | 0.1380 | 0.1430 |
| 6 | 0.1380 | 0.1320 | 0.1440 | 0.1280 | 0.1330 | 0.1380 | 0.1300 |
| 7 | 0.1330 | 0.1400 | 0.1400 | 0.1260 | 0.1400 | 0.1390 | 0.1370 |
| 8 | 0.1340 | 0.1370 | 0.1510 | 0.1250 | 0.1380 | 0.1350 | 0.1320 |
| 9 | 0.1310 | 0.1460 | 0.1470 | 0.1270 | 0.1470 | 0.1460 | 0.1310 |
| 10 | 0.1340 | 0.1430 | 0.1560 | 0.1240 | 0.1430 | 0.1380 | 0.1330 |

第二个实验是在真实数据集上进行的,为了验证本文方法在小样本、非高斯条件下仍然能够取得比较好的效果,我们采用了 UCI 数据库中的 wine 数据集和 vehicle 数据集.Wine 数据集包含 178 个样本、三个类别,每个样本具有 13 个属性.我们在实验中选择 100 个样本作为训练集,另外 78 个样本作为测试集,实验结果如图 2 所示.从图中可以看出本文方法的错误率是最低的,其他三种方法的错误率大体相当.Vehicle 数据

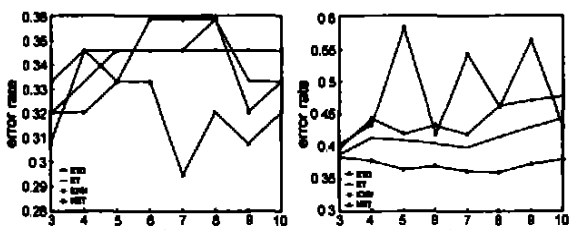


图 2 wine 数据集上的实验结果 图 3 vehicle 数据集上的实验结果

集包含 846 个样本、4 个类别,每个样本具有 18 个属性.在实验中我们选择 200 个样本作为训练集,另取 600 个数据作为测试集,实验结果如图 3 所示,可以看出本文方法的错误率仍然是最低的,而 Zouhal 的优化方法的错误率最高,对于一些 $k$ 值其错误率已经超过 50%.在真实数据集上的实验表明本文方法在小样本、非高斯分布条件下具有比较低的错误率;而 Denoeux 的方法不适用于非正态分布的情况,Zouhal 的优化方法不适用于小样本情况.

在第三个实验中我们分别在上述实验中使用的数据集上测试了本文方法与 Zouhal 的优化方法相应程序的运行速度,Zouhal的优化方法的源程序可以在 Denoeux 的个人网站(http://www.hds.utc.fr/~tdenoeux/)上下载,我们在同样的环境下测试它和我们的方法相应程序的运行速度,实验结果如表 2 所示.可以看出本文方法的时间复杂度比 Zouhal 的优化方法低约 4-8 倍.

表 2 本文方法与 Zouhal 方法相应程序运行速度的比较

| | Clouds 数据 | Clouds 数据 | wine 数据 | vehicle 数据 |
|---|---|---|---|---|
| 训练样本数 | 300 | 800 | 100 | 200 |
| Zouhal 方法 | 0.7094 秒 | 1.6880 秒 | 0.4188 秒 | 0.7437 秒 |
| 本文方法 | 0.0906 秒 | 0.3750 秒 | 0.0541 秒 | 0.0938 秒 |

## 5 结论

相似度随距离变化规律可以用负指数函数表达[3],在这样的模型下我们提出了一种求解相似度参数的方法,为每一模式类求一参考最近邻距离,使其在小错误率意义下将训练样本集中属于该模式类的样本与其他样本分离,然后根据所得参考最近邻距离计算相似度函数参数.该方法适用于训练样本较少、样本非高斯分布情况,降低了分类器的分类错误率,并且提高了相应程序的运行速度.

参考文献:

[1] G Rogova. Combining the results of several neural network classifiers [J]. Neural Networks. 1994, 7(5): 777- 781.

[2] 孙怀江,胡钟山,等. 基于证据理论的多分类器融合方法研究[J]. 计算机学报. 2001. 24(3): 231- 235.

[3] T Denoeux. A k-nearest neighbor classification rule based on Dempster-Shafer theory[J]. IEEE Trans on Systems, Man and Cybernetics. 1995, 25(05): 804- 813.

[4] 朱大奇,于盛林. 基于 D-S 证据理论的数据融合算法及其在电路故障诊断中的应用[J]. 电子学报. 2002, 30(2): 221- 223. Zhu Daqi, Yu Shenglin. Data fusion algorithm based on D-S evidential theory and its application for circuit fault diagnosis[J]. Acta Electronica Sinica, 2002, 30(2): 221- 223(in Chinese).

[5] T Denoeux. A neural network classifier based on Dempster-Shafer theory[J]. IEEE Trans on Systems, Man and Cybernetics A. 2000, 30(2): 131- 150.

[6] S Petit-Renaud, T Denoeux. Nonparametric regression analysis of uncertain and imprecise data using belief functions[J]. International Journal of Approximate Reasoning. 2004, 35(1): 1- 28.

[7] K Fukunaga, T Flick. An optimal global nearest neighbour metric[J]. IEEE Trans on Pattern Recognition and Machine Intelligence. 1984, 6(3): 314- 318.

[8] K Urahama, Y Furukawa. Gradient descent learning of nearest neighbor classifiers with outlier rejection[J]. Pattern Recognition. 1995, 28(5): 761- 768.

[9] L M Zouhal, T Denoeux. An evidence-theoretic k-NN rule with parameter optimization[J]. IEEE Trans on Systems, Man and Cybernetics C. 1998, 28(2): 263- 271.

作者简介:

刘 明 男,1972 年 4 月出生于河北安新,现为北京交通大学信息所博士研究生,主要研究方向为:信息融合,模式识别. E-mail: Liuming@mail.hbu.edu.cn

袁保宗 男,1932 年生于江苏省吴江,北京交通大学信息所教授,名誉所长,博士生导师,目前他的主要研究方向包括语音信号处理、图像处理、计算机视觉、虚拟现实以及人机交互技术等.