Artificial Intelligence and Robotics Research 人工智能与机器人研究, 2019, 8(2), 37-45
Published Online May 2019 in Hans. http://www.hanspub.org/journal/airr
https://doi.org/10.12677/airr.2019.82005

Hans汉斯

A New Target Recognition Method Based on Evidence Theoretic K-NN Rule

Xin Guan, Jing Zhao, Haiqiao Liu

Naval Aeronautical University, Yantai Shandong
Email: 597268914@qq.com, tt86725@163.com, 43458625@qq.com

Received: Feb. 6th, 2019; accepted: Mar. 1st. 2019; published: Mar. 8th, 2019

Abstract

This paper presents a new target recognition method based on evidence theoretic K-NN rule. A training correction step was added on the basis of Zouhal's improved KNN algorithm. Firstly, a reference nearest neighbor distance of every target class should be computed in order to separate samples of one class from other samples with least error rate. Secondly, the initial classification result can be got through the reference nearest neighbor distance and DS evidence theory. Thirdly, setting up confusion matrix *P* and optimizing iteration through neural network to obtain matrix parameters for Zouhal's classification result correction. Finally, the generalization ability of the matrix *P* is verified through multiple data sets, and the feasibility and effectiveness of the new method are verified by comparing with the classification accuracy of the classical algorithm.

Keywords

Evidence Theoretic K-NN Rule, Confusion Matrix, Target Recognition

基于证据K近邻的目标识别新方法

关　欣，赵　静，刘海桥

海军航空大学，山东　烟台
Email: 597268914@qq.com, tt86725@163.com, 43458625@qq.com

收稿日期：2019年2月6日；录用日期：2019年3月1日；发布日期：2019年3月8日

摘　要

本文提出了一种基于证据K近邻的目标识别新方法，在Zouhal改进KNN算法的基础上增加了训练修正步骤。首先，求得每一个目标类别的参考最近邻距离，使训练样本中该目标类别的样本在经验风险最小化的前提下与其他样本完成分离；然后，利用求得的参考最近邻距离和证据理论结合得出初始的识别分类结果；第三，设置混淆矩阵*P*，通过神经网络寻优迭代，获得*P*矩阵参数，用于Zouhal分类结果修正；最后，通过多

文章引用: 关欣, 赵静, 刘海桥. 基于证据 K 近邻的目标识别新方法[J]. 人工智能与机器人研究, 2019, 8(2): 37-45.
DOI: 10.12677/airr.2019.82005

数据集验证了*P*矩阵的泛化能力，通过与经典算法的分类精度对比验证了新方法的可行性和有效性。

关键词

证据K近邻，混淆矩阵，目标识别

1. 引言

证据理论具有较强的理论基础，既能处理随机性所导致的不确定性，又能处理模糊性所导致的不确定性[1] [2]，利用 Dempster 组合规则可以对不同传感器提供的目标识别证据进行空间域判决融合[3]，还可以在时间域对传感器提供的目标识别证据进行时间域融合[4]，证据理论可以依靠证据的积累，不断缩小假设集[5]。K 近邻(K-Nearest Neighbor, KNN)算法是一种比较成熟的分类算法，也是最简单的机器学习算法之一[6]。近年来利用 K 近邻算法进行的识别与分类被广泛应用于各行各业，例如文献[7]将其与支持向量机相融合应用于天气识别领域，文献[8]将其应用于醉酒驾驶识别领域，等等。经典的 K 近邻算法仅适用于边界可分和类分布为椭圆和高斯分布的情况[9]，一旦突破这个前提，K 近邻分类器的分类效果就会不理想。针对这个问题，很多专家学者对经典 K 近邻算法进行了改进。文献[10]中，Zouhal 利用统计学习的方法对经典 K 近邻算法进行了优化，用一个基本概率赋值(Basic Probability Assignment, BPA)函数表示一个测试样本属于某一目标类的可能性，并且将该函数随距离的变化用负指数函数表示。

本文所提出的基于证据 K 近邻的目标识别新方法就是在 Zouhal 方法基础上的改进，当利用求得的参考最近邻距离和证据理论得出初始的识别分类结果后，引入混淆矩阵 ***P***，利用初始分类结果与训练样本真值之间的偏差对结果进行修正，使得分类结果更加准确，识别率提高，然后，通过多数据集验证了 ***P*** 矩阵的泛化能力，通过与经典算法的分类精度对比验证了新方法的可行性和有效性。

2. K 近邻算法

K 近邻(K-Nearest Neighbor, KNN)算法是在 1968 年由 Cover 和 Hart 提出来的[11]，它是一种理论比较成熟的分类算法，也是最简单的机器学习算法之一。

2.1. 算法思路及特点

该算法的思路是：如果一个样本在特征空间中的 K 个最相似(也就是特征空间中的 K 个最近邻)样本中的大多数属于某一识别类，则认为该样本属于此识别类。

在 K-NN 算法中，所选择的邻居都是已经正确分类的对象，该算法在分类决策上只依据最近邻的一个或者几个样本的类别来决定待测样本的类别。

该算法的特点是：①基于实例之间距离和投票表决的分类②适合多分类③大多数情况下比贝叶斯和中心向量法好④当给定训练集、距离度量、K 值及分类决策函数时，结果唯一确定。

2.2. 算法描述

假设训练数据集 $T=\{(x_i,y_i),i=1,2,3,\cdots,N\}$，其中，$x_i\in R^n$ 为实例的特征向量，$y_i\in\{c_i,i=1,2,3,\cdots,K\}$

为实例的分类。根据给定的距离度量 d_{ij} (使用的距离度量不同，K 近邻的结果也会不同，本文取 Euclid 距离)即

$$d_{ij}=\sqrt{\sum_{l=1}^{n}\left(x_i^{(l)}-x_j^{(l)}\right)^2} \tag{1}$$

其中，$x_i,x_j\in R^n$，$x_i=\left(x_i^{(1)},x_i^{(2)},x_i^{(3)},\cdots,x_i^{(n)}\right)$，$x_j=\left(x_j^{(1)},x_j^{(2)},x_j^{(3)},\cdots,x_j^{(n)}\right)$。

在训练集 T 中找出与实例向量 $\boldsymbol{X}$ 最近的 K 个点，涵盖着 K 个点的 $\boldsymbol{X}$ 的邻域记作 $N_K\left(\boldsymbol{X}\right)$，在 $N_K\left(\boldsymbol{X}\right)$ 中根据分类决策规则(如投票表决)决定 $\boldsymbol{X}$ 所属的类别 y，即

$$y=\arg\max_{c_j}\sum_{x_i\in N_K(\boldsymbol{X})}I\left(y_i=c_j\right) \tag{2}$$

其中，$i=1,2,3,\cdots,N$，$j=1,2,3,\cdots,K$，I 为指示函数，即当 $y_i=c_j$ 时，I 为 1，否则为 0。

因此，K-NN 算法中，当训练集、距离度量、K 值及分类规则确定后，对于任意一个输入实例 $\boldsymbol{X}$，它所属的目标类 y 是唯一确定的。

3. 证据 K 近邻规则

K 近邻算法的一个显而易见的改进是对 K 个近邻的贡献进行加权[12]，根据它们相对待识别样本 $\boldsymbol{X}$ 的距离，将较大的权值赋给较近的近邻，因此，引入证据理论中，基本概率赋值函数的规律应该随距离的增大而减小，这里取负指数函数来表达[10]，即

$$\alpha=\alpha_0\mathrm{e}^{-d_k} \tag{3}$$

其中，$k=1,2,3,\cdots,K$，α_0 根据经验设定取 0.95，d_k 为第 k 个近邻到实例向量 $\boldsymbol{X}$ 的距离。则输出的识别分类结果为

$$y=\arg\max_{c_j}\sum_{x_i\in N_K(\boldsymbol{X})}\alpha I\left(y_i=c_j\right) \tag{4}$$

按照上述距离加权的 K 近邻算法是一种非常有效的归纳推理算法，它对训练中的噪声有很好的鲁棒性[13]，并且当给定足够大的训练集合时，它也是非常有效的，本论文将在后续小节进行验证。

基于距离加权的 K 近邻算法引入到证据理论中，可以将权重系数作为证据理论中相关信息的基本概率赋值，则有

$$m\left(y_i\right)=\alpha,\quad \alpha\in[0,1] \tag{5}$$

证据 K 近邻规则的思路可以概括为：利用每个近邻的基本概率赋值作为证据理论的 BPA 赋值，并完成最终分类。即假设现有一个待识别样本 x_s，它的 K 个近邻为 F_s，根据式(3)和(5)可以得到每个近邻的基本概率赋值 m_s，再利用 Dempster 组合规则进行合成，算出 x_s 属于某个目标类的置信度完成最终分类。

4. 新算法

4.1. 新算法的思想

距离 d_k 所包含的信息不仅与其所对应的训练样本中属于目标类 y 的那部分训练样本有关，还与训练样本中不属于目标类 y 的那部分训练样本有关。对于任意一个目标类 y 的样本而言，假设 d_i 是它与目标类 y 内的最近邻之间的距离，若 $d_i\geq d_k$，则认为该样本不属于目标类 y；若 $d_i\leq d_k$，则认为该样本属于目标类 y。证据 K 近邻的目标就是在误分配率最小化的前提下完成样本的分类，因此，在一定程度上，距离信息 d_k 可以看作是一个阈值，以距离信息 d_k 为门限在误分配率最小化的前提下进行样本分

类。这里就存在一个误分配率的问题，它所代表的物理意义是目标类 y 与其他目标类的混淆程度，即使是在其最小化的前提下进行的分类，它也会对最终的结果造成不良的影响，本文就是从该误分配率着手，引入混淆矩阵 $\boldsymbol{P}$，利用初始分类结果与训练样本真值之间的偏差对结果进行修正，使分类结果更加精确。

4.2. 混淆矩阵 P 的定义与求解

混淆矩阵也称作误差矩阵[14]，在人工智能中它被看作是一种可视化工具，可用于监督学习[15]，在图像精度评价中，主要用于比较分类结果和实际测得值[16]，可以把分类结果的精度显示在一个混淆矩阵里面[17]，利用这一思想，本文引入一个混淆矩阵 $\boldsymbol{P}$ 如下：

$$\boldsymbol{P}=\begin{bmatrix} P_{11} & P_{12} & \cdots & P_{1j} \\ P_{21} & P_{22} & \cdots & P_{2j} \\ \vdots & \vdots & \ddots & \vdots \\ P_{i1} & P_{i2} & \cdots & P_{ij} \end{bmatrix} \tag{6}$$

其中，P_{ij} 表示实际属于第 j 个目标类却被误分配给第 i 个目标类的概率。

首先，利用 BP 神经网络对训练样本进行训练，找到目标样本的 K 个近邻，用 $x_k\left(k=1,2,\cdots,K\right)$ 表示，通过预先设定的分类器 C 对这 K 个近邻进行分类输出，其中任意一个样本 x_k 经过分类器输出后得到的结果为 $\boldsymbol{P}_k=\left[P_k(1),P_k(2),\cdots,P_k(C)\right]^{\mathrm{T}}$，$P_k(i)=P\left(L(x_k)=y_i\right)$ 表示通过设定的分类器得到目标样本的近邻 x_k 属于类别 $y_i\left(i=1,2,\cdots,C\right)$ 的概率。假设 x_k 的真实分类为 y_t，则用 $\boldsymbol{T}_k=\left[T_k(1),T_k(2),\cdots,T_k(C)\right]^{\mathrm{T}}$ 表示期望的分类器输出值。

假设共有 C 个目标类记作 $\{y_1,y_2,\cdots y_C\}$，某待测样本 x_s 有 K 个近邻，通过证据 K 近邻规则(见第 2 节)算得待测样本 x_s 归为目标类 $y_i\left(i=1,2,\cdots,C\right)$，并且通过式(3)和(5)得出权重系数 $\boldsymbol{\alpha}$，即

$$\boldsymbol{\alpha}=\left[\alpha_1,\alpha_2,\cdots,\alpha_C\right] \tag{7}$$

然后，在 K 个近邻中找出被分给目标 $y_i\left(i=1,2,\cdots,C\right)$ 类的样本，假设有 N 个，计算这 N 个近邻被修正后的值(即 $\boldsymbol{\alpha P}$)与训练样本真值 T 之间的偏差和(这里取 Euclid 距离)，即

$$\sum_{j=1}^{N} d_j=\left\|\boldsymbol{\alpha P}-\boldsymbol{T}\right\| \tag{8}$$

将式(8)最小化算得 $\boldsymbol{P}$，即为所求混淆矩阵。

4.3. 新算法的步骤

基于上述算法思想及相关基础解析，新算法的步骤可以概括为：

①选取数据集样本进行训练，得出初始目标类和聚类中心；

②利用证据 K 近邻规则计算待测样本的初始分类结果；

③由式(8)最小化算得混淆矩阵 $\boldsymbol{P}$；

④利用混淆矩阵 $\boldsymbol{P}$ 对结果进行修正，得出最终分类结果 $m\left(y_i\right)=\boldsymbol{\alpha P}$，其中 $i=1,2,\cdots,C$。

5. 实验仿真与分析

为了验证第 3 节所说的“当给定足够大的训练集合时本文算法依然非常有效”的说法，这里选取 Satimage 数据集来进行仿真验证和深入分析。为了验证混淆矩阵 $\boldsymbol{P}$ 的泛化能力还选取了其他 6 组数据集进行验证，并将本文方法与经典 Bayes 及经典 ENN 方法进行了对比。

5.1. 仿真步骤

Satimage 数据集共有 6435 个样本，这里将样本一分为二，一半作为训练样本，一半作为测试样本。

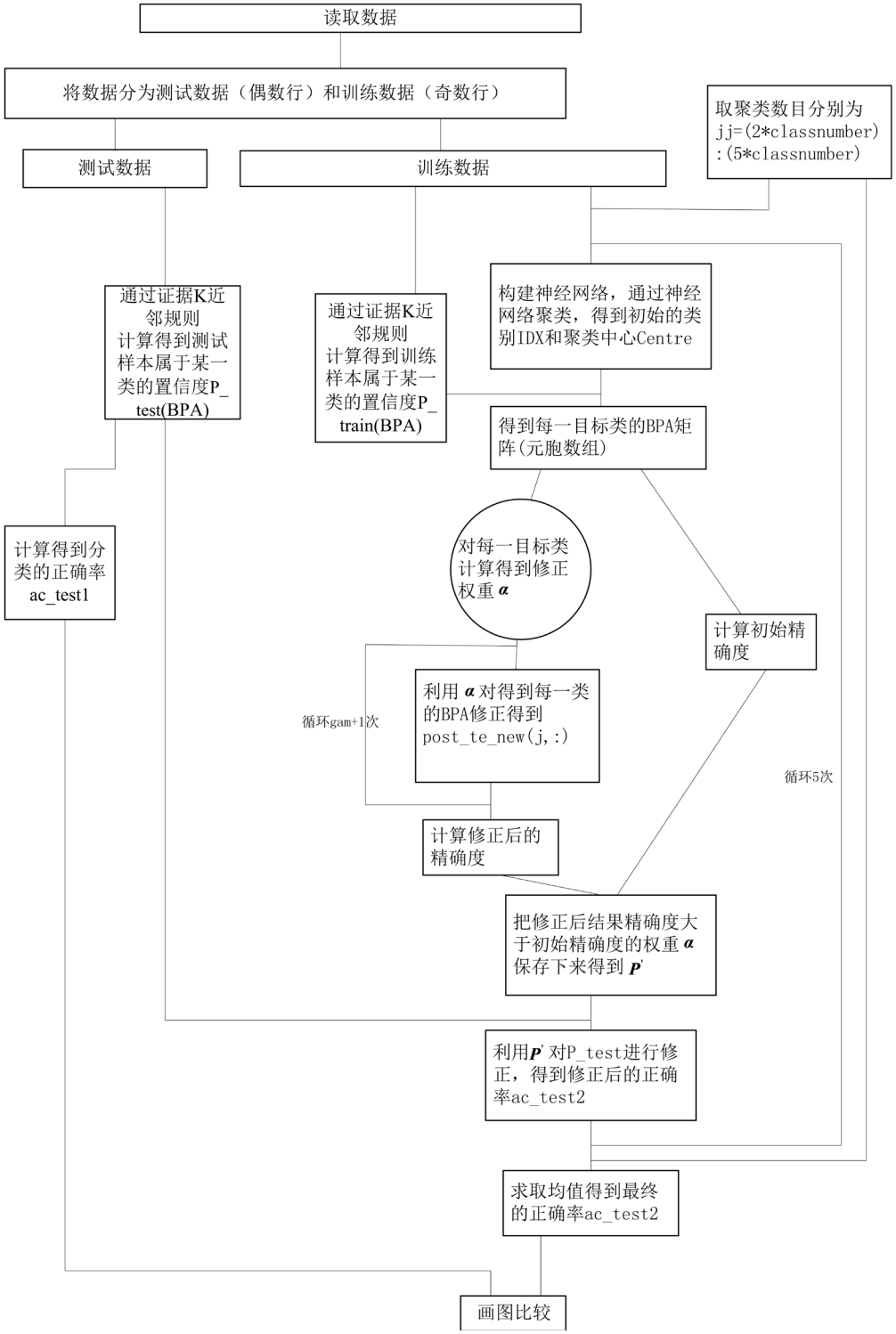

Figure 1. New algorithm MATLAB simulation flow chart
图 1. 新算法 MATLAB 仿真流程图

首先，构建神经网络对训练样本进行训练并得到初始的分类类别和聚类中心；其次，通过证据 K 近邻规则算出训练样本中每一个样本属于某一目标类别的 BPA，得到每一目标类别的 BPA 矩阵；第三，计算每一个目标类的权重系数 α，得出混淆矩阵 $\boldsymbol{P}$ (见 4.2 节)；第四，利用混淆矩阵 $\boldsymbol{P}$ 对初始结果进行修正，计算修正后的识别精确度；第五，重复迭代，将修正后的识别精确度大于初始识别精确度的权重系数 α 保存下来，构成最终的修正矩阵 $\boldsymbol{P}'$；第六，一方面，测试样本根据证据 K 近邻规则算出初始分类结果，另一方面，用训练样本算出的最终的修正矩阵 $\boldsymbol{P}'$ 对测试样本的初始分类结果进行修正，得到修正后的分类结果；最后，将测试样本初始分类识别精确度与修正后的识别精确度进行比较并作图。

Matlab 仿真流程如图 1 所示。

5.2. 实验结果分析

通过本文所提的新算法，利用 Satimage 数据集进行实验仿真后，可以得到初始分类识别精确度与修正后分类识别精确度如表 1 所示，对比图如图 2 所示。

由表 1 及图 2 可以形象直观地看出，新方法在引入混淆矩阵 $\boldsymbol{P}$ 对初始分类结果进行修正后所得到的

Table 1. Comparison of initial classification recognition accuracy and corrected classification identification accuracy
表 1. 初始分类识别精确度与修正后分类识别精确度对比表

聚类数目	修正前精确度	修正后精确度
12	52.43%	81.87%
13	52.43%	81.83%
14	52.43%	81.92%
15	52.43%	81.90%
16	52.43%	82.25%
17	52.43%	82.27%
18	52.43%	82.30%
19	52.43%	82.27%
20	52.43%	82.30%
21	52.43%	83.24%
22	52.43%	83.20%
23	52.43%	83.25%
24	52.43%	83.15%
25	52.43%	83.19%
26	52.43%	83.21%
27	52.43%	83.55%
28	52.43%	83.76%
29	52.43%	83.81%
30	52.43%	84.10%

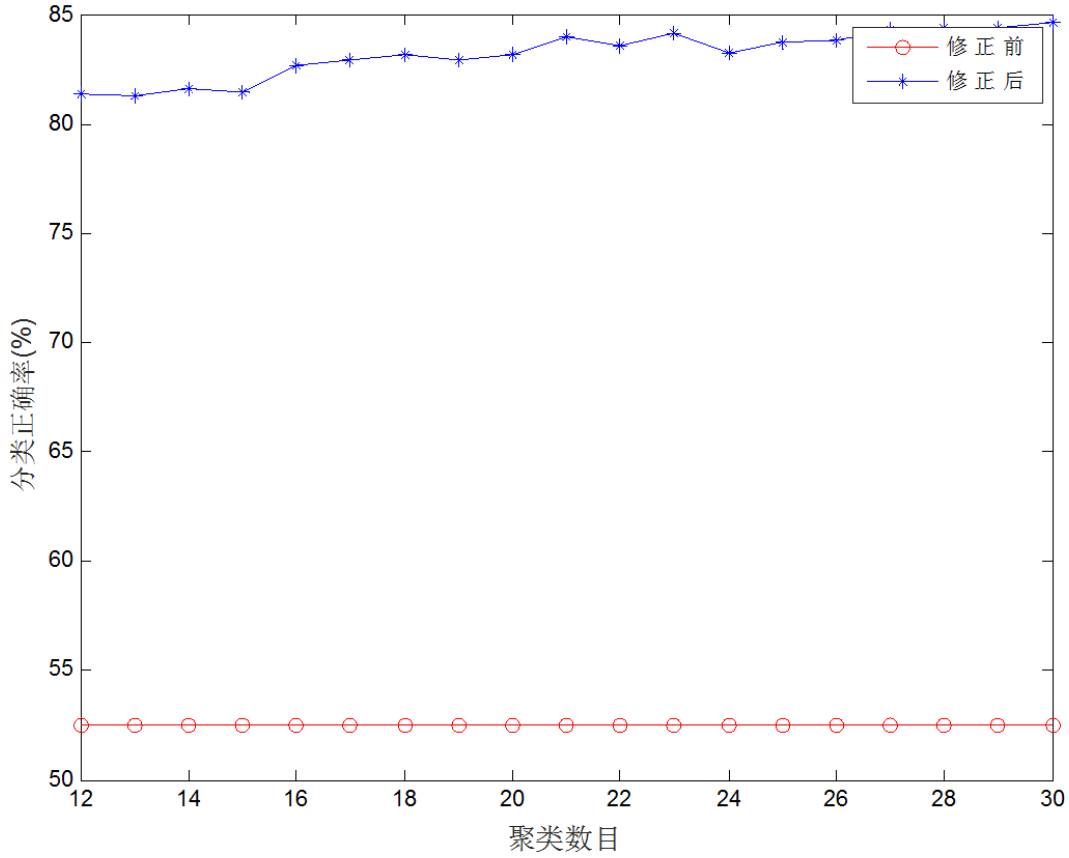

Figure 2. Correction of classification and recognition accuracy before and after correction
图 2. 修正前后分类识别精确度对比图

目标识别分类结果的精确度要远远高于仅利用证据 K 近邻规则所得的目标识别分类结果的精确度。修正前的分类精度因为没有对样本进行训练和修正的过程，所以其精度不会随着聚类数目的增多而发生变化。而本文所提的新方法利用混淆矩阵 ***P*** 对初始分类结果进行不断修正且有对样本进行训练的过程，所以该新方法在聚类数目增多的情况下，分类精度会随之增高。

为了验证混淆矩阵 ***P*** 的泛化能力，随机选取了其他 6 组数据集(Sonar、Vehicle、Page、Vowel、Wine-red 和 Segment)进行验证，并将本文方法与经典 Bayes 及经典 ENN 方法进行了对比。见表 2。

由表 2 可知，***P*** 阵具有很好的泛化能力，对训练样本依赖不大，不光是 Satimage 数据集，其他数据

Table 2. Comparison of classification accuracy of different classification ideas
表 2. 不同分类思想分类精度对比

数据集	Bayes	ENN	本文方法
Sonar	69.23%	73.08%	78.85%
Vehicle	46.93%	45.72%	61.17%
Page	90.51%	89.77%	94.76%
Vowel	71.59%	60.17%	94.44%
Wine-red	44.32%	45.59%	47.12%
Segment	80.32%	71.04%	90.76%
Satimage	80.29%	71.46%	90.78%

集也同样适用。通过实验结果对比可以看出基于证据 K 近邻的目标识别新方法较 Bayes 及经典 ENN 算法分类精度有较大提高，具有现实意义且行之有效。

6. 结论

本文提出了一种基于证据 K 近邻的目标识别新方法，在 Zouhal 改进 KNN 算法的基础上增加了训练修正步骤。首先，求得每一个目标类别的参考最近邻距离，使训练样本中该目标类别的样本在经验风险最小化的前提下与其他样本完成分离；然后，利用求得的参考最近邻距离和证据理论结合得出初始的识别分类结果；第三，设置混淆矩阵 $\boldsymbol{P}$，通过神经网络寻优迭代，获得 $\boldsymbol{P}$ 矩阵参数，用于 Zouhal 分类结果修正；最后，通过多数据集验证了 $\boldsymbol{P}$ 矩阵具有很好的泛化能力，对训练样本依赖不大，且通过实验结果对比可以看出基于证据 K 近邻的目标识别新方法较 Bayes 及经典 ENN 算法分类精度有较大提高，具有现实意义且行之有效。

基金项目

国防科技卓越青年人才基金(2017-JCJQ-ZQ-003)，泰山学者工程专项经费(ts201712072)，国家自然科学基金(61501488)资助课题。

参考文献

[1] 何友, 王国宏, 关欣, 等. 信息融合理论及应用[M]. 北京: 电子工业出版社, 2010.

[2] 郭强, 关欣, 潘丽娜, 孙贵东. 一种基于混合参数和 DSmT 的证据网络多连通结构推理方法[J]. 中国电子科学研究院学报, 2015, 10(1): 67-74.

[3] Yu, X.H., Zhou, Q.-J., Li, Y.-L., An, J. and Liu, Z.-C. (2014) A New Self-Adaptive Fusion Algorithm Based on DST and DSmT. *Proceedings of* 17*th International Conference on Information Fusion*, Salamanca, 7-10 July 2014.

[4] Tan, J.W., Zhan, H., Wen, Y. and Zhan, W.X. (2014) New Method for Multiple Cues Fusion Combined DST and DSmT. *Information Technology Journal*, **13**, 393-396. https://doi.org/10.3923/itj.2014.393.396

[5] Tchamova, A. and Dezert, J. (2012) On the Behavior of Dempster's Rule of Combination and the Foundations of Dempster-Shafer Theory. 6*th IEEE International Conference Intelligent Systems*, Sofia, 6-8 September 2012.

[6] Rahmati, Z., King, V. and Whitesides, S. (2015) Kinetic Reverse *k*-Nearest Neighbor Problem. *Lecture Notes in Computer Science*, **8986**, 307-317.

[7] 张红艳, 李茵茵, 万伟. 改进 K 近邻和支持向量机相融合的天气识别[J]. 计算机工程与应用, 2014, 50(14): 148-151.

[8] 李振龙, 韩建龙, 赵晓华, 等. 基于 K 近邻和支持向量机的醉酒驾驶识别方法的对比分析[J]. 交通运输系统工程与信息, 2015, 15(5): 246-251.

[9] Jiao, L.M. and Pan, Q. (2015) Evidential Editing *K*-Nearest Neighbor Classifier. *Lecture Notes in Computer Science*, **9161**, 461-471. https://doi.org/10.1007/978-3-319-20807-7_42

[10] Zouhal, L.M. and Denoeux, T. (1998) An Evidence-Theoretic k-NN Rule with Parameter Optimization. *IEEE Transactions on Systems*, *Man*, *and Cybernetics*, *Part C* (*Applications and Reviews*), **28**, 263-271. https://doi.org/10.1109/5326.669565

[11] Emrich, T., Kriegel, H.-P., *et al.* (2015) On Reverse-K-Nearest-Neighbor Joins. *GeoInformatica*, **19**, 299-330. https://doi.org/10.1007/s10707-014-0215-5

[12] Dubey, H. and Pudi, V. (2013) Class Based Weighted K-Nearest Neighbor over Imbalance Dataset. *Lecture Notes in Computer Science*, **7819**, 317-328.

[13] Xu, Y.T. and Wang, L.S. (2014) K-Nearest Neighbor-Based Weighted Twin Support Vector Regression. *Applied Intelligence*, **41**, 299-309. https://doi.org/10.1007/s10489-014-0518-0

[14] Lutu, P.E.N. (2011) Using Confusion Matrices and Confusion Graphs to Design Ensemble Classification Models from Large Datasets. *Lecture Notes in Computer Science*, **6862**, 301-315. https://doi.org/10.1007/978-3-642-23544-3_23

[15] Burduk, R. and Trajdos, P. (2013) Construction of Sequential Classifier Using Confusion Matrix. *Lecture Notes in Computer Science*, **8104**, 408-419. https://doi.org/10.1007/978-3-642-40925-7_37

[16] Jiang, N. and Liu, H.B. (2013) Understand System's Relative Effectiveness Using Adapted Confusion Matrix. *Lecture*

Notes in Computer Science, **8012**, 303-311. https://doi.org/10.1007/978-3-642-39229-0_32

[17] Reyes-Vargas, M., Sanchez-Gutierrez, M., Rufiner, L., *et al.* (2013) Hierarchical Clustering and Classification of Emotions in Human Speech Using Confusion Matrices. *Lecture Notes in Computer Science*, **8113**, 170-180. https://doi.org/10.1007/978-3-319-01931-4_22