Computer Science and Application 计算机科学与应用, 2025, 15(5), 735-740
Published Online May 2025 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2025.155145

Hans汉斯

基于熵值法改进的K最近邻算法

王玺帝，谢志坚，王　玲，乔腾飞，和昌伟，刘叶青

河南科技大学数学与统计学院，河南　洛阳

收稿日期：2025年4月25日；录用日期：2025年5月23日；发布日期：2025年5月30日

摘　要

针对传统K最近邻(KNN)算法在处理多维度、多量纲数据时，常因特征分布不均衡及量纲差异导致分类性能下降的问题，本文提出了一种基于熵值赋权的改进KNN算法。该方法融合了熵值赋权与标准化欧氏距离的优点，通过引入信息熵来量化各特征的信息量，并依据其重要性构建自适应权重体系，从而在距离计算中对各特征进行差异化处理，减弱了传统距离度量对高权重特征的过度敏感性。实验环节选取UCI公共数据集中的多个数据集进行测试，结果表明改进算法在大部分时候准确率优于传统KNN算法，且在高维数据集上的提升尤为显著。通过参数寻优确定最佳K值后，该算法能够显著提升分类准确率，有效解决传统KNN算法在特征分布不均和量纲差异下的性能不足问题。

关键词

KNN算法，熵值法，动态加权，二分类问题

An Improved K-Nearest Neighbor Algorithm Based on Entropy Method

Xidi Wang, Zhijian Xie, Ling Wang, Tengfei Qiao, Changwei He, Yeqing Liu

School of Mathematics and Statistics, Henan University of Science and Technology, Luoyang Henan

Received: Apr. 25th, 2025; accepted: May 23rd, 2025; published: May 30th, 2025

Abstract

To address the issue of reduced classification performance in traditional K-Nearest Neighbors (KNN) when dealing with multidimensional and multi-scale data—often caused by imbalanced feature distributions and discrepancies in scales—this paper proposes an improved KNN algorithm based on entropy weighting. This method combines the advantages of entropy weighting and normalized Euclidean distance. By introducing information entropy to quantify the information content of each feature and constructing an adaptive weighting system based on their importance, the approach

文章引用: 王玺帝, 谢志坚, 王玲, 乔腾飞, 和昌伟, 刘叶青. 基于熵值法改进的 K 最近邻算法[J]. 计算机科学与应用, 2025, 15(5): 735-740. DOI: 10.12677/csa.2025.155145

enables differentiated processing of each feature in the distance calculation, thereby reducing the over-sensitivity of traditional distance measures to high-weight features. Experiments conducted on several datasets from the UCI repository demonstrate that the improved algorithm generally achieves higher accuracy than the traditional KNN, with particularly significant improvements on high-dimensional datasets. After determining the optimal K value through parameter tuning, the algorithm substantially enhances classification accuracy, effectively overcoming the shortcomings of traditional KNN in scenarios with imbalanced feature distributions and scale differences.

Keywords

K-Nearest Neighbor Algorithm, Entropy Method, Dynamic Weighting, Binary Classification Problem

1. 引言

在机器学习领域，K 最近邻算法(K-Nearest Neighbors, KNN)是一种经典的监督学习算法，具有简单高效的特性，在文本分类[1]、异常检测[2]、疾病预测[3]等众多领域广泛应用。传统的 KNN 算法在计算样本间相似性时，多采用欧式距离或曼哈顿距离作为度量标准，然而，在处理多维度、多量纲以及特征分布不均的数据时，此类传统度量方法存在明显的局限性。在实际数据中，各特征尺度差异较大往往导致距离计算过分受高尺度特征的支配，从而忽略了尺度较低但具有较高信息价值的特征。并且，传统距离度量方式并未考量特征间诸如方差、信息熵等分布差异，这进一步削弱了算法分类的鲁棒性。因此，优化距离度量公式以提升 KNN 算法的分类性能，已成为一个重要的研究方向。

近年来，各种改进 KNN 算法的策略被相继提出，其中，部分研究聚焦于改进距离度量方法，如采用能够体现特征量之间相关关系的卡方距离作为相似性度量[4]；此外，还有研究通过结合其他算法来提升 KNN 的性能，如结合蜂群算法和模拟退火算法来优化特征选择及权重分配的基于聚类去噪及密度裁剪的改进 KNN 算法[5]，又如结合 BP 神经网络来区分复杂场景下的不同类别以提高土地覆盖分类的准确性的基于 BP 改进的 KNN 算法[6]，这些改进策略不仅丰富了 KNN 算法的应用场景，也显著提升了其在不同领域的分类性能。然而，现有方法在特征权重分配环节，大多依赖先验知识或固定规则，难以灵活适配复杂多变的数据分布。值得注意的是，熵值法在量化特征信息量方面具有显著优势，并且已证实将其应用于聚类算法的距离优化能够有效提升聚类精度[7]，这一思路也为 KNN 分类算法的改进提供了新的可能。

基于上述背景，本文提出将熵值法与标准化欧式距离相结合，构建出一种动态加权的距离度量方法，并将其应用于 KNN 算法之中。通过实验验证，该方法能够自适应地调整不同特征的权重，有效缓解传统距离度量中因量纲差异和分布不均所引发的偏差，同时也显著提升了分类准确率。

2. 相关理论基础

经典 KNN 算法(K 最近邻算法)是一种典型的监督学习算法，通过计算待分类样本与训练集中样本的距离，并依据最近邻居的类别进行多数投票，从而实现分类预测[8]。传统的 KNN 算法通常采用欧式距离或曼哈顿距离作为距离度量标准。然而，这些传统距离度量方法在面对多维度、多量纲以及特征分布不均的数据时存在明显不足，具体表现为：高量纲特征在距离计算中占据过多权重，导致低量纲但信息

量丰富的特征被忽略；同时，传统方法未考虑特征之间的分布差异(如方差、信息熵)，进而降低了算法的鲁棒性。

针对这些问题，本研究引入熵值法[7]对特征进行动态加权，并结合标准化欧式距离构建出一种改进的 KNN 分类算法。

熵值法通过量化各个特征的信息量来实现特征的动态加权。它首先对数据进行标准化处理，以消除不同属性间的量纲差异，然后通过信息熵计算各特征的信息含量，从而确定每个特征对分类任务的重要程度。

具体而言，熵值法用于度量数据特征的信息量，从而确定各特征对分类任务的相对重要性，能够实现权重的动态调整以减少其冗余性，并且可以基于数据本身的统计特性进行客观的权重分配；然而，熵值法也存在一定的局限性，它的计算复杂度较高，且对数据分布的依赖性较强，对噪声也较为敏感，因此在实际应用中需要综合考虑，以充分发挥熵值法的优势并尽量规避其局限性。

熵值法具体实施步骤如下：

(1) 标准化处理数据，消除不同属性量纲的差异：

$$r_{ij} = \frac{x_{ij} - \min\left(x_{\cdot j}\right)}{\max\left(x_{\cdot j}\right) - \min\left(x_{\cdot j}\right)}$$

其中 x_{ij} 是原始数据集中第 i 个样本的第 j 个属性值， r_{ij} 为标准化后的数据。

(2) 计算各特征的信息熵：

$$H_j = -\frac{1}{\ln n}\sum_{i=1}^{n} r_{ij} \ln r_{ij}$$

其中 H_j 表示第 j 个特征的信息熵， n 表示样本总数。

(3) 计算特征权重：

$$\omega_j = \frac{1 - H_j}{\sum_{j=1}^{m}\left(1 - H_j\right)}, 0 \le \omega_j \le 1, \sum_{j=1}^{m} \omega_j = 1$$

(4) 结合权重，采用加权标准化欧式距离进行距离计算：

$$D\left(x_i, x_j\right) = \sqrt{\sum_{k=1}^{m} \omega_k \left(r_{ik} - r_{jk}\right)^2}$$

对一个测试样本用上述公式计算它与各样本的距离，选出与它最近的 K 个近邻，并由这 K 个样本的类别投票决定测试样本的类别。

3. 实验与分析

3.1. 实验目的、数据与方法

实验主要对比经典的 KNN 算法和改进后的 KNN 算法在不同 K 值下的分类准确率。

实验数据选自 UCI 中的二分类数据集(见表 1)，并按 7:3 的比例划分为训练集和测试集。

Table 1. Experimental data
表 1. 实验数据

数据集	样本容量	特征个数	类别
Ionsphere	351	34	2
Sonar	208	60	2

续表

Wdbc	569	31	2
Tictac	958	10	2

3.2. 实验结果与分析

本节对经典 KNN 算法和基于熵值法改进的 KNN 算法在四个数据集(Ionosphere, Sonar, Wdbc, Tictac)上的实验结果进行对比分析。由于在二分类问题中使用偶数 K 值可能会出现两个类别的邻居数量相等，从而导致决策出现平局和不确定性问题，为了规避这一问题，本文均采用奇数作为 K 值，以确保在大多数情况下能够实现多数表决原则，从而提高分类的稳定性和准确性。因此，实验分别选取 K = 1、3、5、7 进行对比分析，具体结果如表 2、表 3 以及图 1 所示，表中每行黑体的数字表示该行的最高准确率。

Table 2. The accuracy rates of the classic KNN with different K values on various datasets
表 2. 经典 KNN 在各数据集上不同 K 值的准确率

数据集 \ K 值	1	3	5	7
Ionosphere	0.858	0.850	**0.868**	0.858
Sonar	**0.841**	0.794	0.778	0.794
Wdbc	0.936	0.941	0.960	**0.965**
Tictac	0.806	**0.889**	0.781	0.840

Table 3. The accuracy rates of the improved KNN with different K values on various datasets
表 3. 改进后的 KNN 在各数据集上不同 K 值的准确率

数据集 \ K 值	1	3	5	7
Ionosphere	**0.877**	0.868	0.850	0.840
Sonar	**0.857**	0.841	0.810	0.746
Wdbc	0.953	0.941	**0.965**	0.960
Tictac	0.792	0.778	0.940	**0.969**

改进的 KNN 算法在四个数据集上的表现总体优于传统 KNN，但对不同 K 值的敏感程度存在差异。对于属性差异不明显或分布较均衡的数据(如 Ionosphere、Wdbc)，K 值较小时改进的 KNN 算法更具优势，但当 K 增大后，改进效果会有所减弱甚至与传统算法相当。对于 Sonar 数据集，当 K 过大时准确率会明显下降，说明高维数据下熵值法的权重调整易受干扰。而在 Tictac 数据集中，由于属性差异较大，改进算法在较大 K 值时体现了更加显著的优势，准确率远高于传统 KNN 算法。

熵值法能够显著提高分类效果，该方法通过信息熵量化了每个特征的信息贡献，从而在距离计算中给予具有较高分类贡献的特征更大的权重，抑制了贡献小甚至具有干扰作用的特征。尤其在特征重要性差异较大的数据集中，分类准确率的提升尤为显著。

整体而言，改进的 KNN 算法在最佳 K 值条件下的分类准确率均可达到或超过传统 KNN 算法，尤其适用于属性差异较大或维度较高的数据集。改进后的算法受到特征维度与噪声干扰以及特征差异度等因素的影响。随着数据维度的提高，特征冗余和噪声影响增大，熵值法权重的性能可能会受到限制。在特征重要性差异明显时，熵值法性能提升幅度较大，在特征分布均衡时提升有限。此外 K 值也是影响算法性能的关键因素。在实际应用中，应结合数据集的特性来选择合适的 K 值，以充分发挥基于熵值法的权重调整优势。

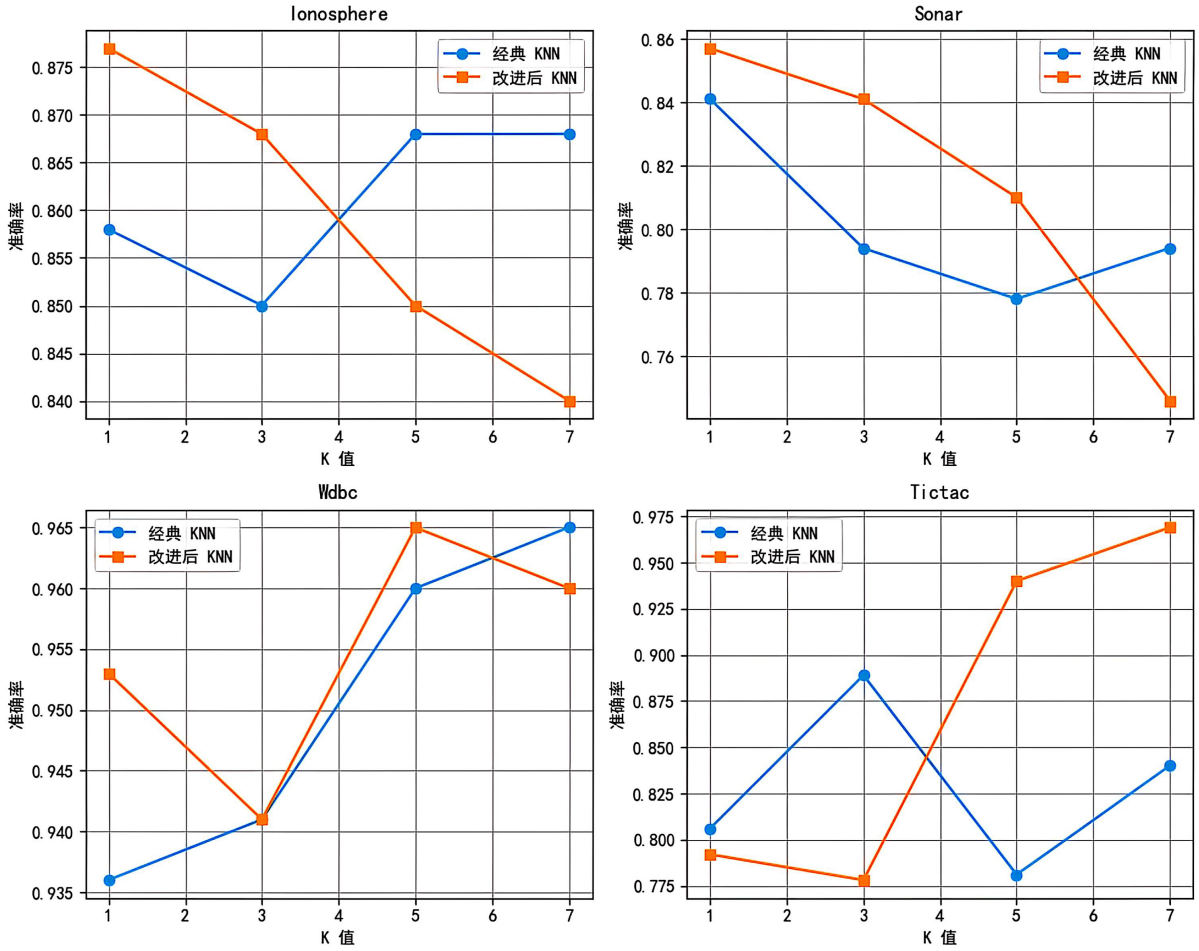

Figure 1. Comparison charts of the accuracy rates of the two algorithms on different datasets under various K values
图 1. 各数据集上两种算法在不同 K 值下准确率的对比图

4. 总结

为解决传统 KNN 在高维、多量纲数据中的性能局限性，本文引入信息熵对距离度量公式中的特征权重进行自适应调整，提出了基于熵值法改进的 KNN 算法。改进后的算法有效降低了传统 KNN 算法对高量纲特征的依赖。实验结果表明，改进算法在高维数据上的分类效果优于传统 KNN，且具有更强的通用性和适应性。

基金项目

河南科技大学大学生创新创业训练计划项目(2024239)。

参考文献

[1] 杨易木. 基于 KNN 算法的电子档案信息文本自动分类方法[J]. 办公自动化, 2025, 30(5): 14-16.

[2] 刘福民, 凌思庆, 于音, 等. 基于 KNN 算法的数控机床加工过程异常检测方法研究[J]. 机床与液压, 2024, 52(21): 168-172.

[3] 梅俊, 陈建敏. 基于 KNN 算法在糖尿病预测中的应用[J]. 电脑与信息技术, 2024, 32(1): 7-9.

[4] 谢红, 赵洪野. 基于卡方距离度量的改进 KNN 算法[J]. 应用科技, 2015, 42(1): 10-14.

[5] 戚孝铭. 基于蜂群算法和改进 KNN 的文本分类研究[D]: [硕士学位论文]. 上海: 上海交通大学, 2013.

[6] 王佃来, 宿爱霞, 刘文萍. 基于 BP 改进的 KNN 算法在北京密云土地覆盖分类中的应用[J]. 科学技术与工程, 2020, 20(23): 9464-9471.

[7] 李婧. 一种改进的最近邻聚类算法[J]. 重庆工商大学学报(自然科学版), 2013, 30(10): 61-63.

[8] 徐彦刚. 数据挖掘算法研究综述[J]. 电脑知识与技术, 2024, 20(24): 64-66+69.