Advances in Applied Mathematics 应用数学进展, 2025, 14(4), 798-809
Published Online April 2025 in Hans. https://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2025.144207

Hans汉斯

基于特征选择和加权的改进条件概率分布距离度量

杨沛融，胡桂开

东华理工大学理学院，江西 南昌

收稿日期：2025年3月26日；录用日期：2025年4月21日；发布日期：2025年4月28日

摘　要

为提高名词性属性实例间差异的识别精度，优化分类算法的准确率，在充分考虑属性间依赖关系下提出了一种基于特征选择和加权的改进条件概率分布距离度量方法。该方法首先利用对称不确定性构建了一个特征选择机制；其次，在此基础上计算属性与类的信息增益率，获得每个属性的权重，并计算加权距离；最后基于K-近邻算法对19个数据集进行仿真实验。结果表明：论文提出的距离度量有效提高了分类算法的性能。

关键词

特征选择，条件概率分布，名词性属性，信息增益率

Improved Conditional Probability Distribution Distance Measurement Based on Feature Selection and Weighting

Peirong Yang, Guikai Hu

School of Science, East China University of Technology, Nanchang Jiangxi

Received: Mar. 26th, 2025; accepted: Apr. 21st, 2025; published: Apr. 28th, 2025

Abstract

To enhance the recognition accuracy of differences between instances of nominal attributes and to optimize the accuracy of classification algorithms, an improved conditional probability distribution

文章引用: 杨沛融, 胡桂开. 基于特征选择和加权的改进条件概率分布距离度量[J]. 应用数学进展, 2025, 14(4): 798-809. DOI: 10.12677/aam.2025.144207

distance measurement based on feature selection and weighting has been proposed, taking into full consideration the dependencies among attributes. Firstly, a feature selection mechanism is constructed by using symmetric uncertainty. Secondly, on this basis, the information gain ratio of attributes and classes is calculated, and the weight of each attribute is obtained. Subsequently, the weighted distance is computed. Finally, simulation experiments are conducted on 19 datasets based on the K-Nearest Neighbors algorithm. The results indicate that the distance measurement proposed in this paper effectively improves the performance of classification algorithms.

Keywords

Feature Selection, Conditional Probability Distribution, Nominal Attribute, Information Gain Ratio

1. 引言

数据挖掘是一个从海量数据中提炼有价值信息和知识的过程。分类作为数据挖掘中的核心方法之一，其目标是构建一个能够根据已有类标签的样本，预测新样本类标签的分类器。目前，常用的分类算法涵盖了基于实例的学习、贝叶斯网络、支持向量机和神经网络等。其中，基于实例的学习(Instance-based learning, IBL) [1]是一种重要的分类学习范式，它在训练阶段保存已知类标签的样本，并在预测阶段通过距离函数来衡量训练样本与新样本的相似度，以此实现对新样本的分类预测。

K-近邻算法(K-Nearest Neighbors, KNN)是 IBL 算法中的一个典型代表。该算法通过距离函数来衡量新实例与训练实例之间的距离，并选择与新实例距离最近的 k 个训练实例。随后，算法统计这些 k 个实例的类标签频率，并将出现频率最高的类标签作为新实例的预测类别。在这个过程中，合适的距离度量对 KNN 算法的性能有显著影响。

距离度量可以根据数据属性的类型分为两大类：一类适用于数值型属性，包括欧氏距离、曼哈顿距离和切比雪夫距离等；另一类则针对名义型属性，这类属性的距离度量更为复杂。由于名义型属性值之间缺乏直接的数值关系，它们之间的关系难以用简单的数学公式直接定义或轻易计算，这使得测量这类属性间的距离成为一个挑战性问题。当全部数据均属于名义型属性时，现有的距离度量方法可被分为三类：基于属性值的距离度量、基于条件概率的距离度量以及基于条件概率分布的距离度量。

其中基于属性值的距离度量中，Aha [2]等人提出的重叠度量(Overlap Metric, OM)是最经典，使用最广泛的方法之一。它通过计算两实例间相同属性值的个数，来衡量实例间的相似度。然而，有学者认为该方法没有考虑属性值中的附加信息(例如，属性值与类之间的关系，属性值彼此之间的信息等)，因此计算粗糙。为解决这一问题，产生了许多基于属性条件概率的距离度量方法，例如，Hindi K [3]考虑测试集的属性值与训练集类标签的关系，提出了反转类指定距离度量(Inverted Specific-Class Distance Measure, ISCDM)。Short R 和 Fukunaga K [4]通过外部分类器计算属性值与类的条件概率，进而考虑最小化有限样本下误分类风险与渐进误分类风险的期望差，提出了 Short-Fukunaga 度量(Short-Fukunaga Metric, SFM)。这类距离度量较好地考虑了属性值的额外信息，但它们的建立均基于数据集中属性间的条件独立性假设，实际数据集中该假设往往不存在。Le Si Quan 和 Ho Tu Bao [5]提出的基于条件概率分布的不相似性度量(Conditional Probability Distribution-Based Dissimilarity with Kullback-Leibler, CPDKL)，较好地缓解了上述

假设。该方法通过计算 Kullback-Leibler (KL)散度来衡量两个实例在特定属性上的条件概率分布(Conditional Probability Distribution, CPD)差异，并累加所有属性上的差异，以确定两个实例间的距离。

CPDKL 主要通过分析属性间的相互关系来衡量实例间的距离。因此，它的效果更依赖于属性间的依赖关系。具体来说，若一个数据集中的属性之间的依赖性越强，那么 CPDKL 在该数据集上的距离度量表现将越出色。基于这个特性，为进一步提高 CPDKL 在分类任务中的性能表现，我们提出了一种特征选择方法。该方法首先采用对称不确定性[6]来量化属性间的依赖关系。随后，选取具有显著依赖关系的属性集计算距离。此外，为突出不同属性对分类任务的贡献，本文将计算属性与类别间的信息增益率，并据此为每个属性分配相应权重。通过此方式，论文构建了一个基于特征选择的属性加权条件概率分布的不相似性度量(Feature Selection Based Attribute-weighted CPD Dissimilarity, FSAWCPD)。本文的创新点如下：(1) 在训练算法之前，利用属性间的依赖关系构建了一种特征选择方法，该方法优化了距离度量的准确性，减少数据维度使算法运行更高效；(2) 在构建基于不同属性差异的距离度量时，通过嵌入属性权重，揭示了各个属性对分类任务差异性的具体影响。

本文结构安排如下：第 2 节给出距离度量的相关的研究知识；第 3 节阐述了本文提出的 FSAWCPD 方法；第 4 节通过 19 个 UCI 数据集对度量算法进行比较分析；第 5 节为全文总结。

2. 相关知识

距离度量是实例学习方法中的核心组成部分，它直接影响分类算法的效率(包括可扩展性、运行时间、计算成本和性能)。本节将介绍几种经典的名词性属性距离度量方法。

若数据集中所有属性都是名词性的，那么假设 x 是训练实例 $\langle x=x_1,\cdots,x_i,\cdots,x_n\rangle$，它由 m 个属性值集合 $\langle a_1(x),\cdots,a_j(x),\cdots,a_m(x)\rangle$ 以及一个类标签 c 组成 $\langle c=c_1,\cdots,c_k,\cdots,c_t\rangle$，$y$ 是测试实例 $\langle y=y_1,\cdots,y_l,\cdots,y_h,h\le n\rangle$，且每个 y 仅由 m 个属性值集合表达。另外，属性值分别归属于属性 A，$\langle A=A_1,\cdots,A_j,\cdots,A_m\rangle$。

2.1. 基于属性值的距离度量

重叠度量通过计算实例 x 与 y 在不同属性上值的差异来评估它们之间的距离或不相似度。定义如下：

$$OM(x,y)=\sum_{j=1}^{m}\left[1-\delta\left(a_j(x),a_j(y)\right)\right], \tag{2.1}$$

其中 $\delta(a,b)$ 为示性函数，当 $a=b$ 时，函数值为 1，当 $a\neq b$ 时，函数值为 0，下文出现的 $\delta(a,b)$ 函数皆由此定义。

2.2. 基于条件概率的距离度量

Short-Fukunaga 度量是一种基于最近邻的距离度量方法，最初设计用于二分类问题(例如，数据集中的类标签为 0 和 1)。随后，学者 Myles 和 Hand [7]将其扩展至多分类场景。本文将重点介绍多分类版本的 Short-Fukunaga 度量，并统一简称为 SFM (下文中的 SFM 均指多分类情况)。具体计算公式如下：

$$SFM(x,y)=\sum_{k=1}^{t}\left|P(c_k\mid x)-P(c_k\mid y)\right|, \tag{2.2}$$

式中，$P(c_k\mid x)$ 和 $P(c_k\mid y)$ 分别表示实例 x 和 y 被分类为 c_k 的条件概率，通常由朴素贝叶斯分类器计算得到。

反转类指定距离度量设计了一种反转类指定条件概率，来探讨类与属性值之间的联系。该度量认为，若一个测试实例 y 在属性 A 上的值在类标签为 $c(x)$ 的训练实例中普遍存在，并且这个值的出现频率较

高，那么 y 与训练实例 x 在该属性上的距离较短(或相似度较高)。具体定义如下：

$$ISCDM(x,y)=\sqrt{\sum_{j=1}^{m}d\left(a_j(x),a_j(y)\right)^2} \tag{2.3}$$

其中，

$$d\left(a_j(x),a_j(y)\right)=\begin{cases}1 & a_j(y)\text{未知}\\ 0 & a_j(y)=a_j(x),\\ 1-P\left(a_j(y)\mid c(x)\right) & \text{其它}\end{cases} \tag{2.4}$$

式(2.4)中，$P\left(a_j(y)\mid c(x)\right)$ 表示训练实例 x 的类标签 $c(x)$ 在第 j 个属性为 $a_j(y)$ 的实例中出现的概率，$a_j(y)$ 为测试实例 y 的第 j 个属性值。计算公式如下：

$$P\left(a_j(y)\mid c(x)\right)=\frac{\sum_{i=1}^{n}\delta\left(c(i),c(x)\right)\delta\left(a_j(i),a_j(y)\right)}{\sum_{i=1}^{n}\delta\left(c(i),c(x)\right)}, \tag{2.5}$$

$c(i)$ 为第 i 个训练实例的类标签。

2.3. 基于条件概率分布的距离度量

基于条件概率分布差异的不相似性度量，是第一个从条件概率分布角度提出的距离度量方法。具体计算步骤如下：

(1) 计算实例 x 和 y 的第 i 个属性值与其它训练实例的第 j 个属性值(不重复，且 $j\neq i$)的条件概率 $p\left(A_j=a_{qj}\mid A_i=a_i(x)\right)$ ($q=1,\cdots,n_j$， n_j 为第 j 个属性中不同值的数量)

$$p\left(A_j=a_{qj}\mid A_i=a_i(x)\right)=\frac{\sum_{l=1}^{n}\delta\left(a_{lj},a_{qj}\right)\delta\left(a_{li},a_i(x)\right)+1}{\sum_{l=1}^{n}\delta\left(a_{li},a_i(x)\right)+n_j}; \tag{2.6}$$

$$p\left(A_j=a_{qj}\mid A_i=a_i(y)\right)=\frac{\sum_{l=1}^{n}\delta\left(a_{lj},a_{qj}\right)\delta\left(a_{li},a_i(y)\right)+1}{\sum_{l=1}^{n}\delta\left(a_{li},a_i(y)\right)+n_j}; \tag{2.7}$$

这 n_j 个条件概率构成了实例 x 和 y 在属性 A_i 下与属性 A_j 的条件概率分布 $cpd\left(A_j\mid A_i=a_i(x)\right)$ 和 $p\left(c_k\mid x\right)$。

(2) KL 散度量化上述两个条件概率分布之间的差异性：

$$pd_{ji}(x,y)=KL\left(cpd\left(A_j\mid A_i=a_i(y)\right),cpd\left(A_j\mid A_i=a_i(x)\right)\right), \tag{2.8}$$

并将此过程作用于其他属性 $A_{j'}$ ($j'=1,\cdots,m-2$， $j'\neq j,i$)对应的条件概率分布，累加得到的结果即为实例 x 与 y 在属性 A_i 上的距离度量。

$$pd_i(x,y)=\sum_{j=1\wedge i}^{m}pd_{ji}(x,y) \tag{2.9}$$

KL 散度是衡量两个概率分布 P 和 Q 差异的指标。假设 P 和 Q 是两个离散随机变量的概率分布，取第 i 个值的概率分别为 P_i 和 Q_i ($i=1,\cdots,n$)，则 KL 散度的计算公式为：

$$KL(P,Q)=\sum_{i=1}^{n}P_i\log_2\frac{P_i}{Q_i} \tag{2.10}$$

(3) 将累加 x 与 y 在各个属性上的距离，从而得到它们之间的总体距离。计算公式为：

$$CPDKL(x,y)=\sum_{i=1}^{m}pd_i(x,y). \tag{2.11}$$

3. 改进的距离度量

3.1. 特征选择

针对属性间的依赖关系，我们研究了如何针对给定属性集选择一个良好的(信息丰富的)属性子集的问题。这是数据挖掘中一个典型的特征选择问题，它的目标是选择一个相关且非冗余的子集[8]。该过程在距离度量的计算之前，因此一定程度地减少了算法运行时间。具体构建思路如下：

第一步，建立两属性关系的评价指标。首先，计算属性 A_i 的熵即属性 A_i 的不确定度：

$$H(A_i)=-\sum_{l=1}^{n_i}P(a_{li})\log_2P(a_{li}), \tag{3.1}$$

其中 a_{li} 表示属性 A_i 的第 l 个属性值，n_i 表示属性 A_i 中不同属性值的个数。在观察到另一个属性 A_j 后，A_i 的熵即不确定度可以被定义为：

$$H(A_i\mid A_j)=-\sum_{l'=1}^{n_j}P(a_{l'j})\sum_{l=1}^{n_i}P(a_{li}\mid a_{l'j})\log_2P(a_{li}\mid a_{l'j}), \tag{3.2}$$

式中 $P(a_{li}\mid a_{l'j})$ 为属性值为 $a_{l'j}$ 的条件下属性值 a_{li} 出现的概率。

随后，计算由 A_i 提供的关于 A_j 的信息，这些信息由信息增益给出：

$$IG(A_j\mid A_i)=H(A_i)-H(A_i\mid A_j), \tag{3.3}$$

这里若 $IG(A_j\mid A_i)>IG(A_j\mid A_{i+1})$，则认为 A_i 比 A_{i+1} 与 A_j 的关系更接近，虽然信息增益能较好地度量两属性的依赖关系，但是它更偏向拥有属性值多的属性[9]，为了解决这一问题，并将指标结果控制在[0, 1]之间，我们计算 A_i 与 A_j 的对称不确定性：

$$SU(A_j,A_i)=2\cdot\frac{IG(A_j\mid A_i)}{H(A_j)+H(A_i)} \tag{3.4}$$

SU 值越大，说明属性 A_i 与 A_j 的依赖关系越强。

第二步，构建一个参数化的特征选择标准，它是对特定目标属性 A_j 采用基于 SU 平均值的启发式方法。这个量的平均值为：

$$mean(SU_{A_j})=\frac{\sum_{i=1\wedge j}^{m}SU(A_j,A_i)}{m-1} \tag{3.5}$$

为确定与目标属性 A_j 具有强依赖关系的属性集合，我们选择满足以下不等式的属性：

$$dependency(A_j)=\left\{A_i\in A\mid A_i\neq A_j\wedge SU(A_j,A_i)\geq\sigma\cdot mean(SU_{A_j})\right\}, \tag{3.6}$$

其中 $\sigma\in[0,1]$ 是一个控制平均值影响的权衡参数，具体由实验得到。

上述启发式方法，保证了每个属性 A_j 都至少有一个与其强依赖的属性 A_i。若对全部的属性 A_i，都有相同的 $SU(A_j,A_i)$，那么对于全部的 A_i 都有 $SU(A_j,A_i)=mean(SU_{A_j})$。此时，$A_j$ 与其它全部属性都有强

依赖关系。反之，至少存在一个属性 A_i，使得 $SU\left(A_j,A_i\right)\geq mean\left(SU_{A_j}\right)$，那么属性 A_i 被选入集合 $dependency\left(A_j\right)$ 中。

3.2. FSAWCPD

本节中，我们将基于上述特征选择设计一个新的距离度量方法，即基于属性加权的条件概率分布的不相似性度量。

为揭示属性与类别之间的关联，同时突出不同属性在分类任务中的重要性，我们采用信息增益率作为属性权值。具体计算步骤如下：

(1) 类似对称不确定性的计算过程，我们首先分别确定类别的熵以及类别在特定属性中的条件熵：

$$H\left(C\right)=-\sum_{k=1}^{t}P\left(c_k\right)\log_2 P\left(c_k\right) \tag{3.7}$$

$$H\left(C\mid A_i\right)=-\sum_{i=1}^{z}P\left(a_i\right)\sum_{k=1}^{t}P\left(c_k\mid a_i\right)\log_2 P\left(c_k\mid a_i\right) \tag{3.8}$$

其次，计算属性相对于类别的信息增益：

$$IG\left(A_i\right)=H\left(C\right)-H\left(C\mid A_i\right) \tag{3.9}$$

(2) 计算每个属性关于类别的信息增益率，以评估它们对类别预测的贡献度，

$$Gain\_ratio\left(A_i\right)=\frac{IG\left(A_i\right)}{-\sum_{i=1}^{n_i}P\left(a_i\right)\log_2 P\left(a_i\right)}. \tag{3.10}$$

将计算出的信息增益率作为权值赋给每个属性 $w_i=Gain\_ratio\left(A_i\right)$。这时，实例间的距离度量 FSAWCPD 可表示为：

$$\begin{aligned} d\left(x,y\right)&=\sum_{i=1}^{m}w_i\cdot pd_i \\ &=\sum_{i=1}^{m}w_i\sum_{j=1\wedge i}^{m}KL\left(cpd\left(A_j\mid A_i=a_i\left(y\right)\right),cpd\left(A_j\mid A_i=a_i\left(x\right)\right)\right). \end{aligned} \tag{3.11}$$

其中 pd_i 的详细计算方法参考公式(2.6)、(2.7)、(2.8)和(2.9)。

具体算法流程如下：

算法：FSAWCPD

输入：训练实例 x，测试实例 y，其中包含 m 个属性($A_1,\cdots,A_m$)；t 个类标签($C_1,\cdots,C_t$)

输出：x 与 y 的距离 $d\left(x,y\right)$

Step 1. 根据特征选择机制(3.6)，筛选与 A_i 强相关的属性集合；

Step 2. 重复步骤 1，直到所有属性筛选完毕，确保每个属性都能得到一个与之强相关的属性集合；

Step 3. 根据公式(3.10)计算每个属性的权重 $w_1,\cdots,w_m$；

Step 4. 根据公式(2.6)和(2.7)，计算 x 与 y 在属性 A_i 下，属性 A_j 中每个值出现的概率，从而得到条件概率分布；

Step 5. 根据公式(2.8)使用 KL 散度量化两分布的差异

Step 6. 重复步骤 4 和步骤 5，针对属性 A_i 的，计算 x 与 y 在除 A_i 外全部属性的条件概率分布差异，并应用公式(2.9)对这些差异求和，即为 x 与 y 在 A_i 中的距离 $pd_i\left(x,y\right)$；

Step 7. 重复步骤 4 至步骤 6，分别计算 x 与 y 在 m 个属性下的距离；

Step 8. 根据公式(3.11)对属性赋权，同时累加 x 与 y 在全部属性下的距离，并返回 $d\left(x,y\right)$。

4. 实验

4.1. 实验设计

本节将对比 FSAWCPD 与 OM、ISCDM、SFM 和 CPDKL 四种距离度量在 KNN 中的性能表现。为降低 KNN 中 k 值选择对模型性能的影响。选用 iris 数据集，针对不同距离度量，评估 KNN 分类器在 k 值从 1 至 30 的范围内的准确率。采用 10 次十折交叉验证，具体结果见图 1。结果表明：当$10 \le k \le 20$时，不同距离度量下，KNN 算法的准确率均趋于稳定。因此，我们选取了一个相对较大的值即 $k = 20$，作为后续实验的固定参数。

进一步地，依据 FSAWCPD 算法的特性，在特征选择时，需要为每个数据集设定一个合适的参数 σ 以优化特征筛选的效果。针对此，本实验采用迭代方法，将 0 到 1 之间的值以步长 0.1 作为参数输入 FSAWCPD，通过对比结果，选出能带来最佳性能的参数值作为最终参数 σ。

实验环境：本文所有实验均在 32GB RAM 和 i9-13905H Quad core CPU @2.40 GHz 的 64 位 Windows 11 系统上，基于 Anaconda3 环境并使用 Python 3.11 编程语言，通过 PyCharm2023.3.2 集成开发环境完成。

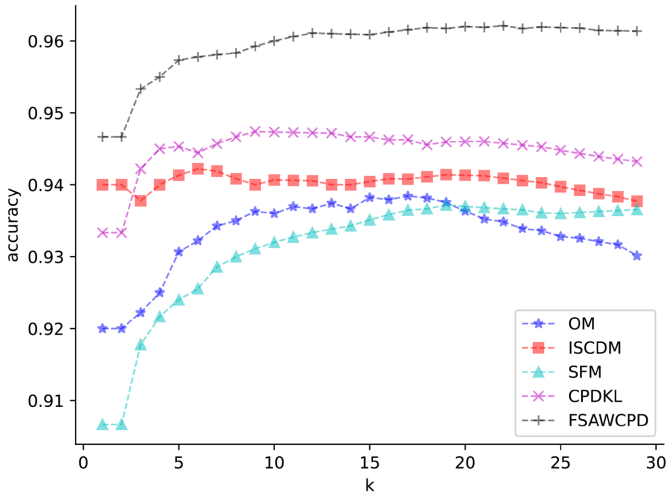

Figure 1. Accuracy of different k values at 5 distances
图 1. 五种距离下不同 k 值的准确率

4.1.1. 数据选取与预处理

为全面评估算法性能，实验从 UCI 机器学习库中挑选了 19 个不同的分类数据集参与实验。表 1 为这些数据集的详细信息，其中“Y”表示数据集中包含缺失值，“N”则表示数据集完整，不存在缺失值。

Table 1. 19 UCI datasets
表 1. 19 个 UCI 数据集

数据集	数量	名词性属性个数	数值属性个数	类标签个数	缺失值(Y/N)
automobile	206	10	15	6	Y
auto-mpg	398	0	7	3	Y
balance-scale	625	4	0	3	N
Breast-C	116	0	9	2	N

续表

breast-w	699	0	9	2	Y
car	1728	7	0	4	N
crx	690	9	6	2	Y
flags	194	30	1	8	N
glass	214	0	9	7	N
heyes-roth	133	0	4	3	N
heart-disease	303	13	0	5	Y
hepaitis	155	1	19	2	Y
iris	150	4	0	3	N
lymphography	148	19	0	4	N
primary-tumor	339	17	0	21	Y
tae	151	2	3	3	N
transfusion	748	0	4	2	N
wine	178	2	11	3	N
zoo	101	17	0	7	N

论文通过 WEKA [10]平台预处理原始数据中数值属性的连续性特征和名词性属性的离散随机性。具体步骤如下：

(1) 缺失值处理：对于名词性属性，用该属性的众数来填补缺失值，以保持数据的一致性；对于数值属性，使用属性的平均值来填补缺失值，以维持数据的连续性。

(2) 离散化处理：将数值属性均匀划分为 10 个区间，使连续的数值属性转换为具体的区间标识，实现了数据的离散化，确保了区间彼此间的独立性。

上述预处理步骤旨在提高实验结果的准确性，同时减少不同数据类型可能带来的干扰。

4.1.2. 评估指标

本节将探讨两种广泛认可的分类性能评估指标：负对数条件似然函数(Negative Conditional Log Likelihood, −CLL) [11]和根相对平方误差(Root Relative Square Error, RRSE) [12]，以评估 KNN 算法输出的类成员概率估计对 FSAWCPD 性能的影响。这两种指标的计算方法如下：

(1) 负对数条件似然函数

$$-CLL(\Theta \mid X) = -\sum_{q=1}^{N} \log \hat{P}\left(c_q \mid x_q\right), \tag{4.1}$$

其中 N 为测试集实例总个数，c_q 是第 q 个测试实例 x_q 的真实类标，$\hat{P}\left(c_q \mid x_q\right)$ 是分类器 Θ 下测试实例 x_q 的真实类标为 c_q 的概率。

(2) 根相对平方误差

$$RRSE = \sqrt{\frac{\sum_{q=1}^{N}\sum_{k=1}^{t}\left(\hat{P}_{qk} - P_{qk}\right)^2}{\sum_{q=1}^{N}\sum_{k=1}^{t}\left(P_{qk} + P_k\right)^2}} \tag{4.2}$$

公式(4.2)中，c_k 为训练集的第 k 个类标签，共包含 t 个不同的类标签；$\hat{P}_{qk}$ 为测试集 x_q 的预测类为 c_k 的概率；P_k 为类 c_k 的先验概率；P_{qk} 为测试集 x_q 的真实类为 c_k 的概率。但是实际情况中，测试集的真实类概率是未知的。所以这里假设 P_{qk} 为：

$$P_{qk}=\begin{cases}1, c_q=c_k\\ 0, c_q\neq c_k\end{cases}, k=1,\cdots,t \tag{4.3}$$

−CLL 与 RRSE 的结果越小，说明性能表现越好。

4.2. 实验结果与分析

本实验对每种度量实施 10 轮独立的十折交叉验证，并根据它们的均值和方差全面评估了准确率(ACC)、负对数条件似然函数(−CLL)以及根相对平方误差(RRSE)。此外，为深入探讨 FSAWCPD 与其他距离度量在性能上的差异，实验运用了双尾配对 t 检验，显著性水平设定为 p = 0.05。结果表格中，特别用“◦”表示 FSAWCPD 显著优于其他方法，用“•”表示 FSAWCPD 显著劣于其他方法。当 FSAWCPD 与其他度量方法无显著差异时，用加粗文字表示。

完整实验结果见表 2~4。表格的末尾两行分别标注为“平均值”和“W/T/L”。“平均值”显示了各距离度量在所有数据集上的平均检测效果。“W/T/L”展示了各度量与 FSAWCPD 在分类性能上的对比情况，即在不同数据集中“胜/平/输”的次数。以表 4 的第 2 列为例，“5/6/8”表示 CPDKL 与 FSAWCPD 的相比，5 次胜出，6 次打成平手，以及 8 次败北。

详尽实验结果如下：

(1) 由表 2 可知，FSAWCPD 在 ACC 上平与胜的总次数明显超过其它距离度量，具体来说，它超过了 CPDKL (12 次)、OM (15 次)、ISCDM (12 次)和 SFM (12 次)；在平均值方面，FSAWCPD 的平均值是 0.71，显著高于 CPDKL 的 0.69。因此，FSAWCPD 在两个维度上均优于 CPDKL，并且在与 OM、SFM 和 ISCDM 的比较中，FSAWCPD 也展现了一定优势。

(2) 在表 3 关于−CLL 的对比分析中，FSAWCPD 对比 CPDKL 和 OM 胜利和打平的次数多于失败的次数。具体而言，FSAWCPD 对比 CPDKL 和 OM 都有 15 次获胜(包含平局)和 4 次失败；从平均值上看，FSAWCPD 优于 CPDKL。

(3) 在表 4 关于 RRSE 的对比分析中，FSAWCPD 与其它距离度量相比，胜利和打平的次数超过失败的次数。具体来看，FSAWCPD 相较于 CPDKL (14 赢(平) 5 输)、OM (14 赢(平) 5 输)、ISCDM (11 赢(平) 8 输)和 SFM (11 赢(平) 8 输)表现更佳；在平均值方面，FSAWCPD 的平均值为 77.69 明显优于 CPDKL 的 78.28。总体而言，FSAWCPD 在 RRSE 指标的两个维度上都显著优于 CPDKL。同时，与其他距离度量方法相比，FSAWCPD 也展现了一定的优势。

综合以上 3 点，论文提出的方法在指标 ACC、−CLL 和 RRSE 上均优于 CPDKL 和 OM。此外，在 ACC 和 RRSE 方面，该方法与 SFM 和 ISCDM 的表现相当，并略具优势。

Table 2. Comparison of ACC between FSAWCPD and CPDKL, OM, ISCDM and SFM
表 2. FSAWCPD 与 CPDKL、OM、ISCDM 和 SFM 的 ACC 比较

数据集	FSAWCPD	CPDKL	OM	SFM	ISCDM
automobile	0.54 ± 0.11	0.52 ± 0.11◦	0.59 ± 0.10•	0.64 ± 0.09•	0.63 ± 0.10•
auto-mpg	0.71 ± 0.06	0.73 ± 0.06•	0.70 ± 0.06◦	**0.72 ± 0.06**	0.69 ± 0.07◦
balance-scale	0.45 ± 0.05	**0.45 ± 0.05**	0.82 ± 0.05•	0.95 ± 0.03•	0.90 ± 0.03•

续表

Breast-C	0.66 ± 0.13	**0.65 ± 0.13**	0.60 ± 0.15◦	**0.67 ± 0.13**	0.63 ± 0.13◦
breast-w	0.97 ± 0.02	0.96 ± 0.02◦	0.94 ± 0.03◦	0.97 ± 0.01•	**0.97 ± 0.02**
car	0.70 ± 0.03	**0.70 ± 0.03**	0.89 ± 0.03•	0.96 ± 0.02•	0.88 ± 0.02•
crx	0.85 ± 0.03	0.84 ± 0.04◦	**0.85 ± 0.03**	0.84 ± 0.04◦	0.86 ± 0.04•
flags	0.65 ± 0.09	0.57 ± 0.10◦	0.56 ± 0.10◦	0.63 ± 0.09◦	**0.65 ± 0.09**
glass	0.62 ± 0.11	0.58 ± 0.11◦	0.55 ± 0.10◦	**0.61 ± 0.11**	**0.61 ± 0.10**
heart-roth	0.59 ± 0.16	0.53 ± 0.13◦	0.55 ± 0.14◦	0.71 ± 0.13•	**0.58 ± 0.14**
heart-disease	0.59 ± 0.08	**0.59 ± 0.09**	**0.58 ± 0.07**	**0.58 ± 0.09**	0.57 ± 0.08◦
hepaitis	0.82 ± 0.09	**0.83 ± 0.09**	**0.81 ± 0.10**	0.85 ± 0.08•	**0.84 ± 0.09**
iris	0.97 ± 0.05	0.94 ± 0.06◦	0.92 ± 0.07◦	0.94 ± 0.06◦	0.94 ± 0.07◦
lymphography	0.81 ± 0.10	**0.82 ± 0.09**	**0.80 ± 0.09**	0.78 ± 0.10◦	0.83 ± 0.09•
primary-tumor	0.44 ± 0.09	**0.45 ± 0.09**	**0.43 ± 0.09**	0.41• ± 0.08◦	**0.46 ± 0.08**
tae	0.49 ± 0.13	**0.50 ± 0.13**	**0.50 ± 0.12**	0.53 ± 0.11•	0.52 ± 0.12•
transfusion	0.78 ± 0.05	0.76 ± 0.05◦	0.77 ± 0.04◦	0.77 ± 0.04◦	0.77 ± 0.05◦
wine	0.96 ± 0.04	0.95 ± 0.04◦	0.94 ± 0.05◦	**0.96 ± 0.04**	**0.95 ± 0.04**
zoo	0.81 ± 0.12	0.82 ± 0.12•	0.84 ± 0.11•	**0.82 ± 0.11**	**0.81 ± 0.11**
平均值	0.71	0.69	0.71	0.75	0.74
W/T/L	—	2/8/9	4/6/9	7/7/6	6/8/5

Table 3. Comparison of −CLL between FSAWCPD and CPDKL, OM, ISCDM and SFM
表 3. FSAWCPD 与 CPDKL、OM、ISCDM 和 SFM 的−CLL 比较

数据集	FSAWCPD	CPDKL	OM	SFM	ISCDM
automobile	24.59 ± 2.55	25.38 ± 2.60◦	23.66 ± 2.57•	19.92 ± 3.11•	20.33 ± 3.48•
auto-mpg	25.13 ± 3.05	24.66 ± 2.88•	25.57 ± 2.97◦	26.02 ± 3.19◦	26.93 ± 3.89◦
balance-scale	61.05 ± 3.82	**61.05 ± 3.82**	42.56 ± 5.16•	13.20 ± 2.93•	35.16 ± 5.35•
Breast-C	7.17 ± 0.92	**7.19 ± 0.77**	7.76 ± 0.77◦	7.85 ± 2.25◦	7.79 ± 1.47◦
breast-w	9.56 ± 2.49	9.31 ± 2.52•	11.14 ± 2.83◦	8.26 ± 2.47•	8.95 ± 2.04•
car	164.77 ± 10.97	**164.77 ± 10.97**	64.88 ± 0.03•	43.03 ± 70•	63.06 ± 4.24•
crx	24.91 ± 4.52	27.63 ± 3.28◦	27.04 ± 3.52◦	25.80 ± 4.53◦	**25.26 ± 4.21**
flags	22.40 ± 2.99	25.85 ± 3.02◦	25.92 ± 3.24◦	21.61 ± 3.87•	20.88 ± 3.66•
glass	21.44 ± 2.80	22.17 ± 2.73◦	23.96 ± 3.12◦	21.03 ± 3.24•	**21.62 ± 3.73**
heart-roth	11.65 ± 1.43	11.90 ± 0.96◦	11.97 ± 1.17◦	7.99 ± 2.58•	9.90 ± 1.35•
heart-disease	30.97 ± 3.78	**30.99 ± 3.80**	**31.19 ± 3.93**	31.45 ± 4.43◦	31.68 ± 4.77◦
hepaitis	5.94 ± 2.17	5.62 ± 1.97•	**5.94 ± 2.15**	5.17 ± 2.13•	5.35 ± 2.29•
iris	3.34 ± 1.24	3.60 ± 1.02◦	5.12 ± 1.27◦	3.62 ± 1.73◦	3.55 ± 1.61◦
lymphography	6.97 ± 2.36	7.22 ± 2.26•	7.63 ± 2.06◦	7.82 ± 2.17◦	6.65 ± 2.53•

续表

primary-tumor	63.67 ± 5.86	64.92 ± 5.71◦	66.38 ± 5.58◦	65.67 ± 5.72◦	61.85 ± 6.22•
tae	16.22 ± 1.35	15.96 ± 1.35•	15.81 ± 1.21•	14.80 ± 2.81•	**16.47 ± 2.56**
transfusion	37.79 ± 4.50	38.91 ± 5.01◦	38.61 ± 5.01◦	38.39 ± 5.01◦	39.16 ± 5.58◦
wine	3.93 ± 1.13	4.22 ± 1.12◦	6.55 ± 0.96◦	3.20 ± 0.04•	3.48 ± 1.21•
zoo	5.81 ± 1.78	5.90 ± 1.80◦	6.11 ± 1.82◦	6.16 ± 1.89◦	5.98 ± 1.83◦
平均值	28.80	29.27	23.57	19.53	21.79
W/T/L	—	4/4/11	4/2/13	10/0/9	10/3/6

Table 4. Comparison of RRSE between FSAWCPD and CPDKL, OM, ISCDM and SFM
表 4. FSAWCPD 与 CPDKL、OM、ISCDM 和 SFM 的 RRSE 比较

数据集	FSAWCPD	CPDKL	OM	SFM	ISCDM
automobile	88.88 ± 5.92	90.17 ± 5.87◦	86.68 ± 5.68•	79.44 ± 7.72•	81.79 ± 9.04•
auto-mpg	81.26 ± 5.29	80.48 ± 5.69•	**81.64 ± 5.02**	82.71 ± 6.54◦	85.05 ± 7.82◦
balance-scale	103.34 ± 3.67	**103.33 ± 3.67**	76.29 ± 3.92•	36.39 ± 7.17•	64.34 ± 4.86•
Breast-C	91.83 ± 8.78	**92.05 ± 7.91**	97.24 ± 8.03◦	95.40 ± 14.99◦	97.09 ± 11.13◦
breast-w	34.99 ± 8.63	34.22 ± 9.00•	41.55 ± 8.77◦	30.72 ± 9.70•	**34.22 ± 10.07**
car	105.04 ± 3.85	**105.04 ± 3.85**	64.88 ± 1.72•	40.84 ± 5.75•	56.88 ± 3.24•
crx	65.45 ± 7.79	69.61 ± 6.19◦	67.99 ± 6.33◦	66.96 ± 8.06◦	**66.17 ± 7.60**
flags	79.24 ± 5.34	86.06 ± 3.72◦	86.09 ± 3.66◦	**78.78 ± 6.06**	77.31 ± 6.81•
glass	83.35 ± 5.73	84.40 ± 5.32◦	87.54 ± 4.71◦	**82.82 ± 6.32**	**84.02 ± 7.18**
heart-roth	89.49 ± 6.23	92.27 ± 3.63◦	92.15 ± 3.77◦	74.27 ± 15.41•	82.84 ± 7.45•
heart-disease	88.26 ± 4.21	**88.02 ± 3.83**	**88.13 ± 3.89**	89.45 ± 4.97◦	90.26 ± 5.50◦
hepaitis	85.29 ± 15.73	83.25 ± 12.91•	**83.60 ± 12.83**	76.53 ± 18.77•	79.77 ± 21.01•
iris	34.01 ± 12.04	35.30 ± 10.84◦	46.34 ± 10.31◦	34.59 ± 15.36◦	35.73 ± 15.09◦
lymphography	69.93 ± 12.14	**70.39 ± 11.73**	72.82 ± 10.21◦	74.77 ± 10.62◦	67.20 ± 14.67•
primary-tumor	89.13 ± 4.59	**89.39 ± 3.89**	90.75 ± 3.87◦	**91.08 ± 3.39**	**89.21 ± 4.41**
tae	97.65 ± 4.91	96.77 ± 4.92•	96.67 ± 4.59•	94.44 ± 10.96•	100.03 ± 9.61◦
transfusion	94.99 ± 4.78	96.90 ± 4.85◦	96.16 ± 5.00◦	95.91 ± 5.42◦	97.11 ± 5.40◦
wine	32.82 ± 11.21	34.57 ± 10.42◦	45.78 ± 7.16◦	26.68 ± 13.29•	29.97 ± 12.83•
zoo	55.45 ± 9.80	55.07 ± 9.53•	54.65 ± 8.92•	56.29 ± 9.94◦	**55.50 ± 9.79**
平均值	77.39	78.28	76.68	68.85	72.34
W/T/L	—	5/6/8	5/3/11	8/3/8	8/5/6

5. 结语

鉴于属性实例间存在的依赖关系，为更精准地度量实例间的距离，本文设计了一种特征选择机制，并据此提出了属性加权条件概率分布的不相似性度量(FSAWCPD)。该方法不仅考虑了属性间的关系，还通过属性加权的方式，强调了各属性对分类任务的重要性。实验结果表明，在 KNN 算法的应用场景下，

FSAWCPD 的性能优于 CPDKL、OM、ISCDM 和 SFM 等传统距离度量。然而，在与 ISCDM 和 SFM 的对比中，FSAWCPD 在负对数似然损失(−CLL)指标上未表现出显著优势。鉴于此，我们计划在未来进一步优化 FSAWCPD，以期在−CLL 指标上实现更出色的性能提升。

基金项目

国家自然科学基金(11661003)，江西省自然科学基金(20192BAB201006)。

参考文献

[1] Ayats, H.A., Cellier, P. and Ferré, S. (2024) Concepts of Neighbors and Their Application to Instance-Based Learning on Relational Data. *International Journal of Approximate Reasoning*, **164**, Article ID: 109059. https://doi.org/10.1016/j.ijar.2023.109059

[2] Aha, D.W., Kibler, D. and Albert, M.K. (1991) Instance-Based Learning Algorithms. *Machine Learning*, **6**, 37-66. https://doi.org/10.1007/bf00153759

[3] El Hindi, K. (2013) Specific-Class Distance Measures for Nominal Attributes. *AI Communications*, **26**, 261-279. https://doi.org/10.3233/aic-130565

[4] Short, R. and Fukunaga, K. (1981) The Optimal Distance Measure for Nearest Neighbor Classification. *IEEE Transactions on Information Theory*, **27**, 622-627. https://doi.org/10.1109/tit.1981.1056403

[5] Quang, L.S. and Bao, H.T. (2004) A Conditional Probability Distribution-Based Dissimilarity Measure for Categorial Data. In: Dai, H., Srikant, R. and Zhang, C., Eds., *Advances in Knowledge Discovery and Data Mining*, Springer, 580-589. https://doi.org/10.1007/978-3-540-24775-3_69

[6] Ienco, D., Pensa, R.G. and Meo, R. (2012) From Context to Distance: Learning Dissimilarity for Categorical Data Clustering. *ACM Transactions on Knowledge Discovery from Data*, **6**, 1-25. https://doi.org/10.1145/2133360.2133361

[7] Myles, J.P. and Hand, D.J. (1990) The Multi-Class Metric Problem in Nearest Neighbour Discrimination Rules. *Pattern Recognition*, **23**, 1291-1297. https://doi.org/10.1016/0031-3203(90)90123-3

[8] Guyon, I. and Elisseeff, A. (2003) An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research*, **3**, 1157-1182.

[9] 龚芳. 反转类指定距离度量的改进及应用研究[D]: [博士学位论文]. 武汉: 中国地质大学, 2021.

[10] 李超群. 名词性属性距离度量问题及其应用研究[D]: [博士学位论文]. 武汉: 中国地质大学, 2012.

[11] Qiu, C., Jiang, L. and Li, C. (2015) Not Always Simple Classification: Learning SuperParent for Class Probability Estimation. *Expert Systems with Applications*, **42**, 5433-5440. https://doi.org/10.1016/j.eswa.2015.02.049

[12] Gong, F., Jiang, L., Zhang, H., Wang, D. and Guo, X. (2020) Gain Ratio Weighted Inverted Specific-Class Distance Measure for Nominal Attributes. *International Journal of Machine Learning and Cybernetics*, **11**, 2237-2246. https://doi.org/10.1007/s13042-020-01112-8