小 型 微 型 计 算 机 系 统
Journal of Chinese Computer Systems
DOI:10.20009/j.cnki.21-1106/TP.2025-0226
2026 年 5 月 第 5 期
Vol.47 No.5 2026

多视图生成与证据理论融合的 k 近邻分类算法

崔金浩[1,2],龚 芳[1,2,3,4],张志强[1,2],赵楠楠[1,2],吕昊东[1,2],梁 超[3,4]

[1](武汉工程大学 计算机科学与工程学院人工智能学院,武汉 430205)
[2](武汉工程大学 智能机器人湖北省重点实验室,武汉 430073)
[3](武汉大学 多媒体网络通信工程湖北省重点实验室,武汉 430072)
[4](武汉大学 国家多媒体软件工程技术研究中心,武汉 430072)
E-mail:fang.gong@wit.edu.cn

摘 要: 在使用 k 近邻算法进行分类任务时,原始数据特征描述不充分和使用多数投票法进行分类决策会严重限制算法的分类效果.为此,本文提出了一种新的多视图生成和证据理论融合的 k 近邻算法,通过更全面地描述数据特征以及更准确的进行分类决策,从而提升 k 近邻算法的分类性能.该算法首先使用超父亲一依赖决策器和随机森林算法对原始属性视图进行分类并生成两个新的标签视图,然后在原始属性视图和两个生成的标签视图上分别构建距离加权的 k 近邻算法,最后通过 D-S 证据理论融合来自不同视图 k 近邻算法的预测结果,从而得到最终的分类结果.实验结果表明,本文提出的算法在分类准确率和根相对平方误差两个指标上均优于传统 k 近邻算法及其他对比算法.

关 键 词: 分类;k 近邻;多视图生成;D-S 证据理论
中图分类号: TP181 **文献标识码:** A **文 章 编 号:** 1000-1220(2026)05-1134-13

K-nearest Neighbor Classification Algorithm Integrating Multi-view Generation and D-S Theory

CUI Jinhao[1,2],GONG Fang[1,2,3,4],ZHANG Zhiqiang[1,2],ZHAO Nannan[1,2],LÜ Haodong[1,2],LIANG Chao[3,4]
[1](School of Computer Science & Engineering Artificial Intelligence,Wuhan Institute of Technology,Wuhan 430205,China)
[2](Hubei Key Laboratory of Intelligent Robot,Wuhan Institute of Technology,Wuhan 430073,China)
[3](Hubei Provincial Key Laboratory of Multimedia Network Communication Engineering,Wuhan University,Wuhan 430072,China)
[4](National Engineering Research Centre for Multimedia Software(NERCMS),Wuhan University,Wuhan 430072,China)

Abstract: In classification tasks utilizing k-nearest neighbors(kNN) algorithm,insufficient feature representation of original data and reliance on majority voting for decision-making can significantly hinder the algorithm's classification performance. To address this issue,this paper proposes an Enhanced k-Nearest Neighbor algorithm(EnDWkNN) based on multi-view generation and evidence theory. This approach aims to enhance the classification performance of kNN by providing a more comprehensive description of data features and making more accurate classification decisions. It first employs multiple super-parent class-dependent estimators along with random forest algorithms to classify the original attribute view,generating two new label views. Then,distance-weighted k-NN algorithms are constructed separately on original attribute view and two generated label views. Finally,Dempster-Shafer(D-S) theory is applied to fuse the predictions from different view-based k-NN algorithms,resulting in an aggregated final classification outcome. Experimental results demonstrate that EnDWkNN outperforms traditional kNN as well as other competitors in terms of both classification accuracy and root relative squared error metrics.

Keywords: classification;k-nearest neighbor algorithm;multi-view generation;D-S theory

0 引 言

k 近邻算法(k-Nearest Neighbor algorithm,kNN)是机器学习中非常经典的非参数学习算法,它通常用于分类任务,在许多应用领域被广泛使用,例如图像识别[1]、文本分类[2]、推荐系统[3,4]、异常检测[5]等.kNN 属于懒惰学习算法的算法,它在训练阶段只存储数据而不做任何计算.当需要预测测试实例类标记时,它才根据存储的训练实例构建关于测试实例的局部分类模型.相比于决策树、朴素贝叶斯分类器等其他主动学习算法在训练阶段构建一个全局分类模型,kNN 往往能够获得更加准确的预测结果.已有研究证明,在满足$(k/N)\to 0$(其中 N 为训练实例总数)且 $N\to\infty$ 的条件下,kNN 算法的分类准确率可以接近最优贝叶斯分类器的两倍[6].

在对测试实例进行预测过程中,kNN 首先利用距离度量函数从训练实例中找到与测试实例最近的 k 个邻居.然后利用众数投票法从这 k 个邻居的类标记中推测出测试实例可能

收稿日期:2025-06-04 收修改稿日期:2025-07-17 基金项目:国家自然科学基金青年项目(62406294)资助;多媒体网络通信工程湖北省重点实验室开放基金项目(2025KFKT17)资助;武汉工程大学第十六届研究生教育创新基金项目(CX2024161)资助. 作者简介:**崔金浩**,男,2001 年生,硕士研究生,CCF 会员,研究方向为距离度量学习、多视图学习;**龚 芳**,女,1991 年生,博士,讲师,CCF 会员,研究方向为机器学习与数据挖掘;**张志强**,男,2001 年生,硕士研究生,研究方向为多视图学习、半监督学习;**赵楠楠**,女,2001 年生,硕士研究生,研究方向为距离度量学习、半监督学习;**吕昊东**,男,2001 年生,硕士研究生,研究方向为距离度量学习、半监督学习;**梁 超**,男,1984 年生,博士,教授,CCF 会员,研究方向为模式识别、多媒体分析与检索.

的类标记. kNN 的具体分类公式如下:

$$c(x)=\underset{c\in C}{\operatorname{argmax}}\sum_{q=1}^{k}\delta(c,c(x_q)) \tag{1}$$

其中,x 表示测试实例,C 表示数据集中类标记的总个数,x_q 表示测试实例的第 q 个近邻. $\delta(\cdot,\cdot)$是一个指示函数,当 $c=c(x_q)$时,$\delta(\cdot,\cdot)=1$,否则 $\delta(\cdot,\cdot)=0$.

传统 kNN 算法虽然具备简单、有效、易理解等优点[7,8],但是还是存在一些问题. 例如,对 k 值选择敏感,如果 k 值过小,算法会对噪声敏感且容易增加过拟合风险. 如果 k 值过大,算法会考虑到更多不同类别的邻居,从而影响分类准确性. 距离度量不准确,kNN 算法依赖距离度量函数来衡量实例间的相似程度. 但距离度量易受属性特征不同量纲和取值范围的影响,且存在属性独立性假设,容易造成 kNN 在具有复杂属性依赖关系和特征尺度的数据集上分类效果较差. 分类决策规则不可靠,现有算法多采用多数投票法来推断待测实例的类标记,这种方法虽然简单有效,但当数据集中不同类别实例数量存在较大差异时,分类结果容易向多数类实例倾斜,导致分类结果偏离真实情况.

尽管学者们分别从 k 值自适应、距离度量优化、分类决策规则升级方面对传统 kNN 算法进行了改进,但现有的改进算法都只考虑了专家定义的原始数据特征. 由于人类认知能力的限制,即使最权威的领域专家,也很难从复杂的实际场景中提取出足够多的判别特征. 而专家定义的原始数据特征通常是易于命名、观察和度量的,例如性别、年龄、职业等. 然而,有助于分类任务的特征有时很难被明确的命名和观察,而原始数据特征的不完全性容易导致模型的分类性能遇到瓶颈. 为了克服传统 kNN 以及相关改进算法只依赖单一的专家定义的数据特征的局限性,本文提出了一种新的 kNN 分类算法模型,称为增强 k 近邻算法(Enhanced Distance Weighted k-Nearest Neighbor Algorithm,EnDWkNN). 该算法通过引入多视图生成实现从原始数据特征中学习和发现能够为分类提供更多判别信息的高级的隐含特征的目的. 具体地,EnDWkNN 借助原始数据特征和新生成的隐含特征构建不同的数据视图,然后在不同视图下构建距离加权的 k 近邻分类模型,最后利用 D-S 证据理论[9,10]来融合来自不同视图下分类模型的预测结果,从而实现更加准确、可靠的分类预测. 本文的贡献如下:

1)通过引入多视图生成来获取更全面的数据特征,为后续 k 近邻分类算法提供了更加丰富的判别信息,突破了传统 kNN 以及相关改进算法只依赖单一的专家定义的数据在不引起较高计算复杂度的同时改善了 kNN 算法的分类性能.

2)使用证据理论融合来自不同视图的 k 近邻算法分类结果. 相比于简单的加权求和融合机制,证据理论通过综合多个视图下测试实例的类概率估计,能够显著提升 kNN 算法的分类性能.

1　相关工作

为了提升传统 kNN 算法的分类性能,学者们从不同方向对其进行了改进. 这一节对这些改进方法进行了全面的介绍.

1.1　分类决策规则升级

分类决策规则升级旨在改进测试实例类标记推理策略,通过更准确地描述测试实例的类别归属,以获得相比传统 kNN 算法具有更好的性能. 传统 kNN 算法采用多数投票法从 k 个近邻的类标记中推理得出测试实例可能的类标记. 这种方法虽然简单,但准确性较低,尤其是在不同类别数量差异较大数据集中,准确性下降明显. 为此,Dudani 等人[11]提出了一种基于距离加权的分类决策规则升级方法,该方法通过向传统多数投票法中引入距离相关权重因子,使模型在统计近邻类别数量的同时,综合考虑各近邻实例与测试实例之间的距离,从而有效提升了 kNN 算法的分类性能. 在此基础上,Gou 等人[12,13]进一步对多种距离加权策略进行了系统研究,验证了不同加权方法对基于距离加权的分类决策规则升级效果的影响. 除此之外,Chen 和 Zhang 等人[14,15]引入模糊理论的思想,提出了一系列基于模糊集合的分类决策规则升级方法. 这些方法通过将模糊隶属度函数应用于近邻实例的权重计算,从而提高了 kNN 算法在处理不确定性和模糊边界实例时的鲁棒性和灵活性,进一步扩展了其适用范围,特别是类间过渡模糊或实例分布不明确的复杂数据集.

1.2　距离度量优化

距离度量优化是指引入属性加权、结构扩展等技术来克服距离度量中属性独立性假设问题,从而提高距离度量准确性,进而改善 kNN 算法在真实应用场景下的分类效果.

属性加权通过给不同属性特征分配不同的权重,以减小冗余或不相关特征对距离度量结果的影响. 现有的属性加权方法大致可以分为两类:过滤式和包裹式. 其中,过滤式属性加权以蒋良孝和李超群等人提出的基于信息增益的加权方法[16]和基于 KL 散度的加权方法[17]最具代表性,它们根据数据集先验知识直接计算权值,然后将这些权值引入值差度量[18,19],从而提升值差度量的度量性能. 而包裹式属性加权则需要根据后续算法的分类效果来不断迭代优化权值. 例如[20]利用差分演化算法来搜索可行域内的最优属性权值,从而提升了反转类指定距离度量的度量性能[21]. 相比于过滤式属性加权,包裹式属性加权因在权值估计中考虑了后续算法的分类性能,因此往往能够获得更加准确的分类结果.

结构扩展通过在朴素贝叶斯网络结构[22]中添加有限的有向边来表示属性之间的依赖关系,然后通过将这种依赖关系引入距离度量,从而放宽距离度量的属性独立性假设. 树扩展的贝叶斯分类器[23]能够从数据集中学到最优的一依赖扩展结构,但学习最优贝叶斯网络容易引入较高的计算复杂度. Webb 等人提出了一种平均一依赖估测器[24],通过为每个属性特征构建一颗相应的一依赖扩展结构,获得好的属性依赖关系表示的同时降低了结构扩展的复杂度. 除此之外,k 依赖的扩展结构[25]为每个属性特征找到 k 个存在依赖关系的相关特征,进一步提升了距离度量对具有复杂属性依赖关系数据集的处理能力. 但由于许多真实数据集并不满足每个属性特征都只拥有 k 相关属性特征条件,导致基于 k 依赖扩展结构改进的距离度量容易产生过拟合问题.

1.3　k 值自适应

k 值自适应通过为每个测试实例找到一个相应的 k 值来构建灵活的局部分类模型,解决了传统 kNN 算法因固定 k 值造成的欠拟合、过拟合等问题. Zhang 等人[26]针对全局固定 k 值无法适应数据局部特性的问题,提出了一种基于索引结构的 k 值自适应方法. 该方法利用稀疏重构技术为每个样本确

定一个对应的最优 k 值,并将该 k 值作为样本的“伪标签”构建索引结构,从而使测试阶段的每个样本都能高效地检索到适配的 k 值,显著提升了 kNN 算法的分类效率. 为应对固定 k 值在数据密度分布不均匀时导致性能下降的挑战,Pan 等人[27]提出了一种局部信息驱动的 k 值自适应方法. 该方法首先为每个待分类样本选取一个较大的初始 k 值获取邻居集合,随后通过分析邻居中不同类别样本的分布情况,评估局部判别能力,并最终选择最具判别力的邻居数量作为该样本的最优 k 值,有效提升了复杂数据分布下 kNN 算法的分类精度. 尽管上述方法在一定程度上缓解了传统 kNN 的局限性,但它们大多依赖单一视角的数据特征,容易受到信息表征不全面的影响,进而限制分类性能. 为此,Fan 等人[28]提出了一种基于多视角学习的 k 值自适应方法. 该方法通过融合多个视角下提取的特征信息,充分挖掘不同视角间的互补性,实现更为准确的 k 值估计与邻居选择,在提升分类性能的同时增强了模型在复杂环境下的鲁棒性. 近年来,另一个受到研究者广泛关注的 k 值自适应思路是将其与集成学习机制相结合. 例如,Ali 等人[29]提出了一种基于扩展邻域规则和特征子集的 k 值自适应集成方法. 该方法通过引入扩展邻域规则与特征子集划分机制,构建多个具有差异化 k 值设定的子分类器,并在决策层进行集成,形成一种多样性与稳定性兼顾的 k 值邻居集成框架,有效提升了 kNN 算法的分类性能,特别适用于高维、异构和复杂分布的数据场景.

上述改进方法从不同角度解决了 kNN 算法分类精度不高的问题. 尽管如此,这些方法都只考虑了专家定义的原始数据特征. 本文认为,大多数实际应用都相当复杂,在这些情况下,实际的类别标签通常取决于数据特征的许多不同方面. 手工制作的原始属性虽然能从单方面反映某些数据特征,但在实际应用中,即使是最聪明的领域专家也很难手动提取出所有所需的数据特征. 因此,提取一些高层次的隐含特征来更全面地描绘数据特征是有意义且必要的.

2 预备知识

2.1 超父类一依赖估测器

超父亲一依赖估测器[30](Super-Parent One-Dependence Estimators,SPODE)是一种改进的一依赖分类器,属于贝叶斯网络分类器的变体. 传统朴素贝叶斯假设所有特征相互之间完全独立,而一依赖分类器放宽了这一假设,允许某些特征之间存在依赖关系. SPODE 进一步优化,选择一个“超父亲节点”,所有其他特征都依赖于该超父节点和类别变量,从而在保证模型计算效率同时提高模型的分类性能.

假设每个实例 x 对应一个向量 $<a_1(x),a_2(x),\cdots,a_m(x)>$ 和一个类标记 $c(x)$. 那么,给定一个类标记未知的测试实例 y,贝叶斯分类器首先计算实例 y 属于每一个类的概率,然后将实例分到概率最大的那一类. 贝叶斯分类器的具体定义如下:

$$c(y)=\underset{c\in C}{\operatorname{argmax}}P(c)P(a_1(y),a_2(y),\cdots,a_m(y)\mid c) \quad (2)$$

然而,完全估测公式(2)中的条件概率是一个 NP-hard 的问题. 为了简化计算,朴素贝叶斯分类器作出了属性独立性假设,即在给定类的情况下,属性之间完全独立. 图 1 给出了朴素贝叶斯网络结构示意图.

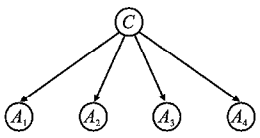


图 1 朴素贝叶斯网络结构示意图

Fig. 1 Schematic diagram of the structure of naive Bayes network

图 1 中 C 表示类变量,$A_i(i=1,2,\cdots,4)$ 表示 4 个属性特征. 类变量与所有属性特征存在关联,但属性与属性之间相互独立. 因为图 1 中只有从类变量指向其他所有属性特征的有向边. 那么,根据属性独立性假设,可得朴素贝叶斯分类器的定义如下:

$$c(y)=\underset{c\in C}{\operatorname{argmax}}P(c)\prod_{i=1}^{m}P(a_i(y)\mid c) \quad (3)$$

尽管朴素贝叶斯分类器表现出了较好的分类性能,但当数据集中属性特征之间存在依赖关系时,朴素贝叶斯分类器的性能就会受到损害. SPODE 利用结构扩展技术对朴素贝叶斯分类器进行改进. 即从所有属性特征中选择一个作为其他属性特征的父亲,然后再朴素贝叶斯网络结构从添加从该属性特征指向其他属性特征的有向边,从而放宽朴素贝叶斯分类器的属性独立性假设. 引入属性依赖关系的 SPODE 定义如下:

$$c(y)=\underset{c\in C}{\operatorname{argmax}}P(a_{ip}(y),c)\prod_{i=1}^{m}P(a_i(y)\mid a_{ip}(y),c) \quad (4)$$

其中,$a_{ip}(y)$ 表示超父亲节点.

2.2 随机森林

随机森林[31,32](Random Forest,RF)是一种采用并行机制的集成学习算法. 它由多个决策树组成,通过投票或平均融合来自不同决策树的预测结果. RF 通过训练实例扰动和属性特征扰动实现好而不同集成学习. 具体来说,RF 首先通过 Bootstrap 采样,从原始训练集中有放回的随机采样得到一个新的训练集,然后利用这个新的训练集来训练单颗决策树. 在单颗决策树构建过程中,采用随机选择策略来获得划分属性特征,在保证单颗决策树的准确性同时增强它的多样性.

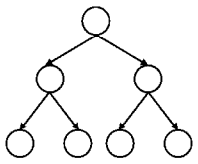


图 2 决策树示意图

Fig. 2 Schematic diagram of decision tree

传统决策树算法通过递归划分数据集构建一个树形模型,从而实现对未知数据的分类或回归预测,如图 2 所示. 算法从根节点开始,根据当前节点的属性特征将数据集划分为若干子集,每个子集对应一个分支. 递归过程持续到满足终止条件为止(如子集纯度达到阈值或特征已用完),最终形成叶节点. 因此,如何选择最优划分特征是决策树算法获得好的性能的关键.

信息增益是衡量某个属性特征划分数据集后,系统不确

定性的减少程度的重要指标,也是传统决策树算法选择划分特征的重要依据. 其具体定义如下:

$$InfoGain(D,A_i) = Ent(D) - \sum_{i=1}^{V}\frac{|D_i|}{|D|}Ent(D_i) \quad (5)$$

其中,D 表示训练集,V 表示属性 A_i 的不同取值个数. $Ent(D)$ 代表训练集 D 未被任何属性特征划分之前的不确定性(或信息熵),其具体计算公式如下:

$$Ent(D) = -\sum_{k-1}^{t}p_k\log_2 p_k \quad (6)$$

其中,p_k 表示属于第 k 个类别的实例数量占总的实例数量的比例. t 是训练集中总的类别个数.

与传统决策树算法不同,RF 在构建单颗决策树时引入了随机选择策略. 即在树的每个节点分裂时,仅从全部属性特征中随机选取 d 个候选特征($d \leqslant m$),然后根据这 d 个候选特征的信息增益选择出最优的划分特征. 解决了传统决策树在相同特征集中选择最优分裂点,容易生成相似的树,导致集成效果有限的问题.

2.3　D-S 证据理论

D-S 证据理论(Dempster-Shafer Theory)是一种用于处理多源决策信息融合问题的方法,在自动驾驶[33]、网络安全入侵检测[34]等领域被广泛使用. 其核心思想是通过信度函数量化多源不确定性证据,然后利用 Dempster 组合规则整合这些不确定性证据,以得到最终的决策结果.

对于一个决策问题,设人们所能认识到的预测结果可以用集合 Θ 表示,那么人们所关心的任一命题都对应于 Θ 的一个子集,而 Θ 被称为识别框架. 如果 A 为识别框架下的一个命题,m 为识别框架下的基本可信度分配,那么就有:

$$\sum_{A\subseteq\Theta} m(A) = 1 \quad (7)$$

信度函数用于量化对某个命题的最低可信度. 它通过整合所有直接支持该命题及其子集的证据,提供一种比传统概率更灵活的不确定性度量. 假设 B 是识别框架下的另一个命题,则命题 A 的信度函数可以被定义如下:

$$Bel(A) = \sum_{\varphi\neq B\subseteq A} m(B) \quad (8)$$

给定两个独立证据源,它们各自的信度函数分别为 Bel_i 和 Bel_j,其对应的基本信度分配分别是 m_i 和 m_j. 那么,Dempster 组合规则就是对两个信度函数求直和:

$$m_{i\oplus j}(A) = m_i(A)\oplus m_j(A) = \frac{1}{1-K}\cdot\sum_{A_i\cap A_j=A} m_i(A_i)m_j(A_j) \quad (9)$$

其中,$(1-K)^{-1}$ 是归一化因子,用于避免证据组合时将非零的信度赋给空集. K 的具体定义如下:

$$K = \sum_{A_i\cap A_j=\varphi} m_i(A_i)m_j(A_j) \quad (10)$$

上述 Dempster 组合规则还可以进一步推广到任意多个独立证据源的情况:

$$\begin{aligned} &m_{1\oplus 2\oplus\cdots\lambda}(A)\\ &= m_1(A)\oplus m_2(A)\oplus\cdots\oplus m_\lambda(A)\\ &= \frac{1}{1-K}\cdot\sum_{A_1\cap A_2\cap\cdots\cap A_\lambda=A} m_1(A_1)m_2(A_2)\cdots m_\lambda(A_\lambda)\end{aligned} \quad (11)$$

2.4　距离加权的 k 近邻算法

距离加权的 kNN 算法[13,35](Distance Weighted k-Nearest neighbor algorihtm,DWkNN)是一种基于分类决策规则升级的 k 近邻改进算法. 与传统 kNN 算法采用多数投票法,通过统计测试实例近邻所属类标记个数来实现推理测试实例的类标记推理不同,DWkNN 聚焦于更加准确的类成员概率估计. 并且在类成员概率估计中引入了距离相关权重因子,使模型在估计类成员概率同时,综合考虑各近邻实例与测试实例之间的距离,从而有效提升 kNN 算法的分类性能. DWkNN 的具体定义如下:

$$P(c|x) = \frac{1+\sum_{q-1}^{k} w_q\cdot\delta(c,c_q)}{\sum_{q=1}^{k} w_q} \quad (12)$$

其中,$P(c|x)$ 是测试实例 x 属于类别 c 的条件概率. w_i 为第 i 个近邻的权值,它是距离的单调递减函数:

$$w_q = \frac{1}{1+d(x,x_q)^2} \quad (13)$$

公式(13)中的 $d(x,x_q)$ 表示测试实例与第 q 个近邻之间的距离. 如果测试实例与这个近邻之间的距离越大,则这个近邻类标记对测试实例类成员概率估计的贡献就越小. 反之,如果测试实例与这个近邻之间的距离越小,则这个近邻类标记对测试实例类成员概率估计的贡献就越大.

由于大多数 k 近邻算法采用的欧式距离只适用于连续数值性属性,而在金融、医疗等许多应用领域的真实数据集中名词性属性广泛存在. 因此,本文选取能够处理名词性属性的反转类指定距离度量[21]来估计两个实例之间的距离. 其具体计算公式如下:

$$d(x,x_q) = \sqrt{\sum_{i=1}^{m} d(a_i(x),a_i(x_q))^2} \quad (14)$$

其中:

$$d(a_i(x),a_i(x_q)) = \begin{cases} 1 & a_i(x)\text{缺失}\\ 0 & a_i(x) = a_i(x_q)\\ 1-P(a_i(x)|c(x_q)) & \text{其他}\end{cases} \quad (15)$$

3　基于多视图生成和证据理论的 kNN 分类算法

EnDWkNN 的总体框架如图 3 所示. 从图 3 可以看出,EnDWkNN 可分为两个阶段:多视图生成和多视图融合. 在多视图生成阶段,由专家定义的数据特征被称为原始属性视图. EnDWkNN 利用原始属性视图构建 m 个差异化的超父亲一依赖估测器和随机森林,并依次利用每个超父亲一依赖估测器和随机森林对原始属性视图中每个训练实例进行分类. 利用所有的预测类标记构建两个标签视图. 在原始属性视图中,假设第 i 个训练实例可以表示成向量 $<a_{i1},a_{i2},\cdots,a_{im},c_i>$,那么,由超父亲一依赖估测器预测类标记构成的标签视图中的第 i 实例可以表示成向量 $<b_{i1},b_{i2},\cdots,b_{im},c_i>$,而由随机森林预测类标记构成的标签视图中的第 i 实例可以表示成向量 $<d_{i1},d_{i2},\cdots,d_{im},c_i>$. EnDWkNN 分别在原始属性视图和两个生成的标签视图上构建距离加权的 k 近邻分类算法,分别记为 M1、M2 和 M3. 在多视图融合阶段,依次用上述 3 个算法模型对每个待测实例进行预测,并利用 D-S 证据理论对预测的类成员概率进行融合,最终得到每个实例的预测类标记.

3.1　多视图生成

如上所述,EnDWkNN 由一个属性视图和两个标签视图

组成. 其中,属性视图由专家定义的属性组成,而两个标签视图则通过自学习机制动态构建. 本小节将深入介绍如何借助自学习过程来生成这两个新的标签视图.

具体而言,如图 3 左上角的矩形所示,EnDWkNN 首先会每个属性特征依次设定为所有其它属性特征的“超父亲节点”,进而构建出 m 个超父亲一依赖估计器,这里用 S 来统一

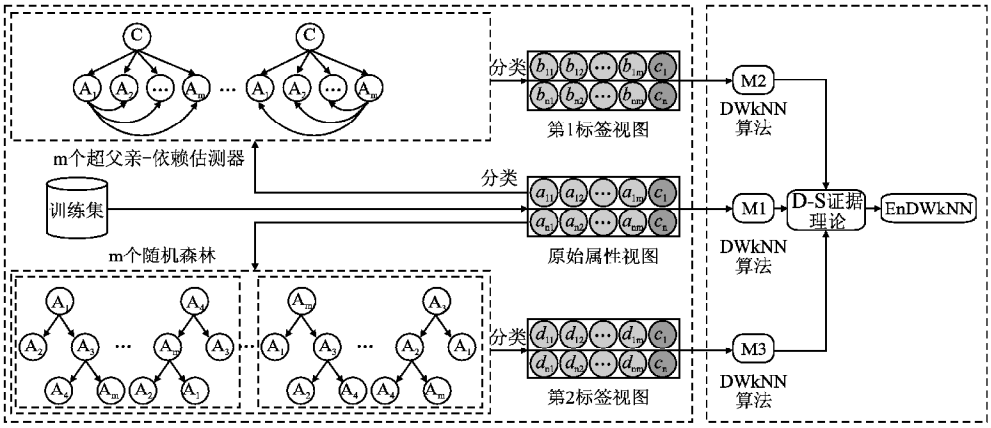


图 3 基于多视图生成和证据理论的 kNN 分类算法

Fig. 3 kNN classification algorithm based on multi-view generation and evidence theory

表示这些估测器集合. 接着,算法利用每个构建的超父亲一依赖估计器,对训练集中每个实例逐一进行分类预测,最终综合所有预测得到的类标签,构建出第 1 个标签视图. 假设第 j 个超父亲一依赖估计器被表示成 S_j,那么,算法用它来预测第 i 个训练示例的类标签,并将得到的预测类标记当作第 i 个实例第 j 个属性特征的取值来构建第 1 个标签视图的计算过程如下:

$$c_1^j(x_i)=\underset{c\in C}{\operatorname{argmax}}P(a_j(x_i),c)\prod_{h=1,h\neq j}^{m}P(a_h(x_i)\mid a_j(x_i),c) \tag{16}$$

$$b_{ij}=c_1^j(x_i) \tag{17}$$

由于每个超父亲一依赖估计器都有一个不同的超父亲节点,因此它们都可以从其独特的角度反映一些特定的数据特征. 除此之外,决策树同样能够在包含较多强相关属性依赖的数据集中表现出良好的分类性能. 为此,EnDWkNN 进一步引入随机森林来构建第二个标签视图,具体操作如图 3 左下角矩形所示. 在这里,算法用 R 来代表每个随机森林模型. 在随机森林的构建过程中,每棵决策树的生成均采用了随机选择策略. 具体而言,在每个节点进行分裂时,算法不会考虑全部属性特征,而是仅从所有属性特征中随机挑选出 $\log_2 m$ 个候选特征. 随后,依据这 $\log_2 m$ 个候选特征的信息增益,从中甄选出最优的划分特征,从而完成决策树的构建.

与第 1 个标签视图类似,在 m 个随机森林模型构建完成后,每个随机森林模型都被用来对每个训练示例进行分类. 假设第 j 个随机森林模型表示为 R_j,它使用式(18)来预测第 i 个训练实例的类标记,然后使用式(19)来获得第 i 个生成实例第 j 个属性特征的取值,从而完成第 2 个标签视图的构建.

$$c_2^j(x_i)=\underset{c\in C}{\operatorname{argmax}}\sum_{i=1}^{T}\delta(c_i,c) \tag{18}$$

$$d_{ij}=c_2^j(x_i) \tag{19}$$

总的来说,专家所定义的原始属性视图仅能从单一角度反映数据的特征. 与之形成鲜明对比的是,新构建的两个标签视图是基于超父亲一依赖估测器与随机森林这两种模型,对每个训练样本进行预测后所得到的类标签而构成的. 这两种模型本身具备从多个不同角度挖掘数据特征的能力. 在预测过程中,每个生成的类标签都全面考量了所有原始属性特征,可将其视为蕴含在更高层次的隐含属性特征. 因此,相较于仅使用单个原始属性视图,联合新构建的两个标签视图可被看作是对训练实例更全面、更高层次的描述,能够为后续分类模型提供更为丰富和有价值的判别信息.

3.2 多视图融合

根据原始属性视图和两个新生成的标签视图构建 3 个不同的 DWkNN 模型 M1、M2 和 M3. 3 个模型对测试实例 x 的类概率估计分别为:

$$\begin{aligned}P(c\mid x)_{M1}&=[p(c_1\mid x)_{M1},p(c_2\mid x)_{M1},\cdots,p(c_t\mid x)_{M1}]\\P(c\mid x)_{M2}&=[p(c_1\mid x)_{M2},p(c_2\mid x)_{M2},\cdots,p(c_t\mid x)_{M2}]\\P(c\mid x)_{M3}&=[p(c_1\mid x)_{M3},p(c_2\mid x)_{M3},\cdots,p(c_t\mid x)_{M3}]\end{aligned} \tag{20}$$

在 D-S 证据理论中,$P(c\mid x)_{M1}$、$P(c\mid x)_{M2}$ 和 $P(c\mid x)_{M3}$ 被分别作为 3 个独立的证据源,式(20)给出了它们的基本信度分配. 基于 D-S 证据理论的多视图融合就是根据 Dempster 组合规则求这 3 个独立证据源基本信度分配的直和. 假设融合 3 个模型对测试实例 x 的类概率估计后得到的最终类条件概率估计为:

$$P(c\mid x)_{Fusion}=[p(c_1\mid x)_{Fusion},p(c_2\mid x)_{Fusion},\cdots,p(c_t\mid x)_{Fusion}] \tag{21}$$

那么,其中的任意元素 $p(c_k\mid x)_{\text{Fusion}}$ 可以通过以下公式求得:

$$p(c_k\mid x)_{Fusion}=\frac{1}{1-K}\cdot p(c_k\mid x)_{M1}p(c_k\mid x)_{M2}p(c_k\mid x)_{M3} \tag{22}$$

其中:

$$1-K=p(c_k\mid x)_{M1}p(c_k\mid x)_{M2}p(c_k\mid x)_{M3} \tag{23}$$

最后,本文提出的 EnDWkNN 算法将融合后的类条件概率估计中取值最大的元素对应的类标签作为测试实例 x 的类标签进行输出:

$$c(x)=\underset{c\in C}{\operatorname{argmax}}P(c\mid x)_{Fusion} \tag{24}$$

综上所述,EnDWkNN 的整个算法流程可以被划分成训练阶段和测试阶段两部分,分别如表 1 和表 2 所示. 训练阶段

的时间复杂度为 $O(md+mu+N+3v)$,其中 d、u 和 v 分别表示训练一个超父亲一依赖估测器、一个随机森林和一个距离加权的 k 近邻算法所需的时间. N 为总的训练实例个数. 测试阶段同样需要使用自学习机制,将测试实例转换到不同的标签视图. 由于该自学习机制使用的分类器在训练阶段已经训练好,因此,它的时间复杂度仅为 $O(2m+t)$,其中 t 表示用于 D-S 证据理论融合所的时间. 总之,EnDWkNN 是一个简单且有效的 k 近邻改进算法模型.

表 1　EnDWkNN 训练阶段算法流程

Table 1　Algorithm flow of EnDWkNN at training stage

输入:训练数据集 D
输出:m 个 SPODE 模型、m 个 RF 模型和 3 个 DWkNN 模型
Step 1. 对每个属性特征 $A_j(j=1,2,\cdots,m)$:
Step 2.　　以 A_j 为超父亲节点构建 SPODE
Step 3. 对第 j 次循环:
Step 4.　　随机选择 $\log_2 m$ 个特征构建随机决策树
Step 5.　　利用随机决策树构建 RF
Step 6. 对每个训练实例 $x_i(i=1,2,\cdots,N)$:
Step 7.　　利用构建的每个 SPODE,依据式(17)预测类标签
Step 8.　　预测得到的类标签当作第 1 标签视图中第 i 个实例的第 j 个属性值
Step 9. 对每个训练实例 $x_i(i=1,2,\cdots,N)$:
Step 10.　利用构建的每个 RF,依据式(19)预测类标签
Step 11.　预测得到的类标签当作第 2 标签视图中第 i 个实例的第 j 个属性值
Step 12. 依据式(12)~式(15),分别在每个视图下构建 DWkNN 模型
Step 13. 返回构建的 m 个 SPODE 模型、m 个 RF 模型和 3 个 DWkNN 模型

表 2　EnDWkNN 测试阶段算法流程

Table 2　Algorithm flow of EnDWkNN at testing stage

输入:m 个 SPODE 模型、m 个 RF 模型、3 个 DWkNN 模型和待测实例 x
输出:待测实例 x 的类标记 $c(x)$
Step 1. 对每个待测实例 x:
Step 2.　　利用构建的每个 SPODE,依据式(17)预测类标签
Step 3.　　预测得到的类标签当作第 1 标签视图中测试实例的第 j 个属性值
Step 4. 对每个待测实例 x:
Step 5.　　利用构建的每个 RF,依据式(19)预测类标签
Step 6.　　预测得到的类标签当作第 2 标签视图中测试实例的第 j 个属性值
Step 7. 使用构建的 DWkNN 模型 M1、M2 和 M3 分别预测 x 的类成员概率估计
Step 8. 通过式(20)~式(24)融合并预测测试实例 x 的类标记 $c(x)$
Step 9. 返回待测实例的类标记 $c(x)$

4　实验结果与分析

4.1　实验设置与数据预处理

本文的所有实验实现和结果分析均在怀卡托数据挖掘实验平台[36](Waikato Environment for Knowledge Analysis,WEKA)上完成,实验环境为 Windows 10 系统,CPU 频率为 3.80 GHz,内存为 16GB. 所有分类结果均采用 10 次十折交叉验证获得. 对于 EnDWkNN 算法,随机森林中决策树的个数 T = 10,距离加权 k 近邻算法的 $k=10$. 注意,随机森林中的决策树默认一直生长,直到没有更多实例可以分裂为止.

为了验证 EnDWkNN 算法的有效性,本文从现有的 k 近邻改进算法中挑选出一部分作为实验比较对象. 它们分别是:

· EkNN[37]:基于证据理论升级分类决策规则的 k-近邻算法,其参数设置为 $\alpha_0=0.95,\beta=1$;

· PNN[38]:基于模糊理论升级分类决策规则的 k-近邻算法,又称伪近邻算法;

· MkNN[39]:基于样本加权升级分类决策规则的 k-近邻算法,又称修正的 k 近邻算法,其参数设置为 H = n/10;

· GRWISCDM[40]:基于距离度量优化的 k 近邻算法,其运用距离度量为增益率加权的反转类指定距离度量;

· MAWVDM[18]:基于距离度量优化的 k 近邻算法,其运用的距离度量为双重属性加权修正的值差度量.

表 3　实验数据集描述

Table 3　Description of the experimental dataset

数据集	实例个数	属性个数	类标个数	缺失值	数值性属性
breast-canner	286	10	2	Y	N
breast-canner-w	699	10	2	Y	N
Car	1728	7	4	N	N
colic. ORIG	368	28	2	Y	Y
colic	368	23	2	Y	Y
credit-a	690	16	2	Y	Y
credit-g	1000	21	2	N	Y
cylinder-bands	512	40	2	Y	N
dermatology	366	35	6	Y	Y
diabetes	768	9	2	N	Y
eucalyptus	736	20	5	Y	Y
eye_movements	10936	28	3	N	Y
grub-damage	155	9	4	N	Y
haberman	306	4	2	N	Y
hayes-roth	132	5	3	N	N
heart-statlog	270	14	2	N	Y
hepatitis	155	20	2	Y	Y
ionosphere	351	35	2	N	Y
iris	150	5	3	N	Y
kr-vs-kp	3196	37	2	N	N
labor	57	17	2	Y	Y
landsat_test	2000	37	6	N	Y
landsat_train	4435	37	6	N	Y
letter	20000	17	26	N	Y
monks	432	7	2	N	N
mushroom	8124	23	2	Y	N
optdigits	5620	65	10	N	Y
pasture	36	23	3	N	Y
pendigits	10992	17	10	N	Y
promoters	106	58	2	N	N
segment	2310	20	7	N	Y
sick	3772	30	2	Y	Y
solar-flare_2	1066	13	6	N	N
sonar	208	61	2	N	Y
spambase	4610	58	2	N	Y
splice	3190	62	3	N	N
vehicle	846	19	4	N	Y
vote	435	17	2	Y	N
vowel	990	14	11	N	Y
waveform-5000	5000	41	3	N	Y

除了与现有的改进算法作比较,本文还分别开展了一组

消融实验和一组 k 值敏感性分析实验，以揭示多视图生成组件与基于 D-S 证据理论的多视图融合组件以及不同 k 取值对算法分类性能的影响.

本文从加州大学欧文分校机器学习库[41]（UC Irvine Machine Learning Repository，UCI）选取了 40 个涉及医疗、金融等不同领域的数据集. 表 3 给出了这些数据集包含实例个数、属性个数、类标个数、缺失值、数值性属性的详细信息. 其中，表格内的“Y”代表数据集存在缺失值或者存在数值性属性，“N”代表数据集不存在缺失值或者不存在数值性属性. 由于实验选择的模型都适用于处理不存在缺失值的名词性属性，致部分数据集不可直接被使用，因此在实验前本文对全部实验数据进行了以下预处理工作，这些数据预处理工作在相关研究中也得到了应用[42]：首先，最小描述长度[43]（Minimum Describe Length，MDL）对所有数值性属性进行离散化处理. 然后使用名词性属性的众数替换了名词性属性的缺失值. 同时，当一个属性的属性值个数等于数据集的实例个数时，这个属性就对分类建模没有任何作用. 因此，本文提前手动删除了这样的属性：“colic. ORIG”数据集中的“Hospital Number”属性以及“zoo”数据集中的“animal”属性.

4.2 评估指标

4.2.1 分类准确率

分类准确率是衡量分类算法整体预测性能最直接、最朴素的指标. 其核心思想是计算模型在测试集（或验证集）上预测正确的样本数量占所有测试样本总数的比例. 该指标反映了模型做出正确分类决策的总体能力，即模型将样本分配到其真实类别的比例有多高. 准确率越高，表明模型在给定数据集上的整体分类性能越好.

分类准确率指标的计算公式如下：

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100 \tag{25}$$

其中，TP 表示模型预测为正类，且真实标签也是正类的样本数. TN 表示模型预测为负类，且真实标签也是负类的样本数. FP 表示模型预测为正类，但真实标签是负类的样本数. FN 表示模型预测为负类，但真实标签是正类的样本数.

4.2.2 根相对平方误差

根相对平方误差（Root Relative Square Error，RRSE）通过计算基分类器在测试实例下的预测概率和真实类标概率之间的差异来评估分类模型预测的不确定性相对于一个简单基准模型的改进程度. 与分类准确率不同的是，根相对平方误差的取值越小，说明分类算法性能越好.

根相对平方误差指标的计算公式如下：

$$RRSE(\Theta|X) = \sqrt{\frac{\sum_{q=1}^{N}\sum_{k=1}^{t}(\hat{P}_{qk} - P_k)^2}{\sum_{q=1}^{N}\sum_{k=1}^{t}(P_{qk} - P_k)^2}} \times 100 \tag{26}$$

其中，N 是测试实例的总数，t 为数据集类标记的总个数，$\hat{P}_{qk}$ 为算法预测测试实例 x_q 的类标记为 c_k 的概率，P_{qk} 为测试实例 x_q 的类标记为 c_k 的真实概率值，P_k 为类标记 c_k 的先验统计概率值. 在分类算法的预测阶段，因为只知道测试实例的真实类标记，不知道测试实例的真实类标记概率，因此将测试实例 x_q 的真实类标记为 c_k 的真实概率 P_{qk} 赋值为 1，否则为 0.

4.3 结果与分析

表 4 和表 5 分别列出了 EnDWkNN 算法与 EkNN、PNN、MkNN、GRWISCDM、MAWVDM 在 40 个数据集上的分类准确率与根相对平方误差的双尾配对 t 检验[44,45] 比较结果. 双

表 4 不同 kNN 改进算法的分类准确率比较

Table 4 Comparison of classification accuracy of different kNN improvement algorithms

数据集	EnDWkNN	EkNN	PNN	MkNN	GRWISCDM	MAWVDM
breast-canner	70.23 ±7.33	71.41 ±7.11	68.08 ±7.85	72.57 ±7.38	70.69 ±7.71	73.79 ±5.68
breast-canner-w	96.15 ±2.27	97.04 ±1.77	94.51 ±2.68 ◦	72.57 ±7.38	96.88 ±1.89	96.84 ±1.84
car	94.38 ±1.66	93.43 ±1.90	91.04 ±2.07 ◦	90.31 ±2.24 ◦	87.31 ±2.08 ◦	92.17 ±3.18 ◦
colic. ORIG	74.03 ±6.58	73.45 ±6.36	74.48 ±7.07	72.31 ±5.93	74.70 ±5.93	75.55 ±6.13
colic	83.01 ±5.99	82.77 ±6.06	77.23 ±5.53 ◦	82.09 ±6.20	82.20 ±5.95	82.85 ±5.56
credit-a	86.16 ±3.76	84.84 ±4.10	80.58 ±4.68 ◦	86.62 ±4.08	86.42 ±3.89	86.22 ±3.50
credit-g	74.91 ±3.55	75.55 ±3.44	70.78 ±3.86 ◦	75.31 ±3.14	73.66 ±3.75	73.53 ±3.68
cylinder-bands	81.69 ±5.45	76.41 ±5.18 ◦	83.37 ±4.66	74.30 ±5.32 ◦	77.39 ±5.45 ◦	74.02 ±5.64 ◦
dermatology	97.62 ±2.41	96.86 ±2.67	97.18 ±2.44	96.72 ±2.77	97.98 ±2.32	96.61 ±2.70
diabetes	75.33 ±3.94	77.38 ±4.43	56.34 ±4.38 ◦	77.85 ±4.31	77.54 ±4.41	78.24 ±4.71 ·
eucalyptus	62.84 ±5.28	61.67 ±5.75	60.03 ±5.92	61.10 ±5.73	65.26 ±5.08	64.63 ±5.76
eye_movements	59.97 ±4.53	58.39 ±4.75	56.28 ±4.55 ◦	57.59 ±4.80	54.65 ±4.77 ◦	52.42 ±4.88 ◦
grub-damage	41.77 ±10.65	44.86 ±10.14	42.17 ±10.46	45.97 ±10.44	47.19 ±12.22	47.54 ±11.22
haberman	74.85 ±5.27	74.65 ±5.40	26.57 ±2.18 ◦	72.62 ±4.97	72.42 ±6.43	73.77 ±5.57
hayes-roth	72.14 ±10.29	77.08 ±8.72	79.03 ±9.61 ·	70.63 ±10.95	71.68 ±9.85	80.32 ±8.54 ·
heart-statlog	84.67 ±6.53	84.48 ±6.90	72.22 ±7.92 ◦	84.19 ±6.51	82.63 ±6.86	84.00 ±6.61
hepatitis	85.77 ±8.39	87.06 ±8.33	83.92 ±9.04	87.71 ±7.66	85.24 ±9.73	82.20 ±8.89
ionosphere	93.47 ±4.12	91.77 ±4.36	93.53 ±3.93	90.66 ±4.60 ◦	91.63 ±4.47	90.65 ±4.18 ◦
iris	94.33 ±5.64	93.40 ±5.72	87.13 ±6.98 ◦	94.33 ±5.14	94.27 ±5.10	94.33 ±5.05
kr-vs-kp	92.07 ±4.34	86.65 ±5.90 ◦	83.29 ±5.80 ◦	82.57 ±6.23 ◦	92.20 ±3.70	91.92 ±4.33
labor	95.43 ±9.54	93.87 ±11.51	92.77 ±11.26	92.57 ±11.45	91.93 ±12.03	85.83 ±12.78 ◦
landsat_test	89.14 ±2.10	86.70 ±2.17 ◦	88.10 ±2.11	86.07 ±2.11 ◦	86.90 ±2.13 ◦	88.10 ±1.96 ◦
landsat_train	88.39 ±4.20	84.82 ±5.04 ◦	87.51 ±4.88	84.05 ±4.97 ◦	84.66 ±5.03 ◦	86.85 ±4.78
letter	78.86 ±3.22	75.73 ±3.34 ◦	80.05 ±3.04	70.06 ±3.38 ◦	73.20 ±3.40 ◦	75.13 ±3.04 ◦
monks	98.94 ±1.56	93.13 ±5.03 ◦	99.61 ±1.19	77.57 ±5.58 ◦	77.15 ±11.7 ◦	76.03 ±9.37 ◦
mushroom	99.85 ±0.40	99.54 ±0.65	99.85 ±0.40	98.31 ±1.60 ◦	99.47 ±0.91	98.15 ±1.51 ◦
optdigits	93.47 ±3.20	93.13 ±3.42	93.40 ±2.91	90.62 ±3.82 ◦	91.07 ±3.63 ◦	93.31 ±3.24
pasture	90.42 ±14.23	89.17 ±14.43	87.00 ±15.14	89.92 ±14.33	90.08 ±14.39	81.50 ±16.00

续表 4

数据集	EnDWkNN	EkNN	PNN	MkNN	GRWISCDM	MAWVDM
pendigits	91.53 ±2.24	88.21 ±3.13◦	91.51 ±2.23	82.17 ±3.45◦	84.81 ±3.21◦	89.85 ±2.66
promoters	91.51 ±8.66	90.26 ±9.16	87.85 ±10.10	89.15 ±9.83	91.21 ±9.68	92.15 ±7.73
segment	88.06 ±6.17	88.88 ±6.03	87.50 ±6.60	87.63 ±6.35	85.20 ±7.06	88.10 ±6.52
sick	97.48 ±2.11	97.03 ±2.23	97.37 ±2.40	96.58 ±2.17	97.09 ±2.11	97.16 ±2.26
solar-flare_2	74.42 ±2.91	72.98 ±3.58	71.24 ±2.53◦	75.27 ±3.49	73.38 ±3.05	73.29 ±3.18
sonar	85.15 ±6.89	84.80 ±7.43	84.96 ±7.31	83.76 ±7.77	80.75 ±9.18◦	75.51 ±9.39◦
spambase	90.15 ±3.84	90.07 ±3.84	90.72 ±4.02	90.09 ±3.85	91.04 ±3.76	87.93 ±4.42
splice	91.44 ±4.59	88.18 ±5.22◦	85.48 ±6.03◦	87.21 ±5.24◦	93.26 ±4.07	93.01 ±4.70
vehicle	73.26 ±3.87	71.80 ±3.43	70.23 ±3.96◦	70.24 ±3.24◦	71.31 ±3.32	67.19 ±3.79◦
vote	95.77 ±2.92	94.39 ±3.30	92.98 ±3.75◦	93.75 ±3.54◦	93.63 ±3.28◦	95.56 ±2.77
vowel	89.97 ±2.99	85.37 ±3.54◦	91.39 ±2.60	72.57 ±4.30◦	68.89 ±4.19◦	65.03 ±4.27◦
waveform-5000	86.68 ±5.76	84.46 ±5.67	81.66 ±5.20◦	80.74 ±5.49◦	82.66 ±5.45◦	80.04 ±5.20◦
Average	84.63	83.54	80.97	81.78	82.24	82.03
W/T/L		9/31/0	16/23/1	16/23/1	13/27/0	13/25/2

表 5　不同 kNN 改进算法的根相对平方误差比较

Table 5　Comparison of root relative rquared error of different kNN improvement algorithms

数据集	EnDWkNN	EkNN	PNN	MkNN	GRWISCDM	MAWVDM
breast-canner	109.09 ±13.57	105.61 ±14.19	122.67 ±15.38◦	99.13 ±12.56·	98.86 ±9.19·	96.05 ±7.52·
breast-canner-w	37.06 ±13.73	32.17 ±12.66	47.73 ±12.48◦	30.92 ±12.97·	34.92 ±8.48	34.94 ±7.50
car	43.17 ±6.17	46.57 ±6.49	62.18 ±7.47◦	55.38 ±3.26◦	63.26 ±2.98◦	60.91 ±4.92◦
colic. ORIG	94.68 ±11.08	95.50 ±11.07	105.80 ±15.12◦	90.28 ±8.55	86.29 ±7.58·	86.40 ±8.31·
colic	77.27 ±14.00	79.36 ±14.20	98.06 ±12.37◦	78.45 ±13.23	75.23 ±9.53	74.54 ±9.16
credit-a	68.67 ±9.94	71.63 ±9.78	88.03 ±10.77◦	65.98 ±9.64	64.84 ±6.10	64.02 ±6.32·
credit-g	100.35 ±6.85	98.35 ±6.59	117.70 ±7.81◦	91.84 ±4.85·	92.76 ±4.58·	92.28 ±4.32·
cylinder-bands	74.19 ±11.02	87.96 ±9.58◦	81.68 ±12.11◦	84.47 ±7.58◦	79.71 ±6.48◦	89.29 ±8.01◦
dermatology	24.57 ±5.66	20.66 ±11.56	21.55 ±15.57	20.54 ±8.94·	49.36 ±1.91◦	48.82 ±2.67◦
diabetes	95.16 ±7.31	96.67 ±9.41	138.46 ±7.13◦	86.12 ±8.66·	83.36 ±5.75·	82.60 ±5.92·
eucalyptus	83.06 ±5.51	87.79 ±6.30◦	100.76 ±7.56◦	83.21 ±5.59	78.16 ±3.23·	78.60 ±2.88·
eye_movements	94.01 ±4.96	99.00 ±5.20◦	115.09 ±6.01◦	93.65 ±4.41	91.77 ±2.84	93.41 ±2.45
grub-damage	112.94 ±10.19	109.27 ±10.52	125.56 ±11.61◦	102.22 ±10.17·	94.46 ±6.56·	95.90 ±5.09·
haberman	106.08 ±10.74	111.81 ±12.16◦	194.24 ±4.01◦	105.76 ±9.25	108.90 ±11.27	97.01 ±5.64·
hayes-roth	80.46 ±15.54	72.22 ±16.43·	77.80 ±20.10	73.19 ±11.39	68.36 ±8.97·	65.31 ±8.71·
heart-statlog	71.44 ±16.06	72.84 ±17.57	104.96 ±15.34◦	70.60 ±15.64	71.38 ±10.88	71.00 ±11.36
hepatitis	80.08 ±28.75	77.47 ±30.29	94.37 ±30.76	74.67 ±27.57	76.36 ±21.67	80.92 ±15.97
ionosphere	45.96 ±16.08	54.14 ±17.09◦	49.04 ±20.25	56.60 ±16.31◦	51.44 ±12.71◦	55.55 ±11.00◦
iris	31.49 ±23.03	33.49 ±23.62	59.13 ±19.16◦	30.09 ±22.49	34.02 ±12.67	33.83 ±12.28
kr-vs-kp	46.77 ±12.78	62.44 ±13.83◦	80.58 ±15.06◦	67.00 ±8.69◦	49.52 ±8.72	50.38 ±8.85
labor	23.22 ±34.90	25.66 ±34.74	32.30 ±46.23	30.16 ±29.24	47.26 ±26.37◦	63.26 ±24.38◦
landsat_test	47.10 ±3.09	53.35 ±4.28◦	53.84 ±4.89◦	53.45 ±4.15◦	62.31 ±1.73◦	68.59 ±1.26◦
landsat_train	51.69 ±5.93	55.63 ±9.61◦	54.23 ±11.10	56.14 ±9.05◦	66.84 ±3.37◦	72.72 ±2.69◦
letter	64.21 ±2.12	59.42 ±3.65·	64.24 ±4.94	64.27 ±3.13	85.19 ±0.72◦	84.61 ±0.69◦
monks	15.13 ±10.20	45.03 ±12.82◦	4.57 ±11.76·	70.63 ±9.05◦	76.87 ±21.87◦	78.87 ±16.48◦
mushroom	4.14 ±5.86	9.13 ±7.95◦	2.66 ±7.24	18.92 ±8.79◦	20.27 ±3.29◦	26.02 ±4.65◦
optdigits	66.72 ±2.13	32.49 ±7.72·	37.39 ±8.69·	37.06 ±6.52·	67.92 ±1.73◦	65.27 ±1.79·
pasture	30.86 ±35.92	37.77 ±40.74	41.43 ±46.53	37.33 ±33.69	43.22 ±18.97	58.18 ±16.58◦
pendigits	41.44 ±3.62	44.37 ±5.61◦	43.08 ±5.76	52.61 ±5.00◦	67.70 ±1.55◦	66.49 ±1.29◦
promoters	50.28 ±15.08	49.74 ±23.12	58.85 ±37.46	53.34 ±16.51	54.57 ±18.13	54.78 ±15.99
segment	43.01 ±9.42	42.36 ±12.27	51.93 ±15.41	45.30 ±9.45	64.52 ±3.44◦	69.15 ±3.13◦
sick	53.62 ±31.28	58.39 ±31.57	58.18 ±43.14	69.43 ±22.01◦	70.17 ±19.89◦	70.39 ±16.52◦
solar-flare_2	70.33 ±3.90	77.61 ±5.19◦	85.49 ±3.75	67.62 ±4.08·	66.99 ±2.08·	66.23 ±1.76·
sonar	66.40 ±17.01	68.66 ±18.29	74.22 ±23.14	66.41 ±15.87	70.38 ±13.63	79.47 ±11.06◦
spambase	56.32 ±12.19	58.64 ±12.35	59.91 ±13.77	57.94 ±11.61	53.24 ±9.50	62.94 ±7.63◦
splice	53.48 ±5.81	52.89 ±12.36	67.16 ±14.83	55.43 ±9.06	46.52 ±7.64·	46.45 ±8.54·
vehicle	72.46 ±4.73	78.21 ±4.88◦	88.91 ±6.03	73.51 ±3.67	72.19 ±2.54	75.09 ±2.37
vote	35.94 ±14.10	43.58 ±13.99◦	51.41 ±17.94	44.30 ±13.35◦	46.84 ±8.65◦	41.03 ±8.56
vowel	43.97 ±5.08	49.44 ±5.65◦	42.95 ±7.00	63.69 ±3.45◦	76.77 ±1.71◦	78.95 ±1.63◦
waveform-5000	55.73 ±10.11	61.17 ±10.75◦	73.45 ±11.06◦	63.78 ±8.86◦	62.71 ±6.53◦	66.29 ±4.66◦
Average	60.56	62.96	73.29	63.54	66.89	68.66
W/T/L		16/21/3	23/15/2	13/19/8	17/14/9	18/10/12

尾配对 t 检验的置信水平为 $p=0.05$. 表中的符号◦和·分别表示本文提出的 EnDWkNN 算法显著优于其他改进的 kNN 算法和本文提出的 EnDWkNN 算法明显差于其他改进的 kNN 算法. 此外,表 4 和表 5 底部汇总了 40 个数据集上分类

准确率和根相对平方误差分值的平均值以及“赢/平/输”(*Win/Tie/Lose*)个数.分类准确率和根相对平方误差分值的平均值提供了算法模型在所有数据集上的总体性能指示.“赢/平/输”(W/T/L)个数展示了 EnDWkNN 算法明显优于,或着持平,或者明显差于其他改进的 kNN 算法的数据集个数.从表 4 和表 5 提供的实验结果不难看出,相比于其他改进的 kNN 算法,本文提出的 EnDWkNN 算法总体性能是最优的.

现将观察到的实验结果总结如下:

1)在分类准确率方面,EnDWkNN(84.63%)在 40 个数据集上的平均分类准确率最高. EnDWkNN 明显优于 EkNN(83.54%)、PNN(80.97%)、MkNN(81.78%)、GRWISCDM(82.24%)、MAWVDM(82.03%). EnDWkNN 在 16 个数据集上明显优于 PNN 和 MKNN,在 13 个数据集上明显优于 GRWISCDM 和 MAWVDM,在 9 个数据集上明显优于 EkNN.

2)在根相对平方误差方面,EnDWkNN(60.56)在 40 个数据集上的平均根相对平方误差最低. EnDWkNN 明显优于 EkNN(62.96)、PNN(73.29)、MkNN(63.54)、GRWISCDM(66.89)、MAWVDM(68.66). EnDWkNN 在 16 个数据集明显优于 EkNN,在 23 个数据集上明显优于 PNN,在 13 个数据集上明显优于 MkNN,在 17 个数据集上明显优于 GRWISCDM,在 18 个数据集上明显优于 MAWVDM.

为了进一步说明本文所提 EnDWkNN 的显著优势具有统计学意义,本文对表 4 和表 5 中的分值做了威尔克森符号秩和检验[46,47].威尔克森符号秩和检验是一种基于符号秩的非参数统计方法.该方法计算配对算法观测值与目标算法测试值的差值,忽略符号对差值的绝对值排序赋秩,再将原始差值的正负符号赋予对应秩次.检验统计量取正秩之和或负秩之和,通过判断正秩之和或负秩之和的显著性,推断是否接受或拒绝目标算法显著优于配对算法的假设.

表 6 和表 7 展示了威尔克森符号秩和检验的统计结果.表中。表示列中的算法模型优于相应的行中的算法模型,而·则表示行中的算法模型优于列中的算法模型.本文将对角线的显著性水平设置为 $p=0.05$,上对角线的显著性水平设置为 $p=0.1$.从统计结果不难看出,在分类准确率方面,EnDWkNN 显著优于 EkNN、PNN、MkNN、GRWISCDM、MAWVDM,在跟相对平方误差方面,EnDWkNN 显著优于 EkNN、PNN、GRWISCDM、MAWVDM.

表 6 EnDWkNN 算法与其他对比算法的分类准确率分值的威尔克森符号秩和检验总体情况

Table 6 Summary of Wilcoxon signed-rank test for the classification accuracy scores of EnDWkNN and other comparison algorithms

	EnDWkNN	EkNN	PNN	MkNN	GRWISCDM	MAWVDM
EnDWkNN	–	·	·	·	·	·
EkNN	◦	–	·	·	·	
PNN	◦	◦	–			
MkNN	◦	◦		–		
GRWISCDM	◦				–	
MAWVDM	◦					–

表 7 EnDWkNN 算法与其他对比算法的 RRSE 分值的威尔克森符号秩和检验总体情况

Table 7 Summary of Wilcoxon signed-rank test for the root relative rquared error scores of EnDWkNN and other comparison algorithms

	EnDWkNN	EkNN	PNN	MkNN	GRWISCDM	MAWVDM
EnDWkNN	–	·	·		·	·
EkNN	◦	–	·			·
PNN	◦	◦	–	◦	◦	
MkNN		◦	·	–		·
GRWISCDM	◦				–	·
MAWVDM	◦			◦		–

4.3.1 消融实验

为了验证本文所提 EnDWkNN 算法中多视图生成组件与多视图融合组件的重要性,本文在上述实验的基础上开展了一组新的消融实验.实验结果如表 8 ~ 表 11 所示.注意,本文把只使用第 1 个标签视图和原始属性视图的算法命名为 EnDWkNN_SPODE,把只使用第一个标签视图和原始属性视

表 8 EnDWkNN 算法的消融实验准确率比较

Table 8 Comparison of accuracy in ablation experiments of EnDWkNN

数据集	EnDWkNN	EnDWkNN_RF	EnDWkNN_SPODE	DWkNN
breast-canner	70.23 ±7.89	71.02 ±7.66	68.93 ±6.92	71.97 ±7.32
breast-canner-w	96.18 ±2.24	96.38 ±2.15	95.89 ±2.37	96.85 ±1.87
car	94.36 ±1.58	91.79 ±1.97◦	93.78 ±1.77	93.85 ±1.75
colic. ORIG	74.13 ±6.06	72.72 ±7.56	74.49 ±5.67	73.70 ±6.64
colic	83.09 ±5.81	81.85 ±6.09	83.50 ±5.46	82.58 ±5.93
credit-a	85.99 ±3.78	84.71 ±4.30	85.67 ±4.02	85.06 ±4.14
credit-g	74.93 ±3.31	74.15 ±3.06	74.78 ±3.63	75.29 ±3.42
cylinder-bands	82.15 ±5.25	81.76 ±5.65	77.41 ±4.95◦	75.98 ±5.38◦
dermatology	97.54 ±2.53	97.71 ±2.34	97.57 ±2.33	96.97 ±2.72
diabetes	74.84 ±4.35	75.10 ±4.01	74.92 ±4.21	77.20 ±4.26
eucalyptus	62.66 ±5.27	62.06 ±5.55	62.46 ±4.83	61.53 ±5.70
eye_movements	60.02 ±4.58	56.17 ±4.65◦	60.44 ±4.42	57.47 ±4.84
grub-damage	42.35 ±10.66	45.40 ±9.15	39.57 ±10.58	47.12 ±10.56
haberman	75.13 ±4.49	74.62 ±4.70	74.32 ±5.24	74.16 ±5.48
hayes-roth	72.58 ±10.85	67.95 ±11.35	73.99 ±10.68	71.78 ±10.09
heart-statlog	84.70 ±5.99	82.74 ±6.70	85.22 ±6.04	83.89 ±6.58
hepatitis	85.64 ±8.79	86.81 ±8.42	84.82 ±9.16	87.18 ±8.38
ionosphere	93.33 ±3.81	92.93 ±3.82	92.96 ±4.30	91.00 ±4.58◦
iris	93.93 ±5.61	93.80 ±6.01	94.07 ±6.13	94.27 ±5.10

续表 8

数据集	EnDWkNN	EnDWkNN_RF	EnDWkNN_SPODE	DWkNN
kr-vs-kp	92.45 ±4.22	86.80 ±5.85◦	94.58 ±3.69	86.33 ±5.48◦
labor	95.40 ±10.00	95.63 ±8.67	94.53 ±9.92	93.60 ±11.15
landsat_test	89.01 ±1.98	87.82 ±2.08◦	89.06 ±1.98	87.07 ±2.16◦
landsat_train	88.41 ±4.18	87.17 ± 4.34	87.83 ±4.22	85.20 ±4.91◦
letter	78.77 ±3.07	75.84 ±3.37◦	79.52 ±2.85	75.80 ±3.33◦
monks	99.19 ±1.56	93.80 ±3.42◦	98.22 ±2.19	86.23 ±5.08◦
mushroom	99.85 ±0.40	99.37 ±0.79	99.83 ±0.50	99.56 ±0.67
optdigits	93.65 ±3.21	93.08 ±3.31	93.26 ±3.05	93.72 ±3.18
pasture	90.42 ±14.23	92.83 ±12.81	90.17 ±14.28	89.92 ±14.33
pendigits	91.32 ±2.39	91.56 ±2.47	89.92 ±2.42◦	88.01 ±3.12◦
promoters	91.59 ±8.43	89.38 ±10.99	92.55 ±8.10	90.05 ±9.13
segment	88.41 ±6.21	89.23 ±5.99	88.54 ±6.36	88.14 ±6.18
sick	97.56 ±2.09	97.61 ±2.15	97.83 ±1.93	96.85 ±2.34
solar-flare_2	74.38 ±3.07	74.64 ±3.24	73.95 ±3.00	74.77 ±3.43
sonar	84.95 ±7.84	85.78 ±7.17	84.00 ±7.80	85.05 ±7.10
spambase	90.28 ±3.86	89.59 ±3.97	90.54 ±3.97	90.13 ±3.94
splice	91.17 ±4.63	90.51 ±4.86	91.97 ±4.36	88.37 ±5.12◦
vehicle	73.46 ±3.87	72.70 ±3.43	73.02 ±3.73	71.97 ±3.54
vote	95.70 ±3.01	94.25 ±3.31	96.09 ±2.97	94.50 ±3.29
vowel	89.95 ±3.01	81.72 ±4.20◦	90.19 ±2.70	79.95 ±3.80◦
waveform-5000	86.68 ±5.37	85.92 ±5.41	85.76 ±5.52	84.64 ±5.18
Average	84.66	83.62	84.40	83.19
W/T/L		(7/33/0)	(2/38/0)	(10/30/0)

表 9　EnDWkNN 算法的消融实验的根相对平方误差比较

Table 9　Comparison of root relative rquared error in ablation experiments of EnDWkNN

数据集	EnDWkNN	EnDWkNN_RF	EnDWkNN_SPODE	DWkNN
breast-canner	109.34 ±14.37	102.09 ±12.85·	108.29 ±11.76	95.82 ±7.61·
breast-canner-w	37.26 ±13.21	34.54 ±11.08	37.84 ±13.44	35.71 ±7.29
car	43.47 ±6.49	50.41 ±5.51◦	44.31 ±5.98	67.00 ±1.85◦
colic. ORIG	94.97 ±10.94	92.46 ±11.27	90.47 ±9.05·	88.80 ±5.13·
colic	77.88 ±13.64	76.65 ±11.18	74.88 ±12.76·	78.01 ±7.21
credit-a	69.24 ±9.84	68.77 ±8.66	68.90 ±9.41	67.19 ±5.81
credit-g	100.50 ±6.48	97.35 ±5.37·	97.32 ±6.53·	90.47 ±3.13·
cylinder-bands	74.28 ±10.50	73.20 ±9.48	80.05 ±8.15◦	84.03 ±3.68◦
dermatology	24.59 ±5.78	42.11 ±3.05◦	42.43 ±2.96◦	73.66 ±1.13◦
diabetes	95.71 ±7.69	91.70 ±7.02	92.59 ±7.49	83.34 ±5.54·
eucalyptus	83.15 ±5.64	79.89 ±4.71·	81.16 ±4.19·	84.41 ±1.87
eye_movements	94.04 ±5.10	93.26 ± 3.91	89.82 ±4.20·	91.18 ±2.05·
grub-damage	112.79 ±10.46	102.35 ±9.03·	109.00 ±9.41·	95.50 ±4.39·
haberman	105.96 ±9.35	102.27 ±8.54·	101.89 ±8.88·	99.31 ±7.79·
hayes-roth	80.46 ±17.01	81.67 ±15.01	74.90 ±15.23	73.50 ±7.44
heart-statlog	71.08 ±15.58	72.62 ±13.62	69.19 ±15.34	69.48 ±10.61
hepatitis	80.32 ±28.88	75.05 ±26.02	79.15 ±26.79	74.98 ±18.05
ionosphere	46.17 ±15.67	49.21 ±12.01	48.01 ±14.87	61.69 ±6.15◦
iris	32.97 ±22.87	34.35 ±20.33	32.92 ±20.49	35.70 ±12.71
kr-vs-kp	46.44 ±12.79	59.25 ±11.53◦	42.86 ±9.83	68.89 ±5.22◦
labor	23.07 ±35.20	26.27 ±30.88	29.57 ±32.48	45.42 ±17.86◦
landsat_test	47.15 ±3.07	58.47 ±1.93◦	56.97 ±1.96◦	81.28 ±0.84◦
landsat_train	51.74 ±5.86	65.66 ±3.51◦	64.76 ±3.33◦	88.12 ± 1.11◦
letter	64.21 ±2.08	80.86 ±1.06◦	78.88 ±1.04◦	93.89 ±0.29◦
monks	13.90 ±10.19	38.12 ±9.00◦	19.47 ±12.61	63.25 ±5.00◦
mushroom	4.20 ±6.05	11.55 ±6.04◦	6.89 ±5.02◦	33.33 ±1.91◦
optdigits	66.71 ±2.07	79.82 ±1.12◦	82.02 ±1.19◦	92.87 ±0.43◦
pasture	31.18 ±35.85	34.13 ±30.23	37.66 ±30.70	51.82 ±17.69◦
pendigits	41.41 ±3.60	57.85 ±2.05◦	57.86 ±1.96◦	86.45 ±0.56◦
promoters	50.52 ±14.80	63.10 ±11.20◦	62.11 ±8.56◦	87.35 ±2.75◦
segment	43.10 ±9.68	52.78 ±5.35◦	52.04 ±5.62◦	78.60 ±1.77◦
sick	53.68 ±30.98	57.97 ±26.92	56.25 ±24.72	79.33 ±13.52◦
solar-flare_2	70.38 ±4.04	66.64 ±3.53·	68.19 ± 3.41·	67.13 ±2.07·
sonar	65.95 ±17.28	62.89 ±15.10	66.04 ±15.49	64.27 ±10.58
spambase	56.19 ±12.54	56.98 ±10.93	52.87 ±10.25·	64.58 ±5.69◦
splice	53.49 ±5.75	66.99 ±3.97◦	69.30 ±3.20◦	93.06 ±1.43◦
vehicle	72.40 ±4.59	69.40 ±3.38·	70.51 ±4.09·	73.26 ±1.98
vote	36.45 ±15.19	42.92 ±12.22◦	34.41 ±14.39	45.45 ±7.65◦
vowel	43.90 ±5.30	60.15 ±3.10◦	50.30 ±3.11◦	81.88 ±0.79◦
waveform-5000	55.77 ±9.76	58.25 ±7.50◦	61.00 ±6.23◦	78.26 ±2.26◦
Average	60.65	64.75	63.58	74.21
W/T/L		(16/17/7)	(13/17/10)	(22/10/8)

表 10 EnDWkNN 算法消融实验的分类准确率分值的威尔克森符号秩和检验总体情况

Table 10 Overall wilcoxon signed-rank test results of classification accuracy scores for ablation experiments of EnDWkNN

	EnDWkNN	EnDWkNN _SPODE	EnDWkNN _RF	DWkNN
EnDWkNN	–	·		·
EnDWkNN_SPODE	◦	–	◦	
EnDWkNN_RF			–	·
DWkNN	◦		◦	–

表 11 EnDWkNN 算法的消融实验的 RRSE 分值的威尔克森符号秩和检验总体情况

Table 11 Overall wilcoxon signed-rank test results of RRSE scores for ablation experiments of EnDWkNN

	EnDWkNN	EnDWkNN _SPODE	EnDWkNN _RF	DWkNN
EnDWkNN	–	·	·	·
EnDWkNN_SPODE	◦	–		·
EnDWkNN_RF			–	·
DWkNN	◦	◦	◦	–

图的算法命名为 EnDWkNN_RF. 只使用专家定义的原始属性视图会令 EnDWkNN 退化为 DWkNN. 从表 8 ~ 表 11 提供的实验结果不难看出,相比于其他变体算法,本文提出的 EnDWkNN 算法性能总体仍然是最优的.

现将观察到的实验结果总结如下:

1)在分类准确率方面,EnDWkNN(84.66%)在 40 个数据集上的平均分类准确率最高. EnDWkNN 明显优于 EnDWkNN_SPODE(83.62%)、EnDWkNN_RF(84.40%)、DWkNN(83.19%). EnDWkNN 在 7 个数据集上明显优于 EnDWkNN_SPODE,在 2 个数据集上明显优于 EnDWkNN_RF,在 10 个数据集上明显优于 DWkNN.

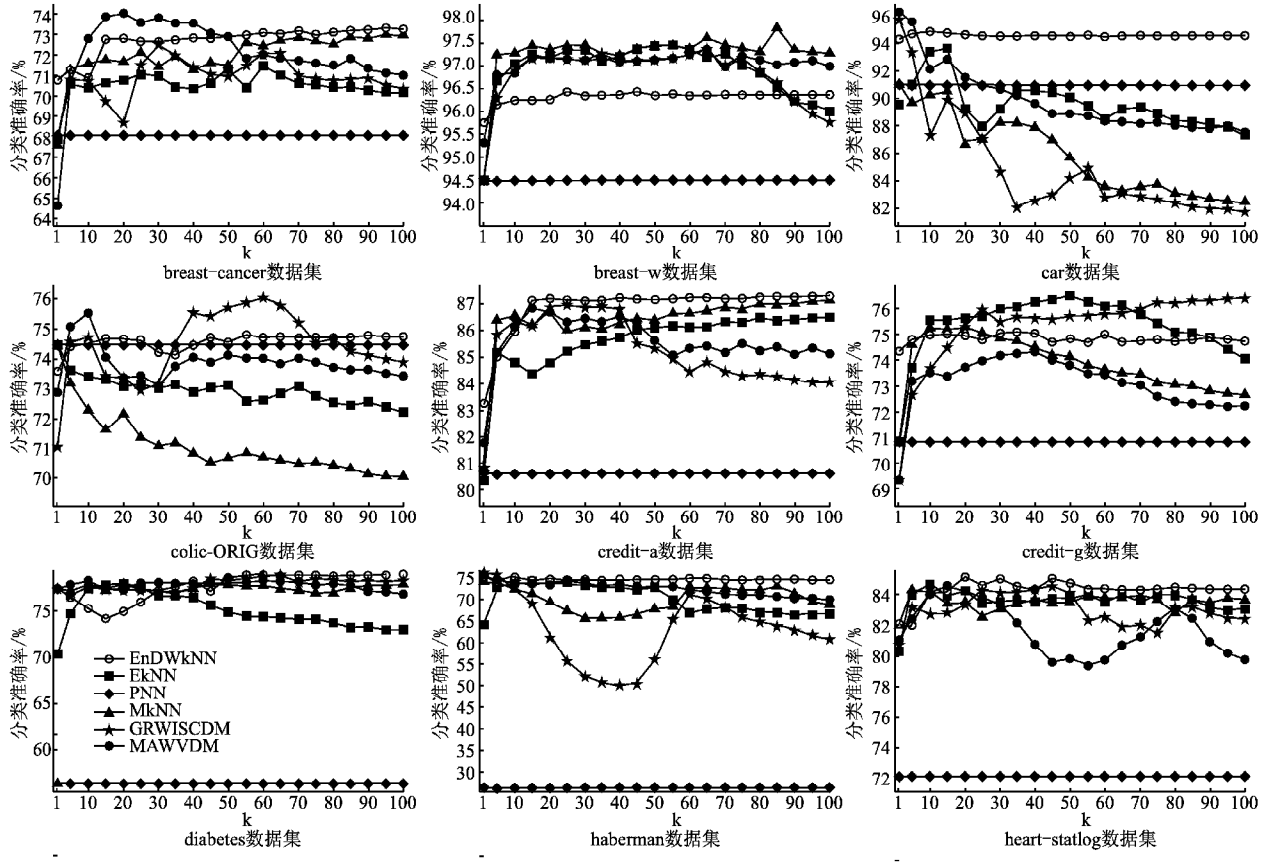
2)在根相对平方误差方面,EnDWkNN(60.65)在 40 个数据集上的根相对平方误差最低. EnDWkNN 明显优于 EnDWkNN_SPODE(64.75)、EnDWkNN_RF(63.58)、DWkNN(74.21). EnDWkNN 在 16 个数据集上明显优于 EnDWkNN_SPODE,在 13 个数据集上明显优于 EnDWkNN_RF,在 22 个数据集上明显优于 DWkNN.

3)根据威尔克森符号秩和检验统计结果,EnDWkNN 在准确率方面要明显优于 EnDWkNN_SPODE、DWkNN. EnDWkNN 在根相对平方误差方面明显优于 EnDWkNN_SPODE、EnDWkNN_RF、DWkNN.

总的来说,表 8 ~ 表 11 的实验结果充分验证了多视图生成组件与基于 D-S 证据理论的多视图融合组件对提升 EnDWkNN 算法性能的必要性.

4.3.2 k 值敏感性分析

考察 k 值变化对改进的 k 近邻算法性能影响是一个关键环节. 通过调整 k 值,可以深入评估改进算法在不同邻近点数量下的稳定性和适应性. 为此,本文进一步考察了所提 EnDWkNN 及其比较对象在上述 40 个 UCI 数据集中性能的变化情况. 由于部分数据集存在类别数量不平衡的问题,PNN 和 MkNN 在 k 取值较大时无法正常运行. 因此,本文实验部分仅展示在 k 取值范围为 1 ~ 100 内能够顺利完成计算的实验结果,并以分类准确率作为主要对比指标进行分析. 如图 4 所示,EnDWkNN 算法在超参数 k 取不同值时,分类准确率保持



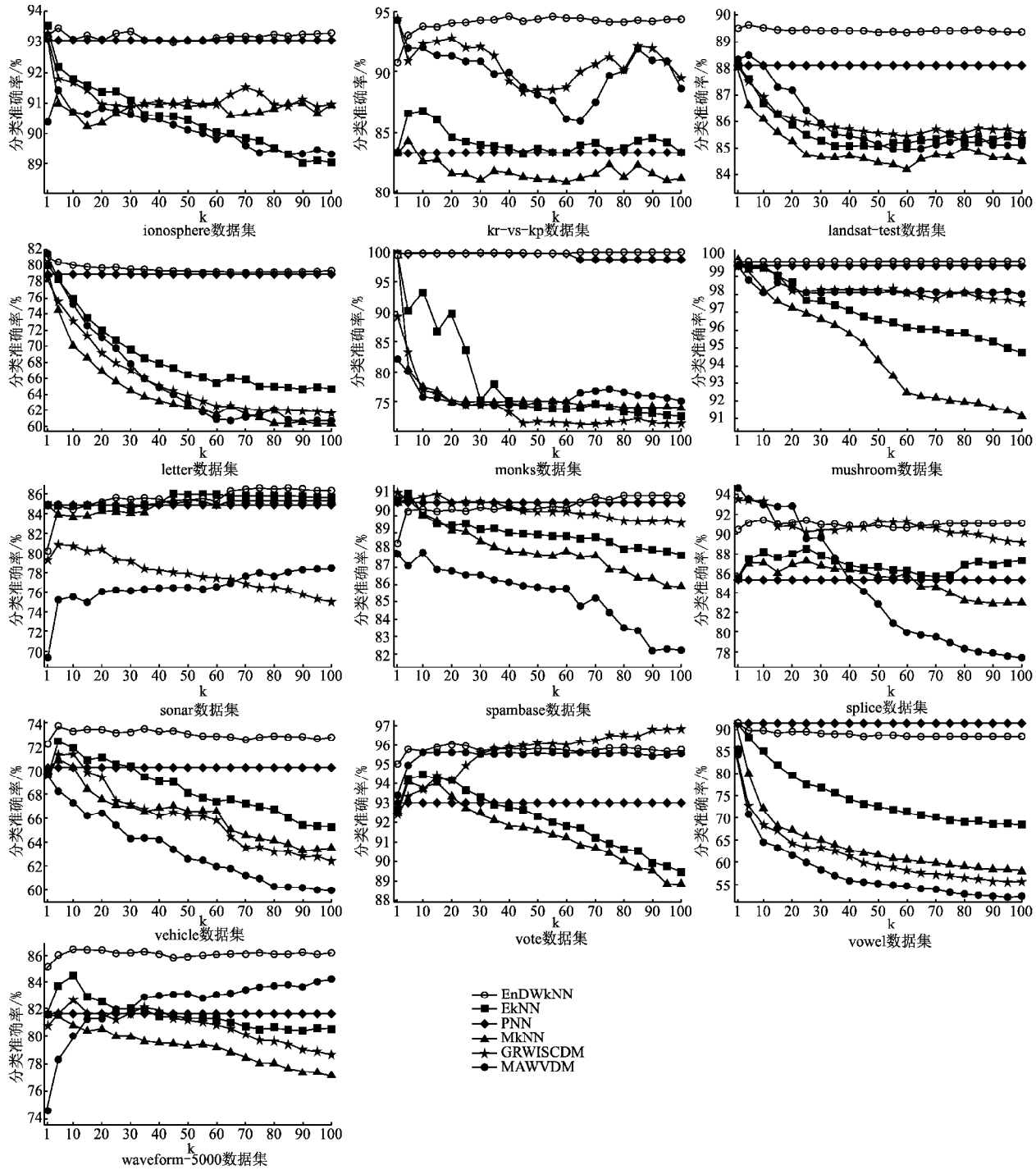


图 4　不同 k 值对算法的分类准确率影响

Fig. 4　Impact of different k values on the classification accuracy of algorithms

在较高水平,起伏更为平稳,表现出显著的稳定性和鲁棒性. 这表明 EnDWkNN 算法在面对近邻实例数量变化时具有较强的适应能力,进一步证明了其在复杂场景下的优势.

5　结　论

本文将多视图生成技术和 D-S 证据理论引入到 kNN 分类中,提出了一种改进的 kNN 分类算法(EnDWkNN). 该算法首先利用原始专家定义的属性视图生成两个标签视图,然后在原始属性视图和新生成的两个标签视图上运行 DWkNN 算法,得到不同视图下测试实例的类概率估计,最后利用 D-S 证据理论整合类概率估计得到测试实例的类标签. 在 40 个 UCI 公开数据集上的实验结果表明,与现有的改进 kNN 算法相比,EnDWkNN 算法无论是在分类准确率方面还是根相对平方误差方面效果都是最好的.

尽管如此,本文所运用的基分类器 DWkNN 采用固定 k 值进行分类决策,导致 EnDWkNN 在真实数据集上的适应性还有待进一步提高. 因此,如何设计灵活有效的自适应算法并将其引入 EnDWkNN 算法是后续研究工作的主要方向. 此

外,如何确定基学习器的个数也是影响集成结果好坏的重要因素,因此,在今后的工作中,后续还将对生成视图的个数的选择问题进行深入研究.

References:

[1] Jin S, Zhang F, Zheng Y, et al. CSKNN: cost-sensitive k-nearest neighbor using hyperspectral imaging for identification of wheat varieties[J]. Computers and Electrical Engineering, 2023, 111: 108896, doi: 10.1016/j.compeleceng.2023.108896.

[2] Zhang J, Li Y, Shen F, et al. Hierarchical text classification with multi-label contrastive learning and KNN[J]. Neurocomputing, 2024, 577: 127323, doi: 10.3390/bdcc7020106.

[3] Nguyen L V, Vo Q T, Nguyen T H. Adaptive KNN-based extended collaborative filtering recommendation services[J]. Big Data and Cognitive Computing, 2023, 7(2): 1-13, doi: 10.3390/bdcc7020106.

[4] Bahrani P, Minaei Bidgoli B, Keshavarz A. A new improved KNN-based recommender system[J]. The Journal of Supercomputing, 2024, 80(1): 800-834.

[5] Guo N, Lin C, Yan H, et al. Real-time pantograph anomaly detection using unsupervised deep learning and K-nearest neighbor classification[J]. IEEE Transactions on Instrumentation and Measurement, 2024, 73: 1-13, doi: 10.1109/tim.2024.330747.

[6] Wagner T. Convergence of the nearest neighbor rule[J]. IEEE Transactions on Information Theory, 1971, 17(5): 566-571.

[7] Uddin S, Haque I, Gide E. Comparative performance analysis of K-nearest neighbour(KNN) algorithm and its different variants for disease prediction[J]. Scientific Reports, 2023, 12(1): 1-11, doi: 10.1038/s41598-022-10358-X.

[8] Halder R K, Uddin M N, Khraisat A. Enhancing K-nearest neighbor algorithm: a comprehensive review and performance analysis of modifications[J]. Journal of Big Data, 2024, 11(1): 1-55.

[9] Shafer G. A mathematical theory of evidence[M]. Princetor: Princeton University Press, 1976.

[10] Kavya R, Jabez C, Subhrakanta P. A new belief interval-based total uncertainty measure for dempster-shafer theory[J]. Information Sciences, 2023, 642: 1-19, doi: 10.1016/j.ins.2023.119150.

[11] Dudani S A. The distance-weighted k-nearest-neighbor rule[J]. IEEE Transactions on Systems, Man, and Cybernetics, 1976, 6(4): 325-327.

[12] Gou J, Du L, Zhang Y, et al. A new distance-weighted k-nearest neighbor classifier[J]. Journal of Information and Computational Science, 2012, 9(6): 1429-1436.

[13] Gou J, Xiong T, Kuang Y. A novel weighted voting for k-nearest neighbor rule[J]. Journal of Computers, 2011, 6(5): 833-840.

[14] Chen J, Li Z. Adaptive condensed fuzzy monotonic K-nearest neighbors for monotonic classification[J]. International Journal of Machine Learning and Cybernetics, 2024, 16(5): 3977-3996.

[15] Zhang Y, Wang C, Huang Y, et al. Adaptive relative fuzzy rough learning for k-nearest classification[J]. IEEE Transactions on Fuzzy Systems, 2024, 32(11): 6267-6276.

[16] Li C, Jiang L, Li H, et al. Attribute weighted value difference metric[C]//Proceedings of 25th International Conference on Tools with Artificial Intelligence, 2013: 575-580.

[17] Jiang L, Li C. Two improved attribute weighting schemes for value difference metric[J]. Knowledge and Information Systems, 2019, 60(2): 949-970.

[18] Li C, Jiang L X, Li H W, et al. Toward value difference metric with attribute weighting[J]. Knowledge and Information Systems, 2017, 50(3): 795-825.

[19] Stanfill C, Waltz D. Toward memory-based reasoning[J]. Communications of the ACM, 1986, 19(12): 1213-1228.

[20] Fang Gong, Xingfeng Guo, Dianhong Wang, et al. Using differential evolution for an attribute-weighted inverted specific-class distance measure for nominal attributes[J]. Data Mining and Knowledge Discovery, 2023, 37(1): 409-433.

[21] El Hindi K. Specific-class distance measures for nominal attributes[J]. AI Communications, 2013, 26(3): 261-279.

[22] Zhang H, Jiang L, Li C. Multi-view attribute weighted naive bayes[J]. IEEE Transactions on Knowledge and Data Engineering, 2023, 35(7): 7291-7302.

[23] Simsek S, Dag A, Coussement K, et al. Decision support framework for misstatement identification in financial reporting: a hybrid tree-augmented Bayesian belief approach[J]. Decision Support Systems, 2025, 189: 114369.

[24] Webb G, Boughton J, Wang Zhihai. Not so naive Bayes: aggregating one-dependence estimators[J]. Machine Learning, 2005, 58(1): 5-24.

[25] Sahami M. Learning limited dependence Bayesian classifiers[C]// Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, 1996: 334-338.

[26] Zhang S, Li X, Zong M, et al. Efficient kNN classification with different numbers of nearest neighbors[J]. IEEE Transactions on Neural Networks and Learning Systems, 2018, 29(5): 1774-1785.

[27] Pan Z, Wang Y, Pan Y. A new locally adaptive k-nearest neighbor algorithm based on discrimination class[J]. Knowledge-Based Systems, 2020, 204: 106185, doi: 10.1016/j.knosys.2020.106185.

[28] Fan Z, Huang Y, Xi C, et al. Multiview adaptive k-nearest neighbor classification[J]. IEEE Transactions on Artificial Intelligence, 2024, 5(3): 1221-1234, doi: 10.1109/access.2024.3392729.

[29] Ali A, Khan Z, Khan D M, et al. An optimal random projection k nearest neighbours nnsemble via extended neighbourhood rule for binary classification[J]. IEEE Access, 2024, 12: 61401-61409.

[30] Keogh E J, Pazzani M J. Learning augmented Bayesian classifiers: a comparison of distribution-based and classification-based approaches[C]//Proceedings of the 7th International Workshop on Artificial Intelligence and Statistics, 1999: 225-230, doi: 10.1016/j.eswa.2023.121549.

[31] Sun Z, Wang G, Li P, et al. An improved random forest based on the classification accuracy and correlation measurement of decision trees[J]. Expert Systems with Applications, 2024, 237: 121549, doi: 10.1016/j.eswa.2023.121549.

[32] Yang Y, Lv H, Chen N. A survey on ensemble learning under the era of deep learning[J]. Artificial Intelligence Review, 2023, 56(6): 5545-5589.

[33] Qiao S, Song B, Fan Y, et al. A fuzzy dempster-shafer evidence theory method with belief divergence for unmanned surface vehicle multi-sensor data fusion[J]. Journal of Marine Science and Engineering, 2023, 11(8): 1-19.

[34] Fei L, Li T, Ding W. Dempster-shafer theory-based information fusion for natural disaster emergency management: a systematic literature review[J]. Information Fusion, 2024, 112: 102585, 10.1016/j.inffus.2024.102585.

[35] Gong F, Wang X, Jiang L, et al. Fine-grained attribute weighted inverted specific-class distance measure for nominal attributes[J]. Information Sciences, 2021, 578: 848-869, doi: 10.1016/j.ins.2021.08.041.

[36] Qamar U, Raza M S. Practical data science with WEKA[M]. Cham: Springer International Publishing, 2023.

[37] Denoeux T. A k-nearest neighbor classification rule based on dempster-shafer theory[J]. IEEE Transactions on Systems, Man, and Cybernetics, 1995, 25(5): 804-813.

[38] Zeng Y, Yang Y, Zhao L. Pseudo nearest neighbor rule for pattern classification[J]. Expert Systems with Applications, 2009, 36(2): 3587-3595.

[39] Parvin H, Alizadeh H, Minaei Bidgoli B. MKNN: modified k-nearest neighbor[C]//Proceedings of the World Congress on Engineering and Computer Science, 2008: 1-4.

[40] Gong F, Jiang L, Zhang H, et al. Gain ratio weighted inverted specific-class distance measure for nominal attributes[J]. International Journal of Machine Learning and Cybernetics, 2020, 11(10): 2237-2246.

[41] Markelle K N, Longjohn R. UCI machine learning repository[EB/OL]. https://archive.ics.uci.edu, 2024.

[42] Zhang W, Jiang L, Li C. ELDP: enhanced label distribution propagation for crowdsourcing[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 47(3): 1850-1862.

[43] Fayyad U, Irani K B. Multi-intervel discretization of continuous-valued attributes for classification learning[C]//13th International Joint Conference on Artificial Intelligence, 1993: 1022-1027.

[44] Xu Y, Zeevi A. Towards optimal problem dependent generalization error bounds in statistical learning theory[J]. Mathematics of Operations Research, 2025, 50(1): 40-67.

[45] Su B, Jiang L, Si S. Confident learning-based noise correction for crowdsourcing[J]. Pattern Recognition, 2025, 169: 111962, doi: 10.1016/j.patcog.2025.111962.

[46] Vierra A, Razzaq A, Andreadis A. Continuous variable analyses: t-test, mann-whitney, Wilcoxon sign rank[M]. Cambridge: Academic Press, 2023.

[47] Rainio O, Teuho J, Klén R. Evaluation metrics and statistical tests for machine learning[J]. Scientific Reports, 2024, 14(1): 1-14.