基于LLaMa3和Choquet积分的最优相似度选择集成学习方法

付超，余良菊，常文军

引用本文

付超，余良菊，常文军. 基于LLaMa3和Choquet积分的最优相似度选择集成学习方法[J]. 计算机科学, 2025, 52(9): 80-87.

FU Chao, YU Liangju, CHANG Wenjun. Selective Ensemble Learning Method for Optimal Similarity Based on LLaMa3 and Choquet Integrals [J]. Computer Science, 2025, 52(9): 80-87.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.250100150

基于 LLaMa3 和 Choquet 积分的最优相似度选择集成学习方法

付 超 余良菊 常文军
合肥工业大学管理学院 合肥 230009
(chaofu@hfut.edu.cn)

摘 要 为了在多分类器集成过程中筛选和赋权存在相关关系的基分类器,提出了一种基于 LLaMa3 和 Choquet 积分的最优相似度选择集成方法(LCOS-SELM)。首先在开源大模型 LLaMa3 的基础上,通过少量标注样本数据进行提示词学习,以实现非结构化文本的关键特征提取。然后,通过 Choquet 积分融合存在相关关系的分类器预测结果,并评估其相关关系以优化分类器选择。最后,采用最优相似度策略学习分类器权重,在确保样本一致性的同时,提升集成方法的性能。将 LCOS-SELM 用于克罗恩病的辅助诊断,以合肥市某三甲医院的 297 份检查报告为基础进行实验,通过与内镜检查报告进行比对,验证了所提方法的有效性。在相同实验条件下将 LCOS-SELM 与单分类器和传统集成模型进行实验对比,结果显示:在相同实验条件下,与单分类器相比,LCOS-SELM 在 Acc,F1 和 AUC 指标上均提升了约 8%;与传统集成模型相比,LCOS-SELM 在 3 个指标上均提升了约 2%,进一步验证了其性能优势。

关键词:选择集成学习;LLaMa3;Choquet 积分;权重学习;相似案例学习
中图分类号 TP391.9

Selective Ensemble Learning Method for Optimal Similarity Based on LLaMa3 and Choquet Integrals

FU Chao, YU Liangju and CHANG Wenjun
School of Management, Hefei University of Technology, Hefei 230009, China

Abstract To screen and weight base classifiers with correlation during the ensemble process of multiple classifiers, this paper proposes a selective ensemble learning method for optimal similarity based on LLaMa3 and Choquet integrals(LCOS-SELM). Leveraging the open-source LLaMa3, this method efficiently achieves key feature extraction from unstructured text through prompt learning, using only a small amount of labeled data. Subsequently, it integrates the prediction results of classifiers with correlated relationships using the Choquet integral, evaluating their correlation to optimize classifier selection. Finally, it employs an optimal similarity strategy to learn classifier weights, ensuring sample consistency while enhancing the performance of the ensemble method. The LCOS-SELM is used for the auxiliary diagnosis of Crohn's disease, using 297 examination reports from a tertiary hospital in Hefei. Experiments are conducted by comparing it with endoscopic examination reports to validate the effectiveness of the proposed method. Under the same experimental conditions, LCOS-SELM demonstrates an approximate 8% improvement in Accuracy (Acc), F1-score, and AUC compared to the single classifier, and an approximate 2% improvement in all three metrics compared to traditional ensemble models, further validating its performance advantage.

Keywords Selective ensemble learning, LLaMa3, Choquet integral, Weight learning, Similar case learning

1 引言

在机器学习领域,选择集成学习作为一种重要的学习算法,其核心思想在于从已有的基分类器中筛选出性能优良的分类器子集,并在此之上构建一个集成分类器,为特定任务提供性能优良的预测服务和决策支持。选择集成学习算法已在多标签分类、图像识别、情感识别等诸多实际场景中得到了广泛应用[1]。

目前,比较流行的选择集成学习算法根据基分类器筛选标准的不同,可分为基于性能的方法和基于多样性的方法[2]。基于性能的方法是根据基分类器在给定的性能指标上的表现对基分类器进行筛选。例如,Ozcift 等[3]提出使用分类准确率、Kappa 误差和接收者操作特征曲线下面积(AUC)3 个性能指标对基分类器进行筛选,基于筛选的基分类器构建的集成模型的平均分类准确率提升了 2.32%。基于多样性的方法则是去除相似性较高的基分类器,从而提升集成预测性能。

到稿日期:2025-01-24 返修日期:2025-05-13
基金项目:国家自然科学基金(72171066,72471073)
This work was supported by the National Natural Science Foundation of China(72171066,72471073).
通信作者:常文军(changwenjun@hfut.edu.cn)

例如，对基分类器进行分簇，通过簇内选择来剔除冗余分类器[4]。然而，随着应用场景的不断拓展，这些方法也面临着一些新的问题与挑战，如模型集成过程中基于相关性协调分析的基分类器筛选难题和权重分配困境等[5]。

针对基于相关性协调分析的基分类器筛选问题，Corriveau等[6]以模糊 ARTMAP 神经网络分类器为研究对象，利用粒子群优化策略和特定实验协议，发现随着搜索空间中基因型多样性的线性增加，决策空间中的分类器多样性也呈现出相同的变化趋势，这有助于协调基分类器相关性以更高效地筛选和组合基分类器。Singh 等[7]提出基于相关性的分类器组合技术，通过计算斯皮尔曼相关系数衡量基分类器之间的相关性，再通过 Fisher 检验筛选出具有显著相关性的分类器对进入后续组合流程，其在不同数据集和交叉验证方案下均能显著提高分类准确率。这些方法侧重于对基分类器预测结果进行线性相关分析，难以捕捉基分类器之间的非线性相关关系，不适用于基分类器间存在非线性相关的场景。

针对基分类器的权重分配问题，Kuncheva[8]提出基于准确率的权重分配方法，即根据基分类器在训练集上的准确率确定其权重；Nanni 等[9]提出基于特征选择的权重分配方法，通过评估不同特征对分类结果的贡献度，为筛选出的重要特征对应的基分类器赋予更高权重；Zeng 等[10]提出基于加权准确性和多样性的权重分配方法，通过计算准确性和多样性的调和值来综合评估基分类器，进而优化基分类器的权重分配过程，其在多数情况下优于仅考虑单指标进行权重分配的传统方法；Zou 等[11]提出多分类问题中的 Fisher 一致损失条件，通过采用梯度下降法最小化基于边际向量的经验损失来确定基分类器权重，为基分类器权重分配提供了有效的损失函数选择。这些方法虽然在一定程度上提高了集成模型的性能，但在保证样本一致性的前提下，如何对基分类器合理赋权以提升集成模型的预测准确性依然面临挑战。

针对上述问题，本文提出了一种基于 LLaMa3 和 Choquet 积分的最优相似度选择集成方法（LCOS-SELM）。首先基于大语言模型 LLaMa3 对非结构化文本进行信息提取；然后，采用 Choquet 积分融合基分类器以识别基分类器间的相关关系，进而优化分类器筛选过程；最后，通过最优相似度策略学习历史样本的基分类器权重，在保证样本一致性的前提下，对存在相关关系的基分类器进行合理赋权以最大化集成模型的预测精度。将 LCOS-SELM 用于克罗恩病（Crohn's disease，CD）的辅助诊断，以合肥市某三甲医院 2021 年 1 月至 2023 年 12 月间的 297 份检查报告为基础进行实验，通过与内镜检查报告进行比对，验证了所提方法的有效性。在相同实验条件下，以单分类器和传统集成模型为基准进行对比实验，结果表明：相比单分类器，所提方法在 Acc，F_1 和 AUC 这 3 个指标上均提升了约 8%；相比传统集成模型，3 个指标均提升了约 2%，进一步验证了 LCOS-SELM 在克罗恩病活动性预测方面的性能优势。

2　理论基础

本章介绍 LCOS-SELM 方法的理论基础，主要包括以下 3 个部分：大语言模型 LLaMa3 的基本原理、Choquet 积分的数学基础，以及选择集成学习的理论基础。

2.1　LLaMa3

LLaMa3（Large Language Model Meta AI 3）是由 Meta 开发的大语言模型，无需大量标注数据就能够从大量非结构化文本中提取关键信息，以支持数据分析、知识获取和决策制定[12]。与 LLaMa3 等大语言模型交互时，通过输入“prompt”引导模型生成特定类型的响应。prompt 分为开放式和封闭式，前者应对宽泛问题，而后者应对特定问题。

一个完整的封闭式 prompt 通常包括 5 个组成部分：指示、上下文、例子、输入和输出。在指示部分可以通过预先规定相应的属性或关键特征，来输出与特定内容相关的文本描述；上下文部分提供背景信息；例子部分用于示范学习；输入和输出部分分别规定待识别的数据格式和其识别结果的格式（如 Excel、JSON、HTML 或指定的其他格式）。典型的 Few-shot prompting 方法通过向模型提供少量（通常为 1 到 10 个）示例，引导其生成响应或执行特定任务，从而增强模型输出的可控性[13]，尤其适合少样本任务。

2.2　Choquet 积分的基础

Choquet 积分是一种处理具有交互作用决策问题的有效方法[14]。

1）λ-模糊测度

定义 1[15]　设 $X=\{x_1,x_2,\cdots,x_n\}$ 为一非空集合，$P(X)$ 为 X 的幂集。给定 $\lambda\in(-1,\infty)$，定义在 X 上的 λ-模糊测度为 $\mu:P(X)\to[0,1]$，满足：

（1）边际条件，$\mu(\varnothing)=0$，$\mu(X)=1$；

（2）单调性，$\forall A,B\in P(X)$，若 $A\subseteq B$，则 $\mu(A)\leqslant\mu(B)$；

（3）$\forall A,B\in P(X)$，$A\cap B=\varnothing$，则 $\mu(A\cup B)=\mu(A)+\mu(B)+\lambda\mu(A)\mu(B)$。

①若 $\lambda=0$，则 μ 为 X 上的可加测度，表示属性集合 A 与 B 无相互关联关系；

②若 $\lambda<0$，则 μ 为 X 上的次可加测度，表示属性集合间存在信息冗余；

③若 $\lambda>0$，则 μ 为 X 上的超可加测度，表示属性集合间存在信息互补。

若非空集 $X=\{x_1,x_2,\cdots,x_n\}$ 是有限的，则 λ-模糊测度 μ 计算为：

$$\mu(A)=\begin{cases}\dfrac{1}{\lambda}\left(\prod\limits_{x_i\in A}\lambda\mu(x_i)\right)-1), & \lambda\neq0\\ \sum\limits_{x_i\in A}\mu(x_i), & \lambda=0\end{cases}\tag{1}$$

根据边际条件 $\mu(X)=1$，式（1）可转化为：

$$\lambda+1=\prod_{i=1}^{n}(1+\lambda\mu(x_i))\tag{2}$$

依此确定唯一的参数 λ。

2）离散 Choquet 积分

定义 2[16]　设 f 为定义在有限集合 $X=\{x_1,x_2,\cdots,x_n\}$ 上的非负实值函数，μ 是定义在 X 上的 λ-模糊测度，则 f 关于模糊测度 μ 的离散 Choquet 积分定义为：

$$C_\mu(f)=\sum_{i=1}^{n}f(x_{(i)})(\mu(A_{(i)})-\mu(A_{(i+1)}))\tag{3}$$

其中，$(\cdot)$ 代表 $f(x_{(i)})$ 的排列，满足 $f(x_{(1)})\leqslant f(x_{(2)})\leqslant\cdots\leqslant f(x_{(n)})$，且 $A_{(i)}=\{x_1,x_2,\cdots,x_n\}$，$A_{(n+1)}=\varnothing$。

2.3 选择集成学习的基础

集成学习是一种通过组合多个弱分类器构建强分类器的方法，旨在提高模型的泛化能力和预测准确性[1]。选择集成学习作为一种特殊的集成方法，通过最优选择策略挑选性能优良的基分类器来构建集成分类器，能够更有效地提升集成效率和性能[1]。

在使用选择集成学习方法提升集成效率的同时，也要考虑选择的基分类器间的多样性，以确保集成性能。基于此，一般从表现良好的原始基分类器中选择[1/3，1/2]的基分类器来构建最终的集成分类器，以在兼顾多样性的同时避免过拟合[2]。以下两类方法常用于考虑多样性的基分类器选择。1）基于多样性度量的方法[17]：利用基分类器的不一致性、Q-统计量、误差比例等多样性度量指标，来估计基分类器之间预测行为的差异。2）可视化分析方法[2]：将基分类器投影到分类器投影空间，空间中的每个点代表一个模型，点之间的距离表示不同基分类器预测行为之间的差异。本文采用第一种方法，其能更有效地通过量化分析评估基分类器的多样性。

3 LCOS-SELM

LCOS-SELM 主要包括数据处理、基分类器筛选、最优相似度集成学习 3 个过程，整体框架如图 1 所示。

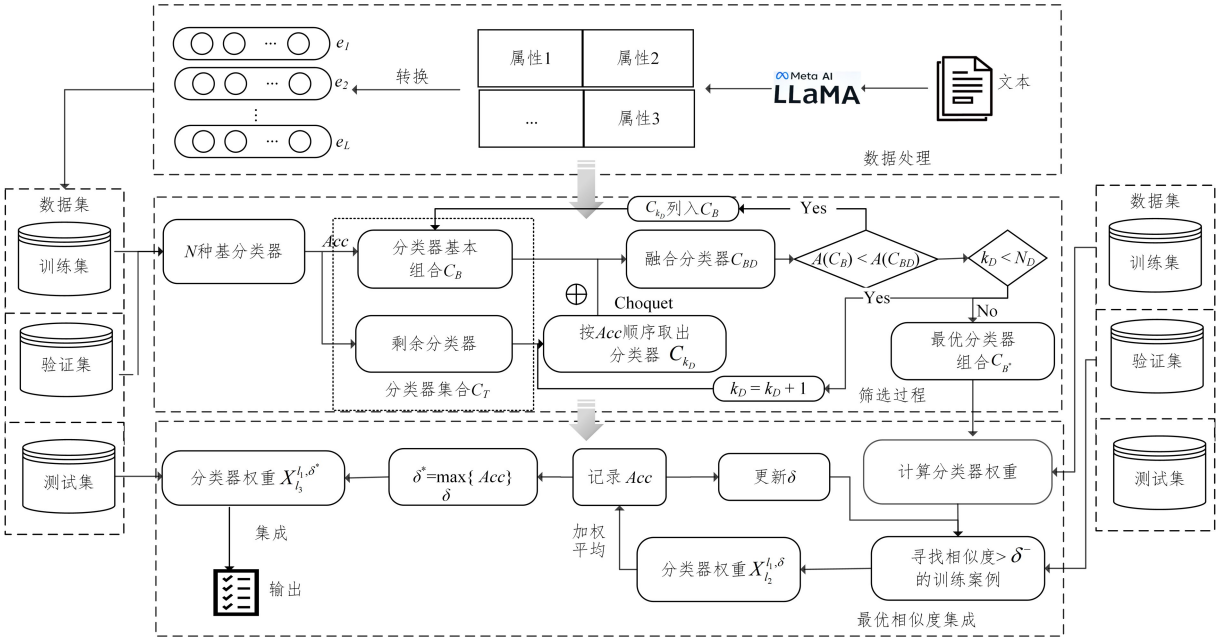

图 1 LCOS-SELM 的框架

Fig. 1 Framework of LCOS-SELM

3.1 数据处理

考虑非结构化文本的处理效率，设计基于 LLaMa3 模型的数据处理流程以生成数据集。

步骤 1 设计 prompt，提取文本中的关键信息。收集需要分析的非结构化文本，将经过预处理的文本采用以下标准格式输入 LLaMa3 模型进行训练[18]。

$$Input=[\mathrm{CLS}]+Token_1+Token_2+\cdots+Token_n+[\mathrm{SEP}] \tag{4}$$

其中，[CLS]为模型的分类标记，[SEP]为句子分隔符，$Token_i$ 为文本中的每个词。

如 2.1 节所述，在指示中给定需要提炼信息的属性，补充背景，提供提取信息的示例，规定输出格式，由此形成 prompt。设计一个迭代过程，使用户能通过 prompt 与 LLaMa3 模型进行交互，直至提取出符合用户要求的属性信息或有效属性信息不再增加时停止迭代。

步骤 2 将属性信息转化为向量。考虑属性信息通常不是连续值且取值之间没有大小关系，因此采用独热编码将文本数据转化为向量。首先，根据属性描述的特点进行类别划分，将分类值映射到整数值。然后，将每个整数值转换为独立的二进制向量，以确保模型正确理解非数值分类的特征，避免误判数值关系[19]。

步骤 3 形成数据集。将步骤 2 中生成的数据收集并整合为特征数据集 X，并按照时间升序排序，而后将 X 划分为训练集 $X_{l_1}=\{x_{l_1}=\{x_{l_1}^i,i=1,\cdots,L\},y_{l_1}\}(l_1=1,\cdots,N_1)$、验证集 $X_{l_2}=\{x_{l_2}=\{x_{l_2}^i,i=1,\cdots,L\},y_{l_2}\}(l_2=1,\cdots,N_2)$ 和测试集 $X_{l_3}=\{x_{l_3}=\{x_{l_3}^i,i=1,\cdots,L\},y_{l_3}\}(l_3=1,\cdots,N_3)$。设数据集中不同类别数目为 m。

3.2 基分类器筛选

基分类器可能由相同数据源训练产生或源自相同分类方法，存在一定的相关性。考虑基分类器间的相关关系，设计基于 Choquet 积分的分类器筛选过程，以得到互补性最高的基分类器组合[20]。

步骤 1 确定基分类器组合策略。使用训练集 X_{l_1} 训练待筛选的 N 个基分类器 $C=(C_1,\cdots,C_n)$，获得其在验证集 X_{l_2} 上的准确率。

$$A(C_k)=\frac{TP+TN}{TP+TN+FP+FN} \tag{5}$$

其中，TP(真正例)表示正类样本被正确分类为正类的数量，TN(真负例)表示负类样本被正确分类为负类的数量，FP(假正例)表示负类样本被错误分类为正类的数量，FN(假负例)

表示正类样本被错误分类为负类的数量。

为了进一步平衡集成性能和泛化能力，将构建集成分类器的基分类器集合 $C_T(|C_T|\in[N/3,N/2])$ 划分为基本组合 C_B 和多样化组合 C_D 两部分，它们包含基分类器数目的比例推荐为 2∶1。具体而言，基本组合 C_B 中包含 C_T 中性能最优的前 2/3 个基分类器，多样化组合 C_D 中则为 C_T 中剩余的基分类器。从 C_D 中依次选取性能最优的基分类器与 C_B 进行组合，通过集成性能的变化来更新 C_B。设 $|C_B|=N_B$，$|C_D|=N_D$，$k_D=1$。

步骤 2　计算基分类器 Choquet 集结的自适应模糊测度。考虑 C_B 中基分类器 $C_{k_B}(k_B=1,\cdots,N_B)$ 和 C_D 中待组合基分类器 C_{k_D} 之间的相关性，利用 Choquet 积分对它们进行集结以形成集成分类器。

利用基分类器 C_{k_B} 和 C_{k_D} 对验证集 X_{l_2} 进行分类，正确分类的概率为：

$$P^{jj}_{l_2,k_B}=P(C_{k_B}(x^i_{l_2})=j\,|\,\{x^i_{l_2},y_{l_2}=j\}\in X_{l_2})=\frac{n^{jj}_{k_B}}{\sum_{d=1}^{m}n^{jd}_{k_B}}\tag{6}$$

$$P^{jj}_{l_2,k_D}=P(C_{k_D}(x^i_{l_2})=j\,|\,\{x^i_{l_2},y_{l_2}=j\}\in X_{l_2})=\frac{n^{jj}_{k_D}}{\sum_{d=1}^{m}n^{jd}_{k_D}}\tag{7}$$

基于 $\gamma^{i_Bj_B}_{l_2}(\gamma^{i_Dj_D}_{l_2})$，$C_{k_B}(C_{k_D})$ 内所有基分类器对验证集 X_{l_2} 的模糊测度一致性校正系数 $\beta_{k_B}(\beta_{k_D})$ 为：

$$\beta_{k_B}=\frac{\sum_{l_2=1}^{N_2}\left(\prod_{i_B,j_B=1}^{N_B}\gamma^{i_Bj_B}_{l_2}\right)}{N_2}\tag{12}$$

$$\beta_{k_D}=\frac{\sum_{l_2=1}^{N_2}\left(\prod_{i_D,j_D=1}^{N_D}\gamma^{i_Dj_D}_{l_2}\right)}{N_2}\tag{13}$$

为了平衡基分类器间的性能差异，综合考虑模糊测度的置信度校正系数和一致性校正系数，定义自适应模糊测度为[16]：

$$\bar{\mu}_{k_B}=\alpha_{k_B}\cdot\beta_{k_B}\cdot\mu_{k_B}\quad(\bar{\mu}_{k_D}=\alpha_{k_D}\cdot\beta_{k_D}\cdot\mu_{k_D})\tag{14}$$

步骤 3　计算当前两种分类器组合的预测精度。首先，基于自适应模糊测度 $\bar{\mu}_{k_B}(\bar{\mu}_{k_D})$，依据式(2)，确定参数 λ；然后，根据式(3)计算关于模糊测度 $\bar{\mu}_{k_B}(k_B=1,\cdots,N_B)$ 的离散 Choquet 积分 $C_{\bar{\mu}_{k_B}}$，得到基分类器 $C_{k_B}(k_B=1,\cdots,N_B)$ 融合的预测类别；最后，依据式(5)得到在验证集 X_{l_2} 上的预测精度 $A(C_B)$。类似地，可以得到基于模糊测度 $\bar{\mu}_{k_B}(k_B=1,\cdots,N_B)$ 和 $\bar{\mu}_{k_D}$ 在验证集 X_{l_2} 上的预测精度 $A(C_{BD})$。

步骤 4　确定最佳基分类器组合。当 $A(C_B)<A(C_{BD})$ 时，将 C_D 中的基分类器 C_{k_D} 列入 C_B 中。设 $k_D=k_D+1$，若 $k_D<N_D$，则转到步骤 2；否则，得到最佳基分类器组合 $C_{B^*}(k_{B^*}=1,\cdots,N_{B^*}=|C_{B^*}|)$。

其中，$n^{jj}_{k_B}(n^{jj}_{k_D})$ 表示验证集中真实标签和基分类器 $C_{k_B}(C_{k_D})$ 的预测结果同为第 j 类的样本数量，$n^{jd}_{k_B}(n^{jd}_{k_D})$ 表示验证集中真实标签为第 j 类且基分类器 $C_{k_B}(C_{k_D})$ 的预测结果为第 d 类的样本数量。基于此，融合基分类器 $C_{k_B}(k_B=1,\cdots,N_B)$ 和 C_{k_D} 的预测结果的 Choquet 积分的模糊测度定义为 $\mu_{k_B}=\sum_{j=1}^{m}P^{jj}_{k_B}/m$ 和 $\mu_{k_D}=\sum_{j=1}^{m}P^{jj}_{k_D}/m$[21]。

考虑不同基分类器之间的性能差异，对以上模糊测度进行自适应修正。设利用基分类器 $C_{k_B}(C_{k_D})$ 对验证集 X_{l_2} 进行分类所得的预测分布为 $P^d_{l_2,k_B}(P^d_{l_2,k_D})(d=1,\cdots,m)$，则模糊测度的置信度校正系数定义为：

$$\alpha_{k_B}=\frac{\sum_{l_2=1}^{N_2}\left(\max_{d\in\{1,\cdots,m\}}\{P^d_{l_2,k_B}\}-P^d_{l_2,k_B}\right)}{N_2}\tag{8}$$

$$\alpha_{k_D}=\frac{\sum_{l_2=1}^{N_2}\left(\max_{d\in\{1,\cdots,m\}}\{P^d_{l_2,k_D}\}-P^d_{l_2,k_D}\right)}{N_2}\tag{9}$$

设 $C_{k_B}(C_{k_D})$ 内基分类器两两之间对验证集 X_{l_2} 的模糊测度一致性程度为 γ^{ij}，其中 $C_{i_B}(C_{i_D})$ 与 $C_{j_B}(C_{j_D})((i_B(i_D),j_B(j_D)=1,2,\cdots,N_B(N_D))$ 表示不同的基分类器，$C_{i_B}(C_{i_D})$ 与 $C_{j_B}(C_{j_D})$ 的输出类别分别为 $d_{i_B}(d_{i_D})$ 和 $d_{j_B}(d_{j_D})$，则 $\gamma^{i_Bj_B}_{l_2}(\gamma^{i_Dj_D}_{l_2})$ 为：

$$\gamma^{i_Bj_B}_{l_2}=\begin{cases}1-\left(\max\limits_{d_{i_B}\in\{1,\cdots,m\}}\{P^{d_{i_B}}_{l_2,i_B}\}-\max\limits_{d_{j_B}\in\{1,\cdots,m\}}\{P^{d_{j_B}}_{l_2,i_B}\}\right), & d_{i_B}\neq d_{j_B}\\ 1, & d_{i_B}=d_{j_B}\end{cases},l_2=1,\cdots,N_2\tag{10}$$

$$\gamma^{i_Dj_D}_{l_2}=\begin{cases}1-\left(\max\limits_{d_{i_D}\in\{1,\cdots,m\}}\{P^{d_{i_D}}_{l_2,i_D}\}-\max\limits_{d_{j_D}\in\{1,\cdots,m\}}\{P^{d_{j_D}}_{j_2,j_D}\}\right), & d_{i_D}\neq d_{j_D}\\ 1, & d_{j_D}=d_{j_D}\end{cases},l_2=1,\cdots,N_2\tag{11}$$

3.3　最优相似度集成学习

针对最佳基分类器组合 C_{B^*}，设计迭代流程确定训练集 X_{l_1} 和验证集 X_{l_2} 之间的最优相似度阈值，在确保训练样本和验证样本一致性的前提下最大化 X_{l_2} 上的预测精度，进而依据最优相似度阈值计算各基分类器的集成权重。

步骤 1　初始化相似度阈值。给定相似度阈值范围 $[\delta^-,1]$，设 $\delta=\delta^-$。

步骤 2　计算训练样本的基分类器权重。设 C_{B^*} 中各基分类器对训练样本 $\{x_{l_1},y_{l_1}\}$ 的预测分布为 $P^d_{l_1,k_{B^*}}(d=1,\cdots,m)$，则各基分类器的集成权重为：

$$w_{l_1,k_{B^*}}=\frac{\max\limits_{d\in\{1,\cdots,m\}}\{P^d_{l_1,k_{B^*}}\}}{\sum_{k_{B^*}=1}^{N_{B^*}}\max\limits_{d\in\{1,\cdots,m\}}\{P^d_{l_1,k_{B^*}}\}},k_{B^*}=1,\cdots,N_{B^*}\tag{15}$$

步骤 3　计算 C_{B^*} 在验证集上的集成预测准确率。给定验证样本 $\{x_{l_2},y_{l_2}\}$，计算它与训练样本 $\{x_{l_1},y_{l_1}\}$ 间的相似度：

$$S^{l_1}_{l_2}=1-\frac{\sqrt{\sum_{i=1}^{L}(x^i_{l_1}-x^i_{l_2})^2}}{\sqrt{2}}\tag{16}$$

显然，$S^{l_1}_{l_2}\in[0,1]$。依据 $S^{l_1}_{l_2}(l_1=1,\cdots,N_1)$ 和 δ，确定验证样本 $\{x_{l_2},y_{l_2}\}$ 的相似训练样本集为 $X^{l_1\cdot\delta}_{l_2}=\{\{x_{l_1},y_{l_1}\}\,|\,S^{l_1}_{l_2}\geqslant\delta,l_1=1,\cdots,N_1\}$。

基于 $S^{l_1}_{l_2}(l_1=1,\cdots,N_1)$，综合形成各基分类器对验证样

本$\{x_{l_2}, y_{l_2}\}$的集成权重：

$$w_{l_2}=\frac{\sum_{l_1\in X_{l_2}^{l_1,\delta}} S_{l_2}^{l_1}\cdot w_{l_1,k_{B^*}}}{\sum_{l_1\in X_{l_2}^{l_1,\delta}} S_{l_2}^{l_1}},l_2=1,\cdots,N_2 \tag{17}$$

显然，对于$w_{l_2}=(w_{l_2}^1, w_{l_2}^2,\cdots, w_{l_2}^{N_{B^*}})$，有$\sum_{i=1}^{N_{B^*}} w_{l_2}^i=1$。基于$w_{l_2}$和$C_{B^*}$对验证集$X_{l_2}$的预测输出，加权集成$C_{B^*}$的输出结果，据式(5)计算得到在验证集上的准确率$Acc_2^{\delta}$。

步骤4 迭代搜寻最优相似度阈值。设置$\delta=\delta+\varepsilon$，其中$\varepsilon$为步长(如0.01)，在每次迭代中更新$\delta$。返回步骤3，直到$\delta$等于1。根据$Acc_2^{\delta}(\delta\in[\delta^-,1])$得到最优相似度阈值$\delta^*=\arg\max\{Acc_2^{\delta},\delta\in[\delta^-,1]\}$。

步骤5 计算基于最优相似度阈值的测试精度。类似地，根据式(16)计算对于每个测试样本$\{x_{l_3}=\{x_{l_3}^i, i=1,\cdots,L\}, y_{l_3}\}$和训练样本之间的相似度$S_{l_3}^{l_1}$；识别相似训练样本集为$X_{l_3}^{l_1,\delta^*}=\{X_{l_1}\mid S_{l_3}^{l_1}\geqslant\delta^*, l_1=1,\cdots,n_1\}$。类似地，使用$X_{l_3}^{l_1,\delta^*}$与$S_{l_3}^{l_1}$，计算$X_{l_3}$每个样本对应的基分类器权重：

$$w_{l_3}=\frac{\sum_{l_1\in X_{l_3}^{l_1,\delta^*}} S_{l_3}^{l_1}\cdot w_{l_1,k_{B^*}}}{\sum_{l_1\in X_{l_3}^{l_1,\delta^*}} S_{l_3}^{l_1}},l_3=1,\cdots,N_3 \tag{18}$$

基于w_{l_3}和C_{B^*}对测试集X_{l_3}的预测输出，加权集成C_{B^*}的输出结果，据式(5)计算得到阈值为δ^*时，在测试集上的预测准确率$Acc_3^{\delta^*}$。

4 实验与分析

利用所提LCOS-SELM进行克罗恩病活动性预测，验证该方法的应用性和有效性。

4.1 数据描述

1)案例背景

克罗恩病是一种炎症性肠病，主要表现为活动期与缓解期交替出现。超声检查为放射科医生提供了一种有效的CD活动性诊断手段，而内镜检查则被视为诊断的金标准[22]。我们与中国安徽合肥一家三甲医院的放射科合作，根据医生的建议以及现有研究结论[23-25]，选取了9个关键属性(用e_i表示，其中$i=1,\cdots,9$)，包括累及肠段、肠壁、血流、弹性、局部回声、网膜、肠壁层次结构、肠管和腹腔。为提高模型运算效率，利用概率分配法[26]，将评估等级定义为$\Omega=\{H_1, H_2, H_3\}=\{[0,1/3],[1/3,2/3],[2/3,1]\}$，不同等级表示所检测属性的恶性程度，从$H_1$到$H_3$的分级表示恶性程度递增。各属性分级评估标准如表1所列。根据临床诊断需求，CD活动性被分为3个显著的活动期，CD活动性分级评估标准如表2所列。通过对检查报告和病理结果进行综合分析，根据部分或全部属性特征，能够全面评估患者的CD总体活动性。

表1 CD活动性属性评估

Table 1 CD activity attribute evaluation

属性	诊断		
	0	1	2
e_1(累及肠段)	局限于直肠，未达乙状结肠	累及左半结肠(脾曲以远)	累及全消化道，包括上消化道、小肠、结肠
e_2(肠壁)	厚度小于3 mm	厚度大于等于3 mm且小于7 mm	厚度大于等于7 mm
e_3(血流)	无血流信号，血流信号未见增多	少量、点条状血流信号，血流信号增多，血流信号较丰富	血流信号丰富
e_4(弹性)	大于等于4 kpa且小于10.5 kpa	大于等于10.5 kpa且小于22.5 kpa	大于等于22.5 kpa
e_5(局部回声)	无异常	回声不均匀、回声降低、回声增强	—
e_6(网膜)	无增厚	小于10 mm	大于等于10 mm
e_7(肠壁层次结构)	无异常	肌层，黏膜层，粘膜层增厚	—
e_8(肠管)	无异常	扩张、变窄	—
e_9(腹腔)	无异常	积液或淋巴结增大	积液且淋巴结肿大

表2 CD活动性评估

Table 2 CD activity evaluation

	诊断		
	0 (临床缓解期)	1 (中度活动期)	2 (重度活动期)
症状	肠道壁轻微增厚，黏膜纹理增加且光滑，无红肿或糜烂，偶见轻度充血，活动性炎症少，黏膜基本正常	肠道壁明显增厚，伴有溃疡和炎症灶，部分肠腔狭窄，黏膜轻度红肿和充血，存在浅表性溃疡或糜烂	肠道壁广泛增厚，严重狭窄或梗阻，可能形成瘘管，黏膜明显红肿、充血和糜烂，观察到较大深度溃疡，病变区域显著

2)文本数据转换

实验收集了中国安徽省合肥市某三甲医院2021年1月至2023年12月期间的297份原始CD病例报告。根据表1中所列的CD活动性分级评估标准，利用LLaMa3模型对297份原始CD病例报告进行关键特征提取。依据医生的专业知识和经验，其中4个属性(累及肠段、肠壁、血流、弹性)为克罗恩病的主要病征，对疾病的诊断具有显著影响。选择提取结果中包含这4个属性的有效文本样本作为有效提取样本，共得到236份包含有效提取结果的病征报告。基于LLaMa3模型的克罗恩病关键特征信息有效提取率为：(有效提取样本数/总样本数)×100%=(236/297)×100%=79.46%。

依据236份病征报告对应的内镜检查报告，236份病征报告中共包含63例临床缓解期样本、62例重度活动期样本、111例重度活动期样本。如前所述，将病征报告中的有效提取结果转换为对应的属性诊断等级。当属性上的有效提取结果为空时，采用恶性程度最低的H_1等级进行填充。依时间升序关系，按照6∶2∶2的比例划分为训练数据集X_{l_1}、验证数据集X_{l_2}和测试数据集X_{l_3}。LLaMa3处理文本数据的流程如图2所示，Prompt的主要参数设置如表3所列。

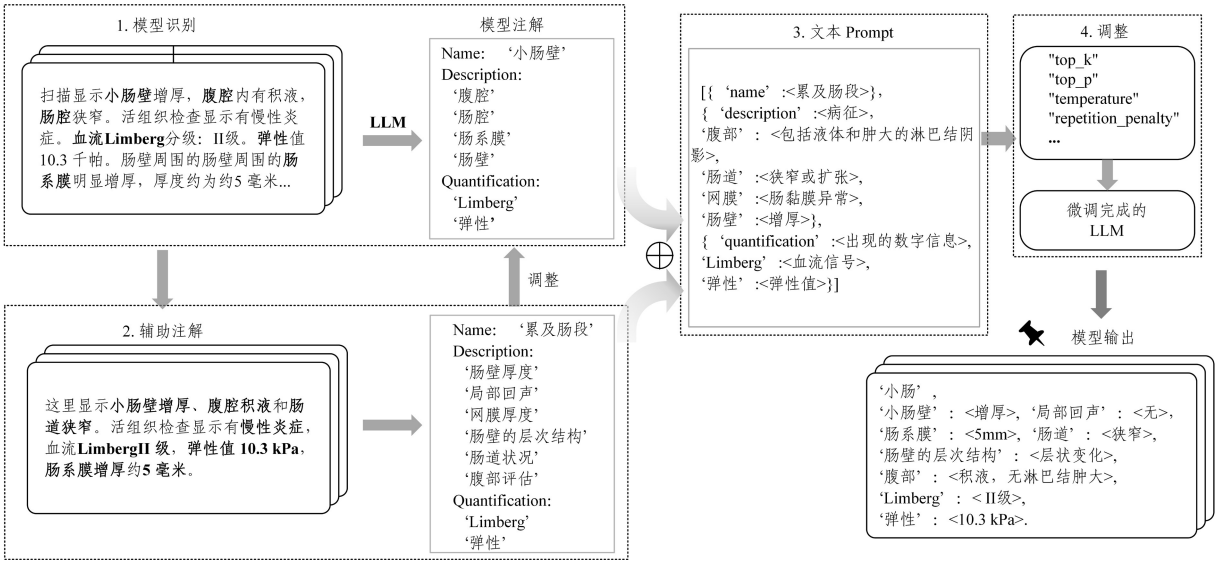

图 2　LLaMa3 数据处理流程

Fig. 2　Processing flow of LLaMa3 data

表 3　Prompt 主要参数

Table 3　Main parameters of Prompt

重要参数	参数设置
max_new_tokens	512
do_sample	True
top_k	50
top_p	0.95
temperature	0.3
repetition_penalty	1.3

4.2　评价指标与实验结果分析

从广泛应用于医疗领域且针对克罗恩病的基分类器中，选择朴素贝叶斯(Naive Bayes，NB)、支持向量机(Support Vector Machine，SVM)、决策树(Decision Tree，DT)、随机森林(Random Forest，DF)、逻辑回归(Logistic Regression，LR)、K 近邻(K-Nearest Neighbors，KNN)、卷积神经网络(Convolutional Neural Network，CNN)、梯度提升树(Gradient Boosting Trees，GBT)作为代表性的基分类器进行实验。

首先，将转换处理后的数据集输入上述 8 种基分类器中进行训练与预测。CD 活动性分级表示病灶的恶性程度，若活动性分级结果与金标准相同，则判定模型预测正确，否则为错误。各指标的计算方法如下：

$$precision=\frac{TP}{TP+FP} \tag{19}$$

$$recall=\frac{TP}{TP+FN} \tag{20}$$

$$F_1=\frac{2\times precision\times recall}{precision+recall} \tag{21}$$

AUC 是指 ROC 曲线下的面积，用于评估分类模型的性能。其取值范围在 0 到 1 之间，值越接近 1 表示模型性能越好。上述 8 种基分类器在验证集上的预测性能结果如表 4 所列。

根据表 4 所列的预测准确率，从 8 种基分类器中挑选基本组合 C_B 与多样化组合 C_D，利用 Choquet 积分方法融合每种组合的预测结果，其中不同分类器组合的参数如表 5 所列，组合后的预测准确率如表 6 所列。可以看出，SVM，RF 与 GBT 组合的预测准确率最高且该组合再与其他 C_{k_D} 组合后预测准确率不再上升，从而可确定 SVM，RF 与 GBT 为最佳基分类器组合。

表 4　不同基分类器的预测性能结果

Table 4　Prediction performance results of different base classifier

模型	Acc	F_1	AUC
SVM	0.8211	0.8218	0.8451
RF	0.8105	0.8122	0.8403
KNN	0.8000	0.7998	0.8184
LR	0.7895	0.7842	0.7984
GBT	0.7790	0.7772	0.7817
DT	0.7684	0.7687	0.8142
CNN	0.7579	0.7652	0.7585
NB	0.7158	0.7047	0.7232

表 5　Choquet 积分融合不同分类器组合的参数

Table 5　Choquet integral fusion of parameters from different classifier combinations

基本组合 C_B	多样化组合 C_D	μ	∞	$\boldsymbol{\beta}$	$\bar{\mu}$
SVM	RF	0.7491	0.6200	0.6158	0.2486
SVM，RF	KNN	0.5584	0.9540	0.5969	0.2058
SVM，RF	LR	0.6045	0.8581	0.6136	0.2300
SVM，RF	GBT	0.8081	0.4802	0.6156	0.1998
SVM，RF，GBT	DT	0.5595	0.8391	0.5713	0.1407
SVM，RF，GBT	CNN	0.5858	0.8156	0.6169	0.1888
SVM，RF，GBT	NB	0.6289	0.7374	0.5656	0.1741

表 6　各分类器组合的预测准确率

Table 6　Prediction accuracy of each classifier combination

基本组合 C_B	多样化组合 C_D	$A(C_B)$	$A(C_{BD})$
SVM	RF	0.8211	0.9153
SVM，RF	KNN	0.9153	0.9110
SVM，RF	LR	0.9153	0.8983
SVM，RF	GBT	0.9153	0.9364
SVM，RF，GBT	DT	0.9364	0.9280
SVM，RF，GBT	CNN	0.9364	0.9068
SVM，RF，GBT	NB	0.9364	0.8983

然后如 3.3 节所述，将最佳分类器组合的 3 个分类器作为集成分类器，并进行权重学习。验证集的预测准确率随相似度阈值变化的曲线如图 3 所示。结果显示，当相似度阈值

为 0.64 时，对应的验证集上的准确率为 0.8936，在验证集上预测准确率达到了最高值，且后续准确率不再提升。根据 3.3 节所述，最优相似度阈值即为 0.64。因此，计算测试集与训练集的相似度时，仅将与测试集样本相似度大于 0.64 的训练集样本加入计算，最终得到测试集的 Acc，F_1，AUC 分别为 0.9128，0.9149，0.9474。如表 4 所列，在该集成模型中，基分类器中预测准确率表现最佳的是 SVM，其 Acc，F_1，AUC 分别为 0.8211，0.8218，0.8451。通过采用集成方法，该模型在 Acc，F_1 和 AUC 这 3 个指标上均有 8%左右的提升，这表明集成模型在性能上有了显著提高。

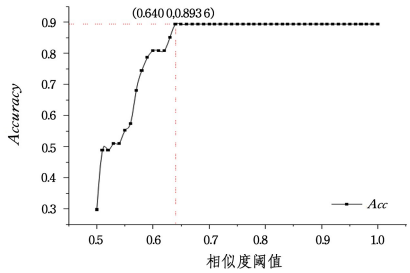

图 3 预测准确率随相似度阈值变化的趋势

Fig. 3 Trend of prediction accuracy with similarity threshold

4.3 对比实验

为比较 LCOS-SELM 模型与不同类型的传统集成模型的性能，进行了统计对比实验。选择单分类器算法多层感知器(MLP)、XGBoost、梯度提升机(GBM)与集成算法 Bagging，使用 236 份有效克罗恩病的病征报告、相同的数据集划分、相同基分类器组成以及相同评价指标，进行 150 次对比实验，实验结果如表 7 所列。

表 7 LCOS-SELM 与其他集成方法的对比结果

Table 7 Comparison results between LCOS-SELM and other integration methods

模型	Acc	F_1	AUC
LCOS-SELM	0.9128±0.0080	0.9149±0.0094	0.9474±0.4050
MLP	0.8632±0.0715	0.8632±0.0882	0.8606±0.0462
XGBoost	0.8105±0.0415	0.8061±0.0419	0.8987±0.0380
GBM	0.8316±0.0414	0.8335±0.0414	0.9166±0.0233
Bagging	0.8920±0.0208	0.8908±0.0212	0.9044±0.0129

基于实验结果，进行以下对比分析。1)对比单分类器算法，LCOS-SELM 模型的平均准确率达到 0.9128，表现最优。其次是多层感知器，其平均准确率为 0.8632，优于 XGBoost 与 GBM，这主要是因为 XGBoost 和 GBM 通常依赖于基于树的模型，可能在处理某些复杂模式时表现不如 MLP。XGBoost 的平均准确率为 0.8105，相对较低，这可能归因于其在处理该特定数据集时，在特征选择或模型复杂度方面存在不足。2)采用 Bagging 的方法集成的平均准确率为 0.8941，相比于单一分类器算法，Bagging 具有显著优势，主要体现在通过减少方差、增强模型稳定性、提升泛化能力和引入多样性等，显著提高了 Bagging 模型的准确率。然而，在 150 次重复实验中，LCOS-SELM 相较于传统 Bagging 方法，在 Acc，F_1 和 AUC 上平均提升了约 2%。进一步采用成对 t 检验[27]，在 Acc 上对 LCOS-SELM 和传统 Bagging 方法进行统计比较。设两种方法的 Acc 值为 A_L^n($n=1,\cdots,150$)和 A_B^n($n=1,\cdots,150$)，原假设为 $A_L^n=A_B^n$，备择假设为 $A_L^n>A_B^n$。检验结果为 $t=7.4700$，$p=6.4523\times10^{-12}<0.0500$，因此备择假设为真，凸显了 LCOS-SELM 的性能优势。

4.2 节和 4.3 节的实验结果表明，LCOS-SELM 在准确率指标上优于单分类器(平均提升幅度大于 8%)和传统集成学习分类器(平均提升幅度大于 2%)，即 LCOS-SELM 能够更准确地进行克罗恩病的活动性预测，为临床医生提供准确、高效的辅助诊断支持。

结束语 本文针对选择集成中相关分类器的筛选与合理赋权提升性能的问题，提出了 LCOS-SELM，并将其应用于克罗恩病的活动性诊断。该方法首先通过本地部署的 LLaMa3 模型对非结构化文本进行提词处理，既在保证数据隐私性的同时节省了人工成本，又解决了少样本训练问题；然后，利用 Choquet 积分融合基分类器的预测结果，识别基分类器之间的相关关系，确保筛选出更为互补的基分类器；最后，采用最优相似度策略进行集成，既保证了样本的一致性，又能根据每个样本的特点灵活调整基分类器权重，从而提高模型的预测准确性。

为验证所提方法的有效性，收集了安徽省某三甲医院 2021 年至 2023 年期间的 CD 诊断数据及相关医疗记录，建立了 CD 活动性预测模型，并将所提方法与其他 4 种集成学习方法进行了对比实验，实验结果表明，LCOS-SELM 显著提高了 CD 活动性预测精度。

未来的研究将重点探讨在保证模型性能的前提下提升集成效率，使模型能更高效地处理大量数据。同时，尝试将该方法应用于其他疾病的辅助诊断。应用该方法时，需要解决多个问题，如提取病征，将病征转化为诊断，筛选合适的基分类器，以及进行大量数据的相似度筛选，以便基于历史数据对相关疾病进行更加合理有效的辅助诊断。

参 考 文 献

[1] ZHANG L, WANG X, MOON W M. PolSAR images classification through GA-based selective ensemble learning [C]// 2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS). 2015: 3770-3773.

[2] ZHOU Z H. Ensemble methods: foundations and algorithms [J]. IEEE Computational Intelligence Magazine, 2013, 8(1): 9-77.

[3] OZCIFT A, GULTEN A. Classifier ensemble construction with rotation forest to improve medical diagnosis performance of machine learning algorithms [J]. Computer Methods and Programs in Biomedicine, 2011, 104(3): 51-443.

[4] LIN C, CHEN W, QIU C, et al. LibD3C: Ensemble classifiers with a clustering and dynamic selection strategy [J]. Neurocomputing, 2014, 123: 35-424.

[5] DONG X, YU Z, CAO W, et al. A survey on ensemble learning [J]. Frontiers of Computer Science, 2019, 14(2): 241-58.

[6] CORRIVEAU G, GUILBAULT R, TAHAN A, et al. Review and study of genotypic diversity measures for real-coded representations [J]. IEEE Transactions on Evolutionary Computation, 2012, 16(5): 695-710.

[7] SINGH P K，SARKAR R，NASIPURI M. Correlation-based classifier combination in the field of pattern recognition [J]. Computational Intelligence，2017，34(3)：839-874.

[8] KUNCHEVA L I. A bound on kappa-error diagrams for analysis of classifier ensembles [J]. IEEE Transactions on Knowledge and Data Engineering，2013，25(3)：494-501.

[9] NANNI L，LUMINI A. Evolved feature weighting for random subspace classifier [J]. IEEE Transactions on Neural Networks and Learning System，2008，19(2)：6-363.

[10] ZENG X D，WONG D F，CHAO L S. Constructing better classifier ensemble based on weighted accuracy and diversity measure [J]. The Scientific World Journal，2014，2014(1)：1-12.

[11] ZOU H，ZHU J，HASTIE T. New multicategory boosting algorithms based on multicategory fisher-consistent losses [J]. Annals of Applied Statistics，2008，2(4)：306-1290.

[12] WIEST I C，FERBER D，ZHU J，et al. Privacy-preserving large language models for structured medical information retrieval [J]. NPJ Digitital Medicine，2024，7(1)：257.

[13] FANG Y，ZHAO X，XIAO W，et al. Few-shot Learning for Heterogeneous Information Networks [J]. ACM Transcation on Information System，2024，42(4)：1-24.

[14] BRANKE J，CORRENTE S，GRECO S，et al. Using Choquet integral as preference model in interactive evolutionary multiobjective optimization [J]. European Journal of Operational Research，2016，250(3)：884-901.

[15] CHOQUET G. Theory of capacities [J]. Annales de l'Institut Fourier，1954，5：131-295.

[16] GRABISCH M，LABREUCHE C. A decade of application of the Choquet and Sugeno integrals in multi-criteria decision aid [J]. Annals of Operations Research，2010，175(1)：90-247.

[17] MURPHY K P. Machine learning—a probabilistic perspective [C]// 2012 Adaptive Computation and Machine Learning Series. 2012：147-182.

[18] VASWANI A，SHAZEER N，PARMAR N，et al. Attention is all you need [C]// Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017：6000-6010.

[19] HARRIS D，HARRIS S. Digital Design and Computer Architecture [M]. Morgan Kaufmann Publishers Inc.，2007：180-200.

[20] BHOWAL P，SEN S，VELASQUEZ J D，et al. Fuzzy ensemble of deep learning models using choquet fuzzy integral，coalition game and information theory for breast cancer histology classification [J]. Expert Systems with Applications，2022，190：116-167.

[21] SIAMI M，NADERPOUR M，LU J. A choquet fuzzy integral vertical bagging classifier for mobile telematics data analysis [C]// 2019 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE). 2019：1-6.

[22] XU X R，XU H X，LIU C，et al. Prospective study on the diagnosis and activity assessment of inflammatory bowel disease using intestinal ultrasound [J]. Journal of Gastroenterology and Hepatology，2013，22(10)：6-1013.

[23] LI Z Y，YANG X L，LI L H，et al. Study on the disease assessment of ulcerative colitis and Crohn's disease using ultrasound contrast quantification [J]. Sichuan Medical Journal，2024，45(4)：62-357.

[24] MEDELLIN-KOWALEWSKI A，WILKENS R，WILSON A，et al. Quantitative contrast-Enhanced ultrasound parameters in Crohn Disease：their role in disease activity determination with ultrasound [J]. AJR American Journal of Roentgenology，2016，206(1)：64-73.

[25] NOVAK K L，KAPLAN G G，PANACCIONE R，et al. A simple ultrasound score for the accurate detection of inflammatory activity in Crohn's Disease [J]. Inflammatory Bowel Diseases，2017，23(11)：2001-2010.

[26] WINSTON W L，GOLDBERG J B. Operations research：applications and algorithms (4th ed) [M]// Belmont，CA：Thomson Brooks Cole，2004：227-261.

[27] STUDEN T. The probable error of a mean [J]. Biometrika，1908，6(1)：1-25.

FU Chao，born in 1978，Ph.D，professor. His main research interest is intelligent decision-making methods and applications.

CHANG Wenjun，born in 1991，Ph.D，lecturer. His main research interest is intelligent decision-making methods and applications.

（责任编辑：柯颖）