基于语言概念格的规则分类模型

李子豪[1],吴甜歌[1],梁海威[2],沙立伟[1],刘 毅[1]

(1. 山东建筑大学 计算机科学与技术学院,山东 济南 250101;2. 山东国智智能科技有限公司,山东 济南 250000)

摘 要: 形式概念分析以一种结构化的方式挖掘具有可解释性的规则,具有广泛的应用前景。基于语言概念格提取具有语义的分类规则,提出一种新颖的基于语言概念格的规则分类模型(LCLC)。首先,经过模糊划分将输入映射到语言术语以构建语言概念形式背景,进而构建语言概念格;其次,提出一种基于语言概念格的分类规则提取算法,根据概念面积和规则权重确定用于分类的规则库;再次,给出LCLC的规则推理机制,即通过输入模糊化后,隶属度加权平均的方式计算规则前件隶属度,同时考虑规则权重以确定分类标签。在8个公开数据集上与机器学习分类以及其他基于规则的分类算法进行对比评估,LCLC的分类精度达到100%,规则数也更为精简,验证了其有效性。

关键词: 形式概念分析;分类规则;语言概念格;知识发现;规则提取

DOI: 10. 11907/rjdk. 251309 扫描二维码阅读全文:

中图分类号: TP311.13;TP18 **文献标识码:** A **文章编号:** 1672-7800(2026)007-0111-08

Rule Classification Model Based on Linguistic Concept Lattice

LI Zihao[1], WU Tiange[1], LIANG Haiwei[2], SHA Liwei[1], LIU Yi[1]

(1. *School of Computer Science and Technology, Shandong Jianzhu University, Ji'nan* 250101, *China*;
2. *Shandong Guozhi Intelligent Technology Co., Ltd., Ji'nan* 250000, *China*)

Abstract: Formal Concept Analysis (FCA) explores interpretable rules in a structured manner and has broad application prospects. This paper proposes a novel rule classification model based on linguistic concept lattice (LCLC), which extracts semantic classification rules based on linguistic concept lattice. First, input data is mapped to linguistic terms via fuzzy partitioning to build a linguistic concept formal context, from which the linguistic concept lattice is derived. Next, a rule extraction algorithm based on linguistic concept lattice is proposed, which determines the rule base for classification by considering the concept area and rule weight. Then, the rule reasoning mechanism of LCLC is introduced, where the antecedent membership degree of the rule is calculated using a weighted average of the membership degrees after fuzzification of the input, while considering the rule weights to determine the classification label. Ultimately, the effectiveness of LCLC was validated through comparative evaluations with machine learning classification and other rule-based classification algorithms across eight public datasets, that LCLC achieved a classification accuracy of 100% and had a more concise set of rules.

Key Words: formal concept analysis; classification rule; linguistic concept lattice; knowledge discovery; rule extraction

0 引言

基于规则的分类器利用明确的规则进行推理,能够将输入数据映射到预定义的类别或标签。其规则可用形式化的"IF-THEN"语句,由于系统决策过程透明,可处理不确定性和模糊性,且具有较强的可解释性和灵活性,广泛适用于需要领域知识的应用场景,如医疗诊断[1-3]和故障诊断[4-6]。同时,形式概念分析(Formal Concept Analysis, FCA)作为一种从形式背景进行数据分析和规则提取的强有力工具,其核心数据结构——概念格能够清晰地表征对象与属性之间的知识关联,并支持有效的知识推理[7-12]。

收稿日期: 2025-06-06

基金项目: 国家自然科学基金项目(62176142)

作者简介: 李子豪(1997-),男,CCF会员,山东建筑大学计算机科学与技术学院硕士研究生,研究方向为智能信息处理;吴甜歌(2000-),女,山东建筑大学计算机科学与技术学院硕士研究生,研究方向为智能信息处理;梁海威(1990-),男,山东国智智能科技有限公司职员,研究方向为智能信息处理;沙立伟(2000-),男,山东建筑大学计算机科学与技术学院硕士研究生,研究方向为智能信息处理;刘毅(1978-),男,硕士,山东建筑大学计算机科学与技术学院副教授,研究方向为机器学习、智能信息处理。本文通讯作者:刘毅。

尤其在处理复杂系统和大规模数据时，基于FCA的规则提取方法因其良好的可解释性，已成为结构化数据挖掘的研究热点。

传统的规则分类器中，规则通常根据领域专家经验知识来定义，虽然满足了分类过程的可信性，但当涉及非常复杂或未知情况时，这种方法的局限性便凸显出来。随着人工智能技术的发展，机器学习和神经网络等自适应学习技术能够根据场景变化自动调整规则和隶属函数，并以数据驱动的方式自动生成规则库[13-16]。然而，这种方法也带来了可解释性下降的问题。

本文从FCA角度，以一种透明可解释的方式生成并筛选较为精简的规则库，用于分类任务。为实现这一目标，提出了基于语言概念格的规则分类模型（Linguistic Concept Lattice，LCLC）。首先，为了处理具有不确定性的数据，利用模糊划分构建形式背景，挖掘语言概念知识；其次，提出一种基于概念稳定性和规则权重的模糊语言分类规则提取算法，计算语言概念格中的稳定性指标，并结合规则权重筛选出重要且有效的规则。这种方法保证了分类规则在精简的同时，能够保持较高的分类准确性，减少了传统方法中可能出现的规则冗余和计算效率低下的问题。本文通过实验与传统分类算法以及其他基于规则的分类算法进行对比，验证了LCLC在分类精度和规则数量方面的优越性。实验表明，LCLC不仅能够提供精确的分类结果，而且能够生成较精简的规则库，从而提高了分类系统的效率和可解释性。

1 相关工作

概念格相关基础知识如下：

定义 1[17] 设 $S=\{s_\alpha | t=-\tau,\cdots,-1,0,1,\cdots,\tau\}$ 是由奇数个语言项组成的有限语言项集，其中 τ 为正整数，若 S 符合以下条件：

（1）有序性：$s_i \geqslant s_j \Leftrightarrow i \geqslant j$；

（2）负算子：$Neg\left(s_{-i}\right)=s_i$，特别地，$Neg\left(s_0\right)=s_0$；

（3）最大算子：$s_i \geqslant s_j \Leftrightarrow \max\left(s_i,s_j\right)=s_i$；

（4）具有最小算子：$s_i \leqslant s_j \Leftrightarrow \min\left(s_i,s_j\right)=s_i$。

则称集合S为下标对称的语言术语集。

定义 2[17] 称三元组 $\left(U,L_{S_\alpha},I\right)$ 为一个语言概念形式背景，$U=\left\{x_1,x_2,\cdots,x_n\right\}$ 是非空有限的对象集，$L_{S_\alpha}=\{l^1_{s_{-\tau}},\cdots,l^1_{s_0},\cdots,l^1_{s_\tau},l^2_{s_{-\tau}},\cdots,l^2_{s_0},\cdots,l^2_{s_\tau},\cdots,l^n_{s_{-\tau}},\cdots,l^n_{s_0},\cdots,l^n_{s_\tau}\}$ 是非空有限的语言概念集，I 为 U 与 L_{S_α} 之间的二元关系，记为 $I\subseteq U\times L_{S_\alpha}$。对于 $x\in U$ 和 $l^i_{s_\alpha}\in L_{S_\alpha}$，如果 $\left(x,l^i_{s_\alpha}\right)\in I$，则称对象 x 拥有语言概念 $l^i_{s_\alpha}$，记为 $xIl^i_{s_\alpha}$。

定义 3[17] 设 $F=\left(U,L_{S_\alpha},I\right)$ 为一个语言概念形式背景，对于任意 $X\subseteq U$ 和 $B_{S_\alpha}\subseteq L_{S_\alpha}$，定义诱导算子“$'$”如下：

$$X'=\left\{l^i_{s_\alpha}\in L_{S_\alpha}\,|\,xIl^i_{s_\alpha},\forall x\in X\right\} \tag{1}$$

$$B'_{S_\alpha}=\left\{x\in U\,|\,xIl^i_{s_\alpha},\forall l^i_{s_\alpha}\in B_{S_\alpha}\right\} \tag{2}$$

其中，X' 表示 X 中所有对象共同拥有的语言概念集，B'_{S_α} 表示具有语言概念集 B_{S_α} 的对象集合。

若 $X'=B_{S_\alpha}$ 且 $B'_{S_\alpha}=X$，则称 (X,B_{S_α}) 是语言概念形式背景 F 中的一个语言概念知识，X 称为语言概念知识 (X,B_{S_α}) 的外延，B_{S_α} 称为语言概念知识 (X,B_{S_α}) 的内涵。$L(U,L_{S_\alpha},I)$ 表示语言概念形式背景中的所有语言概念知识。

对于 $\left(X^1,B^1_{S_\alpha}\right),(X^2,B^2_{S_\alpha})\in L(U,L_{S_\alpha},I)$，定义 $(X^1,B^1_{S_\alpha})$ 与 $(X^2,B^2_{S_\alpha})$ 之间的偏序关系如下：

$$\left(X^1,B^1_{S_\alpha}\right)\leqslant\left(X^2,B^2_{S_\alpha}\right)\Leftrightarrow X^1\subseteq X^2\Leftrightarrow B^1_{S_\alpha}\subseteq B^2_{S_\alpha} \tag{3}$$

$L(U,L_{S_\alpha},I)$ 在偏序关系下生成的完备格称为概念格，它们之间的上确界和下确界分别为：

$$\left(X_1,B^1_{S_\alpha}\right)\vee\left(X_2,B^2_{S_\alpha}\right)=\left(\left(X_1\cup X_2\right)'',B^1_{S_\alpha}\cap B^2_{S_\alpha}\right) \tag{4}$$

$$\left(X_1,B^1_{S_\alpha}\right)\wedge\left(X_2,B^2_{S_\alpha}\right)=\left(X_1\cap X_2,\left(B^1_{S_\alpha}\cup B^2_{S_\alpha}\right)''\right) \tag{5}$$

语言概念格的研究已经相对成熟，但其分类规则研究仍是空白。本文利用语言概念格挖掘具有语义和权重的分类规则，并将其应用于规则分类系统。

2 基于语言概念格的规则分类模型

分类系统框架LCLC如图1所示。模型训练部分主要包括利用模糊化方法计算得到语言概念形式背景，进而构建语言概念格。利用提出的模糊语言分类规则提取算法，计算语言概念格中的稳定性指标，并结合规则权重以提取分类器的模糊语言规则。在模型测试部分，则是利用模糊语言分类规则进行类标签预测。

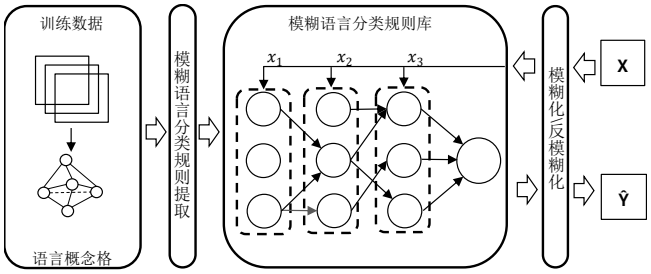

Fig. 1 The classification system framework of LCLC

图1 LCLC的分类系统框架

2.1 模糊化方法计算语言概念形式背景

模糊化是构建形式背景的重要步骤，它将精确的输入数据映射到模糊集，以便在模糊推理过程中使用。常用的模糊化方法包括专家经验知识、定义隶属度函数、自适应算法构造隶属度函数、聚类分析定义模糊集合等方法。为了保证LCLC的可解释性，以便于理解和应用，本文采用经

典的模糊理论方法将其划分为模糊集定义隶属度函数。

为了构建语言概念形式背景进而推理出模糊语言规则,为每个数据类型为连续值的属性选取语言术语集 $S=\{s_{-1}=小, s_0=中, s_1=大\}$。每个属性 l^i 与语言术语集构成3个语言概念 $\{l^i_{s_{-1}}, l^i_{s_0}, l^i_{s_1}\}$,并分别对应3个模糊集。语言概念定义的三角隶属度函数如图2所示,纵坐标为隶属度,横坐标为特征数值。其中:

$$A(x)=\begin{cases}\dfrac{b-x}{b-a}, & a\leqslant x\leqslant b\\ 0, & 其他\end{cases} \tag{6}$$

$$B(x)=\begin{cases}\dfrac{x-a}{b-a}, & a\leqslant x\leqslant b\\ \dfrac{c-x}{c-b}, & b\leqslant x\leqslant c\\ 0, & 其他\end{cases} \tag{7}$$

$$C(x)=\begin{cases}0, & 其他\\ \dfrac{x-b}{c-b}, & b\leqslant x\leqslant c\end{cases} \tag{8}$$

其中,$A(x)$、$B(x)$、$C(x)$分别表示小、中、大3个语言术语的隶属度函数,a、b和c分别表示该特征 l^{ii} 取值的最小值,中间值和最大值,且 $b=\dfrac{a+c}{2}$。分别计算特征 l^{ii} 的取值对于3个隶属函数的隶属度,按最大隶属度为特征取值划分语言概念。

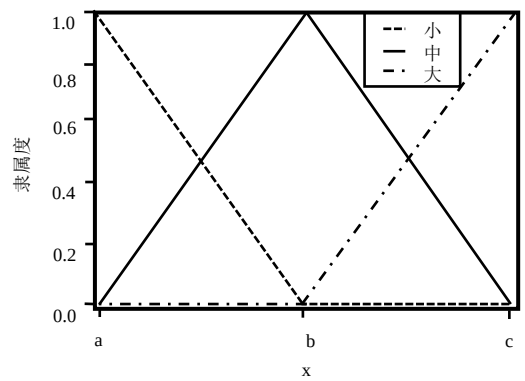

Fig. 2　Triangular membership function

图2　三角隶属函数

例 1　表1是一个二分类的数据集,包含 $U=\{x_1, x_2, x_3, x_4, x_5, x_6\}$ 6个样本、$L=\{l^1, l^2, l^3, l^4\}$ 4个特征属性、$class=\{-1, 1\}$两个类别。

Table 1　Binary classification dataset

表1　二分类数据集

	l^1	l^2	l^3	l^4	class
x_1	4.9	3.5	1.4	0.3	1
x_2	5.5	3.5	4.5	1.1	−1
x_3	7.9	3.8	6.4	2	−1
x_4	5.7	4.4	1.5	0.4	1
x_5	5.4	3.9	1.3	0.3	1
x_6	4.9	3.4	3.3	1	−1

根据上文对隶属度函数的定义,以属性 l^1 为例,在 x_1、x_2、x_3、x_4、x_5、x_6 6个样本中最小值 a、中间值 b,最大值 c 分别为:$a=4.9, b=6.4, c=7.9$。对象 x_1 在属性 l^1 的特征值为4.9,则 $A(4.9)=1, B(4.9)=0, C(4.9)=0$,因而4.9在 l^1 中划分的语言术语为"小",即对象 x_1 拥有语言概念 $l^1_{s_{-1}}$。同理可以计算出每个对象对应的每个语言概念的隶属度,如表2所示。通过最大隶属原则构建出语言概念形式背景 (U, L_{S_α}, I),如表3所示。其中,样本 $U=\{x_1, x_2, x_3, x_4, x_5, x_6\}$,属性 $L=\{l^1, l^2, l^3, l^4\}$,语言术语集 $S_\alpha=\{s_{-1}=小, s_0=中, s_1=大\}$,可以得到对应的语言概念集 $L_{s_\alpha}=\{l^1_{s_{-1}}, l^1_{s_0}, l^1_{s_1}, l^2_{s_{-1}}, l^2_{s_0}, l^2_{s_1}, l^3_{s_{-1}}, l^3_{s_0}, l^3_{s_1}, l^4_{s_{-1}}, l^4_{s_0}, l^4_{s_1}\}$。

Table 2　Membership degree of linguistic concepts

表2　语言概念的隶属度

	$l^1_{s_{-1}}$	$l^1_{s_0}$	$l^1_{s_1}$	$l^2_{s_{-1}}$	$l^2_{s_0}$	$l^2_{s_1}$	$l^3_{s_{-1}}$	$l^3_{s_0}$	$l^3_{s_1}$	$l^4_{s_{-1}}$	$l^4_{s_0}$	$l^4_{s_1}$
x_1	1	0	0	0.8	0.2	0	0.96	0.04	0	1	0	0
x_2	0.6	0.4	0	0.8	0.2	0	0	0.75	0.25	0.06	0.94	0
x_3	0	0	1	0.2	0.8	0	0	0	1	0	0	1
x_4	0.47	0.53	0	0	0	1	0.92	0.08	0	0.88	0.12	0
x_5	0.67	0.33	0	0	1	0	1	0	0	1	0	0
x_6	1	0	0	1	0	0	0.22	0.78	0	0.18	0.82	0

Table 3　Linguistic concept formal context(U, L_{S_α}, I)

表3　语言概念形式背景(U, L_{S_α}, I)

	$l^1_{s_{-1}}$	$l^1_{s_0}$	$l^1_{s_1}$	$l^2_{s_{-1}}$	$l^2_{s_0}$	$l^2_{s_1}$	$l^3_{s_{-1}}$	$l^3_{s_0}$	$l^3_{s_1}$	$l^4_{s_{-1}}$	$l^4_{s_0}$	$l^4_{s_1}$
x_1	1	0	0	1	0	0	1	0	0	1	0	0
x_2	1	0	0	1	0	0	0	1	0	0	1	0
x_3	0	0	1	0	1	0	0	0	1	0	0	1
x_4	0	1	0	0	0	1	1	0	0	1	0	0
x_5	1	0	0	0	1	0	1	0	0	1	0	0
x_6	1	0	0	1	0	0	0	1	0	0	1	0

2.2　规则提取

如何从语言概念格中推理出具有规则权重的分类规则是下文讨论的主要内容。首先,筛选稳定的概念,以概念的内涵作为规则前件,概念外延的对象标签作为规则后件;其次,明确概念外延中的对象标签并不唯一,即一个概念有可能诱导出多个规则。规则权重由概念外延中规则后件的标签对象所占概念外延全部对象的比例而确定。

规则库的第 i 条分类规则表示如下:

$$R_i: If\ A_{S_\alpha}\ \text{then Class is}\ C_i\ \text{with}\ W_i \tag{9}$$

其中,A_{S_α} 为某一语言概念知识的内涵,即语言概念集合,C_i 为语言概念知识外延涵盖的某一类别标签,W_i 为规则权重。分类 R_i 其前件直接对应人类可理解的特征描述(小、中、大及其组合),使得分类决策的依据透明,其规则后件则对应分类结果。

分类规则提取的两个指标如下:

(1)概念面积。概念格蕴含了对象和属性之间的相互影响,每个概念可作为一个聚类,概念内涵作为聚类标签,概念外延是聚类结果。这种聚类方式也存在一定的不妥,当概念内涵中属性的限制过低时,会使概念外延的聚类效

果不佳;当概念内涵中属性的限制过高时,会使概念外延的对象个数太少而使得聚类无意义。

因此,需要一种指标实现对不合理概念聚类的筛选,只利用合理的概念聚类提取分类规则。文献[18]提出概念面积:$S\left(X, B_{S_\alpha}\right) = |X| \times \left|\boldsymbol{B}_{S_\alpha}\right|$以衡量一个概念的稳定性。较大的概念面积可以保障属性个数和对象个数都不会出现太低的情况,即同一类别内具有可观的对象数量且对象间有较高相似度。概念面积可以作为限制不必要规则的参数之一。

(2)规则权重。以概念内涵(*intent*)为规则前件,概念外延(*extent*)的对象类别标签(*lebel*)为规则后件。由于一个概念包含的对象类别并不唯一,因而由概念计算的规则附有权重 W。规则权重为对应概念外延中同一类别的对象个数 ϑ 占概念外延所有对象个数 θ 的比例。权重较低的规则在 LCLC 的规则库中对分类任务参考价值较低,因而规则权重可以作为筛选不必要规则的又一参数。一个概念 *concept* 可以确定至少一条规则 if *concept.intent* then Class *concept.lebel* with $W = \dfrac{\vartheta}{\theta}$。

结合上文给出的规则,提取算法如下:

算法 1 模糊语言分类规则提取算法

输入:语言概念形式背景(U, L_{S_α}, I),概念面积阈值α,规则权重阈值β

输出:模糊语言分类规则集合 R

1. 初始化 $R = \varnothing$

2. 根据(U, L_{S_α}, I)计算语言概念格 *Concepts*

3. 根据概念内涵和概念外延计算概念面积 *concept.area*

4. FOR *concept* in *Concepts* DO

5. IF *concept.area* > α THEN

6. 根据概念外延确定类别标签

7. rule = if *concept.intent* then Class is *concept.lebel* with $W = \dfrac{concept.\vartheta}{concept.\theta}$

8. IF $W > \beta$ THEN:

9. R.add(rule)

10. END

11. END

12. END

13.return R

对算法 1 的时间复杂度分析:由于语言概念格的构建与经典的概念格构造方法类似,因而本文采用的是概念格的并行构造方法[19-20],时间复杂度为 $O(M^2 \cdot N \cdot C)$。算法 1 除计算语言概念格这部分外,其余部分的时间复杂度为 $O(C)$,由于 $O(C) \ll O(M^2 \cdot N \cdot C)$,因而算法 1 的时间复杂度为 $O(M^2 \cdot N \cdot C)$,其中 M 为属性个数,N 为对象个数,C 为概念数。

由此可见,模型训练过程的整体时间复杂度仍由语言概念格的构造过程主导。高维数据下的概念格构造是 FCA 的固有挑战,原因在于:①概念格构造算法随着属性数量 M 与对象数量 N 的增加而复杂度显著增加;②概念数量可能随属性与对象维度的增加而呈指数上升,导致模型整体复杂度激增。因此,在高维数据上的计算效率是需要重点关注的问题。探索更高效的格构造算法、增量更新方法或结合属性约简技术将是未来工作的重要方向,以进一步提升 LCLC 在大规模高维数据集上的实用性。随着高性能计算技术的发展和成熟,训练模型的时间复杂度将得到显著改善。

例 2(续例 1) 根据表 3 语言概念形式背景计算所得语言概念格如图 3 所示,根据算法 1,设概念面积阈值为 6,满足概念面积的 4 个概念为 C,、G、H、J,设规则权重阈值为 0.6,所得到的规则如表 4 所示。以规则 R_1 为例,由概念 C:$(2,6;\ l^1_{s_{-1}}, l^2_{s_{-1}}, l^3_{s_0}, l^4_{s_0})$计算其概念面积为 8 则满足阈值。同时,概念外延的对象 2、6,类别标签均为“-1”,可得规则权重为“1”。规则 $l^1_{s_{-1}}, l^2_{s_{-1}}, l^3_{s_0}, l^4_{s_0} \to -1$,其语义可表述为:如果一个样本属性 l^1 较小,属性 l^2 较小,属性 l^3 一般并且属性 l^4 一般,则该样本为“-1”类。

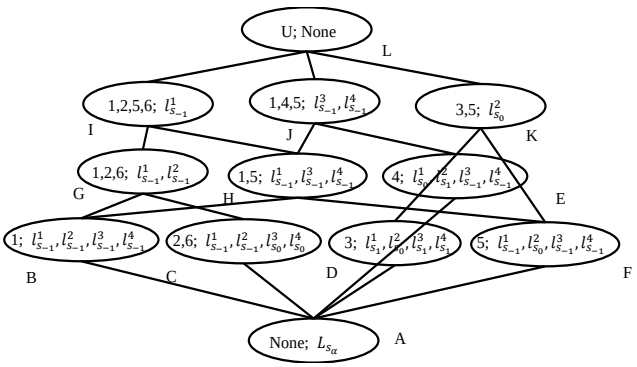

Fig. 3 Linguistic concept lattice

图 3 语言概念格

Table 4 Classification rule base

表 4 分类规则库

概念	概念面积	规则序号	规则	规则权重
C:$(2,6;\ l^1_{s_{-1}}, l^2_{s_{-1}}, l^3_{s_0}, l^4_{s_0})$	8	R_1	$l^1_{s_{-1}}, l^2_{s_{-1}}, l^3_{s_0}, l^4_{s_0} \to -1$	1
G:$(1,2,6;\ l^1_{s_{-1}}, l^2_{s_{-1}})$	6	R_2	$l^1_{s_{-1}}, l^2_{s_{-1}} \to -1$	0.67
H:$(1,5;\ l^1_{s_{-1}}, l^3_{s_{-1}}, l^4_{s_{-1}})$	8	R_3	$l^1_{s_{-1}}, l^3_{s_{-1}}, l^4_{s_{-1}} \to 1$	1
J:$(1,4,5;\ l^3_{s_{-1}}, l^4_{s_{-1}})$	6	R_4	$l^3_{s_{-1}}, l^4_{s_{-1}} \to 1$	1

2.3 类别推理

如何利用训练好的规则进行模糊推理,从而计算分类结果是下文讨论的主要问题。

LCLC 的测试部分可以总结为 3 部分,即模糊化、模糊推理和去模糊化。对于输入的一组特征向量,需要计算每一维度的特征对于 3.1 节中每个特征的“小”“中”“大” 3 个模糊集的隶属度,完成清晰的数值向语言变量的映射,即“模糊化”。

模糊化后的特征向量利用规则库进行推理,从输入的语言变量和隶属度推导出输出部分,该过程的核心是计算每个规则的激活度。本文采用对隶属度加权平均的激活

度计算方式，设所有推理得出的规则为 $Rule = \{R_1, R_2, ..., R_m\}$，输入一组特征向量$\left(l^1, l^2, ..., l^n\right)$，对于规则 R_i: $If\ A_{S_\alpha}$ then Class is C_i with W_i 的激活度计算如下：

$$Activation = \frac{\sum_{l^j_{s_\alpha}}^{A_{S_\alpha}} w_j \times \mu_{l^j_{s_\alpha}}\left(l^j\right)}{\sum_{l^j_{s_\alpha}}^{A_{S_\alpha}} w_j} \tag{10}$$

其中，l^j 为输入特征向量的第 j 个特征，$l^j_{s_\alpha}$ 为规则前件 A_{S_α} 中关于 l^j 的语言概念，$\mu_{l^j_{s_\alpha}}\left(l^j\right)$ 为特征 l^j 对语言概念 $l^j_{s_\alpha}$ 的隶属度，w_j 为特征 l^j 的特征权重，默认值为 1。

采用最大化的去模糊化方法，计算每条规则的激活度和规则权重的乘积，并将其作为规则强度 intensity，从而确定输出类别。

$$Class = \underset{C_i, 0 \leqslant i \leqslant m}{argmax}\left(Activation_i \times W_i\right) \tag{11}$$

例 3（续例 2）　给定一个样本的输入特征向量（l^1: 4.9，l^2: 3.9，l^3: 1.4，l^4: 0.6），根据例 1 的模糊化过程，输入特征向量的模糊化结果如表 5 所示。

Table 5　Fuzzification result of input vector

表 5　输入向量模糊化结果

	l^1: 4.9	l^2: 3.9	l^3: 1.4	l^4: 0.6
s_{-1}	1	0	0.96	0.65
s_0	0	1	0.04	0.35
s_1	0	0	0	0

根据式（6）和式（7）分别计算输入特征向量对于表 4 中 4 条规则的强度，具体如下：

$$R_1: Intensity_1 = W_1 \times \frac{1 + 0 + 0.04 + 0.35}{4} = 0.35 \tag{12}$$

$$R_2: Intensity_2 = W_2 \times \frac{1 + 0}{2} = 0.34 \tag{13}$$

$$R_3: Intensity_3 = W_3 \times \frac{1 + 0.96 + 0.65}{3} = 0.87 \tag{14}$$

$$R_4: Intensity_4 = W_4 \times \frac{0.96 + 0.65}{2} = 0.81 \tag{15}$$

显然，输入的特征向量对 R_3 的强度最高，最终得到的分类结果是类别 1。

3　实验与结果分析

3.1　实验说明

本文利用 PyCharm，在 Python3.11 环境中对算法性能进行评估。使用来自 UCI 的 8 个数据集，包括 Wine 数据集、WDBC 数据集、Forestfires 数据集、Heart 数据集、ILPD 数据集、Seed 数据集、parkinsons 数据集、pima 数据集，8 个数据集的关键信息如表 6 所示。此外，本文对部分数据集进行了标准化处理，即将 Forestfires 数据集简化为二分类问题（着火/非着火，忽略着火面积），并移除 WDBC 数据集中可能引入偏差的性别特征。

在对比实验中，为确保实验可重复性，所有数据划分过程均通过设置全局随机种子实现确定性采样，按照 7∶3 的比例分割训练集与测试集。将 LCLC 与机器学习算法进行比较，讨论其分类精度，并将 LCLC 与其他基于规则的分类算法进行比较，从规则数量和分类精度两方面进行对比分析。

Table 6　Experimental datasets

表 6　实验数据集

Dataset	#Instance	#Attribute	#Classe	Dataset	#Instance	#Attribute	#Classe
Wine	178	13	3	Heart	270	13	2
WDBC	569	30	2	ILPD	583	9	2
Forestfires	517	8	2	Seed	210	7	3
Parkinsons	197	22	2	Pima	768	8	2

3.2　参数分析

LCLC 实验关键参数设置如下：

（1）模糊隶属度函数设置。为每一维的特征固定设置“大”“中”“小”3 个隶属度函数，如图 2 所示。利用模糊化后构建的语言概念形式背景，计算语言概念格。

（2）规则计算参数设置。在由语言概念格计算规则时，概念面积阈值 α 和规则权重阈值 β 是两个重要参数。α 影响规则质量，取值太小会造成规则太过冗余，准确率得不到提升又严重影响计算效率；α 取值太大则会使得规则覆盖不完全，导致分类准确率下降。经过实验可知，筛选概念格中概念面积最大的前一半概念分类效果最佳。由概念计算规则时，存在部分规则权重很低的情况，即使其规则激活度很高，但对分类帮助不大。因此，β 可以进一步筛选重要的规则，实验中通常设置参数 $\beta >= 0.5$。

（3）特征权重设置。LCLC 的规则推理中计算规则激活度时，为每一维度的特征都设置了特征权重，实验中默认设置为 1。

实验中，概念面积阈值 α 和规则权重阈值 β 两个参数对于规则数量和分类精度具有重要影响。参数 α 在 12 个数据集中对概念数的限制如图 4 所示。通过概念面积对概念数折半，保留稳定性较高的概念，同时也保证了下文生成的规则质量。

参数 β 直接影响了规则质量，在概念数折半操作后，分别取 0.5、0.7、0.9 观察规则数量和分类准确率变化，如图 5 和图 6 所示。随着 β 的增大，所有数据集满足的规则数逐渐减少，并且大部分实验数据集的分类精确度随之先升高后降低。由此可见，规则过多导致出现大量冗余，将会降低分类精度，而如果 β 太大将导致有效的分类规则被删除，也会降低分类精度。

α 通过限制概念面积筛选出稳定性高的概念，进而影响规则数量和分类准确率。β 则直接筛选规则，从而在 LCLC 中保留权重较大的规则。实验中 β 越大，满足的规则越少，LCLC 在保证分类准确率的情况下，通过设置 β 可以实现规则库的精简。

3.3　与传统分类器比较

将 LCLC 与传统分类器对比，对比的分类算法包括支持向量机（SVM）、朴素贝叶斯（NB）和随机梯度下降

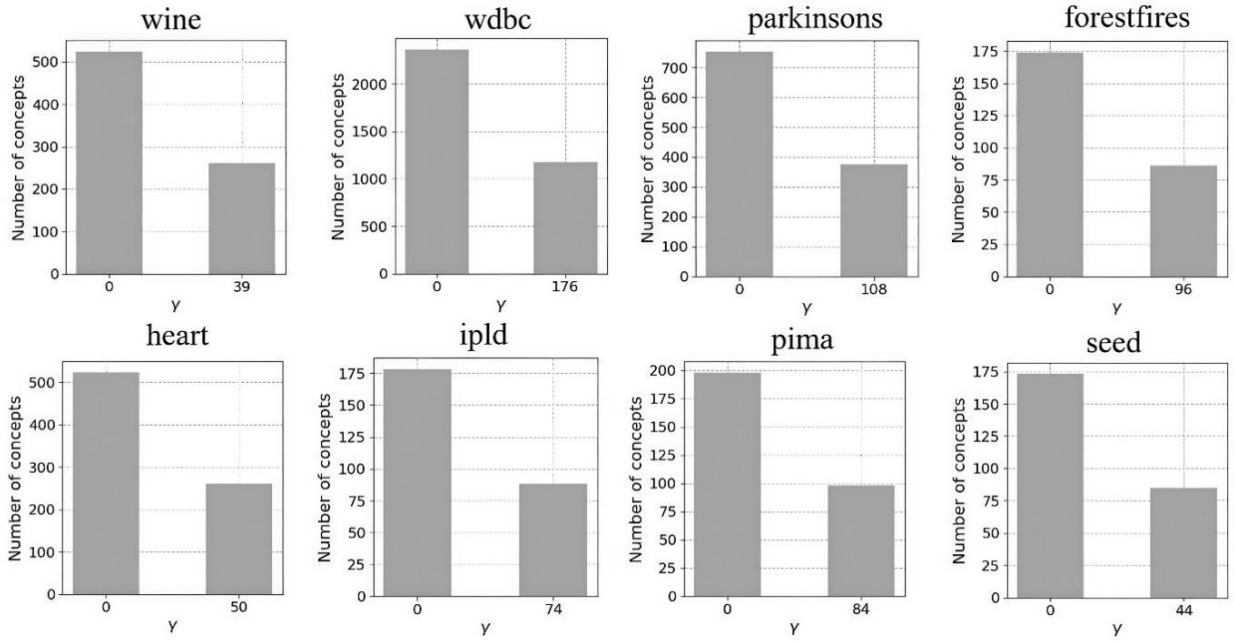

Fig. 4 Variation of concept count with parameters α

图4 概念数随参数α变化

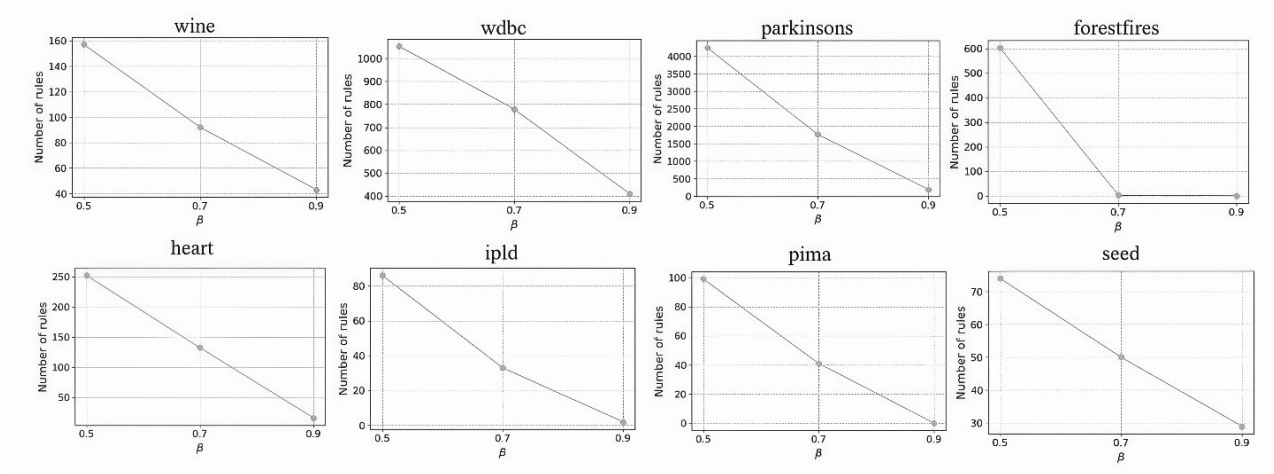

Fig. 5 Variation of the number of rules with parameters β

图5 规则数随参数β变化

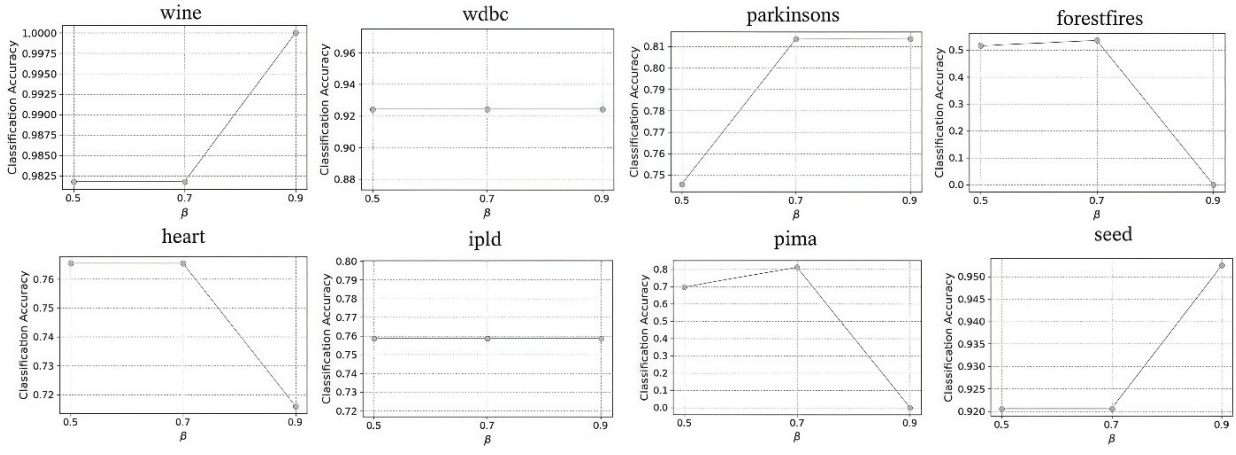

Fig. 6 Variation of classification accuracy with parameters β

图6 分类准确率随参数β变化

(SGD)。传统分类器的参数通常使用其默认值,4个分类器在所有数据集上的分类准确率如表7所示。本文提出的LCLC在实验中8个公开数据集中的5个数据集上达到了最高的分类精度,在Wine数据集的测试中达到了100%的准确率。

Table 7　Comparison of classification accuracy between LCLC and traditional classifiers

表7　LCLC与传统分类器分类精度比较　　(%)

Dataset	SVM	NB	SGD	LCLC(Our)
Wine	66.67	97.15	96.30	**100.00**
WDBC	87.72	91.81	92.40	**96.96**
Forestfires	**59.62**	55.77	51.92	58.59
Heart	59.26	82.72	76.54	**87.65**
ILPD	64.94	63.79	66.09	**75.86**
Seed	85.71	88.89	85.71	**93.65**
Parkinsons	86.44	71.19	**91.53**	83.05
Pima	75.76	77.06	**79.22**	71.86
Average	73.2650	78.5475	79.9638	**83.4525**

LCLC与其他3个传统分类模型的平均分类准确率比较可知,其在保证分类机制可解释性的同时,在分类精度上仍具有优秀表现。

通过上述实验对比可知,LCLC在多数数据集中取得了最佳分类性能。结果表明,基于语言概念格规则提取的规则分类模型为分类问题提供了一种新颖有效且可解释性更高的方法。

3.4　与基于规则的分类器比较

为了进行更加全面的比较,将LCLC与3种基于规则的分类器,从分类准确率和规则数两方面进行比较分析。对比方法分别是基于模糊规则的分类系统(FRBCS)[21]、基于置信规则库的分类系统(BRBCS)[22]和扩展置信规则库(EBRB)[23]。

如表8所示,在实验的8个数据集中,LCLC在5个数据集上的分类准确度更高,并且LCLC的平均分类准确率更优。综合来看,LCLC在分类性能上具有明显优势。

Table 8　Comparison of classification accuracy between LCLC and rule-based classifiers

表8　LCLC与基于规则的分类器分类精度比较　　(%)

Dataset	FRBCS	EBRB	BRBCS	LCLC(Our)
Wine	90.95	93.11	96.26	**100.00**
WDBC	93.16	94.13	94.56	**96.96**
Forestfires	57.05	58.32	**61.96**	58.59
Heart	75.35	82.10	87.45	**87.65**
ILPD	70.69	72.27	71.82	**75.86**
Seed	87.57	90.95	88.57	**93.65**
Parkinsons	83.05	**85.67**	84.90	83.05
Pima	70.20	**72.16**	71.50	71.86
Average	78.5025	81.0888	82.1275	**83.4525**

如表9所示,在实验的8个数据集中,LCLC的规则数量更加精简。结合平均分类准确率,LCLC在wine数据集、heart数据集、ilpd数据集、seed数据集4个数据集中,在实现分类准确率较高的同时,规则也更为精简。

Table 9　Comparison of the number of rules between LCLC and rule-based classifiers

表9　LCLC与基于规则的分类器规则数比较

Dataset	FRBCS	EBRB	BRBCS	LCLC(Our)
Wine	402	455	402	**43**
WDBC	**127**	142	**127**	166
Forestfires	68	563	68	**22**
Heart	176	243	176	**66**
ILPD	59	286	59	**33**
Seed	52	168	52	**40**
Parkinsons	**99**	406	**99**	419
Pima	89	571	89	**23**

为检验模型的显著性,进行统计检验。验证FLRBCS与3个基于规则的分类模型是否存在显著的差异。Friedman检验[41]的相关定义如下:

$$x_F^2=\frac{12H}{G(G+1)}\left(\sum_{i=1}^{G}g_i^2-\frac{G(G+1)^2}{4}\right)\tag{16}$$

其中,G为模型个数,H为数据集个数,g_i表示第i中模型在所有数据集中的平均排名,如表10所示。

Table 10　Ranking of classification accuracy

表10　分类准确率排名

Dataset	FRBCS	EBRB	BRBCS	LCLC(Our)
Wine	4	3	2	**1**
WDBC	4	3	2	**1**
Forestfires	4	3	**1**	2
Heart	4	3	2	**1**
ILPD	4	2	3	**1**
Seed	4	2	3	**1**
Parkinsons	3.5	**1**	2	3.5
Pima	4	**1**	3	2
Average	3.94	2.25	2.25	**1.56**

根据式(8)计算模型的$x_F^2=14.92$,大于自由度为3时的x_F^2临界值11.34,通过x_F^2分布计算,p值=0.002<0.05,说明4种模型性能具有差异性。

综合实验数据可知,LCLC的分类精度在超半数的数据集中更高,并且平均分类精度在对比模型中维持在更高水平;在规则数量比较中,LCLC在超半数的数据集中更加精简。实验结果表明,基于语言概念格规则提取的模糊规则分类系统为分类问题提供了一种新颖有效且可解释性更高的方法。

4　结语

本文提出了基于语言概念格的规则分类模型,该模型利用模糊化的语言术语构建语言概念形式背景,利用概念格聚类的稳定性分析提取用于分类的规则,并且以一种可解释的方法为基于规则的分类系统提供了又一规则提取思路。在8个公开数据集中,将其与传统分类器以及其他基于规则的分类器进行了比较,从而评估了本文模型的有效性。实验结果表明,本文基于语言概念格的规则分类模型在分类性能上优于传统分类器以及其他基于规则的分

类器。

未来研究中,可探索如何在保证分类准确性的前提下,进一步优化规则库,以去除冗余规则,降低计算复杂度。

参考文献:

[1] HAN S T, MOON Y, LEE H, et al. Rule-based continuous line classification using shape and positional relationships between objects in piping and instrumentation diagram [J]. Expert Systems with Applications, 2024, 248: 123366.

[2] DUAN C, LIU Q, WANG J, et al. GWO+ RuleFit: rule-based explainable machine-learning combined with heuristics to predict mid-treatment FDG PET response to chemoradiation for locally advanced non-small cell lung cancer[J]. Physics in Medicine & Biology, 2024, 69(15): 155018.

[3] TOKEDE B, BRANDON R, LEE C T, et al. Development and validation of a rule-based algorithm to identify periodontal diagnosis using structured electronic health record data[J]. Journal of Clinical Periodontology, 2024, 51:547-557.

[4] ZHOU W, LIU X, BAI H, et al. Intelligent medical diagnosis and treatment for diabetes with deep convolutional fuzzy neural networks[J]. Information Sciences, 2024, 677:120802.

[5] CHEN J, CHEN H, SHI J, et al. Factor diagnosis and governance strategies of ship oil spill accidents based on formal concept analysis[J]. Marine Pollution Bulletin, 2023, 196: 115606.

[6] ZHAO B Y, ZHANG Q X, HE W, et al. A deep belief rule base-based fault diagnosis method for complex systems[J]. ISA Transactions, 2024, 150: 77-91.

[7] FENG Z, ZHOU Z, HU C, et al. Fault diagnosis based on belief rule base with considering attribute correlation[J]. IEEE Access, 2017, 6: 2055-2067.

[8] WILLE R. Restructuring lattice theory: an approach based on hierarchies of concept[J]. Ordered Sets, 1982(1):314-339.

[9] GANTER B, WILLE R. Formal concept analysis: mathematical foundations[M]. Berlin: Springer, 1999:157-192.

[10] KIM E H, KIM H G, HWANG S H, et al. Farm: an FCA-based association rule miner[J]. Knowledge-Based Systems, 2015, 85: 277-297.

[11] DOLQUES X, LE B F, HUCHARD M, et al. Performance-friendly rule extraction in large water data-sets with AOC posets and relational concept analysis[J]. International Journal of General Systems, 2016, 45(1-2): 1-24.

[12] FAN M, LUO S, LI J. Network rule extraction under the network formal context based on three-way decision [J]. Applied Intelligence, 2022, 53:5126-5145.

[13] ZARATE L E, DIAS S M, SONG M J. FCANN: a new approach for extraction and representation of knowledge from ANN trained via formal concept analysis[J]. Neurocomputing, 2008, 71(13-15): 2670-2684.

[14] AGHAEIPOOR F, SABOKROU M, FERNÁNDEZ A. Fuzzy rule-based explainer systems for deep neural networks: from local explainability to global understanding[J]. IEEE Transactions on Fuzzy Systems, 2023, 31(9):3069-3080.

[15] CINTRA M E, CAMARGO H A, MONARD M C. Genetic generation of fuzzy systems with rule extraction using formal concept analysis[J]. Information Sciences, 2016, 349:199-215.

[16] BAI Y, LU X. Multiple kernel learning-based rule reduction method for fuzzy modeling[J]. Fuzzy Sets and Systems, 2023, 465:108534.

[17] ZOU L, PANG K, SONG X, et al. A knowledge reduction approach for linguistic concept formal context [J]. Information Sciences, 2020, 524: 165-183.

[18] LIU Z H, ZHAO Q, ZOU L, et al. A heuristic concept construction approach to collaborative recommendation[J]. International Journal of Approximate Reasoning, 2022, 146:119-132.

[19] KRAJCA P, OUTRATA J, VYCHODIL V. Parallel recursive algorithm for FCA[C]//Proceedings of the Sixth International Conference on Concept Lattices and Their Applications, 2008: 71-82.

[20] PACKIARAJ M, KAILASAM S. HyPar-FCA: a distributed framework based on hybrid partitioning for FCA[J]. The Journal of Supercomputing, 2022, 78(10):12589-12620.

[21] CHI Z R, YAN H, PHAM T. Fuzzy algorithms: with applications to image processing and pattern recognition[M]. Singapore: World Scientific, 1996.

[22] JIAO L, PAN Q, DENOEUX T, et al. Belief rule-based classification system: extension of FRBCS in belief functions framework[J]. Information Sciences, 2015, 309: 26-49.

[23] LIU J, MARTINEZ L, CALZADA A, et al. A novel belief rule base representation, generation and its inference methodology [J]. Knowledge-based Systems, 2013, 53:129-141.

(责任编辑:孙 娟)