计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2021,30(7):22−31 [doi: 10.15888/j.cnki.csa.007963]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

融合机器学习与知识推理的可解释性框架①

李迪媛[1], 康达周[2,3]

[1](南京航空航天大学 计算机科学与技术学院/人工智能学院, 南京 211106)
[2](南京航空航天大学 高安全系统的软件开发与验证技术工信部重点实验室, 南京 211106)
[3](软件新技术与产业化协同创新中心, 南京 210023)
通讯作者: 康达周, E-mail: dzkang@nuaa.edu.cn

摘　要: 针对基于规则的可解释性模型可能出现的规则无法反映模型真实决策情况的问题, 提出了一种融合机器学习和知识推理两种途径的可解释性框架. 框架演进目标特征结果和推理结果, 在二者相同且都较为可靠的情况下实现可解释性. 目标特征结果通过机器学习模型直接得到, 推理结果通过子特征分类结果结合规则进行知识推理得到, 两个结果是否可靠通过计算可信度来判断. 使用面向液基细胞学检查图像的融合学习与推理的某类宫颈癌细胞识别案例对框架进行验证, 实验表明, 该框架能够赋予模型的真实决策结果以可解释性, 并在迭代过程中提升了分类精度. 这帮助人们理解系统做出决策的逻辑, 以及更好地了解结果可能失败的原因.

关键词: 可解释性; 机器学习; 知识推理

引用格式: 李迪媛,康达周.融合机器学习与知识推理的可解释性框架.计算机系统应用,2021,30(7):22–31. http://www.c-s-a.org.cn/1003-3254/7963.html

Interpretable Framework for Integrating Machine Learning and Knowledge Reasoning

LI Di-Yuan[1], KANG Da-Zhou[2,3]

[1](College of Computer Science and Technology/College of Artificial Intelligence, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China)
[2](Key Laboratory of Safety-Critical Software, Ministry of Industry and Information Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China)
[3](Collaborative Innovation Center of Novel Software Technology and Industrialization, Nanjing 210023, China)

Abstract: Because the rules of the rule-based interpretability model may fail to reflect the exact decision-making situation of the model, the interpretability framework combining machine learning and knowledge reasoning is proposed. The framework evolves a target-feature result and a reasoning result, which implements interpretability when the two are the same and both reliable. The target-feature result is obtained directly by the machine learning model, while the reasoning result is acquired by sub-feature classification combined with rules for knowledge reasoning. Whether the two results are reliable is judged by calculating their credibility. A particular recognition case of cervical cancer cells for TCT image fusion learning and reasoning is used to verify the framework. Experiments demonstrate that the framework make model's real decisions interpretable and improve classification accuracy during iteration. This helps people understand the logic of the system's decision-making and the reason for its failure.

Key words: interpretability; machine learning; knowledge reasoning

① 基金项目: 十三五装备预研项目 (41402020501, 41402020101)
Foundation item: Pre-research Project of Equipment Foundation of China during 13th Five-Year Plan (41402020501, 41402020101)
收稿时间: 2020-10-21; 修改时间: 2020-11-18; 采用时间: 2020-11-24; csa 在线出版时间: 2021-06-30

机器学习[1] 是计算机基于数据进行和改进预测或行为的一组方法[2], 在效率、规模、可重复性等方面相较人类更加出色. 因此, 利用机器学习技术可以解决现实中很多领域的问题, 如自动驾驶[3,4]、医疗诊断[5,6]、自然语言处理[7] 等.

在很多重要领域, 机器学习结果对最终决策具有重大影响. 例如, 使用机器学习技术实现的宫颈癌细胞图像自动识别系统, 其识别结果能够辅助医师诊断宫颈癌, 这不仅大幅度降低了人工成本, 还提高了识别效率. 然而, 机器学习模型作为缺少可解释性的黑盒, 人们不理解它为什么会做出某种特定的决策, 其输出结果不能让人完全信任. 比如说, 医师很难信任缺少可解释性的宫颈癌细胞图像自动识别系统的结果. 因此, 赋予机器学习系统可解释性非常重要.

可解释性是人们能够理解决策的方法[8]. 在机器学习系统的上下文中, 它是向人类解释或以可理解的术语呈现的能力[9]. 从本质上讲, 可解释性是人类和决策模型之间的接口, 它既是模型的准确代理, 又是人类可以理解的[10]. 可解释性能够让人类明白系统做出决策的逻辑, 还可以帮助人们更好地了解结果可能失败的原因.

机器学习可解释性分为本质可解释性和事后可解释性两类. 本质可解释性意味着机器学习模型自身具有可解释性, 一般在模型较为简单时实现, 例如线性回归模型, 它将目标预测为特征输入的加权和, 所学到的线性关系使解释变得容易[11]; 决策树模型通过遍历决策树的节点 (类别和属性)、根节点到叶子节点的路径 (决策规则), 提供对简单模型决策过程的模拟实现[12]. 事后可解释性是利用可解释性技术来解释复杂机器学习模型, 例如基于个体条件图的可解释性模型, 它为每个实例显示一条线, 该线显示了特征更改时实例的预测如何改变[13]; 基于规则的可解释性模型通过从受训模型中提取解释规则的方式, 提供对复杂模型尤其是黑盒模型的整体决策的逻辑的理解[14]. 该模型旨在以人类可理解的规则对模型做出特定决策的逻辑进行解释, 但是当它的规则或决策出现错误时, 可解释性将无法反映模型的真实决策情况. 为了解决这个问题, 可以思考一下人类是如何进行决策以及对决策结果进行解释的.

很多情况下, 人类利用感知和推理共同完成决策[15]. 比如说, 医师在判断一个宫颈细胞是否发生病变时, 他首先能够根据自己的筛查经验, 对细胞图像展现出来的细胞整体特征进行感知, 得出一个大致的结论. 然后, 医师基于宫颈细胞病变相关的医学知识, 对细胞的每个细胞形态学特征 (例如细胞核大小、核质比高低等) 进行观察, 并结合这些特征和知识推理出另一个结论. 医师会结合、对比两个结论, 得出最终的诊断结果, 并使用相关的医学知识来解释得出此诊断结果的原因. 整个过程如图 1 所示.

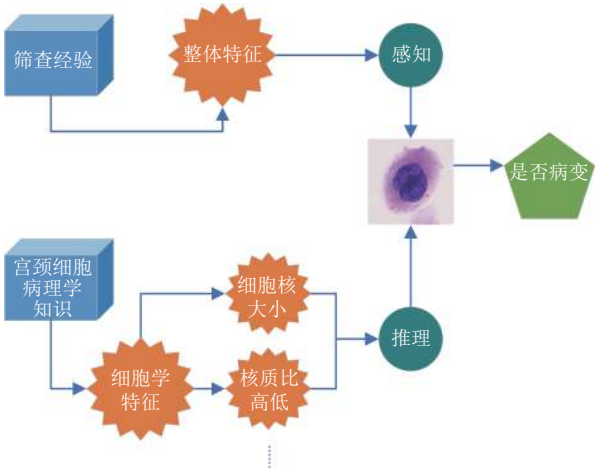


图 1 医师判断宫颈细胞是否病变

基于上述思路, 本文提出了一种融合机器学习和知识推理两种途径的可解释性框架. 它包含两个结果, 一个是由机器学习模型感知整体特征得到的目标特征结果, 另一个是通过结合多个子特征结果和规则进行知识推理得到的推理结果. 框架结合两个结果, 根据它们是否相同、分别是否可靠的不同情况, 来做出相应不同的演进决策. 这使得框架在训练和测试过程中不断提高结果的分类准确率, 同时赋予真实决策结果以可解释性, 在很大程度上解决了机器学习模型缺少可解释性的问题. 本文为衡量推理结果是否可靠, 提出了一种评估方法, 它融合了多个机器学习结果和规则的参数.

本文使用面向液基细胞学检查图像的融合学习与推理的某类宫颈癌细胞识别这一案例, 对融合机器学习和知识推理两种途径的可解释性框架进行了说明和验证.

1 融合机器学习与知识推理的可解释性框架

本文提出的融合机器学习与知识推理的可解释框架, 包含知识推理模块、机器学习模块、知识推理融合机器学习模块, 如表 1 所示. 该框架的示意图如图 2 所示.

其中, 决策目标是指一个系统预期达到的目的, 目标特征是指决策目标的整体特征, 目标特征具有多个

子特征, 它们是专家知识和数据之间的关联特征. 例如, 对于上文提到的诊断宫颈细胞是否病变的例子, 决策目标是识别宫颈细胞图像是否展现出癌变细胞的特征, 目标特征是宫颈细胞的整体特征, 而子特征是细胞形态学特征 (细胞核大小、核质比高低等).

表 1 框架模块表

模块	功能
知识推理模块	构建与决策目标有关的本体库和规则库.
机器学习模块	构建包括一个目标特征分类器和多个子特征分类器的分类器组.
知识推理融合机器学习模块	提取子特征; 支持机器学习结果的知识推理; 演进推理结果和机器学习结果.

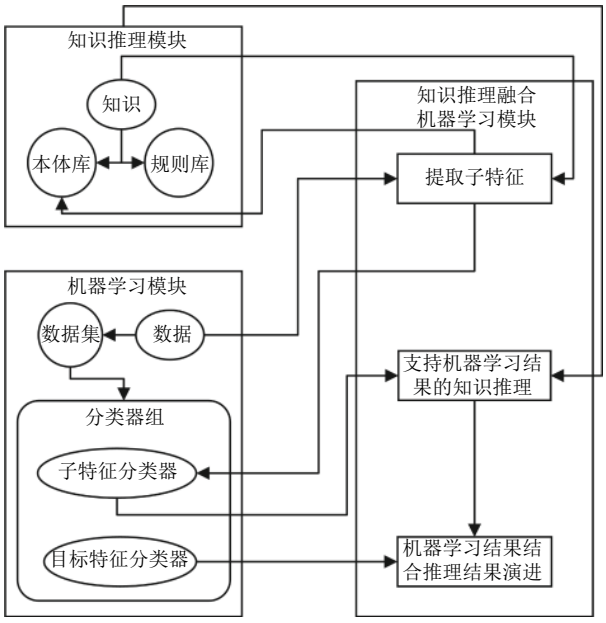


图 2 融合机器学习与知识推理的可解释性框架示意图

1.1 知识推理模块

知识推理模块提供了用于推理决策的领域知识和业务规则, 即决策目标相关的本体库 O 和规则库 K. 根据决策目标相关的领域知识, 通过知识抽取、融合、加工的步骤, 构建用于决策目标的本体库 O, 它表达了与决策目标有关的类和类之间的关系. 本体库 O 支持网络本体语言 (Ontology Web Language, OWL), 其中目标特征类包含子特征类. 将获取到的有关决策目标的专家知识转化为业务规则, 组成规则库 K, 它支持语义网规则语言 (Semantic Web Rule Language, SWRL). 知识推理模块的示意图如图 3 所示.

1.2 机器学习模块

机器学习模块提供了包含一个目标特征分类器 C 和多个子特征分类器 C_1~C_n 的分类器组, 其结果用于推理决策和结果演进. 分类器组通过神经网络组结合数据集 D、D_1~D_n 训练得到. 神经网络组由一个目标特征分类神经网络 N 和 n 个子特征分类神经网络 N_1~N_n 组成. 数据集 D 用于训练 N, D 的数据标注以决策目标为分类标准; 数据集 D_1~D_n 分别用于训练 N_1~N_n, D_1~D_n 的数据标注分别以它们对应的子特征为分类标准. 机器学习模块的示意图如图 4 所示.

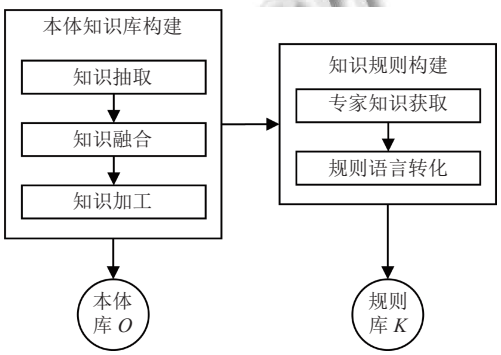


图 3 知识推理模块示意图

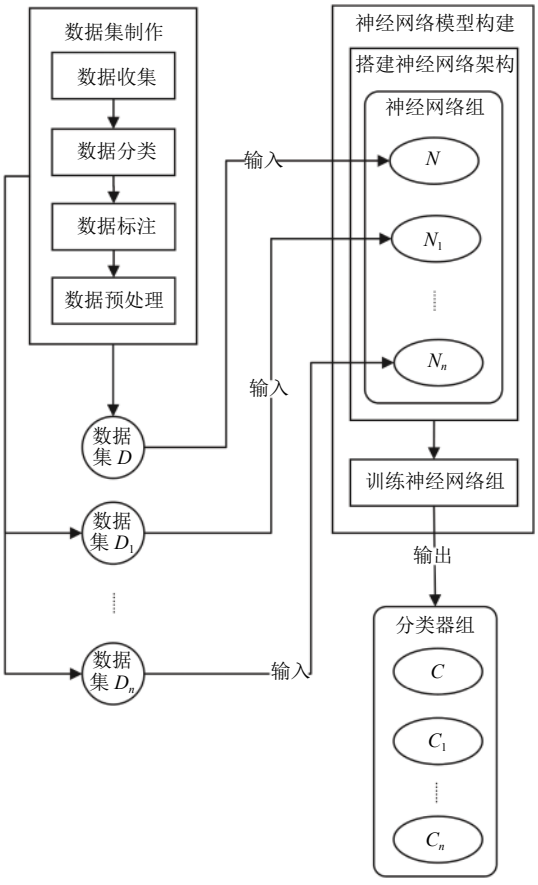


图 4 机器学习模块示意图

1.3 知识推理融合机器学习模块

(1) 提取子特征

基于提取的子特征, 知识推理模块中本体库 O 构建子特征类, 机器学习模块构建子特征分类器. 子特征是与决策目标有关的专家知识和数据之间的关联特征. 其中, 专家知识中高频提及的概念为知识特征, 数据本身的特征为数据特征, 将知识特征与数据特征进行关联和对应, 所重合的特征为子特征. 根据决策目标相关的知识特征和数据特征, 框架提取出 n 个子特征 f_1~f_n. 那么, 第 1.1 节的本体库 O 的子特征类依据 f_1~f_n 构建; 第 1.2 节的数据集 D_1~D_n, 其标注类别分别以 f_1~f_n 为分类标准, 所构建的 n 个子特征分类器 C_1~C_n 也分别以 f_1~f_n 为标准来分类待分类数据.

(2) 支持机器学习结果的知识推理

一个待分类数据 t 经过分类器组, 得到目标特征分类器 C 的分类结果 R_c、子特征分类器 C_1~C_n 的分类结果 R_1~R_n. 将 R_1~R_n 分别映射为本体库 O 中其对应的子特征类的实体数据, 并基于本体库 O 和规则库 K 进行知识推理, 得到推理结果 R_r. 结果 R_c、R_r 都为目标特征结果, 即框架做出数据 t 为 R_c、R_r 的决策. 后续将对两个目标特征结果 R_c 和 R_r 进行演进, 实现结果的可解释性.

(3) 机器学习结果结合推理结果演进

结合目标特征结果 R_c (机器学习结果) 和目标特征结果 R_r (推理结果) 进行演进, 框架根据 R_c 和 R_r 是否相同、R_c 和 R_r 分别是否可靠的不同情况, 做出相应不同的决策. 为衡量结果是否可靠, 本文引进评估结果好坏的指标——可信度. 分别计算目标特征结果 R_c 的可信度 A_{Rc} 和目标特征结果 R_r 的可信度 A_{Rr}, 然后结合两个结果进行演进, 具体情况和每种情况对应的决策如表 2 所示.

表 2 决策表

判别式	R_c和R_r相同	R_c和R_r不同
$A_{Rc}>a$ & $A_{Rr}>a$	框架做出数据t为R_c (R_r)的决策, 并解释结果.	人工选择正确的结果, 优化得出错误结果过程中相关的分类器和规则.
$A_{Rc}>a$ & $A_{Rr}\leq a$	框架做出数据t为R_c的决策, 并更加信任得到R_r过程中的分类器C_1~C_n和规则库K.	框架做出数据t为R_c的决策, 优化得到R_r过程中分类器C_1~C_n和规则库K.
$A_{Rc}\leq a$ & $A_{Rr}>a$	框架做出数据t为R_r的决策, 并更加信任得到R_c的分类器C.	框架做出数据t为R_r的决策, 优化得到R_c的分类器C.
$A_{Rc}\leq a$ & $A_{Rr}\leq a$	框架做出无效结果的决策, 但更加信任得到R_c的分类器C以及得到R_r过程中的分类器C_1~C_n和规则库K.	框架做出无效结果的决策, 优化得到R_c的分类器C的参数以及得到R_r过程中的分类器C_1~C_n和规则库K.

表 2 中, a 为阈值, 它由具体案例所属领域的专家或多次案例实验结果决定, 案例对结果精度要求越严格则阈值越高. 对于任何案例来说, 结果精度要求再低也不能差于在正反类中随机选择一类的情况, 精度要求再高也不可能好于类别全部预测正确的情况, 因此 a 的取值区间在 (0.5,1). 通过对比 A_{Rc}、A_{Rr} 和 a 之间的大小关系, 框架判断 R_c 和 R_r 是否可靠. 在 R_c 和 R_r 相同且两个结果的可信度都较高的情况下, 框架实现可解释性, 即使用子特征分类器 C_1~C_n 的分类结果 R_1~R_n、知识推理使用到的规则库 K 中的规则, 来解释框架做出数据 t 为 R_c (R_r) 的决策的逻辑. 在 R_c 和 R_r 相同且其中一个结果可信度较低的情况下, 框架提升得到低可信度结果的分类器和规则库的信任: 如果低可信度结果为 R_r, 则适当提高证据链中的参数值, 证据链在下一节中描述; 如果低可信度结果为 R_c, 则适当降低阈值 a. 在 R_c 和 R_r 不同且其中一个结果可信度较低的情况下, 框架将优化和改进得到低可信度结果过程中使用到的分类器、规则库.

本文规定可信度 A_{Rc} 为目标特征分类器 C 观察 R_c 的概率值 P 与分类器 C 在验证集上的准确率 Acc 的几何平均值; A_{Rr} 是融合了机器学习结果 R_1~R_n 的参数值 (例如结果概率值、灵敏度) 和规则库 K 的参数值的综合评估值, 具体计算方法在下一节描述. A_{Rc} 考虑了 R_c 本身的可信程度和得到 R_c 的分类器 C 的性能, A_{Rr} 考虑了推理 R_r 过程中使用到的数据 R_1~R_n 本身的可信程度、得到 R_1~R_n 的分类器 C_1~C_n的性能、规则库 K 的可靠度. 因此, 使用可信度衡量结果质量是较为充分的.

2 计算推理结果可信度

推理结果 R_r 是由支持机器学习结果的知识推理得到的. 知识推理过程中规则使用的实体数据, 都是由子特征分类器的结果映射而来, 不一定正确. 因此, R_r 也可能是不正确的. 那么, 如何对融合了多个机器学

习结果和规则的目标特征结果进行评估呢？本文提出了一个定义——结果证据链, 它是有向无环图数据结构. 结果证据链将记录得到 R_r 过程中一些重要的参数值. 本文基于结果证据链的结构, 自底向上地计算 R_r 的可信度 A_{Rr}, 以评估 R_r 是否可靠.

结果证据链是实现可解释性的另一关键部分, 它使得在 R_r 失败时可能追溯到具体的原因.

2.1 结果证据链

定义 1. 结果证据链. 结果证据链是一个有向无环图 (Directed Acyclic Graph, DAG), 记为三元组 $G=<V, E, F>$, 其中顶点集 V 为图中节点的非空集合; 边集 $E \subseteq V \times V$ 为图中有向边的集合, 每一条边用节点对表示为 (x, y), 称 x 为起点, y 为终点; F 是关系的集合, 每一个关系 $F(x, y)$ 对应一个节点对 (x, y) 之间的关系.

结果证据链的节点, 包括推理结果 R_r、子特征分类器结果 R_1~R_n、R_1~R_n 的相关参数、基于的规则库 K、K 的相关参数, 它们是 V 集合的组成元素. 其中, 子特征分类器结果 R_i 的相关参数, 包括子特征分类器 C_i 观察 R_i 的概率值 P_i、其在验证集上的灵敏度 (sensitivity) M_i 和特异度 (specificity) Y_i; 规则库 K 的相关参数, 是人为对 K 可靠性的评估值 K_r, 它的区间是 [0, 1].

本文使用知识图谱 KG 来表示结果证据链 G. 知识图谱通常用于表示和管理知识库[16], 采用三元组描述事实[17]. 本文采用自顶向下的方法建立 KG:

(1) 将推理结果 R_r 与子特征分类器结果 R_1~R_n 之间分别建立三元组$<R_r, F(R_r, R_i), R_i>$, $i=1,\cdots, n$;

(2) 将 R_r 与基于的规则库 K 之间建立三元组$<R_r, F(R_r, K), K>$;

(3) 将子特征分类器结果 R_i 与它的相关参数之间分别建立三元组$<R_i, F(R_i, P_i), P_i>$, $<R_i, F(R_i, M_i), M_i>$, $<R_i, F(R_i, Y_i), Y_i>$;

(4) 将规则库 K 与它的可靠性评估值 K_r 之间建立三元组$<K, F(K, K_r), K_r>$.

表示结果证据链 G 的知识图谱 KG 结构如图 5 所示. 那么, G 中存储了得到 R_r 过程中一些重要的参数值.

2.2 推理结果评估

计算 A_{Rr} 的方法使用了 DS 证据理论的思想, 它是一种不精确推理理论, 被广泛应用于证据 (数据) 合成方面. DS 证据理论首先设辨识框架 θ, 它包含了所有假设; 然后为每一个假设分配概率, 分配函数称为 Mass 函数; 最后基于 Dempster 规则融合结果, 即:

$$m(A)=\begin{cases}0, A=\varnothing \\ \dfrac{\sum_{A_i\cap B_j=A} m_1(A_i)m_2(B_j)}{S}, A\neq\varnothing\end{cases}$$

其中, $S=\sum_{A_i\cap B_j\neq\varnothing} m_1(A_i)m_2(B_j)$为归一化系数[18].

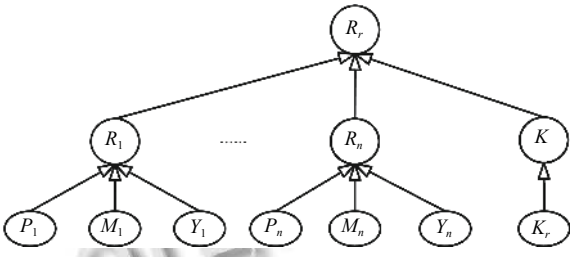


图 5 结果证据链 G 结构图

本文首先计算每个结果的灵敏度和特异度参数值. 假设真阳性数量为 TP, 假阳性数量为 FP, 真阴性数量为 TN, 假阴性数量为 FN, 灵敏度 M 和特异度 Y 的计算公式如下:

$$M=\frac{TP}{TP+FP}$$

$$Y=\frac{TN}{TN+FN}$$

即灵敏度为正确判断阳性的概率, 而特异度为正确判断阴性的概率. 然后, 对 A_{Rr} 进行计算:

(1) 定义 Map 函数来表示每个 R_i 与 R_i 相关参数之间的映射关系, 即 $P_i=m_1(R_i)$、$M_i=m_2(R_i)$、$Y_i=m_3(R_i)$;

(2) 求归一化系数 S:

$$S=\sum_{i=1}^{n}\left(\prod_{j=1}^{3} m_j(R_i)\right) \tag{1}$$

其中, n 为 R_i 的个数;

(3) 融合子特征分类结果 R_1~R_n 的参数, 计算机器学习部分的可信度 A_e:

$$A_e=\frac{1}{S}\sum_{i=1}^{n}\left(W_{fi}\cdot\sum_{j=1}^{3}\left(m_j(R_i)\right)^3\right) \tag{2}$$

其中, W_{fi} 为 R_i 对应子特征 f_i 的权重, 视具体案例而定;

(4) 融合机器学习部分的可信度 A_e、规则库 K 的评估值 K_r, 计算可信度 A_{Rr}:

$$A_{Rr}=\sqrt{A_e\cdot K_r} \tag{3}$$

后续将通过面向液基细胞学检查图像的融合学习与推理的某类宫颈癌细胞识别这一例子, 对可解释框

架进行具体地说明.

3 面向液基细胞学检查图像的融合学习与推理的某类宫颈癌细胞识别

宫颈癌是一个严重的健康问题, 全世界每年有近 50 万妇女患此病[19]. 宫颈癌筛查对于早期预防有着非常重要的作用, 而宫颈鳞状上皮异常对于宫颈癌的诊断有重大意义[20].

3.1 子特征提取

根据宫颈鳞状上皮细胞图像和 ASC-H 细胞形态学的专家知识, 本文提取出了 4 个子特征 f_1~f_4, 如表 3 所示.

表 3 ASC-H 细胞子特征表

子特征名称	子特征符号
细胞大小	f_1
核质比高低	f_2
细胞核大小	f_3
细胞核深染程度	f_4

本文选择对宫颈鳞状上皮异常中的非典型鳞状细胞-不除外高度鳞状上皮内病变 (Atypical Squamous Cells: cannot exclude High-grade squamous intra-epithelial lesion, ASC-H) 细胞进行识别, 以验证可解释框架的可行性. ASC-H 细胞识别框架在识别精度上有所提升, 同时实现了识别结果的可解释性, 在医师使用该识别框架时, 能够根据框架给出的解释选择是否信任结果. 值得一提的是, 宫颈癌筛查的过程中应该避免假阴性, 即避免本来病变的细胞被认为是没有病变的情况. 因此, ASC-H 细胞识别框架将疑似 ASC-H 也作为识别的一类, 以避免漏掉病变细胞.

3.2 本体库和规则库构建

(1) 本体库 O

本文从有关 ASC-H 细胞形态方面的医学知识中抽取出识别 ASC-H 细胞的类和类之间的关系, 并使用 OWL 语言构建 ASC-H 细胞识别本体库 O, 构建平台为 Protégé. 该本体的类信息如表 4 所示.

表 4 中 Cell_size、N/C、Nucleus_size、Hyper-chromatic 为子特征类; ASC-H、Sus-ASC-H、Non-ASC-H 为目标特征类. ASC-H 细胞识别本体的属性信息如表 5 所示.

在 Protégé中为 4 个子特征类添加实例, 实例为每个子特征的类别. 为 Cell_size 添加实体: 中等细胞 (c_l)、小细胞 (c_s), 为 N/C 添加实体: 核质比高 (nc_l)、核质比低 (nc_s), 为 Nucleus_size 添加实体: 细胞核增大 (nu_l)、细胞核正常 (nu_s), 为 Hyper-chromatic 添加实体: 细胞核重度深染 (h_l)、细胞核轻度深染 (h_s). 通过上述步骤, ASC-H 细胞识别本体创建完成.

表 4 ASC-H 细胞识别本体库的类信息表

Class	代表的类
Cell	细胞
Nucleus	细胞核
Cell_size	细胞大小
N/C	核质比高低
Nucleus_size	细胞核大小
Hyperchromatic	细胞核深染程度
Squamous_epithelial_cell	宫颈鳞状上皮细胞
ASC-H	ASC-H
Sus-ASC-H	疑似ASC-H
Non-ASC-H	非ASC-H

表 5 ASC-H 细胞识别本体库的属性信息表

Property	代表的关系
hasProperty	具有属性
donot-hasProperty	不具有属性
is_part_of	是…的一部分

(2) 规则库 K

规则库 K 包括 4 个规则, 由 ASC-H 的细胞形态医学专家知识转化而来.

1) 规则 1. 细胞组成部分的性质也是细胞的性质. SWRL 规则如 rule1 所示:

rule1: is_part_of(?a,?b) ^ hasProperty(?a,?c) -> hasProperty(?b,?c)

对规则 1 的解析如下: is_part_of(?a,?b) 表示 a 是 b 的组成部分; hasProperty(?a,?c) 表示 a 具有 c 性质; hasProperty(?b,?c) 表示 b 具有 c 性质.

2) 规则 2. 细胞形态中, 小细胞、核质比高、细胞核增大、细胞核轻度深染全部符合, 则认为细胞是 ASC-H. SWRL 规则如 rule2 所示:

rule2: Squamous_epithelial_cell(?t) ^ hasProperty (?t, c_s) ^ hasProperty(?t, nc_l) ^ hasProperty(?t, nu_l) ^ hasProperty(?t, h_s) -> ASC-H(?t)

对规则 2 的解析如下: Squamous_epithelial_cell(?t) 表示 t 是 Squamous_epithelial_cell 类的一个实例; hasProperty(?t, c_s) 表示 t 具有小细胞的性质; hasProperty(?t, nc_l) 表示 t 具有核质比高的性质;

hasProperty(?t, nu_l) 表示 t 具有细胞核增大的性质; hasProperty(?t, h_s) 表示 t 具有轻度深染的性质; ASC-H(?t) 表示 t 为 ASC-H 细胞. 文章后面的 SWRL 规则都与规则 2 类似, 因此不再做详细解析.

3) 规则 3. 细胞形态中, 核质比高、细胞核增大有任意一项符合, 则认为细胞是疑似 ASC-H. SWRL 规则如 rule3a、3b 所示:

rule3a: Squamous_epithelial_cell(?t) ^ hasProperty (?t, nc_l) -> Sus-ASC-H(?t)

rule3b: Squamous_epithelial_cell(?t) ^ hasProperty (?t, nu_l) -> Sus-ASC-H(?t)

4) 规则 4. 细胞形态中, 小细胞、核质比高、细胞核增大、细胞核轻度深染全部不符合, 则认为细胞不是 ASC-H. SWRL 规则如 rule4 所示:

rule4: Squamous_epithelial_cell(?t) ^ donot-hasProperty(?t, c_s) ^ donot-hasProperty(?t, nc_l) ^ donot-hasProperty(?t, nu_l) ^ donot-hasProperty(?t, h_s) -> Non-ASC-H(?t)

3.3 数据集和分类器组构建

(1) 数据集

数据集 D 和 D_1~D_4 都由数个大小为 128×128 的宫颈鳞状上皮细胞图像组成. D 的标注类别为细胞类型, D_1~D_4 的标注类别依据子特征 f_1~f_4, 如表 6 所示.

表 6 数据集标注类别表

数据集	标注类别
D	ASC-H Sus-ASC-H Non-ASC-H
D_1	中等细胞 小细胞
D_2	核质比高 核质比低
D_3	细胞核增大 细胞核正常
D_4	细胞核重度深染 细胞核轻度深染

(2) 分类器组

目标特征分类神经网络 N 的架构是任意用于分类的神经网络模型, 本文选取了自己搭建的卷积神经网络 (Convolutional Neural Networks, CNN)、变分自编码器 (Variational Auto-Encoder, VAE)、CNN 经典模型——VGG19 这 3 种模型分别实现 3 种目标特征分类器. 例如, 在使用 VAE 作为 N 的架构时, 其损失函数为 VAE 理论上的损失函数:

$$\mathcal{L}=\frac{1}{2}\sum_{i=1}^{j}\left(\mu_i^2+\sigma_i^2-\log\sigma_i^2-1\right)$$

其中, j 是隐变量的维度, μ、σ^2 是隐变量的变分概率分布的均值和方差. 使用数据集 D 对 N 进行训练: 当经过 30 000 步训练或 $\mathcal{L}$ 达到目标值时, 停止训练并保存当前模型. 该模型为目标特征分类器 C, 它将按照 ASC-H 细胞形态的整体特征的标准来分类细胞.

子特征分类神经网络 N_1~N_4 的架构是任意用于分类的神经网络模型, 本文均选用 CNN 实现 4 个子特征分类器. 子特征分类神经网络 N_1~N_4 均使用交叉熵作为损失函数:

$$\mathcal{L}=-\sum_{i=1}^{k}y_i\log y_i$$

其中, k 是分类的类别数量, y_i 为指示变量 (0 或 1), 如果该类别和样本 i 的类别相同就是 1, 否则是 0. 使用数据集 D_1~D_4 对 N_1~N_4 进行训练: 当经过 10 000 步训练或 $\mathcal{L}$ 达到目标值时, 停止训练并保存当前模型. 这 4 个模型为子特征分类器 C_1~C_4, 它们分别按照细胞大小、核质比高低、细胞核大小、细胞核染色程度的标准来分类细胞.

那么, 分类器组由 1 个目标特征分类器 C 和 4 个子特征分类器 C_1~C_4 组成.

3.4 支持机器学习结果的知识推理

假设将一个待识别细胞图像 t 输入到 ASC-H 细胞识别框架, 它经过分类器组后, 目标特征分类器 C 得到目标特征结果 R_c; 子特征分类器 C_1~C_4 得到 4 个子特征结果 R_1~R_4, 将 R_1~R_4 映射为 ASC-H 细胞识别本体库 O 中对应的子特征类 (Cell_size、N/C、Nucleus_size、Hyperchromatic) 的实体数据, 并基于本体库 O 和规则库 K 进行知识推理, 得到推理结果 R_r. 后续将对两个结果 R_c 和 R_r 进行演进, 实现结果的可解释性.

3.5 机器学习结果结合推理结果演进

(1) 计算可信度

计算目标特征结果 R_c 的可信度 A_{Rc}, 它为分类器 C 观察 R_c 的概率值 P 与 C 在验证集上的准确率 Acc 的几何平均值.

计算目标特征结果 R_r 的可信度 A_{Rr}, 首先要根据 2.1 节的方法构建目标特征结果 R_r 的结果证据链 G. 然后, 根据 2.2 节的计算方法, 基于结果证据链 G 的结

构, 自底向上地计算目标特征结果 R_r 的可信度 A_{Rr}. 对于本 ASC-H 细胞识别案例, 核质比高低、细胞核大小这两个子特征在提取的细胞特征中相对更为重要. 因此, 本文设置 f_1~f_4 的子特征权重值分别为: W_{f1}=0.2, W_{f2}=0.4, W_{f3}=0.3, W_{f4}=0.1.

(2) 分析处理结果

根据 R_c、R_r、A_{Rc}、A_{Rr} 的情况, 框架如 1.5 节所述的方法做出不同的演进决策. 多次实验表明, 将 ASC-H 细胞识别案例的阈值 a 取值为 0.8 最合适.

4 验证 ASC-H 细胞识别框架

验证集由 400 个大小为 128×128 的宫颈鳞状上皮细胞图像组成. 本文实现了 3.3 节中 3 种目标特征分类器. 为方便计算每种分类器的评估值, 将正类设为 Non-ASC-H, 负类设为 Sus-ASC-H 和 ASC-H 的总合. 在经过不同大小的数据集训练后, 每种分类器在验证集上的准确率、$F1$ 值如表 7 所示.

表 7 每种分类器的评估值表

分类器	数据集大小	准确率	$F1$
CNN	1000	0.8025	0.8124
	2000	0.8100	0.8137
	3000	0.8250	0.8069
VAE	1000	0.7525	0.7442
	2000	0.7675	0.7748
	3000	0.7925	0.8101
VGG19	1000	0.8375	0.8346
	2000	0.8450	0.8410
	3000	0.8400	0.8351

通过框架中支持机器学习结果的知识推理方法, 得到验证集里每个样本的推理结果, 并使用文中提出的方法, 将每个样本的目标特征分类器结果结合其推理结果进行演进. 每种目标特征分类器在结合了支持机器学习结果的知识推理方法进行演进后, 准确率和 $F1$ 值均有所提升, 如表 8 所示. 每种分类器在不同数据集大小下演进前后的准确率比较如图 6、图 7、图 8 所示.

实验表明, 文中提出的机器学习结果结合推理结果演进方法总会提升目标特征分类器的性能, 并且在分类器自身精度较低时, 提升的效果更加明显. 当分类器在训练过程中使用的数据量和自身精度都达到了比较饱和的程度时, 演进方法对于提升分类器性能方面的作用会较小. 通过结合推理结果, 演进方法总能将目标特征分类器的一部分错误结果剔除, 并且在不断地使用框架对细胞分类时, 演进过程也在进行迭代, 目标特征分类器将被持续优化.

本文使用 3 个具体的实例来验证 ASC-H 细胞识别框架. 该框架在应用于医学领域时, VAE 为最合适的目标特征分类器, 因为它在分类细胞的同时将细胞样本映射为空间分布, 便于医师在近似分布的细胞图像群中划分细胞类型. 因此, 选择数据集大小为 3000 的 VAE 分类器, 它在验证集上的准确率 Acc=0.7925.

表 8 每种分类器实现演进后的评估值表

分类器	数据集大小	准确率	$F1$
CNN	1000	0.8175	0.8267
	2000	0.8300	0.8333
	3000	0.8400	0.8416
VAE	1000	0.7825	0.7752
	2000	0.7975	0.8039
	3000	0.8175	0.8330
VGG19	1000	0.8475	0.8449
	2000	0.8550	0.8513
	3000	0.8550	0.8505

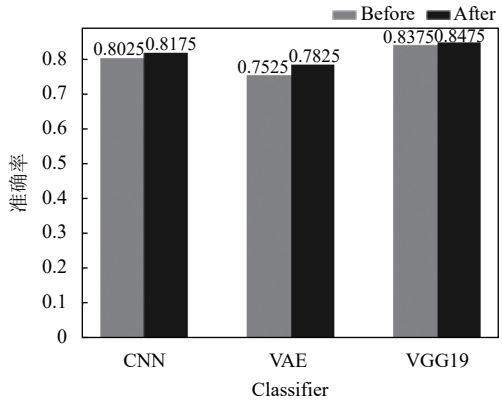


图 6 数据集大小为 1000 时演进前后准确率对比图

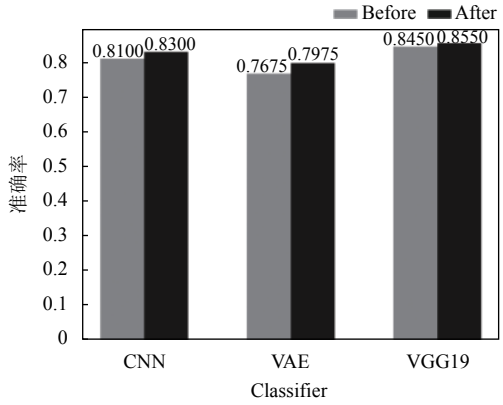


图 7 数据集大小为 2000 时演进前后准确率对比图

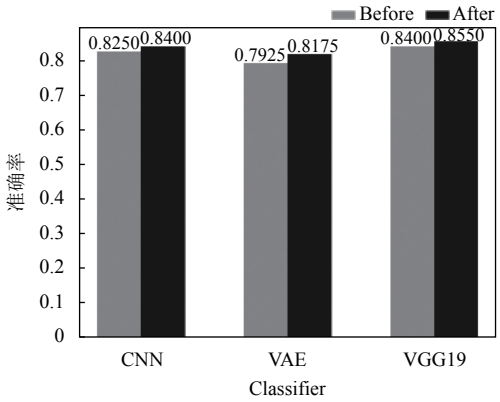
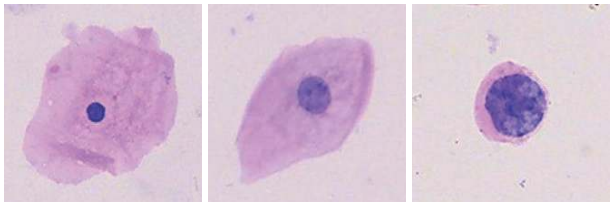


图 8 数据集大小为 3000 时演进前后准确率对比图

实例 a. 将一个待识别细胞图像输入到 ASC-H 细胞识别框架, 细胞图像如图 9(a) 所示. 它经过分类器组后, 分类器 C 得到目标特征结果 R_c 为 Non-ASC-H; 分类器 C_1~C_4 得到的 4 个子特征结果 R_1~R_4 分别为中等细胞、核质比低、细胞核正常、细胞核重度深染, 结合 R_1~R_4、ASC-H 细胞识别本体库 O、规则库 K 的 rule4 规则进行知识推理, 得到推理结果 R_r 为 Non-ASC-H.



(a) 实验 a (b) 实验 b (c) 实验 c

图 9 待分类细胞图像

R_c 的可信度 A_{Rc}=0.84. R_r 的结果证据链 G 的节点值如表 9 所示.

表 9 实例 a 中 R_r 结果证据链 G 的节点值表

i	P_i	M_i	Y_i
1	0.95	0.84	0.73
2	0.86	0.87	0.80
3	0.92	0.88	0.81
4	0.89	0.86	0.77

如第 2.1 节所述, 规则库 K 的参数 K_r 是人为对 K 的评估值. 在本实例中, K_r=0.75. 根据第 2.2 节的计算方法, 计算 R_r 的可信度 A_{Rr}:

1) 使用式 (1) 求归一化系数 S=2.43;

2) 使用式 (2) 求机器学习部分的可信度 A_e=0.77;

3) 使用式 (3) 得到 R_r 的可信度 A_{Rr}=0.76.

R_c 和 R_r 相同, R_c 的可信度 A_{Rc} 高于 0.8, R_r 的可信度 A_{Rr} 低于 0.8. 如 1.5 节所述, 框架做出细胞为 Non-ASC-H 的决策, 并更加信任得到 R_r 过程中的分类器 C_1~C_4 和规则库 K, 因此人工适当地提高结果证据链 G 中较低的参数值, 例如 Y_1、Y_4 和 K_r.

实例 b. 将一个待识别细胞图像输入到 ASC-H 细胞识别框架, 细胞图像如图 9(b) 所示. 框架得到 R_c 为 Sus-ASC-H; R_r 为 Non-ASC-H.

R_c 的可信度 A_{Rc}=0.87. 除 P_i 外, R_r 的结果证据链 G 的节点值因实例 a 人工修改 Y_1、Y_4 和 K_r 而产生变化, 如表 10 所示.

表 10 实例 b 中 R_r 结果证据链 G 的节点值表

i	P_i	M_i	Y_i
1	0.90	0.84	0.75
2	0.88	0.87	0.80
3	0.87	0.88	0.81
4	0.94	0.86	0.80

在本实例中, K_r=0.80. 根据相同计算方法, 得到 S=2.45、A_e=0.76、A_{Rr}=0.78.

R_c 和 R_r 不同, A_{Rc} 高于 0.8, A_{Rr} 低于 0.8. 因此, 框架做出细胞为 Sus-ASC-H 的决策, 并优化得到 R_r 的分类器和规则库的规则. 根据结果证据链 G 记录的参数值, 可以发现基于的规则库的可靠性评估值 K_r 不高, 即规则库可能存在错误; 分类器 C_1 的特异度较低, 即 C_1 正确判断中等细胞的概率偏低. 根据 R_r 失败的原因, 对规则库 K 的规则进行检查, 如有错误进行修正; 对 C_1 进行优化, 以提高框架的分类精度.

实例 c. 将一个待识别细胞图像输入到 ASC-H 细胞识别框架, 细胞图像如图 9(c) 所示. 框架得到 R_c 为 ASC-H; R_r 为 ASC-H.

R_c 的可信度 A_{Rc}=0.85. 除 P_i 外, R_r 的结果证据链 G 的节点值因实例 b 对规则库 K 和分类器 C_1 优化而产生变化, 如表 11 所示.

表 11 实例 c 中 R_r 结果证据链 G 的节点值表

i	P_i	M_i	Y_i
1	0.86	0.86	0.84
2	0.93	0.87	0.80
3	0.92	0.88	0.81
4	0.89	0.86	0.80

在本实例中, K_r=0.85. 根据相同计算方法, 得到 S=2.54、A_e=0.77、A_{Rr}=0.81.

R_c 和 R_r 相同, 且两个结果的可信度都高于 0.8, 框架认为 R_c 和 R_r 都较为可靠. 因此, 框架做出细胞为 ASC-H 的决策, 并使用子特征结果小细胞、核质比高、细胞核增大、轻度深染, 以及规则库 K 的 rule2 规则, 对细胞图像 t 为 ASC-H 这一决策进行解释说明. 因此, 医师可以理解框架将此细胞识别为 ASC-H 的逻辑, 并根据解释来决定是否相信该识别结果.

可以看出, 实例 b 通过结果证据链 G 找到了 R_r 失败的原因, 基于这些原因对相应的部分进行优化后, 实例 c 中框架的分类精度有所提升. 在两个结果都相同且较为可靠时, 框架赋予结果可解释性, 很大程度上解决了规则无法反映模型的真实决策情况的问题.

5 结论与展望

本文提出了一种融合机器学习和知识推理的可解释框架, 该框架在提升分类精度的同时, 实现了机器学习结果的可解释性. 通过面向液基细胞学检查图像的融合学习与推理的某类宫颈癌细胞识别方法对框架进行验证, 说明该方法可靠可行. 所提出的可解释框架对实现机器学习模型的可解释性具有一定参考意义.

参考文献

1 Mitchell TM. Machine learning. New York: McGraw-Hill, 1997.

2 Molnar C. Interpretable Machine Learning. North Carolina: Lulu Press, 2019.

3 Sallab AE, Abdou M, Perot E, *et al*. Deep reinforcement learning framework for autonomous driving. Electronic Imaging, 2017, 2017(19): 70–76. [doi: 10.2352/ISSN.2470-1173.2017.19.AVM-023]

4 Yang YQ, Wu Z, Xu QY, *et al*. Deep learning technique-based steering of autonomous car. International Journal of Computational Intelligence and Applications, 2018, 17(2): 1850006. [doi: 10.1142/S1469026818500062]

5 Gorantla R, Singh RK, Pandey R, *et al*. Cervical cancer diagnosis using CervixNet —A deep learning approach. Proceedings of 2019 IEEE 19th International Conference on Bioinformatics and Bioengineering. Athens, Greece. 2019. 397–404.

6 Zhang L, Lu L, Nogues I, *et al*. DeepPap: Deep convolutional networks for cervical cell classification. IEEE Journal of Biomedical and Health Informatics, 2017, 21(6): 1633–1643. [doi: 10.1109/JBHI.2017.2705583]

7 奚雪峰, 周国栋. 面向自然语言处理的深度学习研究. 自动化学报, 2016, 42(10): 1445–1465.

8 Miller T. Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence, 2019, 267: 1–38. [doi: 10.1016/j.artint.2018.07.007]

9 Doshi-Velez F, Kim B. Towards a rigorous science of interpretable machine learning. arXiv: 1702.08608, 2017.

10 Guidotti R, Monreale A, Ruggieri S, *et al*. A survey of methods for explaining black box models. ACM Computing Surveys, 2018, 51(5): 93.

11 Lipton ZC. The mythos of model interpretability. Communications of the ACM, 2018, 61(10): 36–43. [doi: 10.1145/3233231]

12 Deng HT. Interpreting tree ensembles with inTrees. International Journal of Data Science and Analytics, 2019, 7(4): 277–287. [doi: 10.1007/s41060-018-0144-8]

13 Goldstein A, Kapelner A, Bleich J, *et al*. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. Journal of Computational and Graphical Statistics, 2015, 24(1): 44–65. [doi: 10.1080/10618600.2014.907095]

14 纪守领, 李进锋, 杜天宇, 等. 机器学习模型可解释性方法、应用与安全研究综述. 计算机研究与发展, 2019, 56(10): 2071–2096. [doi: 10.7544/issn1000-1239.2019.20190540]

15 Dai WZ, Xu QL, Yu Y, *et al*. Tunneling neural perception and logic reasoning through abductive learning. arXiv: 1802.01173, 2018.

16 Song Q, Wu YH, Dong XL. Mining summaries for knowledge graph search. Proceedings of 2016 IEEE 16th International Conference on Data Mining. Barcelona, Spain. 2016. 1215–1220.

17 黄恒琪, 于娟, 廖晓, 等. 知识图谱研究综述. 计算机系统应用, 2019, 28(6): 1–12. [doi: 10.15888/j.cnki.csa.006915]

18 韩德强, 杨艺, 韩崇昭. DS 证据理论研究进展及相关问题探讨. 控制与决策, 2014, 29(1): 1–11.

19 Waggoner SE. Cervical cancer. The Lancet, 2003, 361(9376): 2217–2225. [doi: 10.1016/S0140-6736(03)13778-6]

20 劳芝英. HPV、TCT 及阴道镜对宫颈癌筛查的意义. 实用癌症杂志, 2014, 29(7): 826–828, 831. [doi: 10.3969/j.issn.1001-5930.2014.07.032]