差异与学习：模糊系统与模糊推理

GARIBALDI Jonathan M[1]，陈虹宇[2]，李小双[2]
（1. 诺丁汉大学计算机科学学院，英国 诺丁汉郡 NG8 1BB；2. 中国科学院自动化研究所复杂系统管理与控制国家重点实验室，北京 100190）

摘　要：作为一种决策支持系统，模糊系统不仅具有处理不确定性信息的能力，又能够明确表达不确定性知识和推理过程。但现存的一个问题是，对于包括采用模糊方法的系统在内的计算机决策支持系统，目前还未出现能够明确评估系统实际可行性的方法。提出了不可区分性的概念框架，并将其作为评估计算机决策支持系统的关键部分，给出了相关案例研究。案例证明人类专家的评判并非完美，模糊系统能够在技术层面模拟人类的决策，包括人类专家在评判时表现出的差异性。使用模糊方法进行基于知识不确定性的表达与推理是非常必要的，而差异则是学习时不可避免的表现形式，在评估人工智能系统时应接受其不完美的决策。

关键词：人工智能；近似推理；模糊推理系统；模糊集合；人类推理
中图分类号：TP18
文献标识码：A
doi: 10.11959/j.issn.2096−6652.201936

Variation and learning: fuzzy system and fuzzy inference

GARIBALDI Jonathan M[1], CHEN Hongyu[2], LI Xiaoshuang[2]
1. School of Computer Science, University of Nottingham, Nottinghamshire NG8 1BB, UK
2. The State Key Laboratory for Management and Control of Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China

Abstract: As a decision support system, fuzzy system can deal with uncertainty and has a clear representation of uncertainty knowledge and inference process. But one problem that exists is that computerized decision support systems, including systems that use fuzzy methods, do not have a clear assessment method to determine whether they can be allowed to be used in the real world. A conceptual framework of indistinguishable lines as a key component in evaluating computerized decision support systems was proposed, and some case studies were given. The case proves that the performance of human experts is not perfect, and the fuzzy system can simulate human performance at the technical level, including the variation of human experts. In summary, fuzzy methods are necessary for the representation and reasoning of uncertainty of the knowledge-based systems. Variation is an important form of learning. When evaluating AI systems, imperfect performance should be accepted.

Key words: artificial intelligence, approximate reasoning, fuzzy inference system, fuzzy set, human reasoning

1　引言

在几经发展的高峰和低谷后，人工智能（artificial intelligence，AI）再次走到全球计算机科学研究的最前沿[1]。超人类机器智能（即机器智能超越最佳人类表现）的首次出现是在 1997 年国际商业机器公司（international business machines corporation，IBM）的一个名为 Deep Blue 的国际象棋游戏计算机向加里·卡斯帕罗夫（持有官方最高记录的卫冕世界象棋冠军）发起挑战，并在经过 6 场激烈的比赛后以 32:22 将对手击败。2017 年，在 Deep Blue 出现的 20 年后，AlphaGo Master 在中国乌镇的 Future of Go Summit 峰会上以 3:0 击败围棋高手柯洁，再次展现了机器超人类的智能。这个击败柯洁的 AlphaGo Master 由谷歌 DeepMind 公司开发，采用结合蒙特·卡罗树搜索算法和卷积神经网络形

收稿日期：2019–09–30；修回日期：2019–11–22
通信作者：陈虹宇，chenhongyu2017@ia.ac.cn

式的深度学习算法，并通过强化学习方法进行模型训练[2]。显然，在游戏竞赛的场景下评价一个计算机硬件或其软件程序的性能非常简单，例如打败世界冠军就是很好的评价基准，但即便计算机硬件或其软件程序的性能效果能够超越人类智能，其核心算法技术是否达到足以投入实际应用的水准还是未知的。

虽然深度学习这类次级象征性方法取得的突出进展使其成为目前的流行算法，但其缺乏可解释性或决策能力，这为人工智能技术的发展带来了困扰。一个典型的例子就是神经网络的超参数对性能的影响。没有人能够解释清楚为何同样的神经网络结构在不同的参数组合下会得到不同甚至完全不一样的结果。另外，神经网络所提取的用于特定任务的特征或决策逻辑在语义层面是难以理解的，同样也没有人能够完美地诠释神经网络的特征与实际问题之间的联系。

针对特定情境，需要使用基于知识不确定性表示和推理的人工智能方法。基于这一论点，本文的一个主要创新点在于提出可将模糊技术作为一种能够提供必要解释和推理能力的方法，并通过在医学领域的实践证明了其有效性。此外，不完美推理和不完美性能是测试验证这类 AI 算法的基本特征。本文提出了不可区分性的概念框架，并将其作为计算机决策支持系统评估的关键组成部分，用来衡量模型与人类专家的差异，以提升专家系统的决策能力。

本文其余部分的内容如下：第 2 节概述了如何通过图灵测试的形式评估决策支持系统；第 3 节举例说明了推理结果的差异属于人类推理的基本特征，并提出将其纳入计算机专家系统的推理过程中以通过评估测试；第 4 节介绍了模糊专家系统中一些可用于融合差异的现有技术及其使用的相应影响和作用；第 5 节综合分析并讨论了差异与学习的潜在关系；最后概述了未来可能的研究方向并总结了本文的工作。

2　评估人工智能

如何恰当地评估计算机专家系统（或称专家系统）在给定条件下的性能水平？计算机专家系统的性能水平应该达到何种程度才足以投入实际应用？这是重要且关键的问题。

2.1　图灵测试

阿兰·图灵（Alan Turing）的开创性论文“计算机器和智能”[3]中介绍了一种被称为“模仿游戏”的测试，可对“机器能否思考”这一问题进行回答。在游戏的原始版本中（随后经图灵修订）有 2 个游戏主体，一个是男人（A），一个是女人（B），还有第三个人是性别审讯者（C），C 可以向 A 和 B 提出问题以试图识别二者的性别。A 的目标是使 C 做出错误的身份判定，而 B 的目标是帮助 C 进行判定。图灵建议使用计算机替代 A，并提出“游戏修改后，审讯者的判错概率是否与游戏修改前相同”这一问题替代原有的“机器能否思考”问题。

在进一步的阐述中，图灵舍弃了男性和女性的角色，并将游戏任务简化为判定机器和人（通常指仅判定单一性别的人）。模仿游戏的过程如图 1 所示，该过程随后被称为 AI 的“图灵测试”。图灵测试的基本原理即不可区分性。实质上，如果无法将计算机与人类区分开来，那么可认为计算机是智能的，至少与人类的智能程度相同。显然，需注意的是，图灵强调的不可区分性仅限于对话交流，而与物理形态等因素无关。

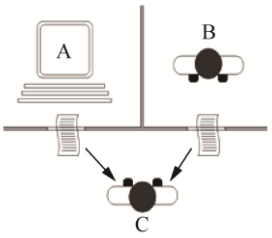

图 1　图灵测试的程式化表示

2.2　评估决策支持系统

那么，针对本节开头提出计算机专家系统的性能评估问题，可给出的相应解决方法为：计算机专家系统应通过与图灵测试相似的流程进行实验评估，即评估观察者是否可以通过向计算机专家系统和人类专家提问来对二者进行区分。如果无法区分计算机专家系统与人类专家，那么无论计算机专家系统达到的绝对性能水平如何，其都已展现出足够的专业知识。

当然，这个测试过程并不完全等同于图灵测试。实际上，由于图灵测试在具体实施中的效果相对较差，故图灵将最初提出的判定性别的测试方案修改为简单地将计算机与人类区分开来。后又进行补充，测试判定应由一个“陪审团”而非仅由一个观察者实施且多次重复。

因此，为使该方法更具体可行，应构建一个由计算机专家系统探究领域中具有挑战性的问题实例组成的测试方案。然后，将问题实例分别提交给计算机专家系统和专家小组（以获得该领域中人类专业知识的代表性答案），得到相应答案后基于统计学知识进行分析比较。如果计算机专家系统和人类专家所给出的答案之间不存在统计学上的差异，则认为计算机专家系统具有足够高的性能水平。

测试要求中的“具有挑战性的问题实例”足以说明测试使用的问题实例的难度水平可体现专家系统的专业水平。例如，如果测试仅向专家系统提出了非常简单的数学问题（例如“2 + 3=？”），那么仅能证明系统能够实施简单的数学运算而不具备数学家的水平（即便参与测试的人类专家小组人员本身就是数学家）。进一步拓展到针对某一组问题，在人类专家能够做出正确率达 100%的判定（即所得的所有答案始终正确）的情况下，计算机专家系统也应能够在同一组问题上具有 100%的准确率。但是，如果提出的一系列问题对于人类专家来说确实具有挑战性，那么人们可能会理所应当地认为专家无法答对所有问题，即其正确率可能会低于 100%。在这种情况下，计算机专家系统可能也只能达到相似的水平。因此，如果人类专家对于测试问题的回答的准确率不能达到 100%，那么就不应该认为该领域的计算机专家系统能够达到 100%，如果能够接受人类专家犯错，那么也应接受计算机专家系统犯错。

3　人类推理中的差异

正如图灵所说的“如果一台机器被认为是绝对正确的，那么它不可能是智能的”，人们不仅应该接受计算机专家系统犯错，还应能预料其表现出一些具有随机元素的行为或者被图灵称作的“部分随机”。虽然人们都喜欢将自己视为理性决策者，但人类确实会犯错。上文提出，计算机专家系统的性能测试应与一组人类专家比照进行。这是因为人类专家可能会表达不同的意见，因此需要一个专家组而非简单地选择个别专家来参与测试。许多情况下的专家间差异现象（不同人类专家之间意见或答案的差异）已被讨论和研究。下面将结合 2 个具体案例进行阐述。

第一个案例是将用于分娩心电图（cardiotocograph，CTG）分析的计算机专家系统与人类专家进行比较，以确定计算机专家系统的表现是否达到可接受的水平[4]。这项研究的研究人员为来自英国各地的 17 位人类专家，该研究分析了他们对具有 CTG 数据的 50 例新生婴儿病例的评判结果，并将决策结果与计算机专家系统给出的结果进行了比较。计算了不同人类专家之间所得结果的一致性（称为“专家间一致性”或“一致性”），发现这一数据为 60％～75%，也就是说，在三分之二的情况下人类专家意见相同。计算机专家系统的性能在一致性方面与人类专家并无明显差异，其与人类专家的一致性约为 68%。在间隔一个月后进行了第二次决策实验，使人类专家能够对自己的评判结果进行比较。实验发现人类专家与自身所得结果的一致性（称为“专家内部一致性”或“一致性”）为 75%～90%。有趣的是，虽然计算机专家系统的专业性比 99%以上的人类专家高得多，但实际上其表现的一致性并非 100%。这是由于该系统具有操作员干预这一部分因素，并且在单一病例中用户输入的差异很小。

这种形式的图灵测试评估是在专家数据健康（Expert DataCare）系统上进行的，这项评估研究招募了 6 名脐酸碱分析专家参加，选择了 50 个具有脐带酸结果且富有挑战性的病例，其中大多数婴儿在出生时出现了不良状况。研究要求人类专家根据婴儿缺氧（出生窒息）的严重程度，对 50 例患者从第 1 例（最差结果）到第 50 例（最佳结果）进行评判和排序。同时，使用 Expert DataCare 系统评估相同的结果，并使用其输出数值对病例进行排序[5]。

结果如图 2（a）所示。在该图中，x 轴显示由 Expert DataCare 系统给出的排序结果，而由每位人类专家给出的同一婴儿的排序结果显示在 y 轴上。例如，x 轴上的位置 1 表示 Expert DataCare 系统根据血气结果排列得出的状况最差的婴儿（即具有最严重窒息情况的婴儿），而 6 位专家分别在该 x 轴排序结果处的 y 轴上绘制圆圈。图 2（a）中，在 x 轴上的位置 1 处，有 6 个圆圈叠加在对应 y 轴的位置 1 处，这说明专家系统和所有 6 位专家均将这个病例标定为婴儿 1，即状况最严重的婴儿。在 x 轴的位置 2 处，有 4 个圆圈叠加在对应 y 轴的位置 2 处，而在位置 4 处和位置 8 处各有一个圆圈。在 x 轴的位置 12 处，所对应的 6 个圆圈分别在 y=9、10、11、14、15 和 16 的位置处。所有 6 位专家和 Expert DataCare 之间的斯皮尔曼等级排列相关性实际上是 0.95，这表示其总体一致性非常高。然而，对于许多病例的状况判定显然存在意见分歧。

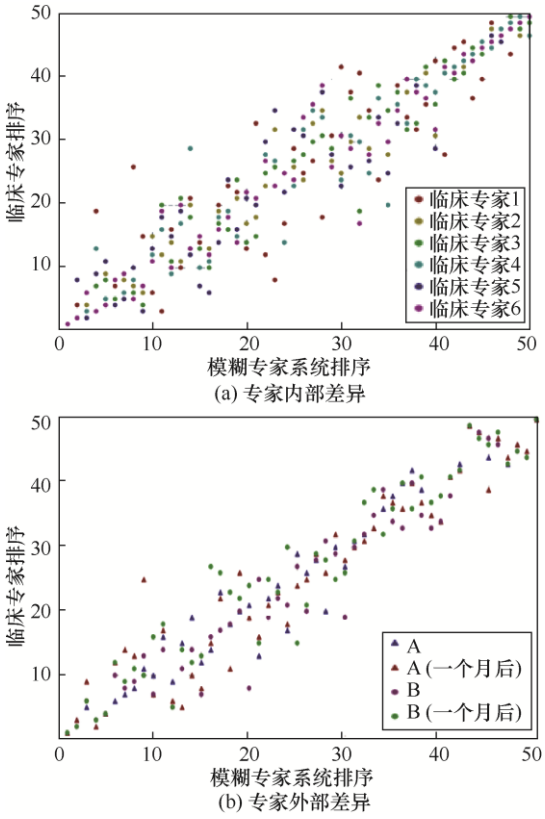

图 2　脐酸碱分析中专家内部及外部差异

在初始实验一个月后，由 6 位临床专家中的 2 位自愿重复进行该实验。同样的 50 例病例被打乱顺序后重新递交给 2 位临床专家。这 2 位临床专家对自己及彼此的原始排序结果都是未知的，并且每次都分别独立对这些病例进行排序，该重复研究的结果如图 2（b）所示。该图的表达方式与图 2（a）相同，x 轴仍然记录 Expert DataCare 系统给出的排序，但不同的是，此次只有 2 位专家参与研究，且分别都进行了 2 次测试。可以看出，专家 A 和专家 B 都将婴儿 1（即 Expert DataCare 系统标记为情况最严重的婴儿）排列为顺序 1。但同时，也可以清楚地看到标记的纵向分离（专家 A 的三角形和专家 B 的圆形）。例如婴儿 8，专家 A 在 2 次测试中分别将其标记为 8 和 12，而专家 B 则分别标记为 9 和 11，即存在明显的专家内部差异。

对图 2（b）进一步观察发现，尽管专家内部也出现了许多差异（实际上，2 位专家在 2 次重复的实验中对约一半的婴儿都给出了不同的答案），但是仍有一些病例的标记前后完全一致。例如，对于病例 1，2 位专家在前后 2 次实验中给出了相同标记，同样，病例 50 每次也均被标记为顺序 50。或许有理由认为观察到的变化或差异与所选病例或问题实例的难度有关。在本研究中，病例 1 的婴儿在分娩结束后死亡，这是唯一死亡的婴儿。从这个角度来看，这个特定的个体情况似乎最严重，因此专家们的标记一致，包括专家 A 和专家 B 在前后 2 次实验中的标记也都证明了这个结论。换句话说，病例或问题越难以解释或解决，可能出现的变化或差异就越多。

然而，对于前文提到的 CTG 专家系统，虽然 Expert DataCare 系统在整体性能及与其他专家的一致性方面几乎没有区别，但在自身一致性方面却表现不同。实际上，对于给定的一组固定输入，人类专家自身存在明显差异，而 Expert DataCare 系统的输出则是完全确定的，能够在实验中表现出零差异性（即自身 100%一致）。

4　建模与度量变化

对于能够通过人类专家不可区分性测试的计算机专家系统来说，加入决策行为的变化是有必要的，问题在于如何引入变化以及能否因此获益。对此本文实施了一项研究，验证能否成功对图 2（a）和图 2（b）中观察到的专家内、外部差异进行建模以评估计算机专家系统的性能以及建模是否会对性能产生影响。

4.1　非平稳模糊集合

用于模拟“中等”身高概念的示例标准模糊集如图 3 所示，x 轴变量与中等身高集合的隶属度之间非线性映射，此映射不存在模糊性。在某种意义上，这种映射应该是精确的，尽管在常规理解下，这一概念并不准确。Zadeh 认识到了标准（type-1）模糊集定义中的这一明显矛盾，并通过引入 type-2 模糊集[6]来解决，其中 x 轴每个值对应的隶属度由 type-1 模糊集给出。

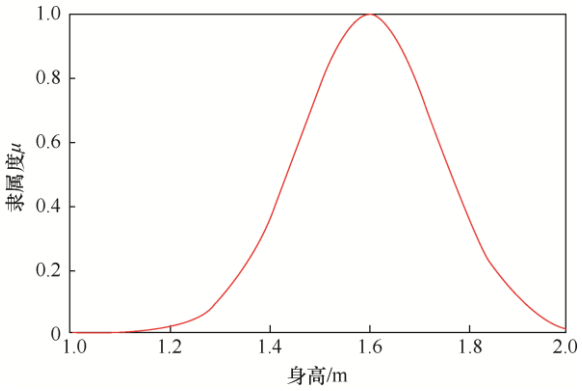

图 3　模拟“中等”身高概念的标准（type-1）模糊集

然而，传统的 type-2 模糊集虽然“模糊”了隶属函数，却没有明确地表示推理的可变性。作为一个思维实验，要求不同的人在 x 轴上分别标定“中等”身高，这使得集合标定位置的观点总是不同的，如图 4 所示。

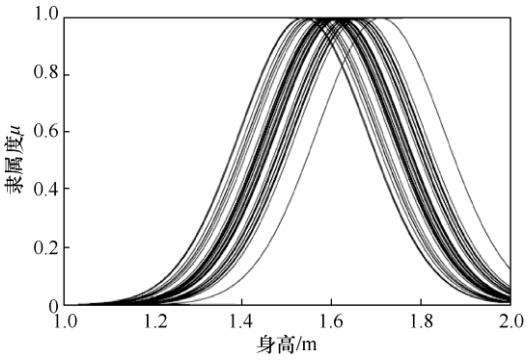

图 4　“中等”身高集合可能的位置

4.2　对脐酸碱评价中的差异建模

如图 2（a）和图 2（b）所示，使用非平稳模糊集的机制描述在脐酸碱评估中观察到的专家内部和外部差异。实验观察 Expert DataCare 系统中模糊集在具有指定变化量（从 1%到 10%）时婴儿的最终排序差异。本实验由 2 个独立的试验组成：一个是找到与观察到的专家间差异的最佳匹配变化量，另一个是找到专家内变化的最佳匹配变化量。这些实验的结果如图 5 所示，其中，在与 x 轴上每个位置对应的 y 轴上绘制的误差线条显示了由包含可变性的 Expert DataCare 系统获取的最小和最大的评估序列。实验发现，当 Expert DataCare 系统基础模糊集移动变化量为总体的 3%时，可获得专家间观点的最佳匹配，如图 5（a）所示。实验还发现，同样基础模糊集 3%的移动变化量也最符合专家内观点的最佳匹配，如图 5（b）所示。

这些实验代表了第一次将包含在模糊集中的差异用于模拟人类推理中观察到的变化。实验可获得 2 个主要结论：一是该方法不仅可用于差异模拟，还可将观察到的观点变化进行量化。实际上，3%的变化（在 x 轴上的位置）能够对内部变化和外部差异进行最佳匹配。可以说，在这一特定背景下观察到的专家内部变化和专家外部差异均为 3%；二是这种变化与将随机的“错误”添加到系统中完全不同。可以看出，具有 3%变化的系统仍对标记在位置 1 的婴儿（情况最严重的婴儿）做出与人类专家相同的判定，且可观察到分布在中间部分的婴儿具有更多的变化或差异。

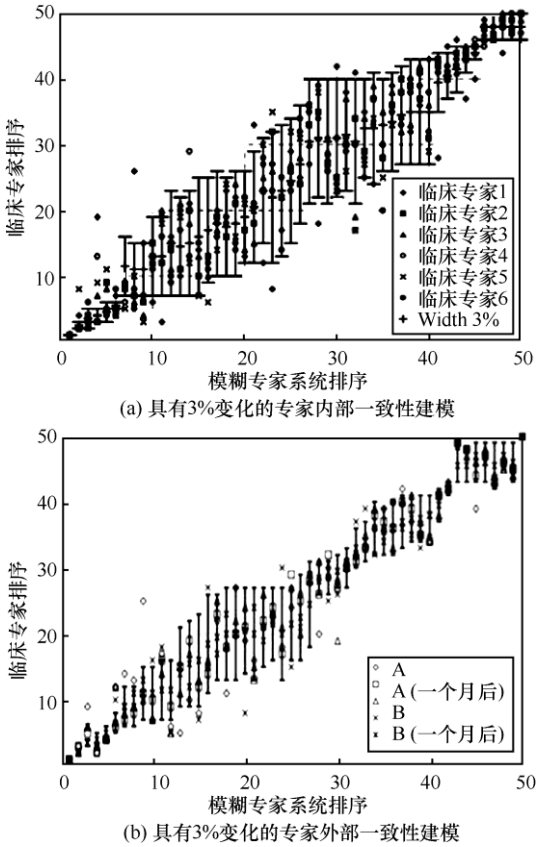

图 5　最佳非平稳差异建模专家内/外部变化

总之，非平稳模糊集的使用使得对人类决策的变化进行建模和量化成为可能。此外，也证明了对构成系统的基础模糊集作一些改变（其基本与决策中所用术语的基本含义对应），能够在最终决策中得到有趣的变化模式。但不同于随机决策，这种改变可与计算机专家系统结合，以模拟人类的变化或差异。

4.3　差异对性能的影响

对决策的差异进行建模及量化后，接下来的问题是这种差异能否以某种方式影响决策的制定。面对毋庸置疑的事实，人类专家的观点通常并不相同，一种常见的策略是咨询至少一位专家的意见（图灵称之为“陪审团”），并以某种方式将所获观点（可能不相同）整合为整体的共识。目前已有多种方法能够使得在多方观点不相同时达成共识，如取平均值（需数字类型的输出结果）、多数投票（可用于分类决策）或其他方法。

对此，本文又进行了另一组实验，以观察模糊专家系统中非平稳模糊集的使用对性能的影响。该实验的实施需针对一个不同的决策问题，其可获得某种形式的目标输出（“正确”或“客观”的决策），以使模糊系统的性能可与之相比，最终选择的决策

问题是乳腺癌治疗适用形式。

如果女性被诊断为患有乳腺癌，则最初的治疗相对简单，即在大多数情况下立即进行乳房肿瘤切除手术，然而后续治疗（医学术语中称为“辅助治疗”）却存在多种不同的选择方案。现代医疗护理具有多种选择，包括药物治疗、放射治疗和化学治疗。近年来，化学治疗的使用仅限于最严重且存在高风险的乳腺癌，因为虽然化学治疗可能效果显著，但其具有高毒性且可能带来明显的副作用。诺丁汉大学医院英国国家医疗服务体系（National Health Service，NHS）信托基金使用的化学治疗临床方案如图 6 所示。

图 6 所示方案多年来一直被诺丁汉大学医院 NHS 信托基金使用，为乳腺癌术后患者所需的后续治疗给出建议，并提供了一组涉及 1 300 多名妇女及其化学治疗方案决策的数据。该临床方案的决策规则被运用于模糊专家系统的实施，仅针对是否使用化学疗法进行决策（即此时忽略其他疗法的使用）。模糊规则库具体如图 7 所示，大致等同于提供化疗建议的临床方案。输入和输出变量中的模糊集由相关专家和领域知识获得。例如，诺丁汉愈后指数（NPI）分为 4 个等级：低、中低、中高和高，以对应图 6 临床方案中的 4 个主要决策类别。此外，构造模糊集的隶属函数使得在临床方案中指定的边界处存在交叉点，例如低集合和中低集合之间的交叉发生在数值 3.0 和 3.1 之间。

在构建了基于该临床方案的模糊专家系统之后，该系统进行了一些简单但彻底的优化确定“最佳”实验设置，以便与实际临床实践中做出的决策达成一致，从而使其性能最优[7]。也就是说，最大化模糊专家系统的化学治疗建议与给予真实患者的实际建议之间具有一致性。

随后进行了一组实验以将基础的 type-1 模糊专家系统与采用非平稳模糊集的替代系统进行比较。实验设置如下：在每种情况下，均创建一个包含非平稳模糊集的比较系统，其中模糊集的位置按论域（如图 4 所示）的固定百分比变化，这个变化百分比为论域的 1%～10%（x 轴）。非平稳模糊集专家系统在每种情况下运行 30 次，获得相应的输出。然后使用多数投票来确定系统的最终建议。所得结果展示在图 8 中，其中系统中存在的变化显示在 x 轴上，同时其与临床实践一致的决策数量显示在 y 轴上。

诺丁汉愈后指数<3.0	无需采取辅助治疗
诺丁汉愈后指数：3.1~3.4 雌激素受体实验结果为阳性 雌激素受体实验结果为阳性	 建议采取激素治疗 若血管浸润测试结果显示血管腔内明确存在肿瘤，建议采取化学治疗
诺丁汉愈后指数：3.4~4.4 雌激素受体实验结果为阳性 雌激素受体实验结果为阳性	 建议采取激素治疗 建议采取化学治疗
诺丁汉愈后指数>4.4 雌激素受体实验结果为阳性 雌激素受体实验结果为阳性	 考虑以下因素，讨论决定是否采取化学治疗： 建议采取化学治疗 1. 年龄<40 2. 血管浸润测试结果显示血管腔内明确存在肿瘤 3. 人类表皮生长因子受体为阳性 4. 雌激素受体实验结果为弱，即<100/300 不建议采取化学治疗 1. 年龄>60 2. 阳性淋巴结数目仅为 1 3. 特殊类型肿瘤 建议采取化学治疗

图 6　诺丁汉大学医院 NHS 信托基金使用的化学治疗临床方案

规则	条件	结果
1	若（诺丁汉愈后指数为低）	则（建议采取化学治疗为“否”）
2	若（诺丁汉愈后指数为中低）且（雌激素受体实验结果为非阴性）	则（建议采取化学治疗为“否”）
3	若（诺丁汉愈后指数为中低）且（雌激素受体实验结果为阴性）	则（建议采取化学治疗为“不确定”）
4	若（诺丁汉愈后指数为中高）且（雌激素受体实验结果为非阴性）	则（建议采取化学治疗为“否”）
5	若（诺丁汉愈后指数为中高）且（雌激素受体实验结果为阴性）	则（建议采取化学治疗为“是”）
6	若（诺丁汉愈后指数为高）且（雌激素受体实验结果为非阴性）	则（建议采取化学治疗为“不确定”）
7	若（诺丁汉愈后指数为高）且（雌激素受体实验结果为非阴性）且（年龄为年轻）	则（建议采取化学治疗为“是”）
8	若（诺丁汉愈后指数为高）且（雌激素受体实验结果为非阴性）且（备管浸润测试结果显示血管腔内明确存在肿瘤为是）	则（建议采取化学治疗为“是”）
9	若（诺丁汉愈后指数为高）且（雌激素受体实验结果为弱）	则（建议采取化学治疗为“是”）
10	若（诺丁汉愈后指数为高）且（雌激素受体实验结果为非阴性）且（年龄为年老）	则（建议采取化学治疗为“否”）
11	若（诺丁汉愈后指数为高）且（雌激素受体实验结果为非阴性）且（阳性淋巴结数目为阴性）	则（建议采取化学治疗为“否”）
12	若（诺丁汉愈后指数为高）且（雌激素受体实验结果为阴性）	则（建议采取化学治疗为“是”）

图 7　模糊规则库

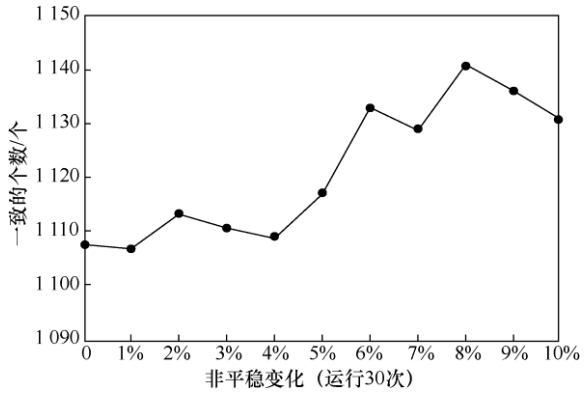

图 8　通过对 30 个非平稳模糊系统集合进行多数投票所获的性能提升示意

从图 8 可以看出，基础“优化”type-1 型系统与临床实践有 1 108 个结果一致，一致性达 84.8%（在 1 306 个案例中占 1 108 个）。相比之下，具有 8%非平稳模糊集变化（在 x 轴上标记为 0.08）的模糊系统在运行 30 次且多数投票的情况下，有 1 141 个一致（占 1 306 个案例的 87.4%），也就是说，与无变化的“优化”系统相比，包含 8%变化的非平稳模糊集专家系统的性能有所提升。

5　差异与学习

根据进一步观察，差异与学习之间存在一定的关联。虽然“学习”这个词的含义似乎很直观，但要精确给出其定义却很困难。两个可用的标准字典定义是：导致行为改变或获得新能力或反应的过程（牛津英语词典）；通过经验改变行为倾向（韦氏词典）。

这些定义都具有一个基本要素——必须对行为作一些改变才能使学习发生。在巴甫洛夫意义上，行为包括采取具体行动或对某种形式（通常是感官输入）的相关刺激作出反应。换句话说，行为是由给定输入的特定输出组成的。如果输出对于给定输入来说是固定的，则行为保持不变。对于将导致行为变化的过程，必须对相同输入给出在某种程度上不同的输出。也就是说，输入-输出映射必须存在变化。如果一个系统（无论是人类还是计算机）总是在给定输入的情况下产生相同的输出，那么就无法进行学习。

因此，系统必须针对相同的输入改变其输出，以便其对自身行为进行更改修正。虽然目前存在多种不同类型的学习（具体可见参考文献[8]），但简单的强化学习可描述为：具有一些被编码为输入–输出映射行为的系统。该系统的行为存在一些随机变化，因此不时会出现微小的行为变化，有时对于同一输入会产生不同的输出。如果行为上的差异可促成结果的改善，则输入–输出映射将持续发生改变。可通过在模糊专家系统中使用非平稳模糊集建模，模糊集的对应术语将具有新的意义，即如果差异能够持续改善性能，那么模糊集的位置也将根据所学经验而不断改变。

由此可知，为进行更充分的学习而采取变化是非常必要的。因此，存在差异是人工智能的关键特性。但获得具有人类差异性的可靠观测数据是极其困难和耗费时间的，在需要高度专业知识和安全核

心系统的领域尤其如此，因为在这些领域中可变性（尤其是犯错误）的概念通常被认为是消极的。人类（即使是专家）会犯错误，但这并不妨碍其工作的开展。但是目前，明确接受这些错误并对其进行衡量似乎是非常具有威胁性的。

当前的社会似乎还未完全接纳机器或者系统犯错，以无人驾驶汽车为例，目前该领域已有许多研究，特别是针对那些导致行人或其他道路使用者死亡的事故的研究。毫无疑问，应当尽量减少受伤或死亡，但是相比之下允许出错率更高的人类开车却可能导致伤害和死亡频发。事实上，根据世界卫生组织的统计数据，2010 年道路交通事故造成全球约 125 万人死亡。假设无人驾驶汽车的应用可将这个数字缩减至百分之一，那么每年将可以挽救超过 100 万人的生命！虽然仍有超过 1 万人会因交通事故死亡，但这种整体性缩减会证明在全球范围内推广无人驾驶汽车是合理的。而追求完美（以及零死亡事故）则可能会延迟这项技术可以提供的根本性安全改进。

6　结束语

本文提出了如何通过差异性来评估人工智能的框架。虽然基于深度学习的神经网络系统似乎提供了目前在计算机系统中可用的最高水平的性能（在需要 AI 技术解决的复杂问题的背景下），但其难以对决策过程进行合理解释。而相比之下，计算机模糊专家系统则能够对所做决策提供一些足以满足更高层次要求的解释。

在正常情况下，即使是最优秀的人类专家也难以达到 100%正确，因此不应期望将计算机决策支持系统投入实际应用后能够实现这个目标。除非接受计算机系统与最优秀的人一样也会犯错误这一观点，否则人们将延迟获得计算机决策支持系统带来的益处。本文探究了计算机决策支持系统与其试图模拟的人类专家之间的不可区分性，通过类似于图灵著名的模仿游戏测试的评估形式，证明应接纳计算机系统的部署以及使用。计算机化的决策支持系统的评估应与人类专家的评估类似。

最后，基于对学习本质的探讨，本文认为差异与学习之间存在必然的联系，只有发生了变化才可以认为计算机系统进行了学习。计算机专家系统在学习的过程中可能存在一些不足，但由于人类专家组成的决策支持系统也会表现出类似的差异和不足，因此要理性、客观、公正地看待计算机决策支持系统的发展，而非一味追求完美。

参考文献：

[1] GARIBALDI J M. The need for fuzzy AI[J]. IEEE/CAA Journal of Automatica Sinica, 2019, 6(3): 610-622.

[2] SILVER D, HUANG A, MADDISON C J, et al. Mastering the game of Go with deep neural networks and tree search[J]. Nature, 2016, 529(7587): 484-489.

[3] TURING A M. Computing machinery and intelligence[J]. Mind, 1950, 59(236): 433-460.

[4] KEITH R D, BECKLEY S, GARIBALDI J M, et al. A multicentre comparative study of 17 experts and an intelligent computer system for managing labour using the cardiotocogram[J]. British Journal of Obstetrics and Gynaecology, 1995, 102(9): 688-700.

[5] GARIBALDI J M, OZEN T. Uncertain fuzzy reasoning: a case study in modelling expert decision making[J]. IEEE Transactions on Fuzzy Systems, 2007, 15(1): 16-30.

[6] ZADEH L A. The concept of a linguistic variable and its application to approximate reasoning[J]. Information Sciences, 1975, 8(8): 199-249.

[7] GARIBALDI J M, ZHOU S, WANG X, et al. Incorporation of expert variability into breast cancer treatment recommendation in designing clinical protocol guided fuzzy rule system models[J]. Journal of Biomedical Informatics, 2012, 45(3): 447-459.

[8] ZUO H, LU J, ZHANG G, et al. Fuzzy transfer learning using an infinite Gaussian mixture model and active learning[J]. IEEE Transactions on Fuzzy Systems, 2019, 27(2): 291-303.

[作者简介]

GARIBALDI Jonathan M（1963– ），男，博士，英国诺丁汉大学计算机科学学院院长，主要研究方向为人类推理的不确定性和差异建模、复杂数据建模和解释。

陈虹宇（1994– ），女，中国科学院自动化研究所复杂系统管理与控制国家重点实验室硕士生，主要研究方向为交通数据分析、社会交通、智能交通。

李小双（1995– ），男，中国科学院自动化研究所复杂系统管理与控制国家重点实验室硕士生，主要研究方向为智能交通系统、强化学习。