人工智能风险分析技术研究进展

陈群[1,2]，陈肇强[1,2]，侯博议[1,2]，王丽娟[1,2]，罗雨晨[1,2]，李战怀[1,2]
1. 西北工业大学计算机学院，陕西 西安 710129；
2. 西北工业大学大数据存储与管理工业和信息化部重点实验室，陕西 西安 710129

摘要
目前基于深度学习模型的预测在真实场景中具有不确定性和不可解释性，给人工智能应用的落地带来了不可避免的风险。首先阐述了风险分析的必要性以及其需要具备的3个基本特征：可量化、可解释、可学习。接着，分析了风险分析的研究现状，并重点介绍了笔者最近提出的一个可量化、可解释和可学习的风险分析技术框架。最后，讨论风险分析的现有以及潜在的应用，并展望其未来的研究方向。

关键词
人工智能；风险分析；不确定性；可解释性

中图分类号：TP391 文献标识码：A doi: 10.11959/j.issn.2096-0271.2020005

Research progress on risk analysis for artificial intelligence

CHEN Qun[1,2], CHEN Zhaoqiang[1,2], HOU Boyi[1,2], WANG Lijuan[1,2], LUO Yuchen[1,2], LI Zhanhuai[1,2]
1. School of Computer Science, Northwestern Polytechnical University, Xi'an 710129, China
2. Key Laboratory of Big Data Storage and Management, Northwestern Polytechnical University, Ministry of Industry and Information Technology, Xi'an 710129, China

Abstract
The predictions of the deep learning models are still uncertain and uninterpretable. As a result, their deployments bring unavoidable risk to business decision making. Firstly, the study on risk analysis was motivated, and the three desirable properties of risk analysis techniques were described: quantifiability, interpretability and learnability. Then the existing work on risk analysis was reviewed, and the newly proposed framework to enable quantifiable, interpretable and learnable risk analysis was introduced. Finally, the existing and potential applications of risk analysis, and its future research direction were discussed.

Key words
artificial intelligence, risk analysis, uncertainty, interpretability

1 引言

随着大数据的发展，以深度学习为代表的人工智能技术在计算机视觉、语音识别、视频处理、自然语言处理等领域得到了广泛的应用。然而，在真实场景中，由于环境的复杂性，基于深度学习模型的预测具有不确定性，在落地时经常带来不可预知的风险。例如，2016年5月，一辆特斯拉Model S电动车在自动驾驶状态下撞上对向正在转弯的卡车，该事故导致特斯拉驾驶员死亡。据特斯拉公司的解释，自动驾驶系统在面对明亮的天空时，没有识别出白色的卡车。基于深度学习模型的人工智能算法的另一个局限是不可解释性，即难以用人们可理解的方式来解释算法做出相应决策的原因。可解释性具有非常重要的意义：一方面，可解释性是保障人工智能安全性的一个重要手段，如果算法能够说明所做决策的依据，人们就可以通过分析依据的合理性和内在逻辑评估算法的安全性；另一方面，可解释性有利于加速推广人工智能的落地应用。人们普遍难以相信一个不可解释的黑盒子模型做出的决策。例如，欧盟提出的《通用数据保护条例》要求算法具备可解释性，数据主体有权获取算法决策的有关解释。电气和电子工程师协会（Institute of Electrical and Electronics Engineers，IEEE）2016年发布的关于人工智能及自动化系统的伦理设计白皮书中，在多个部分提出了对人工智能和自动化系统应有解释能力的要求；美国计算机协会公共政策委员会在2017年初发布的《关于算法透明性和可问责性的声明》提出了7项基本原则，其中之一即“解释”：鼓励使用算法决策的系统和机构对算法过程和特定决策提供解释。

综上所述，目前以深度学习为代表的人工智能技术存在不确定性和不可解释性的问题。因此，能够准确预知人工智能算法在什么情况下可能失效，并提供可解释的原因，是保障人工智能应用安全性的关键。在软件工程中，为确保软件的安全运行，软件测试是其中必不可少的一个环节。软件测试费用达到总开发费用的40%以上。对于某些性命攸关的软件，其测试费用甚至高达其他软件工程阶段费用总和的3~5倍[1]。然而，目前的软件测试只能检测程序的正确性和漏洞，并不能检测人工智能算法预测错误的风险。风险指的是产生损失、造成伤害或者处于不利状况的可能性，该词在实际的生产生活中被广泛运用。在不同的场景中，风险的具体含义通常也不相同。对于人工智能而言，具体的风险包括预测的准确性风险[2-3]、决策的公平性风险[4]以及决策的道德性风险[5]等。本文主要关注人工智能算法的预测准确性风险。

针对目前人工智能技术存在的不确定性和不可解释性问题，高效的风险分析技术需要具备以下3个基本特征。

- 可量化：能够准确分析算法预测错误的可能性，实现风险的量化评估。
- 可解释：可以以人类能理解的方式解释算法预测错误的原因，实现可解释的根因分析。
- 可学习：鉴于深度学习模型的高度复杂性以及其对环境的高度敏感性，风险分析技术需要能够根据实际的应用环境，动态调整其风险模型参数，实现风险评估的自适应性。

2 风险分析技术的研究现状

风险分析在以前的文献中也被称为置

信度评估(confidence ranking)[3]或信任评分(trust scoring)[6],是新兴的研究领域。本节首先回顾与机器学习模型的性能评估和模型可解释性分析相关的工作,说明其与风险分析的区别,然后着重介绍风险分析的研究现状。

2.1 机器学习模型的性能评估

为了预测一个已训练好的机器学习模型在目标数据集上的性能,一个被广泛运用的方法是从目标数据集中随机抽取一部分数据进行人工标注,建立验证集;然后用验证集来评估模型的预测准确度。当目标数据集未知或者无法提供额外的人工标注时,可以在训练数据集上采用交叉验证法(cross-validation),其典型的方式有留一法、十重交叉验证法[7]。需要指出的是,基于验证集的模型准确度预测方法评估的是模型的整体表现,无法评估模型在单个实例上的预测行为。然而,风险分析关注的是单个实例的预测风险,如医学影像分析中单个病人的病情诊断、自动驾驶中某个具体场景的安全分析等。因此,传统的机器学习模型的性能评估方法并不适用于风险分析。

2.2 机器学习模型的可解释性分析

由于深度学习模型的不可解释性,近年来一个热门的研究方向是对黑盒模型进行可解释性分析[8-9]。针对黑盒模型的可解释性研究可分为3类[9]。

(1)模型解释

模型解释即用可解释的、透明的模型来模拟黑盒模型,以此得到一个全局的解释。例如,参考文献[8]提出针对文本分类任务的可解释的数据表示——词袋,然后通过学习一个局部的可解释性模型(比如线性模型)来解释分析算法的预测结果;参考文献[10]提出了基于规则来构建局部的可解释模型,对于符合相同规则的数据,该模型对结果的解释具有一致性,同时排除了不属于规则要求的其他特征的干扰。

(2)输出解释

输出解释即对黑盒模型的输出进行解释分析。例如,参考文献[11]提出利用局部的梯度值来刻画特征的影响,进而解释算法的预测结果;参考文献[12]基于联合博弈论,通过分析每个输入特征的贡献度来解释任意分类器的输出结果;参考文献[13]设计了一个交互式的可视化工具,帮助用户查看每一个具体数据实例的信息。

(3)模型检验

模型检验即对黑盒模型的特性进行解释分析,比如模型的预测行为对输入特征的敏感度、深度神经网络模型中的特定神经元对预测结果的影响等。例如,参考文献[14]提出利用随机森林分类器中的路径信息来指导输入特征的调整,进而改变黑盒模型对某个输入数据的预测结果。需要指出的是,针对机器学习模型的可解释分析的研究旨在提供可解释性的信息,辅助用户对人工智能算法的结果进行分析,但没有提供量化的风险评估。

2.3 风险分析

针对实例的风险分析,最简单的方法就是直接利用模型在每个实例上提供的信息,评估决策的风险。例如,朴素贝叶斯分类器为每个类别标签都提供了相应的类别概率[15],可以天然地作为预测标签的风险度量指标。当分类器的输出不是概率值时,可以采用不同的方法将其转化为类别

概率。例如，基于Platt校准的方法可以将支持向量分类器的输出距离转化为类别概率[16]；softmax函数可以将深度神经网络模型中神经元的输出映射为类别概率[3]。然而，模型本身输出的概率很多时候并不能准确地反映预测的不确定性[3]。为此，有一些工作提出对模型的输出概率进行校准[17-18]。然而，这些校准技术并没有改变实例之间的相对不确定性，而且，当模型的结构复杂或未知时，概率校准极具挑战。对于主流的深度学习模型，直接基于模型输出或校准输出的风险度量方法可解释性差，而且大量的实验表明，其在很多情况下无法获得可靠的风险分析结果[6]。

另外一种方法是通过设计额外的模型来分析原始学习模型在单个实例上的预测行为[6, 19-20]。例如，参考文献[6]提出了一种基于距离的风险评估方法，该方法首先为每种标签构建一个代表该类标签的簇，然后对于给定的任一测试实例，计算该实例与不同机器标签所在簇的距离，最后通过比较这些距离来计算该实例标签的风险。参考文献[19]则针对计算机视觉的应用场景，首先获取部分数据的输入特征和原始模型预测的信息，并以此作为训练数据，然后训练一个支持向量机（SVM）模型或者支持向量回归（SVR）模型，判断新的输入数据不能被原始模型正确处理的风险。严格地说，这种方法是依据输入数据的特征提前拒绝高风险的实例，并没有对原始模型输出的风险进行量化分析。而且，其可解释性也较差。

以上的风险分析方法均根据单一的输出值（如原始学习模型的输出、独立模型的输出）直接度量预测错误的风险。参考文献[21]提出将一个实例的标签概率用一个分布（比如正态分布）来表示，然后借鉴投资风险分析理论中的风险度量指标（如条件在险价值），量化评估标签预测错误的风险。更具体地说，参考文献[22]提出将分类模型的输出作为先验知识，然后通过训练数据获取特征的观测分布，最后利用贝叶斯推理估计实例标签的后验分布。然而，这些方法虽然在风险评估的准确性和可解释性上取得了比之前的方法更好的效果，但仍然无法根据应用环境动态地调整风险模型，即不是可学习的。

3 可量化、可解释和可学习的风险分析框架

笔者在参考文献[23]中提出了一个可量化、可解释和可学习的风险分析框架，并把它成功应用于实体解析的任务中。实体解析旨在识别出关系数据中表示同一个现实世界实体的记录。**图1**所示为一个实体解析的例子，R_1和R_2分别表示文献数据集中的两张数据表，每张表中包含多条记录。对于一个记录对$<r_{1i}, r_{2j}>$，r_{1i}和r_{2j}分别表示R_1和R_2中的一条记录，当且仅当r_{1i}和r_{2j}指向同一篇文章时，称之为“匹配”，否则，称之为“不匹配”。在**图1**的例子中，r_{11}和r_{21}是匹配的，r_{11}和r_{22}是不匹配的。

风险分析框架如**图2**所示，由3个步骤组成：生成风险特征、构建风险模型、训练风险模型。后文将以实体解析为例，阐述每个技术步骤。需要强调的是，这个框架具备很强的通用性，容易被扩展应用于其他一般性分类问题。

3.1 生成风险特征

可解释的风险特征是进行可解释性风险分析的前提。为了有效支持风险分析，风

记录号	标题	作者	地点	年份
r_{11}	Belief Reasoning in MLS Deductive Databases	H Jamil	SIGMOD Conference	1999
r_{12}	Efficient Index Structures for String Databases	T Kahveci, A Singh	VLDB	2001
r_{13}	Multi-Step Processing of Spatial Joins	T Brinkhoff, H Kriegel, R Schneider, B Seeger	SIGMOD Conference	1994
…				

(a) 数据表R_1

记录号	标题	作者	地点	年份
r_{21}	Belief Reasoning in MLS Deductive Databases	HM Jamil	SIGMOD Conference	1999
r_{22}	Reasoning on Regular Path Queries	D Calvanese	SIGMOD RECORD	2003
r_{23}	Efficient Processing of Spatial Joins using R-trees	T Brinkhoff, HP Kriegel, B Seeger	ACM SIGMOD	nan
…				

(b) 数据表R_2

图1 文献数据集中的数据表示例

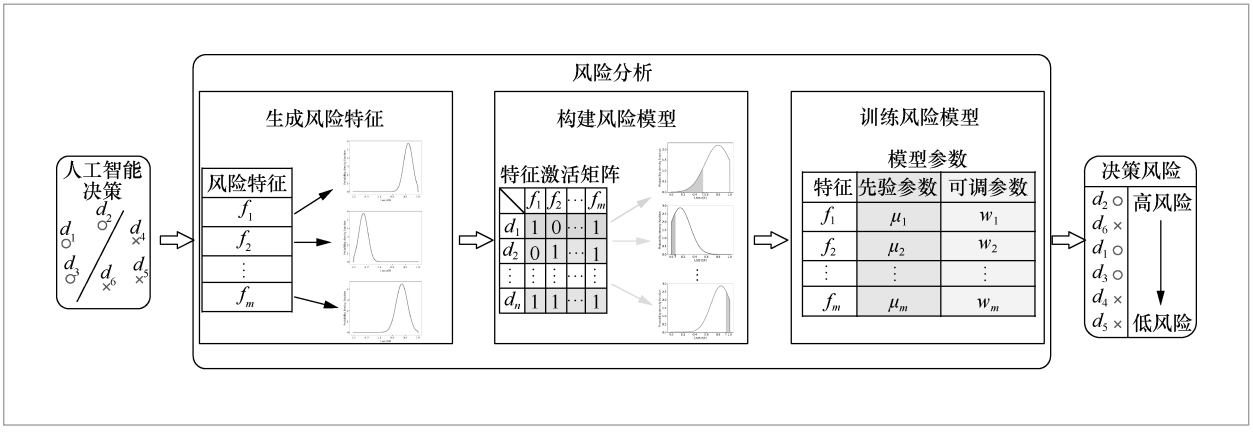


图2 风险分析框架

险特征必须具备以下3个特点：一是可解释的；二是高区分度的，风险特征必须在很大程度上是某一类标签所独有的，对其有明显的指示作用；三是高覆盖率的，风险特征必须被很多个实例共享，只有共享的风险特征才是可学习的。

以实体解析为例，其需要把任一候选记录对标为“匹配”或“不匹配”。规则是一种常见的而且容易被人类理解的知识，因此笔者提出以规则的形式来表达风险特征。具体地说，首先设计能衡量属性值的相似度以及差异度的基本指标，然后在带有真实标签的记录对集合上，以这些基本指标为输入特征，通过生成单边随机森林来获得具有可解释性、高区分度和高覆盖率的规则，得到的规则即风险特征。需要指出的是，单边随机森林中的每一棵树都是单边决策树。传统的双边均

衡的决策树用于判定实例的标签，因此其生成的规则有双向的指示作用。例如，在文献数据集上，“EditDistance(r_{1i}[title], r_{2j}[title])>0.9→equivalent(r_{1i},r_{2j})”作为一个标记规则，其含义如下：如果两条记录r_{1i}和r_{2j}在标题这个属性上的编辑距离相似度大于0.9，那么，这两条记录表示同一篇文章；否则，这两条记录表示不同的文章。与此不同的是，作为风险特征的规则仅具有单边的指示作用。例如，在文献数据集上，规则“r_{1i}[year]≠r_{2j}[year]→inequivalent(r_{1i},r_{2j})”是一个有效的风险规则，因为当两个记录在年份这个属性上的值不一样时，它们表示不同的文章的概率较大。然而，其并不适合作为一个标记规则，因为即便两个记录在年份这个属性上的值一样，它们也很有可能表示不同的文章。

需要强调的是，基于单边决策树的风险特征生成方法具有通用性。对于别的分类问题，只需要设计相应的输入特征即可，整个规则生成过程是同样适用的。例如，在文本分类问题中，提取出的关键词可以作为输入特征[8]；在图像处理中，单个像素通常不具有明确的语义信息，而较大粒度的像素块的语义信息则较为直观[24]；在与抑郁症相关的研究中，面部表情、头部运动以及语气是临床科学家与临床医生们关注的可解释性特点[25]；在嗅觉科学中，一些化学信息（如氢键、芳香环和带电原子等理化性质）为分子科学家们提供了有效的解释信息[26]。

3.2 构建风险模型

基于风险特征提供的信息，风险模型选取合适的风险度量方法来评估人工智能模型的决策风险。受到风险分析在金融领域成功应用的启发，笔者类似地将风险特征提供的信息以概率分布进行表示，然后用风险特征的分布估计目标实例的标签概率分布，最后利用风险度量指标实现风险的量化分析。

在投资组合理论中，一个资产组合的收益概率分布是由资产组合中的每种证券或资产的收益概率分布叠加而成的；通常采用方差、平均绝对离差、半方差、在险价值（value at risk，VaR）、条件在险价值（conditional value at risk）等风险度量指标评估这个投资组合的风险[27]。类似地，针对人工智能模型的风险分析，对于每一个风险特征f_i，假设其蕴含的标签概率为一个服从某种分布（在参考文献[23]中假设的是正态分布，但笔者的方法也同样适用于其他分布）的随机变量。以实体解析为例，对于风险特征f_i，假设其分布的期望为μ_{f_i}，方差为$\sigma_{f_i}^2$，权重为w_i。那么，对于任一记录对d_i，其匹配概率也服从正态分布$\mathcal{N}\left(\mu_i,\sigma_i^2\right)$。如果$d_i$包含$m$个风险特征，那么，其匹配概率的期望可以估计为$\mu_i=\sum_{j=1}^{m}w_j\mu_{f_j}$，方差为$\sigma_i^2=\sum_{j=1}^{m}w_j^2\sigma_{f_j}^2$，即记录对的分布根据风险特征的分布加权叠加来估计。

图3给出了一个当机器标签为“不匹配”时，计算VaR风险指标的示例。指标VaR反映的是在排除掉最坏$(1-\theta)$的情况后，最大可能的损失。在示例中，当θ=0.8时，其在险价值为VaR_1=0.7；当θ=0.95时，其在险价值为VaR_2=0.8。

3.3 训练风险模型

风险模型构建完成后，需要设定可调整的参数，使风险模型能够从观测数据中学习调整风险的评估标准，以期风险模型能够准确地反映人工智能模型在不同环境下的风险。在第3.2节构建的风险模型中，

共有3组参数：风险特征分布的期望、风险特征分布的方差、风险特征的权重。在实践中，可以把风险特征分布的期望当作一种先验知识，由带标签的训练数据通过统计估算出来，而风险特征的权重和方差为待学习参数。

风险模型的训练通过学习排序（learn to rank）技术实现。以实体解析为例，学习排序技术是为了使被错误分类的记录对的风险值能够大于被正确分类的记录对的风险值。给定2个记录对d_i和d_j，假设它们对应的被错误分类的风险值分别为γ_i和γ_j。如果$\gamma_i>\gamma_j$，那么d_i排在d_j前面。然后，采用Logistic函数将它们的风险值映射为d_i排在d_j前面的后验概率：

$$p_{ij}=\frac{e^{(\gamma_i-\gamma_j)}}{1+e^{(\gamma_i-\gamma_j)}} \quad (1)$$

而其目标概率为：

$$\overline{p}_{ij}=0.5\cdot\left(1+g_i-g_j\right) \quad (2)$$

其中，如果记录对d_i被错误分类，那么，风险标签g_i=1，否则，g_i=0。根据定义的记录对排序位置的后验概率和目标概率，在风险模型训练数据D_S上设定目标损失函数为如下的交叉熵损失函数：

$$L(D_S)=\sum_{d_i,d_j\in D_S}[-\overline{p}_{ij}\cdot\ln(p_{ij})-(1-\overline{p}_{ij})\cdot\ln(1-p_{ij})] \quad (3)$$

最后，采用梯度下降的方法，逐渐减小交叉熵损失函数的值直至收敛，从而优化参数。

4 风险分析的应用

风险分析技术不仅可以直接用于评估人工智能算法所作决策的风险，进行风险的根因解释，保障人工智能的安全，还可以用于众包的问题选择和分类的质量控制等任务。另外，风险分析为理解机器学习和人工智能提供了独特的分析视角和手段，潜在地可以影响机器学习的几乎每一个核心环节，包括训练数据的主动选择以及模型的训练等。本章讨论风险分析的一些初步应用以及其潜在的更广泛的应用，并通过它们展望风险分析未来的研究方向。

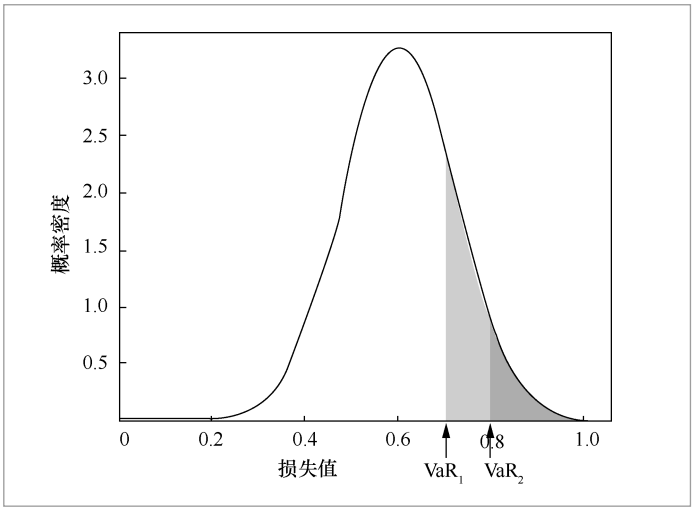


图 3 在险价值的示例

4.1 众包

众包技术旨在将复杂的任务切割并封装为较简单的子任务，通过众包平台，交给非专业或者只有少量专业知识的普通大众来完成[28]。众包通常需要支付酬金给完成任务的人，这会产生人力成本。此外，由于大众的背景知识和认真程度等因素存在差异，他们回答问题的准确性也参差不齐。因此，众包的基本挑战在于从不可靠的答案中推理出准确的答案，并最小化人力成本。为提高答案的准确性，一个典型的做法是将每一个子任务都分配给多个人来完成，然后综合分析返回的多个答案来决定最终的答案。然而，这样冗余的方式也大

大增加了人力成本。因此，在众包应用中，可以先对机器的输出进行风险分析，再根据风险的高低确定不同的人力验证方案，确保更多的人力花费在那些高风险的子任务上，这样就可以有效地平衡整体的准确度与人力开销。

4.2 分类的质量控制

对于分类问题，现有的机器学习模型通常不能保证分类结果的质量。然而，在一些关键的应用领域（如金融欺诈检测和身份识别），经常要求模型的预测结果具有很高的质量，如要求识别的准确率大于一个给定的阈值（如0.99），并且召回率大于一个给定的阈值（如0.99）。在这种情况下，完全基于机器的自动分类往往难以达到设定的质量要求，因此需要人工的介入。在参考文献[29]中，笔者提出了一个人机协作（human and machine cooperation，HUMO）的架构，如**图4**所示，通过人机协作实现分类问题的质量控制。其基本思路是对机器自动分类的结果进行风险分析，将风险较低的实例交由机器自动标注，而将较高风险的实例交给人工验证，这样就可以以少量的人工实现质量控制。在此HUMO的基础上，参考文献[21]进一步提出了一个改进的交互式人机协作框架——r-HUMO（risk-aware HUMO），如**图5**所示。与HUMO静态批量地选择人工工作量不同，r-HUMO通过实时的风险分析渐进地选择人工工作量。相比HUMO，r-HUMO在满足相同质量要求的前提下，能有效地减少所需的人工成本。需要强调的是，虽然参考文献[21]和参考文献[29]的工作针对的是实体解析任务，但是它们提出的框架和技术也能够被扩展应用于其他的通用分类任务。

4.3 训练数据的主动选择

人工智能模型的训练通常需要标注大量的数据。然而，在真实的应用场景中，训练数据的获取往往比较困难且标注的代价较高。因此，需要通过主动学习来减少所需的训练数据。主动学习技术能够选择那些最有助于改善当前模型的数据，并将它们进行人工标注后作为训练数据。相比于随机选择训练数据，主动学习能够有效地

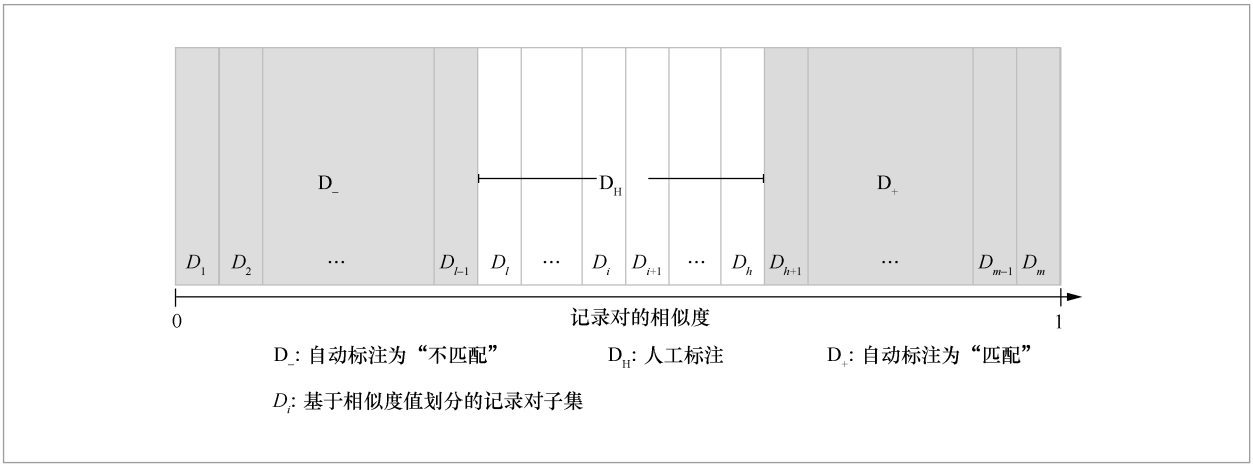


图 4 HUMO 系统框架 [29]

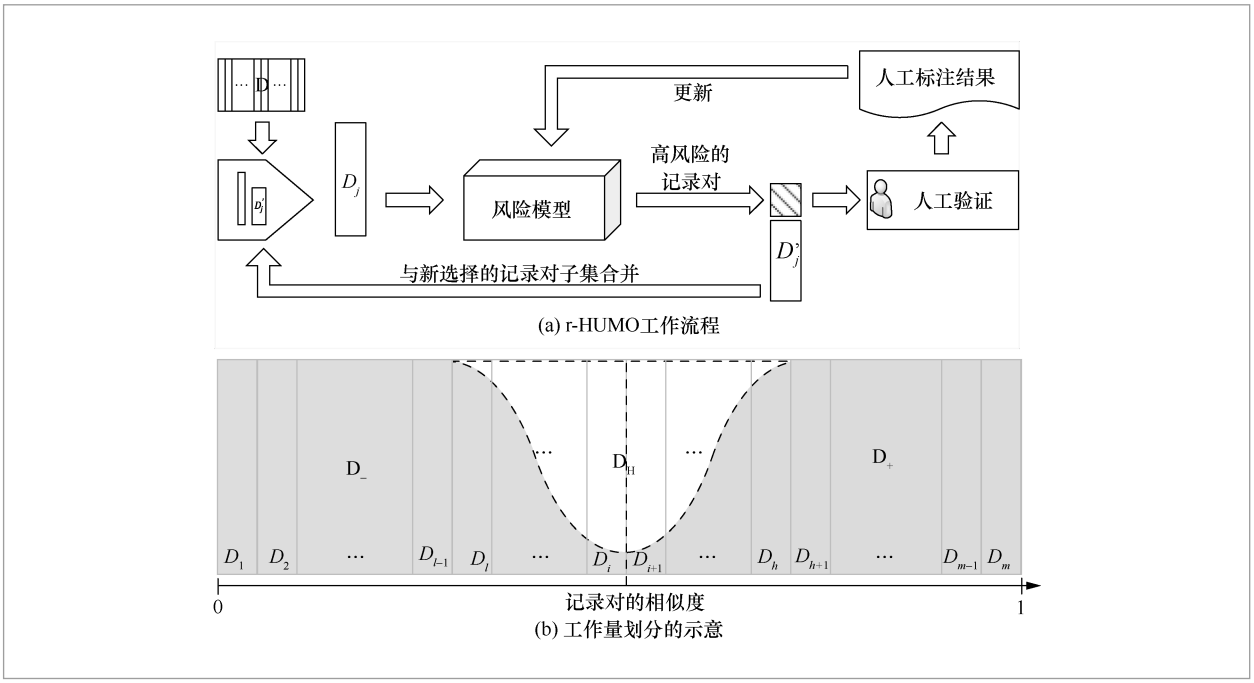


图 5 r–HUMO 系统框架[21]

减少数据的标注成本。目前的主动学习技术主要通过不确定性和代表性等指标或融合了不确定性和代表性的混合指标选择训练数据。不确定性的度量方法包括置信度、离边界的距离、预测类别的熵以及模型委员会的选举等。代表性的度量主要通过计算实例之间的距离实现。有实验表明[30]，当批量选取的训练数据量较大时（如大于1 000），主动学习最好的方法仍是基于模型输出的不确定性。由于风险分析可以更准确地评估不确定性，其自然也可以用于主动学习中训练数据的选取，即每轮都选取风险最高的一组数据来标注。如果可选取的整体数据量有限，则可以综合考虑风险和代表性等指标。

4.4 模型的训练

目前，深度学习模型普遍存在过于乐观的问题[31]，即当目标数据不在模型预测任务的范围内时，模型也可能会给出一个置信度较高的预测结果。参考文献[17]提出，可以通过新增一个离群点检测（outlier exposure）模块改善过于乐观的问题。具体地，在深度学习模型的优化目标函数中新增一个衡量离群数据预测值的损失函数，并增加一个离群数据训练集，通过对模型进行重新训练来改进模型，使新模型能够较好地识别异常数据，并给出较低的预测置信度。然而，目前的方法没有考虑如何提高模型在预期任务上的预测准确度问题。由于风险分析能够评估模型预测的风险并给出解释，那么，它也可以被用来指导模型的设计和训练过程。如何利用风险分析的反馈指导和改善人工智能模型的训练是未来一个非常有价值的研究方向。

5 结束语

当前，基于深度学习的人工智能预测

模型普遍存在不确定性和不可解释性的问题。因此，可量化、可解释和可学习的风险分析技术对保障人工智能的安全至关重要。在本文中，笔者系统地总结了风险分析技术的研究进展，并介绍了一些应用案例，如众包和分类的质量控制等。笔者进一步指出，风险分析为理解人工智能提供了独特的分析视角和手段，其潜在的影响不是局限于目前参考文献中提及的应用案例，而是几乎涉及机器学习的每一个核心环节，包括训练数据的选择和模型的训练等。因此，风险分析是一个非常有价值和前景的研究方向，对推动人工智能的发展具有重要的战略意义。

参考文献:

[1] 齐治昌，谭庆平，宁洪. 软件工程[M]. 北京：高等教育出版社，2001：304-331.
QI Z C，TAN Q P，NING H. Software engineering[M]. Beijing：Higher Education Press，2001：304-331.

[2] AMODEI D，OLAH C，STEINHARDT J，et al. Concrete problems in AI safety[J]. Computer Science，2016，arXiv：1606.06565.

[3] HENDRYCKS D，GIMPEL K. A baseline for detecting misclassified and out-of-distribution examples in neural networks[C]// The 5th International Conference on Learning Representations (ICLR)，April 24-26，2017，Toulon，France. [S.l.:s.n.]，2017：1-21.

[4] CORBETT-DAVIES S，GOEL S. The measure and mismeasure of fairness：a critical review of fair machine learning[J]. Computer Science，2018，arXiv：1808.00023.

[5] MORLEY J，FLORIDI L，KINSEY L，et al. From what to how：an overview of AI ethics tools，methods and research to translate principles into practices[J]. Computer Science，2019，arXiv：1905.06876.

[6] JIANG H，KIM B，GUAN M，et al. To trust or not to trust a classifier[C]// Advances in Neural Information Processing Systems，December 3，2018，Montreal，Canada. [S.l.:s.n.]，2018：5541-5552.

[7] KOHAVI R. A study of cross-validation and bootstrap for accuracy estimation and model selection[C]// The 14th International Joint Conference on Artificial Intelligence，August 20-25，1995，Montreal，Canada. San Francisco：Morgan Kaufmann Publishers Inc.，1995：1137-1143.

[8] RIBEIRO M T，SINGH S，GUESTRIN C. Why should I trust you? explaining the predictions of any classifier[C]// The 22nd ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)，August 24-27，2016，San Francisco，USA. New York：ACM Press，2016：1135-1144.

[9] GUIDOTTI R，MONREALE A，RUGGIERI S，et al. A survey of methods for explaining black box models[J]. ACM Computing Surveys，2019，51(5)：93.

[10] RIBEIRO M T，SINGH S，GUESTRIN C. Anchors：high-precision model-agnostic explanations[C]// The 32nd AAAI Conference on Artificial Intelligence，February 2-7，2018，New Orleans，USA. [S.l.:s.n.]，2018：1527-1535.

[11] BAEHRENS D，SCHROETER T，HARMELING S，et al. How to explain individual classification decisions[J]. Journal of Machine Learning Research，2010，11(11)：1803-1831.

[12] KONONENKO I，STRUMBELJ E. An efficient explanation of individual classifications using game theory[J]. Journal of Machine Learning Research，2010，11(1)：1-18.

[13] AMERSHI S，CHICKERING M，DRUCKER S，et al. Modeltracker：

redesigning performance analysis tools for machine learning[C]// The 33rd Annual ACM Conference on Human Factors in Computing Systems, April 18-23, 2015, Seoul, Korea. New York: ACM Press, 2015: 337-346.

[14] TOLOMEI G, SILVESTRI F, HAINES A, et al. Interpretable predictions of tree-based ensembles via actionable feature tweaking[C]// The 23rd ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), August 13-17, 2017, Halifax, Canada. New York: ACM Press, 2017: 465-474.

[15] KONONENKO I. Semi-naive Bayesian classifier[C]// The 6th European Working Session on Learning, March 6-8, 1991, Porto, Portugal. London: Springer-Verlag, 1991: 206-219.

[16] PLATT J. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods[J]. Advances in Large Margin Classifiers, 1999, 10(3): 61-74.

[17] HENDRYCKS D, MAZEIKA M, DIETTERICH T. Deep anomaly detection with outlier exposure[C]// The 7th International Conference on Learning Representations (ICLR), May 6-9, 2019, New Orleans, USA. [S.l.:s.n.], 2019: 1-18.

[18] OVADIA Y, FERTIG E, REN J, et al. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift[J]. Computer Science, 2019, arXiv: 1906.02530.

[19] ZHANG P, WANG J L, FARHADI A, et al. Predicting failures of vision systems[C]// The IEEE Conference on Computer Vision and Pattern Recognition, June 23-28, 2014, Columbus, USA. Piscataway: IEEE Press, 2014: 3566-3573.

[20] BANSAL A, FARHADI A, PARIKH D. Towards transparent systems: semantic characterization of failure modes[C]// The 13th European Conference on Computer Vision, September 6-12, 2014, Zurich, Switzerland. Heidelberg: Springer, 2014: 366-381.

[21] HOU B Y, CHEN Q, CHEN Z Q, et al. r-HUMO: a risk-aware human-machine cooperation framework for entity resolution with quality guarantees[J]. IEEE Transactions on Knowledge and Data Engineering, 2018: 1-13.

[22] CHEN Z Q, CHEN Q, HOU B Y, et al. Improving machine-based entity resolution with limited human effort: a risk perspective[C]// The International Workshop on Real-Time Business Intelligence and Analytics, August 27, 2018, Rio de Janeiro, Brazil. [S.l.:s.n.], 2018: 1-5.

[23] Chen Z Q, Chen Q, Hou B Y, et al. Towards interpretable and learnable risk analysis for entity resolution[R]. 2019: 1-32.

[24] VENTURA F, CERQUITELLI T. What's in the box? explaining the black-box model through an evaluation of its interpretable features[J]. Computer Science, 2019, arXiv: 1908.04348.

[25] ANIS K, ZAKIA H, MOHAMED D, et al. Detecting depression severity by interpretable representations of motion dynamics[C]// The 13th IEEE International Conference on Automatic Face and Gesture Recognition (FG), May 15-19, 2018, Xi'an, China. Piscataway: IEEE Press, 2018: 739-745.

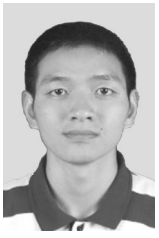
[26] LICON C, BOSC G, SABRI M, et al. Chemical features mining provides new descriptive structure-odor relationships[J]. PLoS Computational Biology, 2019, 15(4): e1006945.

[27] MARKOWITZ H M, BLAY K A. 风险-收益分析: 理性投资的理论与实践(第一卷)[M]. 唐亮, 武微, 译. 北京: 机械工业出版社,

2016: 100-121.
MARKOWITZ H M, BLAY K A. Risk-return analysis: the theory and practice of rational investing(Volume 1) [M]. Translated by TANG L, WU W. Beijing: China Machine Press, 2016: 100-121.
[28] LI G L, WANG J N, ZHENG Y D, et al. Crowdsourced data management: a survey[J]. IEEE Transactions on Knowledge and Data Engineering, 2016, 28(9): 2296-2319.
[29] CHEN Z Q, CHEN Q, FAN F F, et al. Enabling quality control for entity resolution: a human and machine cooperation framework[C]// IEEE 34th International Conference on Data Engineering (ICDE), April 16-19, 2018, Paris, France. Piscataway: IEEE Press, 2018: 1156-1167.
[30] GISSIN D, SHALEV-SHWARTZ S. Discriminative active learning[J]. Computer Science, 2019, arXiv: 1907.06347.
[31] GUO C, PLEISS G, SUN Y, et al. On calibration of modern neural networks[C]// The 34th International Conference on Machine Learning, August 6-11, 2017, Sydney, Australia. [S.l.:s.n.], 2017: 1321-1330.

作者简介

陈群(1976-),男,博士,西北工业大学计算机学院教授、博士生导师,主要研究方向为人工智能风险分析等大数据分析技术。



陈肇强(1988-),男,西北工业大学计算机学院博士生,主要研究方向为人工智能风险分析等大数据分析技术。

侯博议(1990-),男,西北工业大学计算机学院博士生,主要研究方向为人工智能风险分析等大数据分析技术。

王丽娟(1992-),女,西北工业大学计算机学院硕士生,主要研究方向为人工智能风险分析等大数据分析技术。

罗雨晨（1997- ），女，西北工业大学计算机学院硕士生，主要研究方向为人工智能风险分析等大数据分析技术。

李战怀（1961- ），男，博士，西北工业大学计算机学院教授、博士生导师，大数据存储与管理工业和信息化部重点实验室主任，主要研究方向为大数据管理技术、海量信息存储系统等。

收稿日期： 2019-09-12

基金项目： 国家重点研发计划基金资助项目（No.2018YFB1003400）；国家自然科学基金资助项目（No.61732014，No.61672432）；陕西省自然科学基础研究计划基金资助项目（No.2018JM6086）

Foundation Items: The National Key Research and Development Program of China(No.2018YFB1003400), The National Natural Science Foundation of China(No.61732014, No.61672432), The Project Supported by Natural Science Basic Research Plan in Shaanxi Province of China(No.2018JM6086)