计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.201200021

面向机器学习系统的需求建模与决策选择

杨 立[1] 马佳佳[1] 江华禧[1] 马肖肖[1] 梁 赓[1] 左 春[1,2]
1 中国科学院软件研究所 北京 100190
2 中科软科技股份有限公司 北京 100190
(yangli2017@iscas.ac.cn)

摘 要 机器学习支撑的系统应用越来越普遍,但是此类系统的需求通常难以表达完整且可能存在一些难以检测的冲突,使得这些系统通常无法在生产环境中高效满足用户的综合需求。此外,对于在实际场景中使用的机器学习系统,用户信任通常取决于包含可解释性、公平性等非功能需求在内的综合需求的满足程度,且在不同领域内应用机器学习通常有特定的需求,为保证需求描述的质量及实施过程的决策带来了挑战。为解决以上问题,文中提出了一个机器学习系统的需求建模和决策选择框架,包括一个 MLS(Machine Learning Systems)需求概念模型和机器学习管道过程元模型,以及对训练数据集、算法等组件的决策选择方法,旨在规范实际场景中机器学习系统的需求设计、开发和评估。实例研究表明,提出的 MLS 需求描述和实现方法是可行且有效的。

关键词:机器学习系统;需求建模;非功能需求;元模型;决策选择
中图法分类号 TP391

Requirements Modeling and Decision-making for Machine Learning Systems

YANG Li[1], MA Jia-jia[1], JIANG Hua-xi[1], MA Xiao-xiao[1], LIANG Geng[1] and ZUO Chun[1,2]
1 Institute of Software, Chinese Academy of Sciences, Beijing 100190, China
2 Sinosoft. Co, Ltd, Beijing 100190, China

Abstract The application of systems supported by machine learning is becoming more and more common. However, because the requirements of such systems are often difficult to express completely and there may be some conflicts which are hard to detect, these systems usually cannot efficiently meet the comprehensive needs of users in a real application environment. In addition, for Machine Learning Systems (MLS) used in actual scenarios, user trust usually depends on the satisfaction of comprehensive requirements including non-functional requirements such as interpretability and fairness, and application of machine learning in different fields usually has specific needs, which brings challenges to ensure the quality of requirement description and decision-making for implementation process. To solve above-mentioned problems, this paper presents a machine learning system requirements and decision-making framework which includes a concept MLS requirements model and a Meta-Model of MLS pipeline process, as well as decision making method for training datasets and algorithms selection. The purpose is to standardize the design, development and evaluation of requirements for machine learning used in actual scenarios. The case study shows that the proposed MLS requirement description and implementation method is feasible and effective.

Keywords Machine learning systems, Requirements modeling, Non-functional requirements, Meta-model, Decision making

1 引言

机器学习(Machine Learning,ML)描述了一种通过算法从海量数据中“学习”的计算模式,可以解决一些常规软件系统难以解决的问题,如图像识别、医疗辅助诊断、保险定价、情感分析等[1]。不同于单个机器学习算法或者模型,实际应用中的机器学习系统[2](Machine Learning Systems,MLS)指由一个或多个机器学习模型、用于训练模型的数据、与模型交互的界面及文档等组成的完整系统。传统软件系统通常以显式的规则或人机交互来控制系统行为,而 MLS 更多地依赖数据特征和隐性的模型来进行系统行为决策。由于数据属性复杂多变且难以衡量、常规机器学习模型的可解释性差等原因,

到稿日期:2020-09-02 返修日期:2020-10-29 本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:中国科学院战略性先导 A 类专项(XDA20080200);国家重点研发计划项目(2018YFB1005002)
This work was supported by the Strategy Priority Research Program of Chinese Academy of Sciences(XDA20080200) and National Key Research and Development Program of China(2018YFB1005002).
通信作者:左春(zuochun@sinosoft.com.cn)

MLS对传统需求工程提出了新的挑战。

首先，传统软件系统需求很少对不在本系统中流转的数据或者算法提出直接的质量要求，而MLS需求则可能隐含着对训练数据和机器学习算法的质量要求，如公平性要求就限制了将标签分布欠均衡的数据用于模型训练。其次，由于机器学习模型的训练过程通常需要耗费较长时间，且算法调试的复杂性高，MLS需求的不完整性和含糊不清对后续实现过程的影响更大。

相比传统软件系统，MLS开发过程中一旦因为需求问题出现返工，将导致更大的损失和不确定性。因此，清晰、完整地定义MLS需求有助于从软件工程早期阶段确定利益相关者的功能需求及非功能需求(Non-functional Requirements，NFR)，尽可能早地确定满足这些需求的训练数据、算法及实现策略，并对实现结果进行验证，从而对整个MLS的实现过程提供更高水平的质量保障和控制，具有明显的研究意义和实用价值。

从需求角度而言，MLS的行为和结果是否符合人们的预期，可以通过MLS功能需求和NFR的满足程度来反映。然而，相比传统软件系统，MLS的NFR往往更为复杂，难以表达完整，主要体现在如下几个方面：

(1)MLS的NFR往往具有领域特性，如在招生录取领域，关于性别的公平性往往非常重要，而在健康领域，男女之间的差别则是正常的，不需要体现关于性别的公平性；如果涉及到个人信息，隐私保护一般是强需求；在社会计算领域，涉及到国家特征的需求一般都要求公平性等。

(2)MLS的NFR不仅包括常规的正确性，还包括公平性、可解释性、安全性、隐私保护等，不同的NFR在具体实现时可能会依赖数据或者算法的不同特性，如MLS公平性主要依赖训练数据的平衡特性，隐私保护主要依赖数据的加密手段，可解释性主要依赖具体算法实现等。

(3)最近的研究表明[3]，MLS的不同NFR之间往往存在一定的trade-off，如提高安全性肯定会影响效率，隐私保护和公平性也不可能同时达到最优。

因此，在对MLS的需求进行建模时，迫切需要一种方法进行合理描述，利用现有的领域知识和机器学习专业知识对其加工和补全，并有效指导后续的数据选择和算法开发。本文提出了一个MLS需求建模和决策选择框架，首先引入了一个MLS需求目标模型，该模型描述了目标与机器学习过程之间的关系，然后定义了一种决策选择技术，该技术利用目标和NFR需求来评估对训练数据和算法的具体要求，并在给定情况下获得最佳的解决方案。对于分析和设计一个MLS，本文方法可以有效描述和分析针对机器学习特性的非功能性需求，并推荐适合的训练数据和机器学习算法。

本文的主要贡献包括3个方面：

(1)提出了一种MLS需求概念模型，通过建立MLS的NFR、训练数据集、算法目录(Catalog)和特征(Meta-Feature)结构及其映射关系，将用户需求转化为解决方案中的可配置参数。

(2)提出了一个机器学习管道过程元模型，根据上一步提取的训练数据及算法的NFR要求，采用效用函数计算评分的方法来量化评估训练数据和算法选择方法，并推荐可有效辅助算法工程师满足需求的机器学习模型。

(3)通过在一带一路全球新闻情感分析系统中的实际验证表明，提出的MLS需求描述和实现方法是可行的、有效的。

本文第2节介绍了相关工作；第3节介绍了基本定义；第4节对本文方法进行了总体概览描述；第5节描述了机器学习管道流程元模型；第6节介绍了需求描述模型及检测；第7节通过案例分析描述了应用本文方法的具体结果；最后总结全文并提出下一步工作。

2　相关工作

目前有许多工作使用机器学习来改进需求工程任务(如优先级[4]、模型提取[5-6]、需求分类[7])，但是基于MLS的需求工程研究却相对较少，其中数据挖掘参考过程，例如数据库中的知识发现(KDD)过程[8]或跨行业标准的数据挖掘过程(CRISP-DM)[9]，分别提出了两个不同的MLS参考过程模型，但是其对具体步骤缺乏细化分析，更没有涉及到训练数据集和算法选择问题。文献[10]指出，需求问题被列为开发MLS最困难的活动。在常规软件环境中，此活动涉及初始阶段的需求分析和规范，以及最后阶段的验收检查。当面对MLS时，由于无法进行事前估计或无法保证可实现的准确性，现有的流程不一定适用。文献[11]指出了MLS非功能性需求的挑战和研究方向，指出当前的工作仅研究了某项特定的NFR因素(例如隐私与处理时间)，而未能将所有可能的NFR进行统一描述和规划。这是本文力争解决的问题之一。文献[12]提出了一个端到端的监督学习流程，其讨论范围仅限于分类问题。在其他类型系统的需求建模方面，文献[13]研究了基于模型的大数据分析服务组合问题，主要侧重点在于大数据分析和服务计算的结合技术。文献[14]研究了可信软件非功能需求形式化表示与可满足问题。文献[15]研究了中文非功能需求描述的识别与分类方法，为本文研究提供了参考。

尽管机器学习任务存在各种各样的变种，但是现有的需求分析技术和工具往往将机器学习的整个过程假定为具有相似特征的单个功能，因此产生了多种针对机器学习过程的不同视图，进一步使MLS需求描述和实现方式的选择复杂化。同时，面向数据的处理方法的速度往往较慢，并且需要耗时的任务(例如数据标注、模型训练)来应对机器学习对象不断变化的特征。传统的需求工程框架使用通用功能来区分机器学习对象并用来描述MLS需求。但是，通用功能不适用于为机器学习对象组成可演变和可复用的服务组件。因此，为数据预处理、模型训练、模型部署等机器学习过程中不同组件的组合定义动态变化特征，对于MLS的需求描述及其后续实现至关重要。

综上所述，目前仍然缺乏贯穿MLS整个生命周期的细粒度需求建模及决策选择的综合方法。本文在前期工作[16-17]

的基础上，提出了一种通用的 MLS 需求建模和决策选择方法，将面向机器学习的功能需求及 NFR 融合到一个机器学习管道过程元模型中，基于机器学习对象生成变化特征，实现数据选择和算法选择的动态组合。该方法包括：

(1) MLS 需求概念模型，定义了 MLS 需求和约束，包括基于约束概念的验证程序，以及定义一致的描述模型；

(2) 基于 CRISP-DM[9] 扩展的机器学习管道过程元模型，该模型定义了与声明性模型中的需求描述链接的机器学习过程组件组合。

3 机器学习系统需求相关概念

在详细描述具体方法之前，我们先给出一些基本的符号解释和定义。

传统意义上，MLS 的非功能属性(Non-functional Attribute，NFA)通常指正确性和效率等指标，可用于描述 MLS 的 NFR。随着机器学习应用场景的不断扩大，与现实问题的结合更加紧密，安全性、可解释性、公平性等 NFA 渐渐成为研究的热点，目前尚未形成一个公认的完整 NFA 目录框架。图 1 给出了机器学习系统部分常用的 NFA 分类，限于篇幅，关于每一项 NFA 的定义请参见文献[18]，本文不再赘述。

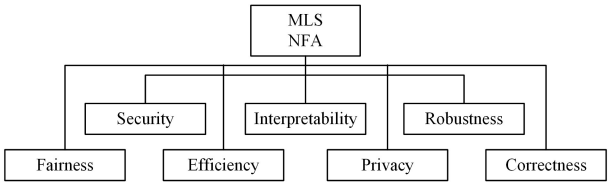

图 1 机器学习系统部分非功能属性分类

Fig. 1 Part of NFA classification of MLS

定义 1(MLS 候选解决方案) 给定领域知识集合 K，MLS 任务 T，G 和 Q 分别为 T 包含的功能性目标和软目标。

如果解决方案 U 满足条件 $U, K \vdash G, Q$，则称其为任务 T 的一个候选解决方案。

下文给出本文讨论的 MLS 需求问题定义，该定义不仅包含如何界定一个 MLS 需求，还包括一个 MLS 建议解决方案的过程。

定义 2(MLS 需求问题) 给定领域知识集合 K，任务集 T，功能目标集 G，软目标集 S，找到所有满足需求问题的候选 MLS 解决方案，并使用效用函数(Utility Function，UF)对候选方案进行评分，以确定建议解决方案。

由于 MLS 的解决方案描述起来通常较为复杂，既包含过程模型，也包含过程中的决策选择。因此，本文首先通过机器学习管道过程元模型对过程进行描述，然后通过不同的效用函数计算元模型中所需的决策参数，从而达到生成建议 MLS 解决方案的目的。

4 本文方法概览

当前的需求分析模型大都基于传统软件过程，对于数据驱动的 MLS 软件过程的研究则相对较少，无法有效地支持 MLS 软件过程中面临的需求建模及决策选择问题，因此，本文提出一种 MLS 需求建模及决策选择方法，该方法的总体概览如图 2 所示。首先，确定 MLS 任务，在领域知识的指导下进行任务分解，将其划分为功能目标和软目标。功能匹配选择过程根据用户请求的功能要求确定机器学习算法的种类(如分类、聚类、回归等)。非功能性匹配过程根据用户的软目标确定非功能需求集合(如公平性、准确性和可解释性等)。其次，根据功能需求和非功能需求分别提取出训练数据特征需求和算法需求，根据数据特征需求及数据集效用函数从数据集目录中挑选出合适的训练数据集。根据算法需求和算法效用函数从算法目录中挑选出合适的机器学习算法。然后，将训练数据集和算法选择结果实例化到管道过程元模型中，算法工程师可根据实例化后的过程模型进行编码、调试、模型生成和测试等工作。最后，进行模型部署并由用户校验需求的满足程度。

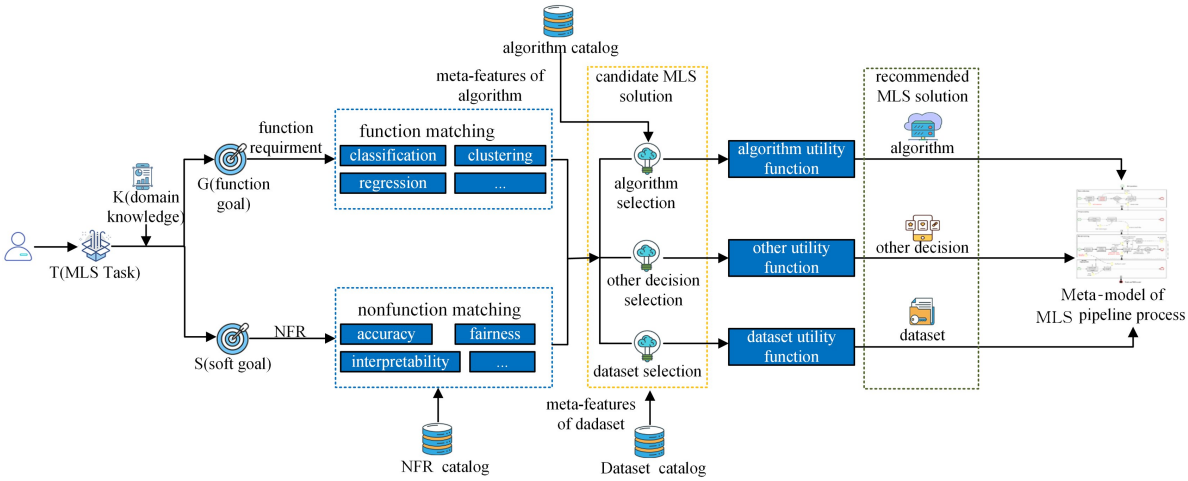

图 2 本文方法的总体概览图

Fig. 2 Overview of proposed method

如图 2 所示，该方法可以针对不同的任务定义不同的效用函数，本文认为其中最为关键的部分是针对训练数据集和机器学习算法的效用函数，下文分别对其进行说明。

本文中使用的数据集效用函数的定义如下：需求描述完

成之后，需要根据面向数据集的 NFR 选择一个合适的训练数据集，数据集 D 的特征可以表示为一个向量 $\boldsymbol{D}^F=(D_1^F,\cdots,D_n^F)$，其中 n 为系统中用到的数据 NFR 的总数，数据 NFR 向量 $\boldsymbol{DNFR}^F=(DNFR_1^F,\cdots,DNFR_n^F)$，$DNFR$ 的权重向量 $\boldsymbol{W}_D=(W_1,\cdots,W_n)$，$\boldsymbol{D}^F$ 在 $\boldsymbol{DNFR}^F$ 上的评分向量 $\boldsymbol{score}_D^F=(score_1^F,\cdots,score_n^F)$，则数据集 D 的效用函数可以表示为：

$$Utility_D=\sum_{i=1}^{n}\boldsymbol{W}_i*\boldsymbol{score}_i^F \tag{1}$$

本文中使用的算法效用函数的定义如下：训练数据集选择完成之后，需要根据面向机器学习算法的 NFR 来选择一个合适的算法对数据进行训练，算法 A 的特征可以表示为一个向量 $\boldsymbol{A}^F=(A_1^F,\cdots,A_m^F)$，其中 m 为系统中用到算法 NFR 的总数，算法 NFR 向量 $\boldsymbol{ANFR}^F=(ANFR_1^F,\cdots,ANFR_m^F)$，$ANFR$ 的权重向量 $\boldsymbol{W}_A=(W_1,\cdots,W_m)$，$\boldsymbol{A}^F$ 在 $\boldsymbol{ANFR}^F$ 上的评分向量 $\boldsymbol{score}_A^F=(score_1^F,\cdots,score_m^F)$，则算法 A 的总评分可以表示为：

$$Utility_A=\sum_{i=1}^{m}\boldsymbol{W}_i*\boldsymbol{score}_i^F \tag{2}$$

5　机器学习管道流程元模型

在参考 CRISP-DM[9] 模型的基础上，本文提出的机器学习管道流程元模型如图 3 所示，其中的决策点用红色表示。

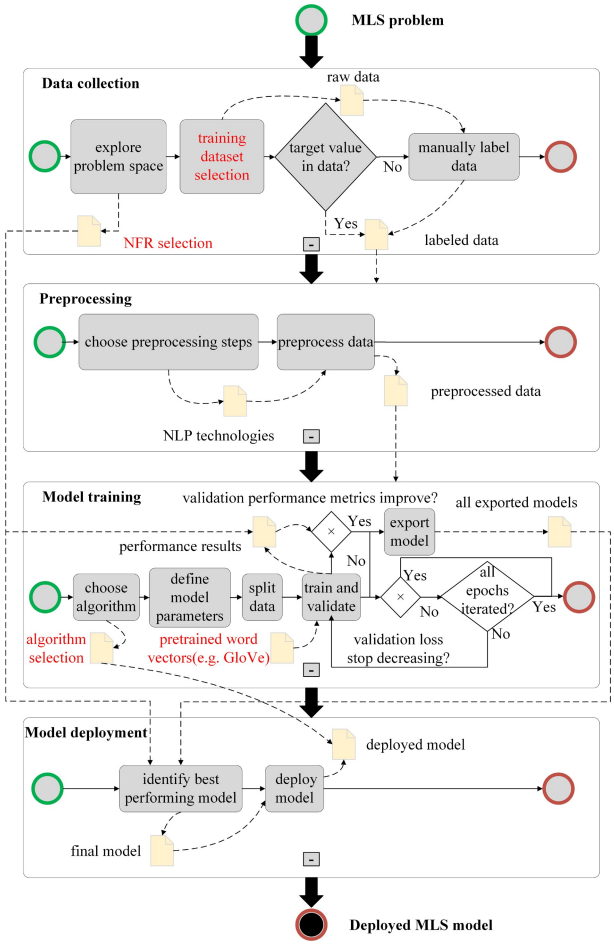

图 3　机器学习系统管道流程元模型（电子版为彩色）

Fig. 3　Meta-model of MLS pipeline process

该元模型包括 4 层：数据收集、数据预处理、模型训练和模型部署。一般情况下，将按以下顺序处理数据源。首先，根据用户需求得到的数据 NFR 需求选择训练数据集，同时判定数据源的种类和数量，为每个数据源收集数据并将其存储在临时数据库中，数据预处理层根据需求对数据源中的数据进行适当的预处理；接下来，模型训练层根据 NFR 评分指标及权重，结合算法 NFR 进行组合分析，并为每个机器学习管道选择适当的机器学习算法，并利用训练数据集进行训练，生成机器学习模型；最后，模型部署层汇总机器学习管道的输出，利用机器学习模型对数据源提供的数据进行分析预测，并将分析结果提供给最终用户。

6　需求描述模型及一致性检测

本文的 MLS 需求描述方法首先定义了需求描述模型，然后说明了需求一致性检测方法，以确保 MLS 需求定义的正确性。

6.1　MLS 需求描述模型

需求描述模型是与数据和算法无关的模型，可让用户定义一套形成 MLS 的需求。需求定义了 MLS 要实现的功能目标 G。软目标集 S 是表示衡量或评估目标的一种方式，而 G 则是衡量目标是否达成的状态表征。例如，用户可以选择情感预测作为功能目标 G，将预测正确率作为衡量目标的非功能属性，而将“预测正确率不小于 90%”作为 MLS 的软目标 S。需求描述模型的定义如下：

$RDM:=(a_i,\omega)$，每个概念 $a_i\in A$，A 为 MLS 领域需求的概念集合。其中 ω 是一组独立的需求，$\omega_i=\{r_i,C_i,pr_i\}$，$r_i$ 是一个子需求，可以进行层次化表达，$C_i=\{c_1,\cdots,c_n\}$ 是 r_i 在需求分析和推理过程中的一组约束，pr_i 是 ω_i 的优先级。例如：$\omega_i=\{$k-Nearest Neighbor，$\{k>10\}$，$1\}$ 对 k-NN 算法需求的参数 k 设置了约束（即 $k>10$），并为其分配最大优先级（即 $pr_i=1$）。

本文涉及如下功能：

数据准备指定了所有旨在为 MLS 准备数据的活动。例如，它描述了如何执行降维，或者定义了如何使用数据脱敏技术（例如非对称加密算法）来保证数据所有者的隐私。

数据表示指定数据的表示方式，并表示每个分析过程的表示选择。例如，可定义数据模型（如文档数据库 MongoDB、图数据库 Neo4j）和存储方式（如云存储、分片等）。

数据预处理指定采用何种方法进行数据抽取、清洗和加工。例如，可定义处理类型（例如实时、近实时、批处理）和预期的等待时间。

模型训练指定机器学习的类型。例如，可定义预期的结果（例如描述性、说明性、预测性）和学习方法（例如监督学习、无监督学习、强化学习等）。

模型部署指定系统部署的环境参数。例如 GPU 参数、内存及存储容量要求等。

结果展示方式指定了如何组织分析结果以进行显示和报告的抽象表示。例如，可定义数据显示类型（例如词云、时间线等）。

MLS 需求描述以 JSON 格式表示，以确保不同步骤之间

具有共同的语义、推理能力和互操作性。

6.2 需求一致性检测

本文的主要目标之一是指导用户定义不同机器学习概念领域的一致需求描述。本文使用约束描述来丰富需求描述模型，将不兼容或相互约束的目标和功能相关联。定义 *attribute* 表示引用过程/部署模型的属性，而 *value* 是给定属性的值(或值的数组)。

$$op::=\langle\mid\rangle\mid=\mid\neq\mid\in\mid\notin$$

$$constraint::=op(attribute,value)$$

设ω_a和ω_b分别为两个需求描述，其中ω_a的需求优先级大于ω_b。接下来引入一种形式化的约束概念，以检测不一致的需求描述并驱动需求描述模型的验证过程。具体来说，将约束定义为 $constraint\{\omega_a,\omega_b\}\in\wp$ 中的关系，$\wp$ 为所有约束的集合。

约束检测过程 $\zeta:\wp\rightarrow\mathcal{R}$，将$\{\omega_a,\omega_b\}\in\wp$中的约束作为输入，并将要在$\omega_b$上执行的规则 $r\in\mathcal{R}$ 作为输出返回，以生成一致的描述模型。其中ω_a是优先级较高的需求描述，并且根据规则 r，ω_a和ω_b之间的不一致性可以通过修改ω_b的范围来解决。具体来说，首先对 MLS 需求进行优先级排序，并通过约束进一步消解冲突。优先级在需求之间分配顺序，以解决 MLS 需求不同阶段的冲突和矛盾。约束在低级别设置的规范中为具有不同能力的用户提供支持，这些设置将应用于执行中的服务参数。约束定义为对象数据结构，即属性的集合。

需要指出的是，不同的抽象级别会导致不同的执行点。当约束被定义为需求描述的一部分时，即$\omega_i=\{r_i,C_i,pr_i\}$，该约束将在需求描述级别执行和解决。

7 案例研究

在中国科学院战略性先导 A 类专项“泛第三极环境变化与绿色丝绸之路建设”的支持下，我们设计并实现了一带一路全球新闻情感分析系统，该系统简要介绍如下：“一带一路”倡议自提出以来，得到了世界上越来越多的国家和国际组织的认同，其相关报道与日俱增，及时获取并掌握这些公开新闻中蕴含的观点对于应对各类问题和制定发展政策具有重大的参考意义。通过对国内外媒体覆盖监测可快速获知“一带一路”的最新相关报道，依托机器学习技术进行智能分析(挖掘热点词、负面新闻等)，同时提供可视化界面辅助用户决策。系统的主要研究问题包括：如何在海量新闻报道中快速全面地监测“一带一路”的相关报道？如何准确地自动提取“一带一路”相关报道的重要信息？如何智能监测“一带一路”相关报道的情感态度？

该系统的主要任务为情感分析，是一种典型的机器学习任务，系统的构建流程如下：首先，选择一组国外权威网站作为收集新闻的数据源，如 BBC、路透社、一带一路新闻网等。其次，对于每个选定的新闻数据源，定义数据预处理流程，以通过使用多种工具和服务来去除无关数据以及没有地理标签的数据。然后，从选定的每个数据集中获取单篇新闻。接下来，选择一个或多个机器学习算法进行数据训练，用于所有单篇新闻的情感分析。最后，在信息提取过程之后，收集算法的不同输出并进行聚合，通过使用各种可视化功能(例如图表、词云、时间线等)进行呈现。系统设计时需考虑的问题包括：

(1)统一的数据预处理机制可能不适用于多种新闻数据源，需要为每个新闻数据源设计一系列单独的数据预处理步骤。

(2)与数据预处理类似，一种算法可能不适用于所有类型的新闻数据源。无论选择何种算法，随着时间和需求的不断演变，原有的模型可能不再满足需求，仍然需要重复训练和算法验证的过程。

(3)由于整个系统涉及的需求较为复杂且不断演化，可能存在需求不一致或者需要权衡的需求。

7.1 需求冲突检测

在一带一路全球新闻情感分析系统中，由于涉及到多语种，对于非英文的新闻文章，在数据预处理步骤中需要先调用翻译引擎将其翻译成英文再进行情感分析，翻译引擎的选择是影响该系统整体性能的关键因素，如国外某翻译引擎精度高，但是由于网络原因，响应速度较低，而国内的翻译引擎响应速度快，但是翻译精度有所欠缺。因此，需要根据需求的不同动态精确地选择翻译引擎。

在描述初始需求时需要同时描述系统总体响应时间(保证用户体验)和翻译引擎返回结果所需时间(用于选择具体的翻译引擎接口)。初始需求可描述为：

$$\omega_a=\{ResponseTime,T_{\text{response}}<5\text{ s},1\}$$

$$\omega_b=\{TranslatorTime,T_{\text{translator}}<2.5\text{ s},3\}$$

根据实际算法测试的结果，最大的模型执行时间为 3 s。在$\wp$中设置约束规则：

$$constraint_1=\{ResponseTime=ForcastingTime+TranslatorTime\}$$

$$constraint_2=\{ForcastingTime\leqslant 3\text{s}\}$$

由于ω_a的优先级高于ω_b，通过执行 $\zeta:\wp\rightarrow\mathcal{R}$ 的需求检测过程，生成规则 $r=T_{\text{translator}}<2\text{ s}$，依据上述结果，$\omega_a$ 与需求ω_b产生冲突，于是系统将ω_b调整为：

$$\omega_b=\{TranslatorTime,T_{\text{translator}}<2\text{ s},3\}$$

这导致采用 Google Translator 不能满足系统需求，在实际场景中采用了符合本需求的 Baidu Translator 引擎进行具体实现。

7.2 训练数据集的选择及预处理

根据用户需求，我们将该系统的 NFR 定义为正确性、可解释性和公平性，由于该系统是一个由用户交互自行判断情感极性的 MLS，可解释性、公平性和正确性的优先级排序为：Interpretability＞Fairness＞Accuracy，按照第 6.1 节中的模型描述方法，有关 NFR 的需求部分可描述为：$\omega_1=\{$Interpretability, Null, 1$\}$，$\omega_2=\{$Fairness, Null, 2$\}$，$\omega_3=\{$Accuracy, Null, 3$\}$。然后将 NFR 分解成为数据 NFR 和算法 NFR，其中 Fairness 主要由数据集 NFR 实现，Interpretability 和 Accuracy 主要由算法 NFR 实现。将上述 3 个 NFR 的权重在数据集和算法效用函数中定义为最高。

本文从数据集目录中选出的候选训练数据集及 NFR 特性如表 1 所列，其中 Time Range 以距当前年份的差距作为评价指标。

表 1　部分候选数据集在不同 NFR 上的特性

Table 1　Different NFR properties of some candidate datasets

Dataset	Relevance	Time Range	Diversity	Fairness	*Data* size
IMDB[19]	Low	N/A	Low	Low	50 000
Yelp[20]	Low	2014	Low	High	30 000
Amazon Electronics[21]	Low	N/A	Medium	High	30 000
Sentiment140[22]	High	N/A	High	Medium	1 600 000
SemEval 2013-2017[23]	High	2013－2017	High	Medium	61 584
MPQA[24]	High	2001－2002	High	High	535
Slovene[25]	Medium	2007－2013	Low	High	10 427

根据需求描述，由领域专家设置各个数据集 NFR 的取值 *score*，如表 2 所列。设置 Relevance，Time Range，Diversity，Fairness，Data size 的权重分别为 9，3，4，10，3。根据式(1)计算的各训练数据集的总评分如表 3 所列。

表 2　数据集 NFR 的 *score* 值

Table 2　*score* of dataset NFR

NFR	Value	*score*
Relevance	Low	30
	Medium	60
	High	90
Time Range	0～5	50
	6～10	40
	＞10	20
	N/A	10
Diversity	Low	20
	Medium	50
	High	80
Fairness	Low	50
	Medium	85
	High	95
Data size	[0，10 000]	10
	(10 000，20 000]	20
	(20 000，30 000]	30
	(30 000，50 000]	50
	(50 000，70 000]	70
	(70 000，1 000 000]	90
	＞1 000 000	100

表 3　部分候选数据集的总评分

Table 3　Overall score of some candidate datasets

Dataset	Overall score
IMDB	35.52
Yelp	52.07
Amazon Electronics	53.10
Sentiment140	79.66
SemEval 2013－2017	**80.69**
MPQA	74.83
Slovene	60.34

本文方法从候选训练数据集中选出评分最高的 SemEval 2013－2017 数据集作为建议使用的训练数据集。使用 n-gram 处理等技术进行数据预处理，并对非英文数据调用翻译接口将其统一翻译成英文。

7.3　机器学习算法选择

本文从算法目录中选出的候选机器学习算法的 NFR 特性如表 4 所列，其中 Flexibility 代表算法实现过程中的特性指标。根据需求描述，由领域专家设置各个算法 NFR 的 *score*，如表 5 所列。

表 4　部分监督学习算法在不同 NFR 上的特性

Table 4　Different NFR properties of some supervised learning algorithms

Algorithm	Accuracy	Training time	Interpretability	Flexibility
Logistic Regression	Low	Fast	Low	Simple to implement
Decision Tree	Low	Fast	High	Simple to implement
Decision Forest	High	Moderate	High	High memory usage
Support Vector Machines	Moderate	Moderate	Low	Large feature sets
Naïve Bayes	Low	Moderate	High	Simple to implement
BI-LSTM (sentence)	High	Slow	High	Customization possible
BI-LSTM (document)	Moderate	Slow	Low	Customization possible

表 5　算法 NFR 的 *score*

Table 5　*score* of algorithms NFR

NFR	Value	*score*
Accuracy	Low	20
	Medium	50
	High	90
Training Time	Slow	60
	Moderate	70
	Fast	80
Interpretability	Low	10
	High	100
Flexibility	Simple to implement	60
	Customization possible	70
	High memory usage	50
	Large feature sets	50

本文设置 Accuracy，Training Time，Interpretability，Flexibility 的权重分别为 10，6，10，4，根据式(2)计算的各训练数据集的总评分如表 6 所列，系统建议句子级的 Bidirectional LSTM 算法作为最符合需求的算法。

表 6　部分监督学习算法的总评分

Table 6　Overall score of some supervised learning algorithms

Algorithm	Overall score
Logistic Regression	34.00
Decision Tree	64.00
Decision Forest	84.00
Support Vector Machines	40.67
Naïve Bayes	62.00
BI-LSTM(sentence)	**84.67**
BI-LSTM(document)	41.33

按照本文方法对一带一路全球新闻情感分析系统建模后的结果如图 4 所示，其中决策后的选择用红色斜体字标出，这样就完成了本文提出的机器学习管道流程元模型到具体 MLS 系统实现流程的实例化过程。该系统已经正式上线运行，支持多语言多新闻源的情感分析任务，满足了新闻情感分析公平性的需求。

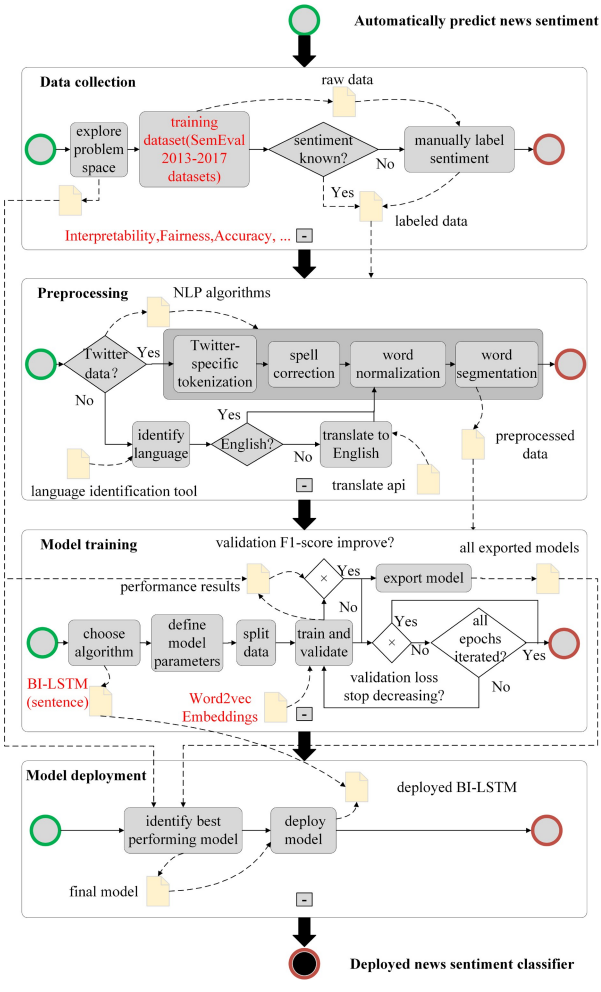

图 4 一带一路全球新闻情感分析系统流程建模结果
(电子版为彩色)

Fig. 4 Process modeling result of Belt and Road Global News Sentiment Analysis System

图 5 给出了系统的分析界面,通过本文方法选择了基于句子级极性标注和词云展现的 Bidirectional LSTM 算法进行具体实现,弥补了传统 LSTM 算法在可解释性方面的不足,同时满足了该系统关于准确性和可解释性的要求。

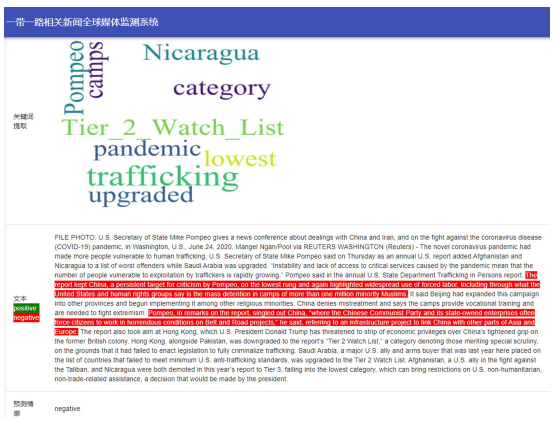

图 5 系统分析界面

Fig. 5 User interface of system analysis

在该案例中,本文方法的优势是使未来的需求变更响应和机器学习关键组件更换都处于一个相对透明的状态,提升了整个系统适应变化的能力和可维护性,同时也可为实现其他类型的 MLS 任务提供实现过程参考。就我们所知,现有的 MLS 参考过程模型[7-8]仅提供了概念模型,无法提供更多实际的具体实现过程帮助。

结束语 MLS 的特点决定了其 NFR 属性相对复杂,不同角色对 MLS 需求的理解存在一定的差距,因此容易造成需求不完整和内在冲突。另外,机器学习系统的开发通常采用自下而上的、技术驱动的方法,其执行工作流程和相应的计算通常对最终用户、开发人员和架构师都是隐藏的,给系统质量带来了隐患。概念模型和元模型的实践证明是提高软件质量的重要途径。本文提出了一种面向机器学习系统的需求建模和决策选择方法,包括一个 MLS 需求概念模型和机器学习管道过程元模型,支持通过需求分解对机器学习过程组件(如训练数据集和算法)进行精确选择,其意义在于在早期需求工程阶段对 MLS 需求进行概念建模与推理,找出满足 MLS 需求的实现策略,避免了可能存在的需求分析与传导不清晰而引起的额外工作量,从而减少了 MLS 整个软件过程的不确定性。在一带一路全球新闻情感分析系统上的实际应用证明了该方法的可行性和有效性。

本文的重点在于为 MLS 的需求描述和决策选择提供一个参考框架,其局限性在于对于机器学习的具体特性或组件(如具体 NFR、算法、数据之间的相互关系)的讨论尚不充分,相关参数的设置目前还需要通过领域专家根据经验和参考文献进行人工设置。下一步我们将研究通过实际测试数据反馈生成参数的方法,来解决现有方法中依赖领域专家设置参数的不足。

未来工作还包括:1)开发机器学习系统需求描述和冲突检测工具;2)将该方法推广到本文场景之外的更多复杂的场景中,同时融入用户偏好等信息进一步提高本文方法的鲁棒性和适应性。

参 考 文 献

[1] ALPAYDIN E. Introduction to machine learning [M]. MIT Press, 2020.

[2] TOMSETT R, BRAINES D, HARBORNE D, et al. Interpretable to whom? A role-based model for analyzing interpretable machine learning systems [J]. arXiv: 1806.07552, 2018.

[3] AGARWAL S. Trade-Offs between Fairness, Interpretability, and Privacy in Machine Learning [D]. Waterloo: University of Waterloo, 2020.

[4] PERINI A, SUSI A, AVESANI P. A machine learning approach to software requirements prioritization [J]. IEEE Transactions on Software Engineering, 2012, 39(4): 445-461.

[5] ARORA C, SABETZADEH M, NEJATI S, et al. An active learning approach for improving the accuracy of automated domain model extraction [J]. ACM Transactions on Software Engineering and Methodology (TOSEM), 2019, 28(1): 1-34.

[6] PUDLITZ F, BROKHAUSEN F, VOGELSANG A. Extraction of system states from natural language requirements [C]// 2019 IEEE 27th International Requirements Engineering Conference (RE). IEEE, 2019: 211-222.

[7] WINKLER J，VOGELSANG A. Automatic classification of requirements based on convolutional neural networks[C]// 2016 IEEE 24th International Requirements Engineering Conference Workshops (REW). IEEE，2016：39-45.

[8] MAIMON O，ROKACH L. Introduction to knowledge discovery and data mining[M]// Data Mining and Knowledge Discovery Handbook. Springer，Boston，MA，2009：1-15.

[9] CHAPMAN P，CLINTON J，KERBER R，et al. CRISP-DM 1.0 Step-by-step Data Mining Guide[OL]. https://www.the-modeling-agency.com/crisp-dm.pdf (accessed on 4 August 2020).

[10] ISHIKAWA F，YOSHIOKA N. How do engineers perceive difficulties in engineering of machine-learning systems? -Questionnaire survey[C]// 2019 IEEE/ACM Joint 7th International Workshop on Conducting Empirical Studies in Industry (CESI) and 6th International Workshop on Software Engineering Research and Industrial Practice (SER&IP). IEEE，2019：2-9.

[11] HORKOFF J. Non-functional requirements for machine learning：Challenges and new directions[C]// 2019 IEEE 27th International Requirements Engineering Conference (RE). IEEE，2019：386-391.

[12] HIRT R，KOEHL N J，SATZGER G. An end-to-end process model for supervised machine learning classification：from problem to deployment in information systems[C]// Designing the Digital Transformation：DESRIST 2017 Research in Progress Proceedings of the 12th International Conference on Design Science Research in Information Systems and Technology. Karlsruhe，Germany. Karlsruher Institut für Technologie (KIT)，2017：55-63.

[13] ARDAGNA C A，BELLANDI V，BEZZI M，et al. Model-based big dataanalytics-as-a-service：take big data to the next level[J/OL]. IEEE Transactions on Services Computing. https://ieeexplore.ieee.org/abstract/document/8319508.

[14] ZHANG X，LI T，WANG X，et al. Formal analysis to non-functional requirements of trustworthy software[J]. Ruan Jian Xue Bao，2015，26(10)：2545-2566.

[15] JIA Y D，LIU L. Recognition and classification of non-functional requirements in Chinese[J]. Ruan Jian Xue Bao，2019，30(10)：3115-3126.

[16] YANG L，LI M，REN J，et al. A Human-in-the-Loop Method for Developing Machine Learning Applications[C]// 2019 6th International Conference on Systems and Informatics (ICSAI). IEEE，2019：492-498.

[17] YANG L，ZUO C，WANG Y G. Research and Implementation of Service Oriented Architecture for Knowledge Discovery[J]. Chinese Journal of Computers，2005，28(4)：445-457.

[18] ZHANG J M，HARMAN M，MA L，et al. Machine learning testing：Survey，landscapes and horizons[J/OL]. IEEE Transactions on Software Engineering. https://ieeexplore.ieee.org/abstract/document/9000651.

[19] MAAS A，DALY R E，PHAM P T，et al. Learning word vectors for sentiment analysis[C]// Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics：Human Language Technologies. 2011：142-150.

[20] TANG D，QIN B，LIU T. Document modeling with gated recurrent neural network for sentiment classification[C]// Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. 2015：1422-1432.

[21] ZHANG X，ZHAO J，LECUN Y. Character-level convolutional networks for text classification[C]// Advances in Neural Information Processing Systems. 2015：649-657.

[22] Sentiment140. For Academics[EB/OL]. http://help.sentiment140.com/for-students/?spm=a2c4e.11153940.blogcont576274.32.1b956c9czbnj9g.

[23] BAZIOTIS C，ATHANASIOU N，CHRONOPOULOU A，et al. Ntua-slp at semeval-2018 task 1：Predicting affective content in tweets with deep attentive rnns and transfer learning[J]. arXiv：1804.06658，2018.

[24] MPQA. MPQA Resources[EB/OL]. http://mpqa.cs.pitt.edu/.

[25] 19Joey85. Sentiment-annotated-news-corpus-and-sentiment-lexicon-in-Slovene[EB/OL]. https://github.com/19Joey85/Sentiment-annotated-news-corpus-and-sentiment-lexicon-in-Slovene.

YANG Li，born in 1978，Ph.D，associate professor，is a member of China Computer Federation. His main research interests include intelligent software engineering and blockchain.

ZUO Chun，born in 1959，professor，Ph.D supervisor. His main research interests include software engineering and so on.