1 AI4PA: Artificial intelligence for program analysis

1.1 NLP

自然语言处理（Natural Language Processing，NLP）主要研究如何让计算机能够理解、生成人类语言，包括早期基于统计方法的文本分类、信息检索等研究，以及近年来基于深度神经网络的机器翻译、问答系统、情感分析等研究。

1.1.1 原理

NLP 的原理包含多方面，如语言学、统计学、计算机科学等，它涉及对语言的结构、语义、语法、语用等方面的研究，以及对大规模语料库的统计分析和模型建立。

语言模型 字符序列的概率分布就是一个最简单的语言模型。例如，在一个网页中，我们可以统计得到 $P(\text{“}the\text{”}) = 0.027$。长度为 n 的书写符号序列称为 n-元组（n-gram）（gram 是希腊语表示“书写”或“字母”的单词的的词根）。n 个字符序列上的概率分布就称为 n 元字符模型。单词模型、音节模型同理。

n 元模型可以定义为 n-1 阶 Markov 链，即 c_i 的概率仅和 c_{i-1}、c_{i-2}，…，$c_{i-(n-1)}$ 有关。

对于一个包含 100 个字符的语言的 3 元字符模型，c_i 的概率有 100 万项参数。这些参数可以通过计数的方式，对包含上千万字符的文本集合进行统计得到。我们把文本集合称作*语料库*。

作为对语言模型的改进，*平滑*是一种调整低频计数的概率的过程，使得语料库中出现概率为零的序列会被赋予一个很小的非零概率值（其他数值会小幅下降以使概率和仍为 1）。

特征向量 与线性代数无关，机器学习中的特征向量是一个特征值向量。假设现在我们有一个*文本分类*的自然语言处理任务，目标是判断一封邮件是否是垃圾邮件。使用 1 元单词模型，语言模型中包含 10 万个单词，那么特征就是词汇表中的单词，即一个 100000 元组，特征的值就是每个单词在邮件中出现的次数，即特征向量 X 长度为 10 万。这种一元的表示形式被称为*词袋*模型。这样我们就可以将垃圾邮件的特征向量用于监督学习。例如，“for cheap”、“You can buy”等 n 元单词序列很像是垃圾邮件的特征。

1.1.2 流程

NLP 的流程一般包含以下几个步骤：

1. **数据收集和预处理**：获取和清洗原始语言数据，如语料库
2. **分析和词法分析**：将原始数据转换为适合模型输入的格式
3. **特征提取**：将文本转换为计算机可以处理的向量形式
4. **模型训练**：训练 NLP 模型
5. **模型评估**：使用验证数据集评估模型的性能
6. **模型应用**：将模型应用于实际问题

1.2 程序作为文本

程序分析指对计算机程序进行自动化地处理，以确认或发现某些程序性质，如正确性、安全性、性能等，程序分析的结果可以用于编译优化、漏洞检查等任务。经典的程序分析技术包括数据流分析、符号执行等。近年来机器学习也被用于一些程序分析任务。

```
#include <stdlib.h>
#include <string.h>
int main()
{
  char *str;
  str = (char *) malloc(15);
  strcpy(str, "abc");
  // free(str);
  return(0);
}
```

假设我们要检查如上代码所示的内存泄露漏洞，语言模型是 C 语言的表达式，包含变量声明、库函数调用、返回等，语料库是正确的程序，那么在 n 元表达式模型中（假设 n 足够大），库函数调用 `malloc` 和 `free` 的特征值应当是相等的，否则就可能存在内存泄露。

我们再看一个例子，类型检查（或类型推断）是编译器的重要组件。

```
// type error 1, C
char x = 'a';
x = "abc";

// type error 2, Java
switch (true) {
  case false: break;
}

// type error 3, Swift
class C {}
Class D {}
let x = true ? C() : D()
```

如上三个例子都包含类型错误。例 1 中，x 赋值为字符串，与其声明的 char 类型不符。例 2 中，switch 不接收布尔类型的变量。例 3 中，因为 C 和 D 没有公共父类型，导致三元运算符的类型不匹配。

简单地，类型检查以一个代码块和一个预期类型作为输入，递归下降地对代码块的子节点进行类型检查，最后输出一个布尔值表示是否存在类型错误。例如，

```
typeCheck(Add(1, '2'), Int)
  = typeCheck(1, Int) && typeCheck('2', Int)
  = true && false
  = false
```

使用机器学习方法，n 元字符模型可以学到字面量如 1、`'2'`、`"abc"` 等的类型，n 元表达式模型可以学到赋值语句、switch 表达式、三元运算符等的类型约束。

甚至对于一些经典类型检查和推断方法中的复杂情形，比如范型类型的类型推断，机器学习方法可以简单地解决。例如，对于 `isEqual<t1, t2>`，n 元字符模型可以学到名为 `isXXX` 的函数大概

率会返回布尔类型。

1.3 编程语言处理

程序语言相对于自然语言有两个特点：（1）更强的结构化，如嵌套和跳转；（2）作为形式语言，有精确定义的语言模型，如语法和语义规则。

Mingzhe Hu
humingzhework@163.com
Hefei, China
July 2023