数据挖掘-SPEIT/SJTU引文预测挑战

E'cole Polytechnique 3月2022日

# 挑战描述

该项目的目标是研究机器学习/人工智能技术并将其应用于链接预测问题。 链路预测是预测网络中两个实体之间存在链路的问题。 这个问题最近在许多领域引起了很多关注。 例如，在社交网络中，人们可能对预测用户之间的友谊联系感兴趣，而在生物学中，预测基因和蛋白质之间的相互作用至关重要。 在此挑战中，您将处理预测一篇研究论文是否引用另一篇研究论文的问题。 更具体地说，您将获得一个由数千篇研究论文组成的引文网络，以及它们的摘要和作者列表。 通常处理问题的管道与任何分类问题中应用的管道相似; 目标是使用边缘信息来学习分类器的参数，然后使用分类器来预测两个节点是否通过边缘链接。

这项挑战是在Kaggle上进行的，Kaggle是一个用于预测建模的平台，公司，组织和研究人员在该平台上发布其数据，来自世界各地的统计人员和数据采矿者竞争以产生最佳模型。 挑战可通过以下链接获得:[https:// www。](https://www.kaggle.com/c/dm-speit-2022) [kaggle.com/c/ dm-speit-2022](https://www.kaggle.com/c/dm-speit-2022)。 要参加挑战，请使用以下链接:[https:](https://www.kaggle.com/t/b2926120ee09442c96dee0b15e1059d6)

[// Www.kaggle.com/t/b2926120ee09442c96dee0b15e1059d6](https://www.kaggle.com/t/b2926120ee09442c96dee0b15e1059d6)。

# 数据集描述

如上所述，您将在引文网络上评估您的方法。 该数据集包含已在机器学习，人工智能，数据挖掘和自然语言处理会议和期刊上发表的论文。 您将获得以下文件 (可在以下网址获得:[https://www.dropbox.com/sh/ fhfjjtk0sr7pmse/AAD4ZEtHv9OI5HfVO22tdMX0a？dl = 0](https://www.dropbox.com/sh/fhfjjtk0sr7pmse/AAD4ZEtHv9OI5HfVO22tdMX0a?dl=0))。

* 1. **edgelist.txt**: 根据在机器学习，人工智能，数据挖掘和自然语言处理场所发表的论文创建的引文网络。 节点对应于论文，而边表示引用关系。 该图是无向的。 因此，如果节点之间存在边*五*和*u*,那么要么纸*五*cites论文*u*或者纸*u*cites论文*五*。该图由138组成*,*499顶点和1*,*091*,*总共955个边缘。 请注意，该图实际上包含更多的边，但是其中一些已被删除，并且要求您在挑战的上下文中识别它们。
  2. **abstracts.txt**: 它包含138的摘要*,*499论文。 文件的每一行都包含论文的ID及其摘要。 下面的字符串用于将ID与摘要分开:*| − − |*。
  3. **作者.Txt**: 此文件包含138的作者*,*499论文。 文件的每一行都包含论文及其作者的ID。 以下字符串用于将ID与作者分开:

*| − − |*。 对于具有多个作者的论文，逗号字符 (*,*) 用于分隔不同的作者。

* 1. **test.txt**: 此文件包含106*,*692有序节点对。 文件的每一行都包含源节点的ID和目标节点的ID。 您的方法的最终评估将在这些节点对上完成，目标是预测每对的两个元素之间是否存在边缘。

# 评价

将使用对数损失度量 (也称为二进制交叉熵损失) 评估模型的性能。 此度量定义为给定概率分类器预测的真实类标签的负对数似然性。 具体地，二进制对数损失定义为:

*L*=*−*1区 (*y*日志 (*p*) + (1*− y*) 日志 (1*− p*))

*我*= 1

*N*

*N*

*我*

*我*

*我*

*我*

哪里*N*是样本数 (即节点对)，*y我*如果两个节点之间有边，则为1，否则为0，并且*p我*是两个节点之间存在边的预测概率。

# 提供源代码

您将获得两个用Python编写的脚本，这将帮助您开始挑战。 第一个脚本 (graph baseline.py) 仅使用基于图的特征和逻辑回归分类器进行预测。 第二个脚本 (text baseline.py) 使用从论文摘要中提取的特征以及逻辑回归分类器。 作为此挑战的一部分，您需要编写自己的代码并构建自己的模型，以预测两个节点是否通过边缘链接。 建议您同时使用图形理论和文本信息。

# 有用的Python库

在本节中，我们简要讨论一些在挑战中有用的工具，并鼓励您使用。

* Python中一个非常强大的机器学习库是scikit-learn[1](#_bookmark0)。 它可以用于预处理步骤 (例如，用于特征选择) 和分类任务 (在scikit-学习)。
* Python中一个非常流行的深度学习库是PyTorch[2](#_bookmark1)。 该库提供了一个简单且用户友好的界面来构建和训练深度学习模型。

1[http:// scikit-learn.org/](http://scikit-learn.org/)2<https://pytorch.org/>

* 由于您将处理表示为图形的数据，因此使用库来管理和分析图形可能很重要。 这样的库的一个例子是NetworkX[3](#_bookmark2)Python库，它将允许您创建、操作和研究的结构和其他几个功能图表。
* 由于您还将处理文本数据，因此自然语言工具包 (NLTK)[4](#_bookmark3)的Python也可以发现有用。
* Gensim[5](#_bookmark4)是一个Python库，用于无监督主题建模和自然语言处理，使用现代统计机器学习。 图书馆提供了学习word和文档的所有必要工具嵌入。

# 关于项目提交的规则和细节

**规则。** 以下规则适用于此挑战 :( i) 每个参与者允许一个帐户 (ii) 每个团队的规模有限制 (最多2名成员)，(iii) 不允许在团队之外私下共享代码，(iv) 每天提交的数量有限制 (每天最多5个条目)，(v) 不允许使用外部数据 (word嵌入除外)。 例如，不允许您使用外部数据来确定一篇论文是否引用了另一篇论文。

**评估和提交。** 您对该项目的最终评估将基于 (1) 您将给出的陈述 (**50**%)，(2) 在您在私人排行榜上的位置以及将要实现的日志丢失 (**20**%)，以及 (3) 关于您解决问题的总体方法和报告质量 (**30**%)。 作为项目的一部分，您必须提交以下内容:

* 一份4-5页的报告，您应该在其中描述您在项目中使用的方法和方法。 由于这是一个真正的分类任务，我们很想知道你如何处理管道的每个部分，例如，你如何创建你的表示，你使用了哪些功能，你使用了哪些分类算法，为什么，你的方法的性能 (日志丢失和训练时间)，这些方法最终没有奏效，但很有趣，总的来说，无论你认为报告什么有趣。
* 一个目录，其中包含您实现的代码 (不是数据，只是代码)。
* 创建一个。包含代码和报告的zip文件，并将其提交给平台。

## 截止日期: 格林尼治标准时间8 23:59 15/04/2022

**演示:**如上所述，将要求您介绍您遵循的方法。 因此，您需要准备一些幻灯片 (使用ppt或您喜欢的任何其他工具)。

## 展示日期: 待定日期

3<http://networkx.github.io/>4<http://www.nltk.org/>

5<https://radimrehurek.com/gensim/>