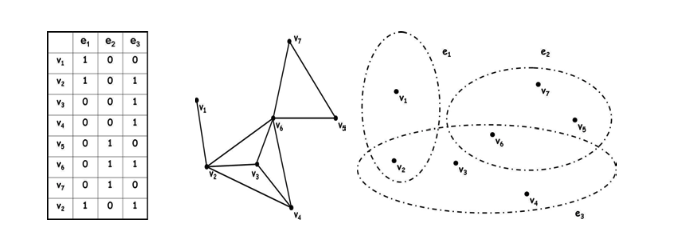
论文阅读笔记一

Learning with Hypergraphs: Clustering, Classification, and Embedding

1. Motivation
2. 通常用图来表示数据之间的关系，但是实际问题中关系不是简单的成对的关系，所以本文提出了一种用超图来表示实际问题中的复杂关系的方法。
3. Contribution
4. 用超图表示多元关系。
5. 利用超图克服多元关系压缩为二元关系的信息损失问题。
6. 本文将频谱聚类的方法应用到超图上。
7. 在频谱超图聚类方法的基础上进一步开发用于超图嵌入和转换分类的算法。
8. Method
9. 基本概念



V为顶点集合，表示对象，E为边集，表示顶点之间的连接关系。

左边的表说明了点与边之间的关系，中间的是一般图，右边的是超图，超图的每个圈圈住的点集合就是超边。

超图强调的是集合的概念，一个超边说明许多点的集合，而对于一般图，一条边说明了两个点之间的关系。超图中的每个圈会有重叠的顶点，重叠部分表示相同的属性，重叠部分越多说明两个样本是同一类的可能性越大。

用δ(e)来表示超边中包含的顶点数，若图中每条超边有一定权重w(e)，则此图为有权超图，用G=(V,E,w)表示。

1. 简单图归一化到超图

对于一个顶点子集S∈V，令Sc表示集合S的compliment。分割超图就是把G=(V,E,w)分割成S和Sc两个部分，如果一条超边同时在集合S和Sc中，则这条边将会被切断。

给定顶点子集S⊂V，定义集合S的超边边界∂S为被分割过的超边集合，定义S中的顶点的度值vol为所有顶点的度的总和，即，此外，∂S可通过顶点个数获得：

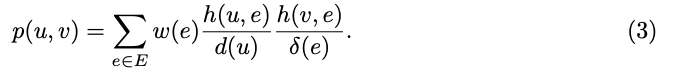


对于vol∂S=vol∂Sc，假定每条超边都是一个全连接子图，每条子超边都被分配一样的权重，当一条边e被剪掉，则子超边被剪掉。统一集合中的点是密集的，两个集合中的点的连接是稀疏的，利用前面的定义，将这种自然区分定义为：



1. 随机游走

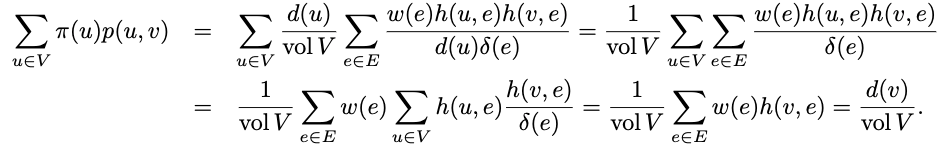
将每个超图与自然随机游走关联，该随机游走具有以下转移规则：给定位置u∈V，首先在所有超边中选择一条超边e，选择边e的概率与权重w(e)成正比，然后随机选择顶点v∈e。这个过程概括了在简单图上定义的自然随机游走。令P表示在此超多随机游走的转移概率矩阵，P的每个记录为：

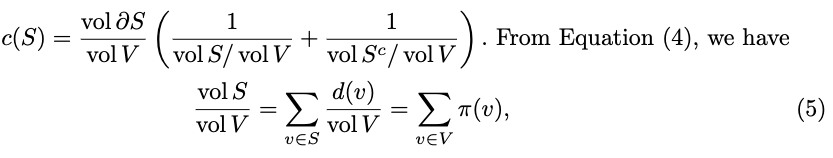


在矩阵中，，随机游走的固定分布π为：



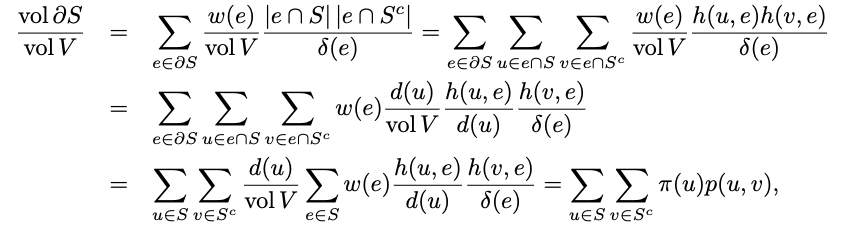
该分布来自于以下推导：





volS/volV表示随机游走在S中的顶点集。

从公式3和4可得，

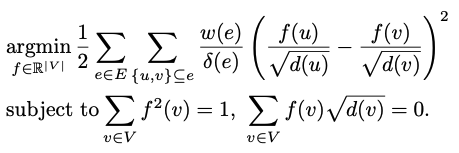


比例vol∂S/volV表示在固定分布下从S到Sc的一跳的概率。

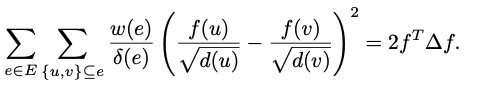
从公式5和公式6，超图规范化切割的规则如下：寻找一个使得随机游走跨越不同聚类的概率尽可能小、而随机游走保持在同一聚类中的概率最大。

1. 频谱超图分割

公式2给出的组合优化问题是一个NPC问题，可以转化为一个实值最优化问题：



定义矩阵，其中I表示单位矩阵，可以证明存在以下关系：



注意∆是半确定的，最小值为0，其特征向量为根号d。顶点集将被分为两个部分，对于一个简单图，所有边的度的矩阵De将减小到2I，因此：



1. 频谱超图嵌入

该部分证明了从松弛中得出的实值优化问题实际上是原始组合优化问题的下限，目前不清楚如何利用多个特征器模拟k路分区，在简单图形的情况下，有很多启发式的方法，可以用到这种情况下。

1. 转导推理(顶点分类)

本节建立了光谱超图分类和嵌入算法，给定超图G=(V,E,w)，子集S∈V中的顶点有{-1,1}的标签，我们的目标是预测其余未标记的顶点的标签。基本上，我们应该尝试为同一超位中包含的所有顶点分配相同的标记。

令表示分类函数，该函数将标签符号f(v)分配给顶点v∈V。给定目标函数Ω(·)，可以选择一个分类函数：



其中Remp(f)表示一个经验损失函数，比如最小平方损失或铰链损耗，μ>0是一个正则化参数。

分类函数为，其中y表示初始标签，ξ 是(0,1)间的参数，

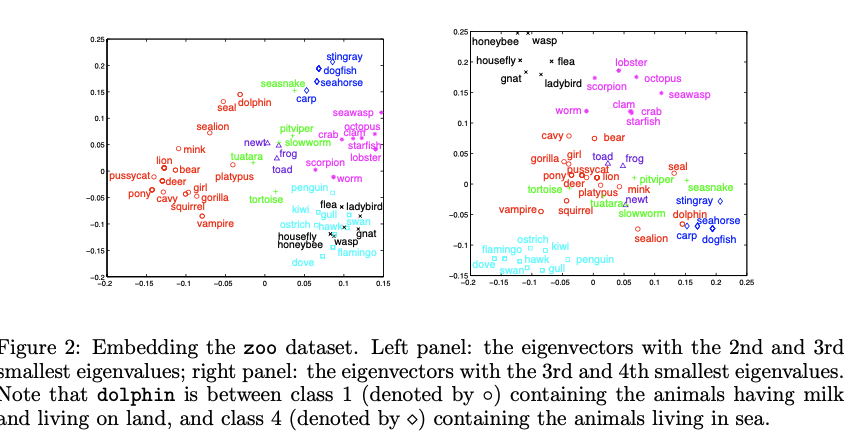
1. Exp

数据集：20个新闻组，来自UCI的机器学习分类数据库。

数据描述：每个实例都由一个或多个属性描述，每个属性只包含少量值，每个值对应于特定类别，

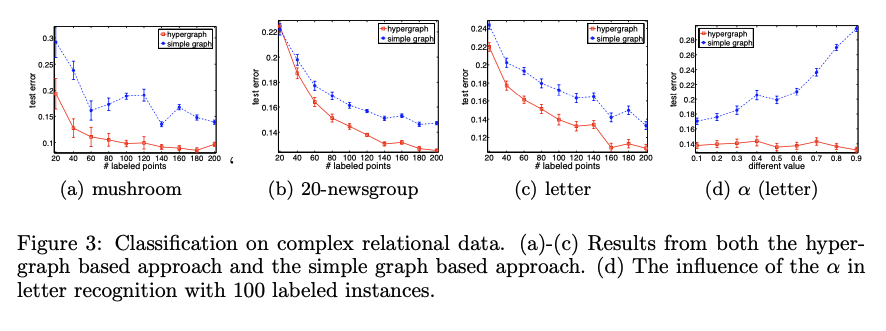
实验描述：实验构建了每个数据集的超图，每个属性值被视为超边，所有超边的权重为1，权重的设置是一个重要的问题，需要额外进行探索。实验也为每个数据集构建了简单图，将基于简单图形的频谱聚类方法作为基线模型。图实质上是通过超图的邻接矩阵定义对象之间的成对关系。

第一个任务是将动物园中的动物数据集嵌入到欧几里得空间，共100个具有17个属性的动物，属性包括头发、羽毛、鸡蛋、牛奶、我把、腿等，这些动物被手动分为7个类别。



第二个任务是对魔鬼数据集进行分类，该数据集包含22个属性（形状、颜色等属性）描述8124个实例，作者删除缺少值的第11个属性，每个实例都标注为可食用或有毒，分别为4208和3916个实例。

第三个任务是在经过修改的20个新闻组数据集上对文本进行分类，该数据集在16242篇文章中具有100个单词的二进制匹配值，文章属于4个不同的主题，每个主题分别有4605、3519、2657和5461篇文章，最后任务是彩瓷字母数据集的字母类别。



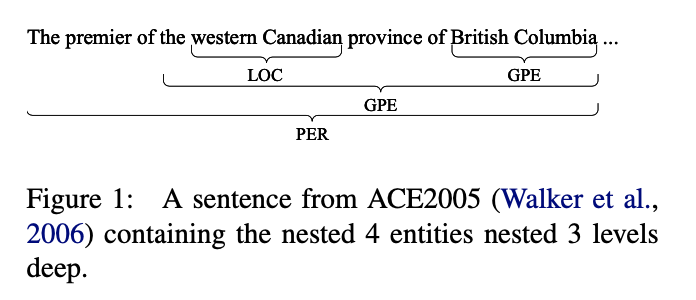
1. Advantage

将频谱聚类推广到超图，开发了超图嵌入和转导推理算法。

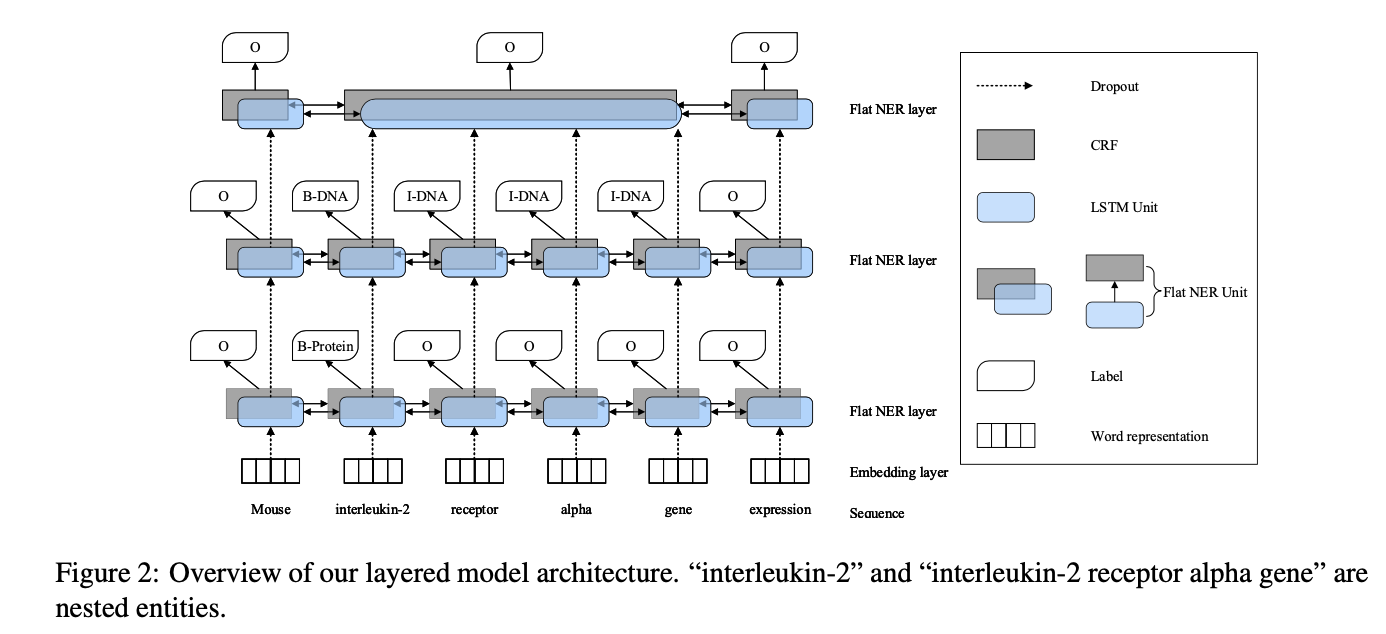
1. Idea

论文阅读笔记二

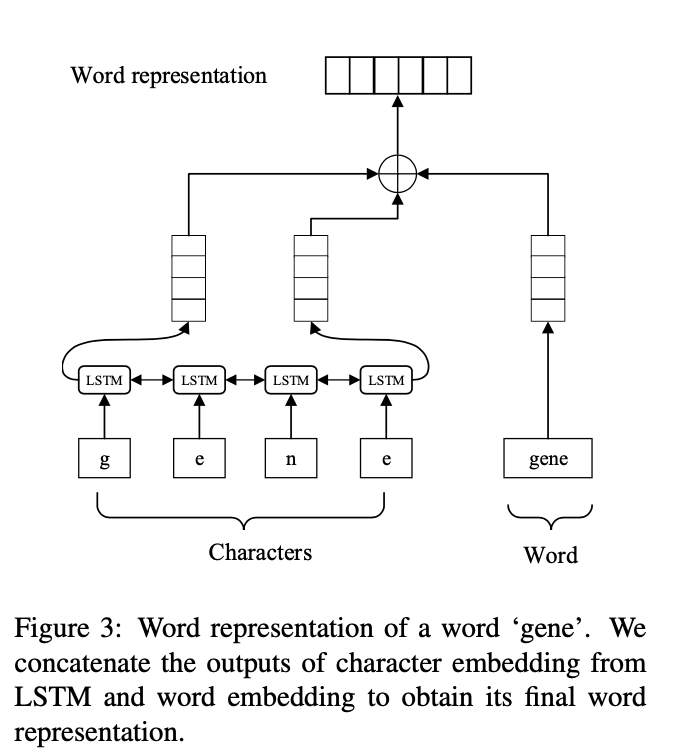
A Neural Layered Model for Nested Named Entity Recognition

1. Motivation
2. 实体嵌套在更长的实体内本看作是嵌套实体，大多数命名实体识别系统只处理flat实体，而忽略嵌套实体，这样不能捕捉文本中细粒度的语义信息。
3. 
4. Contribution
5. 本文提出了一种新型的动态堆叠flat entity layer的神经网络模型，每一个flat NER layer都是基于state-ofthe-art的BiLSTM+CRF层，在当前的flat layer会把已经被识别为实体的token经过LSTM处理的新的表示合并起来，传递给下一个flat layer，这样就使得模型能够以从内到外的方式，充分利用内部实体的编码信息识别出外部的实体，直到没有更外部的实体被识别出来。
6. Method

嵌套命名实体识别模型是由flat NER layers动态连续堆叠而成的，每个flat NER layer含有一个BiLSTM和一个CRF分别进行输入编码和标签预测，整体的模型就是Layered-BiLSTM-CRF，当前flat layer有实体被识别出来时就会引入新的flat layer，不同层之间共享参数，当前flat layer中被识别出来的实体所包含的token的编码表示合并起来作为新的flat layer的输入。如果没有识别出来实体，会停止堆叠，也就完成了所有实体的识别。



由于在第一个NER layer之前没有其他的层，它的输入与其它的flat NER layer不同。用词的character sequence来捕捉词的正交（orthographic）和形态（morphological）特征，把每个词的word embedding和character sequence embedding拼接起来作为每个词的表示，在训练的时候character embedding是随机初始化得到的，而word embeddig是用的预训练的。

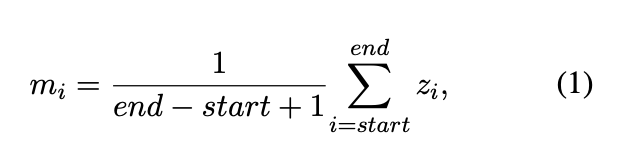


1. 平面实体

每个flat NER layer包含一个LSTM层和一个CRF层，LSTM用来进行序列的双向的上下文的表示，CRF层用LSTM得到的序列表示来进行全局的标签预测。

1. 堆叠实体

为了抽取出更外层的实体，在当前的flat NER layer的上面堆叠一个flat NER layer。把已经识别出来的每个实体所包含的token的编码表示合起来并求其平均。



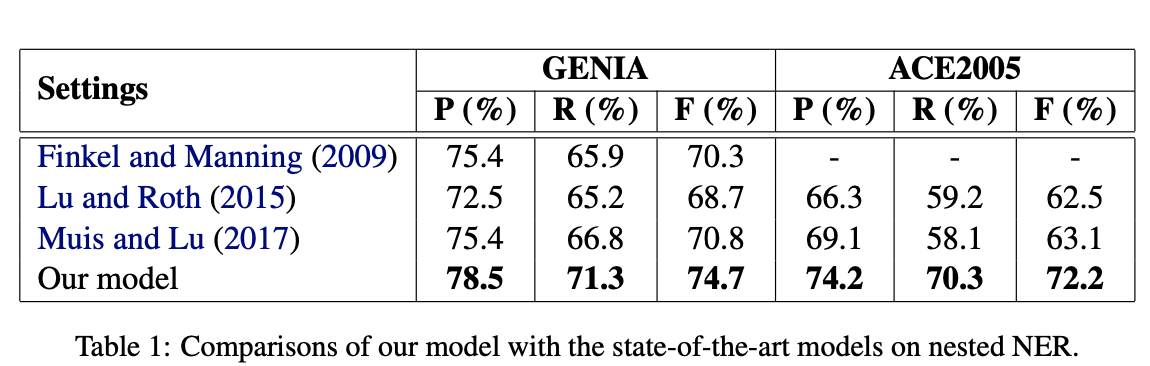
其中，zi表示flat ner层中的第i个词，mi是一个实体的合并表示，start和end是识别出来的实体在flat ner层中的位置。

1. 训练

训练时的gold label是基于BIO标注的，文本模型从内到外识别，故gold label的顺序也是从内到外，这个角detection order rule，对于同一个位置的token不能被多次标位同一个标签称之为non-duplicate rule。如图2中“Mouse interleukin-2 receptor alpha gene expression”的两个gold label sequence是{O, B-Protein, O, O, O, O}和{O,B-DNA, I-DNA, I-DNA, I-DNA, O}。

1. Experiment

在实验阶段，先去除堆叠的层，只保留第一层，在JNLPBA数据集上训练，然后用这些训练作为初始参数在GENIA和ACE05上实验



1. Advantage­­­­

能够逐层识别嵌套实体

具有很好的可迁移性

1. Idea