目录

[1 Introduction 1](#_Toc532304097)

[1.1 问题背景 1](#_Toc532304098)

[1.2相关工作 2](#_Toc532304099)

[1.2.1 知识嵌入 2](#_Toc532304100)

[1.2.2多类聚类 2](#_Toc532304101)

[2 Proposed Methods 3](#_Toc532304102)

[2.1 基于字符级别的卷积神经网络分类方法 3](#_Toc532304103)

[2.1.1 数据初步处理 3](#_Toc532304104)

[2.1.2 CNN卷积神经网络 3](#_Toc532304105)

[2.1.3 数据增强 4](#_Toc532304106)

[2.2基于transE和svm的多分类方法 4](#_Toc532304107)

[2.2.1 transE模型 4](#_Toc532304108)

[2.2.2基于SVM的多类分类 5](#_Toc532304109)

[2.3 新方法：结合字符向量和决策树的多分类方法 9](#_Toc532304110)

[3. Experiments 10](#_Toc532304111)

[3.1 基于字符的CNN方法 10](#_Toc532304112)

[3.2 结合transE以及SVM的多分类实验 11](#_Toc532304113)

[3.3 新方法：结合字符向量和决策树的多分类方法 11](#_Toc532304114)

[4. Conclusions 11](#_Toc532304115)

# 1 Introduction

## 1.1 问题背景

知识库将人类知识组织成结构化的知识系统，人们花费了大量时间来构建了各种结构化的知识库，如WordNet，Freebase等。知识库是推动人工智能学科发展和支撑智能信息化服务应用的重要基础技术。现在，大部分知识往往可以用三元组（实体1，关系，实体2）来表示。但是由这些三元组构建的网络形式主要受到计算效率以及数据稀疏问题。

近年来，以机器学习、深度学习为代表的表示学习开始引起科研界的注意。表示学习旨在将研究对象的语义信息表示为低维向量，进而使用深度学习算法计算其相似度进行多类分类。

## 1.2相关工作

### 1.2.1 知识嵌入

目前知识嵌入方法主要有距离模型、单层神经模型、能量模型、双线性模型、张量神经模型、翻译模型（trans系列）。

知识本身也是文本类型数据，所以也可以用文本分类的方法解决，而文本向量化有独热、TF-IDF、分布式表示等等。

### 1.2.2多类聚类

目前多类聚类方法主要有支持向量机的多类分类、神经网络多类分类、随机森林等。

# 2 Proposed Methods

## 2.1 基于字符级别的卷积神经网络分类方法

基于数据集（h，r，t）都是字符串文本，之前大量研究基于word级别，来生成词向量。比如one-hot、Bag-of-words and its TFIDF、Word2Vec来生成词向量。本方法是基于字符级别生成词向量并送入CNN卷积神经网络训练。

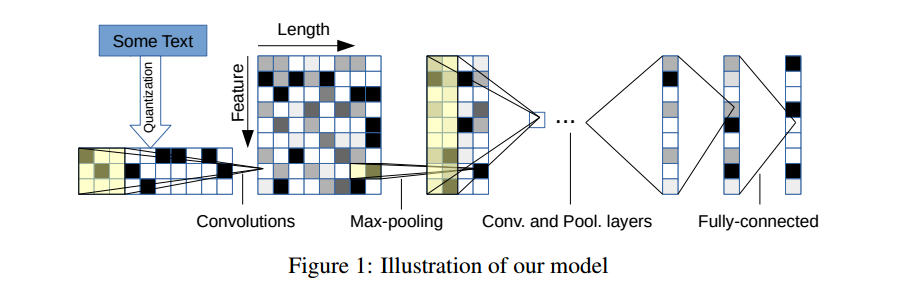
### 2.1.1 数据初步处理

基于（h，r，t）三元组的文本数据，将文件根据关系特征来标注唯一分类标签。数据如下图

|  |  |  |  |
| --- | --- | --- | --- |
| H | T | R | labels |

### 2.1.2 CNN卷积神经网络

我们输入我们的Text，通过one-hot构成约48000 X10维矩阵，矩阵每一位都是字符的表示数字。我们的cnn有6个卷积层以及3个全连接层。



CNN模型经过反向传播方法来不断更新权重。

### 2.1.3 数据增强

我们来谈下如何利用一些简单的转换方法将你现有的数据集变得更大。比如将矩阵某一列交换而不改变标签。这种改变训练数据的数组表征而保持标签不变的方法被称作数据增强技术。这是一种人工扩展数据集的方法。人们经常使用的增强方法包括灰度变化、水平翻转、垂直翻转、随机编组、色值跳变、翻译、旋转等其他多种方法。通过利用这些训练数据的转换方法，我们将获得两倍甚至三倍于原数据的训练样本，以此训练模型更有效。

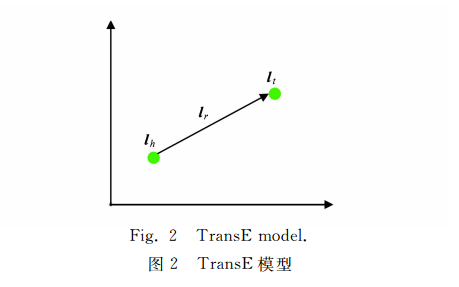
## 2.2基于transE和svm的多分类方法

### 2.2.1 transE模型

在知识图谱的领域，如何将三元组表现成向量也成为我们的关注热点。Mikolov等人提出了word2vec词表示学习模型和工具包。利用该模型，他们发现了词向量空间存在有趣的平移不变现象。例如



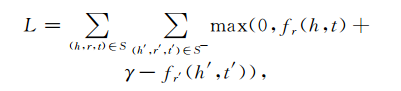
基于上述现象，Bordes等人提出了TransE方法，从而通过（h，t，r）不断的随机打坏h以及t的结构,进而构造出h+t≈r的关系。并用向量表示。



他们定义了如下损失函数



同时也定义了如下优化目标函数，通过不断替换h，t。形成新的h1和t1。进而计算距离并优化



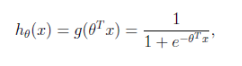
### 2.2.2基于SVM的多类分类

#### 2.2.2.1 SVM的起源：logistic回归

SVM最初用来做二分类问题，本质是寻找一个超平面将二维数据集（Xi，Yi）切分开，这个超平面方程为：

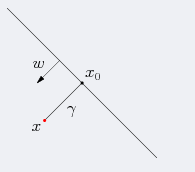


同时Svm的假设函数为，实质上这里就已经决定了我们的分类类别：

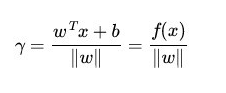


#### 2.2.2.2 几何间隔

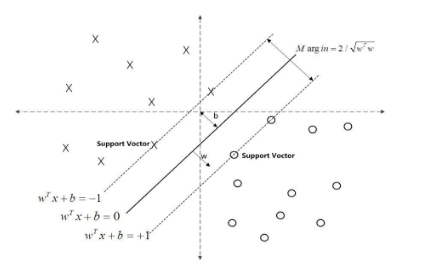
在超平面w\*x+b=0确定的情况下，|w\*x+b|能够表示点x到距离超平面的远近。这有利于我们找出离正负数据集距离最远的超平面。



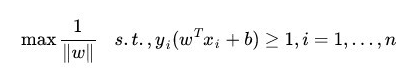
如上图所示，我们需要计算出x离超平面的距离并以最大化该值为目标函数。为此我们引出几何间隔的定义：



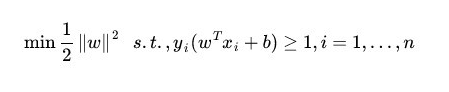
这里的是计算x和超平面距离，现在的问题变成了如何求的最大值，即最优超平面。如下图所示，如果我们初始化f（x）=1，那么而对于所有不是支持向量的点，则显然有https://img-blog.csdn.net/20131111155205109。而虚线上面的点都满足支持向量。



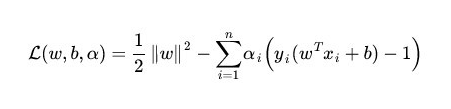
考虑到之前的函数，我们可以得出优化目标函数：



我们可以试着将其转化对下面这个公式的求解：



因为现在的目标函数是二次的，约束条件是线性的，所以它是一个凸二次规划问题。可以总结为：在一定的约束条件下，目标最优，损失最小。（这其实就是高数中的二次求极值问题）我们可以定义拉格朗日函数（通过拉格朗日函数将约束条件融合到目标函数里去来求解）：

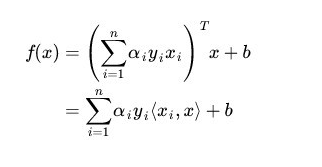


通过拉格朗日函数以及对w、b求偏导（定住a），并结合SMO算法我们求解拉格朗日乘子https://img-blog.csdn.net/20131111195836468。

#### 2.2.2.3 核函数

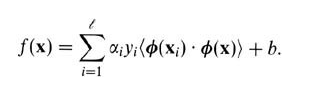
通过上述我们得出了最大几何间隔。但是，我们的SVM还不能处理线性不可分问题，为此我们需要引入核函数的概念。

通过对前面公式的推导，我们可以得出SVM的分类函数为：



我们发现对于新点 x的预测，只需要计算它与训练数据点的内积即可。而对于我们的非支持向量来说，系数a是为0的。所以对新点x的预测只需要和少量的支持向量內积即可。这样，我们就得到一个支持向量机。那么为了处理线性不可分，SVM处理方法是选择一个核函数来在低维计算，再映射到高维空间。

如果有一种方式可以**在特征空间中直接计算内积〈φ(xi · φ(x)〉，**就像在原始输入点的函数中一样，就有可能将两个步骤融合到一起建立一个非线性的学习器，**这样直接计算法的方法称为核函数方法，**那么决策规则可以用训练点和测试点的內积表示：



这里φ是从X到内积特征空间F的映射。

#### 2.2.2.4 基于SVM实现多分类

基于SVM多类分类主要有OAR法、OAO法、DAG法、二叉树法、纠错编码法、模糊SVM法。基于我们对论文《基于支持向量机的多类分类算法综述》的研究，这里对于大量类别的大数据应该是使用DAG-SVMs方法更好，但由于编程能力有限未能实现DAG-SVMs算法。我们这里使用的基于sklearn官网的SVM.SVC方法，本方法使用OAO方法对数据进行多分类。

OAO 法的基本思想是在 N 个类别中的每二类之间均训练一个二类分类器，得到 (N)- 1 N/2 个分类器和分类函数。在测试未知样本时，将其分别代入每个分类函数，对各函数判别结果采用投票的方式记录，得票最多的类别判定为未知样本的类别。这种方法的优点是每个二类分类器只需训练二类样本，简单快速。但需要训练的分类器数量较多，特别是 N 较大时；当不止一类得票最多时，会出现样本的误分类。

## 2.3 新方法：结合字符向量和决策树的多分类方法

2.3.1 对数据集进行去除常用词和停用词操作，再对实体数据进行基于字符级别的向量构建。再利用决策树实现多分类。

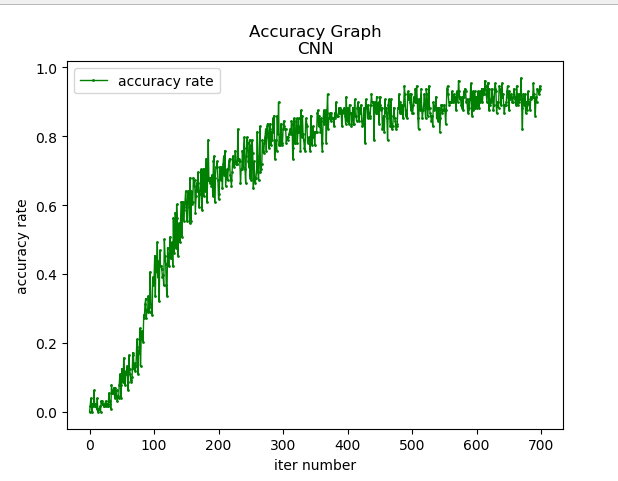
我利用这种方法，得出一个483142\*14维的矩阵，利用决策树方法进行多分类。

# 3. Experiments

## 3.1 基于字符的CNN方法

我使用了老师所给的数据集，在服务器上跑通实验，实验结果如下所示。

在图中可以看出，在迭代700次时，准备率已经可以达到90%，这基于CNN强大的反向传播算法和训练学习。对比下面用TransE模型的方法而言,准确率要高很多，主要是CNN是基于字符细粒度的训练，从而准确率更高。



## 3.2 结合transE以及SVM的多分类实验

经过对数据的测试，我们得出了在测试集的准确率在

48.55%左右。

## 3.3 新方法：结合字符向量和决策树的多分类方法



经过对数据的测试，我们得出了在测试集的准确率在31%左右，效果不是很好。

# 4. Conclusions

对于我整个工作，主要是使用了深度学习CNN、知识表示以及机器学习多分类方法的使用。这些方法都是通过分析问题，查看论文对问题的解决方案，并对比解决方案找出最优，并上网找源码跑实验的方式来完成。

不足之处还是有很多，本问题主要是对知识表示和多类分类两个子问题的求解。对于前者，已有十几种方法将文本转成向量或者张量，但由于精力和时间原因，未能深究原理。对于后者，也已经有了机器学习、深度学习多种方法预测，大方向应是在了解各种算法原理基础上，作出最佳分类结合策略。