目录

[Svm-struct 方法 2](#_Toc360370525)

[基于SVM-Struct 的中文句法分析: 2](#_Toc360370526)

[Svm\_cfg 计算流程 4](#_Toc360370527)

[测试结果: 6](#_Toc360370528)

Svm-struct 方法:

支持向量机SVM(Support Vector Machine) 是新型机器学习方法,具有完备的统计学理论基础,它采用结构风险最小化原则代替传统统计学的基于大样本的风险最小化原则,具有出色的学习和推广能力。 但在实际应用中,大部分要处理的数据(如树形结构、队列结构或网状结构数据等) 都是比较复杂彼此之间存在相互依赖,具有特定结构化关系的,而用传统的支持向量机已经很难去处理这些问题了,因此,结构化支持向量机(SVM2Struct) 根据数据内部的结构性提出结构化函数*ψ( x , y)* 对传统的支持向量机进行了改进,能有效地处理结构化数据。

结构化数据分析问题的目的是要找到样本的输入与输出对之间的一个函数f : X →Y 假定函数f 的形式为:



其中F 是判别函数:  w 是需要训练的权值向量, 是

输入x 和标注y 映射为特征向量的特征函数.

SVM-Struct 通过训练得到w 权向量,使:





定义: 

采用最大间隔法 ,引入松弛变量的最优化问题为:



s.t 



自然语言处理中 的计算额外的复杂所以引入损失函数: 



基于SVM-Struct 的中文句法分析:

目前中文句法分析研究中常采用两类方法,一类是浅层句法分析(基于标注的),也叫部分句法分析或语块分析,主要任务是语块的识别和分析. 另一类是完全句法分析,通过一系列分析过程,最终得到句子的完整的语法树. 本文的研究方法属于完全句法分析,它基于SVM-Struct 方法,主要通过构造结构化函数*ψ*( x , y) ,对中文句法进行分析,最终得到句子的完整语法树。

本文根据中文句子的结构特点构造的结构化函数*ψ*( x , y) 为一列向量,其形式如下:



向量中的元素为语法规则出现的次数。

一个句法分析示例：





一个句子样本x 通过函数 得出与之对应的句法树y。

测试所用到的汉语词类标注及汉语短语标注表1，表2：





　句法树y 中的每个结点对应着语法规则g j 以及权值w j . 句法树y 中所有结点对应的w j 的总和将代表着将样本x 分类为y 的一个评价值. 这个值可以通过函数 得到,其中*ψ( x , y)*列向量中的每个元素对应着句法树y 中每个g j 所出现的次数 . 然后通过式(1) 最大化函数F,就可以得到正确的y。

Svm\_cfg 计算流程

基于SVM-Struct 的中文句法分析算法的主要步骤如下:

1. 输入训练样本( x1 , y1 ) , ⋯, ( xn , yn ) ,设置参数C,ε,并选定损失函数类型



1. 初始化工作集 为空集, i = 1 , ( i = 1 , ⋯, n)

初始化

1. 主循环0
   1. 
      1. 主循环1:
      2. for i =1…n do
         1. 计算 .
         2. 如果 > 

则:



* + - 1. 二次规划更新**** 和 ( 实际程序的实现过程中 的元素增加了一定的数量才执行一次二次规划更新 )



s.t 



* + - 1. End for
    1. 所有的元素不再更新推出主循环1.
  1. 达到精度要求退出迭代(主循环)

计算 采用CKY算法进行求解.

测试结果:

测试所用数据来自北京大学计算语言学研究所公开的微型语料库中的树库样例 (1372 句) ,句长大约都在5 - 20 个词之间.

