FudanNLP 文档

邱锡鹏
计算机学院媒体计算研究所
复旦大学
xpqiu@fudan.edu.cn

2011 年 5 月 4 日

摘要

FudanNLP 主要是为中文自然语言处理而开发的工具包，也包含为实现这些任务的机器学习算法和数据集。本工具包及其包含数据集使用 LGPL3.0 许可证。FudanNLP 是基于 Java 的开源项目，利用统计机器学习和规则方法来处理中文自然语言处理的经典问题，比如：分词、词性标注、句法分析、实体名识别等。

目录

第一章　FudanNLP 结构

1.1　组织结构

FudanNLP 的组织结构可分为 5 层，如图1.1所示。

1. 最底层的操作。比如数据结构、数据表示、数据类型、数据预处理、特征转换等。
2. 结构化机器学习和人工规则框架。涉及到特征抽取，学习算法、推理算法和模型建立等。
3. 可插拔的具体算法。比如分类、聚类、半监督和优化等。
4. 中文自然语言处理应用，比如分词、句法分析等。
5. 信息检索应用，比如文本分类、主题词抽取等。

1.2　文件组织结构

1. /src 源代码，主目录
2. /test 源代码，测试或单元测试
3. /example 源代码，对外 API 使用示例
4. /example-data 使用示例需要的数据
5. /app 基于 FudanNLP 的应用
6. /model 模型文件
7. /doc 项目文档
8. /lab 所有进行中的代码，研究代码等

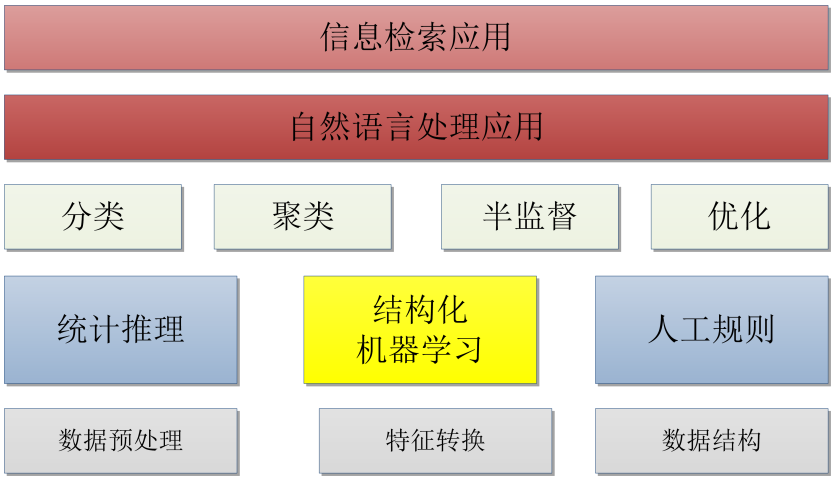

图 1.1: FudanNLP 组织结构图

第二章　结构化机器学习简介

结构化机器学习就是处理的样本 (x,y)，y 不再是离散的类别，而是有结构的，比如序列、树结构等。比如对于序列，$\mathrm{y} = y_1,\ldots,y_L$，$\mathrm{x} = x_1,\ldots,x_L$。$L$ 是序列的长度。这样不能直接用传统的学习方法。因为结构的组合数非常多，需要用推理的方法来求解 y。

结构化机器学习可以分解为下面步骤：

2.1　特征生成

对于一个样本 (x,y)，我们定义 $\Phi(\mathrm{x},\mathrm{y})$ 为特征。

2.2　推理

对于未知的 x，$\hat{\mathrm{y}}$ 可以通过一个得分函数求得，

$$\hat{\mathrm{y}} = \arg\max_{\mathrm{z}} F(\mathrm{w},\Phi(\mathrm{x},\mathrm{z})), \tag{2.1}$$

这里，w 是函数 $F(\cdot)$ 的参数.

函数 $F(\cdot)$ 一般为线性函数或指数函数。

2.3　损失计算

2.4　学习

学习函数 $F(\cdot)$ 的参数 w。根据 $F(\cdot)$ 的形式不同，学习参数的方法也不同，一般为最大似然、最大边际距离或最小均方误差等。

2.4.1　PA 算法

给定一个样本 (x, y)，$\hat{\mathrm{y}}$ 定义为错误标签中得分最高的标签。

$$\hat{\mathrm{y}} = \arg\max_{\mathrm{z} \neq \mathrm{y}} w^T \Phi(\mathrm{x}, \mathrm{z}). \tag{2.2}$$

边际距离 $\gamma(\mathrm{w}; (\mathrm{x}, \mathrm{y}))$ 定义为：

$$\gamma(\mathrm{w}; (\mathrm{x}, \mathrm{y})) = \mathrm{w}^T \Phi(\mathrm{x}, \mathrm{y}) - \mathrm{w}^T \Phi(\mathrm{x}, \hat{\mathrm{y}}). \tag{2.3}$$

我们定义 hinge loss.

$$\ell(\mathrm{w}; (\mathrm{x}, \mathrm{y}) = \begin{cases} 0, & \gamma(\mathrm{w}; (\mathrm{x}, \mathrm{y})) > 1 \\ 1 - \gamma(\mathrm{w}; (\mathrm{x}, \mathrm{y})), & \text{otherwise} \end{cases} \tag{2.4}$$

PA 算法用在线的方式计算更新参数。在第 t 轮, 通过下面公式计算 w_{t+1}：

$$\mathrm{w}_{t+1} = \arg\min_{\mathrm{w} \in \mathbb{R}^n} \frac{1}{2} ||\mathrm{w} - \mathrm{w}_t||^2 + \mathcal{C} \cdot \xi, \\ \text{s.t. } \ell(\mathrm{w}; (\mathrm{x}_t, \mathrm{y}_t)) <= \xi \text{ and } \xi >= 0 \tag{2.5}$$

这里，C 是正的参数来控制松弛变量的作用。

我们用 ℓ_t 来表示 $\ell(\mathrm{w_t}; (x, y))$。如果 $\ell_t = 0$，那么 w_t 满足 Eq. (2.5)。因此我们只关心 $\ell_t > 0$ 的情况。

最终我们得到更新策略为（推导过程可参考 [2]）：

$$\mathrm{w_{t+1}} = \mathrm{w}_t + \bar{\alpha}^* (\Phi(\mathrm{x}, \mathrm{y}) - \Phi(\mathrm{x}, \hat{\mathrm{y}})). \tag{2.6}$$

其中，

$$\bar{\alpha}^* = \min(\mathcal{C}, \bar{\alpha}). \tag{2.7}$$

其中，

$$\bar{\alpha} = \frac{1 - \mathrm{w_t}^T (\Phi(\mathrm{x}, \mathrm{y}) - \Phi(\mathrm{x}, \hat{\mathrm{y}}))}{||\Phi(\mathrm{x}, \mathrm{y}) - \Phi(\mathrm{x}, \hat{\mathrm{y}})||^2}. \tag{2.8}$$

算法如1所示。为了避免过拟合，我们使用平均的策略 [1]。

输入: 训练集: $(\mathrm{x}_n, \mathrm{y}_n), n = 1, \cdots, N$，参数：$\mathcal{C}, K$
输出: w

初始化：cw $\leftarrow 0$,;
for $k = 0 \cdots K-1$ do
  $\mathrm{w}_0 \leftarrow 0$;
  for $t = 0 \cdots T-1$ do
    挑一个样本 $(\mathrm{x}_t, \mathrm{y}_t)$;
    预测：$\hat{\mathrm{y}}_t = \arg\max_{\mathrm{z} \neq \mathrm{y}_t} \langle \mathrm{w}_t, \Phi(\mathrm{x}_t, \mathrm{z}) \rangle$;
    计算 $\ell(\mathrm{w}; (\mathrm{x}, \mathrm{y}))$;
    用 Eq.(2.6) 更新 w_{t+1};
  end
  $\mathrm{cw} = \mathrm{cw} + \mathrm{w}_T$;
end
$\mathrm{w} = \mathrm{cw}/K$;

Algorithm 1: PA 算法

第三章　程序结构

FudanNLP 项目大概结构组织如下：

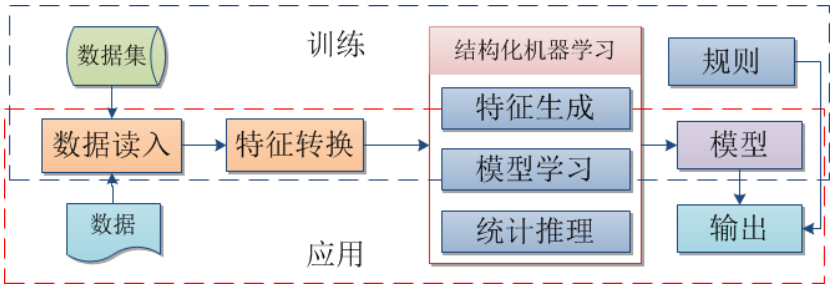

图 3.1: FudanNLP 总体流程图

3.1　数据类型：edu.fudan.ml.types

3.1.1　Instance 类

表示单个样本

3.1.2　InstanceSet 类

样本集合

3.2　数据读取：edu.fudan.ml.data

通过 Reader 接口将原始数据读入，并生产 Instance 对象。Reader 为一个迭代器，依次返回一个 Instance 对象。

3.3　数据特征变换：edu.fudan.ml.pipe

这里进行数据不同形式表示之间的转换。比如从文本到向量的转换。

3.4　机器学习：edu.fudan.ml.classifier

包括分类器和训练器两部分。按照结构化学习的思想分为：特征生成、损失函数和统计推理三部分。

3.4.1　**特征生成**：edu.fudan.ml.feature.generator

3.4.2　**损失函数**：edu.fudan.ml.loss

3.4.3　**统计推理**：edu.fudan.ml.solver

这里对于离散的类别，使用简单遍历计算，然后求最大值得方法。

3.5　结构化机器学习：edu.fudan.ml.struct.*

结构化学习相应的类

3.6　结构化机器学习：edu.fudan.ml.hier.*

层次结构类标签的多任务学习相应的类

3.7　关键词抽取：edu.fudan.nlp.keyword

关键词抽取

3.8　序列标注：edu.fudan.nlp.tag

序列标注以及在此基础之上的分词、词性、实体名识别等。

3.9　句法分析：edu.fudan.nlp.parser

句法分析相关类

3.10　文本分类：edu.fudan.nlp.tc

文本分类相关类

第四章　总结

FudanNLP 是以统计机器学习为基础，并结合人工规则来处理中文自然语言以及信息检索、信息抽取的各种任务。远景目标是实现一个更够实际应用的中文自然语言处理产品，虽然这个目标还比较远。

参考文献

[1] Michael Collins. Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. In Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing, 2002.

[2] Wenjun Gao, Xipeng Qiu, and Xuanjing Huang. Adaptive chinese word segmentation with online passive-aggressive algorithm. In Proc. of CIPS-SIGHAN Joint Conference on Chinese Language Processing, pages 240–244, 2010.

致谢

本项目由以下人员合作完成：计峰、高文君、缪有栋、赵嘉亿、曹零、田乐等。此外，复旦大学计算机学院媒体计算研究所的部分其他研究生和本科生也贡献了部分代码。在此一并感谢大家的工作。