清华大学深圳研究生院

计算智能实验室 周报

姓名	研究方向	报告覆盖时间
李晨辉	NLP（法律文本分类）	2010.3.18-2019.3.24

本周主要完成的工作

1.统计文本分类与情感分类算法学习；

2.EM、SVM 算法复习

1. 语言模型及概率图模型

1.文本分类与情感分类

文本分类

- 文本分类系统类型

  基于知识工程的分类系统

  基于机器学习的分类系统

- 文本表示
  - ✓ 要求

    能够真实地反映文本的内容（主题、领域或结构等）

    对不同的文档具有区分度
  - ✓ 向量空间模型（VSM）

    由文档、特征项、项的权重三部分组成

    各特征项互译（无重复）

    各特征项之间没有顺序关系
- 文本特征选择
  - ✓ 基于文档频率的特征

    从训练预料中统计包含某个特征的文档数量，去掉小于某个阈值的特征（太少没有代表性）和大于某个阈值的特征（太多没有区分度）
  - ✓ 信息增益法

    信息增益的计算方法

    不考虑任何特征时文档的熵值减去考虑该特征文档后文档的熵得到的差值；理论上是最好的方法，但容易出现数据稀疏的问题（增益高的特征出现的频率较低）
  - ✓ 卡方统计量（CHI）

    CHI 衡量的是特征项和文本类别之间的关联程度，并假设特征 ti 和类别 Cj 之间符合具有一阶自由度的卡方分布。特征对于某个类别的 CHI 越高说明其与该类之间的相关性越大

对于多类别：

1.分别计算特征 i 对每个类别的 CHI 值，在整个训练语料上计算该特征对所有类别的最大值，去除统计量低于某个阈值的特征，保留高于给定阈值的特征作为文本特征

2.计算各特征对各类别的 CHI 均值，以这个均值作为各类别的 CHI 值

- ✓ 互信息法 （MI）

  互信息计算

  log(特征与类别的联合概率 / 特征及类别独立的概率分布的乘积)；互信息越大，特征与类别的共现程度越大；特征与类别无关，则其互信息为 0；对于多个类别的处理方法与 CHI 一样的两种方法

- 特征权重的类型

  TF

  TF-IDF

  TFC（归一化 TF）

  ITC

  TF-IWF

  熵权重

- 分类器设计
  - ✓ 朴素贝叶斯分类器

    利用特征项和类别的联合概率来估计给定文档的类别
  - ✓ SVM

    用文本向量特征做分类
  - ✓ kNN（k 近邻）

    给定一个测试文档，系统在训练集中查找距离其最近的 k 个文档，并根据这些邻近文档来给该文档的候选类别评分
  - ✓ 基于神经网络的分类器

    实际效果比 SVM 和 kNN 差，所以并不常用
  - ✓ 线性最小平方拟合法
  - ✓ 决策树分类器
  - ✓ 模糊分类器

    通过计算待分类文本的模糊集与每个文本类别模糊集的关联度 SR 实现
  - ✓ Rocchio 分类器

    首先为每一个训练文本建立一个特征向量，然后使用训练文本的特征向量为每个类建立一个原型向量（类向量）。对于待分类文本计算其与每个类向量之间的距离决定其类别。

- ✓ 基于投票的分类方法
  - ✓ Bagging 算法

    利用从训练集中抽取 R 次文档得到的训练集训练的 R 个分类器，对于新文档用这 R 个分类器分别判断，得到票数最多的类别即为新文本的类别
  - ✓ Boosting 算法

    同样是训练多个分类器，但训练每个分类器的训练集不再是随机抽取，而是由其他分类器给出的”最富信息”的样本点组成。
- ✓ 评价指标

  P、R、F1

  微平均

  不计类别只根据分类对错计算的 P 和 R

  宏平均

  对每个类别分别计算 P、R 然后取平均值

- 情感分类
  - ✓ 按机器学习方法分类

    有监督方法

    半监督方法

    无监督方法
  - ✓ 按研究问题分类

    领域相关

    数据不平衡问题

下周工作主要安排

(1) 完成《统计学习方法》中经典算法的复习；

(2) 复现一种文本分类算法；