基于词频分类器集成的文本分类方法

姜远 周志华

组成

- 问题描述
- 困难点
- 解决方法
- 实现过程
- 表现
- 评价

问题描述

- 高效地分类文本！

文本的特征

高维向量

实时性

困难

计算代价

数据更新

解決方法 (之一)

很菜的分类器

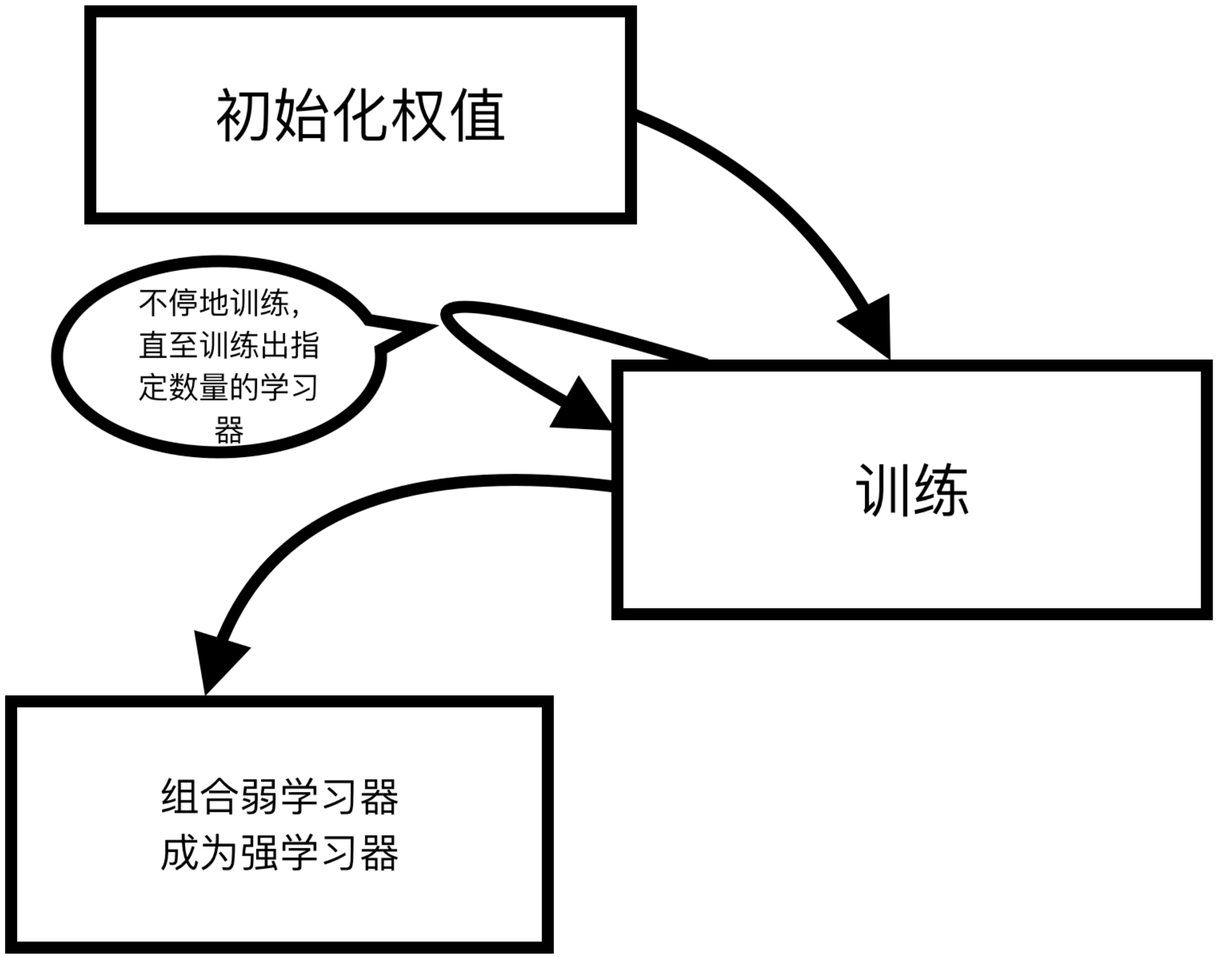
集成

什么是菜?

学习算法对新鲜样本的适应能力

接下来都是实现过程

爱听不听


初始化权值
不停地训练，直至训练出指定数量的学习器
训练
组合弱学习器
成为强学习器

基分类器

因此，我们可以将单词和它出现的频率一起作为一个基分类器. 若文本 $\boldsymbol{d}_i$ 表示为矢量（v_{i1}，…，v_{in}），v_{ik}表示 V 中的第 k 个单词 t_k 在文本$\boldsymbol{d}_i$ 中出现的次数，此时，基分类器被定义为

$$h_{t_k,f}(\boldsymbol{d}_i)=\begin{cases}1, \text{ if } v_{ik}\geqslant f,\\ 0, \text{ if } v_{ik}<f.\end{cases} \tag{2}$$

基分类器（举例）

f(”Taoxianping is smart”)=”Truth”

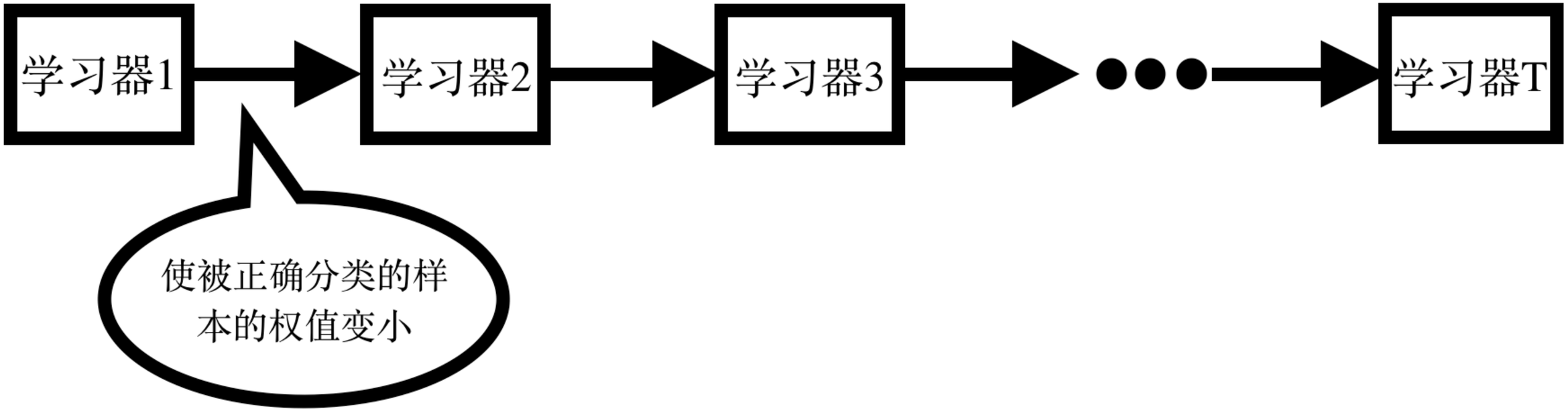
f(“Majun is old”)=“Lie”

V={’Majun’, ’Old’, ’Is’, ’Taoxianping’, ’smart’}

$$h_{majun,0}(d_1) = 1$$
$$h_{majun,0}(d_2) = 1$$
$$h_{majun,1}(d_2) = 1$$
$$h_{majun,2}(d_2) = 0$$

训练细节

每次都选择误差最小的学习器



组合弱学习器

好学生和差学生

$$H(x) = sign(\sum_{t=1}^{T} \alpha_t h_t(x))$$

$$\text{where } \alpha_t = \log(\frac{1}{\beta_t})$$

$$\text{where } \beta_t = \frac{\epsilon_t}{1 - \epsilon_t}$$

Reuters-21578

表现

Table 1 Experimental Data
表 1 实验数据的情况

Class Name	Number of Training Examples	Number of Testing Examples
Earn	2877	1087
acq	1650	719
money-fx	538	179
grain	433	149
crude	389	189
trade	368	118
interest	347	131
wheat	212	71
ship	197	89

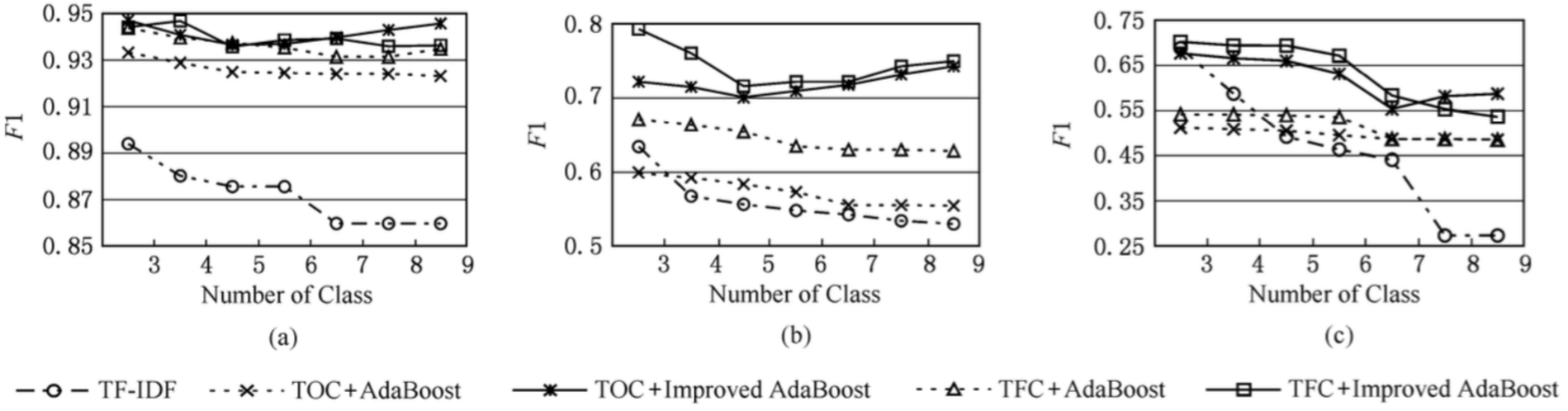


Fig. 1 Comparison of $F1$ values. (a) $F1$ on class "earn"; (b) $F1$ on class "acq"; and (c) $F1$ on class "money-fx".

图 1 $F1$ 值的比较. (a) "earn"类上 $F1$ 值;(b) "acq"类上 $F1$ 值;(c) "money-fx"类上 $F1$ 值

评价

- 性能较好
- 提出了新方法
- 缺乏运行速度比较
- 如何更新

不用谢！