计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2018,27(1):132−136 [doi: 10.15888/j.cnki.csa.006141]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于联合分类器过滤噪声的微博主题发现①

高　森[1,2], 严　曙[1], 崔超远[1], 孙丙宇[1], 汪六三[1]

[1](中国科学院 合肥物质科学研究院 智能机械研究所, 合肥 230031)
[2](中国科学技术大学, 合肥 230026)

摘　要: 伴随着互联网的广泛流行, 以微博为代表的社交网络产生了大量的数据. 从这些数据中挖掘到有用的信息成为当今研究的一项重要方向. 根据微博文本的特点, 本文提出来一种基于联合分类器过滤掉噪声微博, 然后利用LDA 模型进行主题发现. 联合分类器模型是由朴素贝叶斯、支持向量机和决策树三种模型通过简单投票机制结合构成的, 实验结果联合分类器的准确度达到 87%, 显然这种分类方法是可行的, 也是有效的.

关键词: 支持向量机; 朴素贝叶斯; 决策树; 联合分类器; LDA 模型

引用格式: 高森,严曙,崔超远,孙丙宇,汪六三.基于联合分类器过滤噪声的微博主题发现.计算机系统应用,2018,27(1):132–136. http://www.c-s-a.org.cn/1003-3254/6141.html

Microblogging Theme Discovery Based on Combined Classifier Filtering Noise

GAO Sen[1,2], YAN Shu[1], CUI Chao-Yuan[1], SUN Bing-Yu[1], WANG Liu-San[1]

[1](Intelligent Machinery Research Institute, Hefei Institutes of Physical Science, Chinese Academy of Sciences, Hefei 230031, China)
[2](University of Science and Technology of China, Hefei 230026, China)

Abstract: With the popularity of the Internet, microblogging as a representative of the social network has generated a lot of data. Exploring useful information from these data has become an important direction for today's research. According to the characteristics of microblogging text, this paper presents a method based on joint classifier to filter out noise microblogging, and then uses LDA model for subject discovery. The joint classifier model is composed of naive Bayesian, support vector machine and decision tree. The accuracy of the combined classifier is 87%, which can clearly show that this classification method is feasible and effective.

Key words: support vector machine; naive Bayesian; decision tree; joint classifier; LDA model

随着互联网技术和信息技术的迅速发展, 微博等一些社交媒体正改变着人们的生活. 由于微博的广泛流行, 微博中产生大量的数据, 这些数据有着非常大的潜在价值. 目前进行微博主题发现的方法主要是使用LDA 模型[1], 但是由于微博内容比较短, 内容比较随意, 并不是所有的微博内容都是与用户兴趣相关的. 所以可以把微博分为两类: 与用户兴趣相关的微博和与用户兴趣不相关的微博[2]. 与用户兴趣不相关的微博也就是'噪声微博'的存在, 会很大程度上影响微博主题发现的质量. 传统的文本分类方法主要有支持向量机(SVM), 朴素贝叶斯 (Bayes) 和决策树 (Tree) 三种方法, 由于微博文本的特点, 微博文本短而且内容形式随意没有规律, 所以传统的文本分类方法对微博数据分类的效果并不是很好, 所以本文分别采用了 bagging 学习算法来提高分类法的准确率. 即先使用支持向量机, 朴素贝叶斯和决策树三种方法对标注好的微博内容进行训练, 得出三种预测函数序列进行投票, 得到一个分类器, 然后利用这个联合分类器去除噪声微博, 在此基础上再进行 LDA 主题发现[3-9], 发现微博主题分类的质量有很大的提升.

① 基金项目: 中科院 STS 项目 (KFJ-SW-STS-144); 宁夏科技攻关项目 (ZNNFKJ2015-04)
收稿时间: 2017-04-06; 修改时间: 2017-04-26; 采用时间: 2017-04-27; csa 在线出版时间: 2017-12-22

1 基本原理

1.1 朴素贝叶斯分类的基本原理

在分类算法中, 朴素贝叶斯分类 (Bayes Classifier) 因其简单和容易理解的特性, 被广泛使用, 基本思想是: 对于给出的待分类项, 求解在此项出现的条件下各个类别出现的概率, 概率最大的那个类别就认为该项属于这个类别. 正式定义如下:

(1) 设$X=\{a_1,a_2,...,a_m\}$为一个待分类项, 而每个 a 为 x 的一个特征属性.

(2) 有类别集合$C=\{y_1,y_2,...,y_n\}$.

(3) 计算$P(y_1|x),P(y_2|x),...,P(y_n|x)$.

(4) 如果$P(y_k|x)=\max\{P(y_1|x),P(y_2|x),...,P(y_n|x)\}$, 则$x\in y_k$.

假设各个特征属性是条件独立的, 则根据贝叶斯定理可以推导出:

$$P(y_1|x)=\frac{P(x|y_i)P(y_i)}{P(x)}$$

而$P(x|y_i)$可以由已知的分类数据统计获得.

1.2 支持向量机分类的基本原理

支持向量机 (SVM) 是一种十分常用的二分类模型. 支持向量机 (SVM) 的定义是: 给定线性可分训练数据集, 通过间隔最大化或等价的求解相应的凸二次规划问题学习得到的分离超平面为:

$$w^*x+b^*=0$$

以及相应的分类决策函数:

$$f(x)=sign(w^*x+b^*)$$

称为线性可分支持向量机.

1.3 决策树分类的基本原理

决策树二分类器模型是一个典型二叉树结构. 决策树分类器模型经常用来做二分类, 决策树上的非叶节点表示在一个特征属性上的测试, 将测试的数据分为两类. 文本使用信息增益来做决策树的特征属性选择. 信息增益基于香浓的信息论, 找出的属性具有这样的特点: 以属性分裂前后的信息增益比其他属性最大. 这里信息的定义如下:

$$Info(D)=-\sum_{i=1}^{m}p_i\log_2(p_i)$$

1.4 联合分类器原理

联合分类器是三种经典的文本分类方法: 朴素贝叶斯分类器、决策树分类器和支持向量机 (SVM) 分类器结合构成的[10-14]. 先用朴素贝叶斯分类器、决策树分类器和支持向量机 (SVM) 分类器在人工标注好的的微博数据集中抽取 80% 的数据训练生成三个预测函数 Y_B, Y_T, Y_S 分别代表贝叶斯分类器预测函数、决策树分类器预测函数和支持向量机分类器预测函数, Y_U 是联合分类器预测函数. 对于每条微博 w, 使用三个预测函数分别做预测分类. 微博分为两类: 与用户兴趣相关 $c1$ 和与用户兴趣不相关 $c2$. 那么:

$$Y_U(c1|w)=Y_B(c1|w)+Y_T(c1|w)+Y_S(c1|w)$$
$$Y_U(c2|w)=Y_B(c2|w)+Y_T(c2|w)+Y_S(c2|w)$$

Y_B, Y_T, Y_S 取值为 0 或 1. 如果$Y_U(c1|w)>Y_U(c2|w)$, 则将微博 w 分为 $c1$ 类, 否则分为 $c2$ 类.

这样做的原理是:

设 P={$P1$, $P2$, $P3$}分别是三个分类器的准确率, 那么联合分类器的准确率:

$$P_U=C_3^2P^2(1-P)+P^3=3P^2-2P^3$$

对 P_U 求导数$P_U'=6P-6P^2$, 令其等于 0 得 P=0 或 1, 所以可知 P_U 在 $P\in(0,1)$ 的范围内是递增的. 再令 P_U=P 得 P=0、0.5 或 1, 即在 P=0、0.5 和 1 的时候联合分类器的效果和单个分类器的效果一样, 而又因为 P_U 在 $P\in(0,1)$ 的范围内是递增的, 所以在 1>P>0.5 的情况下 P_U>P, 所以要求朴素贝叶斯分类、决策树分类和支持向量机 (SVM) 分类器的准确率达到 0.5 以上, 联合分类器才有效果. 而实验表明本文采用的三个分类器准确率都在 0.5 以上, 所以采用联合分类器进行微博噪声分类是合理而有效的.

1.5 LDA 的主要原理

LDA 模型是一种三层贝叶斯模型, 三层分别为: 单词层、topic 层和文档层. 该模型基于如下假设:

(1) 整个文档集合中存在 k 个互相独立的 topic;

(2) 每一个 topic 是词上的多项分布;

(3) 每一个文档由 k 个 topic 随机混合组成;

(4) 每一个文档是 k 个 topic 上的多项分布;

(5) 每一个文档的 topic 概率分布的先验分布是 Dirichlet 分布;

(6) 每一个 topic 中词的概率分布的先验分布是 Dirichlet 分布.

文档的生成过程如下:

(1) 对于文档集合 M, 从参数为 β 的 Dirichlet 分布中采样 topic 生成 word 的分布参数 φ;

(2) 对于每个 M 中的文档 m, 从参数为 α 的 Dirichlet

分布中采样 doc 对 topic 的分布参数 θ;

(3) 对于文档 m 中的第 n 个词语 W_mn, 先按照 θ 分布采样文档 m 的一个隐含的主题 Z_m, 再按照 φ 分布采样主题 Z_m 的一个词语 W_mn.

LDA 贝叶斯网络结构如图 1 所示.

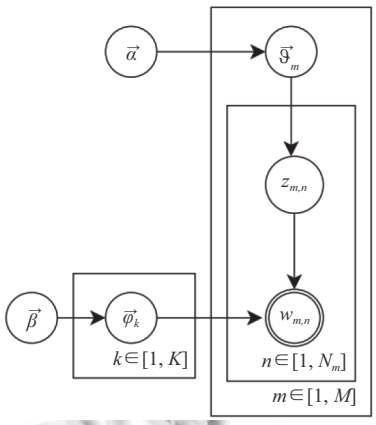

图 1 LDA 贝叶斯网络结构图

因此整个模型的联合分布, 如下:

$$P(Z,W,\theta,\phi|\alpha,\beta)=P(W|\phi,Z)P(Z|\theta)P(\theta)P(\phi)$$

对联合分布求积分, 去掉部分隐变量后:

$$\begin{aligned}P(Z,W|\alpha,\beta)&=\iint P(Z,W,\theta,\phi|\alpha,\beta)d\theta d\phi\\&=\int P(W|\phi,Z)P(\phi)d\phi\int P(Z|\theta)P(\theta)d\theta\\&=\prod_{k=1}^{K}\frac{\Delta(n_k+\beta)}{\Delta(\beta)}\prod_{m=1}^{M}\frac{\Delta(n_m+\alpha)}{\Delta(\alpha)}\end{aligned}$$

通过计算消除中间参数 θ 和 φ, 得到主题的转移概率化为:

$$\begin{aligned}&P(Z_i=k|\overrightarrow{Z_{\neg i}},\overrightarrow{W})\\&\propto\frac{n_{m,\neg i}^{(k)}+\alpha_k}{\sum_{k=1}^{K}(n_{m,\neg i}^{(t)}+\alpha_k)}\frac{n_{k,\neg i}^{(t)}+\beta_t}{\sum_{t=1}^{V}(n_{k,\neg i}^{(t)}+\beta_t)}\end{aligned}$$

2 微博主题发现模型

该模型的主要分为三个阶段: 一是将文本向量化, 将一个字符串转化成向量形式; 二是构造分类器, 本文共采用三个分类器: 朴素贝叶斯、支持向量机和决策树分别对训练集进行训练, 得出三个预测函数进行投票得出一个联合分类器模型; 三是训练 LDA 模型.

2.1 文本向量化

由于微博内容形式比较随意, 而且内容里面穿插着各种表情和 URL 等, 这使文本向量化造成很大的麻烦, 因此文本向量化首先要做的就是去除特殊字符, 包括繁体字、URL、标点符号等. 去除特殊字符之后我们要对微博内容进行分词.

文本向量化的主要流程是:

(1) 去除微博中出现的包括繁体字、URL、标点符号等在内的特殊字符.

(2) 将各文档进行分词, 从字符串转化成单词列表.

(3) 去除停用词, 去掉一些无关主题的词, 以免影响主题发现的质量.

(4) 统计各文档单词, 生成词典.

(5) 利用词典将文档转化成词频表示的向量, 即指向量中的各值对应于词典中对应位置单词在该文档中出现次数.

(6) 再进行进一步处理, 将词频表示的向量转化成 tf-idf 表示的向量.

(7) 由 tf-idf 表示的向量转化成 lsi 表示的向量.

2.2 分类器模型

过滤噪声微博的过程可以看作是一个二分类问题, 即将所有微博分为两个类: 与用户兴趣相关和与用户兴趣不相关. 本文使用了三种经典的文本分类方法和 bagging 学习算法提高分类器的准确率. 先用朴素贝叶斯分类、决策树分类和支持向量机 (SVM) 分类生成三个预测函数. 然后使用简单投票机制, 实行多数服从少数策略组合成联合分类器, 解决噪声微博过滤的问题.

我们将文本向量化后的每一列当作是分类器的特征, 然后直接用文档向量化的结果训练分类器, 得出三个预测函数, 然后将每个预测函数对微博进行投票, 获得两次及以上投票的类别我们就认为是该微博的类别.

2.3 LDA 模型

将得到的联合分类器对未分类的微博进行分类, 筛选出与用户兴趣相关的微博, 然后进行文档向量化, 在进行 LDA 训练, 得出 LDA 模型.

3 实验与结果分析

3.1 实验数据获取及预处理

实验数据是来自 2016 年 5 月 1 日的微博内容, 总共有将近 80 万条微博, 选出前 3000 条微博进行人工标注是否与用户兴趣相关. 结果统计得出与用户兴趣不相关的微博有 1243 条, 与用户兴趣相关的微博的有 1757 条. 然后我们分别对这 3000 条微博做了预处理,

即去除微博中的特殊字符, 然后将微博用结巴分词进行分词, 再将分词后的结果去停用词. 这样我们就把原先的一条微博转换成一个单词列表. 最后随机将标注好的数据集按 8: 2 的比例随机分成训练集和测试集, 这样就完成了数据的预处理过程.

3.2 训练联合分类器模型

在预处理过程中我们获得了训练集数据和测试集数据, 将训练集数据作为数据集来分别训练 Bayes 分类器模型、SVM 分类器模型和 Tree 分类器模型得到三个预测函数, 模型参数都是默认的, 通过投票对微博数据进行分类. 在机器学习中为了评估一个模型的性能, 通常使用 ROC 曲线, ROC 曲线经常用于评价一个二分类器的性能.

图 2 解释了 ROC 曲线中各参数的意义.

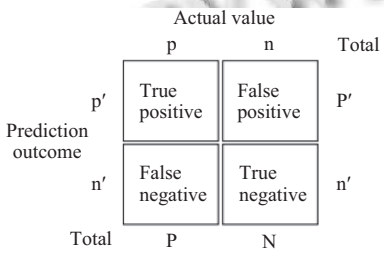

图 2 ROC 曲线参数示意图

ROC 关注两个指标:

TPR = TP/(TP + FN), TPR 代表将正例分为正例的概率;

FPR = FP/(FP + TN), FPR 代表将负例错分为正例的概率.

在 ROC 二维空间图中, FPR 作横坐标, TPR 是纵坐标, 画出的曲线说明了分类器在 P 和 FP 间的权衡. ROC 的主要分析工具是一个画在 ROC 空间的曲线. 对于二值分类问题, 实例的值往往是连续值, 我们通过设定一个阈值, 将实例分类到正类或者负类[11]. 因此我们可以变化阈值, 根据不同的阈值进行分类, 根据分类结果计算得到 ROC 空间中相应的点, 连接这些点就形成 ROC curve. 图 3 是 SVM 分类器和联合分类器的 ROC 曲线.

由图 1 可以看出联合分类器明显比 SVM 分类器分类效果要好上很多, 为了得到联合分类器模型和其他模型更直观的对比, 又算出各个分类器的准确率和召回率如表 1.

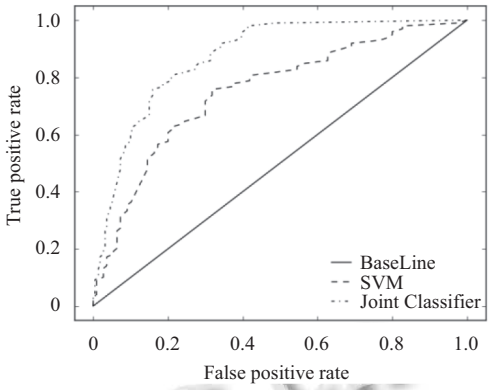

图 3 SVM 分类器和联合分类器在测试集上的 ROC 曲线

表 1 分类器的准确率和召回率

	准确率	召回率
Bayes分类器	0.70	0.83
SVM分类器	0.79	0.78
Tree分类器	0.76	0.64
BST联合分类器	0.87	0.88

由表 1 可知, 联合分类器的分类准确率达到 87%, 几乎可以将噪声微博去除. 我们用联合分类器将 8 万微博重新分类, 获得 53612 条与用户兴趣相关的微博.

3.3 训练 LDA 模型

在训练 LDA 模型中, 主题数的确定是一个非常困难的问题. 目前还没有非常有效的方法来确定主题的数目. 这里我们采用的是根据新浪新闻的分类数目来确定主题数目, 最后把主题数定为是 15 个.

本文先用联合分类器将 8 万微博重新分类, 获得 53612 条与用户兴趣相关的微博, 将这些微博进行去特殊字符分词等预处理, 最后获得是每篇微博都是词的列表. 将得到的数据作为语料库, 统计各文档单词生成词典, 利用词典将文档转化成词频表示的向量, 即指向量中的各值对应于词典中对应位置单词在该文档中出现次数, 再进行进一步处理, 将词频表示的向量转化成 tf-idf 表示的向量, 最后训练 LDA 模型得到 15 个主题. 主题和主题下的词如表 2 所示.

表 2 主题及主题词

主题	主题词				
Topic 1	五一劳动节	火车	劳动者	历史	公园
Topic 2	北京遇上西雅图	不二情书	孤独爱情	电影	客户端
Topic 3	五一	音乐节	武汉	快乐	生活
Topic 4	中国	美国	市场	国家	老板
Topic 5	吃	晚上	健康	睡眠	失眠

3.4 实验结果分析

Bagging 是一种用来提高学习算法准确度的方法, 这种方法通过构造一个预测函数系列, 然后以投票的方式将它们组合成一个预测函数. 通过这种方式把三个原本用于微博分类效果不好的模型组合成一个联合分类器. 通过图 2 和表 1 可以看出联合分类器的分类效果准确率显著提高. 表 2 是主题从 15 个主题中选择的 5 个主题以及每个主题下的前 5 个词. 根据实验结果可以看出去除噪声微博之后每个主题非常明确, 这也进一步说明了去除噪声微博的重要性.

4 结束语

本文先将用户发的微博进行了预处理, 然后进行人工标注. 把标注好的数据作为分类器模型的训练集, 实验表明各个模型的精确率都可达到 70% 以上, 而本文所使用的联合分类器的精确率可以达到 85% 以上, 接下来的 LDA 模型发现的主题效果很好, 更加证明了在进行主题发现之前使用联合分类器对微博内容进行分类, 去除噪声微博的重要性和有效性.

虽然该模型取得了很好的效果, 但是没有对 LDA 模型进行改进. LDA 模型分类的效果不仅取决于训练集的质量, 还有其他的很多方面, 例如分词的效果, 参数的选择等. 下一步将对该问题做更深入的研究.

参考文献

1 王广新. 基于微博的用户兴趣分析与个性化信息推荐[硕士学位论文]. 上海: 上海交通大学, 2013.

2 于洪涛, 崔瑞飞, 董芹芹. 基于遗忘曲线的微博用户兴趣模型. 计算机工程与设计, 2014, 35(10): 3367–3372, 3379. [doi: 10.3969/j.issn.1000-7024.2014.10.006]

3 张培晶, 宋蕾. 基于 LDA 的微博文本主题建模方法研究述评. 图书情报工作, 2012, 56(24): 120–125.

4 刘红兵, 李文坤, 张仰森. 基于 LDA 模型和多层聚类的微博话题检测. 计算机技术与发展, 2016, 26(6): 25–30, 36.

5 柳培林. 基于向量空间模型的中文文本分类技术研究[硕士学位论文]. 大庆: 大庆石油学院, 2006.

6 周茜, 赵明生, 扈旻. 中文文本分类中的特征选择研究. 中文信息学报, 2004, 18(3): 17–23.

7 刘丽珍, 宋瀚涛. 文本分类中的特征选取. 计算机工程, 2004, 30(4): 14–15, 175.

8 刘颖. 用隐马尔柯夫模型对汉语进行切分和标注排歧. 计算机工程与设计, 2001, 22(4): 58–62, 68.

9 李湘东, 高凡, 丁丛. LDA 模型下不同分词方法对文本分类性能的影响研究. 计算机应用研究, 2017, 34(1): 62–66.

10 张红梅, 王利华. 使用否定选择算法改进文本过滤. 计算机工程与科学, 2008, 30(8): 61–64.

11 刘海峰, 刘守生, 姚泽清. 文本分类中基于训练样本空间分布的 K 近邻改进算法. 情报学报, 2003, 32(1): 80–85.

12 李湘东, 巴志超, 黄莉. 基于语料信息度量的文本分类性能影响研究. 情报杂志, 2014, 33(9): 157–162, 180.

13 苑擎飚. 基于决策树中文文本分类技术的研究与实现[硕士学位论文]. 沈阳: 东北大学, 2008.

14 崔建明, 刘建明, 廖周宇. 基于 SVM 算法的文本分类技术研究. 计算机仿真, 2013, 30(2): 299–302.