

基于读者视角的文本情感分类

卢 露¹, 魏登月²

(1 上海电力学院 计算机科学与技术学院, 上海 200090;

2 武汉大学 计算机科学与技术学院, 湖北 武汉 430079)

摘 要: 从读者的角度对文本情感进行分类. 训练样本集以新闻文章作为样本实例, 以文章后读者的投票信息作为样本类别标注的先验知识. 针对该不完备的数据集提出了一种半监督学习的分类模型, 分类方法采用朴素贝叶斯分类法和 EM 算法相结合. 实验证明该方法不仅简单有效, 而且具有较高的分类性能.

关键词: 情感分类; 情感标签; 期望最大化算法; 朴素贝叶斯; 后验概率

中图分类号: TN402

文献标识码: A

文章编号: 1000-7180(2014)10-0122-04

Text Emotion Classification from Reader Perspective

LU Lu¹, WEI Deng-yue²

(1 School of Computer, Shanghai University of Electric Power, Shanghai 200009, China;

2 School of Computer, Wuhan University, Wuhan 430079, China)

Abstract: The past research of text emotion classification mainly face to the author's emotion. But this article classify texts' emotion from the reader's perspective. Using news articles as training sample sets, and using readers' vote information behind news articles as prior knowledge of the sample category. we put forward a semi-supervised classification model according to the incomplete data sets, the classification method using naive bayesian method and in couple with EM algorithm. Experiments show that our method is not only simple and effective, and has higher classification performance.

Key words: Emotion classification; Emotion label; Expectation Maximization algorithm; Naive Bayes classifier; Maximum posteriori estimate

1 引言

文本情感分类的研究分为两类, 一类是分析该文本本身包含的情感或者称为作者的情感 (Write-emotion), 另一类是研究读者阅读文章后触发的情感或者称为读者的情感 (Reader-emotion). 通常文章中包含的作者情感和读者阅读该文章后引发的情感是不一致的^[1]. 比如某政客在自己的博客中发表文章来描述自己的悲惨遭遇, 但这并不一定能引起读者的共鸣, 甚至可能会让某些人感到高兴. 分析文章所能触发的读者情感是重要的, 首先有利于政府部门了解事件引起的网络舆情; 其次在个性化推荐

系统中, 能够根据用户当时的心情推送文章, 为用户提供便捷的浏览体验^[2]; 此外还便于一些新闻站点提供心情中心, 将新闻按读者情感类别归类, 满足用户分心情浏览新闻的需求. 由于一篇文章能够触发读者多种情感, 目前的研究将文章触发的最主要的情感作为该文章的读者情感类, 将其看作一个单分类问题. 文献[3-5]中利用机器学习的方法来建立分类器, 由于缺少必要的语料信息, 必须对训练集中文章的读者情感类别进行手工标注, 耗时费力. 近年来, 随着 Web 2.0 技术的迅猛发展, 带有用户交互特性的网络新闻网站日渐流行起来. 以中新网、新浪网为代表的网站, 允许用户在阅读新闻后投票表达

收稿日期: 2013-12-20; 修回日期: 2014-02-13

基金项目: 国家自然科学基金项目(61272277)

自己的心情,网站通过对投票结果进行统计,进而了解该则新闻如何影响读者的心情.如图1所示为某新闻文章后读者的投票统计结果.读者情感类标签包含感动、同情、无聊、愤怒、搞笑、难过、新奇、温馨等八种情感.

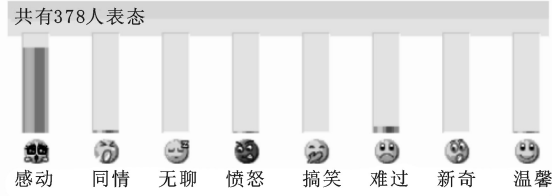


图1 文档后读者的情感标签投票统计

这些用户在一定时间内的投票数据可以为样本集文章类别提供先验知识,即可以认为文章触发的主要情感类别一定是在那些投票数较多的情感标签中产生.这样可以构造概率生成模型建立分类器,模型采用的分类方法为朴素贝叶斯方法,并利用EM算法^[6]来估计模型的参数.实验证明该模型能够在不需要手工标注训练集样本情况下,具有较好的分类效果.

2 半监督情感分类模型

在传统的文本分类中,训练集里每个样本类别需要事先标注,由此建立的分类方法是一种有指导的学习方法.而在本文使用的语料库中,每篇文章的读者情感类别是未知的,只是收集了在一段时间内若干用户阅读完该文章后在网站设定的情感标签上的投票结果作为先验知识.所以本文的方法是一种半指导的分类方法,下面首先给出模型的相关假设:

- 本模型也为单分类分类器,每篇文章只划分到一个读者情感类.
- 选定用户投票数较多的情感标签作为文章的候选读者情感类标签集,集合中的情感类标签可能多于一个.
- 文章的读者情感类一定出自其对应的候选读者情感类标签集.

由假设可知训练集中每个样本实例的类别是未知的,也就是说该训练数据集是不完备的.在处理不完备数据集时,EM算法被证明是有效的方法,本文将它与朴素贝叶斯分类法相结合.文献^[7]已经证明在半监督学习中,这两种方法的结合在文本分类中能够提高分类准确率.不同的是,文献^[7]的训练集中已经具有少量标注的样本,而本文的训练集中样

本均是未标注的.具体分类过程如下:

2.1 模型描述

假设 E 为语料库中包含的 m 个文档集合, $E = \{E_1, E_2, \dots, E_m\}$. E_i 为第 i 个文档. $S = \{s_1, s_2, \dots, s_n\}$ 为大小为 n 的读者情感类标签集. S_i 为文档 E_i 对应的候选读者情感类标签集, $S_i \subseteq S$.

目前能观测到的数据为文档集合 E . 假设该分类模型需要的参数集合为 θ , 该问题转化为 θ 取何值时产生文档集合 E 的概率最大. 在观测数据下模型参数的后验概率为 $P(\theta | E)$. 根据贝叶斯公式:

$$\begin{aligned} \log P(\theta | E) &= \log \frac{p(\theta) | p(E | \theta)}{p(E)} \\ &\propto \log p(\theta) p(E | \theta) \\ &= \log p(\theta) + \log p(E | \theta) \end{aligned} \quad (1)$$

我们用极大后验估计的方法来确定 θ , 即求出 $\log p(\theta) + \log p(E | \theta)$ 极大值对应的 θ . 首先对于 $\log p(E | \theta)$ 可作如下分解:

$$\begin{aligned} \log P(E | \theta) &= \log \prod_{E_i, S_i} P(S_i) P(E_i | S_i) \\ &= \sum_{E_i, S_i} \log P(S_i) P(E_i | S_i) \\ &= \sum_{E_i, S_i} (\log P(S_i) + \log P(E_i | S_i)) \end{aligned} \quad (2)$$

对于文档 E_i , 如果它的读者情感类标签为 $s_k (s_k \in S_i)$, 那么 $P(E_i | S_i)$ 可分解为: $P(E_i | S_i) = P(s_k | S_i) P(E_i | s_k)$, 因为 s_k 是未知的, 可以将其作为隐含的变量. 这里我们利用 EM 算法求出关于隐含变量的期望最大化, 如式(3)所示.

$$\begin{aligned} \log P(\theta) + \sum_{E_i, S_i} (\log P(S_i) + \sum_{s_k} P(s_k | E_i, S_i) \log P(s_k | S_i) P(E_i | s_k)) \end{aligned} \quad (3)$$

因为 $\log P(S_i)$ 为常数, 可重新定义为

$$\log P(\theta) + \sum_{E_i, S_i} \sum_{s_k} P(s_k | E_i, S_i) \log P(s_k | S_i) P(E_i | s_k) \quad (4)$$

假设 $P(E_i | s_k)$ 满足多项式分布^[12], 可以进一步分解为

$$P(E_i | s_k) = P(\omega | E_i) \frac{|E_i|!}{\prod_{\omega} n_{ij}(\omega)!} \prod_{\omega} P(\omega | s_k) n_{ij}(\omega) \quad (5)$$

$n_{ij}(\omega)$ 为词语 ω 在文档 E_i 出现的次数.

对于 $P(\theta)$, 我们使用 Dirichlet 分布作为参数 θ 的先验概率, α 为 Dirichlet 分布的超参数.

$$P(\theta) \propto \prod_{s_k} (P(s_k)^{\alpha-1} \prod_{\omega} P(\omega | s_k)^{\alpha-1}) \quad (6)$$

因此对数似然函数 Q 可写为

$$Q(\theta) = \sum_{s_k} ((\alpha - 1) \log P(s_k) + \sum_{\omega} (\alpha - 1) P(\omega | s_k)) + \sum_{E_i} \sum_{s_k} P(s_k | E_i, S_i) \log P(s_k | S_i) + \sum_{E_i} \sum_{s_k} \sum_{\omega} n_{ij}(\omega) P(s_k | E_i, S_i) \log P(\omega | s_k) + (\text{constant}) \quad (7)$$

2.2 参数估计

EM 算法参数估值分为 E 步骤(expectation) 和 M 步骤(maximization), E 步骤中以参数 θ^k 为条件估计隐藏变量的分布, 在 M 步骤中选取一个新的参数 θ^{k+1} 来使观察到的数据期望最大化. 在每一步过程中似然值只会增大, 可以保证该算法会收敛到局部最大值.

E 步骤中的重要一步是计算隐含变量的后验概率, 这里隐含变量的后验概率 $P(s_k | E_i, S_i)$ 可以通过全概率公式求出:

$$\begin{aligned} P(s_k | E_i, S_i) &= \frac{P(s_k, E_i, S_i)}{\sum_{s_a \in S_i} P(s_a, E_i, S_i)} \\ &= \frac{P(S_i) P(s_k | S_i) P(E_i | s_k, S_i)}{\sum_{s_a \in S_i} P(S_i) P(s_a | S_i) P(E_i | s_a, S_i)} \quad (8) \\ &= \frac{P(s_k | S_i) P(E_i | s_k, S_i)}{\sum_{s_a \in S_i} P(s_a | S_i) P(E_i | s_a, S_i)} \end{aligned}$$

当 s_a 被指定后, E_i 和 S_i 是相互独立的, 这时 $P(E_i | s_a, S_i) = P(E_i | s_a)$. 对于 $P(s_k | S_i)$, $P(s_k | S_i) = \frac{\delta(s_k \in S_i) P(s_k)}{\sum_{s_a} \delta(s_a \in S_i) P(s_a)}$ (9)

δ 函数定义如下:

$$\delta(s_a \in S_i) = \begin{cases} 1 (s_a \in S_i) \\ 0 (s_a \notin S_i) \end{cases} \quad (10)$$

式(9)可进一步写为

$$\begin{aligned} P(s_k | E_i, S_i) &= \frac{\delta(s_k \in S_i) P(s_k) P(E_i | s_k)}{\sum_{s_a \in S_i} \delta(s_a \in S_i) P(s_a) P(E_i | s_a)} \\ &= \frac{\delta(s_k \in S_i) P(s_k) \prod_{\omega} P(\omega | s_k)^{n_{ij}(\omega)}}{\sum_{s_a} \delta(s_a \in S_i) P(s_a) \prod_{\omega} P(\omega | s_a)^{n_{ij}(\omega)}} \quad (11) \end{aligned}$$

在 M 步骤要最大化 $Q(\theta)$, 且该函数的约束项为: $\sum_{\omega} P(\omega | s_k) = 1, \sum_{s_k} P(s_k) = 1$. 构造得拉格朗

日乘子:

$$L(\theta) = Q(\theta) + \sum_{s_k} \beta_k (\sum_{\omega} P(\omega | s_k) - 1) + \gamma (\sum_{s_k} P(s_k) - 1) \quad (12)$$

对每个活动参数求导并使之等于 0, 解方程易得

$$P(\omega | s_k) = \frac{\sum_{E_i} n_{ij}^{(\omega)} P(s_k | E_i, S_i) + (\alpha - 1)}{\sum_{E_i} \sum_{\omega} n_{ij}^{(\omega)} P(s_k | E_i, S_i) + (\alpha - 1)m} \quad (13)$$

$$P(s_k) = \frac{\sum_{E_i} P(s_k | E_i, S_i) + (\alpha - 1)}{\sum_{E_i} \sum_{s_k} P(s_k | E_i, S_i) + (\alpha - 1)n} \quad (14)$$

m, n 分别为文档数和读者情感类别数.

EM 算法运行的初始值设定:

$$P(s_a | E_i, S_i) = \begin{cases} \frac{1}{|S_i|} (s_a \in S_i) \\ 0 (s_a \notin S_i) \end{cases} \quad (15)$$

通过 EM 算法训练出参数后, 可以通过如下公式计算出文档 E_i 的读者情感类别:

$$\begin{aligned} \arg \max_a P(s_a | E_i) &\propto \\ \arg \max_a P(s_a) \prod_{\omega} P(\omega | s_a)^{n_{ij}(\omega)} \quad (16) \end{aligned}$$

3 实验

3.1 数据集和评估方法

我们选择新浪网上的新闻文章作为语料库. 网站为新闻文章提供了八种读者类情感标签(感动、同情、愤怒、搞笑、难过、新奇、温馨、无聊)供投票. 我们下载了 5 600 篇新闻文章以及文章后用户在八种情感类标签上的投票统计数. 为了确保用户的投票过程已经达到稳定状态, 选择发表日期在 2009 年 1 月到 2010 年 3 月之间的文章. 此外在候选读者情感类标签集的选择上作如下处理, 只将大于平均投票数的情感标签才纳入该文章对应的候选读者情感类标签集. 最终有 1 210 篇文章的候选情感标签为 1 个, 其余 4 390 篇文章的候选情感标签为多个.

分类模型的评估指标采用信息检索领域常用的查全率、查准率以及 F 值来作为判断标准, 因为本模型是一个多分类模型, 故采用宏平均 F 值作为最后的分类性能指标(即分别计算各类对应的 F 值, 然后求其平均值). 实验选择 80% 的新闻数据作为训练集, 余下 20% 的数据作为测试集.

3.2 实验结果

EM 算法是一个迭代算法, 首先观察算法是否

能够收敛以及迭代次数对分类准确率的影响.如图2所示,在刚开始运行时,分类准确率是逐渐上升的,在迭代次数为5次左右时,分类准确率达到最大值,迭代次数在15次以后,分类准确率逐渐趋于平稳.说明该算法能够在有限次迭代后收敛,是有效的.而且如果算法能设定合适的迭代次数,在程序运行的过程中强行终止迭代过程,还能获得较好的分类效果.

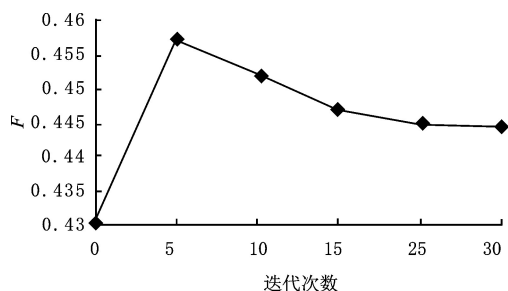


图2 不同迭代次数下的分类准确率宏平均 F 值

其次要验证分类方法的有效性,因为本方法的核心是训练集样本的情感类别标注,事先每个实例的类标签都是未知的,将用户的投票信息作为样本实例类别的先验知识,在构造分类模型时利用 EM 算法推导得出.因此在构造分类器时,我们选择与如下两种训练样本的标注方法作为对比:

(1) 随机从候选情感标签集合中选择一个情感标签作为文章的读者情感标签.(NB+random)

(2) 将文章的候选情感标签等概率分配给该文章.(NB+init)

为方便比较,上述两种方法都同样采用朴素贝叶斯分类法,本方法的迭代次数设为5.如图3、4所示,与其他两种方法相比,本文提出的 NB+EM 算法在八种读者情感类别上的分类准确率较其他两种方法平均能提高5%到10%,说明本方法是有效的.此外还可以看出在愤怒、感动、搞笑、难过这几种情感类的分类准确率较高,主要是因为这几类文章情感比较突出,而且在训练样本集中,这几种类别的文章所对应的候选读者情感类标签集基本只包含唯一的情感标签,更容易学习和区分.

最后对本方法的假设条件进行了分析.在构造本分类模型时,我们假设文章的读者情感类标签一定出自于文章的候选读者情感类标签集,而实际上并不一定如此.因此我们对式(10)进行了修改,将 $\delta(s_a \notin S_i)$ 的值分别设置为不同的值,来观察最后的分类结果,如表1所示.可以看出随着 $\delta(s_a \notin S_i)$ 的取值变大, F 值是不断变小的,当 $\delta(s_a \notin S_i)$ 为0时,

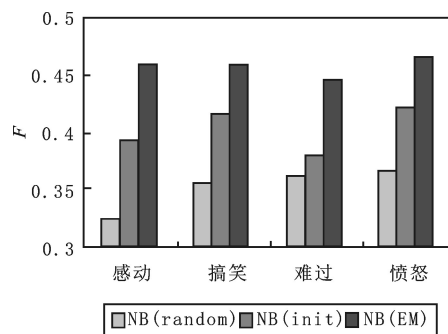


图3 感动、搞笑、难过、愤怒四种情感的分类准确率 F 值

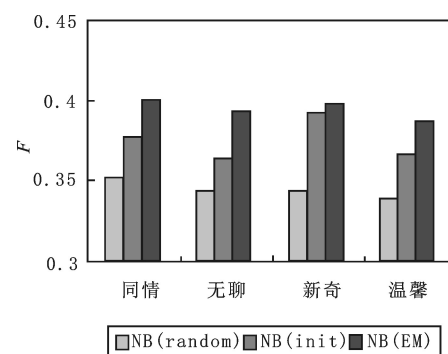


图4 同情、无聊、新奇、温馨四种情感的分类准确率 F 值
可以得到最好的分类结果,也就是说文章的读者情感标签基本上会出自于文章的候选读者情感类标签集,本文的设想是完全合理的.

表1 在不同的 δ 值下对应的宏平均 F 值

$\delta(s_a \notin S_i)$	F-score
0	0.452
0.1	0.349
0.3	0.314
0.5	0.283
0.7	0.262
0.9	0.259
1	0.245

4 结束语

本文主要研究了基于读者情感类的文本分类方法.在建立分类模型时,训练样本集是一种不完备的数据集,每个样本实例的类别是未知的,只是借助新闻网站上提供的用户投票机制作为样本情感类别标注的先验知识.在构建分类器时,以朴素贝叶斯分类法为基础,并结合 EM 算法估计分类模型参数,实验证明能够获得较好的分类效果.在以后的工作中,我们进一步研究该方法的稳定性和可扩展性.

(下转第129页)

5 结束语

本文提出了一种基于整数小波变换和人眼视觉系统的可逆数字水印算法. 具有以下特点: (1) 算法中有效的结合人眼视觉系统, 选择出合适的块进行水印的嵌入, 保证水印嵌入之后的图像具有较好的透明性; (2) 利用 LSB 在两次 IWT 得到的所有中低频子带的系数中嵌入水印, 提高了嵌入量; (3) 整数小波变换具有的计算速度快, 不需要量化等优点, 使得算法复杂度低; (4) 水印的提取以及原始图像的恢复是嵌入的逆过程, 操作简单易行. 本文提出的算法是脆弱水印, 为了扩充应用性, 需要进一步研究鲁棒的可逆数字水印技术.

参考文献:

- [1] 赵学民, 李玲玲. 一种关系数据库的多媒体多重水印[J]. 微电子学与计算机, 2013, 30(9): 122-125.
- [2] 聂廷远, 刘海涛, 周立俭. 利用 LUT 替换的 FPGA 保护数字水印技术[J]. 微电子学与计算机, 2013, 30(1): 21-23.

(上接第 125 页)

参考文献:

- [1] Yang C H, Lin H Y, Chen H H. Writer meets reader: emotion analysis of social media from both the writer's and reader's perspectives[C]// Proceedings of International Conference on Web Intelligence. Milan, Italy: IEEE, 2009: 287-290.
- [2] Cai R, Zhang Z, Wang C, et al. Music sense: contextual music recommendation using emotional allocation modeling [C]// Proc of the ACM International Multimedia Conference and Exhibition Augshorg. Bavaia, Germany: IEEE, 2007: 553 - 556.
- [3] Yang C, Lin K, Chen H-H. Building emotion lexicon from weblog corpora[C]// Proc of 45th Annual Meeting of Association for Computational Linguistics. Prague Czech: IEEE, 2007: 133-136.
- [4] Kevin Hsin-Yih Lin, Changhua Yang. Emotion classification of online news articles from the reader's perspective[C]// proc of the 2008 IEEE/WIC/ACM International Conference on the Web Intelligence and In-

- [3] Tian J. Reversible data embedding using a difference expansion [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2003, 13(8): 890-896.
- [4] Al-Qershi O M, Ee Khoo B. Two-dimensional difference expansion (2D-DE) scheme with a characteristics-based threshold [J]. Signal Processing, 2012, 93(1): 154-162.
- [5] Gu Q L, Gao T G. A novel reversible robust watermarking algorithm based on chaotic system [J]. Digital Signal Processing, 2013, 23 (1): 213-217.
- [6] Jung S W, Ha L T, Ko S J. A new histogram modification based reversible data hiding algorithm considering the human visual system[J]. IEEE Signal Processing Letters, 2011, 18 (2): 95-98.

作者简介:

邱群 女, (1987-), 硕士. 研究方向为图像处理、数字水印.

马小虎 男, (1964-), 博士, 教授, 硕士生导师. 研究方向为数字水印、模式识别、图像处理、计算机图形学.

telligent Agent Technology. Washington: IEEE, 2008: 220-226.

- [5] H Y Lin, H H Chen. Ranking reader emotions using pairwise loss minimization and emotional distribution regression[C]// Proceedings of international Conference on Empirical Methods in Natural Language Processing. [s. l.]: IEEE, 2008: 136-144.
- [6] Dempster A P, Laird N M, Rubin D B. Maximum likelihood from incomplete data via the EM algorithm[J]. Journal of the Royal Statistical Society Series B, 1977, 39(1): 1-38.
- [7] Nigam K, McCallum A, Thrun S, et al. Text classification from labeled and unlabeled documents using EM [J]. Machine Learning, 2000, 39(2): 103 - 134.

作者简介:

卢露 男, (1981-), 博士. 研究方向为数据挖掘、自然语言处理.

魏登月 男, (1984-), 博士研究生. 研究方向为数据挖掘.