基于补全矩阵的多标签相关性情感分类①

许莉莉, 高俊波, 李胜宇

(上海海事大学 信息工程学院, 上海 201306)

摘 要: 目前, 对新闻情感分类问题的研究大部分是从新闻作者的角度进行的, 而对读者反馈的真实情感分析的较少. 本文从读者角度入手进行情感分析研究. 提出一种基于补全矩阵的多标签相关性情感分类模型, 采用 LDA 提取主题表示新闻文本, 然后通过使用标签相关性矩阵对原始的标签矩阵进行补全, 构造了一个增强的补全标签矩阵模型(CM-LDA). 最后通过和原始矩阵的 LDA 模型进行比较, 该模型使最终的多标签分类性能有了明显的提高, 准确率达到了 85.72%.

关键词: 社会新闻; 情感分析; 标签相关性; 补全矩阵-LDA(CM-LDA); 多标签分类

Emotion Classification of Multi-Label Correlation Based on Completion Matrix

XU Li-Li, GAO Jun-Bo, LI Sheng-Yu

(College of Information Engineering, Shanghai Maritime University, Shanghai 201306, China)

Abstract: At present, most of the researches of sentiment classification are carried out from the writer's perspective with quite few analyses from readers. This paper is to study the sentiment analysis from the news readers. A model of multi-label correlation sentiment classification based on completion matrix and LDA is proposed to extract the topic. The original news text is represented with the generated text-subject features, which are taken as the input to a subsequent classifier. Furthermore, the paper constructs a model of enhanced completion label matrix (CM-LDA) by appending the label correlation matrix to the original label matrix. Results show that the accuracy of this approach achieves 85.72% in the multi-label classification task, which outperforms the traditional LDA methods significantly.

Key words: social news; sentiment analysis; label correlation; completion matrix-LDA(CM-LDA); multi-label classification

1 引言

目前, 从作者角度分析, 有关新闻文本的情感分类研究已见成熟. 而从读者角度分析的研究相对较少, 但也逐渐取得了不错的成果. 台湾的 Lin, Chen 等[1]对于这个领域做了较详细的研究. 他们利用 Yahoo!Kimo News 的数据, 研究读者情绪. 研究中选取五种特征: 中文字符二元组, 中文词, 新闻元数据, 词缀相似度和情绪词. 实验结果表明利用 SVM 分类器与词缀相似度和情感词这两种特征结合的方法得到了最好的准确率. Blei 和 McAuliffe[2]引入一个 response 变量因子, 将文档的标签信息加入到构造的统计模型 Supervised LDA 中. 卢露等[3]从读者的角度对新闻文本情感进行分类, 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, 提出了一种半监督学习的分类模型, 实验证明采用 Bayes 方法和 EM 算法相结合具有较高的分类性能. 国内的包胜华、徐生良等[4]考虑了在 LDA 的基础上加上一层"情绪"参数, 将特征词和读者的情感关联起来.

另外, 已有学者研究, 试图将标签之间的相关性影响应用到多标签的分类算法中, 希望来提高分类的性能. 比如, 文献[5,6]利用了标签两两间的的相关性, 使系统的性能有了明显地提高. 然而在现实中, 一个标签可能会与其他标签都有关系. 文献[7]考虑了标签之间所有可能具有的相关性, 即对每个标签都考虑了其它标签可能对它的影响. 文献[8]考虑了标签集中的随机子集之间的相关性. 针对不同的领域提出了多种

① 收稿时间:2016-04-13;收到修改稿时间:2016-05-08 [doi:10.15888/j.cnki.csa.005496]

标签相关性方法的应用, 取得了不同程度的情感分类研究成果. 针对本文的研究领域, 如何将标签相关性有效地应用到多标签分类的问题中, 是当前的重要问题.

针对情感分类和多标签相关性分析的问题, 本文提出了在 LDA 模型能降维使数据处理耗时少的优势上, 利用学习到的标签相关性矩阵, 对多标签标准化处理后得到的原始标签矩阵进行补全, 构造了一个增强的补全标签矩阵模型(CM-LDA), 从而来对读者的情感进行分类, 以期达到更为准确的分类结果, 同时期望能对社会热点新闻的舆情预测分析提供帮助.

2 相关工作

实验采用的数据是从新浪社会热点新闻网站上抓取的, 收集时间为 2014.9-2016.3, 共 8583 篇, 具体信息包括新闻标题、发布时间、文本内容、读者投票数据, 其中读者情绪包括: 感动、震惊、搞笑、难过、新奇、愤怒等 6 种. 为了确保读者的投票过程达到稳定的状态, 在新闻发布一星期后进行采集. 然后进行数据净化处理, 用中科院的中文分词系统 NLPIR 进行分词, 并通过完善停用词表消除样本中的停用词噪声, 得到了 61,505 个不同的词. 本文是在 LDA 主题模型降低文本空间维度的基础上, 提出了一种新的多标签相关性情感分类模型, 具体工作流程如图 1 所示.

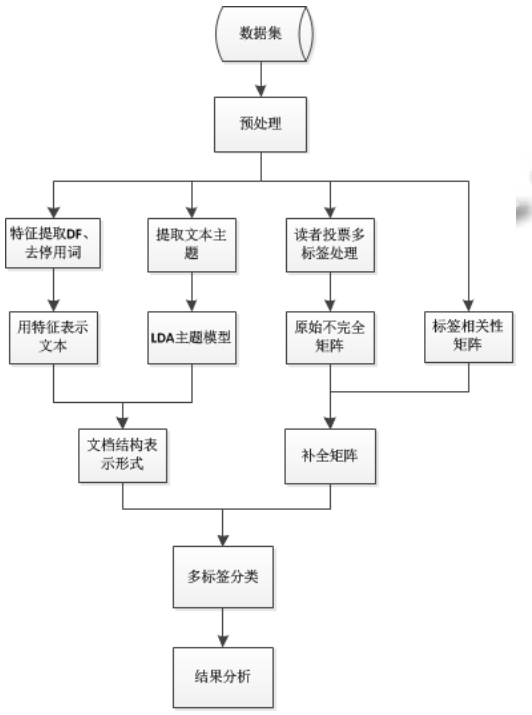

图 1　基于补全矩阵的情感分类模型

2.1 文本预处理

2.1.1 特征词提取

特征词选择的目的是在不损失其他性能的前提下, 有效地选择出较少的特征表示文本. 常用的特征选择方法有: x^2 统计量(CHI)、互信息(MI)、文档频率(DF)等. 本文选取的方法是 DF, 即样本中包含这个特征的所有文本数. DF 是提取出 DF 值达到一定次数的特征, 低于该值的认为影响力小. 然后对提取特征词后的文档按照停用词表去无用词、消除噪声.

2.1.2 LDA 主题模型

在文献[9]的研究中说明了读者情绪与文本主题具有一定的相关性. 于是, 本文就引入了 LDA 模型, 其是一种无监督的机器学习技术, 也是一种典型的词袋模型, 是通过采用 Gibbs Sampling 进行推导的. 而且 LDA 模型能有效的将高维的文本转化到较低的主题维度空间. LDA 主题模型是三层模型: 文本-主题-词.

通过对文本的预处理及 LDA 主题建模将非结构化的文档表示为结构化的数据, 处理之后的文档格式如图 2 所示.

$$D=\{W_1,W_2,...,W_M\}\overset{LDA主题建模}{\Rightarrow}\begin{bmatrix}T_{11} & T_{12} & \Lambda & T_{1P}\\ T_{21} & T_{22} & \Lambda & T_{2P}\\ \mathrm{M} & \mathrm{M} & \mathrm{O} & \mathrm{M}\\ T_{M1} & T_{M2} & \Lambda & T_{MP}\end{bmatrix}$$

图 2　文档集主题概率生成形式

其中, 矩阵 D 中 W_m 表示文档集中第 m 个文档, $w=\{w_1,w_2,\cdots,w_N\}$ 中 w_n 表示文档中的第 n 个单词. 文档--主题概率矩阵中的 M、P 分别表示样本数和主题数, 每行表示一篇文档, 每列表示一个主题, 每个值表示该主题在相应文档中所占的比重, 比重越大认为主题在这篇文本中的相对重要程度越大.

2.2 CM-LDA 建模步骤

2.2.1 多标签处理问题

在早期研究文本分类时, 遇到的歧义性问题[10]提到了多标签学习概念.

定义. 设一个样本的特征向量 $X=(x_1,x_2,...,x_n)$, 其中 $x_l\in R$; 候选标签集合 $L=(l_1,l_2,...,l_m)$, 文本所对应的标签集合 $Y=(l_1,l_2,...,l_p)$, $p\le m$, 即 $Y\subseteq L$; 则包含 t 个样本的多标记数据集合表示如下:

$$D=\{(X_i,Y_j)|1\le i\le t,X_l\in R^n,Y_l\subseteq L\} \quad (1)$$

式(1)中 X_i 表示一个样本向量, Y_l 表示该样本所对应的标签集. 多标签分类算法认为样本中的标签是独立的, 没有考虑它们之间的相关性. 本文提出, 首先利用多组实验统计得到稳定的相关性矩阵, 然后用相关性矩阵对原始不完全的标签矩阵进行补全, 最后采用分类算法RAkEL进行分类. 不需要像王霄等[11]对训练集进行扩展, 而是利用多标签标准化算法对语料的读者投票数据处理, 使每个样本的读者情感标签不超过三个, 为满足该条件本文取阈值为 0.45. 多标签标准化算法如下:

```
读入 N 篇新闻文本的读者投票数据
data = originalData(:,1:6);
mean_row = mean(data,2);
std_row = std(data,0,2);
[row,col] = size(data);
for i = 1:row
        data(i,:) =
(data(i,:)-repmat(mean_row(i),1,col))/std_row(i);
end
threshold = 0.45;
index_1 = find(data > threshold);
data(:,:) = 0;
data(index_1) = 1;
sum_row = sum(data,2);          % sum_row: 每个样本中 1 的个数, 用于检测是否在 1 和 3 之间
若返回两个空矩阵说明每个样本中‘1’的个数都不少于 1 个且不多于 3 个!否则, 则不满足条件, 需要对阈值进行调节!
输出文件.xls, 确定新闻文本的类别标签
```

2.2.2 补全矩阵的 LDA 建模过程

以上 2.1.2 LDA 模型分类的研究中, 完全没有考虑标签之间的相关性. 现实中, 标签之间的关系并不是相互独立的, 而是在一定程度上是有联系的. 本文采用文献[12]介绍的多标签相关性的二阶策略, 考虑标签两两之间的关联关系. 那么, 通过数学统计分析得到了稳定性相关性矩阵, 如何将这种相关性矩阵应用到分类模型中, 能否使分类结果更准确呢?

为了解决这个问题, 本文考虑利用该相关性对初始标签矩阵 Y 进行补全, 从而来得到包含更多标签信息的补全标签矩阵 $\hat{Y} \in R^{n\times c}$, 最后通过对得到的补全标签矩阵建立模型以期达到更好的分类结果. 考虑到不同标签间的共现性和依赖关系, 我们假设补全标签矩阵 $\hat{Y}$ 的构建是由原始得到的不完全标签矩阵 Y 和通过数组实验准确学习的标签相关性矩阵 $S \in R^{c\times c}$ 决定的. 具体如何构造, 是受标签依赖传播思想[13]的启发, 即对不完全标签矩阵 Y 和标签相关矩阵 S 直接进行矩阵相乘来补全增强(如图 6 所示): $\hat{Y} = Y \times S$, 其中补全矩阵 $\hat{Y}$ 中的每个元素 $\hat{Y}_{ij}$ 代表第 i 个样本 x_i 标识为第 j 个标签的置信度, 而且该置信度会被原始矩阵中第 i 个样本 x_i 拥有的其他标签的先验条件所影响, 具体如下:

$$\begin{aligned}\hat{Y} &= Y_{i,1}\times S_{1,j} + Y_{i,2}\times S_{2,j} + \ldots + Y_{i,c}\times S_{c,j} \\ &= \sum_{t=1}^{c} Y_{i,t}\times S_{t,j}\end{aligned} \tag{2}$$

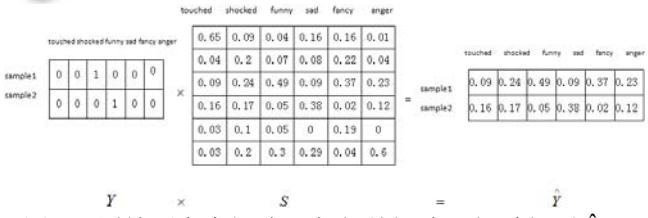

图 3 原始不完全矩阵 Y 与相关矩阵 S 相乘得到 $\hat{Y}$

如果在原始标签矩阵中, 已知样本包含一些与第 i 个标签相关性比较大的其他标签的话, 那么该样本很有可能也包含该标签, 其中会有更多的标签信息, 从而影响实验结果. 然后对补全标签矩阵 $\hat{Y}$ 再次进行多标签标准化操作 $\chi' = \frac{\chi_i - \bar{\chi}}{\sigma}$, 利用相同的文档编号 id, 将 LDA 模型与补全矩阵进行水平串联, 构造 CM-LDA 模型进行多标签分类. 例如 sample1: “印度猴子偷吃捣乱被关进笼子游街示众”, 投票数据显示读者以“搞笑”情绪为主, “新奇”情感次之. 在多标签标准化处理后新闻标签只有“搞笑”, 但通过标签补全后给该新闻又分配了“新奇”的标签. 说明对读者情绪数据进行标签标识时, 标准化算法漏掉了某些重要信息, 这样会影响实验的整体分析效果. 而补全矩阵补全了漏掉的标签, 使文本情绪标签与读者投票数据一致.

3 实验与分析

3.1 多标签分类算法及评估标准

本文多标签分类算法选择 RAkEL 分类器, RAkEL 是构造了 Label PowerSet(LP)[14]分类器的一个集成. LP 算法是将全部标签的每个子集看作是独立的, 这样可

以将多标签分类问题转化. 但是, 测试样本以及所有的子集个数决定了LP算法的计算复杂度. 随着标签数变多, 标签的子集的数目呈指数级增长, 这样 LP 算法的计算复杂度将会很大. RAkEL 算法解决了这一缺陷, 仅仅利用了一部分子集作为标签.

常用的多标签分类评估准则为 Hamming Loss(HL) 、 One-Error(OE) 、 Ranking Loss(RL) 、 Coverage(COV)、Average Precision(AVP). 前四个评估值越小越好, 但最后的 AVP 值越大则分类的准确精度就越高.

3.2 标签相关性统计结果分析

通过上文提到的多标签标准化算法, 对 8583 篇新闻的读者情绪投票数据进行 0 或 1 标签标识之后, 根据不同标签数进行统计的结果如表 1 所示.

表 1 读者情绪标签组合及对应样本数

标签个数	情绪集合	样本数
1	愤怒	2219
	搞笑	1414
	感动	1067
	难过	591
2	搞笑 愤怒	889
	难过 愤怒	467
	震惊 搞笑	190
	感动 难过	158
3	震惊 搞笑 愤怒	66
	搞笑 难过 愤怒	66
	震惊 难过 愤怒	34
	震惊 搞笑 难过	23

从表 1 中统计发现, 搞笑情绪和愤怒情绪组合出现的次数最多, 难过情绪和愤怒情绪次之. 在三个标签的组合中, 震惊情绪、搞笑情绪和愤怒情绪组合也常出现, 对数据的处理也有一定的影响. 可见, 分析标签相关性对情感分类的影响, 对原始标签进行补全方法是可行的.

于是, 对 3188 篇、5235 篇及 8583 篇新闻投票数据分别进行多标签标准化算法处理, 通过多组实验验证, 统计相同标签下不同文本数、哪些标签共同出现在同一篇新闻的文本数, 来学习得到相关性矩阵. 横坐标为读者情感, 纵坐标为共现情绪在包含该情绪的总票数中所占比率, 结果如图 4、图 5、图 6 所示.

从 3 个图中发现, 读者投票数据归一化处理后, 两两标签间的关系是一个稳定的状态, 也就是说从数据中准确地学习的标签相关性矩阵具有一般性, 即标签之间的相关性大小趋于稳定, 如表 2. 每行之和为 1, 除对角线外其他行列对应的值为该列情感占该行情感的共现比率.

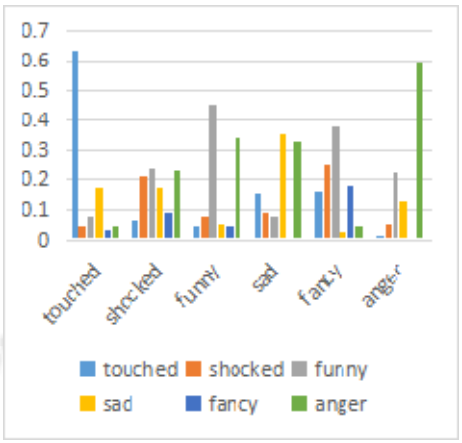

图 4 3188 篇新闻读者情感投票共现统计结果

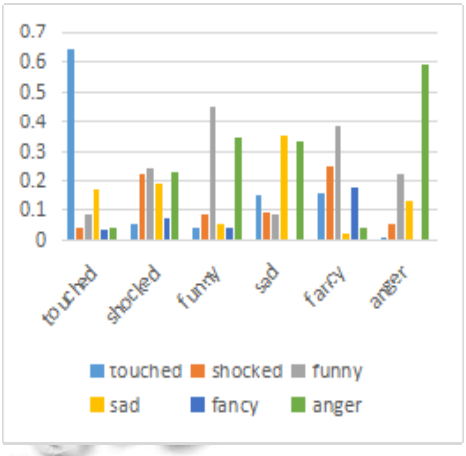

图 5 5235 篇新闻读者情感投票共现统计结果

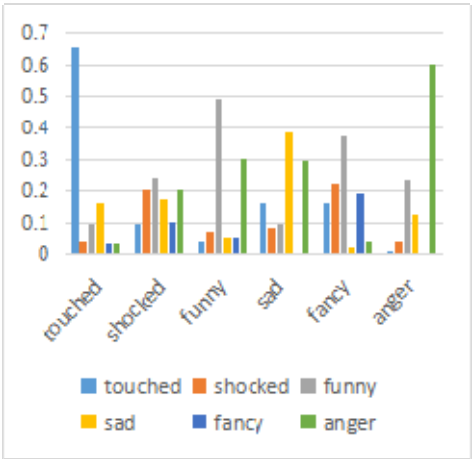

图 6 8583 篇新闻读者情感投票共现统计结果

表 2 读者投票共现统计的相关矩阵

	touched	shocked	funny	sad	fancy	anger
touched	0.65	0.04	0.09	0.16	0.03	0.03
shocked	0.09	0.20	0.24	0.17	0.10	0.20
funny	0.04	0.07	0.49	0.05	0.05	0.30
sad	0.16	0.08	0.09	0.38	0	0.29
fancy	0.16	0.22	0.37	0.02	0.19	0.04
anger	0.01	0.04	0.23	0.12	0	0.6

从上述表 2 中发现, 有些标签两两之间确实存在着联系. 对于包含震惊情绪的新闻总数中, 仅包含震惊情绪的新闻占总数的 20%, 搞笑情绪竟占 24%, 而愤怒情绪也占了 20%. 可见, 在一篇新闻被确定为震惊标签的时候会受到搞笑情绪和愤怒情绪的影响. 愤怒情绪和新奇情绪的共现率就很小, 几乎互不影响. 因此, 标签之间的共现模式一定体现了它们之间所蕴含的某种语义相关性.

3.3 实验结果分析

本文将抓取的新浪社会新闻语料的 7000 篇作为训练集, 剩余的作为测试语料. 实验中, 取 20 到 100, 得到 9 个不同主题数下的 9 种文档表示, 然后使用 RAkEL 来多标签分类. 将补全矩阵模型 CM-LDA 和原始矩阵的 LDA 模型在不同主题数下, 进行分类性能的比较, 加粗表示在该主题数下应用的分类模型性能最好, 实验结果如表 3 所示.

表 3 RAkEL 在不同主题数的 CM-LDA 和 LDA 模型下的分类性能比较

T 主题数	分类模型	HL	OE	COV	RL	AVP
20	LDA	0.1728	0.3378	1.3603	0.1765	0.7714
	CM-LDA	0.1484	0.2674	1.4349	0.1906	0.7887
30	LDA	0.1675	0.3161	1.0099	0.1654	0.7846
	CM-LDA	0.1245	0.1432	1.0298	0.1115	0.8480
40	LDA	0.1676	0.3278	1.2804	0.1642	0.7822
	CM-LDA	0.1262	0.1750	1.1718	0.1374	0.8413
50	LDA	0.1693	0.3150	1.2886	0.1632	0.7873
	CM-LDA	0.1333	0.1036	0.9419	0.0941	0.8520
60	LDA	0.1627	0.3077	1.2567	0.1579	0.7623
	CM-LDA	0.1370	0.0872	0.8124	0.0725	0.8422
70	LDA	0.1619	0.3023	1.2412	0.1551	0.7975
	CM-LDA	0.1365	0.0748	0.7119	0.0612	0.8428
80	LDA	0.1638	0.3192	1.2456	0.1584	0.7890
	CM-LDA	**0.1241**	**0.0663**	**0.5276**	**0.0471**	**0.8572**
90	LDA	0.1618	0.3045	1.2452	0.1577	0.7947
	CM-LDA	0.1291	0.0794	0.5716	0.0794	0.8469
100	LDA	0.1631	0.3112	1.2665	0.1613	0.7905
	CM-LDA	0.1312	0.0739	0.6436	0.0509	0.8422

通过上表 3 可见, 补全矩阵模型 CM-LDA 比原始矩阵的 LDA 模型分类结果准确率高. 尤其是在主题数为 80, 其他参数设置 *alpha* =0.5, *beta* =0.1 时, HL、OE、COV、RL 这四个评估标准均较低, 准确率也达到了 85.72%, 可知该模型 CM-LDA 的综合性能最优. 这说明, 在传统的 LDA 主题模型上, 利用标签相关性对原始不完全的标签矩阵进行补全增强, 获得的 CM-LDA 模型能够改善多标签分类的性能.

图 7 表示在不同主题数下, 多标签分类算法 RAkEL 在原始矩阵的 LDA 和补全矩阵模型 CM-LDA 分类准确率比较, 更清晰地将实验结果展现在了曲线图上. 很明显地看到, 基于 CM-LDA 模型的分类准确率高于原始矩阵的 LDA 模型. 因此, 从整体的分类结果来看, 基于 CM-LDA 模型的读者情感分类方法是可行的, 补全矩阵的应用能提高多标签的分类性能, 同时也为接下来新闻读者情感预测的深入研究做了铺垫.

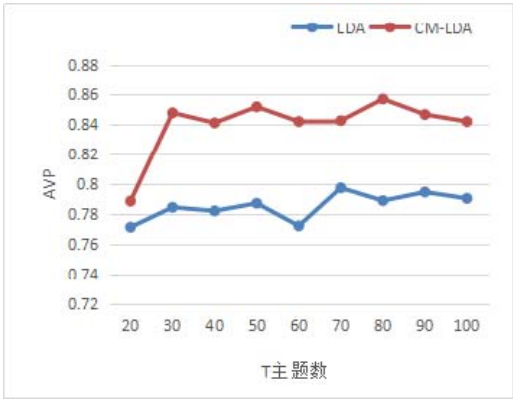

图 7 LDA 和 CM-LDA 模型下的 RAkEL 算法分类准确率对比

4 结语

通过在真实多标签数据集(即新浪社会新闻语料)上的实验中, 学习到相关性矩阵, 然后对原始不完全矩阵进行补全获得的补全矩阵, 不仅验证了标签间相关关系的语义合理性, 而且也证明了利用本文提出的 CM-LDA 模型进行多标签分类的正确性和实用性, 这对热点问题或突发事件做好社会舆情预测的研究是很有价值的. 而如何学习主题与标签之间的联系以及进一步获得更高的读者情感预测的准确率, 则是将来需要进一步研究的方面.

参考文献

1 Lin KHY, Yang C, Chen HH. Emotion classification of online news articles from the reader's perspective. Proc. 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT '08). 2008. 220–226.

2 Blei DM, Mcauliffe JD. Supervised topic models. Advances in Neural Information Processing Systems, 2010, 3: 327–332.

3 卢露,魏登月.基于读者视角的文本情感分类.微电子学与计算机,2014,(10):122–125.

4 Bao S, Xu S, Zhang L, et al. Mining social emotions from affective text. IEEE Trans. on Knowledge & Data Engineering, 2011, 24(99): 1658–1670.

5 Zhang ML, Zhou ZH. Multilabel neural networks with applications to functional genomics and text categorization. IEEE Trans. on Knowledge & Data Engineering, 2006, 18(10): 1338–1351.

6 Fürnkranz J, Hüllermeier E, Mencía EL, et al. Multilabel classification via calibrated label ranking. Machine Learning, 2008, 73(2): 133–153.

7 Read J, Pfahringer B, Holmes G, et al. Classifier chains for multi-label classification. Machine Learning, 2011, 85(3): 254–269.

8 Tsoumakas G, Katakis I, Vlahavas I. Random k-labelsets for multilabel classification. IEEE Trans. on Knowledge & Data Engineering, 2010, 23(7): 1079–1089.

9 叶璐.新闻文本的读者情绪自动分类方法研究[硕士学位论文].哈尔滨:哈尔滨工业大学,2012.

10 Schapire RE, Singer Y. BoosTexter: A boosting-based system for text categorization. Machine Learning, 2000, 39(2-3): 135–168.

11 王霄,周李威,陈耿,等.一种基于标签相关性的多标签分类算法.计算机应用研究,2014,31(9):2609–2612.

12 胡春安,范丽文,毛伊敏.HPDBSCAN:高效的不确定数据处理算法.计算机工程与设计,2013,34(3):1044–1049.

13 Pizzuti C. A multi-objective genetic algorithm for community detection in networks. 21st International Conference on Tools with Artificial Intelligence(ICTAI '09). IEEE. 2009. 379–386.

14 Tsoumakas G, Katakis I. Multi-label classification: An overview. International Journal of Data Warehousing & Mining, 2009, 3(3): 1–13.