DOI:10.13718/j.cnki.xsxb.2018.09.012

基于不平衡情感分类的 Lasso-Lars 特征选择方法研究①

万会芳[1], 闵 兰[2], 舒 畅[2]

成都理工大学 管理科学学院, 成都 610059

摘要: 基于 Lasso 回归和支持向量机分类器, 首先利用 Lasso 回归具有变量筛选的特点, 过滤部分不重要的特征, 然后利用支持向量机分类器做情感提取. 在某化妆品品牌的评论数据实验中, 利用基础情感词典和领域情感词典构建待选择高维特征集, 通过对比特征选择前后的 G-means, 精确度和召回率等, 均取得显著效果.

关 键 词: 不平衡情感分类; 特征选择; Lasso

中图分类号: TP391 **文献标志码**: A **文章编号**: 1000-5471(2018)09-0074-05

情感分析(sentiment analysis, SA)是对带有情感色彩的主观性文本进行分析、处理、归纳和推理的一种技术[1]. 情感分析需要用到的技术和方法通常来自于文本挖掘、信息检索、机器学习、自然语言处理以及统计学等方面[2]. 情感分析可以分为篇章、语句和词语 3 个层次的问题; 或者分为情感分类、情感检索、情感抽取等子问题[3]. 文献[4]利用自然语言处理的方法, 对中文网络评论语句进行语义极性分析和观点抽取, 提出了一种计算上下文极性的算法, 并与手工标注结果比较, 验证了算法的合理性. 文献[5]基于情感词典 HowNet 提出了语义相关场和语义相似度两种方法计算词语的情感类型, 判别准确率可达 80%以上. 文献[6]分析了产品评论和新闻评论的研究进展, 总结了中文文本情感分析的难点和未来的研究方向.

目前大部分情感分类问题并没有考虑分类样本的平衡性问题, 而对于不平衡数据的特征选择该如何去研究, 是一个重要的问题. 对于情感分类中的不平衡数据特征选择, 这方面的研究较少. 主要在推荐系统、过滤系统、问答系统等领域有广泛的应用[7-9]. 套索(least absolute shrinkage and selection operator, Lasso)方法是对非负绞除法的改进[8], 在参数估计的过程中, 同时实现了变量的选择. 本文基于 Lasso 方法在高维数据特征选择上的优越性, 将其推广到不平衡情感分类中, 研究 Lasso 方法对于不平衡情感数据分类的有效性.

情感分类的任务主要分为主、客观分类和褒贬分类, 本文主要研究情感分类中的主观信息的褒贬情感极向分类. 文献首次提出不平衡情感分类问题, 指出虽然情感分类提出多年, 但大多是针对平衡数据分类, 并就此提出一种基于聚类的欠采样技术来解决不平衡问题, 取得了初步效果. 文献[10]最早研究了中文不平衡情感分类问题, 提出了基于欠采样和多分类算法的集成学习框架. 笔者通过查阅文献发现, 对于中文不平衡情感分类的研究较少, 而针对不平衡情感分类中的特征选择问题更少, 基于此本文尝试将 Lasso 方法, 引入到情感分类特征选择中, 通过具体实验分析, 验证本文方法的有效性. 支持向量机[11](support vector machine, SVM)自提出以来被广泛应用于机器学习、数据挖掘、情感分析等领域, 是最常用的分类模型, 在以往关于支持向量机的研究中均表明, 其是最好的分类器.

① 收稿日期: 2017-12-10

作者简介: 万会芳(1966-), 女, 硕士, 副教授, 主要从事数据挖掘、模糊数学研究.

1　Lasso 方法

一般的线性回归模型：

$$\boldsymbol{Y}_i = x'_i\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i,\ i = 1,2,\cdots,n \tag{1}$$

其中：$\boldsymbol{\beta}$ 是 $p_n \times 1$ 维列向量，$\boldsymbol{\varepsilon}_i$ 是独立同分布的且各观测值看作是独立的，同时假设 x_{ij} 是经过标准化之后得到的，即

$$1/n\sum\nolimits_i x_{ij} = 0,\ 1/n\sum\nolimits_i x_{ij}^2 = 1 \tag{2}$$

当训练集维数和样本量几乎相等或者超过样本量时，有些回归系数是稀疏的即有些元素为 0，传统的最小二乘法不再适用，这就需要正则化方法或者惩罚方法来代替. 常用的包括岭回归和 Lasso 方法，但是，通过岭回归得到的模型包含全部的特征，不能进行特征选择. 和岭回归不同的是 Lasso 方法采用 L_1 范数 $\sum_{j=1}^{p_n} |\boldsymbol{\beta}_j|$，而岭回归是 L_2 范数 $\sum_{j=1}^{p_n}\boldsymbol{\beta}_j^2$. Lasso 方法的参数估计

$$\begin{cases}\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}(\lambda) = \arg\min\limits_{\beta}\sum\nolimits_{i=1}^{n}(\boldsymbol{Y}_i - \boldsymbol{x}'_i)^2 + \lambda\sum\nolimits_{j=1}^{p}|\boldsymbol{\beta}_j| \\ s.t.\ \sum\nolimits_j |\boldsymbol{\beta}_j| \leq \lambda\end{cases} \tag{3}$$

其中 $\lambda \in [0, +\infty)$ 是调和参数，对它的控制将会使回归系数总体变小. 若令 $\hat{\beta}_j^0$ 是回归参数的最小二乘估计值，就会使一些回归系数缩小并趋于 0，有些甚至会等于 0. 第一部分表示模型拟合的优良性，第二部分表示对参数的惩罚. 调和系数 λ 越小，模型的惩罚力度就越小，保留的特征就越多；反之特征减少. 因此，Lasso 方法经常被用来特征选择，它有两个优点. 其一，Lasso 方法在特征选择的过程中是连续的且很稳定；其二，Lasso 方法对于高维数据的时间复杂度较低.

2　实　　验

本文的目的不仅是验证 Lasso 方法在不平衡情感分类中的精度，而且比较了不同分类器在情感分类中的实验结果. 因为在实际应用中，不仅要考虑分类精度问题，也要考虑时间复杂度的问题. 第一小节介绍本文的实验设置部分，包括数据的介绍及数据的整理；第二小节介绍评价分类准确性的方法；第三小节详细描述实验结果及分析.

2.1　实验设置

本文的语料来源于数据交易公司数据堂提供的情感分析语料，是化妆品品牌的评论数据，并标注有情感类别，数据集描述见表 1. 其中正实例表示积极情感，负实例表示消极情感.

表 1　数据集描述

语料来源	样本量 / 个	正实例 / 个	负实例 / 个	正负实例数比值 /%
化妆品	100033	96415	3618	26.65

实验环境为 Intel Core(TM) i5－4200H CPU @ 2.8GHz，12.0GB 内存. 实验语言为 python，调用了机器学习库 scikit-learn. 实验中，首先对样本数据做分词及去停用词处理，选用的工具是 R 的 Rwordseg 中文分词工具包，这是一个 R 环境下的中文分词工具，引用了 Ansj 包，Ansj 是一个开源的 java 中文分词工具，基于中科院的 ICTCLAS 中文分词算法，采用隐马尔科夫模型(HMM). Rwordseg 的优点是分词准确且速度快，并且可以导入自定义的词库，尤其是搜狗的细胞词库，所以 Rwordseg 在中文分词方面有很强的优势. 然后将每条评论看作一个文本，对其分词之后得到的词称为项(Term)，基于向量空间模型[12](vector space model，VSM)，将文本内容转化为可以计算的空间向量模型.

2.2　不平衡数据分类的评价方法

对于分类问题来说，传统的方法是通过计算分类的准确性来评估模型的精确性. 但对于不平衡数据来说，单靠准确性已经不能满足要求. 为了分析不平衡情感分类的准确性，本文使用了准确性、敏感度、特异

性、G-means 以及 F-value 作为指标. 对于不平衡数据的预测结果可以分为 4 类(如表 2 所示),被称为模糊矩阵. 实际正类、预测正类记为 TP,实际正类、预测负类记为 FN,实际负类、预测正类记为 FP,实际负类、预测负类记为 TN,每一类的样本数目记作 N_{TP},N_{FP},N_{FN},N_{TN}.

表 2 混淆矩阵

	预测正类	预测负类
实际正类	TP	FN
实际负类	FP	TN

评价平衡数据分类准确性一般用 P_{acc} 来表示,计算方法为 $P_{acc}=(N_{TP}+N_{TN})/(N_{TP}+N_{FP}+N_{FN}+N_{TN})$ 称为精确度. 文献[13] 提出的 *G-means* 被用来衡量不平衡数据的分类准确性,它表示的是少数类分类精度和多数类分类精度的几何平均值. *G-means* 越大表示分类精度越高,且只有少数类和多数类分类精度都高的时候整体的分类精度才会很高.

$$G\text{-}means = \sqrt{P_{sens} \times P_{spec}} \tag{4}$$

其中

$$P_{sens} = N_{TP}/(N_{TP}+N_{FN}) \tag{5}$$

$$P_{spec} = N_{TN}/(N_{TN}+N_{FP}) \tag{6}$$

分别表示正实例和负实例的分类精度,称为敏感度和特异性. *F-value* 的计算公式为:

$$F\text{-}value = 2N_{TP}^2/(2N_{TP}^2 + N_{TP} \cdot N_{FN} + N_{FN} \cdot N_{FP}) \tag{7}$$

2.3 实验结果

实验结果部分主要分析本文提出的 Lasso 方法在不平衡情感分类特征选择中的有效性. 对比了信息增益、文档频率、卡方统计量以及 Lasso 方法在 4 种实验语料中的分类精度,主要是敏感度、特异性以及 *G-means* 的比较,目的是说明哪种方法可以取得最优分类精度的特征子集.

首先使用 python 编程,引入 Lars 算法,输出 Lasso 系数解路径,见图 1. 图中每条线代表每个变量的系数变化轨迹. 由解路径可以看出,当调和参数趋向于 0 的时候,特征基本被保留了下来,而当增大的时候,模型选择能力增强,系数被压缩,直到部分系数变为 0. 表 3 是 Lasso 方法的参数估计值,其中非零特征代表最终选择的特征,即最优特征子集. 对于不平衡情感分类影响显著的特征被保留了下来,而影响不显著的特征系数被压缩到 0 而达到选择的目的.

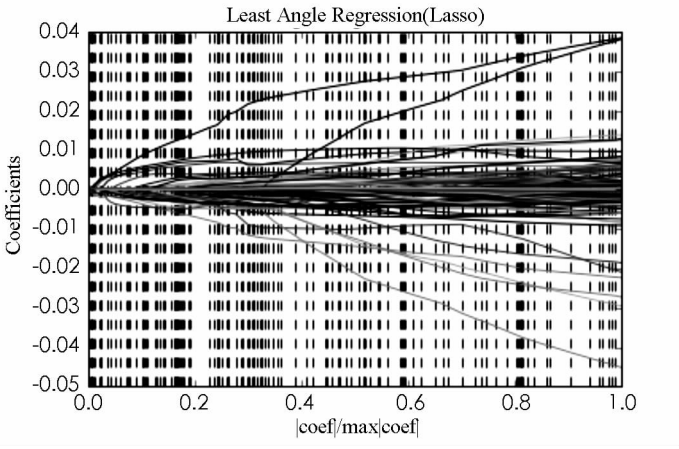

图 1 Lasso 系数解路径图

基于 Lasso 特征选择方法从原始的 102 个特征中选择了 44 个非零特征,然后利用传统的支持向量机分类模型,做情感分类.

根据分类得到的混淆矩阵,计算敏感度、特异性以及 *G-means*,比较不同特征选择方法的分类效果,结果见表 4,结果保留 3 位小数.

根据实验结果易见,Lasso 特征选择方法的 *G-means* 为 0.924,而分类精度也达到了 93.7%,可以看出其精度均高于 IG,DF,CHI 特征选择方法.

图 2 和图 3 中的垂直虚线表示的是交叉验证选择的惩罚值,图 2 的值为 0.032 648,图中的值是对其取以 10 为底的负对数得到的即 1.486 14. 图 3 的惩罚值及其负对数分别为 0.327 42 和 1.484 884;其中的斜虚线代表的是测试集上的均方误差随着交叉验证中 alpha 取值的不同的变化轨迹,实线代表的是其平均值.

表 3　Lasso 系数表

特征	系数	特征	系数	特征	系数	特征	系数	特征	系数	特征	系数
x_1	−0.002 59	x_{18}	−0.000 89	x_{35}	0	x_{52}	0	x_{69}	0.004 17	x_{86}	0
x_2	0	x_{19}	−0.002 31	x_{36}	0	x_{53}	0	x_{70}	0.000 98	x_{87}	−0.004 93
x_3	0.000 56	x_{20}	−0.000 22	x_{37}	0	x_{54}	0	x_{71}	0.002 94	x_{88}	0
x_4	−0.000 03	x_{21}	0.000 7	x_{38}	0	x_{55}	0	x_{72}	0	x_{89}	−0.003 9
x_5	0.006 86	x_{22}	0	x_{39}	0.000 71	x_{56}	0	x_{73}	0	x_{90}	−0.003 39
x_6	−0.001 45	x_{23}	0	x_{40}	−0.003 55	x_{57}	0	x_{74}	0	x_{91}	−0.002 53
x_7	−0.003 68	x_{24}	0	x_{41}	0.002 56	x_{58}	0	x_{75}	0	x_{92}	0
x_8	−0.002 32	x_{25}	0	x_{42}	−0.001 8	x_{59}	0.002 1	x_{76}	0	x_{93}	0
x_9	0	x_{26}	0.003 79	x_{43}	0	x_{60}	−0.000 1	x_{77}	0	x_{94}	0
x_{10}	−0.001 46	x_{27}	0	x_{44}	−0.002 85	x_{61}	−0.000 72	x_{78}	0.001 35	x_{95}	0
x_{11}	0.002 15	x_{28}	−0.005 66	x_{45}	0	x_{62}	0	x_{79}	0.008 92	x_{96}	−0.004 42
x_{12}	0	x_{29}	−0.009 46	x_{46}	0	x_{63}	0	x_{80}	0	x_{97}	0
x_{13}	0	x_{30}	−0.002 66	x_{47}	0	x_{64}	0	x_{81}	0	x_{98}	0
x_{14}	−0.004 71	x_{31}	0	x_{48}	0	x_{65}	0	x_{82}	0	x_{99}	0
x_{15}	0	x_{32}	0.018 67	x_{49}	−0.000 19	x_{66}	0.002 8	x_{83}	0	x_{100}	0
x_{16}	0	x_{33}	0.007 48	x_{50}	0	x_{67}	0	x_{84}	0	x_{101}	0
x_{17}	0	x_{34}	−0.000 26	x_{51}	0.000 54	x_{68}	−0.000 15	x_{85}	−0.003 57	x_{102}	0

表 4　化妆品评论语料实验结果

方法	精确度	敏感度	特异性	*F-value*	*G-means*
IG	0.857	0.877	0.833	0.871	0.855
DF	0.860	0.880	0.835	0.873	0.857
CHI	0.834	0.799	0.900	0.863	0.848
Lasso	0.937	0.911	0.938	0.947	0.924

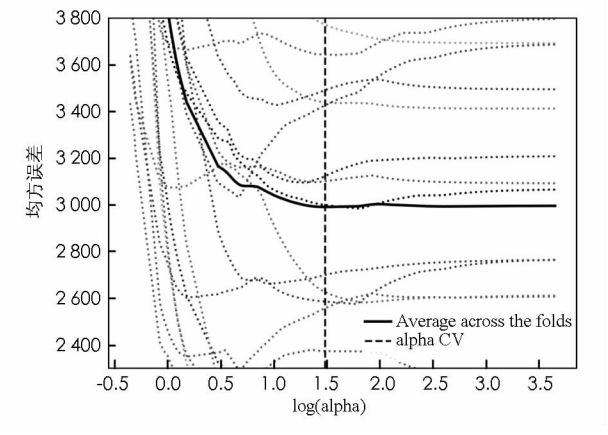

图 2　均方误差基于 Lars 算法的 Lasso 特征选择变化图

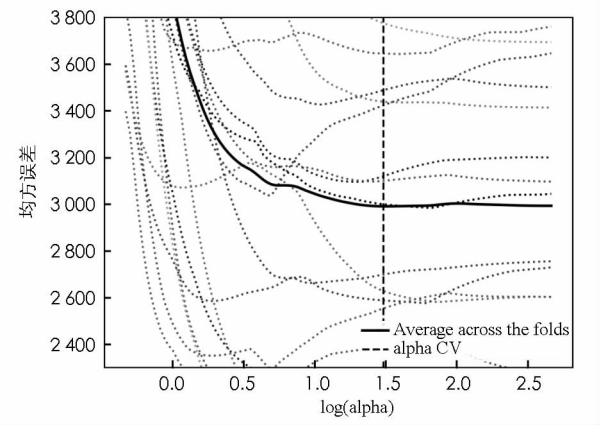

图 3　均方误差不做特征选择的变化图

4　结束语

本文介绍了一种用于不平衡情感分类的特征选择方法，目的是将 Lasso 方法引入到处理不平衡情感数据特征选择中，基于传统特征选择方法受限于不平衡情感特征选择、分类精度不高等问题，尝试引入 Lasso

方法. 实验结果表明，该方法提高了情感分类的准确性. 本文的基础工作是研究如何将 Lasso 方法应用到不平衡情感分类，但还有一些值得继续深入的问题，比如对于更高维的特征选择的对比研究，如果扩大情感词典，待选择特征子集数量增加，结果又会是什么，这也是本文今后的研究方向.

参考文献：

[1] 张 林，钱冠群，樊卫国，等. 轻型评论的情感分析研究 [J]. 软件学报，2014，25(12)：2790－2807.
[2] 张紫琼，叶 强，李一军. 互联网商品评论情感分析研究综述 [J]. 管理科学学报，2010，13(6)：84－96.
[3] 赵妍妍，秦 兵，刘 挺. 文本情感分析 [J]. 软件学报，2010，21(8)：1834－1848.
[4] 娄德成，姚天昉. 汉语句子语义极性分析和观点抽取方法的研究 [J]. 计算机应用，2006，26(11)：2622－2625.
[5] 朱嫣岚，闵 锦，周雅倩，等. 基于 HowNet 的词汇语义倾向计算 [J]. 中文信息学报，2006，20(1)：14－20.
[6] 魏 韡，向 阳，陈 千. 中文文本情感分析综述 [J]. 计算机应用，2011，31(12)：3321－3323.
[7] MEDHAT W，HASSAN A，KORASHY H. Sentiment Analysis Algorithms and Applications：A Survey [J]. Ain Shams Engineering Journal，2014，5(4)：1093－1113.
[8] TIBSHIRANI R. Regression Shrinkage and Selection via the Lasso [J]. Journal of the Royal Statistical Society，1996，58(1)：267－288.
[9] WANG Z，SHOUSHAN L I，ZHU Q，et al. Chinese Sentiment Classification on Imbalanced Data Distribution [J]. Journal of Chinese Information Processing，2012，26(3)：33－32.
[10] 王志昊，王中卿，李寿山，等. 不平衡情感分类中的特征选择方法研究 [J]. 中文信息学报，2013，27(4)：113－118.
[11] TONG S，KOLLER D. Support Vector Machine Active Learning with Applications to Text Classification [J]. Journal of Machine Learning Research，2001，2(1)：999－1006.
[12] 许建豪. 采用向量空间模型的个性化信息检索方法 [J]. 华侨大学学报(自然科学版)，2016，37(2)：175－178.
[13] BRADLEY A P. The Use of the Area Under the ROC Curve in the Evaluation of Machine Learning Algorithms [J]. Pattern Recognition，1997，30(7)，1145－1159.

Feature Selection in Imbalanced Sentiment Classification：A Method Using Lasso-Lars

WAN Hui-fang[1]， MIN Lan[2]， SHU Chang[2]
College of Management Science，Chengdu University of Technology，Chengdu 610059，China

Abstract：The characteristics of textual emotion analysis are usually of high dimension and sparseness. Lasso has a simple and efficient trait in feature selection. This paper introduces the Lasso regression into the unbalanced emotion analysis and achieves remarkable results. Applying emotional analysis in e-commerce plays an important role in improving product quality and improving service，which attracts many researchers and has high research value. In fact，the number of positive comments on e-commerce data generally exceeds the number of bad reviews. If the feature selection is not reasonable，it is easy to ignore the bad reviews，and the bad reviews are the key to analyzing the problems. Based on the Lasso regression and SVM classifier，this paper first uses Lasso regression to filter the features that have variable screening，filters some unimportant features，and then makes use of SVM classifier to extract the emotion. In a cosmetic brand's reviewing data experiment，the basic emotion dictionary and domain sentiment lexicon are used to construct the high-dimensional feature set to be selected，and the significant effects are achieved by comparing G-means before and after feature selection，accuracy and recall.

Key words：imbalanced sentiments classification；feature selection；Lasso

责任编辑 张 枸