DOI:10.13546/j.cnki.tjyjc.2019.10.016

基于再缩放策略的逻辑回归算法及其应用

李琼阳
（许昌学院 数学与统计学院，河南 许昌 461000）

摘 要：文章提出了一种利用过滤式属性筛选-去相关的方法选择建模属性，极大程度地削弱属性之间的共线性程度。逻辑回归算法在进行类别判定时默认以0.5为判定阈值，但此种判定方式在不平衡数据集上的效果不理想，利用再缩放策略研究逻辑回归算法在不平衡数据集上的应用，并在某运营商提供的垃圾短信用户行为消费特征样本数据上进行实证分析。结果表明，经由过滤式属性筛选-去相关选取建模变量之后，基于再缩放策略的逻辑回归学习器具有良好的准确率和普适性。

关键词：过滤式属性筛选；属性相关；不平衡数据集；逻辑回归
中图分类号：TP391.7　　**文献标识码**：A　　**文章编号**：1002-6487(2019)10-0072-03

0 引言

逻辑回归算法是目前公认的一种简单有效的分类算法，它是一种基于概率的分类方法，也是应用最广的模型。但共线性问题的存在，极有可能会使得回归系数不显著，增大了估计参数的均方误差和标准误，还可能使回归系数的方向相反[1]，对模型造成极大的干扰，影响模型的分类准确率。此外，逻辑回归算法有一个基本假设，即默认以0.5为类别判定阈值，此种假设往往适用于不同类别训练样例数相当或稍有差别的情况，但若各类样例数量相差悬殊，训练出来的模型将无实际意义。如训练样本集中，有998个反例，但正例只有两个，那么学习方法只需返回一个永远将新样本预测为反例的学习器，就能达到99.8%的精确度，但这样的学习器没有任何价值，因为它不能预测出任何正例。

基于逻辑回归算法存在的这两个问题，不少学者致力于对其进行改进，以提高算法的分类准确率和普适性。改进之处主要体现在两个方面：(1)削弱建模属性之间的共线性程度；(2)提高其在不平衡样本集上的适用性。

为了解决逻辑回归算法在实际应用过程中面临的这两个问题，在以往研究文献基础之上[1-8]，本文首先构造属性筛选器，有效解决共线性的问题。其次，基于再缩放策略，改进逻辑回归算法的分类阈值，以此削弱不平衡样本集的影响，提高模型的准确率。最后，在某运营商提供的垃圾短信用户行为消费特征样本数据上进行实证分析。

1 构造属性筛选器

1.1 过滤式属性选择

过滤式属性选择，是当前较为流行的一种属性选择方式，该方法首先从建模数据的属性集合中选择合适的建模属性，继而把样本投入模型进行训练[9]。

Relief是比较典型的过滤式特征选择方式，该方式设计了一个与特征空间维数相同的“相关统计量”来衡量对应的建模特征对目标变量的重要程度。该统计量是一个向量，向量的每一个分量均与一个初始建模特征相互对应，通过设置一个阈值 τ，然后在向量中选择比 τ 大的分量所对应的初始建模特征即可；也可指定合适的特征个数 k，然后选择该向量中分量最大的前 k 个特征。其算法流程如下：

(1)寻找近邻：给定训练集 $\{(x_1,y_1),(x_2,y_2),\cdots,(x_m,y_m)\}$，对每个示例 x_i，Relief先在 x_i 的同类样本中寻找最近邻 $x_{i,nh}$，称为“猜中近邻”($near-hit$)，再从 x_i 的异类样本中样本中寻找最近邻 $x_{i,nm}$，称为“猜错近邻”($near-miss$)。

(2)计算每个属性对应的相关统计量：相关统计量对应于属性 j 的分量为：

$$\delta^j=\sum_i -diff(x_i^j,x_{i,nh}^j)^2+diff(x_i^j,x_{i,nm}^j)^2 \quad (1)$$

其中，x_a^j 表示一个样本 x_a 在属性 j 上的值，$diff(x_a^j,x_b^j)$ 取决于属性 j 的类型：若属性 j 为离散型，则 $x_a^j=x_b^j$ 时，$diff(x_a^j,x_b^j)=0$，否则为1；若属性 j 为连续型，则 $diff(x_a^j,x_b^j)=|x_a^j-x_b^j|$，注意 x_a^j、x_b^j 已规范化到[0,1]区间。

(3)对基于不同样本得到的估计结果进行平均，就得到各属性的相关统计量，某个分量值越大，那么其对应属性的分类能力就越强。

由此可知，*Relief* 是专门为二分类问题设计的，其扩展变体 *Relief-F* 能处理多分类问题。

假定数据集 D 中的样本来自 $|\gamma|$ 个类别。对于示例 x_i，若它属于第 k 类 ($k\in\{1,2,\ldots,|\gamma|\}$)，则 *Relief-F* 先在

基金项目：许昌学院科研基金资助项目(2019YB029)
作者简介：李琼阳(1991—)，女，河南漯河人，硕士，讲师，研究方向：数据挖掘、经济信息管理。

第 k 类的样本中寻找到 x_i 的最近示例 $x_{i,nh}$，并将其作为猜中近邻，然后在第 k 类之外的每个类中找到 x_i 的最近示例作为猜错近邻 $x_{i,l,nh}(l=1,2,\cdots,|\gamma|;l\neq k)$。于是，向量中对应于初始建模属性 j 的分量表达式为：

$$\delta^j=\sum_i -diff(x_i^j,x_{i,nh}^j)^2+\sum_{l\neq k}(p_l^* diff(x_i^j,x_{i,nm}^j)^2) \quad (2)$$

其中，P_l 代表第 l 类样本在整个建模数据集合 D 中的占比。

基于以上分析，过滤式属性选择的方法既可以适用于多分类模型，又能较好地处理连续属性和离散属性，是进行属性选择时较为理想的一种方法。

1.2 属性约减

共线性问题的存在会对逻辑回归算法造成极大的干扰，但在实际应用过程中，参与建模的变量之间往往会存在一定程度的相关性。相关系数即是反映两个变量之间相关程度的一个重要度量，计算公式如下：

$$\rho_{A_i,A_j}=\frac{\mathrm{cov}(A_i,A_j)}{\sqrt{D(A_i)}\cdot\sqrt{D(A_j)}} \quad (3)$$

一般情况下，$0.4\leqslant\rho_{A_i,A_j}\leqslant0.6$，认为 A_i 与 A_j 之间中等程度相关；$0.6\leqslant\rho_{A_i,A_j}\leqslant0.8$，认为 A_i 与 A_j 之间强相关；$0.8\leqslant\rho_{A_i,A_j}\leqslant1$，认为 A_i 与 A_j 之间极强相关。

为了尽可能避免属性之间的共线性问题，有必要对变量进行约减。约减规则如下：

（1）根据过滤式属性筛选原理筛选一批对目标变量影响程度较大的自变量，通常选择累计占比前80%的变量。

（2）计算筛选出的自变量之间的相关系数，一般当 $\rho_{A_i,A_j}>0.5$ 时，即可认为 A_i 与 A_j 之间有较强的相关性，不宜全部进入模型。

（3）若 $\rho_{A_i,A_j}>0.5$，且变量 A_i 对目标变量的影响力大于变量 A_j，则只选择变量 A_i 参与建模。

综上，训练样本集中的属性经由属性筛选器处理后，基本上可以解决共线性的问题。

2 基于再缩放策略的逻辑回归算法

2.1 逻辑回归算法

逻辑回归算法是一种广义的线性回归分析模型，它是一种基于概率的分类方法，通过模型返回的概率值来判断某种情况发生的可能性大小。

在二分类问题中，逻辑回归的统计学表达式为：

$$In(\frac{p}{1-p})=\beta_0+\beta_1a_1+\beta_2a_2+\cdots+\beta_na_n \quad (4)$$

$$P=\frac{e^{(\beta_0+\beta_1a_1+\beta_1a_1+\cdots\beta_na_n)}}{1+e^{(\beta_0+\beta_1a_1+\beta_1a_1+\cdots\beta_na_n)}} \quad (5)$$

其中，P 代表该问题发生的概率，$1-P$ 代表该问题不发生的概率，$a_1,a_2,\cdots,a_n$ 是参与建模的 n 个属性。$\beta_0,\beta_1,\cdots,\beta_n$ 为回归方程的系数。

在线性分类学习方法中都有一个基本假设，即不同类别的样本例数相当，如果不同类别的训练样例数目稍有差别，通常影响不大。逻辑回归也是一个线性分类器，当对新样本 X 进行分类时，实际上是利用预测出的概率值 P 与一个阈值进行比较，例如通常在 $P>0.5$ 时判定为正例；否则为反例。P 实际上表达了为正例的可能性，几率 $\frac{P}{1-P}$ 则反映了正例可能性与反例可能性之比值。一般默认的阈值为0.5，恰好表明分类器认为真实正、反例可能性相同，即分类器的决策规则为：若 $\frac{P}{1-P}>1$，则预测为正例。

2.2 基于再缩放策略的逻辑回归算法

不平衡数据集是指在一个分类问题中，各类样本数量相差悬殊的一种数据集。此类数据集在经济社会中十分常见，如信用欺诈、肿瘤诊断、垃圾短信识别等。

通常情况下，逻辑回归算法认为当预测的概率值 $P>0.5$ 时，即将新样本预测为正例。但此种做法在不平衡数据集上效果十分不理想。为此，令 m^+ 表示正例数目，m^- 表示反例数目，则该样本集的观测几率即为 $\frac{m^+}{m^-}$。由于通常假设训练集是真实样本的无偏采样，因此观测几率就代表了真实几率。故而据此提出了一种不平衡学习的基本策略——“再缩放”，若 $\frac{P}{1-P}>\frac{m^+}{m^-}$，则预测为正例[9]。

3 实证分析

3.1 数据的收集和处理

某运营商提供了用户行为消费特征样本数据共67085条，其中垃圾短信用户样本数据664条，用1标识，代表正例样本；正常用户66421条，用0标识，代表负例样本，正负样本比例约为1:100。数据集中有当月消费额、品牌、通话时长、发送短信条数、短信回复率、账户余额、是否为垃圾短信用户等共有56个属性。下面将利用基于属性筛选-再缩放策略的逻辑回归算法来进行建模。

3.2 属性筛选

在垃圾短信用户识别过程中，是否是垃圾短信用户是目标变量。利用过滤式属性筛选的计算公式，计算除目标变量外的其他55个变量的影响力。根据前述的规则进行属性约减，最终选定9个变量参与建模，各建模变量对目标变量的影响力和相关系数如表1和下页表2所示。

表1 建模变量的影响力

指标	影响力
消费额	1.84
返还比	1.18
充值次数	1.05
入网时长	0.98
余额	0.62
短信回复率	0.46
是否集团用户	0.36
漫游时长	0.22
是否省外漫游	0.12

表2　　相关系数矩阵

	消费额	返还比	充值次数	入网时长	余额	短信回复率	是否集团用户	漫游时长	是否省外漫游
消费额	1.00	–0.24	0.49	–0.08	0.16	0.33	0.04	0.28	0.28
返还比	–0.24	1.00	–0.33	0.33	–0.11	–0.11	–0.11	–0.06	–0.08
充值次数	0.49	–0.33	1.00	–0.35	0.23	0.22	–0.13	0.14	0.13
入网时长	–0.08	0.33	–0.35	1.00	–0.21	–0.03	0.01	0.02	0.01
余额	0.16	–0.11	0.23	–0.21	1.00	0.07	–0.04	0.03	0.04
短信回复率	0.33	–0.11	0.22	–0.03	0.07	1.00	0.06	0.18	0.09
集团客户	0.04	–0.11	–0.13	0.01	–0.04	0.06	1.00	0.10	–0.01
漫游时长	0.28	–0.06	0.14	0.02	0.03	0.18	0.10	1.00	0.05
是否省外漫游	0.28	–0.08	0.13	0.01	0.04	0.09	–0.01	0.05	1.00

由参与建模的9个变量的相关系数矩阵可以看出，变量之间基本相互独立。

3.3　模型构建及对比检验

在数据集中按照7:3的比例进行分层随机等比例抽样，划分训练样本与检验样本。训练集中有46495个负样本，465个正样本，测试集中有19926个负样本，199个正样本，正负样本的比例约为1:100。依据不平衡策略，将逻辑回归算法的分类阈值设定为0.01。分别利用基于再缩放策略的逻辑回归算法和传统的逻辑回归算法建立模型，二者在训练集和测试集的建模结果如表3所示。

表3　　模型改进前后效果对比

	改进的逻辑回归		传统的逻辑回归	
	训练集	测试集	训练集	测试集
(0,0)	38065	16390	46495	19926
(0,1)	8430	3536	0	0
(1,0)	120	45	465	199
(1,1)	345	154	0	0
查全率(%)	74.2	77.4	0	0
准确率(%)	81.5	82	99	99

由表3可以看出，传统的逻辑回归算法在不平衡数据集上，虽然准确率比较高，但没有任何现实意义，因为这样的学习器无法识别出来正样本。而改进后的逻辑回归算法，虽然整体准确率低于传统的逻辑回归算法，但其查准率却有了质的飞跃，说明基于再缩放策略的逻辑回归算法在不平衡数据集上具有一定的优越性的。

4　总结

逻辑回归算法是目前应用最广泛的分类算法之一，本文在逻辑回归的基础上，提出了一种解决共线性问题的方法，并基于再缩放策略，改善其在不平衡数据集上的适用性。实证结果表明，基于再缩放策略的逻辑回归算法在不平衡数据集上表现出了显著的优越性。

参考文献：

[1]陶然. Logistic模型多重共线性问题的诊断及改进[J]. 统计与决策, 2008,(15).

[2]张凤莲. 多元线性回归中多重共线性问题的解决办法探讨[D].广州：华南理工大学硕士论文, 2010.

[3]满敬銮, 杨薇. 基于多重共线性的处理方法[J]. 数学理论与应用, 2010,(2).

[4]郭媛媛. 基于核主成分回归的多重共线性消除问题研究[D].唐山：河北联合大学硕士论文, 2014.

[5]赵东波. 线性回归模型中多重共线性问题的研究[D].锦州：渤海大学硕士论文, 2017.

[6]王鹏. 面向不平衡数据分类问题的核逻辑回归算法的设计与实现[D].西安：西安电子科技大学硕士论文, 2015.

[7]郭华平, 董亚东, 邬长安等. 面向类不平衡的逻辑回归方法[J]. 模式识别与人工智能, 2015, 28(8).

[8]Zeng Z Q, Qun W U, Liao B S, et al. A Classfication Method for Imbalance Data Set Based on Kernel SMOTE[J]. Acta Electronica Sinica, 2009, 37(11).

[9]周志华. 机器学习[M]. 北京：清华大学出版社, 2016.

（责任编辑/浩　天）

Rescaling Strategy-Based Logistic Regression Algorithm and Its Application

Li Qiongyang

(College of Mathematics and Statistics, Xuchang University, Xuchang Henan 461000, China)

Abstracts: This paper proposes a method of filtering attributes and de-correlation to select modeling attributes, greatly reducing the degree of collinearity between attributes. By default, the logistic regression algorithm takes 0.5 as the threshold when making a category decision, but the effect of such a decision on the imbalanced data set is not satisfactory. Therefore, the paper uses the rescaling strategy to study the application of logistic regression algorithm in unbalanced data sets, and makes an empirical analysis on the sample data of consumer behavior characteristics of spam SMS provided by an operator. The results show that after using method of filtering attributes and de-correlation to select modeling variables, logistic regression learning tool based on rescaling strategy has good accuracy and universality.

Key words: filtered attribute selection; attribute correlation; imbalance data set; logical regression