DOI:10.13774/j.cnki.kjtb.2013.12.020

基于高维数据的集成逻辑回归分类算法的研究与应用

毛 林[1,2],陆全华[1,2],程 涛[1,2]
(1. 泰州市农业物联网工程技术中心,江苏 泰州 225300;
2. 江苏畜牧兽医职业技术学院,信息工程系,江苏 泰州 225300)

摘 要:针对逻辑回归分类模型,提出基于高维数据的集成逻辑回归分类算法,该算法随机抽取多个特征集,并针对各个特征集构建多个回归模型。并最终针对多个逻辑回归模型结果,利用集成学习方法进行最终预测。实验结果表明,集成逻辑回归分类算法具有很高的预测精度,与传统算法相比有明显的提高。

关键词:高维数据;集成学习;逻辑回归;分类算法;随机子空间

中图分类号:TP3 **文献标识码**:A **文章编号**:1001-7119(2013)12-0064-03

The Research and Application of Ensemble Logistic Regression Classification Algorithm Based on High Dimensional Aata

Mao Lin[1,2],*Lu Quanhua*[1,2],*Cheng Tao*[1,2]
(1. Agricultural IOT Engineering Technology Center of Taizhou, Taizhou 225300, China;
2. Jiangsu Animal Husbandry & Veterinary College, Taizhou 225300, China)

Abstract: In this paper, focusing on logistic regression model, we propose ensemble logistic regression classification model based on high dimensional data. We selected features from the whole features set, and we predict the classification results with ensemble learning method. The experimental results show that the prediction accuracy of ensemble logistic regression classification algorithm is high, and has obvious promotion comparing with traditional algorithm.

Key words: high dimensional data; ensemble learning; logistic regression; classification algorithm,random subspace

随着信息化社会的高速发展,信息多元化成为主要发展模式。人们使用更多的特征属性描述数据信息,通常某一数据记录使用成千上万的特征描述。但是,在数据挖掘领域中,高维的数据信息记录中往往存在着不相关或冗余的特征属性,影响模型预测的准确性。例如,在构建逻辑回归模型中,如果加入某些对该模型无意义或者不相关的属性,那么,构建的模型将影响最终的预测结果。本文针对为高维数据构建逻辑回归模型[1]问题,提出了基于集成学习方法[2]的逻辑回归分类算法。通过实验结果表明,该算法可以有效地提高逻辑回归模型的预测精度。

1 逻辑回归模型

Logistic regression(逻辑回归)[3]是比较常用的机器学习方法,用于估计某种事物的可能性。比如某用户购买某商品的可能性,某病人患有某种疾病的可能性[4],以及某广告被用户点击的可能性等。逻辑回归延伸了多元线性回归思想,即因变量是二值的情形,自变量为$x_1,x_2,x_3,\cdots,x_k$。逻辑回归是用来测量分类结果与因变量之间的关系。逻辑回归模型的最终结果为0,1分类结果。其中1表示属于该类,0表示不属于该类别。逻辑回

收稿日期:2012-12-24
基金项目:江苏省农委重大攻关项目(2130109)。
作者简介:毛林(1968-),男,江苏靖江人,硕士,讲师,研究方向:物联网应用与系统集成。

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

表 1 实验数据描述

Table 1 Experimental Data Description

数据名称	样本个数	特征属性个数	属性类型	分类个数
S1	98,000	125,000	实数	2
S2	89,796	255,000	实数	2
S3	78,698	234,789	实数	2
S4	89,657	256,980	实数	2
S5	90,879	231,000	实数	2
S6	89,076	234,567	实数	2

表2 第一组实验预测精度结果比较

Table 2 The first experimental prediction accuracy results comparison

数据名称	ES_LR 算法预测精度	Trad_LR 预测精度
D1	0.768	0.569
D2	0.689	0.675
D3	0.879	0.769
D4	0.745	0.721
D5	0.578	0.545
D6	0.676	0.654

表3 第二组实验预测精度结果比较

Table 3 The second experimental prediction accuracy results comparison

数据名称	ES_LR 算法预测精度	Trad_LR 预测精度
D1	0.787	0.569
D2	0.712	0.675
D3	0.889	0.769
D4	0.785	0.721
D5	0.598	0.545
D6	0.686	0.654

归方程表示如下：

$$f(x,\beta)=\frac{e^{\beta_0\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k}}{1+e^{\beta_0\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k}}$$

其中，$x_1,x_2,x_3,\cdots,x_k$表示k个特征属性值，$\beta_1,\beta_2,\beta_3,\cdots,\beta_k$表示特征属性参数值。逻辑回归方程表示某一样本属于该分类的概率值。

构建逻辑回归模型的主要目的是求解逻辑回归方程中的参数$\beta_0,\beta_1,\beta_2,\beta_3,\cdots,\beta_k$。下面，概要的介绍求解逻辑回归方程参数的方法。

首先，知道一个样本属于正类的概率值为$f(x,\beta)$，那么，该样本属于负类的概率值为$1-f(x,\beta)$。所以，在样本和概率已知的条件下，该样本属于正类的概率可以表示为：

$$P(y|x,\beta)=\begin{cases}f(x,\beta) & if\ y=1\\ 1-f(x,\beta) & if\ y=0\end{cases}$$

从上面的表达式可以推断出在参数为β逻辑回归模型已知的条件下，训练集X发生的可能性以及对数可能性可以表示为：

$$L(X,y,\beta)=\prod_{i=1}^{N} f(x_i,\beta)^{y_i}(1-f(x_i,\beta))^{1-y_i}$$

$$\ln L(X,y,\beta)=\sum_{i=1}^{N}(y_i\ln(f(x_i,\beta))+(1-y)\ln(1-f(x_i,\beta)))$$

其中，x_i表示训练集X的第i个样本，y_i表示第i个样本的分类结果，N表示样本的数目。当训练样本以及样本的分类结果已知的条件下，可以通过计算最大对数可能性[5]$\ln L(X,y,\beta)$求解逻辑回归模型参数。

本文使用Newton-Raphson算法求解$\ln L(X,y,\beta)$的最大值，Newton-Raphson更新公式可以写为如下：

$$\beta^{update}=\beta+(X^TDX)^{-1}X^T(y-f(X,\beta))$$

其中，D表示对角矩阵，$d_{ii}=f(x_i,\beta)(1-f(x_i,\beta))$。

2 集成逻辑回归算法

针对高维数据问题，本文提出了基于随机子空间的集成逻辑回归算法。下面，详细地介绍集成逻辑回归算法的主要步骤：

首先，使用随机抽样方法从特征集合（$F=\{f_1,f_2,\cdots,f_m\}$）中抽取多个随机子空间特征集$S=\{F_1,F_2,..,F_k\}$。其中，每个特征子集都是由部分特征属性组成的。设置一个条件判断，满足特征集中不存在相同的特征子集。对于给定特征空间即M维特征集，随机子空间方法选择W个特征集，那么存在C_M^W个可能的子空间。因为K远远小于C_M^W，所以，使用该便捷的随机方法探求随机性。

第二步，完成随机子空间特征集选择后，针对不同的特征子空间分别构建多个逻辑回归模型，$L_1,L_2,L_3,\cdots,L_k$。

第三步，得到多个逻辑回归模型后，针对某一预测样本，会产生k个预测结果值，即$P_1,P_2,P_3,\cdots,P_k$。

第四步，得到的k个分类结果后，使用投票方法，选择分类结果中最多的那个作为最终的分类结果。

3 实验

实验部分本文设置了多组实验，通过改变训练模型的个数，以及随机子空间的大小（即每个子空间抽取特征的个数）测试集成逻辑回归算法的预测精度。本文从网站中获取了6组高维数据，表1描述了6组高维数据的信息，通过本文提出的基于随机子空间的集成逻辑回归算法（ES_LR）与传统的逻辑回归算法(Trad_LR)进行对比，传统的逻辑回归算法即运用所有的特征属性构建逻辑回归模型。下面详细地描述和分析实验结果。

第一组实验，设置随机子空间的大小为10，训练的逻辑回归模型个数为10，本文提出算法的预测精度与传统逻辑回归模型的预测精度比较实验结果如表2所

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

表4 第三组实验预测精度结果比较

Table 4 The third experimental prediction accuracy results comparison

数据名称	ES_LR(20)预测精度	ES_LR(30)预测精度	ES_LR(40)预测精度	Trad_LR 预测精度
D1	0.802	0.812	0.812	0.569
D2	0.742	0.756	0.775	0.675
D3	0.901	0.913	0.912	0.769
D4	0.798	0.812	0.814	0.721
D5	0.612	0.645	0.675	0.545
D6	0.693	0.714	0.754	0.654

示。

通过第一组实验的结果可以看出，本文提出的算法在不同的数据集下完成逻辑回归模型构建并预测的精度与传统算法的预测精度有所提高，但是，针对某些数据集的预测精度提高得并不是很明显。

第二组实验，设置随机子空间的大小为20，训练的逻辑回归模型个数为10，本文提出算法的预测精度与传统逻辑回归模型的预测精度比较实验结果如表3所示。

通过第二实验的结果可以看出，本文提出的算法在不同的数据集下完成逻辑回归模型构建并预测的精度与传统算法的预测精度有进一步地提高，同时，通过改变子空间大小，即增大子空间中特征属性的个数，ES_LR算法比第一组实验有了进一步地提高。

第三组实验中，改变子空间的大小，固定训练模型的个数为10，分别设置子空间的大小为30，40，50。实验结果如表4所示。

结合第三组实验与第二组实验结果，可以看出，随着子空间大小的增大，集成子空间逻辑回归算法的预测精度逐渐增大，即 随着子空间的增大，模型预测精度增大。但是，当子空间增大到一定个数是，预测精度维持在一定范围内，不能无限增大。总体来说，本文提出的算法与传统算法相比，在预测精度上有着明显的提高。

总之，子空间集成逻辑回归算法与传统的算法相比在预测精度上有明显提高。同时，通过大量的实验可以看出，本文提出的子空间逻辑回归算法可以通过改变子空间的大小和训练模型的个数，即增大子空间的大小和增多训练模型的个数，增大算法的最终预测精度。

4 总结

高维数据的发展趋势给逻辑回归模型的构建带了巨大的挑战，由于在高维属性中存在不相关或者冗余的特征属性，将严重影响模型的预测精度。本文针对逻辑回归模型的构建问题，提出了基于随机抽象构建多个逻辑回归模型，并进一步利用集成学习方法进行模型预测的算法，本文提出的子空间逻辑回归算法可以通过改变子空间的大小和训练模型的个数，即增大子空间的大小和增多训练模型的个数，增大算法的最终预测精度。总体来说，该算法可以有效地提高模型的预测精度。

参考文献：

[1] 王国强, 赵克勤, 郑选军. 天气预报多元回归模型中模糊因子的集对分析[J].科技通报,2004,20 (2): 151–155.

[2] Menard, Scott.Applied logistic regression nalysis. Vol. 106 [M]. Sage Publications, Incorporated, 2001.

[3] Breiman,Leo.Random forests[J].Machine earning,2001,45 (1): 5–32.

[4] Liu, Yong, and Xin Yao. Ensemble learning via negative correlation[J].Neural Networks,1999,12 (10): 1399–1404.

[5] 于玲, 吴铁军.集成学习: Boosting 算法综述[J].模式识别与人工智能,2004,17 (1): 52–59.

(上接第 63 页)

结果表明，该算法能够大幅度提高了数据库查询效率，在数据库处理中有着广泛的应用前景。

参考文献：

[1] 韩梅.数据库管理系统查询优化技术研究[D].郑州 :中国人民解放军信息工程大学电子技术学院,2004:31–36.

[2] 曹阳,方强,王国仁,等. 基于遗传算法的多连接表达式并行查询优化[J].软件学报,2002.13(2):250–25.

[3] 许峰,杨敏,王志坚. 基于遗传算法的多数据源连接查询优化方法[J].计算机工程与应用,2006,13:145–148.

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net