基于云平台的逻辑回归模型构建算法的设计与实现

俞庆生
(佛山职业技术学院,广东 佛山 528137)

摘 要:逻辑回归模型作为分类算法已经被广泛应用到许多领域。近年来随着信息领域的高速发展,海量规模数据成为发展的主要趋势。传统构建逻辑回归模型的算法在大规模数据下不能有效地构建逻辑回归模型。针对海量数据,本文提出了高效的分布式逻辑回归模型构建算法。该算法是基于云计算平台,能够快速、高效地完成逻辑回归分类模型的构建。实验表明,本文提出的算法具有很好的加速比以及可扩展性。

关键词:逻辑回归;云计算;分类;可扩展性

中图分类号:TP3 文献标识码:A 文章编号:1001-7119(2013)06-0137-03

The Design and Implementation of Cloud Platform Based Building Algorithm for Logistic Regression Model

Yu Qingsheng
(Foshan Polytechnic, Guangdong 528137, China)

Abstract: Logistic regression model is applied into many areas as a classification algorithm. Recently, with the highly development of information field, high scale data is becoming the main development trend. Traditional building algorithms for logistic regression models could not build logistic regression models for huge scale data effectively. In this paper, focusing on huge scale data, we propose a high efficient distributed building algorithm for logistic regression models, and it is based on cloud computing platform, and it could build logistic regression classification models fast and efficiently. The experimental results show that the proposed algorithm in this paper has good speed-up and scalability.

Key words: logistic regression; cloud computing; classification; scalability

0 引言

在数据挖掘、机器学习领域,分类算法是比较常见的算法。其中,逻辑回归模型可以用来对数据进行分类分析,广泛应用在各个领域中。随着现代科学技术的进步,各个领域都产生了大量的数据样本,以供分析研究。数据规模的庞大,对传统构建逻辑回归模型[1]的算法产生了巨大的挑战[2,3]。传统算法不能高效地构建模型,甚至会不能构建模型。针对海量数据构建逻辑回归模型问题,本文提出了分布式、高效的逻辑回归模型构建算法,该算法是基于云计算平台[4],可以有效地完成模型的构建,具有很好的加速比以及可扩展性。

1 逻辑回归模型

逻辑回归模型是用于预测二进制分类结果的分类模型,被广泛应用在诸多领域[5]。下面详细介绍逻辑回归模型的定义及解法。

假设训练集为T,对于数据集T中的每个记录t_i,分类结果是1或者0。其中,1代表属于正类,0代表不属于某类,也就是说属于负类。本文希望创建一个逻辑回归模型,该模型可以用来预测一个实验记录属于正类或者负类,即一个实验结果或者属于正集或者属于负集。

收稿日期:2012-09-07
基金项目:广东省高新技术产业化攻关项目(00595750177068012)。
作者简介:俞庆生(1956-),男,武汉人,硕士,副教授,主要研究方向:云计算,群体智能,工业控制。

表 1 实验数据信息

Table 1 Experimental data information

数据集	记录条数	属性个数
S1	100,000	20,000
S2	500,000	20,000
S3	1,000,000	40,000
S4	2,000,000	40,000

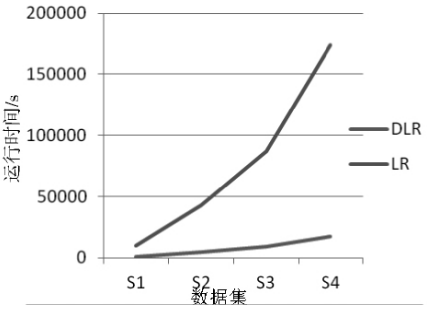

图1 算法运行时间对比图

Fig. 1 The running time comparison figure of algorithms

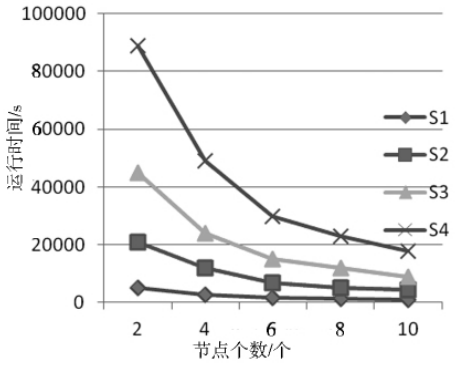

图2 可扩展测试实验结果

Fig.2 Scalability test experimental results

应用逻辑回归模型进行预测主要分为两个步骤[6]，第一步是运用训练集构建逻辑回归模型，第二步是针对测试集对分类结果进行预测。逻辑回归方程如下：

$$f(x,\beta)=\frac{e^{\beta_0+\beta_1x_1+\cdots+\beta_kx_k}}{1+e^{\beta_0+\beta_1x_1+\cdots+\beta_kx_k}}$$

其中，$\beta=(\beta_0,\beta_1,\beta_2,\cdots,\beta_k)^T$ 是逻辑回归模型的参数。x_i是记录x的第i个特征值，其中x是数据集向量 $(x_1,x_2,x_3,\cdots,x_k)$（k是x中含有特征属性的个数）。构建逻辑回归模型的目的是计算出逻辑回归方程中的参数。假设一个记录属于正类的概率是$f(x,\beta)$，那么属于负类的概率就是$1-f(x,\beta)$。在x和β已知的条件下，y的概率可以写成如下等式：

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

可以将上式改写为如下表达：

$$P(y|x,\beta)=f(x,\beta)^y(1-f(x,\beta))^{1-y}$$

整个训练数据集的可能性和对数可能性表示为如下：

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

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

其中，$x^{(i)}$为数据集X的第i个记录，y_i是第i个记录的分类结果，N是训练集中的记录条数。求解逻辑回归模型参数可以通过计算$\ln P(X,y,\beta)$的最大值来求得逻辑回归模型参数，计算参数β满足下面的等式：

$$\beta=\max\sum_{i=1}^{N}(y_i\ln f(x^{(i)},\beta)+(1-y_i)\ln(1-f(x^{(i)},\beta)))$$

求解参数可以转换为极大似然估计问题，本文的求解逻辑回归参数的方法是基于梯度的求解极大似然估计的算法，基于梯度的方法是通过多次迭代直至收敛，求得参数值。基于梯度的更新等式可以写为如下：

$$\beta^{i+1}=\beta^i+\xi\frac{\partial}{\partial\beta}\ln P(X,y,\beta)\Rightarrow\beta^{i+1}=\beta^i+\xi X^T(y-f(X,\beta))$$

2 分布式逻辑回归模型构建算法

数据规模的增大，给求解逻辑回归参数的算法带来了大量的困难。由于训练集规模的增大，导致梯度求解算法中X值过大，造成内存不足，运算速度慢等问题。针对上面的难点，提出了基于云计算平台的分布式算法，现有的流行Hadoop分布式计算平台，可以提供可扩展性良好的平台，有效地完成模型的构建。Hadoop是一个能够对大量数据进行分布式处理的软件框架[7]。Hadoop 是以一种可靠、高效、可伸缩的方式进行处理的。Hadoop是一个能够让用户轻松架构和使用的分布式计算平台。下面，详细地介绍基于云计算平台的分布式逻辑回归模型构建算法的步骤：

第一步，初始化逻辑回归模型参数β，将训练数据分块分配到各个运算节点中。

第二步，在各个运算节点中，计算对应的样本的$y-f(X,\beta)$ 同时，计算计算节点中含有的样本$x(i)$中各个特征值与$y-f(X,\beta)$的乘积，即 $(x_1^0(y-f(X,\beta)),x_2^0(y-f(X,\beta)),x_3^0(y-f(X,\beta)),\cdots,x_n^0(y-f(X,\beta)))$。其中，$n$表示训练集中特征属性的个数。

第三步，各个节点计算按照第二步计算完毕后，将节点的运算结果合并，对各个节点的向量结果做和相加。最后得到的行向量求转置即为本次循环中$X^T(y-f(X,\beta))$ 的值。计算到本次循环的新参数β^{new}。

第四步，判断参数是否收敛，如果收敛，那么结束循环 如果不收敛，则重复执行第二步，第三步，其中的参数β为更新的参数β^{new}。

下面，为基于MapReduce的分布式逻辑回归模型构建算法的伪代码：

Mapper Procedure

Input: Training data X;
Output: 矩阵中间值
for each $x^{(i)}$ in X
compute $x^0(y-f(x^0,\beta))$;
store intermediate value (i, $x^0(y-f(x^0,\beta))$);
end for
Output intermediate value
Reducer Procedure:
Input: Intermediate Value
Output: $X^T(y-f(X,\beta))$
A = Sum(intermediate Value);
Output(A);
Driver Procedure:
Initiate β,ε;
Transmit Training Data, β,ε;
Start Mapper;
Start Reducer;
$\beta^{new}=\beta+\xi X^T(y-f(X,\beta))$
if $|\beta^{new}-\beta|<\Theta$
stop compute;
return 0;
else
Go to 3,4,5

3　实验

在实验部分，通过针对不同大规模数据集构建逻辑回归模型，验证算法的有效性。本文的实验数据集是针对大规模数据的，表1总结了实验数据的信息。

实验分两个部分，第一部分测试分布式逻辑回归模型构建算法与传统模型构建算法在执行时间上的比较。针对4组数据，分布式算法在10个运算节点上执行，查看其执行时间与传统算法之间的加速比。图1 为两个算法在四组实验数据上的执行时间。

通过图2可以看出，传统算法在执行效率上比分布式逻辑回归模型构建算法低很多，基于云平台的分布式逻辑回归算法具有很高的加速比。因此，从这部分实验可以得出：本文提出的基于云平台的分布式逻辑回归模型构建算法在执行效率上比传统算法快很多，具有明显的加速比。

第二部分实验测试分布式算法是否具有良好的可扩展性，由于数据规模的增大，本文希望设计的算法在增大数据规模的时候，仍具有很好的性能，即 :能够有效地处理大规模数据，而不是随着数据规模的增大，影响算法的执行甚至不能执行。这部分实验，通过改变数据集的大小，测试分布式算法在不同节点个数的条件下的执行时间。在图2 中展示了实验结果。

通过图2 可以看出，针对同一个数据集，随着节点的个数的增加，分布式逻辑回归模型构建算法的执行时间减少。可以看出，分布式算法即使面对规模大的数据，也可以通过增加运算节点的个数，降低运行的时间，从而有效处理大规模数据。从图2还可以看出，随着数据规模的增多，算法执行时间呈现等比例增长趋势，因而算法具有很好的可扩展性。

4　总结

逻辑回归模型作为一种分类方法被广泛应用在各个领域，随着信息化的进步，数据规模呈现指数级增长，提出有效地处理海量规模的模型构建算法成为当今社会的发展趋势。本文提出了基于云计算平台的分布式构建逻辑回归模型的算法。该算法是应用Hadoop分布式计算框架，与传统模型构建算法具有很好的加速比，同时，具有很好的可扩展性。

参考文献:

[1] 农秀丽, 彭展声.非齐次等式约束线性回归模型回归系数的综合条件岭估计[J].科技通报，2012,28 (2): 4-6.

[2] P Komarek and A Moore. Making logistic regression a core data mining tool with tr-irls [C]//.Proceedings of the 5th International Conference on Data Mining Machine Learning, 2005 :4.

[3] Lin C, Weng R, Keerthi S.Trust region newton methods for large-scale logistic regression[C]//.Proceedings of the 24th international conference on Machine learning, ACM,2007: 561 - 568.

[4] 苏汉宸，李红燕，苗高杉，等. PTLR :云计算平台上处理大规模移动数据的置信域逻辑回归算法[J].计算机研究与发展，2010 :414-419.

[5] Lin, C., Weng, R., Keerthi, S.: Trust region newton methods for large-scale logistic regression [C]//.Proceedings of the 24th international conference on Machine learning, ACM,2007: 561 - 568.

[6] Jun Liu, Jianhui Chen, Jieping Ye.Large-scale sparse logistic regression [C]//. Proc of ACM SIGKDD 2009.New York:ACM,2009:547-556.

[7] Tom White. Hadoop: The Definitive Guide [M]. O'Reilly Media, Inc. 2005.