文章编号:1006-9348(2016)03-0292-04

云计算环境下大数据合理分流技术研究与仿真

王 欣[1],周晓梅[2]

(1. 南京工业大学浦江学院,江苏 南京 211134;2. 中国传媒大学南广学院,江苏 南京 211172)

摘要:对云计算环境下的大数据进行准确分流能够提高云计算的服务质量。传统的数据分流方法无法避免云计算环境下大数据复杂性和高动态变化性带来的影响,降低了数据分流的准确率。提出一种改进K均值聚类算法的数据分流方法。对数据进行特征提取,在此过程中通过降维处理加快了特征提取的速度;利用K均值算法进行数据特征聚类,在特征聚类的过程中不断调整数据特征的聚类中心,最终得到准确的数据分流结果。仿真结果表明,利用改进算法能够提高云计算环境下的大数据分流的准确率,提高了数据分流效率。

关键词:云计算环境;数据分流;聚类

中图分类号:TP391.9 **文献标识码**:B

Research and Simulation on Big Data Reasonable Splitting Technology in Cloud Computing Environment

WANG Xin[1], ZHOU Xiao-mei[2]

(1. Pujiang College, Nanjing University of Technology, Nanjing Jiangsu 211134, China;
2. Nanguang College, China University of Communication, Nanjing Jiangsu 211172, China)

ABSTRACT: To accurately split the big data in cloud computing environment, can improve the quality of service of cloud computing. A data splitting method based on improved K-mean clustering algorithm was proposed in the paper. The data features were extracted, and in the process of feature extraction, the feature extraction speed was accelerated by dimension reduction processing. K-means algorithm was used for data feature clustering. In the process of feature clustering, we can adjust the cluster center of data characteristics, and finally get the accurate results of data splitting. The simulation results show that the improved algorithm can increase the accuracy of big data splitting in the cloud computing environment and the efficiency of data splitting.

KEYWORDS: Cloud computing environment; Data splitting; Clustering

1 引言

随着网络技术的迅猛发展,云计算技术应运而生[1]。云计算是对互联网络的比喻性说法,云计算的核心是互联网络中大量的计算资源、存储资源和应用软件资源通过互联网络汇聚起来[2],形成巨大的计算能力、存储能力和服务能力,用户根据需要从网络中获取响应的资源[3]。在云计算环境下,存在着规模庞大的数据,对云计算环境下大数据进行合理分流能够提高用户获得资源的效率[4],使客户感受到更好的云计算服务质量。云计算环境下的大数据中往往包含有大量的冗余数据,具有高度复杂性、动态变化性等特点,如果直接进行数据分流[5],会降低客户获取云计算资源的效率。因此,如何对云计算环境下大数据进行合理分流,已经成为云计算领域一个难点,引起了重视[6]。

业界很多学者针对云计算下大数据分流的问题,已经提出了一些数据分流方法[7]。目前,云计算环境下大数据的分流方法主要包括基于支持向量机算法的数据分流方法、基于聚类算法的数据分流方法和基于决策树算法的数据分流算法。其中最常用的是基于支持向量机算法的数据分流方法[8]。由于云计算环境下大数据的分流方法在提高云计算服务质量方面具有无可替代的作用,因此该课题拥有极为广阔的发展前景,并成为很多学者重点研究的课题[9]。

利用传统算法在进行云计算环境下大数据分流的过程中,需要对每一条数据进行分类,由于云计算环境下的数据具有高度的复杂性和动态变化性[10],使得传统的分类方法无法对其进行准确分类,降低了数据分流的准确性,降低了客户获得云计算资源的效率。

针对上述传统算法的缺陷,提出一种基于改进K均值聚

基金项目:江苏省高校自然科学研究面上项目(15KJD520005)
收稿日期:2015-03-10

类算法的数据分流方法。对数据进行特征提取,在此过程中通过降维处理加快了特征提取的速度;利用 K 均值算法进行数据特征聚类,在特征聚类的过程中不断调整数据特征的聚类中心,最终得到准确的数据分流结果。仿真实验表明了改进算法体现在数据分流方面的优势。

2 数据分流的有关原理

利用支持向量机算法能够对云计算环境下的大数据进行合理分流,其核心思想是,首先构建大数据的最优分类平面,然后将待分流的大数据从高维空间映射到低维空间,实现非线性分类问题到线性分类问题的变换,最后利用核函数对数据分流问题的线性问题进行求解,最终实现云计算环境下大数据的分流。具体方法如下所述:

设置大数据样本集合为$\{x_i,y_i\},i=1,2\cdots,n,x\in R^d,y_i\in\{1,-1\}$为数据分流的标识号,数据分流的线性判别函数能够描述为:

$$f(x)=w\cdot x+b \tag{1}$$

将数据分流问题进行归一化处理,能够得到下述公式:

$$w\cdot x+b=\pm 1 \tag{2}$$

上述的数据分流问题能够用一个存在约束条件的非线性规划问题进行表达:

$$y_i(w\cdot x_i)+b\geq 0 \tag{3}$$

对上述公式进行运算可以得到数据分流问题的 *Wolfe* 对偶函数如下所述:

$$Maximizew(\alpha)=\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i=1}^{n}\alpha_i\alpha_j y_i y_j x_i x_j \tag{4}$$

公式(4)的约束条件能够描述为:

$$\sum_{i=1}^{n}\alpha_i y_i=0,\alpha_i\geq 0,i=1,2,\cdots,n \tag{5}$$

通过 *Wolfe* 对偶函数能够使支持向量机应用到非线性的大数据分流问题中。对于非线性的大数据分流问题,需要通过核函数在高维空间内转化为带有约束条件的二次函数,其公式如下所述:

$$MaximizeQ(\alpha)=\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i=1}^{n}\alpha_i\alpha_j y_i y_j K(x_i,x_j) \tag{6}$$

这样能够获得大数据的最优分流函数:

$$g(x)=\mathrm{sgn}((fx))=\mathrm{sgn}\{\sum_{j=1}^{n}\alpha_j y_j k(x,x_i)+b\} \tag{7}$$

假设需要云计算环境下的大数据有 N,可以记为$\{S_1,S_2,\cdots,S_N\}$,大数据分流的问题实质上是一个多目标分类的问题,而支持向量机能够通过建立二分类模型实现对大数据的准确分类,因此需要将多个二分类的支持向量机进行组合构建多目标分流模型,实现大数据的多目标分类,这样可以降低运算量,同时,若要增加分流任务,只需加入一个新的两分类支持向量机。

云计算环境下的大数据分流的实现过程如下所述:

1) 采集云计算环境下待分类的样本数据;

2) 对样本数据进行归一化处理,降低数据分类的运算量;

3) 提取样本数据的特征信息,并通过支持向量机从高维空间向低维空间进行映射,将非线性分流问题转化为线性分流问题,并降低维度;

4) 将获取的样本数据特征作为支持向量机的训练样本进行训练;

5) 设置有 N 个支持向量机的数据分流器,$f_i,i=1,2,\cdots N$,将数据分流问题转换为两种类型,一种类型是正样本,另一种类型是负样本,数据分流器的输出分别为 +1,-1;

6) 将数据样本的训练样本集利用支持向量机进行学习,并进行相关参数的优化,获得数据分流模型;

7) 将每一条数据输入到 N 个分流器中,若 f_i 输出值为 +1,则该条数据可以被归于第 i 类;若同时有多个分流器的输出值都为 +1,则需要利用欧氏距离法计算该条数据与各个分流器之间的距离,然后将其划分为与之距离值最小的分流器之中。假设全部的数据分流器的输出值都为 -1,则可认为本次分流错误,分流结束。

根据上面阐述的方法,构建大数据的最优分类平面,利用支持向量机将非线性分流问题变换为线性分流问题,最后利用核函数对数据分流问题的线性问题进行求解,最终实现云计算环境下大数据的分流。

3 基于协作分流算法的大数据分流方法原理

传统的大数据分流方法需要对每一条数据进行分类,无法适应云计算环境下的数据具有高度的复杂性和动态变化性的特点,造成数据分流的准确性降低。为此,提出一种基于改进 K 均值聚类算法的数据分流方法。

3.1 提取数据特征

对数据进行准确分流的过程中,关键的步骤是提取数据的特征,并将数据特征构成特征集合。设置数据的特征数目为 n,数据的数目为 p。利用这两个参数能够构建数据的特征矩阵,其形式为 $C_{n\times p}=\{c_{jk}\}_{n\times p}$。数据特征矩阵中的每一个元素都能够描述数据的一个特征参数,每一列的元素都能够描述数据特征的分量,则有:

$$C_{n\times p}=\begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1p} \\ c_{21} & c_{22} & \cdots & c_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ c_{n1} & c_{n1} & \cdots & c_{np} \end{bmatrix} \tag{8}$$

上述数据特征矩阵具有较高的维度,这就使得数据特征提取的过程中具有复杂的运算过程,导致数据特征提取的准确性降低,因此,需要对其进行降维处理,其公式如下所述:

$$C_{n\times p}=W_{n\times s}\cdot U_{s\times s}\cdot X^T_{s\times p} \tag{9}$$

其中,n 为数据的全部特征数目,p 为一个数据的特征分量。上述降维处理的过程需要满足下述条件:

$$T=diag(\varepsilon_1,\varepsilon_2,\cdots,\varepsilon_s),s\leq\min(n,p) \tag{10}$$

其中,ε_j 是数据特征提取过程中的与数据有关的系数。

数据特征矩阵能够利用下述公式进行简化处理:

$$C_l = \sum_{j=1}^{l} w_j \varepsilon_j \tag{11}$$

若 $l \to rank(B)$，则通过运算能够得到 C_l 的值，即数据的特征矩阵。

根据上面阐述的方法，对云计算环境下的数据进行降维处理，提取数据的特征，为数据的准确分流提供了准确的数据基础。

3.2 数据分流的实现

K 均值聚类算法是一种有效的数据分流方法，利用它能够对云计算环境下的大数据进行准确分流。具体方法是，对数据特征进行准确分类，将数据分为不同的类别，实现了云计算环境下大数据的准确分流。这种数据分流方法，具有简单可靠、效率高、准确性高的特点，因此在数据分流方面得到广泛的应用。

利用 K 均值聚类算法进行云计算环境下的大数据分流，需要首先得到数据的初始聚类中心，在数据分流的过程中对聚类中心进行更新，适应云计算环境下数据具有高度的动态变化的特点。具体的数据分流方法如下所述：

设置云计算环境下数据特征构成的聚类中心的数据为 l，数据特征的数目为 p，在 p 个数据特征中选取 l 个特征作为初始的聚类中心，每个聚类中心代表一类数据。通过运算能够得到其它 $p-l$ 个数据特征到初始聚类中心的距离，并将这些数据特征分配到最近的聚类中心中，这样就实现了全部数据特征的分配。

利用下述公式能够将数据特征划分为 L 个不同的类别：

$$F = \sum_{j=1}^{l} \sum_{y \in D_j} |y - \overline{y_j}|^2 \tag{12}$$

其中，y 为数据特征，$\overline{y_j}$ 为待分流数据特征集合 D_j 均值化处理的结果，F 为数据特征误差的方差之和。通常情况下，F 的值越小，表明数据特征聚类的效果越好，这也表明数据分流的效果越理想。

根据上述方法，能够将云计算环境下的数据特征划分为 L 个不同类别，$T_j(j = 1,2,\cdots,l)$，聚类中心 D_j 能够描述数据特征构成的集合 T_j，数据特征集合能够用 $T = \{Y\}$ 描述。设置任意两个数据特征 Y 和 Z，则这两个数据特征之间的欧氏距离为 $e(Y,Z)$。

通过迭代处理，能够对云计算环境下的大数据进行准确分流，具体的分流过程如下所述：

1）设置云计算环境下的数据的初始聚类中心为 $TD_q = \{D_j\}$，对数据特征进行聚类处理，能够将其划分为 l 个聚类中心，该运算过程需要满足下述条件：

$$\begin{aligned} &k = 1,2,\cdots,l \\ &T_k = \{Y \mid e(Y,D_k) \leq e(Y,D_j), j \neq k\} \end{aligned} \tag{13}$$

2）对数据特征进行迭代处理获得新的数据特征集合 TD_{q+1}；

3）若 $q = 0$，则数据特征聚类中心为 TD_0；

4）对数据特征聚类中心进行迭代运算，能够得到新的数据特征聚类中心 TD_{q+1}；

5）通过运算获得数据分流的误差方差，若数据分流的误差方差的值足够小，则结束数据分流，此时获得准确的数据分流结果；否则，$q+1 -> q$，并返回到过程(2)，继续进行数据分流。

根据上面阐述的方法，提取云计算环境下的数据特征，构成数据特征集合，利用 K 均值聚类算法进行聚类，在迭代处理的过程中不断更新聚类中心，很好的适应了云计算环境下数据动态变化的特点，获得准确的数据分流结果。

4 算法实验结果比较与分析

为了验证改进算法在云计算环境下大数据分流方面的有效性，需要进行一次仿真实验。

4.1 实验环境设置

利用仿真软件编写了云计算环境下的大数据分流的仿真平台，利用传统算法与改进算法进行数据分流方面的性能比较。数据分流仿真实验用的计算机的硬件配置为，*CPU* 为 *AMD Phenom II X4* 910，内存为 *4G GGR3* 1333，硬盘为 *500Gb*，操作系统为 *Window* 7，传统的分流算法采用的是支持向量机的分流方法。在数据分流仿真实验的过程中，忽略了数据传输过程中可能存在的丢包错误。为了证明改进算法在数据分流方面的有效性，仿真实验首先对用户获取云计算资源的网络延迟进行比较，仿真实验的参数设置能够用下表 1 进行描述：

表 1 实验参数设置

参数	设置值
网络半径	*100KM*
数据传输距离	*200KM*
数据特征数目	37 个
数据的数量	10 万条
数据传输速率	*100bit/s*
数据类型	5 种

待分流的 5 种数据类型分别为政治、经济、文化、体育和宗教，数据的训练样本和测试样本能够用下表进行描述：

表 2 数据分流实验中的样本数据

类别	训练样本	测试样本
政治	1 万	0.5 万
经济	2 万	0.3 万
文化	1.5 万	0.2 万
体育	2 万	1 万
宗教	1 万	0.5 万

数据分流的准确性能够衡量不同算法在数据分流方面的性能，其公式如下所述：

$$t = \frac{y_t}{y_t + y_f} \tag{14}$$

其中，y_t 为数据分流的准确性，y_f 为数据分流的错误性。

4.2 实验结果比较及分析

利用传统算法和改进算法进行云计算环境下的大数据分流实验，不同算法的数据分流准确率能够用下表 3 进行描述：

表 3 不同算法数据分流准确率

类别	传统算法	改进算法
政治	90.5%	98.7%
经济	89.8%	99.6%
文化	92.4%	97.5%
体育	87.4%	98.3%
宗教	92.1%	99.2%

根据上表 3 中的实验结果能够得知，改进算法的数据分流的准确率远远高于传统的支持向量机数据分流算法，主要原因是，传统的数据分流方法没有充分考虑云计算环境下的数据具有高度的复杂性和高动态变化性的特点，导致数据特征提取的准确性较差，降低了数据分流的准确率。而改进算法能够对数据特征进行准确提取，并且在数据特征聚类的过程中，能够不断调整数据特征的聚类中心，很好的适应了云计算环境下的数据复杂性和高动态变化性的特定，最终获得准确的数据分流效果，这充分体现出改进算法在数据分流方面的优势。

在上述实验过程中，不同算法进行数据分流过程中的数据延迟时间能够用下表 4 进行描述：

表 4 不同算法的数据延迟时间比较

算法	数据分流时间
传统算法	32ms
改进算法	17ms

根据上表 4 中的实验数据能够得知，改进算法的平均数据延迟时间比传统算法低 15ms，这是因为传统算法在数据特征提取的过程中没有降低数据的维数，导致大量冗余特征数据被分类，延长了数据特征提取的时间，降低了数据分流的效率，最终延迟了数据发送的时间。而改进算法由于能在数据特征提取的过程中有效降低数据特征的维数，删除了大量冗余的特征信息，提高了数据特征提取的效率，加快了数据分流的速度，这也提高了数据分流的准确性。因此，与传统算法相比，改进算法在数据分流的过程中，能够极大的提高数据分流的准确性和效率，在云计算环境下的大数据分流方面相对传统算法具有明显的优越性。

5 结束语

针对传统算法的缺陷，提出一种基于改进 K 均值聚类算法的数据分流方法。对数据进行特征提取，利用 K 均值算法进行数据特征聚类，在特征聚类的过程中不断调整数据特征的聚类中心，获得准确的数据分流效果。仿真实验表明，在云计算环境下大数据分流方面，改进算法相对传统算法具有明显的优势。取得了令人满意的效果。

参考文献：

[1] 姚宏，白长敏，胡成玉，曾德泽，梁庆中. 移动数据分流研究综述[J]. 计算机科学，2014，41(B11)：182-186.

[2] 胡晓敏. 无线传感器网络 Agent 数据分流策略[J]. 新型工业化，2013，(4)：2-6.

[3] 程思霖. 基于四网协同下的 WLAN 发展策略及数据分流研究[J]. 电信工程技术与标准化，2013，(10)：20-23.

[4] 王强，杨宏. WLAN 接入 GPRS 实现数据业务分流方案研究[J]. 邮电设计技术，2013，(4)：38-41.

[5] 尹粤宁. 基于宽带远程接入服务器的成分流量分析[J]. 中国新通信，2015，17(4)：64-65.

[6] 成国营，王艳. 无线传感器网络中多移动 Agent 协同控制数据分流方法[J]. 计算机应用，2015，35(4)：910-915.

[7] 周宏旭，陈涛. 移动大数据应用助力实现 G/T 双网分流[J]. 通信世界，2014，(32)：40-41.

[8] 席兵，韩盈盈. GPRS 网络 Gb/Gn 接口数据过滤与分流的实现[J]. 广东通信技术，2014，34(7)：11-14.

[9] 刘龙庚，罗光春. 大数据通信中带宽优化技术仿真[J]. 计算机仿真，2014，31(9)：225-228.

[10] 李亚，刘伟. 基于多点映射分解的网络突变流量分解仿真[J]. 计算机仿真，2013，30(9)：298-301.

[作者简介]

王 欣(1985-)，男(汉族)，江苏南京人，硕士，讲师，主要研究方向：云计算与大数据、人工智能、高等教育；

周晓梅(1980-)，女(汉族)，江苏南京人，硕士，主要研究方向：数据库。