(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 107818347 A
(43)申请公布日 2018.03.20

(21)申请号 201711095943.7

(22)申请日 2017.11.08

(71)申请人 千寻位置网络有限公司
地址 200433 上海市杨浦区军工路1436号64幢一层J165室

(72)发明人 万景琨

(74)专利代理机构 上海市海华永泰律师事务所 31302
代理人 包文超

(51)Int.Cl.
G06K 9/62(2006.01)

权利要求书1页 说明书5页 附图1页

(54)发明名称

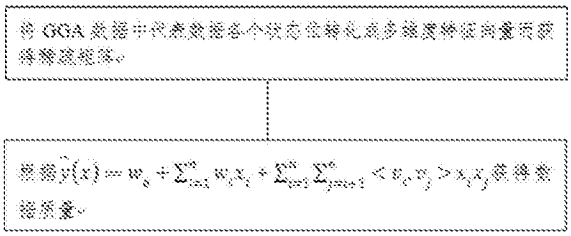
GGA数据质量的评定预测方法

(57)摘要

本发明公开基于GGA数据质量的快速评定预测方法。该方法包括如下步骤:步骤一:将GGA数据中代表数据各个状态位转化成多维度特征向量而获得稀疏矩阵;步骤二:根据

$$\tilde{y}(x)=w_0+\sum_{i=1}^{n}w_i x_i+\sum_{i=1}^{n}\sum_{j=i+1}^{n}\langle v_i,v_j\rangle x_i x_j$$

获得数据质量,其中,$\tilde{y}(x)$代表定位准确度,x_i为稀疏矩阵的中任意GGA数据在维度i特征值,x_j为稀疏矩阵的中任意GGA数据在维度j特征值,w_0和w_i表示权重因子,$\langle v_i,v_j\rangle$是因子之间相互影响程度。该方法能提高预测准确度、降低存储空间和提高了运行效率。



CN 107818347 A

1.一种GGA数据质量的评定预测方法，其特征是：该方法包括如下步骤：

步骤一：将GGA数据中代表数据各个状态位转化成多维度特征向量而获得稀疏矩阵；

步骤二：根据$\hat{y}(x) := w_0 + \sum_{i=1}^{n} w_i x_i + \sum_{i=1}^{n} \sum_{j=i+1}^{n} \langle v_i, v_j \rangle x_i x_j$获得数据质量，其中，$\hat{y}(x)$代表定位准确度，$x_i$为稀疏矩阵的中任意GGA数据在维度i特征值，$x_j$为稀疏矩阵的中任意GGA数据在维度j特征值，$w_0$和$w_i$表示权重因子，$\langle v_i, v_j \rangle$是因子之间相互影响程度。

2.如权利要求1所述的对GGA数据质量的评定预测方法，其特征是：采用采用PCA矩阵分解技术对$\langle v_i, v_j \rangle$通过降低维度进行估计。

3.如权利要求1或3所述的对GGA数据质量的评定预测方法，其特征是：所述w_0、w_i和$\langle v_i, v_j \rangle$通过如下方式求得：

定义误差函数$loss(y, \hat{y}) = (y - \hat{y})^2$；

对误差函数中w_0、w_i和M求导数可以得到损失函数最小极值，则有：

$$\frac{\partial L}{\partial \theta} := \begin{cases} 1 & \theta = 0 \\ x_l & \theta = w_l, l = 1,2 \ldots n \\ x_l \sum_{s=1 \,\&\&\, s \neq l}^{n} m_{sj} x_s & \theta = m_{lj}, l = 1,2 \ldots n, j = 1,2 \ldots k \end{cases};$$

对于给定的x_i矩阵和选取的内部因子估算长度k，采用梯地下降法找出损失函数局部最小值时的w_0、w_i和M的取值，每次循环迭代得到一组w_0、w_i和M，通过该值计算所述误差函数的误差值，两次误差值小于等于正实数a时所取得的w_0、w_i和M就是所得到的最优值。

4.如权利要求4所述的GGA数据质量的评定预测方法，其特征是：对于已有的GGA数据打标后的训练集，默认w_0和w_i初始值都为0，M取服从标准正态分布的随机取值矩阵。

5.如权利要求4所述的GGA数据质量的评定预测方法，其特征是：所述a取值在[0.0001，0.01]之间。

GGA数据质量的评定预测方法

技术领域

[0001] 本发明涉及软件开发领域，尤其涉及对GGA数据质量的评定预测方法。

背景技术

[0002] 随着大数据时代的到来，如何从海量的数据中对影响定位准确的因素进行预测是当前的一个研究热点。很多公司经常采用的逻辑和线性回归作为一般性的预测和分类的方法。该方法不需要考虑被预测项目的内容就能够为目标用户提供新的预测内容，因此，在电子商务和社交网络等互联网应用中尤为常见。但是，随着数据规模的不断增大，大数据彰显出的数据量大、数据多样性、信息量增长速度过快、维度增多造成稀疏矩阵现象严重，数据质量参差不齐等特征导致用户-项目评分数据的维度急剧增高，并且，用户进行评分的项目很少，传统的回归预测技术面临着数据稀疏、算法计算复杂度较高、预测准确率低的问题。

发明内容

[0003] 本发明解决的问题是现有数据预测计算复杂度高、准确度低的问题。

[0004] 为解决上述问题，本发明在位置定位领域，差分账户数据的挖掘并不多见。本发明结合相似性查询算法的有意义的研究内容和位置数据的关联特点性，摒弃传统只在数据维度较低时性能良好基于回归的划分预测算法，而采用专门解决稀疏矩阵场景和内在因素叠加的FM算法。基于该思路，本发明提供一种GGA数据质量的快速评定预测方法。该方法包括如下步骤：步骤一：将GGA数据中代表数据各个状态位转化成多维度特征向量而获得稀疏矩阵；步骤二：根据 $\widehat{y}(x) := w_0 + \sum_{i=1}^{n} w_i x_i + \sum_{i=1}^{n} \sum_{j=i+1}^{n} \langle v_i, v_j \rangle x_i x_j$ 获得数据质量，其中，$\widehat{y}(x)$ 代表定位准确度，x_i 为稀疏矩阵的中任意GGA数据在维度i特征值，x_j 为稀疏矩阵的中任意GGA数据在维度j特征值，w_0 和 w_i 表示权重因子，$\langle v_i, v_j \rangle$ 是因子之间相互影响程度。

[0005] 与现有技术相比，本发明至少具有以下优点：

[0006] (1)本方法的降低了存储空间，传统的矩阵降维远远小于输入海量信息的数量和其特征维度的数量，进一步提高空间复用率和节省实现和存储的空间复杂度。

[0007] (2)本发明提高了运行效率，通过上述定义所描述，方法可采用并行计算，同时支持横向扩展，时间复杂度可控并且不随输入信息的暴增而无序增长。

[0008] (3)本发明提高预测的准确度，通过对特征向量的内部因子的相互影响建模来解决传统机器学习算法不能解决的内部因子相互影响的问题，提高准确度和特征向量选取时候尽量相互独立的要求限制。

附图说明

[0009] 图1是本发明GGA数据质量的评定预测方法的流程图。

具体实施方式

[0010] 为详细说明本发明的技术内容、构造特征、所达成目的及功效，下面将结合实施例并配合附图予以详细说明。

[0011] 请参阅图1，本发明GGA数据质量的评定预测方法包括如下步骤：

[0012] 步骤一：将GGA数据中代表数据各个状态位转化成多维度特征向量而获得稀疏矩阵。具体的，该步骤详述如下：将GGA数据格式抽象出影响定位本身的特征向量，根据取值转换成特征值类型。数据源表如下表所示。

[0013]

搜星能力（某个时间段搜星个数）	网络通讯情况（单位毫秒）	登录时长(单位秒)	登录次数(一天登录次数)	质量状况
3	200	2000	4	1
4	100	4000	6	3
2	140	1000	23	2
3	12	1234	34	0

[0014] 上述数据源表展示了一般GGA中信息提取加上可能影响GGA数据质量的其他特征向量列表，转化为特征值类型以后如下表：

[0015]

搜星能力=3	搜星能力=4	搜星能力=2	网络通讯情况=200	网络通讯情况=100	网络通讯情况=140	网络通讯情况=12	登录时长=2000	登录时长=4000	登录时长=1000	登录时长=1234	登录次数=4	登录次数=6	登录次数=23	登录次数=34	质量状况
1	0	0	1	0	0	0	1	0	0	0	1	0	0	0	1

[0016]

0	1	0	0	1	0	0	0	1	0	0	0	1	0	0	3
0	0	1	0	0	1	0	0	0	1	0	0	0	1	0	2
1	0	0	0	0	0	1	0	0	0	1	0	0	0	1	0

[0017] 上述表格中，每个特征值和特征向量组成一个新的特征维度，对于任意一行源数据，符合其中一个组合特征维度的标记为1，否则为0，以此类推将源数据横向展开。展开以后建立的稀疏矩阵表示任意GGA在各个维度的标记状况。质量状况是人为根据已有的训练

数据打标所得的结论，不属于特征向量范围，其通过枚举数字0-2来判定GGA数据状况，这里约定0表示数据状况最好，1情况良好，2情况一般，3最差，以此类推

[0018] 步骤二：对数据质量的状况可以根据因子分解机原理，将其与n个特征维度的关系记：

[0019] $$\hat{y}(x) := w_0 + \sum_{i=1}^{n} w_i x_i + \sum_{i=1}^{n} \sum_{j=i+1}^{n} \langle v_i, v_j \rangle x_i x_j$$

[0020] 而获得数据质量，其中，式子左边$\hat{y}(x)$代表定位准确度，x_i可以理解为步骤1中得到的稀疏矩阵的中任意GGA数据在维度i的特征值，x_j为稀疏矩阵的中任意GGA数据在维度j的特征值。w_0，w_i可以理解为权重因子，权重因子决定各个由第一步提取的各个特征向量对GGA数据质量的影响度。其中w_0代表初始因子，模型中假设即使没有任何明确的特征向量来影响GGA的数据质量，$\langle v_i, v_j \rangle$是因子之间相互影响程度，由于在特征向量的取舍之间不可能保证特征特征向量的选取相互独立，即特征向量之间是很有可能相互影响，其复杂的因子相互影响模型会造成GGA本身质量的预测和判定准确度，所以通过引入$\langle v_i, v_j \rangle$向量表示任意两个向量i，j之间的相互影响度，考虑到特征向量的维度会非常多，则引入的$\langle v_i, v_j \rangle$将是非常巨大的矩阵，这会造成很大的计算资源消耗也不能达到实效性的要求。但我们注意到，引入$\langle v_i, v_j \rangle$向量是比较稀疏的矩阵，传统矩阵计算的方法可以采用矩阵分解求近似解的方法得到维度之间的相互影响关系。

[0021] 具体地，对于n个特征向量的相互影响度，其向量矩阵为n×n，由于考虑稀疏矩阵的计算并行性，可以采用PCA矩阵分解技术对n×n矩阵降低维度进行估计，即n×n矩阵降维为n×k的矩阵来估算分析，其中k远远小于n。假设采用PCA分析得到对$\langle v_i, v_j \rangle$的估计结果为w_{ij}。

[0022] 对每一个特征分量引入辅助向$M=(m_{i1}, m_{i2}, \cdots m_{ik})$，利用M向量对$\langle v_i, v_j \rangle$进行估计

[0023] $$M = \begin{bmatrix} m_{11} \cdots m_{1k} \\ \cdots \quad \cdots \quad \cdots \\ m_{n1} \cdots m_{nk} \end{bmatrix}_{n \times k}$$

[0024] $W_{ij}=MM^T$

[0025] 这就对应了一种矩阵的分解。对值的限定，对模型的表达能力有一定的影响。

[0026] 由于是预测模型，对于任意已经达标的GGA数据，必然有误差。误差模型定义根据最小二乘法定义：

[0027] $loss(y, \hat{y}) = (y-\hat{y})^2$

[0028] 其中，y代表已有GGA数据中打标的值，$\hat{y}$代表根据同样的GGA根据上述模型的特征值所判定的质量情况，即$\hat{y}$为模型中的$\hat{y}(x)$。将$\hat{y}(x)$代入误差模型得到误差函数：

$$loss(y,\hat{y}) = \left(y - \left(w_i + \sum_{i=1}^{n} w_i x_i + \sum_{i=1}^{n} \sum_{j=i+1}^{n} \langle v_i, v_j \rangle x_i x_j\right)\right)^2$$

[0029]
$$= \left(y - \left(w_i + \sum_{i=1}^{n} w_i x_i + \sum_{i=1}^{n} \sum_{j=i+1}^{n} M x_i x_j\right)\right)^2$$

[0030] 基于维度特征和GGA质量的关系模型，需要确定w_0、w_i和$\langle v_i, v_j \rangle$来完善预测模型，同样对误差函数中w_0、w_i和M求导数可以得到损失函数最小极值，由于y是固定已知值，对损失函数求导数转化成为对$\hat{y}(x)$求导，则有：

$$\frac{\partial L}{\partial \theta}$$

[0031]
$$= \begin{cases} 1 & \theta = 0 \\ x_i & \theta = w, i = 1,2 \ldots n \\ x_i \sum_{s=1 \& s \neq i}^{n} m_{sj} x_s & \theta = m, i = 1,2 \ldots n, j = 1,2 \ldots k \end{cases}$$

[0032] 上式分别得出$\hat{y}(x)$对w_0、w_i和M求导情况，其中k表示M中的内部影响因子维度，k一般取小于n的整数。

[0033] 3)对于给定的xi矩阵和选取的内部因子估算长度k，采用梯地下降法找出损失函数局部最小值时的w_0、w_i和M的取值，同时给定一个梯度下降的变化值a来判定最后算法的结束条件，一般a取正实数，考虑到其误差范围使用的场景不同，一般a取值在[0.0001,0.01]之间，例如0.0001、0.0003、0.0004、0.0005、0.0006、0.0007、0.0008、0.003、0.006、0.008、0.009、0.0096、0.01等，具体地，a的值可以理解为对GGA质量误差所能接受的偏差范围。

[0034] 下面，以具体例子说明上述计算过程：

[0035] 对于已有的GGA数据打标后的训练集，默认w_0和w_i初始值都为0，M取服从标准正态分布的随机取值矩阵。对于任意训练集的GGA转换的稀疏矩阵中值x：

[0036] $$w_0 = w_0 - r \frac{\partial L}{\partial W_0} = w_0 - r \cdot l$$

[0037] $$w_i = w_i - r \frac{\partial L}{\partial w_i} = w_i - r \cdot x_i$$

[0038] $$m_{ij} = m_{ij} - r \cdot \frac{\partial L}{\partial m_{ij}} = m_{ij} - r \cdot x_i \cdot \sum_{s=1, s \neq i}^{n} m_{sj} x_s$$

[0039] 其中，r是梯度下降法的步长，r越大最小值学习过程中下降越厉害，r一般取任意

小实数(比如0.00001)。mij表示在n×k矩阵M中,下标为i,j取值,其中i在1...n中取值,j在1...k中取值,同时,s的取值不和i(值域也在1...n)中取值相等。每次循环迭代得到一组w_0、w_i和M,通过这组值对所述误差函数计算一次误差值,相邻两次迭代的误差值小于等于a即可判定算法结束。最后取得的w_0、w_i和M就是所得到的最优值。

[0040] 此方法对于大量GGA数据判定数据质量方面得到大量应用。通过在有限内存中将海量GGA数据进行平行分析,再借助数据矩阵分析和降维的方法对海量数据进行分析处理。本方法成功在分钟级别处理5TBGGA数据的海量质量的自动打标和分析。

[0041] 与现有技术相比,本发明至少具有如下特点:

[0042] (1)本方法的降低了存储空间,采用矩阵奇异值分解的技术将高维海量矩阵分解成可计算和维护的小维度矩阵,进一步提高空间复用率和节省实现和存储的空间复杂度。

[0043] (2)本发明提高了运行效率,通过上述定义所描述,由于采用奇异矩阵分解技术,一个高维度大矩阵分解成一定数量的小矩阵,将多个高维矩阵计算分解成更多的小型矩阵的计算,方便工程实现方面采用并行计算的方法,同时支持横向扩展,时间复杂度可控并且不随输入信息的暴增而无序增长。

[0044] (3)本发明提高预测的准确度,通过对特征向量的内部因子的相互影响建模$<v_i,v_j>$来解决传统机器学习算法不能解决的内部因子相互影响的问题,提高准确度和特征向量选取时候尽量相互独立的要求限制。

[0045] (4)本发明以PCA压缩矩阵方法存储稀疏矩阵,极大的降低存储的空间耗损,提高存储效率,解决了传统差分账户分析时需要较大的内存空间去存储差分账户的行为特征向量的问题。

[0046] 综上所述,本发明提供的方法及系统在GGA数据质量自动,打标,判定和学习领域有十分广阔的应用前景。

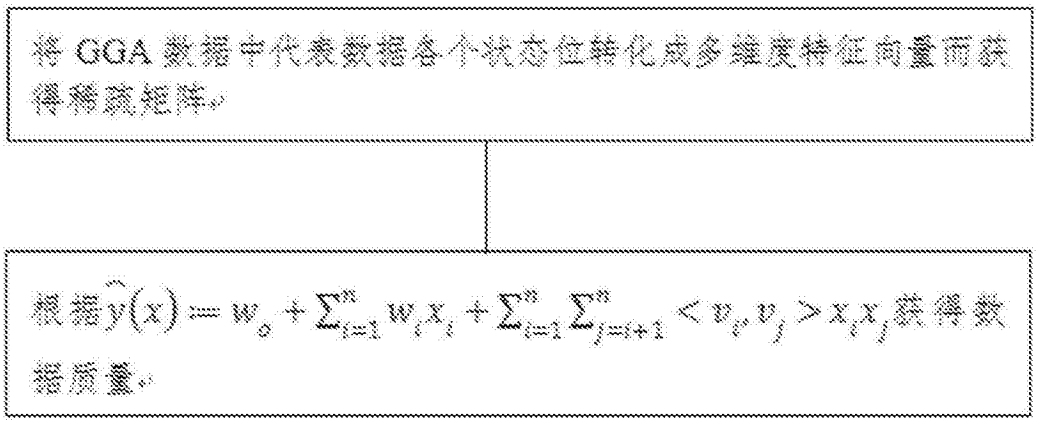

将GGA数据中代表数据各个状态位转化成多维度特征向量而获得稀疏矩阵。
根据$\hat{y}(x) := w_0 + \sum_{i=1}^{n} w_i x_i + \sum_{i=1}^{n} \sum_{j=i+1}^{n} < v_i, v_j > x_i x_j$获得数据质量。

图1