(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 106022360 B
(45)授权公告日 2019.02.01

(21)申请号 201610316111.2

(22)申请日 2016.05.11

(65)同一申请的已公布的文献号
申请公布号 CN 106022360 A

(43)申请公布日 2016.10.12

(73)专利权人 蒋林智
地址 611731 四川省成都市高新区(西区)西源大道2006号电子科技大学计算机科学与工程学院

(72)发明人 蒋林智 王晓芳

(74)专利代理机构 成都行之专利代理事务所(普通合伙) 51220
代理人 温利平

(51)Int.Cl.
G06K 9/62(2006.01)
H04L 9/00(2006.01)
H04L 29/08(2006.01)

(56)对比文件
CN 103095733 A,2013.05.08,
CN 105488422 A,2016.04.13,
CN 104967516 A,2015.10.07,
CN 103607409 A,2014.02.26,
Zvika Brakerski等.Efficient Fully Homomorphic Encryption from (Standard) LWE.《2011 52nd Annual IEEE Symposium on Foundations of Computer Science》.2011,
冯岩盛.云计算中序列比较的外包方案的研究.《中国优秀硕士学位论文全文数据库 信息科技辑》.2014,(第12期),
Zvika Brakerski等.Fully Homomorphic Encryption from Ring-LWE and Security for Key Dependent Messages.《ResearchGate》.2011,
汤殿华等.基于RLWEDE的全同态加密方案.《通信学报》.2014,第35卷(第1期),

审查员 刘素兵

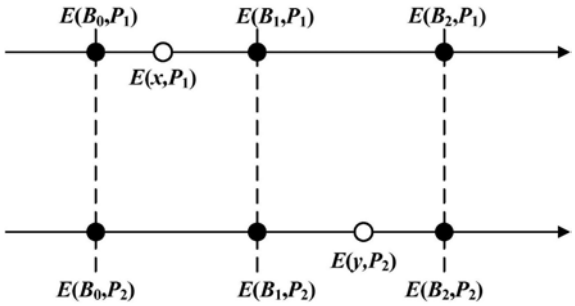
权利要求书1页 说明书5页 附图2页

(54)发明名称

一种统计学习中基于密文的数据二分类方法

(57)摘要

本发明公开了一种统计学习中基于密文的数据二分类方法,利用基于RLWE的全同态加密方法具有的数据与密文的一致性,根据需要加密并在云服务器中进行比较的两个数据,建立一个递增的序列,然后,用两组加密参数分别对其进行加密,并上传至云服务器;在云服务器中,根据比较的数据的密文在构建序列的密文中的位置关系,判断出两个需要比较的数据的大小,并利用这个判断结果完成加密数据的比较计算,从而实现了云服务器中,两个加密数据的比较。在本发明中使用划分的界(Bounds)来划分密文空间,将密文划分成不同的区间,实现了全同态密文的比较计算,同时,避免了用户隐私的泄露。



CN 106022360 B

1.一种统计学习中基于密文的数据二分类方法，其特征在于，二分类模型训练和分类中，加密数据的比较包括以下步骤：

(1)、用户本地对需要加密并在云服务器中进行比较的两个数据x、y按照一个递增的顺序排序，然后，在排好的两个数据之间随机选择一个数据B_1，同时，随机选择一个比两个数据中较小数据小的数据B_0，随机选择一个比两个数据中较大数据大的数据B_2构成一个递增的序列B_0、B_1、B_2，数据B_i，i＝0,1,2，作为比较的界；

用户本地首先采用基于RLWE的全同态加密方法，分别使用两组加密参数P_1，P_2对数据B_i进行加密，得到加密序列$E(B_i,P_1)$、$E(B_i,P_2)$，i＝0,1,2，其中，E表示基于RLWE的全同态加密；然后，将加密序列$E(B_i,P_1)$、$E(B_i,P_2)$，i＝0,1,2，上传到云服务器；

(2)、用户本地对需要在云服务器进行比较的两个数据x、y，使用两组加密参数P_1，P_2分别对数据x、y进行加密，得到加密数据$E(x,P_1)$、$E(y,P_2)$，并上传至云服务器；

(3)、在云服务器中，根据加密数据$E(x,P_1)$找到距离最近的两个界B_p、B_{p+1}，即加密数据$E(x,P_1)$与加密数据$E(B_p,P_1)$、加密数据$E(B_{p+1},P_1)$差值最小，这样得到加密数据$E(x,P_1)$位于加密数据$E(B_p,P_1)$、加密数据$E(B_{p+1},P_1)$之间；同理，能够得到加密数据$E(y,P_2)$位于加密数据$E(B_q,P_2)$、加密数据$E(B_{q+1},P_2)$之间；

(4)、在云服务器中，根据加密数据$E(x,P_1)$、加密数据$E(y,P_2)$所处的加密序列$E(B_i,P_1)$、$E(B_i,P_2)$，i＝0,1,2，的位置关系，判断出用户本地数据x、y的大小，并利用这个判断结果完成加密数据的比较计算。

2.根据权利要求1所述的二分类方法，其特征在于，所有的加密数据需要转换为整数，然后编码成多项式，以满足加密、解密和全同态运算的需要。

3.根据权利要求1所述的二分类方法，其特征在于，所述的基于RLWE的全同态加密方法中采用的安全参数λ为134。

4.根据权利要求1所述的二分类方法，其特征在于，所述的二分类模型为感知器模型、Boosting模型或支持向量机模型。

一种统计学习中基于密文的数据二分类方法

技术领域

[0001] 本发明属于统计学习技术领域，更为具体地讲，涉及一种统计学习中基于密文的数据二分类方法。

背景技术

[0002] 统计学习是以数据为研究目标，提取数据特征并发现数据中相关知识的一个技术领域，其已经广泛应用于数据分析和预测中，包括医疗诊断、垃圾邮件过滤、模式识别和金融预测分析等工程和金融领域。统计学习包括三个阶段：数据存储、数据训练和数据分类，其中二分类是统计学习的核心基础内容之一，它可以应用在垃圾邮件过滤和广告推荐系统等领域，包括感知器分类模型、支持向量机模型以及Boosting模型。其中感知器分类模型是Frank Rosenblatt在1957年提出的分类模型，属于典型的二分类方法，是以一定的样本特征向量和权重构建的线性预测模型；支持向量机模型是利用特征向量的间隔最大化求解的线性化分类模型；Boosting模型将弱分类器组合产生一个更强的分类器，这些二分类模型仍然有广泛的应用。

[0003] 在统计学习过程中，二分类模型的训练需要消耗大量的存储和计算资源。由于云计算具有高效的资源利用率、存储空间大、强大的计算能力和廉价的投资，在外包存储和计算方面受到了广大用户的欢迎。将云计算平台运用于统计学习，即用户上传训练和加密数据到云服务器，云服务器执行数据训练和分析，可以满足二分类模型的训练需要消耗大量的存储和计算资源的问题。

[0004] 然而，数据包括用户信息以及其他敏感信息的安全是统计学习采用云服务的最大挑战和主要障碍。尽管我们可以加密训练数据集和测试数据集，然后上传到云服务器，这样传输、存储的数据是安全的，然而如何使用加密的训练数据得到加密的二分类模型是现有技术需要解决的问题。

发明内容

[0005] 本发明的目的在于克服现有技术的不足，提供一种统计学习中基于密文的数据二分类方法，以实现加密(密文)数据在云服务器中进行二分类模型的训练，以及新加密数据在二分类模型中的分类。

[0006] 为实现上述发明目的，本发明统计学习中基于密文的数据二分类方法，其特征在于，二分类模型训练和分类中，加密数据的比较包括以下步骤：

[0007] (1)、用户本地对需要加密并在云服务器中进行比较的两个数据x、y按照一个递增的顺序排序，然后，在排好的两个数据之间随机选择一个数据B_1，同时，随机选择一个比两个数据中较小数据小的数据B_0，随机选择一个比两个数据中较大数据大的数据B_2构成一个递增的序列B_0、B_1、B_2，数据B_i，i＝0,1,2，作为比较的界；

[0008] 用户本地首先采用基于RLWE(Ring Learning With Errors，环上错误学习)的全同态加密方法，分别使用两组加密参数P_1，P_2对数据B_i进行加密，得到加密序列$E(B_i,P_1)$、E

(B_i,P_2)，i＝0,1,2，其中，E表示基于RLWE的全同态加密；然后，将加密序列$E(B_i,P_1)$、$E(B_i,P_2)$，i＝0,1,2，上传到云服务器；

[0009] (2)、用户本地对需要在云服务器进行比较的两个数据x、y，使用两组加密参数P_1，P_2分别对数据x、y进行加密，得到加密数据$E(x,P_1)$、$E(y,P_2)$，并上传至云服务器；

[0010] (3)、在云服务器中，根据加密数据$E(x,P_1)$找到距离最近的两个界B_p、B_{p+1}，即加密数据$E(x,P_1)$与加密数据$E(B_p,P_1)$、加密数据$E(B_{p+1},P_1)$差值最小，这样得到加密数据$E(x,P_1)$位于加密数据$E(B_p,P_1)$、加密数据$E(B_{p+1},P_1)$之间；同理，能够得到加密数据$E(y,P_2)$位于加密数据$E(B_q,P_2)$、加密数据$E(B_{q+1},P_2)$之间；

[0011] (4)、在云服务器中，根据加密数据$E(x,P_1)$、加密数据$E(y,P_2)$所处的加密序列$E(B_i,P_1)$、$E(B_i,P_2)$，i＝0,1,2，的位置关系，判断出用户本地数据x、y的大小，并利用这个判断结果完成加密数据的比较计算。

[0012] 本发明的目的是这样实现的。

[0013] 借助云服务器(云计算平台)，统计学习要求用户上传训练数据和测试数据到云端，云服务器根据一定的策略训练这些数据集，得到相应的统计学习模型。然后，云服务器利用得到的分类模型可以根据新的数据实现分析和预测。统计学习经常使用损失函数来度量预测误差，使用符号函数(Sign Function)来做分析和预测，通常可以将两种计算转换为数据的比较计算。

[0014] 本发明统计学习中基于密文的数据二分类方法，利用基于RLWE的全同态加密方法具有的数据与密文的一致性，根据需要加密并在云服务器中进行比较的两个数据，建立一个递增的序列，然后，用两组加密参数分别对其进行加密，并上传至云服务器，同时，该两组加密参数分别对两个需要在云服务器中进行比较的两个数据进行加密，并上传至云服务器；在云服务器中，根据比较的数据的密文在构建序列的密文中的位置关系，判断出两个需要比较的数据之间的大小关系，从而实现了云服务器中，两个加密数据的比较。在本发明中使用划分的界(Bounds)来划分密文空间，将密文划分成不同的区间，实现了全同态密文的比较计算。同时，由于B_i是随机的，同时也进行了加密，在云服务器中，只反应出一个递增的序列，这样其作为一个参照，对加密数据进行比较，避免了用户隐私的泄露。

附图说明

[0015] 图1是本发明统计学习中基于密文的数据二分类方法一种具体方式流程图；

[0016] 图2是本发明步骤(1)中数B_i作为界的选取示意图；

[0017] 图3是本发明步骤(3)、(4)比较加密数据的位置关系示意图。

具体实施方式

[0018] 下面结合附图对本发明的具体实施方式进行描述，以便本领域的技术人员更好地理解本发明。需要特别提醒注意的是，在以下的描述中，当已知功能和设计的详细描述也许会淡化本发明的主要内容时，这些描述在这里将被忽略。

[0019] 全同态加密方法可以实现用户的隐私(用户信息以及其他敏感信息)保护，但是大部分的全同态加密方法效率低下，不能满足实际应用的需求。基于环上错误学习(RLWE，即Ring Learning With Errors)可以看作是分圆域上多项式的计算，支持对密文的有限次加

法和乘法的有限次计算，因此基于RLWE的全同态加密方法比其他的全同态加密方法更有效率。有鉴于此，本发明中用户本地对数据进行加密采用基于RLWE的全同态加密方法。

[0020] 此外，为了进一步提高安全性，在本实施例中，基于RLWE的全同态加密方法中采用的安全参数λ为134。

[0021] 图1是本发明统计学习中基于密文的数据二分类方法一种具体方式流程图。

[0022] 在本实施例中，如图1所示，本发明统计学习中基于密文的数据二分类方法包括以下步骤：

[0023] S1：用户本地对需要加密的，并在云服务器中进行比较的两个数据x、y按照一个递增的顺序排序，然后，在排好的两个数据之间随机选择一个数据B_1，同时，随机选择一个比两个数据中较小数据小的数据B_0，随机选择一个比两个数据中较大数据大的数据B_2构成一个递增的序列B_0、B_1、B_2，数据B_i，i＝0,1,2，作为比较的界；

[0024] 用户本地首先采用基于RLWE的全同态加密方法，分别使用两组加密参数P_1，P_2对数据B_i进行加密，得到加密序列(密文)$E(B_i,P_1)$、$E(B_i,P_2)$，i＝0,1,2，其中，E表示基于RLWE的全同态加密；然后，将加密序列$E(B_i,P_1)$、$E(B_i,P_2)$上传到云服务器。

[0025] 在本实施例中，如图2所示，数据B_1随机地位于两个数据x、y之间，数据B_0随机在小于两个数据中较小数据中选择，数据B_2随机在大于两个数据中较大数据中选择。在图2中，数据x小于数据y，因此，随机选择一个小于数据x的数据作为B_0，随机选择一个大于数据y的数据作为B_2，这样，得到一个递增序列B_0、B_1、B_2，数据B_i，i＝0,1,2，作为比较的界。反之，数据x大于数据y，其构建递增序列B_0、B_1、B_2方法相同，只不过数据x、y的位置对调了一下而已。

[0026] 此外，分别用加密参数P_1，P_2对B_0、B_1、B_2进行加密，得到加密数据$E(B_0,P_1)$、加密数据$E(B_1,P_1)$、加密数据$E(B_2,P_1)$，以及加密数据$E(B_0,P_2)$、加密数据$E(B_1,P_2)$、加密数据$E(B_2,P_2)$，其中，E表示基于RLWE的全同态加密；然后，将加密序列$E(B_i,P_1)$、$E(B_i,P_2)$，i＝0,1,2，上传到云服务器。

[0027] S2：用户本地对需要在云服务器进行比较的两个数据x、y，使用两组加密参数P_1，P_2分别对数据x、y进行加密，得到加密数据$E(x,P_1)$、$E(y,P_2)$，并上传至云服务器；

[0028] S3、在本实施例中，如图3所示，在云服务器中，根据加密数据$E(x,P_1)$找到距离最近的两个界B_p、B_{p+1}，即加密数据$E(x,P_1)$与加密数据$E(B_p,P_1)$、加密数据$E(B_{p+1},P_1)$差值最小，这样得到加密数据$E(x,P_1)$位于加密数据$E(B_p,P_1)$、加密数据$E(B_{p+1},P_1)$之间；同理，能够得到加密数据$E(y,P_2)$位于加密数据$E(B_q,P_2)$、加密数据$E(B_{q+1},P_2)$之间。在本实施例中，数据x小于数据y，p＝0，而q＝1，即加密数据$E(x,P_1)$位于加密数据$E(B_0,P_1)$、加密数据$E(B_1,P_1)$之间，加密数据$E(y,P_2)$位于加密数据$E(B_1,P_2)$、加密数据$E(B_2,P_2)$之间。反之，数据x大于数据y，则p＝1，而q＝0，即加密数据$E(x,P_1)$位于加密数据$E(B_1,P_1)$、加密数据$E(B_2,P_1)$之间，加密数据$E(y,P_2)$位于加密数据$E(B_0,P_2)$、加密数据$E(B_1,P_2)$之间。

[0029] S4：在云服务器中，根据加密数据$E(x,P_1)$、加密数据$E(y,P_2)$所处的加密序列$E(B_i,P_1)$、$E(B_i,P_2)$，i＝0,1,2，的位置关系，判断出用户本地数据x、y的大小，并利用这个判断结果完成加密数据的比较计算。在本实施例中，如图3所示，q大于p，则数据数据y大于数据x，反之，则数据y小于数据x。

[0030] 在本发明中，所有的加密数据需要转换为整数，然后编码成多项式，以满足加密、解密和全同态运算的需要。

[0031] 下面就本发明在三种二分类模型中的应用做一个详细说明。

[0032] 二分类模型是给定训练数据集，对函数f进行训练，其中函数f可以表示为：

[0033] $$f=\left\{(x_i,y_i)\,|\,x_i\in R^n,y_i\in\{-1,1\},i=1,...,m\right\},x_i\in f_+\Leftrightarrow y_i=1,x_i\in f_-\Leftrightarrow y_i=-1;$$

[0034] 目标是找到函数f：使得$f(x)\geqslant 0,x\in f_+$；$f(x)<0,x\in f_-$。

[0035] 最简单的二分类函数为线性函数：$f(x)=w^Tx+b$，用来实现对数据集的分类。

[0036] 一、基于密文的感知器模型(对偶形式)

[0037] 在模型训练阶段，我们必须计算：

[0038] $$y_i\left(\sum_{j=1}^{N}\alpha_j y_j x_j gx_i+b\right)\leq 0。$$

[0039] 我们可以将上式转换为：

[0040] $$y_i\sum_{j=1}^{N}\alpha_j y_j x_j gx_i\leq -y_i b。$$

[0041] 这样根据本发明中的加密数据比较方法，我们可以实现：

[0042] $$E\left(y_i\sum_{j=1}^{N}\alpha_j y_j x_j gx_i\right)\leq E\left(-y_i b\right)$$

[0043] 的比较计算。

[0044] 其中：

[0045] $$y_i\sum_{j=1}^{N}\alpha_j y_j x_j gx_i$$

[0046] 相当于本发明中$x,-y_ib$相当于本发明中的y。因此，云服务器可以对训练数据集不断的进行比较，从而得到加密的二分类感知器模型：

[0047] $E(f_{\pm}(x))=E(sign(w^TX+b))$。

[0048] 当使用加密的感知器模型对新的数据进行分析和预测时，云服务器根据加密数据(密文)进行比较为：

[0049] 比较$E\left(\sum_{j=1}^{N}\alpha_j y_j x_j+b\right)$和E(0)的大小。

[0050] 从而，云服务器实现对新数据(加密状态下)的二分类。

[0051] 二、基于密文的Boosting模型

[0052] Boosting模式下主要的计算是$\alpha_mG_m(x)$求和以及sign符号函数的密文计算。在训练和分类阶段，云服务器可以直接执行全同态加密计算$E(\alpha_mG_m(x))=E(\alpha_m)gE(G_m(x))$，符号函数sign计算转换成：

[0053] $$E\left(G(x)\right)=E\left(sign(f(x))\right)=E\left(sign\left(\sum_{m=1}^{m}\alpha_m G_m(x)\right)\right):$$

[0054] 即：

[0055] $E\left(\sum_{m=1}^{m}\alpha_m G_m(x)\right)$与E(0)的比较，从而实现模型的训练和加密数据的二分类。

[0056] 三、基于密文的支持向量机(ESVM)模型

[0057] 在支持向量机的训练阶段，我们可以将$w^Tx+b=0$转换成：$\sum\alpha_i^* y_i(xgx_i)+b^*=0$。根据训练数据$x_i$，$y_i$，云服务器计算：$E\left(\sum\alpha_i^* y_i(xgx_i)+b^*\right)=E(0)$。从而得到加密的支持向量机模型：$E[f(x)]=E\left[sign\left(\sum\alpha_i^* y_i(xgx_i)+b^*\right)\right]$。

[0058] 在利用获得加密模型对新的加密数据进行分析预测时，将$E[f(x)]=E\left[sign\left(\sum\alpha_i^* y_i(xgx_i)+b^*\right)\right]$转化为$E\left(\sum\alpha_i^* y_i(xgx_i)+b^*\right)$和E(0)比较计算，从而实现对新加密数据的分类。

[0059] 尽管上面对本发明说明性的具体实施方式进行了描述，以便于本技术领域的技术人员理解本发明，但应该清楚，本发明不限于具体实施方式的范围，对本技术领域的普通技术人员来讲，只要各种变化在所附的权利要求限定和确定的本发明的精神和范围内，这些变化是显而易见的，一切利用本发明构思的发明创造均在保护之列。

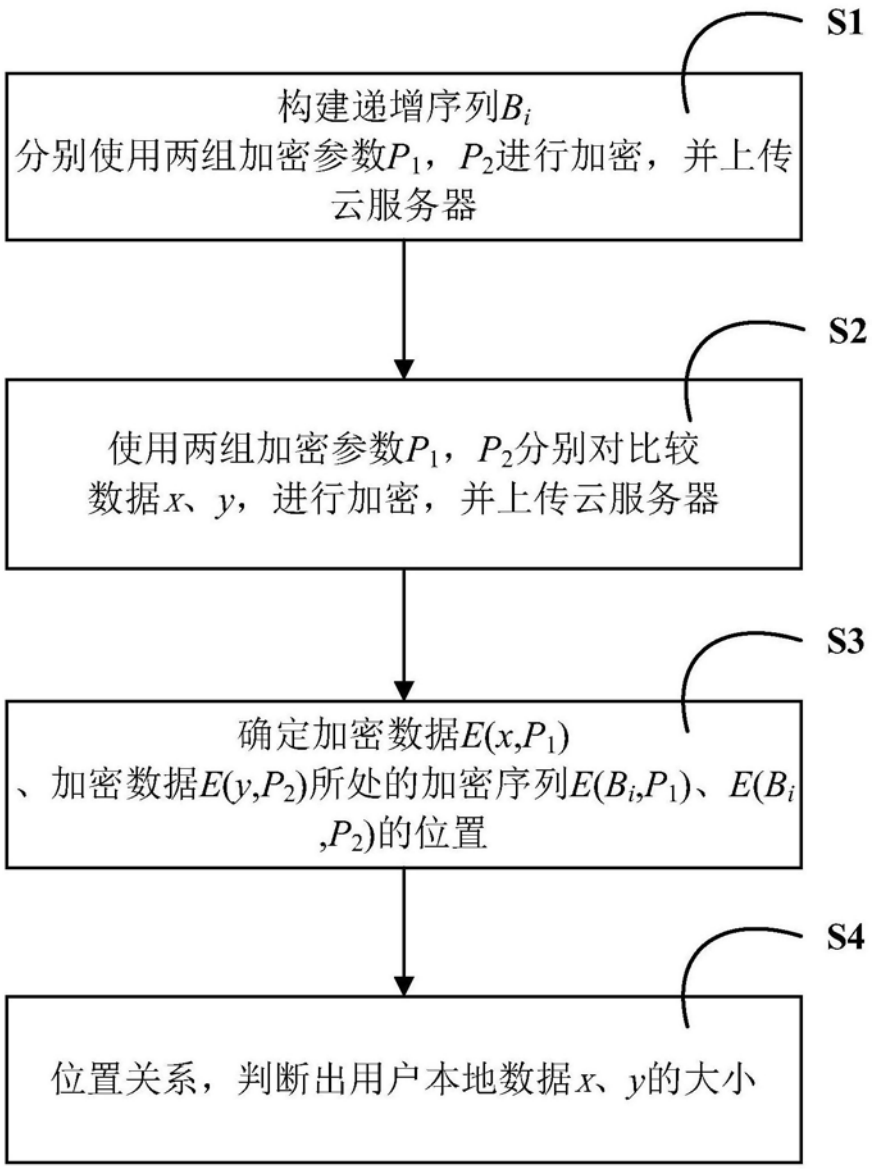

S1
构建递增序列B_i
分别使用两组加密参数P_1，P_2进行加密，并上传云服务器
S2
使用两组加密参数P_1，P_2分别对比较数据x、y，进行加密，并上传云服务器
S3
确定加密数据$E(x,P_1)$、加密数据$E(y,P_2)$所处的加密序列$E(B_i,P_1)$、$E(B_i,P_2)$的位置
S4
位置关系，判断出用户本地数据x、y的大小

图1

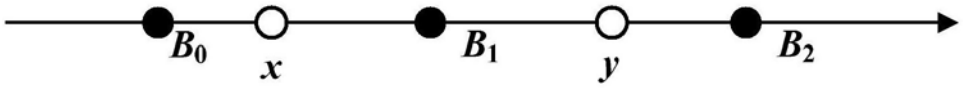

B_0
x
B_1
y
B_2

图2

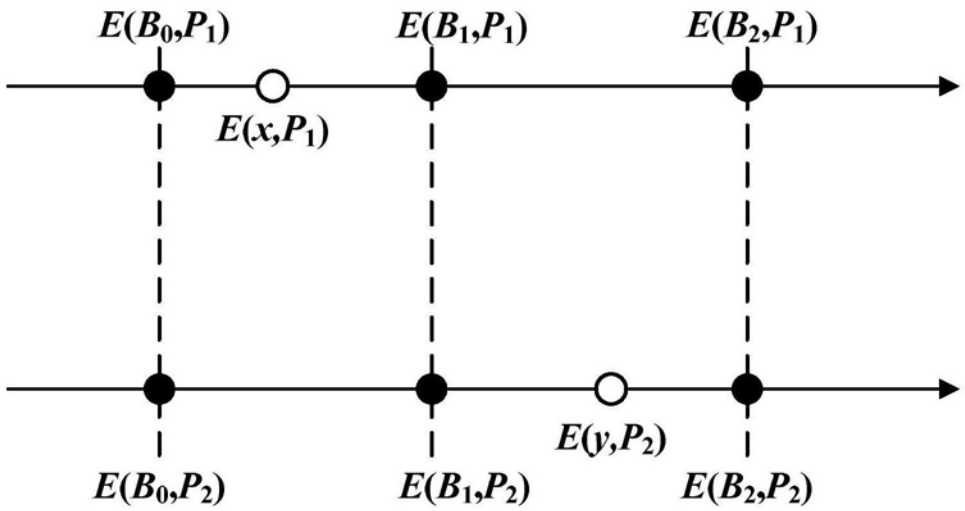


图3