(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 105989095 B
(45)授权公告日 2019.09.06

(21)申请号 201510076329.0

(22)申请日 2015.02.12

(65)同一申请的已公布的文献号
申请公布号 CN 105989095 A

(43)申请公布日 2016.10.05

(73)专利权人 香港理工大学深圳研究院
地址 518000 广东省深圳市南山区高新园南区粤兴一道18号香港理工大学产学研大楼205室

(72)发明人 史文中 张安舒

(74)专利代理机构 深圳中一专利商标事务所 44237
代理人 张全文

(51)Int.Cl.
G06F 16/2458(2019.01)

(56)对比文件
CN 101799810 A,2010.08.11,
CN 101937447 A,2011.01.05,
CN 101667197 A,2010.03.10,
李德仁 等.论空间数据挖掘和知识发现.《武汉大学学报(信息科学版)》.2001,第26卷(第6期),第491-493页.

审查员 乔晋

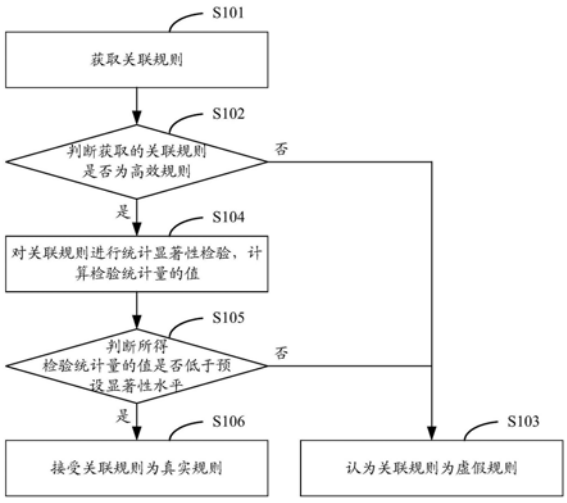
权利要求书4页 说明书15页 附图3页

(54)发明名称

顾及数据不确定性的关联规则显著性检验方法及装置

(57)摘要

本发明适用于数据挖掘技术领域,提供了顾及数据不确定性的关联规则显著性检验方法及装置。所述方法包括:获取关联规则,并判断获取的所述关联规则是否为高效规则;若所述关联规则不为所述高效规则,则认为所述关联规则为虚假规则;若所述关联规则为所述高效规则,则对所述关联规则进行统计检验,并判断所得检验统计量的值是否低于预设显著性水平,若是,则接受所述关联规则为真实规则;若否,则认为所述关联规则为虚假规则。本发明基于统计健全检验法,能将族错误率控制在较低水平;修正随机数据误差对所述统计检验运算的影响,由此显著恢复由于随机数据误差引起的统计检验结果中真实规则的丢失,大大提高了关联规则挖掘结果的可靠性。



CN 105989095 B

1.一种顾及数据不确定性的关联规则显著性检验方法,其特征包括:

获取关联规则,并判断获取的所述关联规则是否为高效规则;

若所述关联规则不为所述高效规则,则认为所述关联规则为虚假规则;

若所述关联规则为所述高效规则,则对所述关联规则进行统计检验,并判断所得检验统计量p的值是否低于预设显著性水平,若是,则接受所述关联规则为真实规则;若否,则认为所述关联规则为虚假规则;所述统计检验涉及的每一个数据模式为若干数据项的集合,每个数据项指的是数据中一个属性中的一个类别,每个属性的误差概率分布为已知;

所述对所述关联规则进行统计检验,计算检验统计量的值包括:

对所述统计检验涉及的每一个数据模式,将其中指定数据项所对应的属性的误差概率分布表达为误差矩阵,所述误差矩阵包括指定属性的全部k个类别之间的误差分布,其中,指定属性指的是所述指定数据项对应的属性,k为大于1的整数;

根据所述误差矩阵,对数据误差的传播进行建模,得到所述k个类别的观测支持度分布期望及方差;

根据所估计的k个类别的观测支持度分布以及所述误差矩阵,计算所述k个类别的真实支持度估计值;

以c_i表示所述统计检验涉及的数据模式中的指定数据项,将所述k个类别中的每个类别与所述数据模式中除c_i以外的所有数据项求并集,得到k个并集,其中包含c_i的并集即为所述数据模式;根据所述k个类别的真实支持度估计值,以及k个并集在数据中的支持度观测值,计算所述数据模式的真实支持度估计值;

根据所述统计检验所涉及数据模式的真实支持度估计值,计算所述统计检验的第一参数估计真值、第二参数估计真值、第三参数估计真值以及第四参数估计真值,以对第一参数观测值、第二参数观测值、第三参数观测值以及第四参数观测值受到数据误差的影响进行修正;

根据所述第一参数估计真值、第二参数估计真值、第三参数估计真值以及第四参数估计真值计算所述检验统计量p的值;

所述检验统计量p由费氏精确检验得出,检验中的四个关键计算参数a,b,c,d为:

$a=s(X\cup\{y\})$

$b=s(X)-s(X\cup\{y\})$

$c=s(X-\{x_m\})\cup\{y\}-s(X\cup\{y\})$

$d=s(X-\{x_m\})-s((X-\{x_m\})\cup\{y\})+s(X\cup\{y\})$

其中a表示第一参数,b表示第二参数,c表示第三参数,d表示第四参数,x_m为被检验是否冗余的项,$x_m\in X$,s表示各数据模式的观测支持度,a～d的真值为a_0,b_0,c_0,d_0,其中所述真值a_0,b_0,c_0,d_0无随机数据误差影响,根据

$$\begin{aligned}a&=s(X\cup\{y\})\\b&=s(X)-s(X\cup\{y\})\\c&=s(X-\{x_m\})\cup\{y\}-s(X\cup\{y\})\\d&=s(X-\{x_m\})-s((X-\{x_m\})\cup\{y\})+s(X\cup\{y\})\end{aligned}$$

的各关键计算参数的内容可变化I和c_i的值,将

$\hat{E}(c_i,I,P,z)=\hat{s_0}(I\cup\{c_i\})=\sum_{j=1}^{k}\left(p_{ij}^{-i}\left(s(I\cup\{c_j\})-z\left(\sum_{l=1}^{k}p_{jl}(1-p_{jl})s(I\cup\{c_l\})\right)^{1/2}\right)\right)$应用于a～d，以获得估计真值$\hat{a_0}$，$\hat{b_0}$，$\hat{c_0}$，$\hat{d_0}$。

2.如权利要求1所述的方法，其特征在于，在所述根据所述统计检验所涉及数据模式的真实支持度估计值，计算第一参数估计真值、第二参数估计真值、第三参数估计真值以及第四参数估计真值时，所述方法还包括：

使用经过随机化处理的数据进行模拟的关联规则提取，求出使所述统计检验的族错误率小于指定上限的最佳参数修正量，其中，所述最佳参数修正量为非负数；

将所述最佳参数修正量用于计算所述第一参数估计真值以及第四参数估计真值；

将所述最佳参数修正量的相反数用于计算所述第二参数估计真值以及第三参数估计真值。

3.如权利要求2所述的方法，其特征在于，在所述求出使所述统计检验的族错误率小于指定上限的最佳参数修正量的过程中，所述方法还包括：

对数据中每个属性在所有记录中的类别进行n次随机排列，其中，n为大于1的整数；

对每一次随机排列，从随机排列后的数据中获取关联规则，取参数修正量z为0，对获取的所述关联规则进行统计检验，并逐渐增大z值，直至所有所述关联规则均被判定为虚假规则，并记录此时的z值；

将n次数据随机排列所得到的n个z值中最大者作为所述最佳参数修正量。

4.如权利要求2所述的方法，其特征在于，在所述根据所述统计检验所涉及数据模式的真实支持度估计值，计算第一参数估计真值、第二参数估计真值、第三参数估计真值以及第四参数估计真值时，所述方法还包括：

根据c_i在所述关联规则中所处的的位置，获取与所述位置对应的修正数学式计算所述第一参数估计真值、第二参数估计真值、第三参数估计真值以及第四参数估计真值。

5.如权利要求1所述的方法，其特征在于，所述根据所述第一参数估计真值、第二参数估计真值、第三参数估计真值以及第四参数估计真值计算所述检验统计量p的值，其具体过程为：

将所述第一参数估计真值、第二参数估计真值、第三参数估计真值以及第四参数估计真值用于健全统计检验法，计算所述检验统计量p的值。

6.一种顾及数据不确定性的关联规则显著性检验装置，其特征包括：

高效规则判断单元，用于获取关联规则，并判断获取的所述关联规则是否为高效规则；

虚假规则判定单元，用于若所述关联规则不为所述高效规则，则认为所述关联规则为虚假规则；

检验单元，用于若所述关联规则为所述高效规则，则对所述关联规则进行统计检验，并判断所得检验统计量p的值是否低于预设显著性水平，若是，则接受所述关联规则为真实规则；若否，则认为所述关联规则为虚假规则；所述统计检验涉及的每一个数据模式为若干数据项的集合，每个数据项指的是数据中一个属性中的一个类别，每个属性的误差概率分布为已知；

所述检验单元包括检验统计量值计算子单元，所述检验统计量值计算子单元具体用于：

对所述统计检验涉及的每一个数据模式，将其中指定数据项所对应的属性的误差概率分布表达为误差矩阵，所述误差矩阵包括所述指定属性的全部k个类别之间的误差分布，其中，指定属性指的是所述指定数据项对应的属性，k为大于1的整数；

根据所述误差矩阵，对数据误差的传播进行建模，得到所述k个类别的观测支持度分布期望及方差；

根据所估计的k个类别的观测支持度分布以及所述误差矩阵，计算所述k个类别的真实支持度估计值；

以c_i表示所述统计检验涉及的数据模式中的指定数据项，将所述k个类别中的每个类别与所述数据模式中除c_i以外的所有数据项求并集，得到k个并集，其中包含c_i的并集即为所述数据模式；根据所述k个类别的真实支持度估计值，以及k个并集在数据中的支持度观测值，计算所述数据模式的真实支持度估计值；

根据所述统计检验所涉及数据模式的真实支持度估计值，计算所述统计检验的第一参数估计真值、第二参数估计真值、第三参数估计真值以及第四参数估计真值，以对第一参数观测值、第二参数观测值、第三参数观测值以及第四参数观测值受到数据误差的影响进行修正；

根据所述第一参数估计真值、第二参数估计真值、第三参数估计真值以及第四参数估计真值计算所述检验统计量p的值；

所述检验统计量p由费氏精确检验得出，检验中的四个关键计算参数a,b,c,d为：

$a=s(X\cup\{y\})$

$b=s(X)-s(X\cup\{y\})$

$c=s(X-\{x_m\})\cup\{y\}-s(X\cup\{y\})$

$d=s(X-\{x_m\})-s((X-\{x_m\})\cup\{y\})+s(X\cup\{y\})$

其中a表示第一参数，b表示第二参数，c表示第三参数，d表示第四参数，x_m为被检验是否冗余的项，$x_m\in X$，s表示各数据模式的观测支持度，a～d的真值为a_0,b_0,c_0,d_0，其中所述真值a_0,b_0,c_0,d_0无随机数据误差影响，根据

$$\begin{aligned}a&=s(X\cup\{y\})\\b&=s(X)-s(X\cup\{y\})\\c&=s(X-\{x_m\})\cup\{y\}-s(X\cup\{y\})\\d&=s(X-\{x_m\})-s((X-\{x_m\})\cup\{y\})+s(X\cup\{y\})\end{aligned}$$

的各关键计算参数的内容可变化I和c_i的值，将

$$\hat{E}(c_i,I,P,z)=\hat{s}_0(I\cup\{c_i\})=\sum_{j=1}^{k}\left(p_{ij}^{-i}\left(s(I\cup\{c_j\})-z\left(\sum_{l=1}^{k}p_{jl}(1-p_{jl})s(I\cup\{c_l\})\right)^{1/2}\right)\right)$$

应用于a～d，以获得估计真值$\hat{a}_0$，$\hat{b}_0$，$\hat{c}_0$，$\hat{d}_0$。

7.如权利要求6所述的装置，其特征在于，所述装置还包括检验参数修正单元，所述检验参数修正单元用于：

使用经过随机化处理的数据进行模拟的关联规则提取，求出使所述统计检验的族错误

率小于指定上限的最佳参数修正量，其中，所述最佳参数修正量为非负数；

将所述最佳参数修正量用于计算所述第一参数估计真值以及第四参数估计真值；

将所述最佳参数修正量的相反数用于计算所述第二参数估计真值以及第三参数估计真值。

8.如权利要求7所述的装置，其特征在于，所述装置还包括最佳参数修正量确定单元，所述最佳参数修正量确定单元用于：

对数据中每个属性在所有记录中的类别进行n次随机排列，其中，n为大于1的整数；

对每一次随机排列，从随机排列后的数据中获取关联规则，取参数修正量z为0，对获取的所述关联规则进行统计检验，并逐渐增大z值，直至所有所述关联规则均被判定为虚假规则，并记录此时的z值；

将n次数据随机排列所得到的n个z值中最大者作为所述最佳参数修正量。

9.如权利要求7所述的装置，其特征在于，所述检验参数修正单元还用于：

根据c_i在所述关联规则中所处的的位置，获取与所述位置对应的修正数学式计算所述第一参数估计真值、第二参数估计真值、第三参数估计真值以及第四参数估计真值。

10.如权利要求6所述的装置，其特征在于，所述检验统计量值计算子单元具体用于：

将所述第一参数估计真值、第二参数估计真值、第三参数估计真值以及第四参数估计真值用于健全统计检验法，计算所述检验统计量p的值。

顾及数据不确定性的关联规则显著性检验方法及装置

技术领域

[0001] 本发明属于数据挖掘技术领域，尤其涉及顾及数据不确定性的关联规则显著性检验方法及装置。

背景技术

[0002] 关联规则挖掘旨在提取数据库中所有符合给定兴趣度指标的规则，是数据挖掘中的一大研究课题。关联规则挖掘尤其适合探索现代数据库中复杂且多角的关系，目前已广泛应用于研究与实践中的数据分析与决策支持。

[0003] 提升关联规则挖掘价值的关键在于获取可靠的结果，即发现有助于决策的真实规则，并避免表达数据中并不存在的虚假规则，以防误导用户做出错误决策。数据库中的项目很可能组合成数以万计甚至亿计的潜在规则，因此，挖掘结果中通常包含大量的虚假规则，这已成为关联规则挖掘结果可靠性的关键阻碍因素。另外，关联规则挖掘所用数据中普遍存在的误差是数据不确定性的一大来源。误差从源数据传播到关联规则挖掘中的每一个阶段，导致结果中真实规则的丢失和虚假规则的增加。

[0004] 最初的关联规则研究提出了采用支持度(support)和可信度(confidence)两个基本的兴趣度指标来衡量关联规则的价值。后续研究又提出了采用其它指标值与支持度、可信度结合来衡量关联规则的价值。每条关联规则中的指标值由该关联规则及其相关模式在数据库中的数量计算得来。若指标值高于(有时是低于)给定的阈值，则认为该关联规则为真实规则，否则认为该关联规则为虚假规则。这些单一阈值的兴趣度指标可能有效地减少虚假规则，但所采用的阈值通常难以通过科学推导确定，也缺少普适的经验值，而是由用户主观给定。因此，所采用的阈值很可能并不合理，很可能导致不能有效滤除虚假规则，或者误删过多的真实规则。综上，采用该方法筛选出的关联规则的可靠性较低。

[0005] 对关联规则的统计检验是一类重要的避免虚假规则的方法。在这类方法中，若关联规则对给定兴趣度指标的符合程度不具有统计显著性，则认为其为虚假规则，并将其滤除。无论是全体数据还是抽样数据，都是现实世界的有限次表达，可以看作现实的“有限样本”。在数据中，一条关联规则之所以符合给定的兴趣度指标，可能并非由于相应的关联在现实中确实符合该兴趣度指标，而仅出自现实在数据中进行有限次表达(即采样)的偶然，此时该规则为虚假规则。因此，很多研究利用统计检验来滤除虚假规则。以零假设为例，检验的结果为一概率值p表示零假设成立时，该关联规则得到数据中观测到的兴趣度指标值的可能性，也就是该关联规则为虚假规则的可能性。当p小于给定的显著性水平α，如0.05时，则接受该关联规则为真实规则，反之则认为该关泽规则为虚假规则并将其删除。

[0006] 统计检验可以显著减少虚假规则，但很难将其基本消除。显著性水平α指的是每条通过检验的关联规则为虚假规则的概率。若n条关联规则被同时检验，则接受至少一条虚假规则的可能性，即族错误率将远远大于α。即使α和n值较小，族错误率仍然接近100％，即结果中几乎必然有虚假规则。这个问题可以用多重比较的Bonferroni修正来解决。最直接的办法是，要将族错误率控制在α，则将检验每条关联规则的显著性水平设为κ＝α/n。但此法

收效不佳，所得结果中通常仍然包含多条虚假规则。这是因为被检验的关联规则一般已经过支持度等兴趣度指标的初步筛选，因而比其他关联规则更倾向于通过检验。

[0007] 统计健全检验成功地将族错误率控制在很低的水平，如5％。该方法针对只含一个项目y的关联规则后件Y＝{y}，这也是常见的实际情况，对每一条规则X→y，X＝{x_1...x_n}，检验其是否符合以下条件，且符合程度具有统计显著性：

[0008] $\forall m=1...n,\ \Pr(y|X)>\Pr(y|X-\{x_m\})$。

[0009] 也就是说，X中每一个项目都使y发生的可能性更大，X中没有冗余项目。对于$\forall m=1...n,\ \Pr(y|X)>\Pr(y|X-\{x_m\})$的假设检验，其零假设为Pr(y|X)＝Pr(y|X-{x_m})，即X→y在数据中呈现为高效规则仅仅出于偶然，而非出自项目x_m与关联规则中其他项目的真实关联。

[0010] 费氏精确检验（Fisher exact test）是最适合检验$\forall m=1...n,\ \Pr(y|X)>\Pr(y|X-\{x_m\})$的方法，步骤如下。令a,b,c,d为数据D中含有以下模式的记录数量：

[0011] a＝|D|×Pr(X∪{y})

[0012]
$$\begin{aligned} b&=|D|\times\Pr(X\cup\neg\{y\})\\ c&=|D|\times\Pr((X-\{x_m\})\cup\neg\{x_m\}\cup\{y\})\\ d&=|D|\times\Pr((X-\{x_m\})\cup\neg\{x_m\}\cup\neg\{y\}) \end{aligned}$$ ，

[0013] 其中|D|为数据中记录的总数，¬指数据中不含此项目，如b为包含X中所有项目，且不包含y的记录数量。该检验的p值为

[0014] $$p=\sum_{i=0}^{\min(b,c)}\frac{(a+b)!(c+d)!(a+c)!(b+d)!}{(a+b+c+d)!(a+i)!(b-i)!(c-i)!(d+i)!}$$。

[0015] 在统计健全检验法中，Bonferroni修正不使用待检测规则的数量n，而取显著性水平κ＝α/s，s为数据中所有项目排列组合出的潜在规则的总数。如有20个数据项，规定X中至多有4个项目，则$s=C_{20}^{1}\times C_{20-1}^{1}$（X包含一个项目）$+C_{20}^{2}\times C_{20-2}^{1}$（X包含两个项目）$+C_{20}^{3}\times C_{20-3}^{1}$（X包含三个项目）$+C_{20}^{4}\times C_{20-4}^{1}$（X包含四个项目）$=100700$。只需少量的数据项，s就达到数以万计甚至亿计，导致κ值极小。实验证明，采用该κ值能发现相当大比例的真实规则，而族错误率可低至不到1％。

[0016] 统计健全检验法是目前避免虚假规则最有效的方法，可将族错误率控制在很低的水平。然而，当数据有误差时，统计健全检验法会同时造成大量真实规则的丢失，而数据误差在关联规则挖掘中是非常普遍的。除系统误差外，数据误差多随机发生，与数据项没有关联，因此会弱化数据项之间的关联，导致很多原本能被发现的真实规则无法通过检验而丢失，严重影响关联规则挖掘结果的可靠性。

[0017] 现有的顾及数据不确定性的关联规则挖掘方法主要针对不确定数据库的数据结构，即对每一记录或数据项赋予概率值，表示该记录或数据项的不确定程度。如医学实验中，患者甲10天中有7天头痛，则记录条“甲”的“头痛”属性值为“有”，其概率值为0.7。然而，

这些研究不适用于解决随机数据误差对关联规则统计检验的影响。这些研究通常将误差列为数据不确定性的一大来源，但对数据项赋予固定概率值的模型与数据误差的随机发生的表现相去甚远。现有技术均采用基于固定概率值的不确定数据结构，而无一针对数据误差的随机性进行建模。

[0018] 综上，现有的统计健全检验法能有效避免虚假规则，但在存在数据误差时，会明显导致真实规则的丢失。

发明内容

[0019] 鉴于此，本发明实施例提供了一种顾及数据不确定性的关联规则显著性检验方法及装置，以解决现有的统计健全检验法在存在数据误差时导致真实规则大量丢失的问题。

[0020] 一方面，本发明实施例提供了一种顾及数据不确定性的关联规则显著性检验方法，包括：

[0021] 获取关联规则，并判断获取的所述关联规则是否为高效规则；

[0022] 若所述关联规则不为所述高效规则，则认为所述关联规则为虚假规则；

[0023] 若所述关联规则为所述高效规则，则对所述关联规则进行统计检验，并判断所得检验统计量p的值是否低于预设显著性水平，若是，则接受所述关联规则为真实规则；若否，则认为所述关联规则为虚假规则；所述统计检验涉及的每一个数据模式为若干数据项的集合，每个数据项指的是数据中一个属性中的一个类别，每个属性的误差概率分布为已知；

[0024] 所述对所述关联规则进行统计检验包括：

[0025] 对所述统计检验涉及的每一个数据模式，将其中指定数据项所对应的属性的误差概率分布表达为误差矩阵，所述误差矩阵包括指定属性的全部k个类别之间的误差分布，其中，指定属性指的是所述指定数据项对应的属性，k为大于1的整数；

[0026] 根据所述误差矩阵，对数据误差的传播进行建模，得到所述k个类别的观测支持度分布期望及方差；

[0027] 根据所估计的k个类别的观测支持度分布以及所述误差矩阵，计算所述k个类别的真实支持度估计值；

[0028] 以c_i表示所述统计检验涉及的数据模式中的指定数据项，将所述k个类别中的每个类别与所述数据模式中除c_i以外的所有数据项求并集，得到k个并集，其中包含c_i的并集即为所述数据模式；根据所述k个类别的真实支持度估计值，以及k个并集在数据中的支持度观测值，计算所述数据模式的真实支持度估计值；

[0029] 根据所述统计检验所涉及数据模式的真实支持度估计值，计算所述统计检验的第一参数估计真值、第二参数估计真值、第三参数估计真值以及第四参数估计真值，以对第一参数观测值、第二参数观测值、第三参数观测值以及第四参数观测值受到数据误差的影响进行修正；

[0030] 根据所述第一参数估计真值、第二参数估计真值、第三参数估计真值以及第四参数估计真值计算所述检验统计量p的值。

[0031] 第二方面，本发明实施例提供了一种顾及数据不确定性的关联规则显著性检验装置，包括：

[0032] 高效规则判断单元，用于获取关联规则，并判断获取的所述关联规则是否为高效

规则；

[0033] 虚假规则判定单元，用于若所述关联规则不为所述高效规则，则认为所述关联规则为虚假规则；

[0034] 检验单元，用于若所述关联规则为所述高效规则，则对所述关联规则进行统计检验，并判断所得检验统计量p的值是否低于预设显著性水平，若是，则接受所述关联规则为真实规则；若否，则认为所述关联规则为虚假规则；所述统计检验涉及的每一个数据模式为若干数据项的集合，每个数据项指的是数据中一个属性中的一个类别，每个属性的误差概率分布为已知；

[0035] 所述检验单元包括检验统计量值计算子单元，所述检验统计量值计算子单元具体用于：

[0036] 对所述统计检验涉及的每一个数据模式，将其中指定数据项所对应的属性的误差概率分布表达为误差矩阵，所述误差矩阵包括所述指定属性的全部k个类别之间的误差分布，其中，指定属性指的是所述指定数据项对应的属性，k为大于1的整数；

[0037] 根据所述误差矩阵，对数据误差的传播进行建模，得到所述k个类别的观测支持度分布期望及方差；

[0038] 根据所估计的k个类别的观测支持度分布以及所述误差矩阵，计算所述k个类别的真实支持度估计值；

[0039] 以c_i表示所述统计检验涉及的数据模式中的指定数据项，将所述k个类别中的每个类别与所述数据模式中除c_i以外的所有数据项求并集，得到k个并集，其中包含c_i的并集即为所述数据模式；根据所述k个类别的真实支持度估计值，以及k个并集在数据中的支持度观测值，计算所述数据模式的真实支持度估计值；

[0040] 根据所述统计检验所涉及数据模式的真实支持度估计值，计算所述统计检验的第一参数估计真值、第二参数估计真值、第三参数估计真值以及第四参数估计真值，以对第一参数观测值、第二参数观测值、第三参数观测值以及第四参数观测值受到数据误差的影响进行修正；

[0041] 根据所述第一参数估计真值、第二参数估计真值、第三参数估计真值以及第四参数估计真值计算所述检验统计量p的值。

[0042] 与现有技术相比，本发明实施例的有益效果是：基于统计健全检验法，在将族错误率控制在较低水平的前提下，修正随机数据误差对统计检验运算的影响，由此显著恢复由于随机数据误差引起的统计检验结果中真实规则的丢失，大大提高了关联规则挖掘结果的可靠性。

附图说明

[0043] 为了更清楚地说明本发明实施例中的技术方案，下面将对实施例或现有技术描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图仅仅是本发明的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动性的前提下，还可以根据这些附图获得其他的附图。

[0044] 图1是本发明实施例提供的顾及数据不确定性的关联规则显著性检验方法的实现流程图；

[0045] 图2是本发明实施例提供的顾及数据不确定性的关联规则显著性检验方法步骤S104的具体实现流程图；

[0046] 图3是本发明实施例提供的顾及数据不确定性的关联规则显著性检验方法中用σ(s(cj))和z控制确定 $\hat{E}(s(c_j))$ 时高估E(s(cj))的概率为任意值的示意图；

[0047] 图4是本发明实施例提供的顾及数据不确定性的关联规则显著性检验装置的结构框图。

具体实施方式

[0048] 为了使本发明的目的、技术方案及优点更加清楚明白，以下结合附图及实施例，对本发明进行进一步详细说明。应当理解，此处所描述的具体实施例仅仅用以解释本发明，并不用于限定本发明。

[0049] 图1示出了本发明实施例提供的顾及数据不确定性的关联规则显著性检验方法的实现流程图，参照图1：

[0050] 在步骤S101中，获取关联规则；

[0051] 在步骤S102中，判断获取的所述关联规则是否为高效规则，若否，执行步骤S103；若是，执行步骤S104；

[0052] 在步骤S103中，认为所述关联规则为虚假规则；

[0053] 在步骤S104中，对所述关联规则进行统计显著性检验，计算检验统计量的值；

[0054] 在步骤S105中，判断步骤S104所得检验统计量的值是否低于预设显著性水平，若是，执行步骤S106；若否，执行步骤S103；

[0055] 在步骤S106中，接受所述关联规则为真实规则。

[0056] 在本发明实施例中，逐个获取待检验的关联规则。对于获取的每一个关联规则，首先判断该关联规则是否为高效规则。若该关联规则不为高效规则，则认为该关联规则为虚假规则，并删除该关联规则。若该关联规则为高效规则，则进一步对该关联规则的高效性进行统计检验，判断所得统计量的值是否低于预设显著性水平，若是，接受该关联规则为真实规则；若否，认为该关联规则为虚假规则，并删除该关联规则。在所有关联规则检验完成后，向用户展示所有真实规则。其中，预设显著性水平α可以为0.05，在此不作限定。

[0057] 图2示出了本发明实施例提供的顾及数据不确定性的关联规则显著性检验方法步骤S104的具体实现流程图，参照图2：

[0058] 在步骤S201中，对所述统计检验涉及的每一个数据模式，将其中指定数据项所对应的属性的误差概率分布表达为误差矩阵，所述误差矩阵包括指定属性的全部k个类别之间的误差分布，其中，指定属性指的是所述指定数据项对应的属性，k为大于1的整数。

[0059] 在本发明实施例中，将数据视为分类数据。分类数据是关联规则挖掘中最常用的两种数据之一，另一种最常用的事务数据很容易转换为分类数据，而定量数据通常先分类为分类数据再用于关联规则挖掘。

[0060] 作为本发明的一个实施例，指定属性a有k个类别1，…，k，用数据项c_1，…，c_k表示。当一条记录中a的真实分类为j时，a的值被记录为i的概率为p_{ij}，i，j∈[1，k]，则a的误差矩阵为

[0061] $$\mathbf{P}=\begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1k} \\ p_{21} & p_{22} & \cdots & p_{2k} \\ \cdots & \cdots & \cdots & \cdots \\ p_{k1} & p_{k2} & \cdots & p_{kk} \end{pmatrix}$$

[0062] P主对角线上的元素表示i＝j,即正确记录各分类的概率,其他元素均为各种数据与真实分类不符,即误差发生情况的概率。根据不确定关联规则挖掘的常用简化假设——各数据项的不确定概率表现相互独立,正确或错误记录a属性值的各种情况,其可能性在所有记录中相同,与记录中其他属性的值无关。因此,可以用单一的P对a在全体数据中的误差传播进行建模。

[0063] 在步骤S202中,根据所述误差矩阵,对数据误差的传播进行建模,得到所述k个类别的观测支持度分布期望及方差。

[0064] 对表示类别i的数据项c_i,其观测支持度$s(c_i)$为数据中包含c_i的记录条数,而其真实支持度$s_0(c_i)$为实际包含c_i的记录条数,在现实中不可知。$s(c_i)$与$s_0(c_i)$的差异即为随机数据误差的影响。对a的真值为j的$s_0(c_j)$条记录,每条记录中a的值被误记录为i是一个概率为p_{ij}的伯努利实验。因此,数据中a的真值为j,而记录值为i的记录条数$s(c_j\rightarrow c_i)$服从二项分布:$s(c_j\rightarrow c_i)\sim B(s_0(c_j),p_{ij})$。由于关联规则挖掘中$s_0(c_j)$,$s_0(c_j)p_{ij}$和$s_0(c_j)(1-p_{ij})$均较大,该分布可近似为正态分布:$s(c_j\rightarrow c_i)\sim N(s_0(c_j)p_{ij},s_0(c_j)p_{ij}(1-p_{ij}))$。因$s(c_i)=\sum_{j=1}^{k}s(c_j\rightarrow c_i)$,而$s(c_1\rightarrow c_i),\cdots,s(c_k\rightarrow c_i)$相互独立,因此$s(c_i)$也近似服从正态分布,该分布的期望和方差为

[0065] $$E(s(c_i))=\sum_{j=1}^{k}p_{ij}s_0(c_j)$$

[0066] $$\sigma^2(s(c_i))=\sum_{j=1}^{k}p_{ij}(1-p_{ij})s_0(c_j)$$

[0067] 所有k个类别的观测支持度分布期望可以合写为

[0068] $$\begin{pmatrix} E(s(c_1)) \\ \vdots \\ E(s(c_k)) \end{pmatrix}=\begin{pmatrix} p_{11} & \cdots & p_{1k} \\ \vdots & \ddots & \vdots \\ p_{k1} & \cdots & p_{kk} \end{pmatrix}\begin{pmatrix} s_0(c_1) \\ \vdots \\ s_0(c_k) \end{pmatrix}$$

[0069] $E(S(a))=PS_0(a)$

[0070] $$\mathbf{\Sigma(S}(a))=\begin{pmatrix} \sigma(s(c_1)) \\ \vdots \\ \sigma(s(c_k)) \end{pmatrix}=\begin{pmatrix} (p_{11}(1-p_{11}))^{\frac{1}{2}} & \cdots & (p_{1k}(1-p_{1k}))^{\frac{1}{2}} \\ \vdots & \ddots & \vdots \\ (p_{k1}(1-p_{k1}))^{\frac{1}{2}} & \cdots & (p_{kk}(1-p_{kk}))^{\frac{1}{2}} \end{pmatrix}\begin{pmatrix} (s_0(c_1))^{\frac{1}{2}} \\ \vdots \\ (s_0(c_k))^{\frac{1}{2}} \end{pmatrix}$$

[0071] 在步骤S203中,根据所估计的k个类别的观测支持度分布以及所述误差矩阵,计算所述k个类别的真实支持度估计值。

[0072] 在步骤S204中,以c_i表示所述统计检验涉及的数据模式中的指定数据项,将所述k

个类别中的每个类别与所述数据模式中除c_i以外的所有数据项求并集，得到k个并集，其中包含c_i的并集即为所述数据模式；根据所述k个类别的真实支持度估计值，以及k个并集在数据中的支持度观测值，计算所述数据模式的真实支持度估计值。

[0073] $$\begin{pmatrix} E(s(c_1)) \\ \vdots \\ E(s(c_k)) \end{pmatrix} = \begin{pmatrix} p_{11} & \cdots & p_{1k} \\ \vdots & \ddots & \vdots \\ p_{k1} & \cdots & p_{kk} \end{pmatrix} \begin{pmatrix} s_0(c_1) \\ \vdots \\ s_0(c_k) \end{pmatrix}$$

[0074] $E(S(a)) = PS_0(a)$

[0075] 等同于$S_0(a) = P^{-1}E(S(a))$。观测支持度分布期望$E(S(a))$的值由P和$S_0(a)$决定，$S_0(a)$为现实中未知的所有类别的真实支持度，因此观测支持度分布期望$E(S(a))$也未知。若能确定观测支持度分布期望$E(S(a))$的观测支持度分布期望估计值$\hat{\mathbf{E}}(\mathbf{S}(a))$，则可得真实支持度$S_0(a)$的真实支持度估计值$\hat{\mathbf{S}}_0(a)$：

[0076] $$\hat{\mathbf{S}}_0(a) = \mathbf{P}^{-1}\hat{\mathbf{E}}(\mathbf{S}(a)).$$

[0077] 展开$\hat{\mathbf{S}}_0(a) = \mathbf{P}^{-1}\hat{\mathbf{E}}(\mathbf{S}(a))$并取其第i行，可得类别i的真实支持度估计值$\hat{s}_0(c_i)$：

[0078] $$\hat{s}_0(c_i) = \sum_{j=1}^{k} p_{ij}^{-1}\hat{E}\left(s(c_j)\right)$$

[0079] 其中p_{ij}^{-1}为P^{-1}在(i,j)位置上的元素值。

[0080] 根据对$s_0(c_i)$进行估值的目的不同，$\hat{E}\left(s(c_j)\right)$大于或小于实际$E(s(c_j))$的概率，也即$E(s(c_j))$被高估或低估的概率，可能需要为(0,1)间的任意值。对此，可取$\hat{E}\left(s(c_j)\right) = s(c_j) - z\sigma\left(s(c_j)\right)$，z为常量，此时我们将$s(c_j)$视为$E(s(c_j)) + z\sigma(s(c_j))$，而事实上$s(c_j) > E(s(c_j)) + z\sigma(s(c_j))$的概率为$1-\Phi(z)$，Φ为标准正态分布的累计密度函数。$\hat{E}\left(s(c_j)\right)$大于实际$E(s(c_j))$，即$E(s(c_j))$被高估的情况等同于$s(c_j) > E(s(c_j)) + z\sigma(s(c_j))$，其概率也为$1-\Phi(z)$，如图3所示。

[0081] 将$\hat{s}_0(c_i) = \sum_{j=1}^{k} p_{ij}^{-1}\hat{E}\left(s(c_j)\right)$中$\hat{E}\left(s(c_j)\right)$替换为$s(c_j) - z\sigma(s(c_j))$，再用$\sigma^2(s(c_i)) = \sum_{j=1}^{k} p_{ij}(1-p_{ij})s_0(c_j)$代换$\sigma(s(c_j))$，有

[0082] $$\hat{s}_0(c_i) = \sum_{j=1}^{k}\left(p_{ij}^{-1}\left(s(c_j) - z\left(\sum_{l=1}^{k} p_{jl}(1-p_{jl})s_0(c_l) \right)^{1/2} \right) \right)$$

[0083] $S_0(c_1)$也是未知的真值，应替换为估计值$\hat{s}_0(c_l)$：

[0084] $$\hat{s}_0(c_i)=\sum_{j=1}^{k}\left(p_{ij}^{-1}\left(s(c_j)-z\left(\sum_{l=1}^{k}p_{jl}(1-p_{jl})\hat{s}_0(c_l)\right)^{1/2}\right)\right)$$

[0085] 对全部类别的真实支持度估计值 $\hat{s}_0(c_1),\cdots,\hat{s}_0(c_k)$ 各写出形如 $\hat{s}_0(c_i)=\sum_{j=1}^{k}\left(p_{ij}^{-1}\left(s(c_j)-z\left(\sum_{l=1}^{k}p_{jl}(1-p_{jl})\hat{s}_0(c_l)\right)^{1/2}\right)\right)$ 的等式，所有等式联立可解出 $\hat{s}_0(c_1),\cdots,\hat{s}_0(c_k)$。但此解法比较繁琐，且仅需一个 $\hat{s}_0(c_i)$ 时也必须解出全部 $\hat{s}_0(c_1),\cdots,\hat{s}_0(c_k)$ ，浪费运算时间。事实上，$\hat{s}_0(c_i)=\sum_{j=1}^{k}\left(p_{ij}^{-1}\left(s(c_j)-z\left(\sum_{l=1}^{k}p_{jl}(1-p_{jl})\hat{s}_0(c_l)\right)^{1/2}\right)\right)$ 右侧的 $\hat{s}_0(c_l)$ 可以用观测支持度s(c$_l$)来近似，这对所得 $\hat{s}_0(c_i)$ 的影响很小：

[0086] $$\hat{s}_0(c_i)=\sum_{j=1}^{k}\left(p_{ij}^{-1}\left(s(c_j)-z\left(\sum_{l=1}^{k}p_{jl}(1-p_{jl})s(c_l)\right)^{1/2}\right)\right)。$$

[0087] 在步骤S205中，根据所述统计检验涉及的数据模式的真实支持度估计值，计算所述统计检验的第一参数估计真值、第二参数估计真值、第三参数估计真值以及第四参数估计真值，以对第一参数观测值、第二参数观测值、第三参数观测值以及第四参数观测值受到数据误差的影响进行修正。

[0088] 令I为a以外的N个属性的集合，先将I视为无随机发生的数据误差，若存在误差则将各个存在误差的数据项比照c_i逐一处理。设I∪{c_i}的不含c_i误差的真实支持度为s_0(I∪{c_i})，而观测支持度为s(I∪{c_i})。基于各数据项不确定概率表现相互独立的假设，若将 $\hat{s}_0(c_i)=\sum_{j=1}^{k}\left(p_{ij}^{-1}\left(s(c_j)-z\left(\sum_{l=1}^{k}p_{jl}(1-p_{jl})s(c_l)\right)^{1/2}\right)\right)$ 中的c_i替换为I∪{c_i}，等式同样成立。因此，记由P和z确定的、s(I∪c_i)的估计真值为 $\hat{E}(c_i,I,\mathbf{P},z)$，有

[0089] $$\begin{aligned}\hat{E}(c_i,I,\mathbf{P},z)&=\hat{s}_0(I\cup\{c_i\})\\&=\sum_{j=1}^{k}\left(p_{ij}^{-1}\left(s(I\cup\{c_j\})-z\left(\sum_{l=1}^{k}p_{jl}(1-p_{jl})s(I\cup\{c_l\})\right)^{1/2}\right)\right)。\end{aligned}$$

[0090] 费氏精确检验中的四个关键计算参数a,b,c,d可改写为

[0091] a＝s(X∪{y})

[0092] b＝s(X)-s(X∪{y})

[0093] c＝s((X-{x_m})∪{y})-s(X∪{y})，

[0094] d＝s(X-{x_m})-s(X)-s((X-{x_m})∪{y})+s(X∪{y})

[0095] 其中a表示第一参数，b表示第二参数，c表示第三参数，d表示第四参数，x_m为被检验是否冗余的项，x_m∈X，s表示各数据模式的观测支持度。设a～d的真值(无随机数据误差

影响)为a_0,b_0,c_0,d_0,根据 $\begin{aligned} a &= s(X\cup\{y\}) \\ b &= s(X)-s(X\cup\{y\}) \\ c &= s((X-\{x_m\})\cup\{y\})-s(X\cup\{y\}) \\ d &= s(X-\{x_m\})-s(X)-s((X-\{x_m\})\cup\{y\})+s(X\cup\{y\}) \end{aligned}$ 所示的各关键计算参数的内容,可变化I和c_i的值,将

$$\hat{E}(c_i,I,\mathbf{P},z)=\hat{s}_0(I\cup\{c_i\})=\sum_{j=1}^{k}\left(p_{ij}^{-1}\left(s(I\cup\{c_j\})-z\left(\sum_{l=1}^{k}p_{jl}(1-p_{jl})s(I\cup\{c_l\})\right)^{1/2}\right)\right)$$

应用于a～d,得其估计真值$\hat{a}_0$,$\hat{b}_0$,$\hat{c}_0$,$\hat{d}_0$。$\hat{a}_0\sim\hat{d}_0$受误差的影响小于a～d,故使用$\hat{a}_0\sim\hat{d}_0$代替a～d计算检验值,可使检验结果更加准确。

[0096] 在步骤S206中,根据所述第一参数估计真值、第二参数估计真值、第三参数估计真值以及第四参数估计真值计算所述检验统计量p的值,即在计算检验统计量 $p=\sum_{i=0}^{\min(b,c)}\frac{(a+b)!(c+d)!(a+c)!(b+d)!}{(a+b+c+d)!(a+i)!(b-i)!(c-i)!(d+i)!}$ 时,使用$\hat{a}_0\sim\hat{d}_0$的值代替a～d。

[0097] 本发明实施例提供了基于统计健全检验法的修正方法,根据统计学原理和误差传播定律,建立数学模型来描述随机数据误差在统计检验中的传播,直至对统计检验所用的关键计算参数(第一参数、第二参数、第三参数以及第四参数)的影响。根据所建立的数学模型以及已知的随机数据误差水平可以得到关键计算参数的修正量,即相对于存在随机数据误差的数据中的观测值而言,关键计算参数的估计真值。关键计算参数的估计真值比观测值更接近真值,因此用关键计算参数的估计真值代替观测值计算检验值,可以使计算结果更加准确,有利于增加真实规则。

[0098] 优选地,步骤S205中在所述根据所述统计检验所涉及数据模式的真实支持度估计值,计算第一参数估计真值、第二参数估计真值、第三参数估计真值以及第四参数估计真值时,所述方法还包括:

[0099] 使用经过随机化处理的数据进行模拟的关联规则提取,求出使所述统计检验的族错误率小于指定上限的最佳参数修正量,其中,所述最佳参数修正量为非负数;

[0100] 将所述最佳参数修正量用于计算所述第一参数估计真值以及第四参数估计真值;

[0101] 将所述最佳参数修正量的相反数用于计算所述第二参数估计真值以及第三参数估计真值。

[0102] 计算第一参数估计真值、第二参数估计真值、第三参数估计真值以及第四参数估计真值时,还需要根据用户要求的所述统计检验错误接受虚假规则的风险上限值(即指定上限),确定一最佳参数修正量。确定最佳参数修正量后,应将最佳参数修正量用于计算所述第一参数估计真值以及第四参数估计真值,而将最佳参数修正量的相反数用于计算所述第二参数估计真值以及第三参数估计真值。

[0103] 由 $p=\sum_{i=0}^{\min(b,c)}\frac{(a+b)!(c+d)!(a+c)!(b+d)!}{(a+b+c+d)!(a+i)!(b-i)!(c-i)!(d+i)!}$ 可知,当a、d值增大或b、c值减小时,p值减小,导致真实规则和虚假规则均更可能通过检验。为了不增加虚假规则,最佳参数修正量不能令a、d增大或b、c减小,因此应使用非负的最佳参数修正量,并用

$\hat{E}(c_i, I, \mathbf{P}, z)$修正a、d，用$\hat{E}(c_i, I, \mathbf{P}, -z)$修正b、c。

[0104] 使用经过随机化处理的数据进行模拟的关联规则提取，求出最佳参数修正量，使在所述统计检验错误接受虚假规则的风险小于用户要求上限的前提下，统计检验有能力发现最多的正确规则。

[0105] 优选地，在所述求出使所述统计检验的族错误率小于指定上限的最佳参数修正量的过程中，所述方法还包括：

[0106] 对数据中每个属性在所有记录中的类别进行n次随机排列，其中，n为大于1的整数；

[0107] 对每一次随机排列，从随机排列后的数据中获取关联规则，取参数修正量z为0，对获取的所述关联规则进行统计检验，并逐渐增大z值，直至所有所述关联规则均被判定为虚假规则，并记录此时的z值；

[0108] 将n次数据随机排列所得到的n个z值中最大者作为所述最佳参数修正量。

[0109] 等式 $\hat{E}(c_i, I, \mathbf{P}, z) = \hat{s}_0(I \cup \{c_i\}) = \sum_{j=1}^{k}\left(p_{ij}^{-1}\left(s(I \cup \{c_j\}) - z\left(\sum_{l=1}^{k} p_{jl}(1-p_{jl}) s(I \cup \{c_l\}) \right)^{1/2} \right) \right)$

中的最佳参参数修正量z是控制统计检验关键计算参数修正程度的关键。z值越小，修正程度越大，使修正检验有能力发现更多真实规则，但也增大了过度修正的可能和最终产生虚假规则的风险。如果能分析得出族错误率和z值之间的定量关系，就可以根据用户给定的族错误率上限，直接确定所需的z值。但族错误率和z值的关系极度复杂，受到误差分布和数据本身的诸多不确定因素影响，几乎不可能将这些影响全部定量化，而对任何一种影响估计得很不准确，就无法确定合理的z值。由于难以对确定修正参数所需的z值进行上述定量分析，在本发明实施例中使用以下模拟法作为替代方案来确定z值，使真实规则得到最大程度的增加，同时族错误率不超过用户给定的指定上限r_{max}。模拟法步骤如下：

[0110] 第一步，对数据表中每一列即每一属性，将该列所有属性值随机重新排序；

[0111] 第二步，使用关联规则挖掘算法提取步骤一所得随机化数据中的关联规则，用修正方法检验所得关联规则，先取z＝0，逐渐增加z值，直到所有关联规则都被拒绝，即不能通过检验；

[0112] 第三步，将第一步和第二步重复n次，找到n次中最大的令所有关联规则被拒绝的z值。

[0113] 第一步所得的随机化数据中，各数据项支持度(数量)与实际数据相同，但失去了所有数据项间的关联。因此，从随机化数据中发现的任何关联规则均为虚假规则。除失去关联外，随机化数据保存了实际数据中的其他特征，这些特征可以用来模拟族错误率和z值关系的诸多不确定影响因素。因此，将第三步所得的最大z值用于检验从实际数据中提取的关联规则，族错误率应与模拟过程中的值处于同一水平。

[0114] 循环数n由r_{max}确定。每个循环可以看作无限种数据随机化可能情况中的一个抽样，如果每次随机化后检验中接受至少一条虚假规则的概率为r_{max}，则在n个“抽样”循环中，接受不多于一条虚假规则的概率为

[0115] $$\begin{aligned}\Pr(K\le 1)&=\Pr(K=0)+\Pr(K=1)\\&=C_n^0 r_{\max}{}^0(1-r_{\max})^{n-0}+C_n^1 r_{\max}{}^1(1-r_{\max})^{n-1}\\&=(1-r_{\max})^n+nr_{\max}(1-r_{\max})^{n-1}\end{aligned},$$

[0116] K表示接受虚假规则的数量。所需n值为令Pr(K≤1)≤0.5的最小正整数，也就是说，当数据误差在模拟中呈现平均程度的影响(概率为0.5)时，族错误率不高于r_{max}。当给定r_{max}为0.05时，所需循环数为n＝34。虽然z值可以使检验拒绝所有规则，但z值再减少一个递增时的最小单位量，就会产生虚假规则，因此计算中应包括Pr(K＝1)。

[0117] 需要说明的是，模拟法中检验结果的族错误率取决于r_{max}，而非检验所用的预设显著性水平κ。不过，因为取预设显著性水平κ＝α/s和采用模拟法的目的均为使族错误率低于用户给定的上限(r_{max}或α)，因此，r_{max}和α一般应取相同的值，如0.05。

[0118] 在步骤S205所述根据所述统计检验所涉及数据模式的真实支持度估计值，计算第一参数估计真值、第二参数估计真值、第三参数估计真值以及第四参数估计真值时，所述方法还包括：

[0119] 根据有误差的数据项c_i在所述关联规则中的位置不同，采取不同的修正数学式计算所述第一参数估计真值、第二参数估计真值、第三参数估计真值以及第四参数估计真值。

[0120] 对规则X→y，误差可能发生在三种位置：x_m，y或某个x_m以外的项目$x_e\in X$。这三种情况下，$\hat{a}_0\sim\hat{d}_0$需要三套不同的公式化表示。

[0121] 当误差项c_i在关联规则中的位置为$c_i=x_m$时：

[0122] $$\hat{a}_0=\hat{E}(c_i,(X-\{x_m\})\cup\{y\},\mathbf{P},z),$$

[0123] $$\hat{b}_0=\hat{E}(c_i,X-\{x_m\},\mathbf{P},-z)-\hat{E}(c_i,(X-\{x_m\})\cup\{y\},\mathbf{P},-z),$$

[0124] $$\hat{c}_0=a+c-\hat{a}_0,$$

[0125] $$\hat{d}_0=b+d-\hat{b}_0。$$

[0126] 当误差项c_i在关联规则中的位置为$c_i=y$时：

[0127] $$\hat{a}_0=\hat{E}(c_i,X,\mathbf{P},z),$$

[0128] $$\hat{b}_0=a+b-\hat{a}_0,$$

[0129] $$\hat{c}_0=\hat{E}(c_i,X-\{x_m\},\mathbf{P},-z)-\hat{E}(c_i,X,\mathbf{P},-z),$$

[0130] $$\hat{d}_0=c+d-\hat{c}_0。$$

[0131] 当误差项c_i在关联规则中的位置为$c_i=x_e$，$x_e\in X-\{x_m\}$时：

[0132] $$\hat{a}_0=\hat{E}(c_i,(X-\{x_e\})\cup\{y\},\mathbf{P},z),$$

[0133] $$\hat{b}_0=\hat{E}(c_i,X-\{x_e\},\mathbf{P},-z)-\hat{E}(c_i,(X-\{x_e\})\cup\{y\},\mathbf{P},-z),$$

[0134] $$\hat{c}_0=\hat{E}(c_i,(X-\{x_m\}-\{x_e\})\cup\{y\},\mathbf{P},-z)-\hat{E}(c_i,(X-\{x_e\})\cup\{y\},\mathbf{P},-z),$$

[0135]

$$\hat{d}_0 = \hat{E}\left(c_i, X-\{x_m\}-\{x_e\}, \mathbf{P}, z\right) - \hat{E}\left(c_i, X-\{x_e\}, \mathbf{P}, z\right) - \hat{E}\left(c_i, (X-\{x_m\}-\{x_e\})\cup\{y\}, \mathbf{P}, z\right) + \hat{E}\left(c_i, (X-\{x_e\})\cup\{y\}, \mathbf{P}, z\right)$$

[0136] 最后，使用第一参数估计真值、第二参数估计真值、第三参数估计真值以及第四参数估计真值取代原统计检验中的四个关键参数值，计算检验统计量p的值，以修正数据误差对所得p值的影响。

[0137] 进一步地，所述根据所述第一参数估计真值、第二参数估计真值、第三参数估计真值以及第四参数估计真值计算所述检验统计量p的值，其具体过程为：

[0138] 将所述第一参数估计真值、第二参数估计真值、第三参数估计真值以及第四参数估计真值用于健全统计检验法，计算所述检验统计量p的值。

[0139] 本发明实施例提供的顾及数据不确定性的关联规则显著性检验方法能明显提高关联规则挖掘结果的可靠性，在随机数据误差存在的普遍情况下，增加真实规则，严格控制虚假规则，使挖掘结果在数据分析和决策支持中更具价值。

[0140] 本发明实施例基于独创误差传播模型的统计检验参数修正，可以减少随机数据误差对统计检验计算结果的影响，弥补高达近60％由于随机数据误差造成的真实规则损失。最有实际意义的关联规则往往对误差非常敏感，此时本发明实施例就尤其有效。同时，使用模拟过程控制修正程度的机制，使虚假规则数量接近统计健全检验法达到的极低水平(族错误率<5％)，明显优于绝大部分其他滤除虚假规则的方法(减少虚假规则比例，但族错误率仍接近100％)。

[0141] 本发明实施例已在合成数据和真实数据实验中得到验证和应用。合成数据试验的数据为计算机根据预先设计的、已知的真实规则生成，因此可以明确判断检验结果中的真实与虚假规则。在低至2％，高至36％记录包含误差的多种误差水平，以及多种数据量的情况下，运用本发明实施例提供的修正方法均比原始统计健全检验法发现更多的真实规则。修正方法的效果可以用恢复率来表示：恢复率＝(修正方法发现的真实规则数-原始方法发现的真实规则数)/(无随机误差数据中发现的真实规则数-原始方法发现的真实规则数)×100％。原始方法和修正方法均指应用于有随机数据误差的情况。在各误差水平下，修正方法的平均恢复率约为58％。修正方法得到的虚假规则虽也高于原始方法，但平均族错误率仅为2％，最差情况即最高误差水平下也不过5％。增加的真实规则与虚假规则数量比例约为130:1。

[0142] 真实数据实验的数据为土地利用和人口、收入等社会经济指标在1985～1999年的变化。真实数据中的真实规则未知，而模拟实验证明，统计健全检验从无误差数据中发现的真实规则族错误率不到1％，因此借用无误差数据中发现的关联规则作为真实规则，来评估原始方法和修正方法用于有误差数据的结果。在多种误差水平下，修正方法均发现更多的真实规则。其中，包含两个年份土地利用变化(利用类型不同)的规则最有实际意义，但仅有约100条，且对误差非常敏感。原始方法导致45％～85％此类真实规则的丢失，而修正方法发现的真实规则为原始方法的2～4倍。现实中的关联规则挖掘经常与本实验相似：最重要的规则数量稀少，且对误差敏感，因此修正方法具有很高的潜在实用价值。

[0143] 应理解，在本发明实施例中，上述各过程的序号的大小并不意味着执行顺序的先后，各过程的执行顺序应以其功能和内在逻辑确定，而不应对本发明实施例的实施过程构

成任何限定。

[0144] 本发明实施例基于统计健全检验法，在将族错误率控制在较低水平的前提下，修正随机数据误差对统计检验运算的影响，由此显著恢复由于随机数据误差引起的统计检验结果中真实规则的丢失，大大提高了关联规则挖掘结果的可靠性。

[0145] 图4示出了本发明实施例提供的顾及数据不确定性的关联规则显著性检验装置的结构框图，该装置可以用于运行图1或图2所述的顾及数据不确定性的关联规则显著性检验方法。为了便于说明，仅示出了与本发明实施例相关的部分。参照图4，所述装置包括：

[0146] 高效规则判断单元41，用于获取关联规则，并判断获取的所述关联规则是否为高效规则；

[0147] 虚假规则判定单元42，用于若所述关联规则不为所述高效规则，则认为所述关联规则为虚假规则；

[0148] 检验单元43，用于若所述关联规则为所述高效规则，则对所述关联规则进行统计检验，并判断所得检验统计量p的值是否低于预设显著性水平，若是，则接受所述关联规则为真实规则；若否，则认为所述关联规则为虚假规则；所述统计检验涉及的每一个数据模式为若干数据项的集合，每个数据项指的是数据中一个属性中的一个类别，每个属性的误差概率分布为已知；

[0149] 检验单元43包括检验统计量值计算子单元431，检验统计量值计算子单元431具体用于：

[0150] 对所述统计检验涉及的每一个数据模式，将其中指定数据项c_i所对应的属性的误差概率分布表达为误差矩阵，所述误差矩阵包括所述指定属性的全部k个类别之间的误差分布，其中，指定属性指的是所述指定数据项对应的属性，k为大于1的整数；

[0151] 根据所述误差矩阵，对数据误差的传播进行建模，得到所述k个类别的观测支持度分布期望及方差；

[0152] 根据所估计的k个类别的观测支持度分布以及所述误差矩阵，计算所述k个类别的真实支持度估计值；

[0153] 以c_i表示所述统计检验涉及的数据模式中的指定数据项，将所述k个类别中的每个类别与所述数据模式中除c_i以外的所有数据项求并集，得到k个并集，其中包含c_i的并集即为所述数据模式；根据所述k个类别的真实支持度估计值，以及k个并集在数据中的支持度观测值，计算所述数据模式的真实支持度估计值；

[0154] 根据所述统计检验所涉及数据模式的真实支持度估计值，计算所述统计检验的第一参数估计真值、第二参数估计真值、第三参数估计真值以及第四参数估计真值，以对第一参数观测值、第二参数观测值、第三参数观测值以及第四参数观测值受到数据误差的影响进行修正；

[0155] 根据所述第一参数估计真值、第二参数估计真值、第三参数估计真值以及第四参数估计真值计算所述检验统计量p的值。

[0156] 优选地，根据实行检验统计量值计算子单元431的需求，所述装置还包括检验参数修正单元44，检验参数修正单元44用于：

[0157] 使用经过随机化处理的数据进行模拟的关联规则提取，求出使所述统计检验的族错误率小于指定上限的最佳参数修正量，其中，所述最佳参数修正量为非负数；

[0158] 将所述最佳参数修正量用于计算所述第一参数估计真值以及第四参数估计真值；

[0159] 将所述最佳参数修正量的相反数用于计算所述第二参数估计真值以及第三参数估计真值。

[0160] 根据实行检验参数修正单元44的需求，所述装置还包括最佳参数修正量确定单元45，最佳参数修正量确定单元45用于：

[0161] 对数据中每个属性在所有记录中的类别进行n次随机排列，其中，n为大于1的整数；

[0162] 对每一次随机排列，从随机排列后的数据中获取关联规则，取参数修正量z为0，对获取的所述关联规则进行统计检验，并逐渐增大z值，直至所有所述关联规则均被判定为虚假规则，并记录此时的z值；

[0163] 将n次数据随机排列所得到的n个z值中最大者作为所述最佳参数修正量。

[0164] 进一步地，所述检验参数修正单元44还用于：

[0165] 根据c_i在所述关联规则中所处的的位置，获取与所述位置对应的修正数学式计算所述第一参数估计真值、第二参数估计真值、第三参数估计真值以及第四参数估计真值。

[0166] 进一步地，检验统计量值计算子单元431在检验参数修正单元44、所述装置还包括最佳参数修正量确定单元45的辅助下，获取所述第一参数估计真值、第二参数估计真值、第三参数估计真值以及第四参数估计真值后，检验统计量值计算子单元431还用于：

[0167] 将所述第一参数估计真值、第二参数估计真值、第三参数估计真值以及第四参数估计真值用于健全统计检验法，计算所述检验统计量p的值。

[0168] 本领域普通技术人员可以意识到，结合本文中所公开的实施例描述的各示例的单元及算法步骤，能够以电子硬件、或者计算机软件和电子硬件的结合来实现。这些功能究竟以硬件还是软件方式来执行，取决于技术方案的特定应用和设计约束条件。专业技术人员可以对每个特定的应用来使用不同方法来实现所描述的功能，但是这种实现不应认为超出本发明的范围。

[0169] 所属领域的技术人员可以清楚地了解到，为描述的方便和简洁，上述描述的装置和单元的具体工作过程，可以参考前述方法实施例中的对应过程，在此不再赘述。

[0170] 在本申请所提供的几个实施例中，应该理解到，所揭露的装置和方法，可以通过其它的方式实现。例如，以上所描述的装置实施例仅仅是示意性的，例如，所述单元的划分，仅仅为一种逻辑功能划分，实际实现时可以有另外的划分方式，例如多个单元或组件可以结合或者可以集成到另一个系统，或一些特征可以忽略，或不执行。另一点，所显示或讨论的相互之间的耦合或直接耦合或通信连接可以是通过一些接口，单元的间接耦合或通信连接，可以是电性，机械或其它的形式。

[0171] 所述作为分离部件说明的单元可以是或者也可以不是物理上分开的，作为单元显示的部件可以是或者也可以不是物理单元，即可以位于一个地方，或者也可以分布到多个网络单元上。可以根据实际的需要选择其中的部分或者全部单元来实现本实施例方案的目的。

[0172] 另外，在本发明各个实施例中的各功能单元可以集成在一个处理单元中，也可以是各个单元单独物理存在，也可以两个或两个以上单元集成在一个单元中。

[0173] 所述功能如果以软件功能单元的形式实现并作为独立的产品销售或使用时，可以

存储在一个计算机可读取存储介质中。基于这样的理解，本发明的技术方案本质上、或者说对现有技术做出贡献的部分、或者该技术方案的部分可以以软件产品的形式体现出来，该软件产品存储在一个存储介质中，包括若干指令用以使得一台计算机设备(可以是个人计算机，服务器，网络设备等)执行本发明各个实施例所述方法的全部或部分步骤。而前述的存储介质包括：U盘、移动硬盘、只读存储器(ROM，Read-Only Memory)、随机存取存储器(RAM，Random Access Memory)、磁碟或者光盘等各种可以存储程序代码的介质。

[0174] 以上所述，仅为本发明的具体实施方式，但本发明的保护范围并不局限于此，任何熟悉本技术领域的技术人员在本发明揭露的技术范围内，可轻易想到的变化或替换，都应涵盖在本发明的保护范围之内。因此，本发明的保护范围应以所述权利要求的保护范围为准。

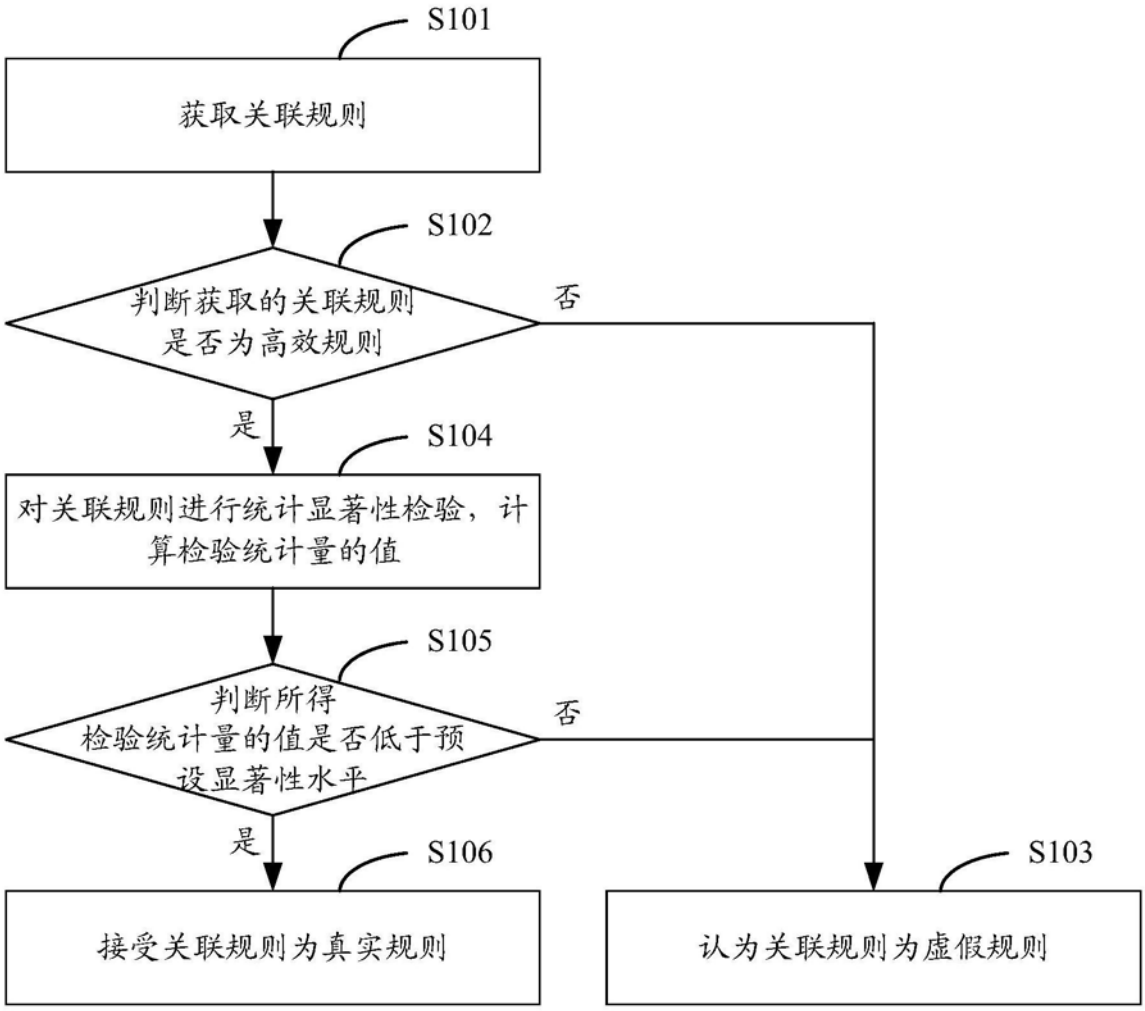

S101
获取关联规则
S102
判断获取的关联规则是否为高效规则
否
是
S104
对关联规则进行统计显著性检验，计算检验统计量的值
S105
判断所得检验统计量的值是否低于预设显著性水平
否
是
S106
接受关联规则为真实规则
S103
认为关联规则为虚假规则

图1

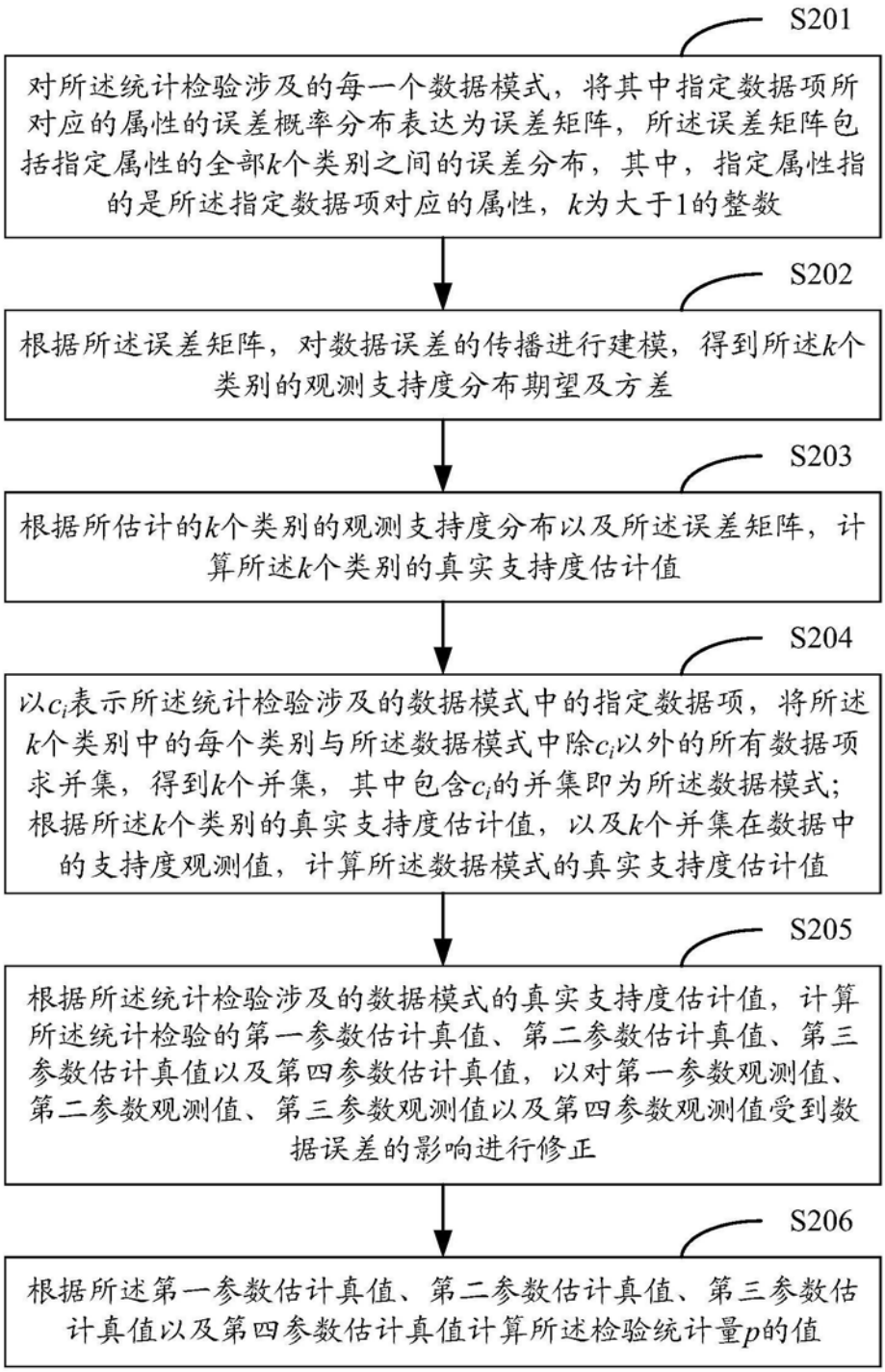

S201
对所述统计检验涉及的每一个数据模式，将其中指定数据项所对应的属性的误差概率分布表达为误差矩阵，所述误差矩阵包括指定属性的全部k个类别之间的误差分布，其中，指定属性指的是所述指定数据项对应的属性，k为大于1的整数
S202
根据所述误差矩阵，对数据误差的传播进行建模，得到所述k个类别的观测支持度分布期望及方差
S203
根据所估计的k个类别的观测支持度分布以及所述误差矩阵，计算所述k个类别的真实支持度估计值
S204
以c_i表示所述统计检验涉及的数据模式中的指定数据项，将所述k个类别中的每个类别与所述数据模式中除c_i以外的所有数据项求并集，得到k个并集，其中包含c_i的并集即为所述数据模式；根据所述k个类别的真实支持度估计值，以及k个并集在数据中的支持度观测值，计算所述数据模式的真实支持度估计值
S205
根据所述统计检验涉及的数据模式的真实支持度估计值，计算所述统计检验的第一参数估计真值、第二参数估计真值、第三参数估计真值以及第四参数估计真值，以对第一参数观测值、第二参数观测值、第三参数观测值以及第四参数观测值受到数据误差的影响进行修正
S206
根据所述第一参数估计真值、第二参数估计真值、第三参数估计真值以及第四参数估计真值计算所述检验统计量p的值

图2

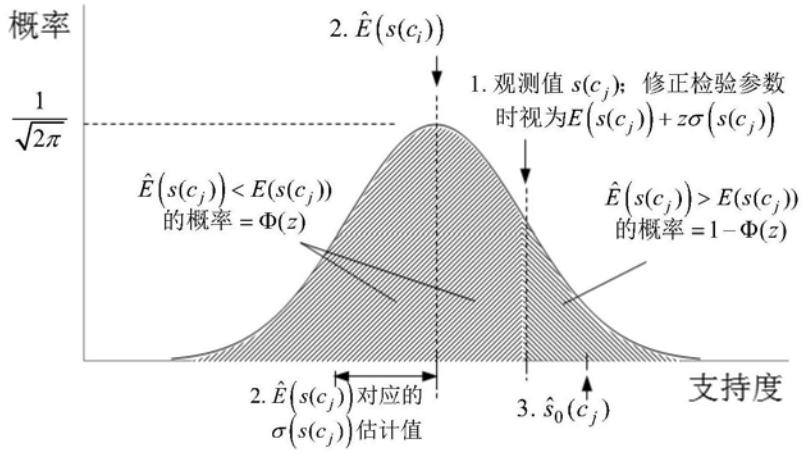

概率
$\frac{1}{\sqrt{2\pi}}$
2. $\hat{E}(s(c_i))$
1. 观测值 $s(c_j)$；修正检验参数时视为$E(s(c_j))+z\sigma(s(c_j))$
$\hat{E}(s(c_j)) < E(s(c_j))$ 的概率 $=\Phi(z)$
$\hat{E}(s(c_j)) > E(s(c_j))$ 的概率 $=1-\Phi(z)$
2. $\hat{E}(s(c_j))$ 对应的 $\sigma(s(c_j))$ 估计值
3. $\hat{s}_0(c_j)$
支持度

图3

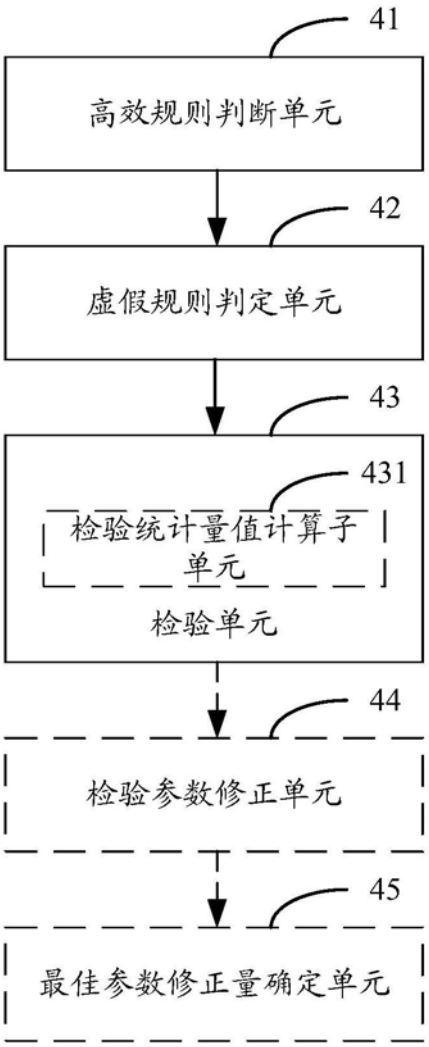

41
高效规则判断单元
42
虚假规则判定单元
43
431
检验统计量值计算子单元
检验单元
44
检验参数修正单元
45
最佳参数修正量确定单元

图4