



(12)发明专利申请

(10)申请公布号 CN 106933779 A

(43)申请公布日 2017. 07. 07

(21)申请号 201710142741.7

(22)申请日 2012.08.17

(30)优先权数据

13/213,780 2011.08.19 US

(62)分案原申请数据

201280048043.6 2012.08.17

(71)申请人 哈佛蒸汽锅炉检验和保险公司

地址 美国康涅狄格

(72)发明人 R·B·琼斯

(74)专利代理机构 中国国际贸易促进委员会专

利商标事务所 11038

代理人 边海梅

(51)Int.Cl.

G06F 17/18(2006.01)

G01N 33/00(2006.01)

权利要求书5页 说明书17页 附图8页

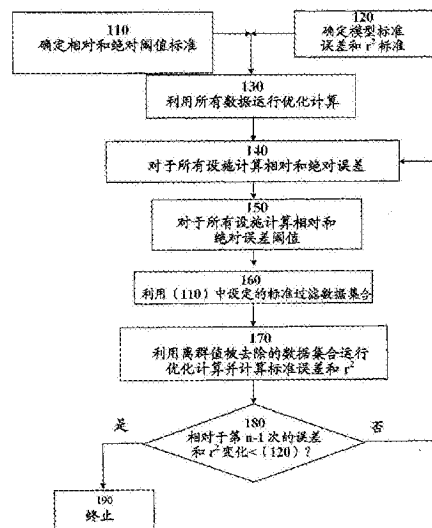
(54)发明名称

动态离群值偏倚减少系统和方法

(57)摘要

本申请涉及动态离群值偏倚减少系统和方法。这里描述了用于进行数据过滤以减少功能性和趋势线离群值偏倚的系统和方法。通过客观统计方法从数据集合中去除离群值。基于绝对误差、相对误差或两者确定偏倚。根据数据、模型系数或趋势线计算来计算误差值。当误差值大于或等于用户提供的标准时，去除离群数据记录。对于优化方法或其他迭代计算，去除的数据在每次迭代中被重应用到模型以计算新结果。利用完整数据集的模型值，计算新误差值并且重应用离群值偏倚减少过程。以迭代方式对于模型系数和离群值去除数据最小化整体误差，直到达到用户定义的误差改进限度为止。经过滤的数据可用于验证、离群值偏倚减少和数据质量操作。

数据离群值偏倚减少方法



1. 一种计算机实现的方法,所述方法包括以下步骤:

由被专门编程的计算系统以电子方式接收包括针对至少一个目标变量收集的所有实际数据值的目标数据集合;

由所述被专门编程的计算系统基于所述目标数据集合生成随机数据集合;

由所述被专门编程的计算系统以电子方式接收被用于确定一个或多个离群值的多个偏倚标准值;

由所述被专门编程的计算系统使用所述目标数据集合和模型针对所述偏倚标准值中的每个偏倚标准值生成多个离群值偏倚减少目标数据集合;

由所述被专门编程的计算系统使用所述随机数据集合和所述模型针对所述偏倚标准值中的每个偏倚标准值生成多个离群值偏倚减少随机数据集合;

由所述被专门编程的计算系统计算用于所述离群值偏倚减少目标数据集合中的每一个的至少一个目标误差值以及用于所述离群值偏倚减少随机数据集合中的每一个的至少一个随机误差值;

由所述被专门编程的计算系统计算用于所述离群值偏倚减少目标数据集合中的每一个的至少一个目标相关值以及用于所述离群值偏倚减少随机数据集合中的每一个的至少一个随机相关值;

由所述被专门编程的计算系统基于用于所述离群值偏倚减少目标数据集合中的每一个的所述至少一个目标误差值和所述至少一个目标相关值来构建第一偏倚标准曲线;

由所述被专门编程的计算系统基于用于所述离群值偏倚减少随机数据集合中的每一个的所述至少一个随机误差值和所述至少一个随机相关值来构建第二偏倚标准曲线;以及

由所述被专门编程的计算系统比较所述第一偏倚标准曲线和所述第二偏倚标准曲线以确定被用于开发所述模型的所述目标数据集合的可行性。

2. 如权利要求1所述的计算机实现的方法,其中所述随机数据集合包括在所述模型的多个预测值的范围内生成的多个随机数据值。

3. 如权利要求1所述的计算机实现的方法,还包括,由所述被专门编程的计算系统将所述第一偏倚标准曲线与所述第二偏倚标准曲线的比较翻译为指示被用于开发所述模型的所述目标数据集合的可行性的自动化建议消息。

4. 如权利要求1所述的计算机实现的方法,其中所述至少一个目标误差值是标准误差,并且其中所述至少一个目标相关值是确定系数值。

5. 如权利要求1所述的计算机实现的方法,其中,由所述被专门编程的计算系统使用所述目标数据集合和所述模型针对所述偏倚标准值中的每个偏倚标准值生成多个离群值偏倚减少目标数据集合包括:

针对所述偏倚标准值中的每个偏倚标准值:

由所述被专门编程的计算系统通过将所述模型应用到所述目标数据集合来生成用于所述目标数据集合的多个模型预测值;

由所述被专门编程的计算系统使用所述目标数据集合和所述模型预测值来计算多个误差值;

由所述被专门编程的计算系统将所述误差值与对应的偏倚标准值进行比较;

由所述被专门编程的计算系统移除所述目标数据集合内的离群值以形成从所述误差

值与所述对应的偏倚标准值的比较确定的对应的离群值偏倚减少目标数据集合;以及

由所述被专门编程的计算系统优化所述模型以基于所述对应的离群值偏倚减少目标数据集合来形成更新的模型。

6. 如权利要求5所述的计算机实现的方法,其中所述误差值包括多个相对误差值和多个绝对误差值。

7. 如权利要求5所述的计算机实现的方法,其中所述被专门编程的计算系统使用所述目标数据集合针对所述偏倚标准值中的每个偏倚标准值生成多个离群值偏倚减少目标数据集合还包括:

针对所述偏倚标准值中的每个偏倚标准值:

由所述被专门编程的计算系统将所述误差值与预定义的终止标准进行比较以确定优化所述模型的终止;以及

当所述误差值与所述预定义的终止标准的比较不表示优化所述模型的终止时,由所述被专门编程的计算系统通过将所述更新的模型应用到所述目标数据集合来生成用于所述目标数据集合的多个第二模型预测值。

8. 如权利要求1所述的计算机实现的方法,其中,由所述被专门编程的计算系统比较所述第一偏倚标准曲线和所述第二偏倚标准曲线以确定被用于开发所述模型的所述目标数据集合的可行性包括:

由所述被专门编程的计算系统基于所述至少一个目标误差值确定所述第一偏倚标准曲线上的第一偏倚标准值;

由所述被专门编程的计算系统基于所述至少一个随机误差值确定所述第二偏倚标准曲线上的第二偏倚标准值;以及

由所述被专门编程的计算系统将所述第一偏倚标准值与所述第二偏倚标准值进行比较,其中所述至少一个目标误差值和所述至少一个随机误差值是相同的。

9. 如权利要求1所述的计算机实现的方法,还包括:

通过比较若干迭代来确定所述离群值偏倚减少对每个偏倚标准值的影响以优化用于所述偏倚标准值中的每个偏倚标准值的所述更新的模型;以及

比较用于所述偏倚标准值中的每个偏倚标准值的所述至少一个目标相关值的差别。

10. 如权利要求1所述的计算机实现的方法,其中所述随机数据集合包括基于所述目标数据集合的所有随机数据值,并且其中由所述被专门编程的计算系统使用所述随机数据集合和所述模型针对所述偏倚标准值中的每个偏倚标准值生成多个离群值偏倚减少随机数据集合包括:

对于所述偏倚标准值中的每个偏倚标准值:

由所述被专门编程的计算系统通过将所述模型应用到所述随机数据集合来生成用于所述随机数据集合的多个模型预测值;

由所述被专门编程的计算系统使用所述随机数据集合和所述模型预测值来计算多个误差值;

由所述被专门编程的计算系统将所述误差值与对应的偏倚标准值进行比较;

由所述被专门编程的计算系统移除所述随机数据集合内的离群值以形成从所述误差值与所述对应的偏倚标准值的比较确定的对应的离群值偏倚减少随机数据集合;以及

由所述被专门编程的计算系统优化所述模型以基于所述对应的离群值偏倚减少随机数据集合来形成更新的模型。

11. 一种对于评估开发模型中使用的数据集合的可行性而被专门化的系统,所述系统包括:

服务器,所述服务器包括处理器和非暂态存储子系统;

由所述非暂态存储子系统存储的数据库,所述数据库包括目标数据集合,所述目标数据集合包括针对至少一个目标变量收集的所有实际数据值;

由所述非暂态存储子系统存储的计算程序,所述计算程序包括指令,所述指令当由所述处理器执行时,使得所述系统对于评估用于开发所述模型的所述目标数据集合的可行性而被专门化以至少执行以下操作:

从所述目标数据集合生成随机数据集合;

获得被用于确定一个或多个离群值的偏倚标准值集合;

针对所述偏倚标准值集合中的一个或多个偏倚标准值执行对所述目标数据集合的动态离群值偏倚减少以生成一个或多个离群值偏倚减少目标数据集合;

针对所述偏倚标准值集合中的一个或多个偏倚标准值执行对所述随机数据集合的动态离群值偏倚减少以生成一个或多个离群值偏倚减少随机数据集合;

计算用于所述一个或多个离群值偏倚减少目标数据集合的目标误差值集合以及用于所述一个或多个离群值偏倚减少随机数据集合的随机误差值集合;

计算用于所述离群值偏倚减少目标数据集合的目标相关系数集合以及用于所述离群值偏倚减少随机数据集合中的随机相关系数集合;

从所述一个或多个偏倚标准值、所述目标误差值集合、所述随机误差值集合、所述目标相关系数集合和所述随机相关系数集合构建用于所述目标数据集合的第一偏倚标准曲线和用于所述随机数据集合的第二偏倚标准曲线;以及

比较所述第一偏倚标准曲线和所述第二偏倚标准曲线以确定被用于开发所述模型的所述目标数据集合的可行性。

12. 如权利要求11所述的系统,其中所述指令当由所述处理器执行时,使得对于评估用于开发所述模型的所述目标数据集合的可行性而被专门化的所述系统将所述第一偏倚标准曲线与所述第二偏倚标准曲线的比较翻译为指示被用于开发所述模型的所述目标数据集合的可行性的自动化建议消息。

13. 如权利要求11所述的系统,其中所述指令当由所述处理器执行时,使得对于评估用于开发所述模型的所述目标数据集合的可行性而被专门化的所述系统通过至少执行以下操作来针对所述偏倚标准值集合中的一个或多个偏倚标准值执行对所述目标数据集合的动态离群值偏倚减少以生成一个或多个离群值偏倚减少目标数据集合:

针对所述一个或多个偏倚标准值中的每个偏倚标准值:

通过将所述模型应用到所述目标数据集合来生成用于所述目标数据集合的多个模型预测值;

计算从所述目标数据集合和所述模型预测值确定的多个误差值;

将所述误差值与对应的偏倚标准值进行比较;

移除所述目标数据集合内的离群值以形成从所述误差值与所述对应的偏倚标准值的

比较确定的对应的离群值偏倚减少目标数据集合;以及

优化所述模型以形成从所述对应的离群值偏倚减少目标数据集合确定的更新的模型。

14. 如权利要求13所述的系统,其中所述指令当由所述处理器执行时,使得对于评估用于开发所述模型的所述目标数据集合的可行性而被专门化的所述系统通过至少执行以下操作来针对所述偏倚标准值集合中的一个或多个偏倚标准值执行对所述目标数据集合的动态离群值偏倚减少以生成一个或多个离群值偏倚减少目标数据集合:

针对所述一个或多个偏倚标准值中的每个偏倚标准值:

将所述误差值与预定义的终止标准进行比较以确定优化所述模型的终止;以及

当所述误差值与所述预定义的终止标准的比较不表示优化所述模型的终止时,通过将所述更新的模型应用到所述目标数据集合来生成用于所述目标数据集合的多个第二模型预测值。

15. 如权利要求11所述的系统,其中所述指令当由所述处理器执行时,使得对于评估用于开发所述模型的所述目标数据集合的可行性而被专门化的所述系统通过至少执行以下操作来比较所述第一偏倚标准曲线和所述第二偏倚标准曲线以确定被用于开发所述模型的所述目标数据集合的可行性:

确定所述第一偏倚标准曲线上的与所述目标误差值集合中的第一目标误差值对应的第一偏倚标准值;

确定所述第二偏倚标准曲线上的与所述随机误差值集合中的第一随机误差值对应的第二偏倚标准值;以及

将所述第一偏倚标准值与所述第二偏倚标准值进行比较,其中所述第一目标误差值和所述第一随机误差值是相同的。

16. 如权利要求11所述的系统,其中所述指令当由所述处理器执行时,使得对于评估用于开发所述模型的所述目标数据集合的可行性而被专门化的所述系统通过比较若干迭代来确定所述离群值偏倚减少对每个偏倚标准值的影响以优化用于所述偏倚标准值中的每个偏倚标准值的所述模型;以及比较所述目标相关系数集合的差别。

17. 如权利要求11所述的系统,其中所述指令当由所述处理器执行时,使得对于评估用于开发所述模型的所述目标数据集合的可行性而被专门化的所述系统通过至少执行以下操作来针对所述偏倚标准值集合中的一个或多个偏倚标准值执行对所述目标数据集合的动态离群值偏倚减少以生成一个或多个离群值偏倚减少目标数据集合:

针对所述一个或多个偏倚标准值中的每个偏倚标准值:

通过将所述模型应用到所述目标数据集合来生成用于所述目标数据集合的多个模型预测值;

计算从所述目标数据集合和所述模型预测值确定的多个误差值;

将所述误差值与对应的偏倚标准值进行比较;

移除所述目标数据集合内的离群值以形成从所述误差值与所述对应的偏倚标准值的比较确定的对应的离群值偏倚减少目标数据集合;以及

优化所述模型以形成从所述对应的离群值偏倚减少目标数据集合确定的更新的模型。

18. 一种对于评估用于开发模型的数据集合的可行性而被专门化的装置,所述装置包括:

耦接到非暂态计算机可读介质的处理器,其中所述非暂态计算机可读介质包括指令,所述指令当由所述处理器执行时,使得所述装置至少执行以下操作:

接收包括针对至少一个目标变量收集的所有实际数据值的目标数据集合;

基于所述目标数据集合生成随机数据集合;

接收被用于确定一个或多个离群值的多个偏倚标准值;

通过对所述目标数据集合应用数学模型和动态离群值偏倚减少来产生与所述偏倚标准值相关联的多个离群值偏倚减少目标数据集合;

通过对所述随机数据集合应用所述数学模型和所述动态离群值偏倚减少来产生与所述偏倚标准值相关联的多个离群值偏倚减少随机数据集合;

计算用于所述离群值偏倚减少目标数据集合中的每一个的至少一个目标误差值以及用于所述离群值偏倚减少随机数据集合中的每一个的至少一个随机误差值集合;

计算用于所述离群值偏倚减少目标数据集合中的每一个的至少一个目标相关值以及用于所述离群值偏倚减少随机数据集合中的每一个的至少一个随机相关值;

基于用于所述离群值偏倚减少目标数据集合中的每一个的所述至少一个目标误差值和所述至少一个目标相关值来构建第一偏倚标准曲线;

基于用于所述离群值偏倚减少随机数据集合中的每一个的所述至少一个随机误差值和所述至少一个随机相关值来构建第二偏倚标准曲线;以及

比较所述第一偏倚标准曲线和所述第二偏倚标准曲线以确定被用于开发所述数学模型的所述目标数据集合的可行性。

19. 如权利要求18所述的装置,其中所述指令当由所述处理器执行时,使得所述装置通过至少执行以下操作来通过对所述目标数据集合应用数学模型和动态离群值偏倚减少来产生与所述偏倚标准值相关联的多个离群值偏倚减少目标数据集合:

针对所述一个或多个偏倚标准值中的每个偏倚标准值:

通过将所述数学模型应用到所述目标数据集合来生成用于所述目标数据集合的多个模型预测值;

计算从所述目标数据集合和所述模型预测值确定的多个误差值;

将所述误差值与对应的偏倚标准值进行比较;

移除所述目标数据集合内的离群值以形成从所述误差值与所述对应的偏倚标准值的比较确定的对应的离群值偏倚减少目标数据集合;以及

优化所述数学模型以形成从所述对应的离群值偏倚减少目标数据集合确定的更新的数学模型。

20. 如权利要求18所述的装置,其中所述指令当由所述处理器执行时,使得所述装置通过至少执行以下操作来比较所述第一偏倚标准曲线和所述第二偏倚标准曲线以确定被用于开发所述模型的所述目标数据集合的可行性:

确定所述第一偏倚标准曲线上的与所述至少一个目标误差值对应的第一偏倚标准值;

确定所述第二偏倚标准曲线上的与所述至少一个随机误差值对应的第二偏倚标准值;

以及

将所述第一偏倚标准值与所述第二偏倚标准值进行比较,其中所述至少一个目标误差值和所述至少一个随机误差值是相同的。

动态离群值偏倚减少系统和方法

[0001] 本申请是申请日为2012年8月17日、发明名称为“动态离群值偏倚减少系统和方法”的中国专利申请201280048043.6的分案申请。

[0002] 相关申请的交叉引用

[0003] 此国际专利申请要求2011年8月19日递交的标题为“Dynamic Outlier Bias Reduction System and Method”、序列号为13/213,780的美国非临时专利申请的权益和优先权,这里通过引用将该美国申请全部并入。

[0004]

技术领域

[0005] 本发明涉及对数据的分析,其中离群元素被从分析开发中去除(或过滤掉)。分析可与简单统计量的计算或者在其开发中涉及使用数据的数学模型的更复杂操作有关。离群数据过滤的目的可以是执行数据质量和数据验证操作,或者计算能够应用于后续分析、回归分析、时间序列分析中的代表性标准、统计量、数据群组或者用于数学模型开发的合格数据。

背景技术

[0006] 在标准或数据驱动模型开发中去除离群数据是分析前工作的一个重要部分,用来确保从底层数据开发出有代表性的且合理的分析。例如,为二氧化碳(CO₂)、臭氧(O₃)、水蒸气(H₂O)、氢氟碳(HFC)、全氟化碳(PFC)、含氯氟烃(CFC)、六氟化硫(SF₆)、甲烷(CH₄)、一氧化二氮(N₂O)、一氧化碳(CO)、氧化氮(NO_x)以及非甲烷挥发性有机化合物(NMVOC)排放的温室气体标准开发公正的基准要求在标准开发中使用的所收集的工业数据表现出某些属性。少数几个工业场所的极好或极坏的性能不应当导致为其他场所计算的标准值的偏倚。可以判断将这种性能结果包括在标准计算中是不合理或者没有代表性的。过去,经由一种要求主观输入的半定量过程来去除性能离群值(outlier)。本系统和方法是一种数据驱动的方案,该方案把此任务作为模型开发的一个组成部分来执行,而不是在分析前或模型开发前阶段执行。

[0007] 偏倚(bias)的去除可以是一个主观过程,其中以某种形式记录认为恰当的理由以证实数据变化。然而,任何形式的离群值去除都是一种带着改变计算结果的潜在可能性的数据审查。这种数据过滤可能减少或不减少计算中的偏倚或误差,并且本着完全分析公开的精神,严格的数据去除指导方针和去除离群值的证明材料需要与分析结果包括在一起。因此,在现有技术中需要提供一种新的系统和方法,用于利用一种对于数据质量操作、数据验证、统计计算或数学模型开发等等有用的动态统计过程来客观地去除离群数据偏倚。离群值偏倚去除系统和方法还可用于将数据分组成代表性类别,其中数据被应用到为每个群组定制的数学模型开发。在优选实施例中,系数被定义为数学模型中的乘法和加法因子以及在性质上非线性的其他数值参数。例如,在数学模型 $f(x, y, z) = a*x + b*y^c + d*\sin(ez) + f$ 中,a、b、c、d、e和f全都被定义为系数。这些项的值可以固定或者是数学模型的开发的一部

分。

发明内容

[0008] 优选实施例包括一种用于减少离群值偏倚的由计算机实现的方法,包括以下步骤:选择偏倚标准;提供数据集合;提供模型系数的集合;选择目标值的集合;(1)为完整数据集生成预测值的集合;(2)为数据集生成误差集合;(3)基于误差集合和偏倚标准生成误差阈值的集合;(4)由处理器基于误差集合和误差阈值的集合生成经审查的数据集合;(5)由处理器生成新模型系数的集合;(6)使用新模型系数的集合,重复步骤(1)–(5),除非满足了审查性能终止标准。在优选实施例中,可基于数据集合和模型系数的集合来生成预测值的集合。在优选实施例中,误差集合可包括基于预测值的集合和目标值的集合生成的绝对误差的集合和相对误差的集合。在另一实施例中,误差集合可包括作为预测值的集合与目标值的集合之间的差异计算出的值。在另一实施例中,生成新系数的集合的步骤还可包括最小化预测值的集合与实际值的集合之间的误差的集合的步骤,这可利用线性或非线性优化模型来实现。在优选实施例中,审查性能终止标准可基于标准误差(standard error)和确定系数(coefficient of determination)。

[0009] 另一实施例包括一种用于减少离群值偏倚的由计算机实现的方法,包括以下步骤:选择误差标准;选择数据集合;选择实际值的集合;选择模型系数的初始集合;基于完整数据集和模型系数的初始集合生成模型预测值的集合;(1)对于完整数据集,基于模型预测值和实际值的集合生成误差的集合;(2)对于完整数据集,基于误差的完整集合和误差标准生成误差阈值的集合;(3)生成离群值被去除的数据集合,其中过滤基于完整数据集和误差阈值的集合;(4)基于经过滤的数据集合和先前系数的集合生成新系数的集合,其中新系数的集合的生成由计算机处理器执行;(5)基于经过滤的数据集合和新模型系数的集合生成离群值偏倚减少模型预测值的集合,其中离群值偏倚减少模型预测值的集合的生成由计算机处理器执行;(6)基于模型预测值和实际值的集合生成模型性能值的集合;在用新系数的集合替代来自先前迭代的系数的集合的同时,重复步骤(1)–(6),除非:满足性能终止标准;并且将模型预测值的集合存储在计算机数据介质中。

[0010] 另一实施例包括一种用于减少离群值偏倚的由计算机实现的方法,包括以下步骤:为设施选择目标变量;选择目标变量的实际值的集合;为设施识别与目标变量有关的多个变量;为设施获得数据集合,该数据集合包括多个变量的值;选择偏倚标准;选择模型系数的集合;(1)基于完整数据集和模型系数的集合生成预测值的集合;(2)基于预测值的集合和实际值的集合生成审查模型性能值的集合;(3)对于目标变量,基于预测值的集合和实际值的集合生成误差集合;(4)基于误差集合和偏倚标准生成误差阈值的集合;(5)由处理器基于数据集合和误差阈值的集合生成经审查的数据集合;(6)由处理器基于经审查的数据集合和模型系数的集合生成新模型系数的集合;(7)由处理器基于数据集合和新模型系数的集合生成新预测值的集合;(8)基于新预测值的集合和实际值的集合生成新审查模型性能值的集合;使用新系数的集合,重复步骤(1)–(8),除非满足了审查性能终止标准;以及将新模型预测值的集合存储在计算机数据介质中。

[0011] 另一实施例包括一种用于减少离群值偏倚的由计算机实现的方法,包括以下步骤:为设施选择目标变量,其中目标变量是关于工业设施的度量,与其生产、金融性能或排

放有关;为设施识别多个变量,其中多个变量包括:设施的影响目标变量的多个直接变量;以及设施的经变换变量的集合,每个经变换变量是影响目标变量的至少一个直接设施变量的函数;选择误差标准,包括:绝对误差,以及相对误差;为设施获得数据集合,其中数据集合包括多个变量的值;选择目标变量的实际值的集合;选择模型系数的初始集合;基于完整数据集合和模型系数的初始集合生成模型预测值的集合;基于模型预测值的集合和实际值的集合生成误差的完整集合,其中相对误差是利用如下公式计算的:相对误差 $m = ((\text{预测值}_m - \text{实际值}_m) / \text{实际值}_m)^2$,其中“m”是参考编号,并且其中绝对误差是利用如下公式计算的:绝对误差 $m = (\text{预测值}_m - \text{实际值}_m)^2$;基于模型预测值的集合和实际值的集合生成模型性能值的集合,其中整体模型性能值的集合包括:第一标准误差,以及第一确定系数;(1)对于完整数据集合,基于模型预测值和实际值的集合生成误差的集合;(2)对于完整数据集合,基于误差的完整集合和误差标准生成误差阈值的集合;(3)通过去除具有大于或等于误差阈值的误差值的数据生成离群值被去除的数据集合,其中过滤是基于完整数据集合和误差阈值的集合的;(4)通过利用线性优化模型和非线性优化模型中的至少一个,最小化预测值的集合与实际值的集合之间的误差,来基于离群值被去除的数据集合和模型系数的集合生成离群值偏倚减少模型预测值的集合,其中新模型预测值的生成是由计算机处理器执行的;(5)基于离群值被去除的数据集合和先前系数的集合生成新系数的集合,其中新系数的集合的生成是由计算机处理器执行的;(6)基于新预测模型值的集合和实际值的集合生成整体模型性能值的集合,其中模型性能值的集合包括:第二标准误差,以及第二确定系数;在用新系数的集合替代来自先前迭代的系数集合的同时,重复步骤(1)-(6),除非:满足了性能终止标准,其中性能终止标准包括:标准误差终止值和确定系数终止值,并且其中满足性能终止标准包括:标准误差终止值大于第一和第二标准误差之间的差异,并且确定系数终止值大于第一和第二确定系数之间的差异;以及将新模型预测值的集合存储在计算机数据介质中。

[0012] 另一实施例包括一种用于减少离群值偏倚的由计算机实现的方法,包括以下步骤:选择误差标准;选择数据集合;选择实际值的集合;选择模型预测值的初始集合;基于模型预测值的集合和实际值的集合确定误差的集合;(1)基于误差的完整集合和误差标准确定误差阈值的集合;(2)生成离群值被去除的数据集合,其中过滤基于数据集合和误差阈值的集合;(3)基于离群值被去除的数据集合和先前模型预测值生成离群值偏倚减少模型预测值的集合,其中离群值偏倚减少模型预测值的集合的生成由计算机处理器执行;(4)基于新模型预测值的集合和实际值的集合确定误差的集合;在用新模型预测值的集合替代来自先前迭代的模型预测值的集合的同时,重复步骤(1)-(4),除非:满足了性能终止标准;以及将离群值偏倚减少模型预测值的集合存储在计算机数据介质中。

[0013] 另一实施例包括一种用于减少离群值偏倚的由计算机实现的方法,包括以下步骤:为设施选择目标变量;为设施识别多个变量,其中多个变量包括:设施的影响目标变量的多个直接变量;以及设施的经变换变量的集合,每个经变换变量是影响目标变量的至少一个直接设施变量的函数;选择误差标准,包括:绝对误差,以及相对误差;获得数据集合,其中数据集合包括多个变量的值,以及选择目标变量的实际值的集合;选择模型系数的初始集合;通过将模型系数的集合应用到数据集合来生成模型预测值的集合;基于模型预测值的集合和实际值的集合确定性能值的集合,其中性能值的集合包括:第一标准误差,以及

第一确定系数；(1) 对于完整数据集合，基于模型预测值的集合和实际值的集合生成误差的集合，其中相对误差是利用如下公式计算的：相对误差_m = ((预测值_m - 实际值_m) / 实际值_m)²，其中“m”是参考编号，并且其中绝对误差是利用如下公式计算的：绝对误差_m = (预测值_m - 实际值_m)²；(2) 对于完整数据集合，基于误差的完整集合和误差标准生成误差阈值的集合；(3) 通过去除具有大于或等于误差阈值的集合的误差值的数据来生成离群值被去除的数据集合，其中过滤是基于数据集合和误差阈值的集合的；(4) 基于离群值被去除的数据集合和先前系数的集合生成新系数的集合；(5) 通过利用线性优化模型和非线性优化模型中的至少一个，最小化预测值的集合与实际值的集合之间的误差，来基于离群值被去除的数据集合和新模型系数的集合生成离群值偏倚减少模型预测值的集合，其中模型预测值的生成是由计算机处理器执行的；(6) 基于离群值偏倚减少模型预测值的集合和实际值的集合生成经更新的性能值的集合，其中经更新的性能值的集合包括：第二标准误差，以及第二确定系数；在用新系数的集合替代来自先前迭代的系数集合的同时，重复步骤(1) - (6)，除非：满足了性能终止标准，其中性能终止标准包括：标准误差终止值，以及确定系数终止值，并且其中满足性能终止标准包括：标准误差终止值大于第一和第二标准误差之间的差异，并且确定系数终止值大于第一和第二确定系数之间的差异；以及将离群值偏倚减少因子的集合存储在计算机数据介质中。

[0014] 另一实施例包括一种用于评估数据集合用于开发模型的可行性的由计算机实现的方法，包括以下步骤：提供包括多个数据值的目标数据集合；基于目标数据集合生成随机目标数据集合；选择偏倚标准值的集合；由处理器基于数据集合和每个所选择的偏倚标准值生成离群值偏倚减少目标数据集合；由处理器基于随机数据集合和每个所选偏倚标准值生成离群值偏倚减少随机数据集合；对于离群值偏倚减少数据集合和离群值偏倚减少随机数据集合计算误差值的集合；对于离群值偏倚减少数据集合和离群值偏倚减少随机数据集合计算相关系数的集合；基于所选偏倚标准值和相应的误差值和相关系数为数据集合和随机数据集合生成偏倚标准曲线；以及将数据集合的偏倚标准曲线与随机数据集合的偏倚标准曲线相比较。离群值偏倚减少目标数据集合和离群值偏倚减少随机目标数据集合是利用动态离群值偏倚去除方法生成的。随机目标数据集合可包括从多个数据值的范围内的值开发出来的随机化数据值。另外，误差值的集合可包括标准误差的集合，并且其中相关系数的集合包括确定系数值的集合。另一实施例还可包括以下步骤：基于将目标数据集合的偏倚标准曲线与随机目标数据集合的偏倚标准曲线相比较，生成关于目标数据集合支持所开发的模型以及所开发的模型支持目标数据集合的可行性的自动化建议。可基于分析者选择的参数——例如相关系数阈值和/或误差阈值——来生成建议。另外一个实施例还包括以下步骤：提供实际数据集合，其包括与模型预测值相对应的多个实际数据值；基于实际数据集合生成随机实际数据集合；由处理器基于实际数据集合和每个所选偏倚标准值生成离群值偏倚减少实际数据集合；由处理器基于随机实际数据集合和每个所选偏倚标准值生成离群值偏倚减少随机实际数据集合；对于每个所选偏倚标准，基于离群值偏倚减少随机目标数据集合和离群值偏倚减少随机实际数据生成随机数据图线；对于每个所选偏倚标准，基于离群值偏倚减少目标数据集合和离群值偏倚减少实际目标数据集合生成真实数据图线；以及将与每个所选偏倚标准相对应的随机数据图线与真实数据图线相比较。

[0015] 优选实施例包括一种系统，其包括：服务器，其包括：处理器，以及存储子系统；由

存储子系统存储的数据库,其包括:数据集合;以及由存储子系统存储的计算机程序,其包括当被执行时使得处理器进行以下操作的指令:选择偏倚标准;提供模型系数的集合;选择目标值的集合;(1)为数据集合生成预测值的集合;(2)为数据集合生成误差集合;(3)基于误差集合和偏倚标准生成误差阈值的集合;(4)基于误差集合和误差阈值的集合生成经审查的数据集合;(5)生成新模型系数的集合;以及(6)使用新模型系数的集合,重复步骤(1)-(5),除非满足了审查性能终止标准。在优选实施例中,可基于数据集合和模型系数的集合来生成预测值的集合。在优选实施例中,误差集合可包括基于预测值的集合和目标值的集合生成的绝对误差的集合和相对误差的集合。在另一实施例中,误差集合可包括作为预测值的集合与目标值的集合之间的差异计算出的值。在另一实施例中,生成新系数的集合的步骤还可包括最小化预测值的集合与实际值的集合之间的误差的集合的步骤,这可利用线性或非线性优化模型来实现。在优选实施例中,审查性能终止标准可基于标准误差和确定系数。

[0016] 本发明的另一实施例包括一种系统,其包括:服务器,其包括:处理器,以及存储子系统;由存储子系统存储的数据库,其包括:数据集合;以及由存储子系统存储的计算机程序,其包括当被执行时使得处理器进行以下操作的指令:选择误差标准;选择实际值的集合;选择系数的初始集合;从数据集合和系数的初始集合生成模型预测值的完整集合;(1)对于完整数据集合,基于模型预测值和实际值的集合生成误差的集合;(2)对于完整数据集合,基于误差的完整集合和误差标准生成误差阈值的集合;(3)生成离群值被去除的数据集合,其中过滤基于完整数据集合和误差阈值的集合;(4)基于离群值被去除的数据集合和系数的集合生成离群值偏倚减少模型预测值的集合,其中离群值偏倚减少模型预测值的集合的生成由计算机处理器执行;(5)基于离群值被去除的数据集合和先前系数的集合生成新系数的集合,其中新系数的集合的生成由计算机处理器执行;(6)基于离群值偏倚减少模型预测值和实际值的集合生成模型性能值的集合;在用新系数的集合替代来自先前迭代的系数的集合的同时,重复步骤(1)-(6),除非:满足性能终止标准;并且将整体离群值偏倚减少模型预测值的集合存储在计算机数据介质中。

[0017] 另一实施例包括一种系统,其包括:服务器,其包括:处理器,以及存储子系统;由存储子系统存储的数据库,其包括:设施的目标变量;目标变量的实际值的集合;设施的与目标变量有关的多个变量;设施的数据集合,该数据集合包括多个变量的值;以及由存储子系统存储的计算机程序,其包括当被执行时使得处理器进行以下操作的指令:选择偏倚标准;选择模型系数的集合;(1)基于数据集合和模型系数的集合生成预测值的集合;(2)基于预测值的集合和实际值的集合生成审查模型性能值的集合;(3)对于目标变量,基于预测值的集合和实际值的集合生成误差集合;(4)基于误差集合和偏倚标准生成误差阈值的集合;(5)基于数据集合和误差阈值的集合生成经审查的数据集合;(6)基于经审查的数据集合和模型系数的集合生成新模型系数的集合;(7)基于数据集合和新模型系数的集合生成新预测值的集合;(8)基于新预测值的集合和实际值的集合生成新审查模型性能值的集合;使用新系数的集合,重复步骤(1)-(8),除非满足了审查性能终止标准;以及将新模型预测值的集合存储在存储子系统中。

[0018] 另一实施例包括一种系统,其包括:服务器,其包括:处理器,以及存储子系统;由存储子系统存储的数据库,其包括:设施的数据集合;以及由存储子系统存储的计算机程

序,其包括当被执行时使得处理器进行以下操作的指令:确定目标变量;识别多个变量,其中多个变量包括:设施的影响目标变量的多个直接变量;以及设施的经变换变量的集合,每个经变换变量是影响目标变量的至少一个直接变量的函数;选择误差标准,包括:绝对误差,以及相对误差;选择目标变量的实际值的集合;选择系数的初始集合;基于数据集合和系数的初始集合生成模型预测值的集合;基于模型预测值的集合和实际值的集合生成误差的集合,其中相对误差是利用如下公式计算的:相对误差 $m = ((\text{预测值}_m - \text{实际值}_m) / \text{实际值}_m)^2$,其中“m”是参考编号,并且其中绝对误差是利用如下公式计算的:绝对误差 $m = (\text{预测值}_m - \text{实际值}_m)^2$;基于模型预测值的集合和实际值的集合确定性能值的集合;其中性能值的集合包括:第一标准误差,以及第一确定系数;(1)基于模型预测值和实际值的集合生成误差的集合;(2)对于完整数据集合,基于误差的完整集合和误差标准生成误差阈值的集合;(3)通过过滤掉具有误差阈值的集合以外的误差值的数据来生成离群值被去除的数据集合,其中过滤是基于数据集合和误差阈值的集合的;(4)通过利用线性优化模型和非线性优化模型中的至少一个,最小化模型预测值的集合与实际值的集合之间的误差,来基于离群值被去除的数据集合和系数的集合生成新模型预测值的集合,其中离群值偏倚减少模型预测值的生成是由计算机处理器执行的;(5)基于离群值被去除的数据集合和先前系数的集合生成新系数的集合,其中新系数的集合的生成是由计算机处理器执行的;(6)基于新预测模型值的集合和实际值的集合生成性能值的集合,其中模型性能值的集合包括:第二标准误差,以及第二确定系数;在用新系数的集合替代来自先前迭代的系数集合的同时,重复步骤(1)–(6),除非:满足了性能终止标准,其中性能终止标准包括:标准误差,以及确定系数,并且其中满足性能终止标准包括:标准误差终止值大于第一和第二标准误差之间的差异,并且确定系数终止值大于第一和第二确定系数之间的差异;以及将新模型预测值的集合存储在计算机数据介质中。

[0019] 本发明的另一实施例包括一种系统,其包括:服务器,其包括:处理器,以及存储子系统;由存储子系统存储的数据库,其包括:数据集合,由存储子系统存储的计算机程序,其包括当被执行时使得处理器进行以下操作的指令:选择误差标准;选择数据集合;选择实际值的集合;选择模型预测值的初始集合;基于模型预测值的集合和实际值的集合确定误差的集合;(1)基于误差的完整集合和误差标准确定误差阈值的集合;(2)生成离群值被去除的数据集合,其中过滤基于数据集合和误差阈值的集合;(3)基于离群值被去除的数据集合和模型预测值的完整集合生成离群值偏倚减少模型预测值的集合,其中离群值偏倚减少模型预测值的集合的生成由计算机处理器执行;(4)基于离群值偏倚减少模型预测值的集合和实际值的相应集合确定误差的集合;在用离群值偏倚减少模型预测值的集合替代模型预测值的集合的同时,重复步骤(1)–(4),除非:满足了性能终止标准;以及将离群值偏倚减少因子的集合存储在计算机数据介质中。

[0020] 本发明的另一实施例包括一种系统,其包括:服务器,其包括:处理器,以及存储子系统;由存储子系统存储的数据库,其包括:数据集合,由存储子系统存储的计算机程序,其包括当被执行时使得处理器进行以下操作的指令:确定目标变量;为设施识别多个变量,其中多个变量包括:设施的影响目标变量的多个直接变量;以及设施的经变换变量的集合,每个经变换变量是影响目标变量的至少一个主设施变量的函数;选择误差标准,包括:绝对误差,以及相对误差;获得数据集合,其中数据集合包括多个变量的值,以及选择目标变量的

实际值的集合;选择系数的初始集合;通过将模型系数的集合应用到数据集合来生成模型预测值的集合;基于模型预测值的集合和实际值的集合确定性能值的集合,其中性能值的集合包括:第一标准误差,以及第一确定系数;(1) 基于模型预测值的集合和实际值的集合确定误差的集合,其中相对误差是利用如下公式计算的:相对误差 $k = ((\text{预测值}_k - \text{实际值}_k) / \text{实际值}_k)^2$,其中“k”是参考编号,并且其中绝对误差是利用如下公式计算的:绝对误差 $k = (\text{预测值}_k - \text{实际值}_k)^2$;(2) 对于完整数据集合,基于误差的集合和误差标准确定误差阈值的集合;(3) 通过去除具有大于或等于误差阈值的误差值的数据来生成离群值被去除的数据集合,其中过滤是基于数据集合和误差阈值的集合的;(4) 基于离群值被去除的数据集合和先前系数的集合生成新系数的集合;(5) 通过利用线性优化模型和非线性优化模型中的至少一个,最小化预测值的集合与实际值的集合之间的误差,来基于离群值被去除的数据集合和系数的集合生成离群值偏倚减少模型值的集合;(5) 基于离群值偏倚减少模型预测值的集合和实际值的集合确定经更新的性能值的集合,其中经更新的性能值的集合包括:第二标准误差,以及第二确定系数;在用新系数的集合替代来自先前迭代的系数集合的同时,重复步骤(1)–(5),除非:满足了性能终止标准,其中性能终止标准包括:标准误差终止值,以及确定系数终止值,并且其中满足性能终止标准包括:标准误差终止值大于第一和第二标准误差之间的差异,并且确定系数终止值大于第一和第二确定系数之间的差异;以及将离群值偏倚减少因子的集合存储在计算机数据介质中。

[0021] 另一实施例包括一种用于评估数据集合用于开发模型的可行性的系统,其包括:服务器,其包括:处理器,以及存储子系统;由存储子系统存储的数据库,其包括:目标数据集合,其包括多个模型预测值;由存储子系统存储的计算机程序,其包括当被执行时使得处理器进行以下操作的指令:生成随机目标数据集合;选择偏倚标准值的集合;基于目标数据集合和每个所选择的偏倚标准值生成离群值偏倚减少数据集合;基于随机目标数据集合和每个所选偏倚标准值生成离群值偏倚减少随机目标数据集合;对于离群值偏倚减少目标数据集合和离群值偏倚减少随机目标数据集合计算误差值的集合;对于离群值偏倚减少目标数据集合和离群值偏倚减少随机目标数据集合计算相关系数的集合;对于每个所选偏倚标准,基于相应的误差值和相关系数,对于目标数据集合和随机目标数据集合生成偏倚标准曲线;以及将目标数据集合的偏倚标准曲线与随机目标数据集合的偏倚标准曲线相比较。处理器利用动态离群值偏倚去除方法来生成离群值偏倚减少目标数据集合和离群值偏倚减少随机目标数据集合。随机目标数据集合可包括从多个数据值的范围内的值开发出来的随机化数据值。另外,误差值的集合可包括标准误差的集合,并且相关系数的集合包括确定系数值的集合。在另一实施例中,该程序还包括在被执行时使得处理器进行以下操作的指令:基于将目标数据集合的偏倚标准曲线与随机目标数据集合的偏倚标准曲线相比较,生成自动化建议。可基于分析者选择的参数——例如相关系数阈值和/或误差阈值——来生成建议。在另外一个实施例中,系统的数据库还包括实际数据集合,其包括与模型预测值相对应的多个实际数据值,并且程序还包括当被执行时使得处理器进行以下操作的指令:基于实际数据集合生成随机实际数据集合;基于实际数据集合和每个所选偏倚标准值生成离群值偏倚减少实际数据集合;基于随机实际数据集合和每个所选偏倚标准生成离群值偏倚减少随机实际数据集合;对于每个所选偏倚标准,基于离群值偏倚减少随机目标数据集合和离群值偏倚减少随机实际数据生成随机数据图线;对于每个所选偏倚标准,基于离群值

偏倚减少目标数据集合和离群值偏倚减少实际目标数据集合生成真实数据图线;以及将与每个所选偏倚标准相对应的随机数据图线与真实数据图线相比较。

附图说明

- [0022] 图1是示出数据离群值识别和去除方法的实施例的流程图。
- [0023] 图2是示出用于数据质量操作的数据离群值识别和去除方法的实施例的流程图。
- [0024] 图3是示出用于数据验证的数据离群值识别和去除方法的实施例的流程图。
- [0025] 图4是用于实现本发明的方法的说明性节点。
- [0026] 图5是数据集合的定量评估的说明性图线。
- [0027] 图6A和6B是图5的数据集合的定性评估的说明性图线,分别示出了整个数据集合的随机化和真实数据集合。
- [0028] 图7A和7B是图5的数据集合的定性评估的说明性图线,分别示出了在将30%的数据作为离群值去除之后的随机化和真实数据集合。
- [0029] 图8A和8B是图5的数据集合的定性评估的说明性图线,分别示出了在将50%的数据作为离群值去除之后的随机化和真实数据集合。

具体实施方式

[0030] 以下公开提供了许多不同实施例,或者示例,用于实现用于访问和管理结构化内容的系统和方法的不同特征。描述了组件、过程和实现方式的具体示例来帮助阐明本发明。这些只是示例,而并不打算偏离权利要求中所述地限制本发明。公知的元素在没有详细描述的情况下给出,以免以不必要的细节模糊本发明的优选实施例。大多数情况下,省略了获得对本发明的优选实施例的完整理解所不必要的细节,因为这样的细节在相关领域的普通技术人员的技能范围内。

- [0031] 动态离群值偏倚减少的一个实施例的数学描述如下所示:
- [0032] 术语:
- [0033] $\hat{X}$ -所有数据记录的集合: $\hat{X} = \hat{X}_k + \hat{X}_{ck}$,其中:
- [0034] $\hat{X}_k$ -第k次迭代的接受数据记录的集合
- [0035] $\hat{X}_{ck}$ -第k次迭代的离群(去除)数据记录的集合
- [0036] $\hat{Q}_k$ -对于 $\hat{X}_k$ 计算出的模型预测值的集合
- [0037] $\hat{Q}_{ck}$ -对于数据记录 $\hat{X}_{ck}$ 的离群模型预测值的集合
- [0038] $\hat{A}$ -模型所基于的实际值(目标值)的集合
- [0039] $\hat{\beta}_{k \rightarrow k+1}$ -作为使用 $\hat{X}_k$ 的模型计算的结果计算出的第k+1次迭代时的模型系数的集合
- [0040] $M(\hat{X}_k : \hat{\beta}_{k \rightarrow k+1})$ -从 $\hat{X}_k$ 存储模型得出的和用户提供的系数 $\hat{\beta}_{k \rightarrow k+1}$ 产生 $\hat{Q}_{k+1}$ 的模型计算
- [0041] C-用户提供的误差标准(%)
- [0042] $\Psi(\hat{Q}_k, \hat{A})$ -误差阈值函数
- [0043] $F(\Psi, C)$ -误差阈值(E)

[0044] $\hat{\Omega}_k$ -迭代终止标准,例如迭代计数、 r^2 、标准误差等等

[0045] 初始计算, $k=0$

[0046] 初始步骤1:利用初始模型系数估计 $\hat{\beta}_{0 \rightarrow 1}$,通过将模型应用到完整数据集合来计算初始模型预测值:

[0047] $\hat{Q}_1 = M(\hat{X} : \hat{\beta}_{0 \rightarrow 1})$

[0048] 初始步骤2:计算初始模型性能结果:

[0049] $\hat{\Omega}_1 = f(\hat{Q}_1, \hat{A}, k=0, r^2, \text{标准误差}, \text{等等})$

[0050] 初始步骤3:计算(一个或多个)模型误差阈值:

[0051] $E_1 = F(\Psi(\hat{Q}_1, \hat{A}), C)$

[0052] 初始步骤4:过滤数据记录以去除离群值:

[0053] $\hat{X}_1 = \{ \forall x \in \hat{X} \mid \Psi(\hat{Q}_1, \hat{A}) < E_1 \}$

[0054] 迭代计算, $k>0$

[0055] 迭代步骤1:通过将模型应用到接受数据集合来计算预测值:

[0056] $\hat{Q}_{k+1} = M(\hat{X}_k : \hat{\beta}_{k \rightarrow k+1})$

[0057] 迭代步骤2:计算模型性能结果:

[0058] $\hat{\Omega}_{k+1} = f(\hat{Q}_{k+1}, \hat{A}, k, r^2, \text{标准误差}, \text{等等})$

[0059] 如果达到终止标准,则停止,否则去到步骤3:

[0060] 迭代步骤3:利用当前模型为去除的数据 $\hat{X}_{Ck} = \{ \forall x \in \hat{X} \mid x \notin \hat{X}_k \}$ 计算结果:

[0061] $\hat{Q}_{Ck+1} = M(\hat{X}_{Ck} : \hat{\beta}_{k \rightarrow k+1})$

[0062] 迭代步骤4:计算模型误差阈值:

[0063] $E_{k+1} = F(\Psi(\hat{Q}_{k+1} + \hat{Q}_{Ck+1}, \hat{A}), C)$

[0064] 迭代步骤5:过滤数据记录以去除离群值:

[0065] $\hat{X}_{k+1} = \{ \forall x \in \hat{X} \mid \Psi(\hat{Q}_{k+1} + \hat{Q}_{Ck+1}, \hat{A}) < E_{k+1} \}$

[0066] 动态离群值偏倚减少的一个实施例的另一数学描述如下所示:

[0067] 术语:

[0068] $\hat{X}$ -所有数据记录的集合: $\hat{X} = \hat{X}_k + \hat{X}_{Ck}$,其中:

[0069] $\hat{X}_k$ -第k次迭代的接受数据记录的集合

[0070] $\hat{X}_{Ck}$ -第k次迭代的离群(去除)数据记录的集合

[0071] $\hat{Q}_k$ -对于 $\hat{X}_k$ 计算出的模型预测值的集合

[0072] $\hat{Q}_{Ck}$ -对于 $\hat{X}_{Ck}$ 的离群模型预测值的集合

[0073] $\hat{A}$ -模型所基于的实际值(目标值)的集合

[0074] $\hat{\beta}_{k \rightarrow k+1}$ -作为使用 $\hat{X}_k$ 的模型计算的结果计算出的第k+1次迭代时的模型系数的集合

[0075] $M(\hat{X}_k : \hat{\beta}_{k \rightarrow k+1})$ —从 $\hat{X}_k$ 存储模型得出的和用户提供的系数 $\hat{\beta}_{k \rightarrow k+1}$ 的产生 $\hat{Q}_{k+1}$ 的模型计算

[0076] C_{RE} —用户提供的相对误差标准 (%)

[0077] C_{AE} —用户提供的绝对误差标准 (%)

[0078] $RE(\hat{Q}_k + \hat{Q}_{CK}, \hat{A})$ —所有数据记录的相对误差值

[0079] $AE(\hat{Q}_k + \hat{Q}_{CK}, \hat{A})$ —所有数据记录的绝对误差值

[0080] P_{RE_k} —第k次迭代的相对误差阈值, 其中

[0081] $P_{RE_k} = \text{百分位}(RE(\hat{Q}_k + \hat{Q}_{CK}, \hat{A}), C_{RE})$

[0082] P_{AE_k} —第k次迭代的绝对误差阈值, 其中

[0083] $P_{AE_k} = \text{百分位}(AE(\hat{Q}_k + \hat{Q}_{CK}, \hat{A}), C_{AE})$

[0084] $\hat{\Omega}_k$ —迭代终止标准, 例如迭代计数、 r^2 、标准误差等等

[0085] 初始计算, $k=0$

[0086] 初始步骤1: 利用初始模型系数估计 $\hat{\beta}_{0 \rightarrow 1}$, 通过将模型应用到完整数据集来计算初始模型预测值:

[0087] $\hat{Q}_1 = M(\hat{X} : \hat{\beta}_{0 \rightarrow 1})$

[0088] 初始步骤2: 计算初始模型性能结果:

[0089] $\hat{\Omega}_1 = f(\hat{Q}_1, \hat{A}, k=0, r^2, \text{标准误差}, \text{等等})$

[0090] 初始步骤3: 计算模型误差阈值:

[0091] $P_{RE_1} = \text{百分位}(RE(\hat{Q}_1, \hat{A}), C_{RE})$

[0092] $P_{AE_1} = \text{百分位}(AE(\hat{Q}_1, \hat{A}), C_{AE})$

[0093] 初始步骤4: 过滤数据记录以去除离群值:

[0094] $\hat{X}_1 = \left\{ \forall x \in \hat{X} \mid \begin{Bmatrix} RE(\hat{Q}_1, \hat{A}) \\ AE(\hat{Q}_1, \hat{A}) \end{Bmatrix} < \begin{pmatrix} P_{RE} \\ P_{AE} \end{pmatrix}_1 \right\}$

[0095] 迭代计算, $k>0$

[0096] 迭代步骤1: 通过将模型应用到离群值被去除的数据集合来计算模型预测值:

[0097] $\hat{Q}_{k+1} = M(\hat{X}_k : \hat{\beta}_{k \rightarrow k+1})$

[0098] 迭代步骤2: 计算模型性能结果:

[0099] $\hat{\Omega}_{k+1} = f(\hat{Q}_{k+1}, \hat{A}, k, r^2, \text{标准误差}, \text{等等})$

[0100] 如果达到终止标准, 则停止, 否则去到步骤3:

[0101] 迭代步骤3: 利用当前模型为去除的数据 $\hat{X}_{CK} = \{ \forall x \in \hat{X} \mid x \notin \hat{X}_k \}$ 计算结果:

[0102] $\hat{Q}_{CK+1} = M(\hat{X}_{CK} : \hat{\beta}_{k \rightarrow k+1})$

[0103] 迭代步骤4: 计算模型误差阈值:

$$[0104] \quad P_{RE_{k+1}} = \text{百分位}(RE(\bar{Q}_{k+1} + \bar{Q}_{Ck+1}, \bar{A}), C_{RE})$$

$$[0105] \quad P_{AE_{k+1}} = \text{百分位}(AE(\bar{Q}_{k+1} + \bar{Q}_{Ck+1}, \bar{A}), C_{AE})$$

[0106] 迭代步骤5:过滤数据记录以去除离群值:

$$[0107] \quad \hat{X}_{k+1} = \left\{ \forall x \in \hat{X} \mid \begin{Bmatrix} RE(\bar{Q}_{k+1} + \bar{Q}_{Ck+1}, \bar{A}) \\ AE(\bar{Q}_{k+1} + \bar{Q}_{Ck+1}, \bar{A}) \end{Bmatrix} < \begin{pmatrix} P_{RE} \\ P_{AE} \end{pmatrix}_{k+1} \right\}$$

[0108] 递增k并去到迭代步骤1。

[0109] 在从当前审查的数据集计算新模型系数的每次迭代之后,从先前迭代去除的数据加上当前审查的数据被重组合。此组合涵盖了完整数据集中的所有数据值。随后将当前模型系数应用到完整数据集以计算预测值的完整集合。对于预测值的完整集合计算绝对和相对误差,并且计算新的偏倚标准百分位阈值。通过去除绝对或相对误差大于阈值的所有数据值来创建新的审查数据集,然后向新审查的数据集应用非线性优化模型来计算新的模型系数。此过程使得所有数据值在每一次迭代中都能够被审查以查明其是否可能包括在模型数据集中。随着模型系数收敛在最佳拟合数据的值上,在先前迭代中排除的一些数据值有可能将被包括入后续迭代中。

[0110] 在一个实施例中,GHG排放的变动可导致对排放结果的过高估计或过低估计,这导致了模型预测值中的偏倚。这些非工业影响——例如环境条件和计算过程中的误差——可使得关于特定设施的结果与类似设施的极为不同,除非去除了模型预测值中的偏倚。模型预测值中的偏倚也可由于独特的操作条件而存在。

[0111] 如果分析者确信一设施的计算有差错或者拥有独特的、情有可原的特性,则可以通过简单地将设施的数据从计算中去除来去除偏倚。然而,当测量来自许多不同的公司、地区和国家的设施性能时,数据细节的精确先验知识是不现实的。因此,任何基于分析者的数据去除过程都具有向模型结果添加无事实根据的、无数据支持的偏倚的潜在可能。

[0112] 在一个实施例中,动态离群值偏倚减少被应用到使用数据和规定的整体误差标准来确定被从模型系数计算中去除的统计离群值的过程。这是一个数据驱动的过程,其利用由数据产生的全局误差标准来识别离群值,其中该误差标准例如使用百分位函数。动态离群值偏倚减少的使用不限于模型预测值中的偏倚的减少,并且其在此实施例中的使用只是说明性和示范性的。动态离群值偏倚减少也可用于例如从任何统计数据集合中去除离群值,包括用于算术平均、线性回线和趋势线的计算,但不限于这些计算。离群设施仍从计算结果中排名,但离群值不用在被应用来计算模型系数或统计结果的经过滤的数据集合中。

[0113] 常用于去除离群值的标准过程是计算数据集合的标准偏差(σ)并且简单地将例如在均值的 2σ 间隔外的所有数据定义为离群值。此过程具有一般在实践中不能测试的统计假设。在本发明的实施例中应用的动态离群值偏倚减少方法描述在图1中概略示出,其使用了相对误差和绝对误差两者。例如,对于一设施“m”,

$$[0114] \quad \text{相对误差}_m = ((\text{预测值}_m - \text{实际值}_m) / \text{实际值}_m)^2 \quad (1)$$

$$[0115] \quad \text{绝对误差}_m = (\text{预测值}_m - \text{实际值}_m)^2 \quad (2)$$

[0116] 在步骤110中,分析者规定误差阈值标准,该误差阈值标准将定义要从计算中去除的离群值。例如,使用百分位运算作为误差函数,可以设定相对和绝对误差的80%的百分位值。这意味着,对于相对误差小于第80百分位值的数据值和对于绝对误差计算小于第80百

分位值的数据值将被包括,而剩余的值被去除或被认为是离群值。在此示例中,一个数据值要避免被去除,则该数据值必须小于相对和绝对误差两者的第80百分位值。然而,可以独立地改变相对和绝对误差的百分位阈值,并且在另一实施例中,可以只使用百分位阈值中的一个。

[0117] 在步骤120中,规定模型标准误差和确定系数(r^2)百分比变化标准。虽然这些统计量的值在模型与模型之间将有变化,但在前迭代过程中的百分比变化可被预设例如5%。这些值可用于终止迭代过程。另一个终止标准可以是简单的迭代计数。

[0118] 在步骤130中,执行优化计算,其为每个设施产生模型系数和预测值。

[0119] 在步骤140中,利用式(1)和(2)计算所有设施的相对和绝对误差。

[0120] 在步骤150中,向在步骤140中计算出的数据应用具有在步骤110中规定的阈值标准的误差函数以确定离群值阈值。

[0121] 在步骤160中,过滤数据以仅包括如下设施:对于这些设施,相对误差、绝对误差或者这两个误差(这取决于所选择的配置)小于在步骤150中计算出的误差阈值。

[0122] 在步骤170中,仅利用离群值被去除的数据集合来执行优化计算。

[0123] 在步骤180中,将标准误差和 r^2 的百分比变化与在步骤120中规定的标准相比较。如果百分比变化大于标准,则通过返回到步骤140来重复该过程。否则,在步骤190中终止迭代过程,并且完成了从这个动态离群值偏倚减少标准过程计算出的结果模型。模型结果被应用到所有设施,无论其当前迭代过去去除或许可数据状态如何。

[0124] 在另一实施例中,该过程开始于选择某些迭代参数,具体而言:

[0125] (1) 绝对误差和相对误差百分位值,其中一者、另一者或两者可用在迭代过程中,

[0126] (2) 确定系数(也称为 r^2)改进值,以及

[0127] (3) 标准误差改进值。

[0128] 该过程开始于原始数据集合、实际数据的集合以及用于基于原始数据集合计算预测值的至少一个系数或因子。系数或系数的集合将被应用到原始数据集合以创建预测值的集合。系数的集合可包括但不限于标量、指数、参数和周期函数。随后将预测数据的集合与实际数据的集合相比较。基于预测数据与实际数据之间的差异来计算标准误差和确定系数。与数据点中的每一个相关联的绝对和相对误差被用于基于用户选择的绝对和相对误差百分位值来去除数据离群值。对数据排名是不必要的,因为落在与绝对和/或相对误差的百分位值相关联的范围之外的所有数据都被从原始数据集合中去除。使用绝对和相对误差来过滤数据是说明性的并且仅用于示范,因为可以仅用绝对误差或仅用相对误差来者用另一函数来执行该方法。

[0129] 与在用户选择的百分位范围内的绝对和相对误差相关联的数据是离群值被去除的数据集合,并且该过程的每次迭代将具有其自己的经过滤的数据集合。这个第一离群值被去除的数据集合被用于确定将与实际值相比较的预测值。通过优化误差来确定至少一个系数,然后将这些系数用于基于第一离群值被去除的数据集合生成预测值。离群值偏倚减少的系数充当用来从一次迭代向下一次迭代传递知识的机制。

[0130] 在创建第一离群值被去除的数据集合之后,计算标准误差和确定系数并将其与原始数据集合的标准误差和确定系数相比较。如果标准误差的差异和确定系数的差异都低于其各自的改进值,则该过程停止。然而,如果改进标准中的至少一个未得到满足,则该过程

继续另一次迭代。将标准误差和确定系数用作对迭代过程的检查只是说明性和示范性的，因为该检查可仅利用标准误差或仅利用确定系数、利用不同的统计检查或者利用某种其他的性能终止标准（例如迭代次数）来执行。

[0131] 假设第一次迭代未能满足改进标准，则第二次迭代开始，将第一离群值偏倚减少数据系数应用到原始数据以确定预测值的新集合。然后再次处理原始数据，从而在使用第一离群值被去除的数据集合系数的同时，对于数据点建立绝对和相对误差，并且对于原始数据集合建立标准误差和确定系数值。然后过滤数据以形成第二离群值被去除的数据集合并基于第二离群值被去除的数据集合来确定系数。

[0132] 然而，第二离群值被去除的数据集合不一定是第一离群值被去除的数据集合的子集，并且其与离群值偏倚减少模型系数的第二集合、第二标准误差和第二确定系数相关联。一旦确定了这些值，就将第二标准误差与第一标准误差相比较，并且将第二确定系数再次与第一确定系数相比较。

[0133] 如果（标准误差和确定系数的）改进值超过这些参数的差异，则该过程将结束。如果否，则另一迭代将开始，再次处理原始数据；这一次使用第二离群值偏倚减少系数来处理原始数据集合并生成预测值的新集合。基于用户为绝对和相对误差选择的百分位值的过滤将创建第三离群值被去除的数据集合，该集合将被优化以确定第三离群值偏倚减少系数的集合。该过程将继续，直到满足误差改进或其他终止标准为止（例如收敛标准或规定的迭代次数）。

[0134] 此过程的输出将是系数或模型参数的集合，其中系数或模型参数是数学值（或值的集合），例如但不限于用于比较数据的模型预测值、线性方程的斜率和截距值、指数或者多项式的系数。动态离群值偏倚减少的输出本身不是输出值，而是将会修改数据以确定输出值的系数。

[0135] 在图2中所示的另一实施例中，将动态离群值偏倚减少作为一种数据质量技术应用来评估数据的一致性和准确性以核实数据对于特定的用途是适当的。对于数据质量操作，该方法可不涉及迭代过程。在此过程期间可与动态离群值偏倚减少并行地使用其他数据质量技术。该方法被应用到给定数据集合的算术平均计算。数据质量标准对于此示例是连续的数据值被包含在某个范围内。从而，值间隔得太开的任何值都将构成质量不佳的数据。然后构造函数的连续值的误差项，并且向这些误差值应用动态离群值偏倚减少。

[0136] 在步骤210中，按任何顺序列出初始数据。

[0137] 步骤220构成对数据集执行的函数或操作。在此实施例示例中，函数和操作是对数据的升序排名，然后是连续的算术平均计算，其中每一行对应于该行和该行上方的所有数据的平均。

[0138] 步骤230利用来自步骤220的结果的连续值从数据计算相对和绝对误差。

[0139] 步骤240允许分析者输入期望的离群值去除误差标准（%）。质量标准值是基于步骤220中的数据来自步骤230中的误差计算的结果值。

[0140] 步骤250示出数据质量离群值过滤数据集。如果相对和绝对误差超过步骤240中给出的规定误差标准，则去除具体的值。

[0141] 步骤260示出了完整的数据集和离群值被去除的数据集之间的算术平均计算比较。分析者像所有应用的数学或统计计算中那样进行最终步骤，判断所识别出的离群值被

去除的数据元素是否实际质量不佳。动态离群值偏倚减少系统和方法消除了分析者直接去除数据的情况,但最佳实践指导方针建议分析者就实际相关性审查并检查结果。

[0142] 在图3中所示的另一实施例中,应用动态离群值偏倚减少作为一种数据验证技术,该技术测试数据集合的合理准确性以确定数据对于特定的用途是否适当。对于数据验证操作,该方法可不涉及迭代过程。在此示例中,动态离群值偏倚减少被应用到两个数据集合之间的皮尔逊相关系数的计算。皮尔逊相关系数对于数据集合中的与其他数据点相对不同的值可以敏感。就这个统计量验证数据集合对于确保结果代表大部分数据表明什么而不是极值的影响是重要的。数据验证过程对于此示例是连续的数据值被包含在规定范围内。从而,值间隔得太开(例如在规定范围以外)的任何值将表示质量不佳的数据。这是通过构造函数的连续值的误差项来完成的。动态离群值偏倚减少被应用到这些误差值,并且离群值被去除的数据集合是经验证的数据。

[0143] 在步骤310中,按任何顺序列出配对的数据。

[0144] 步骤320对于数据集中的每个有序对计算相对和绝对误差。

[0145] 步骤330允许分析者输入期望的数据验证标准。在示例中,选择90%相对和绝对误差阈值。步骤330中的质量标准值输入是步骤320中所示的数据的结果绝对和相对误差百分位值。

[0146] 步骤340示出了离群值去除过程,其中利用相对和绝对误差值两者都超过与在步骤330中输入的用户选择百分位值相对应的值这个标准从数据集中去除可能无效的数据。在实践中可以使用其他误差标准,并且当如此示例中所示那样应用多个标准时,可以应用误差值的任何组合来确定离群值去除规则。

[0147] 步骤350计算经验证数据和原始数据值统计结果。在此情况下,皮尔逊相关系数。随后分析者就实际相关性审查这些结果。

[0148] 在另一实施例中,动态离群值偏倚减少用于执行整个数据集合的验证。选择标准误差改进值、确定系数改进值和绝对和相对误差阈值,然后根据误差标准来过滤数据集合。即使原始数据集合具有高质量,也仍将会有一些数据会具有落在绝对和相对误差阈值以外的误差值。因此,确定对数据的任何去除是否必要,是重要的。如果离群值被去除的数据集合在第一次迭代之后通过了标准误差改进和确定系数改进标准,则原始数据集合已得到验证,因为经过滤的数据集合产生了太小以至于不被认为有意义(例如在所选择的改进值以下)的标准误差和确定系数。

[0149] 在另一实施例中,动态离群值偏倚减少用于提供关于数据离群值去除的迭代如何影响计算的洞悉。提供图线或数据表格来允许用户观察随着每次迭代被执行,数据离群值去除计算的进展。这个分步方案使得分析者能够观察计算的能够向结果添加价值和知识的独特属性。例如,速度和收敛性质可指示动态离群值偏倚减少对计算多维数据集合的代表性因子的影响。

[0150] 作为说明,考虑对87个记录的质量不佳数据集合的线性回归计算。回归的方程的形式是 $y = mx + b$ 。表1示出了5次迭代的迭代过程的结果。注意,利用95%的相对和绝对误差标准,在3次迭代中实现收敛。可以观察到回归系数的变化,并且动态离群值偏倚减少方法基于79个记录缩减了计算数据集合。相对低的确定系数($r^2 = 39\%$)表明,应当测试更低($< 95\%$)的标准以研究对 r^2 统计量和对计算出的回归系数的额外离群值去除效果。

[0151] 表1:动态离群值偏倚减少示例:95%的线性回归

[0152]

迭代	N	误差	r^2	m	b
0	87	3.903	25%	-0.428	41.743
1	78	3.048	38%	-0.452	43.386
2	83	3.040	39%	-0.463	44.181
3	79	3.030	39%	-0.455	43.630
4	83	3.040	39%	-0.463	44.181
5	79	3.030	39%	-0.455	43.630

[0153] 在表2中,利用80%的相对和绝对误差标准示出了应用动态离群值偏倚减少的结果。注意,离群值误差标准中的15个百分点(95%到80%)的变化产生了 r^2 的35个百分点(39%到74%)的增加,并且许可的数据有35%的额外减少(包括79个记录到包括51个记录)。分析者在分析过程中可使用回归线的变化的图形视图,其中带有离群值被去除的数据和表1和2的数值结果,以将离群值被去除的结果传达给更宽的受众并且提供关于数据可变性对分析结果的效果的更多洞悉。

[0154] 表2:动态离群值偏倚减少示例80%的线性回归

[0155]

迭代	N	误差	r^2	m	b
0	87	3.903	25%	-0.428	41.743
1	49	1.607	73%	-0.540	51.081
2	64	1.776	68%	-0.561	52.361
3	51	1.588	74%	-0.558	52.514
4	63	1.789	68%	-0.559	52.208
5	51	1.588	74%	-0.558	52.514

[0156] 如图4中所示,用于执行该方法的系统的一个实施例包括计算系统。硬件包括处理器410,处理器410包含充足的系统存储器420以执行所需的数值计算。处理器410执行驻留在系统存储器420中的计算机程序以执行该方法。视频和存储控制器430可用于使能显示器440的操作。系统包括用于数据输入的各种数据存储设备,例如软盘单元450、内部/外部盘驱动器460、内部CD/DVD 470、磁带单元480和其他类型的电子存储介质490。上述数据存储设备只是说明性和示范性的。这些存储介质用于将数据集合和离群值去除标准输入到系统中,存储离群值被去除的数据集合,存储计算出的因子,以及存储系统产生的趋势线和趋势线迭代图。计算可应用统计软件包或者可根据例如利用Microsoft Excel以电子数据表格格式输入的数据执行。计算是利用为公司特定的系统实现设计的定制软件程序或者利用兼容Excel或其他数据库和电子数据表程序的市售软件来执行的。系统还可与专有的或公共的外部存储介质300接口以与其他数据库链接来提供要用于动态离群值偏倚减少系统和方法计算的数据。输出设备可以是用于经由内联网或因特网将计算工作表和其他由系统产生的图线和报告发送到管理人员或其他人员的电信设备510、打印机520、与作为输入设备450、460、470、480、490提及的那些类似的电子存储介质以及专有的存储数据库530。这里使用的这些输出设备只是说明性和示范性的。

[0157] 如图5、6A、6B、7A、7B、8A和8B中所示,在一个实施例中,动态离群值偏倚减少可用于基于与基准数据集的误差和相关性相比的数据集合的数据值的误差和相关性来定量地和定性地评估数据集合的质量,其中基准数据集由从适当范围内开发出来的随机数据值构成。在一个实施例中,可以指定误差为数据集合的标准误差,并且可以指定相关性为数据集合的确定系数(r^2)。在另一实施例中,可以指定相关性为肯德尔等级相关系数,通常称为肯德尔 τ 系数。在另外一个实施例中,可以指定相关性为斯皮尔曼等级相关系数,或者斯皮尔曼 ρ 系数。如上所述,动态离群值偏倚减少用于系统地去被识别为离群值、不代表所描述的底层模型或过程的数据值。通常,离群值与相对少量的数据值相关联。然而,在实践中,数据集可能不知不觉地被伪造值或随机噪声所污染。图5、6A、6B、7A、7B、8A和8B的图示说明了可如何应用动态离群值偏倚减少系统和方法来识别数据不支持底层模型的情形。通过去除如下数据值来执行离群值减少:对于这些数据值,在模型预测值与实际数据值之间计算出的相对和/或绝对误差大于基于百分位的偏倚标准,例如80%。这意味着,如果相对或绝对误差百分位值大于与第80百分位相关联的百分位阈值(80%的数据值具有小于此值的误差),则去除数据值。

[0158] 如图5中所示,真实模型开发数据集和在实际数据集的范围内开发的随机值的数据集两者被比较。因为在实践中,分析者通常不具有关于任何数据集污染的先验知识,所以这种实现必须来自于观察使用动态离群值偏倚减少系统和方法的若干个模型计算的迭代结果。图5示出了两个数据集的示范性模型开发计算结果。相对于表示模型解释了多少数据变动的确定系数(%)或 r^2 ,绘出作为模型未解释的误差量的度量的标准误差。每个点旁边的百分位值表示偏倚标准。例如,90%表示相对或绝对误差值大于第90百分位的数据值作为离群值被从模型中去除。这相当于每次迭代去除具有最高误差的10%的数据值。

[0159] 如图5所示,对于随机和真实数据集模型两者,通过增大偏倚标准来减少误差,即,对于两个数据集都改进标准误差和确定系数。然而,随机数据集的标准误差是真实模型数据集的两倍或三倍那么大。分析者可使用例如80%确定系数要求作为确定模型参数的可接受精度水平。在图5中,对于随机数据集在70%的偏倚标准下实现80%的 r^2 ,而对于真实数据在大约85%的偏倚标准下实现80%的 r^2 。然而,随机数据集的相应标准误差超过真实数据集的两倍那么大。从而,通过以不同的偏倚标准系统地运行模型数据集分析并且以代表性伪造数据集重复这些计算并且如图5中所示绘出结果,分析者可以评估数据集合的可接受偏倚标准(即,去除的数据值的可接受百分比),并且相应地评估整体数据集质量。另外,这种系统性的模型数据集分析可用于自动提出关于数据集用来基于可配置参数集合开发模型的可行性的建议。例如,在利用动态离群值偏倚去除为数据集开发模型的一个实施例中,在不同的偏倚标准下计算出的模型数据集和代表性伪造数据集的误差和相关系数值可用于自动提出关于以下可行性的建议:数据集在支持所开发的模型方面的可行性,以及本质而言,所开发的模型在支持数据集方面的可行性。

[0160] 如图5中所示,对于若干种情况观察这些模型性能值的行为提供了用于确定数据值是否代表被建模的过程的定量基础。例如,参考图5,100%偏倚标准下(即,无偏倚减少)真实数据集的标准误差对应于大约65%偏倚标准下(即,具有最高误差的35%的数据被去除)随机数据集的标准误差。这种发现支持了数据未受污染的结论。

[0161] 除了上述由图5的说明性图线促进的定量分析以外,动态离群值偏倚减少还可同

样地——甚至更强大地——用在主观过程中来帮助评估数据集的质量。这是通过对于离群值和包括的结果两者对照在给定实际目标值的情况下的数据绘出模型预测值来完成的。

[0162] 图6A和6B对于图5中的真实和随机曲线两者的100%点示出了这些图线。图6A中的大分散与任意的目标值和由此造成的模型不能拟合此故意随机性相一致。图6B与实际数据收集一致且共通,因为模型预测和实际值更多地聚集在模型预测值等于实际目标值的线(以下称为实际=预测线)附近。

[0163] 图7A和7B示出了来自图5中的70%点的结果(即,30%的数据被作为离群值去除)。在图7A和7B中,离群值偏倚减少被示为去除了最远离实际=预测线的点,但在图7A和7B之间模型准确度的大变动表明此数据集代表了被建模的过程。

[0164] 图8A和8B示出了来自图5中的50%点的结果(即,50%的数据被作为离群值去除)。在此情况下,大约一半的数据被识别为离群值,并且甚至是在这样的大变动被从数据集中去除的情况下,图8A中的模型仍没有紧密地描述随机数据集。考虑到每种情况中去除的数据,实际=预测线附近的一般变动与图6A和7A中的大致相同。图8B示出了在去除变动性的50%的情况下,模型能够产生紧密匹配实际数据的预测结果。除了对图5中所示的性能标准的分析以外,对这些类型的可视图线的分析也可被分析者用于在实践中就模型开发评估实际数据集的质量。虽然图5、6A、6B、7A、7B、8A和8B示出了其中分析基于与各种偏倚标准值相对应的性能标准趋势的可视图线,但在其他实施例中,分析可基于与偏倚标准值相对应的其他变量,例如与分析者选择的各种偏倚标准相对应的模型系数趋势。

[0165] 对本发明的优选实施例的以上公开和描述是对其的例示和说明,并且本领域技术人员将会理解,在不脱离本发明的范围的情况下,可对所例示的系统和方法的细节进行各种改变。

数据离群值偏倚减少方法

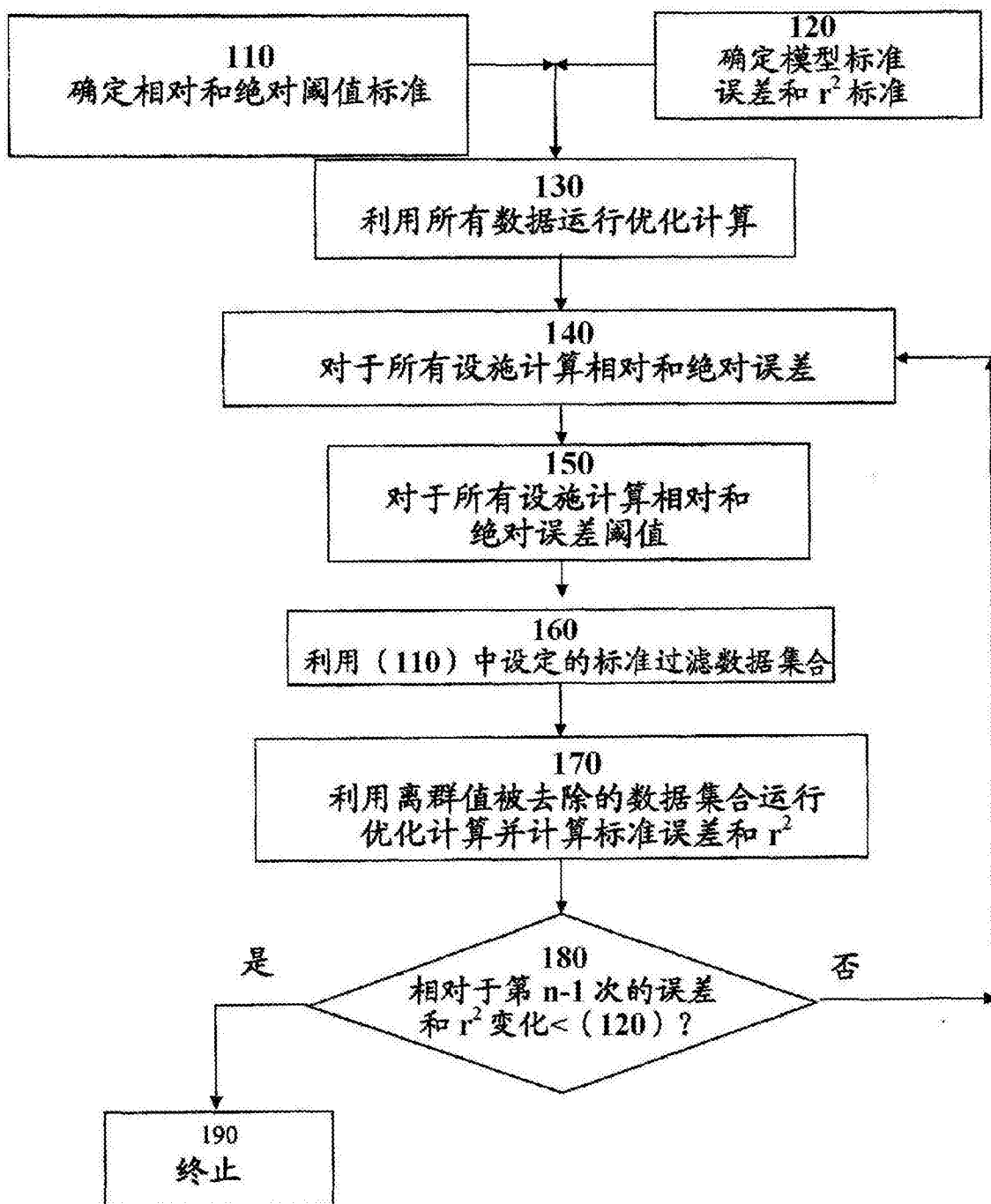


图1

步骤210	步骤220		步骤230		步骤240			步骤250		步骤260
初始数据	排名数据	排名函数计算	误差计算		基于百分位的质量标准			经质量筛选的数据		完整数据函数值
		函数:(算术平均)	相对误差	绝对误差	相对误差	质量标准%	质量标准值	许可/去除(1/0)	经验证的数据	离散值偏倚减少函数值
0.576	0.030	0.030	0.94336	0.000		90%	0.215	0		0.546
0.339	0.088	0.059	0.28177	0.001	绝对误差	90%	0.003	0		
0.433	0.153	0.091	0.20717	0.001				1	0.153	
0.403	0.255	0.132	0.05772	0.002				1	0.255	
0.567	0.290	0.163	0.03192	0.001				1	0.290	
0.701	0.339	0.193	0.02442	0.001				1	0.339	
0.514	0.403	0.223	0.01393	0.001				1	0.403	
0.280	0.433	0.249	0.00738	0.001				1	0.433	
0.255	0.441	0.270	0.00815	0.000				1	0.441	
0.765	0.514	0.295	0.00708	0.001				1	0.514	
0.618	0.567	0.320	0.00448	0.001				1	0.567	
0.775	0.576	0.341	0.00392	0.000				1	0.576	
0.909	0.618	0.362	0.00364	0.000				1	0.618	
0.088	0.668	0.384	0.00271	0.000				1	0.668	
0.153	0.684	0.404	0.00211	0.000				1	0.684	
2.000	0.701	0.423	0.00203	0.000				1	0.701	
0.744	0.744	0.442	0.00166	0.000				1	0.744	
0.684	0.765	0.460	0.00130	0.000				1	0.765	
0.668	0.775	0.476	0.00206	0.000				1	0.775	
0.441	0.909	0.498	0.02066	0.000				1	0.909	
0.030	2.000	0.569		0.005				0		

图2

步骤 310		步骤 320		步骤 330		步骤 340		步骤 350	
初始数据值		误差计算		数据验证标准		经验证的数据集		结果比较	
数据集合 1	数据集合 2	相对误差	绝对误差	质量标准 %	质量标准值	许可/去除 (1/0)	经质量筛选的数据值	高群值偏离减少系数	原始皮尔逊系数
68,991	26,105	0.386	1.8E+09	90%	0.945	1	68,991	0.933	0.915
8,446	8,997	0.004	3.0E+05	90%	8,439,995	1	8,446		
2,833	3,890	0.164	1.3E+06			1	2,833		
3,581	7,722	1.338	1.7E+07			0			
3,021	1,040	0.430	3.8E+06			1	3,021		1,040
8,693	6,904	0.042	3.2E+06			1	8,693		6,904
2,788	4,654	0.448	3.5E+06			1	2,788		4,654
1,625	878	0.211	5.6E+05			1	1,625		878
1,133	822	0.075	9.6E+04			1	1,133		822
1,635	2,715	0.436	1.2E+06			1	1,635		2,715
2,375	4,630	0.902	5.1E+06			1	2,375		4,630
1,044	1,493	0.195	2.0E+05			1	1,044		1,493
1,308	619	0.278	4.8E+05			1	1,308		619
1,581	3,695	1.870	4.6E+06			0			
1,265	1,292	0.000	6.8E+02			1	1,265		1,292
1,047	1,686	0.371	4.1E+05			1	1,047		1,686
723	1,360	0.777	4.1E+05			1	723		1,360
3,213	1,218	0.385	4.0E+06			1	3,213		1,218
1,475	2,047	0.151	3.3E+05			1	1,475		2,047
1,460	2,813	0.856	1.8E+06			1	1,460		2,813
980	377	0.379	3.6E+05			1	980		377
1,157	500	0.322	4.3E+05			1	1,157		500
1,166	757	0.123	1.7E+05			1	1,166		757
1,628	89	0.894	2.4E+06			1	1,628		89
1,466	1,291	0.014	3.1E+04			1	1,466		1,291
5,596	8,450	0.263	8.2E+06			1	5,596		8,450
4,181	4,030	0.001	2.3E+04			1	4,181		4,030
1,659	2,290	0.145	4.0E+05			1	1,659		2,290
3,072	3,576	0.027	2.5E+05			1	3,072		3,576
860	479	0.197	1.5E+05			1	860		479
1,752	280	0.706	2.2E+06			1	1,752		280
1,786	2,146	0.041	1.3E+06			1	1,786		2,146
1,511	577	0.383	8.7E+05			1	1,511		577
2,467	2,815	0.020	1.2E+05			1	2,467		2,815

图3

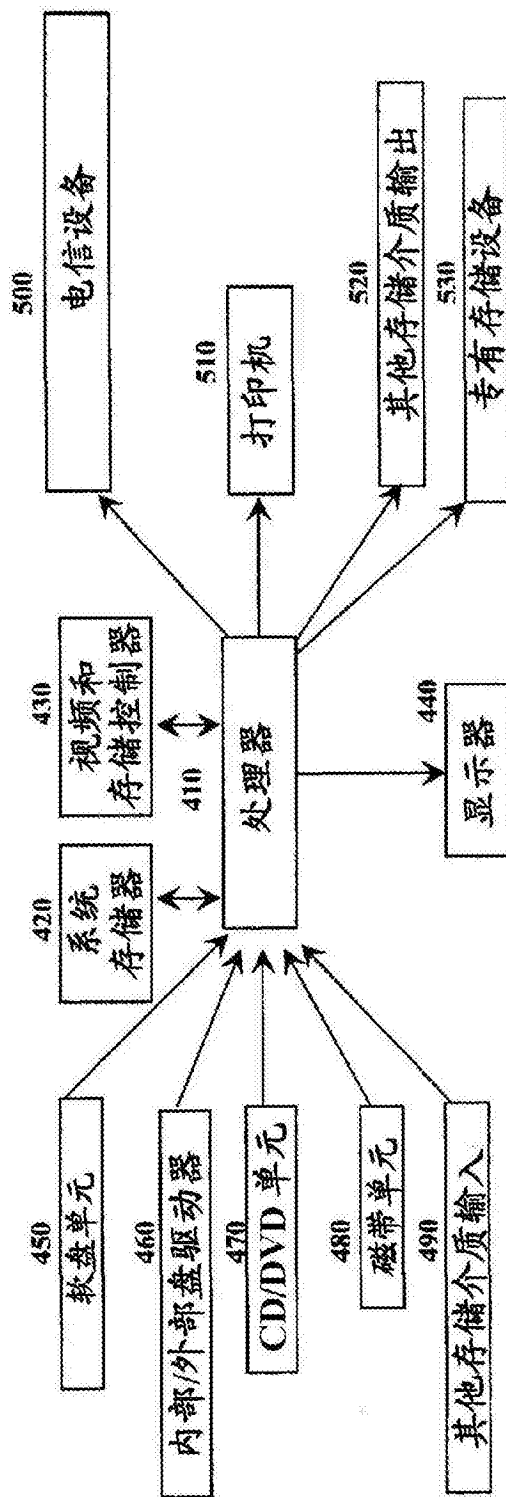


图4

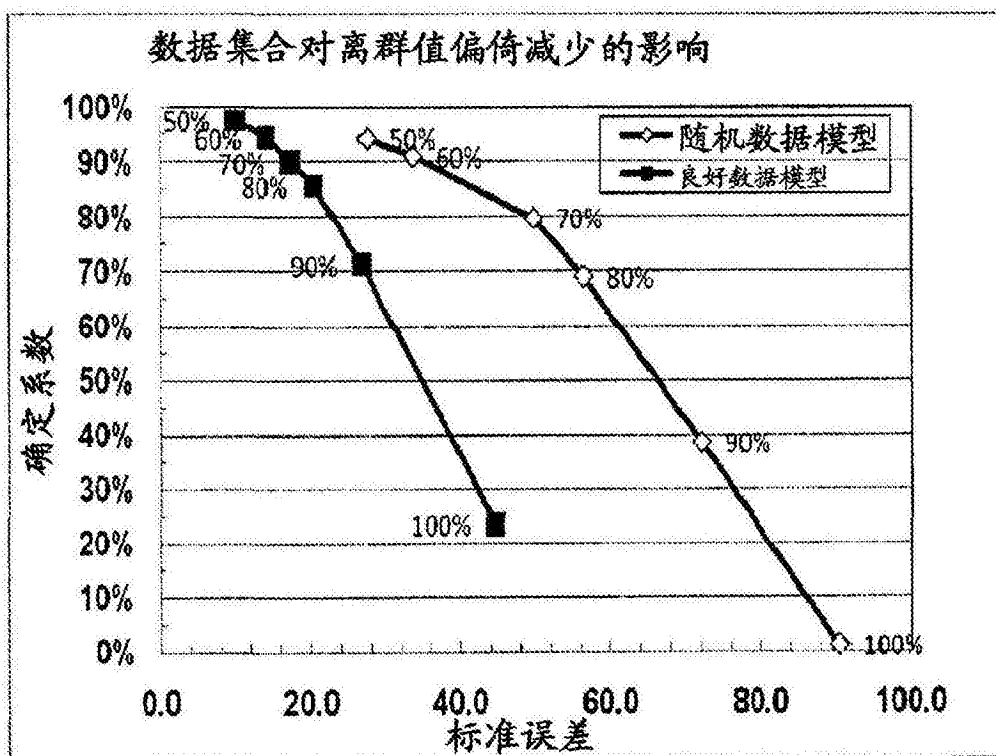


图5

随机数据集合：没有离群值偏倚减少

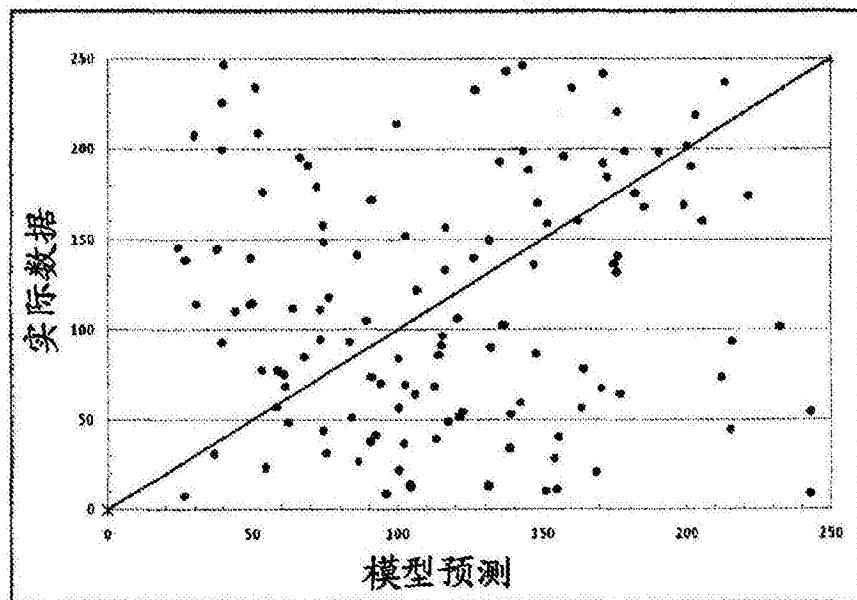


图6A

真实模型数据集合: 没有离群值偏倚减少

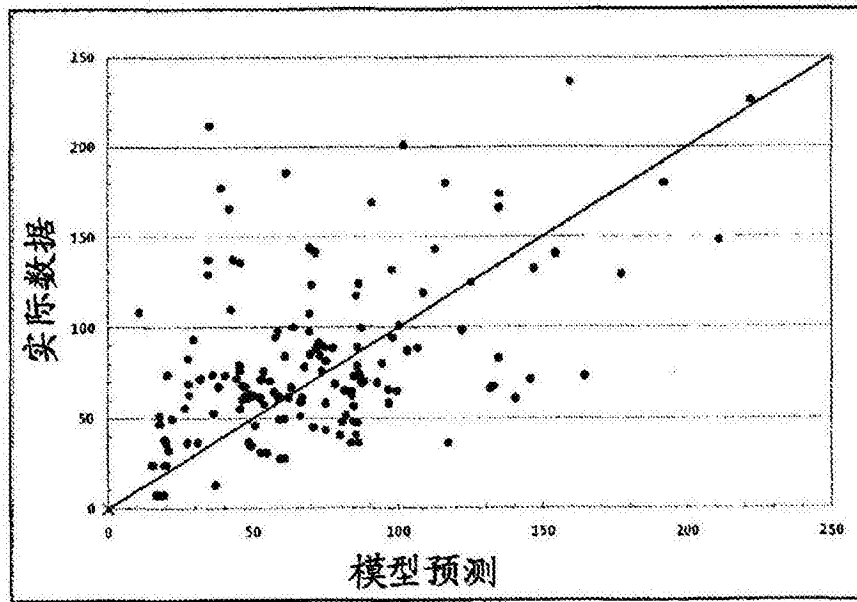


图6B

随机数据集合: 离群值偏倚减少 (70%)
~30%的数据作为离群值被去除

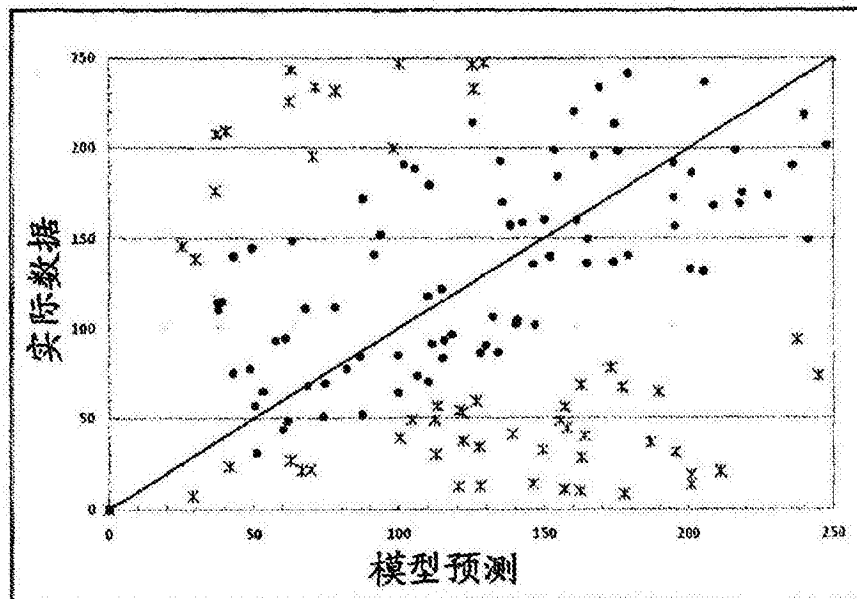


图7A

真实模型数据集合：离群值偏倚减少（70%）
~30%的数据作为离群值被去除

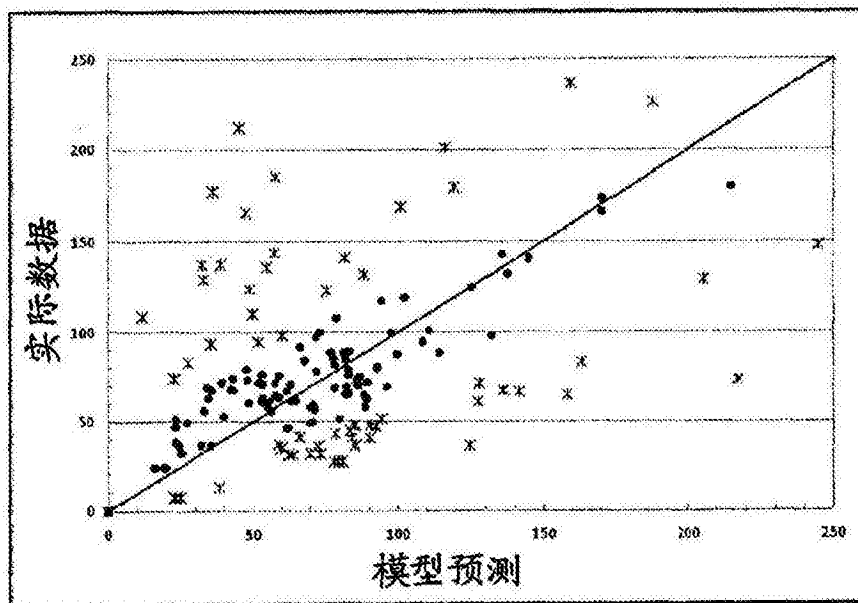


图7B

随机数据集合：离群值偏倚减少（50%）
~50%的数据作为离群值被去除

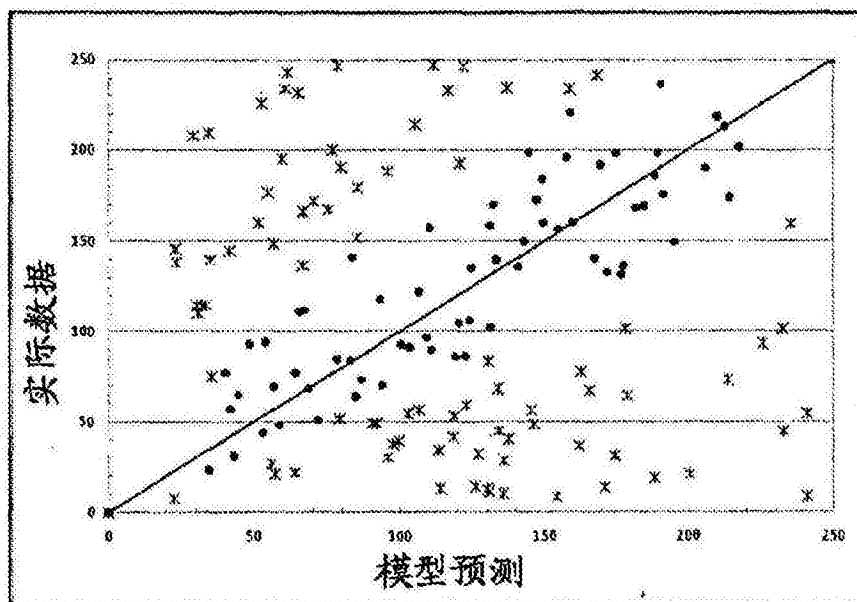


图8A

真实模型数据集：离群值偏倚减少（50%）
~50%的数据作为离群值被去除

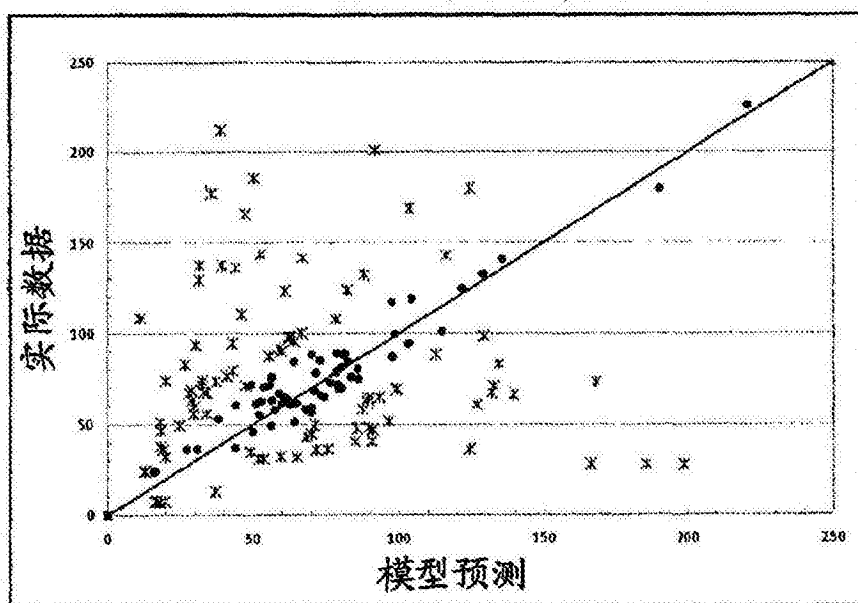


图8B