说明书摘要

本发明公开了一种惰性降维方法、系统、计算机设备及存储介质，方法包括：获取高维数据；计算高维数据中的各个属性下对应样本数据的方差，并将各个属性按照其对应的方差大小进行排序；按照帕累托法则从按方差排序后的属性中确定第一组属性；基于主成分分析法对按方差排序后的属性中的其余属性进行主成分计算以得到主成分属性，并对主成分属性按照其对应特征值大小进行排序；按照帕累托法则从按主成分排序后的主成分属性中确定第二组属性；判断确定的各组属性的数量之和是否满足预设条件，并基于判断结果确定最终的降维数据。通过本发明的方案，能够极大程度地保留高维数据原始属性信息以及提取高维数据信息量，得到完整的降维数据。

摘要附图

S10

获取高维数据，所述高维数据的每一行表示样本数据，每一列表示样本属性

S20

计算所述高维数据中的各个属性下对应样本数据的方差，并将各个属性按照其对应的方差大小进行排序

S30

按照帕累托法则从按方差排序后的属性中确定第一组属性

S40

基于主成分分析法对所述按方差排序后的属性中的其余属性进行主成分计算以得到主成分属性，并对所述主成分属性按照其对应特征值大小进行排序

S50

按照帕累托法则从按主成分排序后的主成分属性中确定第二组属性

S60

判断上述确定的各组属性的数量之和是否满足预设条件，并基于判断结果确定最终的降维数据

权利要求书

1．一种惰性降维方法，其特征在于，包括：

获取高维数据，所述高维数据的每一行表示样本数据，每一列表示样本属性；

计算所述高维数据中的各个属性下对应样本数据的方差，并将各个属性按照其对应的方差大小进行排序；

按照帕累托法则从按方差排序后的属性中确定第一组属性；

基于主成分分析法对所述按方差排序后的属性中的其余属性进行主成分计算以得到主成分属性，并对所述主成分属性按照其对应特征值大小进行排序；

按照帕累托法则从按主成分排序后的主成分属性中确定第二组属性；

判断上述确定的各组属性的数量之和是否满足预设条件，并基于判断结果确定最终的降维数据。

2．根据权利要求1所述的方法，其特征在于，所述将各个属性按照其对应的方差大小进行排序包括：

将各个属性按照其对应的方差大小从大到小进行排序。

3．根据权利要求2所述的方法，其特征在于，所述按照帕累托法则从按方差排序后的属性中确定第一组属性包括：

按照帕累托法则将按方差排序后的属性中前20%的属性确定为第一组属性。

4．根据权利要求1所述的方法，其特征在于，所述对所述主成分属性按照其对应特征值大小进行排序，包括：

对所述主成分属性按照其对应特征值大小从大到小进行排序；

按照帕累托法则从按主成分排序后的主成分属性中确定第二组属性包括：

按照帕累托法则将所述按主成分排序后的主成分属性中前20%的属性确定为第二组属性。

5．根据权利要求1所述的方法，其特征在于，所述基于判断结果确定最终的降维数据包括：

响应于上述确定的各组属性的数量之和满足预设条件，组合各组属性下的数据以得到降维数据。

6．根据权利要求1所述的方法，其特征在于，进一步包括：

响应于上述确定的各组属性的数量之和不满足预设条件，返回基于主成分分析法对所述按方差排序后的属性中的其余属性进行主成分计算的步骤以对主成分排序后的除第二组数据以外的其余属性进行主成分计算，直到上述确定的各组属性的数量之和满足预设条件。

7．根据权利要求1所述的方法，其特征在于，所述获取高维数据之后，方法进一步包括：

对所述高维数据进行缺失值处理、异常值处理、重复值处理以及变量编码处理，以得到处理后的高维数据。

8．一种惰性降维系统，其特征在于，包括：

获取模块，所述获取模块配置为获取高维数据，所述高维数据的每一行表示样本数据，每一列表示样本属性；

方差排序模块，所述方差排序模块配置为计算所述高维数据中的各个属性下样本数据的方差，并将各个属性按照其对应的方差大小进行排序；

确定模块，所述确定模块配置为按照帕累托法则从按方差排序后的属性中确定第一组属性；

主成分计算模块，所述主成分计算模块配置为基于主成分分析法对所述按方差排序后的属性中的其余属性进行主成分计算以得到主成分属性，并对所述主成分属性按照其对应特征值大小进行排序；

所述确定模块还配置为按照帕累托法则从按主成分排序后的主成分属性中确定第二组属性；

判断降维模块，所述判断降维模块配置为判断上述确定的各组属性的数量之和是否满足预设条件，并基于判断结果确定最终的降维数据。

9．一种计算机设备，包括：

至少一个处理器；以及

存储器，所述存储器存储有可在所述处理器上运行的计算机程序，其特征在于，所述处理器执行所述程序时执行如权利要求1至7任意一项所述的方法的步骤。

10．一种计算机可读存储介质，所述计算机可读存储介质存储有计算机程序，其特征在于，所述计算机程序被处理器执行时执行如权利要求1至7任意一项所述的方法的步骤。

说明书

一种惰性降维方法、系统、计算机设备及存储介质

技术领域

本发明涉及数据科学技术领域，尤其涉及一种惰性降维方法、系统、计算机设备及存储介质。

背景技术

数据降维是一种常见的数据处理方法，常见于数学建模过程。在数据维度很高时，数学建模需要庞大的计算资源，所以希望可以将高维数据压缩至低维数据，且不影响原始数据的质量，在加速计算的情况下，达到预期的效果。

数据降维的问题在于降维后的数据会失去数据原本的含义，数据属性含糊，甚至无法直观地解释。另外，原始数据属性本身具有意义，属性之间的重要性或许不同，而单纯地叠加数据内部信息量去降维会造成数据失真，导致预测偏差。

发明内容

有鉴于此，本发明提出了一种惰性降维方法、系统、计算机设备及存储介质，改善了传统数据降维中仅对数据进行单纯叠加导致数据失真、预测偏差大等问题。

基于上述目的，本发明实施例的一方面提供了一种惰性降维方法，具体包括如下步骤：

获取高维数据，所述高维数据的每一行表示样本数据，每一列表示样本属性；

计算所述高维数据中的各个属性下对应样本数据的方差，并将各个属性按照其对应的方差大小进行排序；

按照帕累托法则从按方差排序后的属性中确定第一组属性；

基于主成分分析法对所述按方差排序后的属性中的其余属性进行主成分计算以得到主成分属性，并对所述主成分属性按照其对应特征值大小进行排序；

按照帕累托法则从按主成分排序后的主成分属性中确定第二组属性；

判断上述确定的各组属性的数量之和是否满足预设条件，并基于判断结果确定最终的降维数据。

在一些实施方式中，所述将各个属性按照其对应的方差大小进行排序包括：

将各个属性按照其对应的方差大小从大到小进行排序。

在一些实施方式中，所述按照帕累托法则从按方差排序后的属性中确定第一组属性包括：

按照帕累托法则将按方差排序后的属性中前20%的属性确定为第一组属性。

在一些实施方式中，所述对所述主成分属性按照其对应特征值大小进行排序，包括：

对所述主成分属性按照其对应特征值大小从大到小进行排序；

按照帕累托法则从按主成分排序后的主成分属性中确定第二组属性包括：

按照帕累托法则将所述按主成分排序后的主成分属性中前20%的属性确定为第二组属性。

在一些实施方式中，所述基于判断结果确定最终的降维数据包括：

响应于上述确定的各组属性的数量之和满足预设条件，组合各组属性下的数据以得到降维数据。

在一些实施方式中，方法进一步包括：

响应于上述确定的各组属性的数量之和不满足预设条件，返回基于主成分分析法对所述按方差排序后的属性中的其余属性进行主成分计算的步骤以对主成分排序后的除第二组数据以外的其余属性进行主成分计算，直到上述确定的各组属性的数量之和满足预设条件。

在一些实施方式中，所述获取高维数据之后，方法进一步包括：

对所述高维数据进行缺失值处理、异常值处理、重复值处理以及变量编码处理，以得到处理后的高维数据。

本发明实施例的另一方面，还提供了一种惰性降维系统，包括：

获取模块，所述获取模块配置为获取高维数据，所述高维数据的每一行表示样本数据，每一列表示样本属性；

方差排序模块，所述方差排序模块配置为计算所述高维数据中的各个属性下样本数据的方差，并将各个属性按照其对应的方差大小进行排序；

确定模块，所述确定模块配置为按照帕累托法则从按方差排序后的属性中确定第一组属性；

主成分计算模块，所述主成分计算模块配置为基于主成分分析法对所述按方差排序后的属性中的其余属性进行主成分计算以得到主成分属性，并对所述主成分属性按照其对应特征值大小进行排序；

所述确定模块还配置为按照帕累托法则从按主成分排序后的主成分属性中确定第二组属性；

判断降维模块，所述判断降维模块配置为判断上述确定的各组属性的数量之和是否满足预设条件，并基于判断结果确定最终的降维数据。

本发明实施例的又一方面，还提供了一种计算机设备，包括：至少一个处理器；以及存储器，所述存储器存储有可在所述处理器上运行的计算机程序，所述计算机程序由所述处理器执行时实现如上方法的步骤。

本发明实施例的再一方面，还提供了一种计算机可读存储介质，计算机可读存储介质存储有被处理器执行时实现如上方法步骤的计算机程序。

本发明至少具有以下有益技术效果：通过对原始数据进行方差计算以及帕累托法则对原始数据按照方差计算选取第一组数据，并通过递归主成分分析法对剩余数据进行计算以及通过帕累托法则对主成分分析计算的数据进行多次选取，极大程度地保留了高维数据的原始属性信息和极大程度地提取了高维数据信息量，得到了高质量的降维数据。

附图说明

为了更清楚地说明本发明实施例或现有技术中的技术方案，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的前提下，还可以根据这些附图获得其他的实施例。

图1为本发明提供的惰性降维方法的一实施例的框图；

图2为本发明提供的惰性降维系统的一实施例的示意图；

图3为本发明提供的计算机设备的一实施例的结构示意图；

图4为本发明提供的计算机可读存储介质的一实施例的结构示意图。

具体实施方式

为使本发明的目的、技术方案和优点更加清楚明白，以下结合具体实施例，并参照附图，对本发明实施例进一步详细说明。

需要说明的是，本发明实施例中所有使用“第一”和“第二”的表述均是为了区分两个相同名称非相同的实体或者非相同的参量，可见“第一”“第二”仅为了表述的方便，不应理解为对本发明实施例的限定，后续实施例对此不再一一说明。

基于上述目的，本发明实施例的第一个方面，提出了一种惰性降维方法的实施例。如图1所示，其包括如下步骤：

S10、获取高维数据，所述高维数据的每一行表示样本数据，每一列表示样本属性；

S20、计算所述高维数据中的各个属性下对应样本数据的方差，并将各个属性按照其对应的方差大小进行排序；

S30、按照帕累托法则从按方差排序后的属性中确定第一组属性；

S40、基于主成分分析法对所述按方差排序后的属性中的其余属性进行主成分计算以得到主成分属性，并对所述主成分属性按照其对应特征值大小进行排序；

S50、按照帕累托法则从按主成分排序后的主成分属性中确定第二组属性；

S60、判断上述确定的各组属性的数量之和是否满足预设条件，并基于判断结果确定最终的降维数据。

为了解决现有技术中存在的数据降维后属性含糊、数据失真和预测偏差等问题，本发明实施例提出一种高保留且解释性强的数据降维方案，即，一种惰性降维方法。通过方差计算测度各个属性的信息量，通过帕累托法则和递归的主成分分析法，极大程度地保留原始属性信息，并穷尽提取高维数据的信息量。基于此，可得到高质量的降维数据，该降维数据在投入实际使用时，能够得到更为准确的预测结果。

步骤S10中，获取需要进行降维处理的高维数据，假设高维数据**X**的维度为n•m，n为样本数量，m为属性数量，降维的目的是减小m。

步骤S20中，计算高维数据中各个属性下对应样本数据的方差，方差大小测度属性信息量的大小，将属性按照方差的大小从大到小进行排序。以此极大程度地保留高维数据的原始属性信息，避免对高维数据进行降维时数据过分失真。

步骤S30中，帕累托法则，即二八法则，用20%的数据可代表80%的数据。帕累托法则认为：原因和结果、投入和产出、努力和报酬之间本来存在着无法解释的不平衡。一般来说，投入和努力可以分为两种不同的类型：多数，它们只能造成少许的影响；少数，它们造成主要的、重大的影响。80/20关系提供了一个较好的基准。一个典型的模式表明，80%的产出源自20%的投入；80%的结论源自20%的起因；80%的收获源自20%的努力。

因此，基于帕累托法则，将前20%的属性确定为第一组属性，即0.2m。

步骤S40中，主成分分析法（Principal Component Analysis，简称PCA）通过构建一种被称为主成分的变量，并将所用到的所有向量映射到由主成分变量构建的空间上去，从而能够减少数据的维度。

在高维数据中，将上述基于帕累托法则确定的第一组属性以外的80%属性，通过主成分分析法（Principal Component Analysis，简称PCA）进行计算，得到80%的主成分属性，即0.8m。主成分分析法利用属性之间的协方差矩阵，通过矩阵正交分解技术，得到特征值和特征值向量，将特征值向量作用于属性上，得到主成分属性。特征值大小表示其包含的信息量的多少。将得到的80%的主成分属性按特征值大小从大到小进行排序，保证了排在前面的主成分属性为包含信息量多的主成分。

步骤S50中，和步骤S20一样，按照帕累托法则，将前20%的主成分属性确定为第二组属性，即0.8•0.2m。

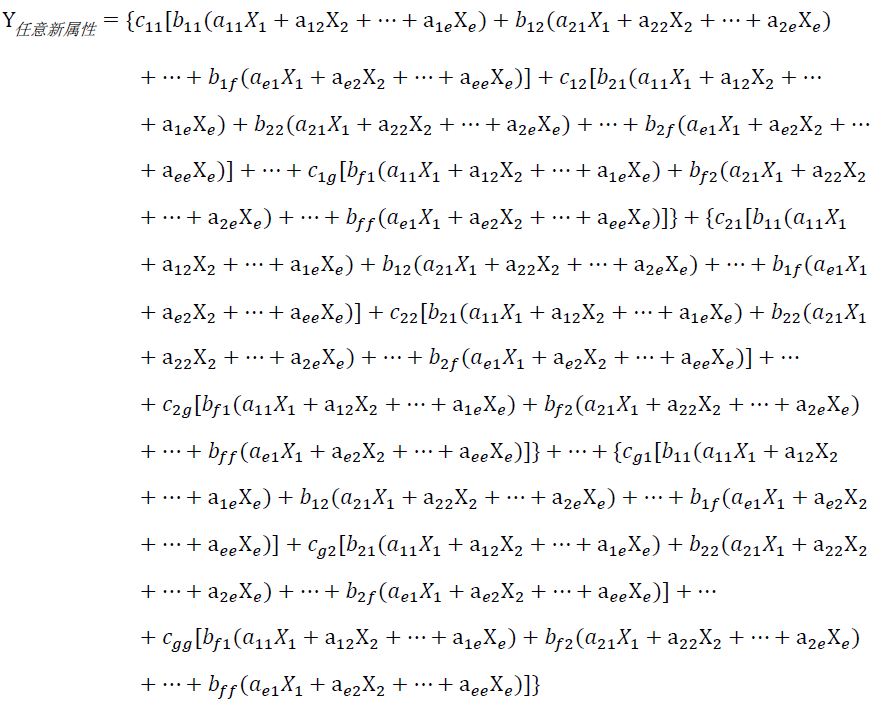
步骤S60中，判断之前确定的各组属性的数量之和是否降到指定数量，在本实施例中，即0.8m+0.8•0.2m是否大于阈值。

若是此时各组属性的数量之和大于阈值，则将之前各组的属性组合，得到降维后的数据；

若是此时各组属性的数量之和不大于阈值，返回步骤S40继续对确定第二组属性后的其余数据进行主成分计算，得到新的主成分属性，直到此时各组属性的数量大于阈值结束，得到降维后的数据。

本实施例在步骤S60中基于判断条件判断是否递归（循环执行）S40、S50的步骤，根据定义阈值的不同，递归次数不等。PCA降维过程中的矩阵分解，即正交化的过程，特征值的大小代表新一代主成分所含有上一代主成分的信息量，可通过PCA计算保证降维过程中的信息完整度。其中，递归降维后期是计算主成分，而主成分又是原始属性的线性转换，那么偏后（第一次主成分计算之后）的计算过程每个新一代主成分涉及到呈指数形式增长的复杂度，为保证计算速度，优先递归次数，即循环次数优选为1~5次。

下面展示递归3次的计算过程：



其中，表示第一次主成分计算的系数矩阵，表示第二次主成分计算的系数矩阵，表示第三次主成分计算的系数矩阵。其中，e=0.8m，f=0.8•0.8m，g=0.8•0.8•0.8m，分别代表对应组次主成分计算的属性数量。

本发明实施例的惰性降维方法，在惰性降维过程中，通过对原始数据进行方差计算以及帕累托法则对原始数据按照方差计算选取第一组数据，并通过递归主成分分析法对剩余数据进行计算以及通过帕累托法则对主成分分析计算的数据进行多次选取，在保留原始属性信息基础上，极大程度地提取高维数据信息量，得到高质量的降维数据。

下面对本发明实施例的降维方法的应用场景进行举例说明。

应用场景一、银行信贷模型

在银行信贷业务中，可通过算法模型来衡量贷款客户的信用风险，判断是否应该借钱给该客户，例如逻辑回归模型。但是银行业务的数据量庞大，高维度金融数据屡见不鲜，数学建模时会消耗大量的计算资源、存贮空间和时间，而且高维度数据建模的效果也未必好，因此，降维操作就显得非常必要。可是，降维后的数据属性，是不具备实际意义的，信贷人员使用模型解释贷款客户是否符和贷款要求时会陷入困境，所以可将本发明实施例的惰性降维方法应用于构建信贷模型，解决此困境。表1为银行业务数据属性基本说明，银行业务的属性数据较多，一般多余20个，本实施例仅以20个属性进行说明使用。

表1

|  |  |  |
| --- | --- | --- |
| 属性 | 名称 | 备注 |
| X1 | 服务年数 |  |
| X2 | 月收入 |  |
| X3 | 家属数量 |  |
| X4 | 可用额度比值 |  |
| X5 | 是否理财客户 |  |
| X6 | 抵押贷款 |  |
| X7 | 目前贷款数量 |  |
| X8 | 目前贷款总额 |  |
| X9 | 贷记卡额度占用率 |  |
| X10 | 合同利率 |  |
| X11 | 合同期限 |  |
| X12 | 用户负债率 |  |
| X13 | 主担保方式 | 1：信用担保类型；2：保证担保类型；3：抵押担保类型；4：质押担保类型 |
| X14 | 贷款近6个月平均应还款 |  |
| X15 | 贷款预期30-59天次数 |  |
| X16 | 贷款预期60-89天次数 |  |
| X17 | 贷款预期90天以上次数 |  |
| X18 | 贷前6个月月均贷方发生金额 |  |
| X19 | 贷前6个月月均贷方发生笔数 |  |
| X20 | 贷前6个月存款日均金额 |  |
| Y | 是否违约 | 标签(0，1) |

基于惰性数据降维的信贷模型流程如下：

1）从银行数据库中获取数据**X**；

2）清洗数据，缺失值、异常值、重复值、变量编码，保证不增加属性；

3）计算银行业务数据各属性的方差，并将其从大到小排序；

4）按照帕累托法则，将前20%的属性确定为第一组属性，即4个；

5）将其余80%的属性应用PCA进行降维，得到80%的主成分属性，即16个，将得到16个主成分属性按特征值大小排序；

6）按照帕累托法则，将前20%的主成分属性确定为第二组属性，即3个；将其余13个主成分属性应用PCA降维，得到新主成分属性，即13个；

7）递归步骤5）和6），直至各组属性的数量相加达到阈值，例如原始属性数量的1/2或2/3或1/3，本实施例假设返回步骤两次达到阈值。

此时，得到降维数据属性为第一组属性(4)+第二组属性(3)+第三组属性(2)+第四组属性(2)=11。

假设A1+A2+A3+A4+B1+B2+B3+C1+C2+D1+D2=11，其中**A**属性为原始属性，因信息量高而保留下来，一般情况下**B**，**C**，**D**信息量依次递减,为后文表示方便，A1+A2+A3+A4+B1+B2+B3+C1+C2+D1+D2—> Z1+Z2+Z3+Z4+Z5+Z6+Z7+Z8+Z9+Z10+Z11。

此时，数据维度下降约50%，但不限于此，仍可计算信息量的留存，理论上可穷尽，取决于降维的终止条件。

8）建立信贷模型

a）常规线性建模方式：

其中，为线性模型的系数，M表示中间变量，Y为模型的激活结果。

b）确定模型的损失函数为交叉熵：

其中，**w**为待求解参数向量，n为业务样本数量，为业务样本的实际标签，为基于业务数据的计算标签，表示业务样本的取值，为惩罚项。

c）使用优化方法最小化损失函数

使用该降维数据，训练信贷模型，即，最小化损失函数。

d）评估信贷模型

利用精确度、召回率和AUC面积等评价指标，来评估银行信贷模型。

9）模型上线使用

应用场景二、降维聚类

降维聚类可以作为一种独立的业务方法存在，或应用于数据处理中段，其降维的意义在于缩减计算量，以kmeans聚类为例，确定聚类中心后每次聚类都需要计算所有点到聚类中心的距离，且这一过程要重复多次，若数据的维度过高，非常容易陷入维度灾难。在使用聚类时，都会先对数据进行降维处理。聚类算法在某些情况下也需要考虑属性重要性度量，不可将所有属性混为一谈，而常见的降维方法默认数据为统一度量，而丢掉了属性这部分信息。因此基于本发明实施例的惰性降维方法，得到降维数据后，再进行聚类，可以实现更好的聚类效果。降维聚类的应用，递归降维的次数可适量减少，因为偏后的递归降维意义甚小。

数据降维聚类的具体过程如下：

1）获取高维数据**X**，**X**∈Rn·m，n为样本数量，m为属性数量；

2）清洗数据，缺失值、异常值、重复值、变量编码，保证不增加属性；

3）按照帕累托法则，将前20%的属性确定为第一组属性，即0.2m；

4）将其余80%的属性应用PCA进行降维，得到80%的主成分属性，即0.8m，将得到的0.8m主成分属性按特征值大小从大到小排序；

5）按照帕累托法则，将前20%的主成分属性确定为第二组属性，即0.8•0.2m；将其余主成分属性应用PCA降维，得到新主成分属性，即0.8•0.8个；

6）递归步骤4）和5），直至各组属性的数量相加达到阈值，例如原始属性数量的1/2或2/3或1/3，本实施例假设返回步骤两次达到阈值。

由此得到的降维数据为0.2m(第一属性)+0.8·0.2m(第二属性)+ 0.8·0.8·0.2m(第三属性)+0.8·0.8·0.8·0.2m(第四属性)=0.5904m

7）Kmeans聚类

(a) 随机选取K个初始中心，,…**;**

(b) 定义损失函数，其中，表示样本**x**第i个属性，表示中心第i个维度；

(c) 令t=1，2，3…，重复步骤(c1、c2)，直到J收敛或主动停止：

(c1) —>，其中，表示第t次迭代的聚类中心，表示第t次聚类计算样本所划分到的类中心；

(c2) —>**，**其中，b表示该类内样本数量，表示表示第t+1次迭代的聚类中心，表示所在的类中心，为待优化参数。

8）聚类效果评价

聚类的效果可根据簇内与簇间数据的稠密程度度量，此处为方便展示，引入轮廓系数：

样本与其自身所在簇中其它样本的相似度记为，其中，表示簇内的任意两样本，b1表示簇内样本的数量，c表示类中心。

样本与最近簇中的样本相似度记为，其中，表示各相邻簇中的任意两样本，b1表示第一簇内样本的数量，b2表示第二簇内样本的数量，c表示类中心。

单个样本的轮廓系数，s的值越接近1越好。

10）完成降维聚类

以上为本发明实施例的应用场景，除此之外，本发明提供的数据惰性降维方法还可以在系统降噪、系统推荐中使用，但不限于此。降维数据均在极大程度地保留了高维数据的原始属性信息的基础上，极大程度地提取了高维数据信息量，保证了降维数据的完整性。因此将该降维数据应用于各类场景的模型时，可以提高模型预测的准确性。

在一些实施方式中，所述将各个属性按照其对应的方差大小进行排序包括：

将各个属性按照其对应的方差大小从大到小进行排序。

具体的，各个属性下样本数据的方差测度属性对应信息量的大小，将属性按照方差的大小从大到小进行排序，即属性以信息量递减的方式呈现在高维数据中，以此保证后续通过帕累托法则确定第一组属性时，能够尽量保留原始属性信息，避免获得的降维数据产生数据失真、属性模糊等问题。

在一些实施方式中，所述按照帕累托法则从按方差排序后的属性中确定第一组属性包括：

按照帕累托法则将按方差排序后的属性中前20%的属性确定为第一组属性。

在一些实施方式中，所述对所述主成分属性按照其对应特征值大小进行排序，包括：

对所述主成分属性按照其对应特征值大小从大到小进行排序；

按照帕累托法则从按主成分排序后的主成分属性中确定第二组属性包括：

按照帕累托法则将所述按主成分排序后的主成分属性中前20%的属性确定为第二组属性。

具体的，本发明实施例中特征值大小对应其包含信息量的大小，将主成分属性按特征值大小从大到小进行排序，保证了排在前面的主成分属性为包含信息量多的主成分，以通过帕累托法则极大程度地提取高维数据信息量、保证数据信息的完整性，较传统的高维数据降维方法，本发明实施例的数据降维方案，作为一种惰性降维方法，明显提高了高维数据降维时留存的信息量。

在一些实施方式中，所述基于判断结果确定最终的降维数据包括：

响应于上述确定的各组属性的数量之和满足预设条件，组合各组属性下的数据以得到降维数据。

在一些实施方式中，方法进一步包括：

响应于上述确定的各组属性的数量之和不满足预设条件，返回基于主成分分析法对所述按方差排序后的属性中的其余属性进行主成分计算的步骤以对主成分排序后的除第二组数据以外的其余属性进行主成分计算，直到上述确定的各组属性的数量之和满足预设条件。

具体的，本实施例中，以此时确定的各组属性的数量之和是否大于阈值为条件判断是否返回主成分计算步骤以循环执行该步骤及其之后的步骤，以此时确定的各组属性的数量之和大于阈值为终止条件停止循环，将之前各组的属性组合，得到降维数据。由于，循环次数增加，会导致计算量递增，因此为兼顾计算速度与降维效果，循环次数优选为1~5次。

在一些实施方式中，所述获取高维数据之后，方法进一步包括：

对所述高维数据进行缺失值处理、异常值处理、重复值处理以及变量编码处理，以得到处理后的高维数据。

基于同一发明构思，根据本发明的另一个方面，如图2所示，本发明的实施例还提供了一种惰性降维系统，包括：

获取模块110，所述获取模块110配置为获取高维数据，所述高维数据的每一行表示样本数据，每一列表示样本属性；

方差排序模块120，所述方差排序模块120配置为计算所述高维数据中的各个属性下样本数据的方差，并将各个属性按照其对应的方差大小进行排序；

确定模块130，所述确定模块130配置为按照帕累托法则从按方差排序后的属性中确定第一组属性；

主成分计算模块140，所述主成分计算模块140配置为基于主成分分析法对所述按方差排序后的属性中的其余属性进行主成分计算以得到主成分属性，并对所述主成分属性按照其对应特征值大小进行排序；

所述确定模块130还配置为按照帕累托法则从按主成分排序后的主成分属性中确定第二组属性；

判断降维模块150，所述判断降维模块150配置为判断上述确定的各组属性的数量之和是否满足预设条件，并基于判断结果确定最终的降维数据。

基于同一发明构思，根据本发明的另一个方面，如图3所示，本发明的实施例还提供了一种计算机设备30，在该计算机设备30中包括处理器310以及存储器320，存储器320存储有可在处理器上运行的计算机程序321，处理器310执行程序时执行如上的方法的步骤。

其中，存储器作为一种非易失性计算机可读存储介质，可用于存储非易失性软件程序、非易失性计算机可执行程序以及模块，如本申请实施例中的所述惰性降维方法对应的程序指令/模块。处理器通过运行存储在存储器中的非易失性软件程序、指令以及模块，从而执行装置的各种功能应用以及数据处理，即实现上述方法实施例的惰性降维方法。

存储器可以包括存储程序区和存储数据区，其中，存储程序区可存储操作系统、至少一个功能所需要的应用程序；存储数据区可存储根据装置的使用所创建的数据等。此外，存储器可以包括高速随机存取存储器，还可以包括非易失性存储器，例如至少一个磁盘存储器件、闪存器件、或其他非易失性固态存储器件。在一些实施例中，存储器可选包括相对于处理器远程设置的存储器，这些远程存储器可以通过网络连接至本地模块。上述网络的实例包括但不限于互联网、企业内部网、局域网、移动通信网及其组合。

基于同一发明构思，根据本发明的另一个方面，如图4所示，本发明的实施例还提供了一种计算机可读存储介质40，计算机可读存储介质40存储有被处理器执行时执行如上方法的计算机程序410。

最后需要说明的是，本领域普通技术人员可以理解实现上述实施例方法中的全部或部分流程，可以通过计算机程序来指令相关硬件来完成，程序可存储于一计算机可读取存储介质中，该程序在执行时，可包括如上述各方法的实施例的流程。其中，程序的存储介质可为磁碟、光盘、只读存储记忆体（ROM）或随机存储记忆体（RAM）等。上述计算机程序的实施例，可以达到与之对应的前述任意方法实施例相同或者相类似的效果。

本领域技术人员还将明白的是，结合这里的公开所描述的各种示例性逻辑块、模块、电路和算法步骤可以被实现为电子硬件、计算机软件或两者的组合。为了清楚地说明硬件和软件的这种可互换性，已经就各种示意性组件、方块、模块、电路和步骤的功能对其进行了一般性的描述。这种功能是被实现为软件还是被实现为硬件取决于具体应用以及施加给整个系统的设计约束。本领域技术人员可以针对每种具体应用以各种方式来实现的功能，但是这种实现决定不应被解释为导致脱离本发明实施例公开的范围。

以上是本发明公开的示例性实施例，但是应当注意，在不背离权利要求限定的本发明实施例公开的范围的前提下，可以进行多种改变和修改。根据这里描述的公开实施例的方法权利要求的功能、步骤和/或动作不需以任何特定顺序执行。上述本发明实施例公开实施例序号仅仅为了描述，不代表实施例的优劣。此外，尽管本发明实施例公开的元素可以以个体形式描述或要求，但除非明确限制为单数，也可以理解为多个。

应当理解的是，在本文中使用的，除非上下文清楚地支持例外情况，单数形式“一个”旨在也包括复数形式。还应当理解的是，在本文中使用的“和/或”是指包括一个或者一个以上相关联地列出的项目的任意和所有可能组合。

所属领域的普通技术人员应当理解：以上任何实施例的讨论仅为示例性的，并非旨在暗示本发明实施例公开的范围（包括权利要求）被限于这些例子；在本发明实施例的思路下，以上实施例或者不同实施例中的技术特征之间也可以进行组合，并存在如上的本发明实施例的不同方面的许多其它变化，为了简明它们没有在细节中提供。因此，凡在本发明实施例的精神和原则之内，所做的任何省略、修改、等同替换、改进等，均应包含在本发明实施例的保护范围之内。

说明书附图

S10

获取高维数据，所述高维数据的每一行表示样本数据，每一列表示样本属性

S20

计算所述高维数据中的各个属性下对应样本数据的方差，并将各个属性按照其对应的方差大小进行排序

S30

按照帕累托法则从按方差排序后的属性中确定第一组属性

S40

基于主成分分析法对所述按方差排序后的属性中的其余属性进行主成分计算以得到主成分属性，并对所述主成分属性按照其对应特征值大小进行排序

S50

按照帕累托法则从按主成分排序后的主成分属性中确定第二组属性

S60

判断上述确定的各组属性的数量之和是否满足预设条件，并基于判断结果确定最终的降维数据

图1



图2



图3



图4