(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 105718600 A
(43)申请公布日 2016.06.29

(21)申请号 201610130663.4

(22)申请日 2016.03.08

(71)申请人 上海晶赞科技发展有限公司
地址 200072 上海市闸北区共和新路912号1501-5室

(72)发明人 汤奇峰 薛守辉

(74)专利代理机构 上海翰信知识产权代理事务所(普通合伙) 31270
代理人 张维东

(51)Int.Cl.
G06F 17/30(2006.01)
G06F 17/24(2006.01)
G06N 99/00(2010.01)
G06Q 30/02(2012.01)

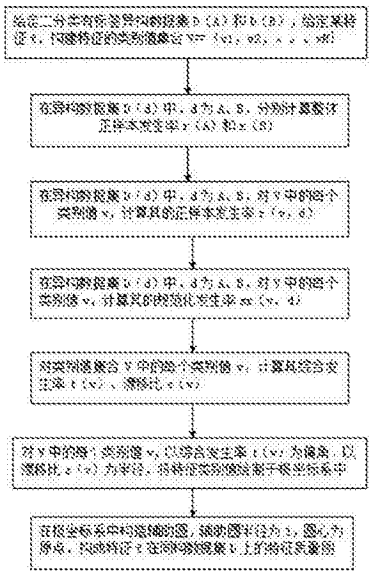
权利要求书2页 说明书7页 附图3页

(54)发明名称

一种异构数据集特征质量可视化方法

(57)摘要

一种异构数据集特征质量可视化方法，通过对异构的训练集和验证集特征分布进行统计，引入特征离散值的发生率，采用异构方法在极坐标系中对特征集合以及特征类别值集合进行可视化，通过计算类别值的正样本发生率、规范化发生率、漂移比、综合发生率，以漂移比为半径、综合发生率为偏角，在极坐标中绘制特征质量图。根据特征可视化方法帮助解决有监督学习中典型的四个特征工程问题：特征评估、特征归因、特征选择、特征改进。本发明使有监督机器学习模型面对领域迁移学习问题或者同领域但数据分布有趋势性漂移时，能够克服训练集和测试集分布差异问题，进而可以进行有效的特征评估、特征归因、特征选择，甚至通过改善特征以提升模型效果。



CN 105718600 A

1.一种异构数据集特征质量可视化方法,其特征在于,至少包括以下步骤:

步骤1,给定二分类有标签异构数据集D(A)和D(B),给定某特征f,构建特征的类别值集合V={v1,v2,。。。vN};

步骤2,在异构数据集D(d)中,d为A、B,分别计算整体正样本发生率r(A)和r(B),计算公式为r(d)=pos(d)/ins(d),pos(d)为异构数据集D(d)中的正样本总数、ins(d)为异构数据集D(d)的样本总数;

步骤3,在异构数据集D(d)中,d为A、B,对V中的每个类别值v,计算其的正样本发生率r(v,d),计算公式为r(v,d)=pos(v,d)/ins(v,d),其中pos(v,d)、ins(v,d)分别为D(d)中包含v的正样本数量和样本总数;

步骤4,在异构数据集D(d)中,d为A、B,对V中的每个类别值v,计算其的规范化发生率sr(v,d),计算公式为:sr(v,d)=r(v,d)/r(d);

步骤5,对类别值集合V中的每个类别值v,计算其综合发生率t(v)、漂移比s(v),计算公式为:t(v)=sr(v,A)+sr(v,B),即类别值v在D(A)和D(B)上的规范化发生率求和;s(v)=sr(v,B)/sr(v,A),即v在D(A)和D(B)上的规范化发生率求比率;

步骤6,对V中的每个类别值v,以综合发生率t(v)为偏角、以漂移比s(v)为半径,将特征类别值绘制于极坐标系中,极坐标p(v)=(t(v),s(v));

步骤7,在极坐标系中构造辅助圆,辅助圆半径为1,圆心为原点,构成特征f在同构数据集D上的特征质量图。

2.根据权利要求1所述的一种异构数据集特征质量可视化方法,其特征在于,所述步骤1中,“构建特征的类别值集合”的方法为:

步骤1.1,判断特征f是否为数值特征,若是,则将特征f采用公式int(log2(c))离散化成类别值,其中c为特征f的特征值,int表示取整,log2表示以2为底取对数,进而在给定数据集上可以得到类别值集合为V0;若不是,则执行步骤1.2;

步骤1.2,设置一阀值,将样本数量少于此阀值的类别值归为一个类别值中,将类别值集合为V0转化为特征类别值集合V。

3.根据权利要求1所述的一种异构数据集特征质量可视化方法,其特征在于,其特征评估流程至少包括以下步骤:

步骤1,对异构数据集D,给定特征集合F,需要选择的特征数量N。

步骤2,计算特征集合F中每个特征在可视化过程中的各项指标数据,包括发生率、规范化发生率、漂移比、综合发生率,构成指标集M;并绘制特征集合F中的每个特征的特征质量图,构成图形集G。

步骤3,根据指标集M和图形集G,对特征集合F中的特征的稳定性和相关度进行评估,得到特征评估结论。

步骤4,根据指标集M和图形集G,判断预测模型的效果瓶颈是特征稳定度还是特征相关度,得到特征归因结论。

步骤5,根据指标集M和图形集G,从特征集合F中选择出来前N个质量好的特征,构成特征选择结果集。

步骤6,根据指标集M和图形集G,对特征集合中部分相关度好、类别值多但稳定性差的特征进行特征改进,采用将具有相近综合发生率的类别值进行聚类,使整体类别数量减少

的同时，提高特征整体稳定性，形成特征改建议。

步骤7，综合特征评估结论、特征归因结论、特征选择结果集、特征改建议构成特征评估报告。

一种异构数据集特征质量可视化方法

技术领域

[0001] 本发明涉及机器学习领域，尤其涉及一种异构数据集特征质量可视化方法。

背景技术

[0002] 近年来，随着大数据行业的发展，很多行业都产生了海量数据，数据种类、数据规模和数据维度都在不断膨胀。为了从大量数据中发现知识和价值，机器学习算法在工业界的应用越来越广泛。除了数据样本不断膨胀，数据特征种类和维度也在迅猛增长，特征维度可以达到千万甚至更大。

[0003] 海量的特征会给后续机器学习算法在可扩展性和效果方面带来一些问题，影响效果的主要原因有两个方面：1)大量特征与预测目标无关或相关程度较低，即特征相关度(FRS,Feature Relevance Score)较差；2)部分特征与预测目标相关程度较高，但其在训练集和测试集(或训练阶段和应用阶段)的分布差异显著，即特征稳定程度(FSL,Feature Stability Level)较差。

[0004] 在有监督学习领域，特征工程是非常重要的环节，而特征工程要解决的问题可以分为：特征评估、特征归因、特征选择和特征改进。传统的特征选择方法，对特征质量的评估往往只考了到特征相关度，例如特征与标签的互信息，而没有将特征稳定性和特征相关性作为一个二元指标进行量化研究或可视化分析。因此本发明既考虑特征相关度、同时兼顾特征稳定度，通过极坐标系将两者构成的指标二元组进行可视化。本发明的特征质量(FQ,Feature Quality)，具体指特征相关度和特征稳定度构成的二元组或者其所表达的特征对于特定预测模型的重要程度。

[0005] 本发明适的领域包括：1)迁移学习，训练集和测试集是跨行业或跨领域的情况；2)非迁移学习，训练集和测试集，不同时间的数据集分布差异较大的情况。

[0006] 在传统的机器学习框架下，学习的任务是在给定充分训练数据的基础上学习一个分类模型或回归模型，然后利用学习到的模型来对测试集样本进行分类或预测。然而实际应用中，经常会看到新的领域涌现，比如从传统的新闻，到网页、图片、博客、播客等，这新的领域或数据集往往缺乏标注；另一方面，传统的机器学习假设训练数据和测试数据服从相同的分布，而实际情况下，这种同分布假设并不满足。因此，如何利用现有的大量有标注、但不通分布的训练数据，迁移知识，用来帮助学习，是迁移学习需要解决的问题。

[0007] 迁移学习(Transfer Learning)的目标是将从一个环境中学习的知识用来帮助新环境的学习任务。迁移学习的重要特点是不对训练集和测试集做同分布假设，即两个数据集是异构的。通常迁移学习中的训练集数据和特征会非常多，因此仅从特征的角度要让从训练集的学习到的模型能够有效预测测试集，就需要对大量特征进行评估、选择，以选出哪些分布变化较小、且与预测目标相关的特征集合。

[0008] 例如在广告转化率模型中，往往从行业数据学习模型，来预测行业内的某个客户的广告是否会转化；或者从一个行业训练模型，来预测类似行业的广告是否会转化。类似这样的迁移学习问题，很有必要通过特征可视化方法来进行特征评估、特征归隐、特征选择和

特征改进。

[0009] 另外，在非迁移学习领域也有类似的训练集和测试集异构的情况，比如广告转化率模型中某客户的平时数据集和节日数据集，如果用平时数据集预测节日数据集可能就导致预测不准问题，因此也是本专利提到的“异构数据集”问题。

发明内容

[0010] 本发明的目的是为了解决现有技术的不足，提供一种异构数据集的极坐标可视化的特征评估与特征选择方法，不仅可以增加对预测问题的直观理解、产生解释性强的特征评估报告，还可以根据特征评估报告进行特征选择和特征改进，以使后续的监督式机器学习模型面对异构数据集时仍可以克服特征不稳定性带来的不利影响，进行更有效的学习。本发明适用于下列情况：1)异构数据集假设下，训练集和测试集产生机理不同、产生领域不同或者有着层级关系，包括典型的迁移学习；2)同构数据集假设下，数据本身随时间产生周期性或非周期性漂移的情况；3)同构数据集假设下，数据本身具有内生性波动，即本质随机性，表现在部分特征分布的方差较大的情况；4)同构数据集假设下，数据分布不变化，即训练集和测试集同分布的情况。

[0011] 本发明的目的是通过以下技术方案实现的：

[0012] 一种异构数据集特征质量可视化方法(Heterogeneous Dataset Feature Quality Visualization，以下简称HeDFQV)，至少包括以下步骤：

[0013] 步骤1，给定二分类有标签异构数据集D(A)和D(B)，给定某特征f，构建特征的类别值集合V＝{v1，v2，。。。vN}；

[0014] 步骤2，在异构数据集D(d)中，d为A、B，分别计算整体正样本发生率r(A)和r(B)，计算公式为r(d)＝pos(d)/ins(d)，pos(d)为异构数据集D(d)中的正样本总数、ins(d)为异构数据集D(d)的样本总数；

[0015] 步骤3，在异构数据集D(d)中，d为A、B，对V中的每个类别值v，计算其的正样本发生率r(v，d)，计算公式为r(v，d)＝pos(v，d)/ins(v，d)，其中pos(v，d)、ins(v，d)分别为D(d)中包含v的正样本数量和样本总数；

[0016] 步骤4，在异构数据集D(d)中，d为A、B，对V中的每个类别值v，计算其的规范化发生率sr(v，d)，计算公式为：sr(v，d)＝r(v，d)/r(d)，其中r(v，d)为数据集D(d)上v的发生率，r(d)为D(d)上整体发生率；

[0017] 步骤5，对类别值集合V中的每个类别值v，计算其综合发生率t(v)、漂移比s(v)，计算公式为：t(v)＝sr(v，A)+sr(v，B)，即类别值v在D(A)和D(B)上的规范化发生率求和；s(v)＝sr(v，B)/sr(v，A)，即v在D(A)和D(B)上的规范化发生率求比率；

[0018] 步骤6，对V中的每个类别值v，以综合发生率t(v)为偏角、以漂移比s(v)为半径，将特征类别值绘制于极坐标系中，极坐标p(v)＝(t(v)，s(v))；

[0019] 步骤7，在极坐标系中构造辅助圆，辅助圆半径为1，圆心为原点，构成特征f在同构数据集D上的特征质量图，完成对特征f的可视化。

[0020] 上述的一种异构数据集特征质量可视化方法，其中，所述步骤1中，“构建特征的类别值集合”的方法为：

[0021] 步骤1.1，判断特征f是否为数值特征，若是，则将特征f采用公式int(log2(c))离

散化成类别值，其中c为特征f的特征值，int表示取整，log2表示以2为底取对数，进而在给定数据集上可以得到类别值集合为V0；若不是，则执行步骤1.2；

[0022] 步骤1.2，设置一阀值，将样本数量少于此阀值的类别值归为一个类别值中，将类别值集合为V0转化为特征类别值集合V。

[0023] 上述的一种异构数据集特征质量可视化方法，其中，其特征评估流程(异构数据集特征可视化评估流程，Homogeneous Dataset Feature Evaluation Pipeline，以下简称HeDFEP)至少包括以下步骤：

[0024] 步骤1，对异构数据集D，给定特征集合F，需要选择的特征数量N。

[0025] 步骤2，计算特征集合F中每个特征在可视化过程中的各项指标数据，包括发生率、规范化发生率、漂移比、综合发生率，构成指标集M；并绘制特征集合F中的每个特征的特征质量图，构成图形集G。

[0026] 步骤3，根据指标集M和图形集G，对特征集合F中的特征的稳定性和相关度进行评估，得到特征评估结论。

[0027] 步骤4，根据指标集M和图形集G，判断预测模型的效果瓶颈是特征稳定度还是特征相关度，得到特征归因结论。

[0028] 步骤5，根据指标集M和图形集G，从特征集合F中选择出来前N个质量好的特征，构成特征选择结果集。

[0029] 步骤6，根据指标集M和图形集G，对特征集合中部分相关度好、类别值多但稳定性差的特征进行特征改进，采用将具有相近综合发生率的类别值进行聚类，使整体类别数量减少的同时，提高特征整体稳定性，形成特征改建议。

[0030] 步骤7，综合特征评估结论、特征归因结论、特征选择结果集、特征改建议构成特征评估报告。

[0031] 本发明可以根据特征质量图判断特征的整体漂移比和整体相关度，具体方法为：根据特征的类别值点集在图中的分布形状，判断该特征的整体质量，当点集分布角坐标方向越散时，特征相关度越好，当点集在轴坐标方向分布越接近标准圆时，特征稳定性越好；一般来说，点集分布会分散在单位圆内外。

[0032] 特征类别值点集在特征质量中的分布形状模式一般有四种，可根据这四种模式判断该特征整体质量(参见图3)：

[0033] (1)长瘦弧模式，即点集基本上分布在圆周上或圆周附近，且在圆周角度方向分布较散，形状类似圆周上的一段长且瘦的弧线，这种特征具有“强相关强稳定”的特征，特征质量最好，参见图3(左上)；

[0034] (2)长胖弧模式，即点集分布在轴向距离圆周较远，且在圆周角度方向分布较散，形状类似圆周上的一段长且胖的弧线，这种特征具有“强相关弱稳定”的特征，特征质量一般，参见图3(右上)；

[0035] (3)短瘦弧模式，即点集分布在轴向距离圆周较较近的区域内，且在圆周角度方向分布较集中，形状类似圆周上的一段短且瘦的弧线，这种特征具有“弱相关强稳定”的特征，特征质量一般，参见图3(左下)；

[0036] (4)射线模式，即点集分布在轴向距离圆周较远，且在圆周角度方向分布较集中，形状类似从原点射出的射线线段，这种特征具有“弱相关弱稳定”的特征，特征质量最差，参

见图3(右下)。

[0037] 还可以通过每个特征类别值的极坐标半径和偏角来衡量每个类别值的漂移比和综合发生率,方法为:根据特征类别值点的位置判断,当点在圆周上时,表示特征发生率无漂移;点有可能分布在单位圆内外,当点在圆周外时,表示该特征类别值的漂移比大于1,即测试集规范化发生率大于训练集,反之,如果如果点在圆周内时,表示该特征类别至的票一笔小于1,即测试集规范化发生率小于训练集;当点的偏角越大,其综合发生率越大。

[0038] 综上所述,与现有技术相比,本发明有以下优点和有益效果:

[0039] 1、本发明提出的特征极坐标可视化方法,首次将特征类别值进行可视化,以二维图形的方式可视化包含相关度和综合发生率两个维度的特征类别指标,将特征的类别值(数值特征需要离散化成类别值)映射成极坐标系中的点,进而根据点的轴向坐标分布判断特征类别值稳定度或漂移比,通过点的角度坐标判断特征类别值的发生率相对于均值的高低水平。

[0040] 2、本发明提出的特征极坐标可视化方法,首次将特征质量进行可视化,以二维图形的方式可视化包含相关度和稳定度两个维度的特征质量,将特征类别集合(或数值特征离散化形成的类别集合)映射成极坐标系中的点集,并提出“特征质量图四种模式判断准则”,进而根据点集的整体形状判断特征相关度、特征稳定度构成的特征质量。

[0041] 3、本发明提出的特征极坐标可视化方法,首次将采用可视化方法对特征进行研究,包括基于特征质量图特征评估方法、基于特征质量图的特征归因方法、基于特征质量图特征选择方法、基于特征质量图特征改进方法。

[0042] 4、本发明提出的特征极坐标可视化方法和极坐标可视化特征评估流程,一方面可以增加对预测问题的直观理解、产生解释性强的特征评估报告,加深对建模问题的理解深度,帮助人工特征选择和特征改进工作,另一方面根据特征评估报告进行特征选择和特征改进,以使后续的监督式机器学习模型面对异构数据集时仍可以克服特征不稳定性带来的不利影响,进行更有效的学习。

附图说明

[0043] 图1是本发明一种异构数据集特征质量可视化方法的流程图。

[0044] 图2是本发明实施例1的特征质量图。

[0045] 图3是本发明的特征质量图分布的四种模式图。

具体实施方式

[0046] 实施例1

[0047] 表1

[0048]

PART 1	pos(v,d)		ins(v,d)		r(v,d)		sr(v,d)			
v	d=A	d=B	d=A	d=B	d=A	d=B	d=A	d=B	t(v)	s(v)
1	99	80	1136	787	0.0871	0.1017	0.8063	1.0924	1.8987	1.3549
2	112	46	938	649	0.1194	0.0709	1.1047	0.7617	1.8664	0.6895
3	119	85	1518	1049	0.0783	0.0810	0.7247	0.8708	1.5955	1.1983
4	206	93	1583	1096	0.1301	0.0849	1.2040	0.9119	2.1159	0.7574
5	112	78	1079	747	0.1038	0.1044	0.9603	1.1222	2.0825	1.1686
6	120	74	1139	783	0.1062	0.0945	0.9825	1.0157	1.9982	1.0338
7	173	186	1325	918	0.1306	0.1144	1.2080	1.2292	2.4372	1.0176
PART 2	pos(d)		ins(d)		r(d)					
	d=A	d=B	d=A	d=B	d=A	d=B				
summary	941	561	8706	6029	0.1081	0.0931				

[0049] 本实施例为某广告主转化率模型，广告主所在行业为电商行业，训练集D(A)为电商行业的样本数据集，简称“行业数据”，测试集D(B)为该广告主的样本数据集，简称“公司数据”。本实施例中，一种异构数据集特征质量可视化方法HeDFQV的步骤如下：

[0050] 步骤1，给定二分类有标签异构数据集D(A)和D(B)，给定某特征f，f为dayofweek，即周几，V＝{1,2,3,4,5,6,7}分别表示周一到周日。构建特征的类别值集合V＝{v1,v2,。。。vN}。

[0051] 步骤2，在D(d)中，d为A、B，分别计算整体正样本发生率r(A)和r(B)，计算公式为r(d)＝pos(d)/ins(d)，pos(d)为正样本总数、ins(d)为样本总数；则训练集A的整体样本发生率为r(A)＝941/8706＝0.1081，训练集B的整体样本发生率r(B)＝561/6029＝0.0931。

[0052] 步骤3，在D(d)中，d为A、B，对类别值集合V中的每个类别值v，计算其的正样本发生率r(v,d)，计算公式为r(v,d)＝pos(v,d)/ins(v,d)，其中pos(v,d)、ins(v,d)分别为D(d)中包含v的正样本数量和样本总数。

[0053] 例如v＝1，d＝A，即计算特征类别值“周一”的正样本发生率，r(1,A)＝pos(1,A)/ins(1,A)＝99/1136＝0.0871，其他数据请参见表1的r(v,d)列。

[0054] 步骤4，在D(d)中，d为A、B，对类别值集合V中的每个类别值v，计算其的规范化发生率sr(v,d)，计算公式为：sr(v,d)＝r(v,d)/r(d)。

[0055] 例如v＝1，d＝B，即计算特征类别至“周一”的规范化发生率，sr(1,B)＝r(1,B)/r(A)＝0.1017/0.0931＝1.0924，其他数据请参见表1sr(v,d)列。

[0056] 步骤5，对V中的每个类别值v，计算其综合发生率t(v)、漂移比s(v)，t(v)＝sr(v,A)+sr(v,B)，即v在D(A)和D(B)上的规范化发生率求和，例如，t(1)＝sr(1,A)+sr(1,B)＝0.8063+1.0924＝1.8987，其他数据请参见表1的t(v)列；s(v)＝sr(v,B)/sr(v,A)，即v在D(A)和D(B)上的规范化发生率求比率，例如，s(1)＝sr(1,B)/sr(1,A)＝1.0924/0.8063＝1.3549请参见表1的s(v)。

[0057] 步骤6，对V中的每个类别值v，以综合发生率t(v)为偏角、以漂移比s(v)为半径，将特征类别值绘制于极坐标系中的一个点；即极坐标p(v)＝(t(v),s(v))，请参见图2，其中，P点为角度坐标t(v)最小的点，此时v＝3，即特征类别值“周三”在极坐标中为点P。

[0058] 步骤7，构造辅助圆，方法为构造单位标准圆，进而构成特征f在数据集D上的特征质量图(Feature Quality Graph，简称FQG)，完成对特征f的可视化。

[0059] 实施例2

[0060] 本实施例为某广告主转化率模型，假设广告主所在行业为电商行业，令训练集D(A)为电商行业的样本数据集，简称“行业数据”，测试集D(B)为该广告主的样本数据集，简称“公司数据”。本实施例中，异构数据集特征可视化评估流程HeDFEP的步骤如下：

[0061] 步骤1，给定异构数据集D(A)和D(B)，给定特征集合F，需要选择的特征数量N。特征集合F包括两个特征{hourofday，dayofweek}分别表示几点和星期几，需要选择特征数量为N＝1。

[0062] 步骤2，分别计算特征hourofday和dayofweek的各项指标数据，包括类别值数量、发生率、规范化发生率、漂移比、综合发生率等，构成指标集M；绘制特征hourofday和dayofweek的特征质量图，构成图形集G。

[0063] 步骤3，根据指标集M和图形集G对特征集F中的特征进行评估得到的特征评估结论为：特征hourofday稳定性较差、相关度较好；特征dayofweek稳定性较好、相关度较好。

[0064] 步骤4，根据指标集M和图形集G，通过判断hourofday和dayofweek两个特征的特征质量图，发现应用这两个特证的预测模型的效果瓶颈主要是hourofday这个特征的稳定性不好，构成特征归因结论。

[0065] 步骤5，根据指标集M和图形集G，选择出整体表现较好的前N＝1个特征为dayofweek，因为其稳定度和相关度都较好，构成特征选择结果集。

[0066] 步骤6，根据指标集M和图形集G，针对表现不好的特征hourofday，通过对其进行聚类，例如按时段聚成以下几类：深夜时段MN{0-6}，早晨时段M{7-10}，中午时段N{11-14}，下午时段AN{15-18}，晚间时段E{19-23}；则新的特征类别值集合为{MN，M，N，AN，E}共5个值，构成特征改进建议。

[0067] 步骤7，综合特征评估结论、特征归因结论、特征选择结果集、特征改建议构成特征评估报告。

[0068] 综上所述，与现有技术相比，本发明有以下优点和有益效果：

[0069] 1、本发明提出的特征极坐标可视化方法，首次将特征类别值进行可视化，以二维图形的方式可视化包含相关度和综合发生率两个维度的特征类别指标，将特征的类别值(数值特征需要离散化成类别值)映射成极坐标系中的点，进而根据点的轴向坐标分布判断特征类别值稳定度或漂移比、特征类别值发生率变化方向，通过点的角度坐标判断特征发生率相对于均值的高低水平。

[0070] 2、本发明提出的特征极坐标可视化方法，首次将特征质量进行可视化，以二维图形的方式可视化包含相关度和稳定度两个维度的特征质量，将特征类别集合(或数值特征离散化形成的类别集合)的每个类别影射成极坐标系中的点集，并提出特征质量图四种特征模式判断方法，进而根据点集的整体形状判断特征相关度、特征稳定度和特征质量。

[0071] 3、本发明提出的特征极坐标可视化方法，首次将采用可视化方法对特征进行研究，包括基于特征质量图特征评估方法、基于特征质量图的特征归因方法、基于特征质量图特征选择方法、基于特征质量图特征改进方法。

[0072] 4、本发明提出的特征极坐标可视化方法和极坐标可视化特征评估流程，一方面可以增加对预测问题的直观理解、产生解释性强的特征评估报告，加深对建模问题的理解深度，帮助人工特征选择和特征改进工作，另一方面根据特征评估报告进行特征选择和特征改进，以使后续的监督式机器学习模型面对异构数据集时(也可适用于同构数据集)，仍可

以克服特征不稳定性带来的不利影响，进行更有效的学习。

[0073] 以上所述的实施例仅用于说明本发明的技术思想及特点，其目的在于使本领域内的技术人员能够了解本发明的内容并据以实施，不能仅以本实施例来限定本发明的专利范围，即凡依本发明所揭示的精神所作的同等变化或修饰，仍落在本发明的专利范围内。

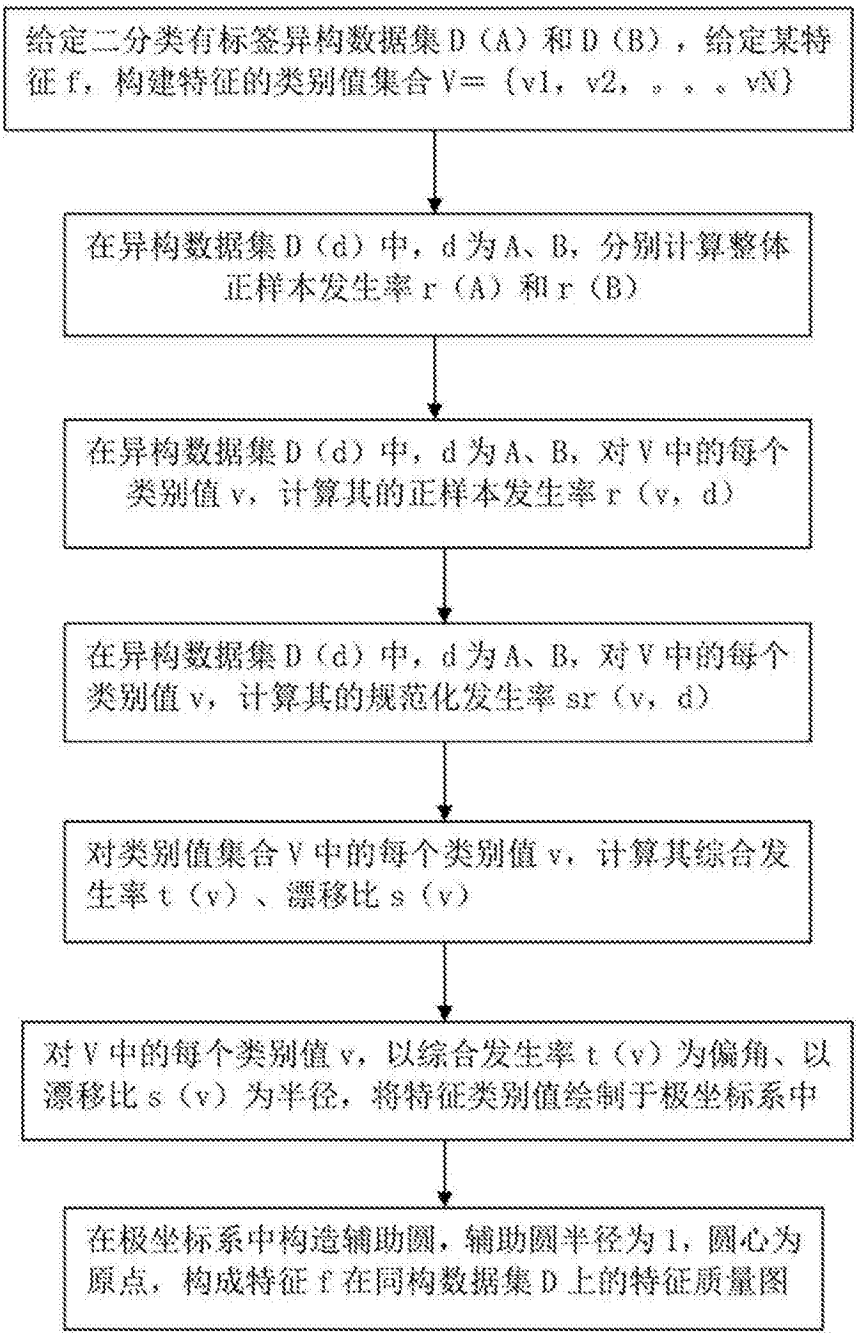

给定二分类有标签异构数据集D（A）和D（B），给定某特征f，构建特征的类别值集合V＝｛v1，v2，。。。vN｝
在异构数据集D（d）中，d为A、B，分别计算整体正样本发生率r（A）和r（B）
在异构数据集D（d）中，d为A、B，对V中的每个类别值v，计算其的正样本发生率r（v，d）
在异构数据集D（d）中，d为A、B，对V中的每个类别值v，计算其的规范化发生率sr（v，d）
对类别值集合V中的每个类别值v，计算其综合发生率t（v）、漂移比s（v）
对V中的每个类别值v，以综合发生率t（v）为偏角、以漂移比s（v）为半径，将特征类别值绘制于极坐标系中
在极坐标系中构造辅助圆，辅助圆半径为1，圆心为原点，构成特征f在同构数据集D上的特征质量图

图1

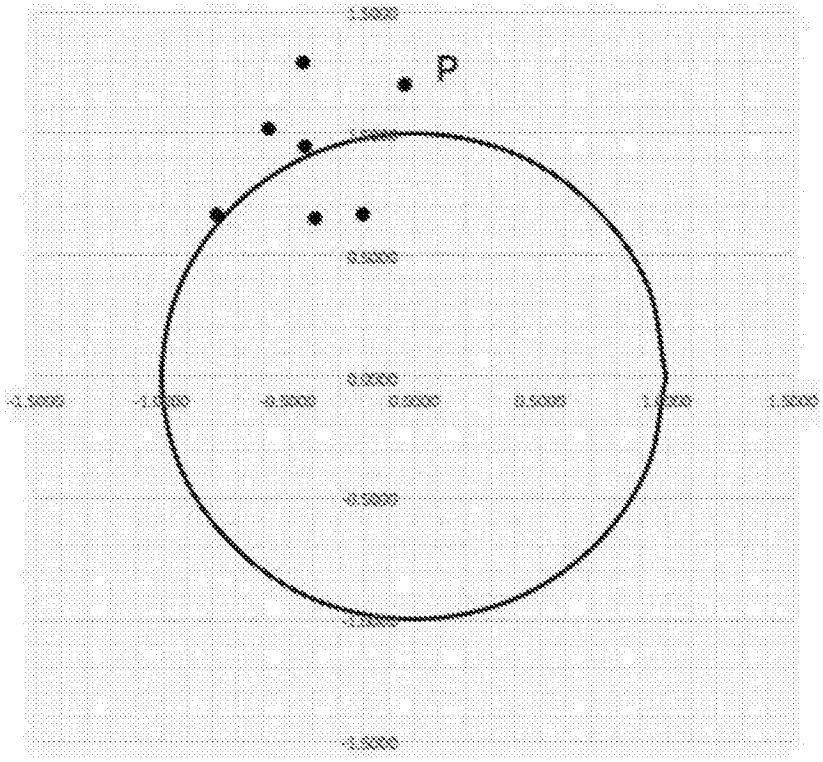


图2

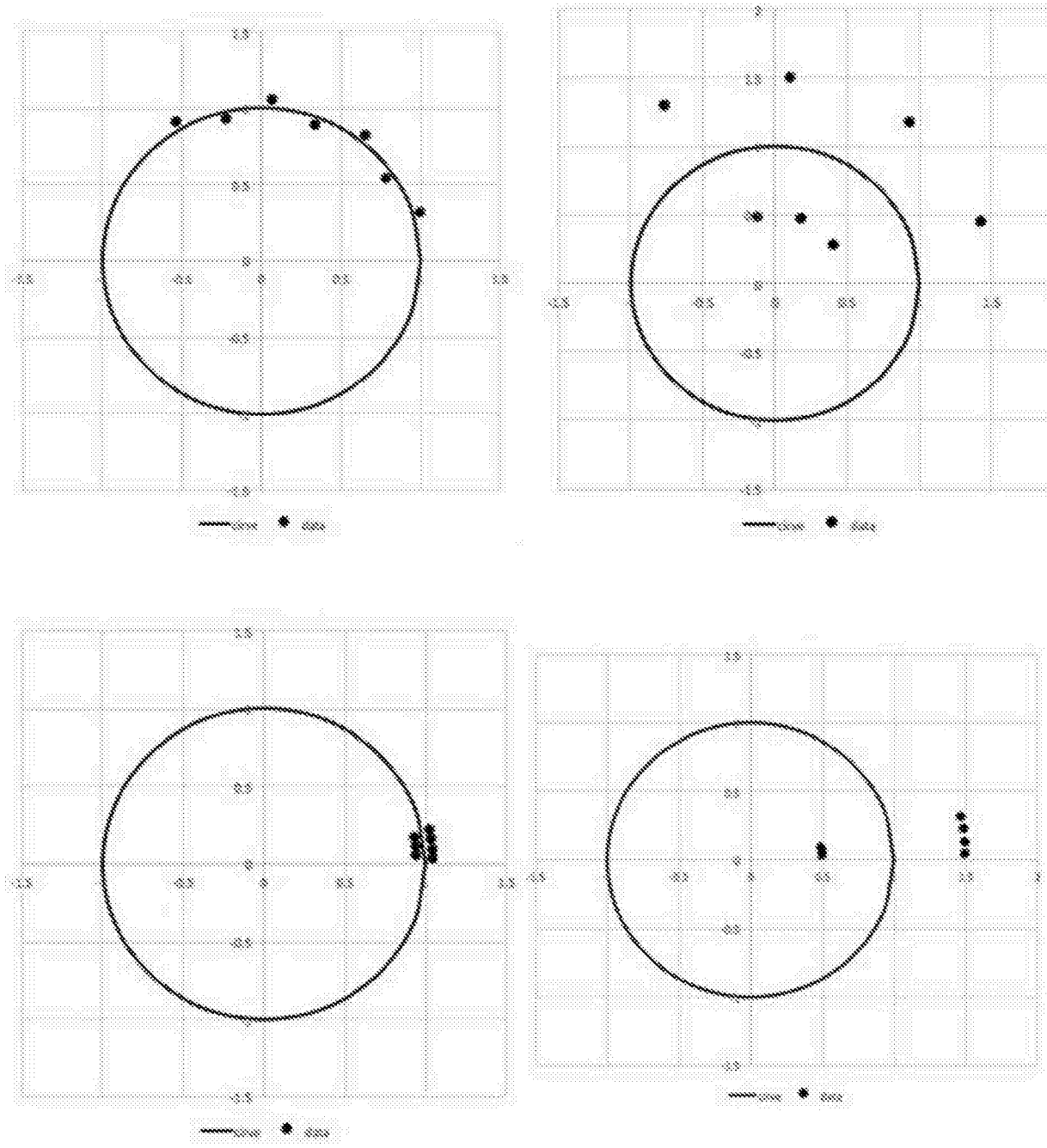


图3