

(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国际局

(43) 国际公布日
2025 年 1 月 9 日 (09.01.2025)



(10) 国际公布号
WO 2025/007412 A1

- (51) 国际专利分类号:
G06Q 10/04 (2023.01)
- (21) 国际申请号: PCT/CN2023/116931
- (22) 国际申请日: 2023 年 9 月 5 日 (05.09.2023)
- (25) 申请语言: 中文
- (26) 公布语言: 中文
- (30) 优先权:
202310807852.0 2023年7月4日 (04.07.2023) CN
- (71) 申请人: 江南大学 (JIANGNAN UNIVERSITY) [CN/
CN]; 中国江苏省无锡市蠡湖大道 1800
号, Jiangsu 214122 (CN)。
- (72) 发明人: 钱鹏江 (QIAN, Pengjiang); 中国江苏省无锡
市蠡湖大道 1800 号, Jiangsu 214122 (CN)。 王志

煌 (WANG, Zhihuang); 中国江苏省无锡市蠡湖大道 1800 号, Jiangsu 214122 (CN)。 王士同 (WANG, Shitong); 中国江苏省无锡市蠡湖大道 1800 号, Jiangsu 214122 (CN)。 蒋亦樟 (JIANG, Yizhang); 中国江苏省无锡市蠡湖大道 1800 号, Jiangsu 214122 (CN)。 方伟 (FANG, Wei); 中国江苏省无锡市蠡湖大道 1800 号, Jiangsu 214122 (CN)。 宋威 (SONG, Wei); 中国江苏省无锡市蠡湖大道 1800 号, Jiangsu 214122 (CN)。 姚健 (YAO, Jian); 中国江苏省无锡市蠡湖大道 1800 号, Jiangsu 214122 (CN)。 张欣 (ZHANG, Xin); 中国江苏省无锡市蠡湖大道 1800 号, Jiangsu 214122 (CN)。 陈爱国 (CHEN, Aiguo); 中国江苏省无锡市蠡湖大道 1800 号, Jiangsu 214122 (CN)。 顾逸 (GU, Yi); 中国江苏省无锡市蠡湖大道 1800 号, Jiangsu 214122 (CN)。

(54) **Title:** OBJECT PREDICTION MODEL CONSTRUCTION METHOD AND PREDICTION METHOD IN MULTI-CENTER SMALL-SAMPLE SCENARIO

(54) 发明名称: 多中心小样本场景下目标预测模型构建方法和预测方法

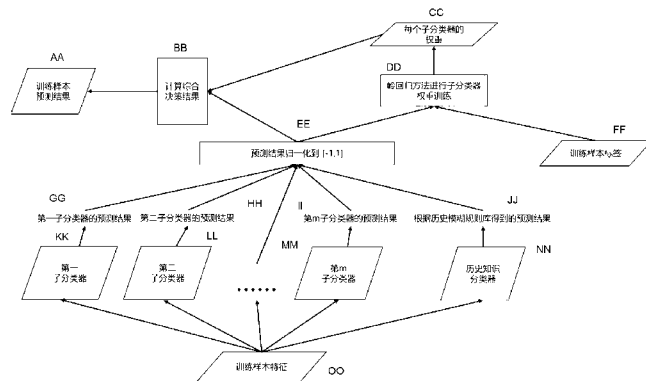


图 1

- AA Prediction results of the trained samples
BB Calculate an integrated decision-making result
CC Weight of each sub-classifier
DD Perform weight training on each sub-classifier by using a ridge regression method
EE Normalize the prediction results to [-1, 1]
FF Labels of the trained samples
GG A prediction result of the first sub-classifier
HH A prediction result of the second sub-classifier
II A prediction result of the mth sub-classifier
JJ Prediction results obtained according to a historical fuzzy rule library
KK First sub-classifier
LL Second sub-classifier
MM mth sub-classifier
NN Historical knowledge classifier
OO Features of trained samples

(57) **Abstract:** The present invention relates to an object prediction model construction method and prediction method in a multi-center small-sample scenario. On the basis of the idea of transfer learning, prediction is directly performed on a training set of a new node by using knowledge of trained nodes, and the difference between the new node and the trained nodes is reflected by means of an incorrectly predicted sample, which difference is used as supplementary knowledge, thus quickly acquiring model knowledge of the new node, and avoiding training the new node from the very beginning each time; and finally, parallel integration of incremental sub-classifiers

(74) 代理人: 苏州市中南伟业知识产权代理事务所(普通合伙)(**CENTRAL SOUTH WELL INTELLECTUAL PROPERTY OFFICE**); 中国江苏省苏州市吴中区石湖西路188号吴中万达广场A座苏州大学国家科技园2303, Jiangsu 215000 (CN)。

(81) 指定国(除另有指明, 要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CV, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IQ, IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, MG, MK, MN, MU, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW。

(84) 指定国(除另有指明, 要求每一种可提供的地区保护): ARIPO (BW, CV, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SC, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, ME, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

本国际公布:

— 包括国际检索报告(条约第21条(3))。

is realized by using a ridge regression method, thereby greatly reducing deployment time and costs. The generalization of a model is ensured by means of historical knowledge sharing and a knowledge discard mechanism, and therefore with the support of sufficient shared knowledge, even for a node with a small sample size, a relatively good classification effect can be achieved. An improvement is performed on the basis of a zero-order TSK fuzzy system having high interpretability, all of the model parameters can be embodied in the form of a rule, and high semantic transparency is achieved; thus, a user can be assisted in analyzing and understanding data in practical applications.

(57) 摘要: 本发明涉及一种多中心小样本场景下目标预测模型构建方法和预测方法。通过结合迁移学习的思想直接用已训练节点的知识对新节点的训练集进行预测, 借助预测错误样本体现新节点与已训练节点的差异, 将其作为补充知识, 以此快速获取新节点的模型知识, 避免每次都对新节点从头训练; 最后使用岭回归方法实现增量式子分类器的并行集成, 大大节省了部署时间和成本。通过共享历史知识和知识丢弃机制保证模型的泛化性, 在足够的共享知识的支持下, 对于小样本量的节点也可以取得较好的分类效果。基于具有高度可解释性的零阶TSK模糊系统进行改进, 所有的模型参数都可以通过规则的形式体现, 有着高度的语义透明性, 在实际应用中更能帮助用户分析和理解数据。

多中心小样本场景下目标预测模型构建方法和预测方法

技术领域

本发明涉及计算机技术领域，尤其是指一种多中心小样本场景下目标预测模型构建方法和预测方法。

背景技术

随着计算机算力的提升以及大数据时代的到来，机器学习成为数据分析的有力工具，对人类生活质量的提高做出了突出贡献。然而，由于数据的隐私性，不同机构、部门之间存在数据壁垒，导致海量数据难以实现共享，多中心数据的部署和联合训练成了难题。为此，谷歌提出联邦学习（federated learning）技术，通过在本地阶段对各自的数据进行存储和训练，仅与中心服务器交互模型更新，有效保障了用户的隐私安全。目前，虽然该领域已经存在诸多算法，但大部分算法基于深度学习网络提出，目前的深度学习方法参数多，模型大，数据泄露风险高，部署困难，虽然分类精度已经达到十分理想的程度，但是大多都是“黑盒模型”，缺乏语义可解释性，在高风险领域达不到安全性的要求，对于要求高度安全性和可解释性的机构并不适用。而传统机器学习算法，如邻近算法(K-NearestNeighbor,KNN)、支持向量机(Support Vector Machine,SVM)、朴素贝叶斯法(Naive Bayes, NB)、隐含狄利克雷分布(Latent Dirichlet Allocation, LDA)、分类与回归树(Classification and Regression tree, CART)等，对于多中心数据也容易因数据漂移和小样本问题而导致准确率不高，没有深度学习分类效果好，也无法保证多中心数据隐私问题。

发明内容

为解决上述技术问题，本发明提供了一种多中心小样本场景下目标预测模型构建方法和预测方法。

一种多中心小样本场景下目标预测模型构建方法，包括以下步骤：

S1：在具有 m 个数据中心的多个中心训练数据集中输入训练样本的特征，
5 使用多个子分类器以及历史知识分类器分别对训练样本进行预测，所述历史知识分类器为融合了所有子分类器规则的分类器；整合所有预测结果共有 $m+1$ 维，每个维度代表一个分类器的预测结果，大于 0 则判断为正类，小于 0 则为负类；

S2：将预测结果向量归一化到 $[-1,1]$ 之间；

10 S3：以归一化后的预测向量为预测特征，结合训练样本对应的训练标签，使用岭回归方法训练每个子分类器的权重；

S4：结合预测向量和训练得到的分类器权重得到综合预测结果；

S5：对综合预测结果进行分类，得到与所述训练标签相同的结果。

优选的，所述权重训练的目标函数以及权重参数的求解公式为：

$$15 \quad J = \frac{1}{2} \|PW - Y\|_2^2 + \frac{1}{2} \lambda \|W\|_2^2,$$

$$P_g = (\lambda I + P^T P)^{-1} P^T Y,$$

其中， P 为预测向量， W 为要训练的子分类器权重向量， Y 为对应训练样本的真值标签， λ 为结构风险项的正则化参数， P_g 为通过岭回归方法计算得到的后件参数， I 为单位矩阵。

20 优选的，第一子分类器的训练包括两个阶段：基础阶段和学习阶段。

优选的，所述基础阶段包括：借助经典 0 阶 TSK 来获取初始模糊规则参数 c 、 P_g 、 δ 以及错误分类样本信息 $errors$ ，使用高斯核函数作为隶属度函数；

其中, c 和 δ 为前件参数, c 是输入数据 Data 借助 FCM 聚类得到的聚类中心, δ 是根据输入数据 Data、FCM 聚类得到的 Data 对聚类中心的隶属度矩阵 U 以及 c 得到的隶属度函数核宽度。

优选的, 所述学习阶段包括: 先对 c 和 P_g 执行知识丢弃处理, 然后进行知识整理得到最佳知识组合, 最后结合输入数据 Data 计算出当前模糊规则对应的后件知识, 得到预测结果。

优选的, 所述知识丢弃处理包括: 首先, 确定知识丢弃比例 r_1 和知识混乱特征比例 r_2 , 根据 r_1 随机选择出要进行知识丢弃的规则, 根据 r_2 随机选择要进行知识混乱处理的特征列 $lose_columns$; 对于每一条要进行知识丢弃的规则, 随机确定要进行知识遗忘或知识混乱; 要进行知识遗忘的规则直接将 c 和 P_g 置空; 对于要进行知识混乱处理的规则, 将该规则的 $lose_columns$ 列以其他规则的相对应列值进行替换。

优选的, 所述知识整理包括: 将丢弃处理后的 c 和 P_g 以及错误分类样本信息 $errors$ 一起作为候选知识, 从中选出最佳知识组合作为当前阶段的预备知识, 所述错误分类样本信息 $errors$ 包括分别对应 c 和 P_g 的样本特征和样本标签。

优选的, 结合输入数据 Data 计算出当前模糊规则对应的后件知识, 得到预测结果具体包括: 结合输入数据 Data 和预备知识中的 c 计算出隶属度函数核宽度 δ ; 利用前件参数 c 和 δ 以及对应后件参数的历史迁移知识, 通过岭回归的方法, 将经验风险项、结构风险项以及历史后件参数迁移知识正则化项作为目标函数计算出当前模糊规则对应的后件知识, 目标函数表达式以及后件参数求解公式如下:

$$J = \frac{1}{2} \|X_g P_g - Y\|_2^2 + \frac{1}{2} \lambda_1 \|P_g\|_2^2 + \frac{1}{2} \lambda_2 \|P_g - P_{g_history}\|_2^2,$$

$$P_g = \left((\lambda_1 + \lambda_2) I + X_g^T X_g \right)^{-1} \left(X_g^T Y + \lambda_2 P_{g_history} \right),$$

其中, X_g 是高斯核映射后训练样本的输入向量, $P_{g-history}$ 历史后件参数迁移知识。

优选的, 每个子分类器训练之后都将产生的模糊规则更新加入到历史模糊规则库中, 以便后续子分类器的继续使用。

5 一种多中心小样本场景下目标预测方法, 步骤包括:

S1: 采用如上所述的多中心小样本场景下目标预测模型构建方法构建预测模型;

S2: 在所述预测模型中输入当前目标中与训练样本特征对应的特征;

S3: 所述预测模型根据所述特征得出该当前目标的预测结果。

10 本发明的上述技术方案相比现有技术具有以下优点:

1、本发明所述的多中心小样本场景下目标预测模型构建方法结合迁移学习的思想直接用已训练节点的知识来对新节点的训练集进行预测, 借助预测错误样本来体现新节点与已训练节点的差异, 将其作为补充知识, 以此快速获取新节点的模型知识, 避免每次都对新节点从头训练, 最后使用岭回归方法即可实现增量式子分类器的并行集成, 大大节省了部署时间和成本。

2、本发明所述的多中心小样本场景下目标预测模型构建方法通过共享历史知识和知识丢弃机制来保证模型的泛化性, 在足够的共享知识的支持下, 对于小样本量的节点也可以取得较好的分类效果。

3、所述的多中心小样本场景下目标预测模型构建方法基于具有高度可解释性的零阶 TSK 模糊系统进行改进, 所有的模型参数都可以通过规则的形式体现, 有着高度的语义透明性, 在实际应用中更能帮助用户分析和理解数据。

4、本发明所述的多中心小样本场景下目标预测模型构建方法很好地保证了不同节点间的样本不会相互泄露, 并且在实验中表现出了更好的分类性能。

附图说明

为了使本发明的内容更容易被清楚的理解，下面根据本发明的具体实施例并结合附图，对本发明作进一步详细的说明。

图 1 是本发明多中心小样本场景下目标预测模型构建方法整体模型架构。

图 2 是本发明多中心小样本场景下目标预测模型构建方法第一子分类器
5 训练细节。

图 3 是本发明多中心小样本场景下目标预测模型构建方法子分类器的协同训练过程。

图 4 是本发明多中心小样本场景下目标预测模型构建方法实验流程图。

具体实施方式

10 下面结合附图和具体实施例对本发明作进一步说明，以使本领域的技术人员可以更好地理解本发明并能予以实施，但所举实施例不作为对本发明的限定。

如图 1 所示，本发明提供一种多中心小样本场景下目标预测模型构建方法，包括以下步骤：

15 S1：在具有 m 个数据中心的 m 个中心训练数据集中输入训练样本的特征，使用多个子分类器以及历史知识分类器分别对训练样本进行预测，所述历史知识分类器为融合了所有子分类器规则的分类器；整合所有预测结果共有 $m+1$ 维，每个维度代表一个分类器的预测结果，大于 0 则判断为正类，小于 0 则为负类。

20 S2：将预测结果向量归一化到 $[-1,1]$ 之间，如此操作可保证量纲统一。

S3：以归一化后的预测向量为预测特征，结合训练样本对应的训练标签，使用岭回归方法训练每个子分类器的权重，权重训练的目标函数以及权重参数的求解公式为：

$$J = \frac{1}{2} \|PW - Y\|_2^2 + \frac{1}{2} \lambda \|W\|_2^2$$

$$P_g = (\lambda I + P^T P)^{-1} P^T Y,$$

其中, P 为预测向量, W 为要训练的子分类器权重向量, Y 为对应训练样本的真值标签, λ 为结构风险项的正则化参数, P_g 为通过岭回归方法计算得到
5 的后件参数, I 为单位矩阵。通过上述公式利用岭回归的方法可以迅速得到权重 W 的解。

S4: 结合预测向量和训练得到的分类器权重计算综合预测结果。

S5: 对于预测结果进行分类, 若大于 0 则为正类, 反之为负类。

在一个具体的实施例中, 第一子分类器作为后续子分类器快速训练的保
10 障, 第一子分类器应该选择数据量相对充足的数据中心进行训练。如图 2 所示, 第一子分类器的训练包括两个阶段: 基础阶段和学习阶段。

在基础阶段中, 借助经典 0 阶 TSK 模糊系统 (Takagi-Sugeno-Kang, TSK) 的一阶形式和零阶形式来获取初始模糊规则参数 c 、 P_g 、 δ 以及错误分类样本信息 $errors$, 使用高斯核函数作为隶属度函数。其中, c 和 δ 为前件参数, c
15 是输入数据 Data 借助模糊 C 均值算法 (fuzzy c-means, FCM 聚类) 得到的聚类中心, δ 是根据输入数据 Data、FCM 聚类得到的 Data 对聚类中心的隶属度矩阵 U 以及 c 得到的隶属度函数核宽度。

在学习阶段中, 包括以下步骤:

S1: 先对 c 和 P_g 执行知识丢弃处理,

20 S2: 然后进行知识整理得到最佳知识组合,

S3: 最后结合输入数据 Data 计算出当前模糊规则对应的后件知识, 得到预测结果。

在 S1 中, 知识丢弃机制模拟人脑的遗忘特性, 包括知识遗忘和知识混乱,

对应于知识完全遗忘和知识部分特征遗忘两种情况。首先，确定知识丢弃比例 r_1 和知识混乱特征比例 r_2 ，根据 r_1 随机选择出要进行知识丢弃的规则，根据 r_2 随机选择要进行知识混乱处理的特征列 *lose_columns*。对于每一条要进行知识丢弃的规则，随机确定要进行知识遗忘或知识混乱。要进行知识遗忘的规则直接将 c 和 P_g 置空即可；对于要进行知识混乱处理的规则，将该规则的 *lose_columns* 列以其他规则的相对应列值进行替换，达到知识错乱的效果。

在 S2 中，将丢弃处理后的 c 和 P_g 以及错误分类样本信息 *errors* 一起作为候选知识，从中选出最佳知识组合作为当前阶段的预备知识，错误分类样本信息 *errors* 包括分别对应 c 和 P_g 的样本特征和样本标签。知识整理涉及的距离计算是前件知识 c 以及 *errors* 的特征列构成的中心集中的中心的距离。知识的选择遵循两个原则：高可靠性和低相似性。

首先，为了衡量候选中心的可信度，需要计算每个候选中心邻域内的训练样本数，然后将值映射到[0,1]。邻域中的训练样本越多，候选中心的可信度就越高。邻域半径是所有候选中心之间的欧几里得距离之和的平均值。此外，通过计算候选中心与代表中心集中最近的中心之间的欧几里得距离来估计每个候选中心和代表中心集之间的相似度，并试图确保所选中心之间的相似度较低。同样，相似性值也需要映射到[0,1]。通过计算映射后的可靠性和相似性的指标之和来反映每个中心的重要程度，以此为依据选出最终的最佳中心组合，最后将这些中心以及这些中心相对应的后件知识作为选中的最佳知识组合。

在 S3 中，结合输入数据 *Data* 和预备知识中的 c ，即可计算出隶属度函数核宽度 δ 。此时，利用前件参数 c 和 δ 以及对应后件参数的历史迁移知识，通过岭回归的方法，将经验风险项、结构风险项以及历史后件参数迁移知识正则化项作为目标函数即可快速计算出当前模糊规则对应的后件知识，目标函数表达式以及后件参数求解公式如下：

$$J = \frac{1}{2} \|X_g P_g - Y\|_2^2 + \frac{1}{2} \lambda_1 \|P_g\|_2^2 + \frac{1}{2} \lambda_2 \|P_g - P_{g_history}\|_2^2$$

$$P_g = \left((\lambda_1 + \lambda_2) I + X_g^T X_g \right)^{-1} \left(X_g^T Y + \lambda_2 P_{g_history} \right)$$

其中， X_g 是高斯核映射后训练样本的输入向量， $P_{g_history}$ 历史后件参数迁移知识。由此，借助历史知识和错误分类样本知识实现了知识的进一步泛化学习，
5 不断得到更加稳健的新知识。

在一个具体的实施例中，多个子分类器的协同训练过程如图 3 所示，每个子分类器训练之后都将产生的模糊规则更新加入到历史模糊规则库中，以便后续子分类器的继续使用。对于第一个子分类器之外的其他子分类器，不再需要基础阶段的训练，而是直接使用当前历史模糊规则库的知识对本中心
10 数据样本进行预测，得到错误分类样本信息 errors 后，便可以像第一个子分类器一样进行学习阶段的训练，大大节省了训练时间，也更好地利用了历史知识，避免小样本场景下基础阶段无法得到较好知识的情况。

本发明还提供一种多中心小样本场景下目标预测方法，步骤包括：

S1：采用如上所述的多中心小样本场景下目标预测模型构建方法构建预
15 测模型；

S2：在所述预测模型中输入当前目标中与训练样本特征对应的特征；

S3：所述预测模型根据所述特征得出该当前目标的预测结果。

本发明的多中心小样本场景下目标预测模型构建方法和预测方法可以应用于交通情况预测、疾病辅助诊断等场景。

20 当交通堵塞时，如果没有提前派遣足够的交警前往疏通交通，导致交通事故和违章事件的发生，给人们的生命财产安全带来严重威胁。具体的，进行交通情况预测时，以不同城市作为数据中心，各数据中心的交通堵塞时间，堵塞地点等由于各种原因会有一定的不同，但又有相似的趋势。各城市分别

采集其多年内的交通数据作为训练特征共同采用本申请所述的多中心小样本场景下目标预测模型构建方法进行拥堵情况预测模型构建，实现不同的数据中心的交通数据在不对其他数据中心泄露的情况下实现抽象模糊规则的共享，各中心可将当前交通数据作为特征输入至该拥堵情况预测模型，得到当前拥堵情况的预测结果。该预测结果借鉴其他中心的堵塞特征趋势，从而更加合理地判断哪些时段更容易出现堵塞高峰，以便及时地派遣交警队员前往保障交通，保证人们正常通行和行车安全。

在医疗领域，医生通常无法提供大量的标注数据供人工智能模型训练，大部分情况下只能得到少量的有标注样本，而小样本训练得到的预测结果通常是比较差的。对于不同医院的医疗数据，虽然由于地区等原因数据分布会有所不同，但还是存在相似的趋势，对彼此模型的训练有着很大的指导意义。然而，为了患者的隐私考虑，医疗数据一般不能向外公开，因此虽然多个医院可以拥有大量的数据但彼此却不能通用。本申请所述的多中心小样本场景下目标预测模型构建方法可以提取出每个医院的抽象的模糊规则，在保证患者隐私的前提下，每个医院都可以在其他医院模糊规则的指导下开发出更加合理的预测模型，成为医生诊断的辅助手段，大大减少了医生的工作量，提高了患者的就医效率，使得不同医院的数据都得到了充分的利用。

本申请以 Kaggle 网站上的 personal-key-indicators-of-heart-disease 数据集 (<https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease>) 作为实验数据集。由于原数据集数据量过多，为了模拟多中心数据集的小样本节点场景，实验从中选取了部分样本进行试验。首先，根据数据集中的人种特征列，分别取白种人，黑种人以及亚洲人作为三个中心来制作数据集，选取的数据样本特征列信息完整，并将字符型特征列以离散数字代替以便后续计算，最终选中的每个样本都包括 16 个特征列和 1 个标签列。实验中，以 4:1 的比例划分训练集和测试集，并采用五折交叉验证确定参数，对数据特征进行归一化操作，将特征映射到 $[-1,1]$ 的范围内。下表记录了数据集

的具体划分情况：

	训练集(正类)	训练集(负类)	测试集(正类)	测试集(负类)
白种人	400 个	400 个	100 个	100 个
黑种人	16 个	16 个	4 个	4 个
亚洲人	16 个	16 个	4 个	4 个
总计	432 个	432 个	108 个	108 个

其中，正类代表患了心脏病的样本数据，负类代表未患心脏病的样本数据。

具体实验步骤如图 4 所示，包括：

- 5 S1：数据预处理。根据人种，分别从白种人，黑种人以及亚洲人三个中心中选择特征信息完整的样本用于制作数据集，并将字符型特征列以离散数字代替以便后续计算，最终选中的每个样本都包括 16 个特征列和 1 个标签列，其中，标签列用于计算各子分类器的权重或验证本申请多中心小样本场景下目标预测模型构建方法预测数据的准确性。实验中，为每个中心以 4：1 的比例随机划分训练集和测试集，并对数据特征进行归一化操作，将特征映射到
- 10 [-1,1]的范围内。

S2：制作多中心数据集。根据处理好的数据，把每个中心的训练集和测试集分别用.mat 格式保存，并额外将所有中心的训练集和所有中心的测试集也用.mat 格式保存起来，以便最后训练每个子分类器的权重。

- 15 S3：模型训练与参数范围调整。由于网络针对特定的数据集会有不同效果的偏差，所以要进行网格寻优，主要作用是特定网络针对特定的数据集取到最佳参数。以五折交叉验证的方式从参数寻优范围中寻找最优参数，通过不断更新当前最优参数直至迭代结束，得到最终的最优超参数。在这一过程中，通过多次实验观测模型结果，结合模型在训练集和测试集上的表现，判

断模型是否达到最佳拟合，逐步确定模型参数的最佳寻优范围。在上述数据集中，该算法在寻优时的相关参数具体如下：

参数	值	参数	值
数据量较大的数据中心的聚类中心数寻优范围(本实验指中心 1)	[5,7,9,11,13,15]	后件参数的结构风险项正则化系数寻优范围	[0.001,0.01,0.1,1,10,100,1000]
数据量较小的数据中心的聚类中心数寻优范围(本实验指中心 2 和中心 3)	[2, 3, 4, 5]	子分类器权重的结构风险项正则化系数寻优范围	[0.001,0.01,0.1,1,10,100,1000]
隶属度函数的核宽度调节系数寻优范围	[0.001,0.01,0.1,1,10,100,1000]	知识混乱特征比例	0.1
历史后件知识迁移项正则化系数寻优范围	[0.001,0.01,0.1,1,10,100,1000]	知识丢弃比例	0.1

S4：再次训练。使用网格寻优查找到的最优超参数来重新进行模型训练。

S5：模型测试。利用测试集对已经训练好的模型进行测试，观察测试结果是否达到预期。

S6：评价指标。通过上述步骤已经得到测试集的测试结果，利用公式计算多个评价指标，本次实验采用的评价指标为准确率，精确率，召回率，F1score 以及 AUC。

本实施例将本申请提供的多中心小样本场景下目标预测模型构建方法与
 10 现有经典的传统机器学习算法进行对比，对比算法包括：线性判别分析(Linear Discriminant Analysis, LDA)、k 近邻 (K-Nearest Neighbor, KNN)、朴素贝叶斯 (Naïve Bayes, NB)、分类回归树算法 (classification and regression tree,

CART)、支持向量机 (support vector machines, SVM) 以及 TSK 模糊系统 (Takagi-Sugeno-Kang Fuzzy-System, TSK-FS) 的一阶形式和零阶形式。下表记录了这些算法十次实验五项分类评价指标的均值和方差 (为了便于观察结果, 所有结果均保留 4 位小数):

		准确率	精确率	召回率	F1 分数	AUC
LDA	均值	0.6671	0.6672	0.6676	0.6674	0.7546
	方差	0.0114	0.0149	0.0029	0.0082	0.0014
KNN	均值	0.6607	0.6805	0.6056	0.6408	0.7206
	方差	0.0069	0.0058	0.0159	0.0105	0.0065
CART	均值	0.6806	0.6628	0.7361	0.6975	0.7265
	方差	0.0146	0.0160	0.0049	0.0111	0.0124
SVM	均值	0.6824	0.6841	0.6787	0.6813	0.7534
	方差	0.0088	0.0135	0.0045	0.0048	0.0058
NB	均值	0.6667	0.6765	0.6389	0.6571	0.7503
	方差	1.1703e^{-16}	1.1703e^{-16}	1.1703e^{-16}	1.1703e^{-16}	0
1-TSK-FS	均值	0.6704	0.6772	0.6565	0.6666	0.7471
	方差	0.0053	0.0050	0.0102	0.0064	0.0019
0-TSK-FS	均值	0.5602	0.3258	0.2685	0.2944	0.6961
	方差	0.0683	0.3435	0.2830	0.3103	0.0019
本申请	均值	0.6903	0.6912	0.6889	0.6897	0.7591
	方差	0.0091	0.0131	0.0248	0.0111	0.0055

使用 AUC 值作为模型的评价标准是因为很多时候 ROC 曲线并不能清晰的说明哪个分类器的效果更好，而作为一个数值，对应 AUC 更大的分类器效果更好。从表中可以看出，在准确率、精确率和 AUC 这三个指标上，本申请提供的多中心小样本场景下目标预测模型构建方法都比其他算法有着更好的表现，在召回率和 F1 分数上与其他传统算法也是可比较的，因此，从分类性能上看，本申请提供的多中心小样本场景下目标预测模型构建方法有着比其他传统算法更好的表现。

另外，考虑到算法的可解释性，实验中 1-TSK-FS、0-TSK-FS 以及本申请提供的多中心小样本场景下目标预测模型构建方法这三个基于模糊系统的规则复杂度比较如下表所示：

	前件特征个数	后件参数个数	平均规则数	规则复杂度
0-TSK-FS	16	1	6.5	214.5
1-TSK-FS	16	16	18.5	888
本申请	16	1	22.2	732.6

由上表可知，虽然 0-TSK-FS 比本申请提供的多中心小样本场景下目标预测模型构建方法有着更低的规则复杂度，但本申请提供的多中心小样本场景下目标预测模型构建方法的分类性能明显远远好于 0-TSK-FS。而 1-TSK-FS 从分类性能上和规则复杂度上都不如本申请提供的多中心小样本场景下目标预测模型构建方法优秀。综上所述，本申请提供的多中心小样本场景下目标预测模型构建方法的分类性能和可解释性比这些传统算法都更好。

在模型正式训练时，通过观察每次训练结果不断调整交叉验证的参数寻优范围，避免过复杂的参数模型导致过拟合的发生，提高模型泛化能力。在进行多次训练后，逐步确立一个稳定、性能良好的模型。在实际应用中，只要有一部分机构有着充足的样本数据，对于样本量较少的机构也可以在不侵犯其他机构数据隐私的情况下快速，有效地进行本地部署，获得具有较好分

类能力、泛化能力以及可解释性的本地模型投入使用，协助多机构更加高效地完成联合训练。

显然，上述实施例仅仅是为清楚地说明所作的举例，并非对实施方式的限定。对于所属领域的普通技术人员来说，在上述说明的基础上还可以做出
5 其它不同形式变化或变动。这里无需也无法对所有的实施方式予以穷举。而由此所引申出的显而易见的变化或变动仍处于本发明创造的保护范围之内。

权 利 要 求 书

1.一种多中心小样本场景下目标预测模型构建方法，其特征在于，包括以下步骤：

S1：在具有 m 个数据中心的 m 个中心训练数据集中输入训练样本的特征，
5 使用多个子分类器以及历史知识分类器分别对训练样本进行预测，所述历史知识分类器为融合了所有子分类器规则的分类器；整合所有预测结果共有 $m+1$ 维，每个维度代表一个分类器的预测结果，大于 0 则判断为正类，小于 0 则为负类；

S2：将预测结果向量归一化到 $[-1,1]$ 之间；

10 S3：以归一化后的预测向量为预测特征，结合训练样本对应的训练标签，使用岭回归方法训练每个子分类器的权重；

S4：结合预测向量和训练得到的分类器权重得到综合预测结果；

S5：对综合预测结果进行分类，得到与所述训练标签相同的结果；

其中，第一子分类器的训练包括基础阶段和学习阶段；其他子分类器的
15 训练基于第一子分类器的训练，包括学习阶段；所述基础阶段包括：借助经典 0 阶 TSK 来获取初始模糊规则参数 c 、 P_g 、 δ 以及错误分类样本信息 $errors$ ，使用高斯核函数作为隶属度函数；其中， c 和 δ 为前件参数， c 是输入数据 Data 借助 FCM 聚类得到的聚类中心， δ 是根据输入数据 Data、FCM 聚类得到的 Data 对聚类中心的隶属度矩阵 U 以及 c 得到的隶属度函数核宽度。

20 2.根据权利要求 1 所述的多中心小样本场景下目标预测模型构建方法，其特征在于，采用岭回归方法进行权重训练的目标函数以及权重参数的求解公式为：

$$J = \frac{1}{2} \|PW - Y\|_2^2 + \frac{1}{2} \lambda \|W\|_2^2,$$

$$P_g = (\lambda I + P^T P)^{-1} P^T Y,$$

其中, P 为预测向量, W 为要训练的子分类器权重向量, Y 为对应训练样本的真值标签, λ 为结构风险项的正则化参数, P_g 为通过岭回归方法计算得到的后件参数, I 为单位矩阵。

5 3.根据权利要求 1 所述的多中心小样本场景下目标预测模型构建方法, 其特征在于, 所述学习阶段包括: 先对 c 和 P_g 执行知识丢弃处理, 然后进行知识整理得到最佳知识组合, 最后结合输入数据 Data 计算出当前模糊规则对应的后件知识, 得到预测结果。

10 4.根据权利要求 3 所述的多中心小样本场景下目标预测模型构建方法, 其特征在于, 所述知识丢弃处理包括: 首先, 确定知识丢弃比例 r_1 和知识混乱特征比例 r_2 , 根据 r_1 随机选择出要进行知识丢弃的规则, 根据 r_2 随机选择要进行知识混乱处理的特征列 $lose_columns$; 对于每一条要进行知识丢弃的规则, 随机确定要进行知识遗忘或知识混乱; 要进行知识遗忘的规则直接将 c 和 P_g 置空; 对于要进行知识混乱处理的规则, 将该规则的 $lose_columns$ 列以
15 其他规则的相对应列值进行替换。

20 5.根据权利要求 4 所述的多中心小样本场景下目标预测模型构建方法, 其特征在于, 所述知识整理包括: 将丢弃处理后的 c 和 P_g 以及错误分类样本信息 $errors$ 一起作为候选知识, 从中选出最佳知识组合作为当前阶段的预备知识, 所述错误分类样本信息 $errors$ 包括分别对应 c 和 P_g 的样本特征和样本标
20 签。

25 6.根据权利要求 5 所述的多中心小样本场景下目标预测模型构建方法, 其特征在于, 结合输入数据 Data 计算出当前模糊规则对应的后件知识, 得到预测结果具体包括: 结合输入数据 Data 和预备知识中的 c 计算出隶属度函数核宽度 δ ; 利用前件参数 c 和 δ 以及对应后件参数的历史迁移知识, 通过岭回归的方法, 将经验风险项、结构风险项以及历史后件参数迁移知识正则化项作

为目标函数计算出当前模糊规则对应的后件知识，目标函数表达式以及后件参数求解公式如下：

$$J = \frac{1}{2} \|X_g P_g - Y\|_2^2 + \frac{1}{2} \lambda_1 \|P_g\|_2^2 + \frac{1}{2} \lambda_2 \|P_g - P_{g_history}\|_2^2,$$

$$P_g = \left((\lambda_1 + \lambda_2) I + X_g^T X_g \right)^{-1} \left(X_g^T Y + \lambda_2 P_{g_history} \right),$$

- 5 其中， X_g 是高斯核映射后训练样本的输入向量， $P_{g_history}$ 历史后件参数迁移知识。

7.根据权利要求 1 所述的多中心小样本场景下目标预测模型构建方法，其特征在于，每个子分类器训练之后都将产生的模糊规则更新加入到历史模糊规则库中，以便后续子分类器的继续使用。

- 10 8.一种多中心小样本场景下目标预测方法，其特征在于，步骤包括：

S1:采用如权利要求 1-7 任意一项所述的多中心小样本场景下目标预测模型构建方法构建预测模型；

S2：在所述预测模型中输入当前目标中与训练样本特征对应的特征；

S3：所述预测模型根据所述特征得出该当前目标的预测结果。

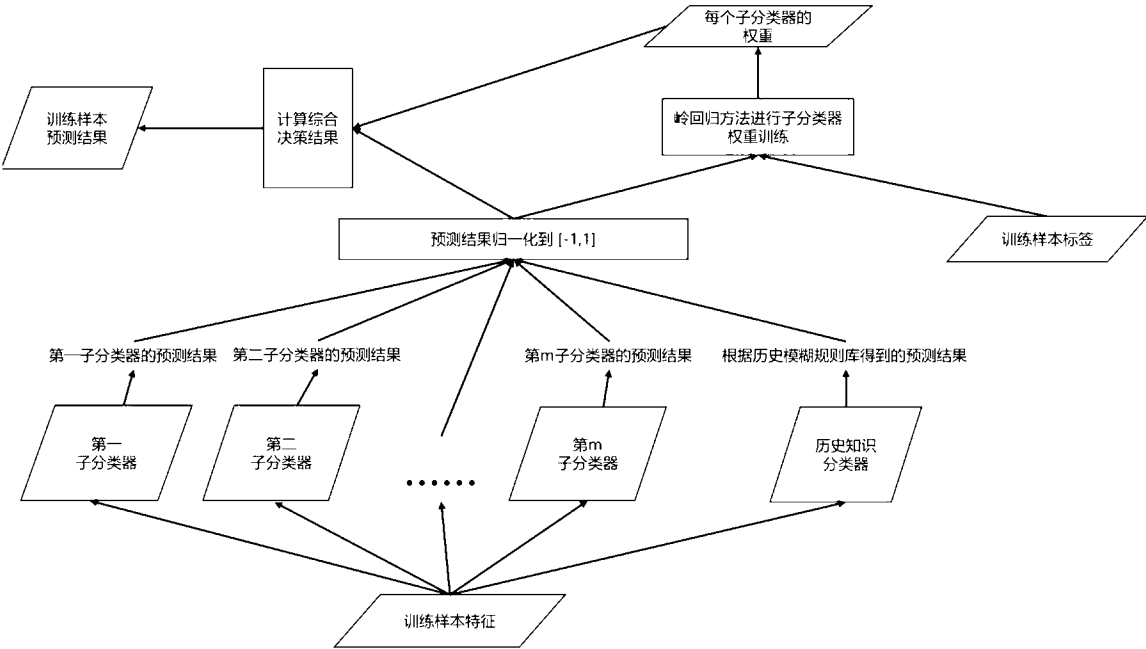


图 1

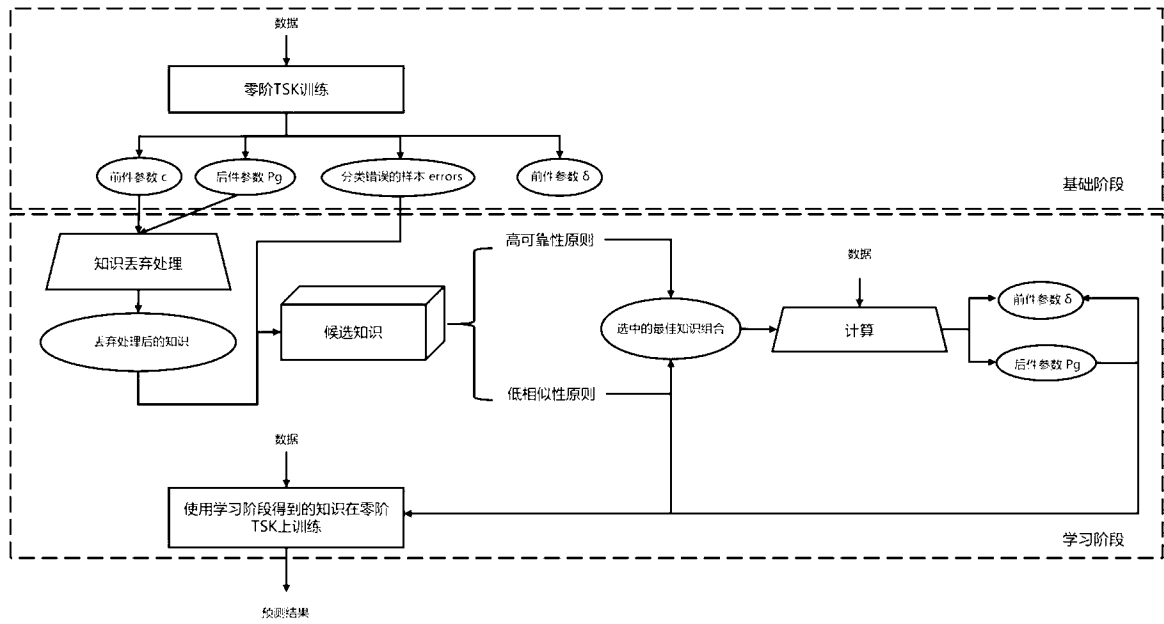


图 2

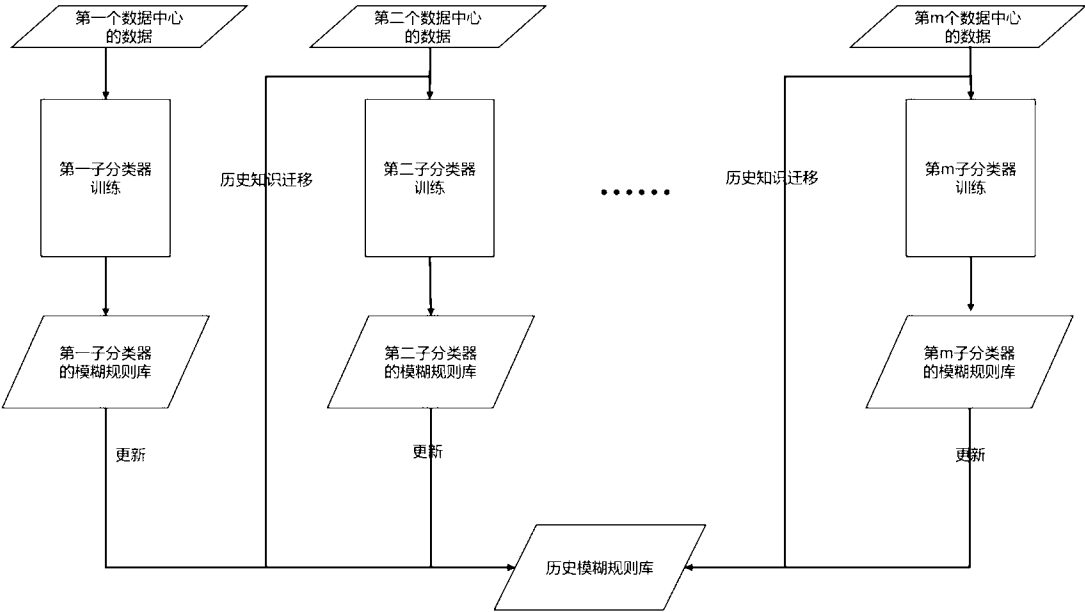


图 3

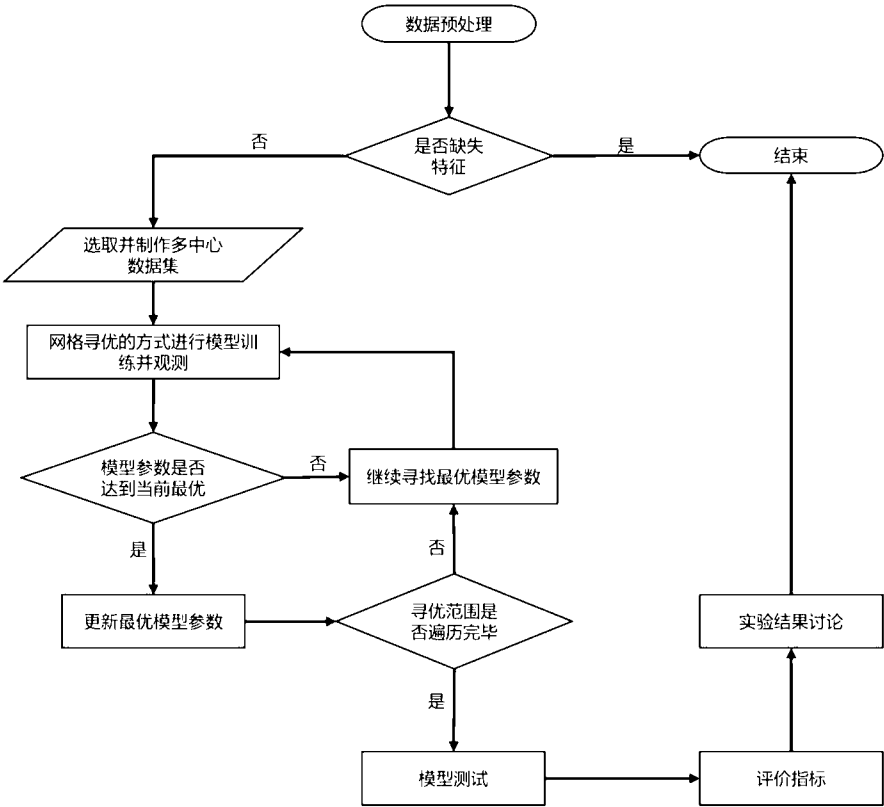


图 4

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2023/116931

A. CLASSIFICATION OF SUBJECT MATTER

G06Q 10/04(2023.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

IPC: G06N,G06F,G06Q,G06V

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

VEN, CNABS, CNTXT, WOTXT, EPTXT, USTXT, CNKI, IEEE: 小样本, 质心, 模型, 预测, 特征, 训练, 分类器, TSK, 隶属度, 模糊, 规则, 参数; sparse sample, centroid, models, predict+, features, train, classifier, TSK, membership, fuzzy, rule, parameter

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	CN 114298082 A (HUZHOU UNIVERSITY) 08 April 2022 (2022-04-08) entire document	1-8
PX	CN 116596161 A (JIANGNAN UNIVERSITY) 15 August 2023 (2023-08-15) claims 1-8	1-8
A	CN 110298382 A (HUZHOU UNIVERSITY) 01 October 2019 (2019-10-01) entire document	1-8
A	CN 115186731 A (HUZHOU UNIVERSITY) 14 October 2022 (2022-10-14) entire document	1-8
A	CN 115186798 A (HUZHOU UNIVERSITY) 14 October 2022 (2022-10-14) entire document	1-8
A	JP 2023011313 A (JVC KENWOOD CORP.) 24 January 2023 (2023-01-24) entire document	1-8



Further documents are listed in the continuation of Box C.



See patent family annex.

* Special categories of cited documents:

“A” document defining the general state of the art which is not considered to be of particular relevance

“D” document cited by the applicant in the international application

“E” earlier application or patent but published on or after the international filing date

“L” document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

“O” document referring to an oral disclosure, use, exhibition or other means

“P” document published prior to the international filing date but later than the priority date claimed

“T” later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

“X” document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

“Y” document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

“&” document member of the same patent family

Date of the actual completion of the international search

17 December 2023

Date of mailing of the international search report

22 December 2023

Name and mailing address of the ISA/CN

China National Intellectual Property Administration (ISA/
CN)
China No. 6, Xitucheng Road, Jimenqiao, Haidian District,
Beijing 100088

Authorized officer

Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2023/116931

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN	114298082	A	08 April 2022	None	
CN	116596161	A	15 August 2023	None	
CN	110298382	A	01 October 2019	None	
CN	115186731	A	14 October 2022	None	
CN	115186798	A	14 October 2022	None	
JP	2023011313	A	24 January 2023	None	

A. 主题的分类

G06Q 10/04(2023.01)i

按照国际专利分类(IPC)或者同时按照国家分类和IPC两种分类

B. 检索领域

检索的最低限度文献(标明分类系统和分类号)

IPC: G06N,G06F,G06Q,G06V

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库(数据库的名称, 和使用的检索词 (如使用))

VEN, CNABS, CNTXT, WOTXT, EPTXT, USTXT, CNKI, IEEE: 小样本, 质心, 模型, 预测, 特征, 训练, 分类器, TSK, 隶属度, 模糊, 规则, 参数; sparse sample, centroid, models, predict+, features, train, classifier, TSK, membership, fuzzy, rule, parameter

C. 相关文件

类型*	引用文件, 必要时, 指明相关段落	相关的权利要求
A	CN 114298082 A (湖州师范学院) 2022年4月8日 (2022 - 04 - 08) 全文	1-8
PX	CN 116596161 A (江南大学) 2023年8月15日 (2023 - 08 - 15) 权利要求1-8	1-8
A	CN 110298382 A (湖州师范学院) 2019年10月1日 (2019 - 10 - 01) 全文	1-8
A	CN 115186731 A (湖州师范学院) 2022年10月14日 (2022 - 10 - 14) 全文	1-8
A	CN 115186798 A (湖州师范学院) 2022年10月14日 (2022 - 10 - 14) 全文	1-8
A	JP 2023011313 A (JVC KENWOOD CORPORATION) 2023年1月24日 (2023 - 01 - 24) 全文	1-8

☐ 其余文件在C栏的续页中列出。

☒ 见同族专利附件。

★ 引用文件的具体类型:

“A” 认为不特别相关的表示了现有技术一般状态的文件

“D” 申请人在国际申请中引证的文件

“E” 在国际申请日的当天或之后公布的在先申请或专利

“L” 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件 (如具体说明的)

“O” 涉及口头公开、使用、展览或其他方式公开的文件

“p” 公布日先于国际申请日但迟于所要求的优先权日的文件

“T” 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件

“X” 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性

“Y” 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性

“&” 同族专利的文件

国际检索实际完成的日期

2023年12月17日

国际检索报告邮寄日期

2023年12月22日

ISA/CN的名称和邮寄地址

中国国家知识产权局
中国北京市海淀区蓟门桥西土城路6号 100088

授权官员

李福永

电话号码 (+86) 010-53961311

PCT/ISA/210 表(第2页) (2022年7月)

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2023/116931

检索报告引用的专利文件			公布日 (年/月/日)	同族专利	公布日 (年/月/日)
CN	114298082	A	2022年4月8日	无	
CN	116596161	A	2023年8月15日	无	
CN	110298382	A	2019年10月1日	无	
CN	115186731	A	2022年10月14日	无	
CN	115186798	A	2022年10月14日	无	
JP	2023011313	A	2023年1月24日	无	