基于 PCA-RANSAC 的混凝土抗压强度数据清洗设计与实现

姜万顺[1]，肖丙刚[1,*]，王奕鹏[1]，张亮亮[2]，赵华[2]，冯兰洲[2]，徐昌杰[2]，赖德发[2]

[1] 中国计量大学信息工程学院，浙江杭州 310018
[2] 中才邦业(杭州)智能技术有限公司，浙江杭州 310000

摘要： 混凝土抗压强度预测所使用的数据在采集的过程中容易产生异常值，这会对模型的准确性产生一定的影响，因此需要通过数据清洗来剔除异常数据。本文针对该问题，引入随机采样一致性算法，并结合主成分分析降维方法，提出了一种改进的异常数据清洗方法。首先验证了随机采样一致性算法的数据清洗模型的有效性，并与常用的孤立森林算法和 K-means 算法在公开数据集上比较性能，随机采样一致性算法模型的各性能指标均明显领先，进一步，利用主成分分析对数据降维，并优化采样点的选取规则，形成主成分分析-随机采样一致性算法数据清洗模型。实验表明，主成分分析-随机采样一致性算法性能指标与随机采样一致性算法相比无明显下降，但迭代次数减少至 492 次，耗时减少至 196ms，分别为传统随机采样一致性算法的 4.45%和 5.14%。所提出的改进方法缓解了随机采样一致性算法的易波动性，极大程度上减少了算法的迭代次数和耗时，清洗效果显著。

关键词： 混凝土抗压强度；数据清洗；随机采样一致性算法；主成分分析

DOI: 10.57237/j.cst.2023.04.004

Design and Implementation of Concrete Compressive Strength Data Cleaning Based on PCA-RANSAC Algorithm

Jiang Wan-shun[1], Xiao Bing-gang[1], Wang Yi-peng[1], Zhang Liang-liang[2], Zhao Hua[2], Feng Lan-zhou[2], Xu Chang-jie[2], Lai De-fa[2]

[1]School of Information Engineering, China Jiliang University, Hangzhou 310018, China
[2]SINOMA Bonyear (Hangzhou) Intelligent Technology Co. Ltd., Hangzhou 310000, China

Abstract: The data used for concrete compressive strength prediction is prone to produce outliers in the process of collection, which will have a certain impact on the accuracy of the model, so it is necessary to remove the abnormal data through data cleaning. In this paper, to address this problem, the random sampling consistency algorithm is introduced, and combined with the principal component analysis dimensionality reduction method, an improved anomalous data cleaning method is proposed. Firstly, the effectiveness of the data cleaning model of the random sampling consistency algorithm is verified, and the

基金项目: 2022 年度杭州市重大科技创新项目 (2022AIZD0085); 2022 年度杭州市重大科技创新项目 (2022AIZD0016).

*通信作者：肖丙刚, bgxiao@cjlu.edu.cn

收稿日期: 2023-09-08; 接受日期: 2023-10-26; 在线出版日期: 2023-10-28

performance is compared with the commonly used isolated forest algorithm and K-means algorithm on the public dataset, and the performance indexes of the random sampling consistency algorithm model are obviously leading, and furthermore, the data are downgraded by using the principal component analysis and the selection rules of the optimization of the sampling points, so as to form the principal component analysis-random sampling consistency algorithm data cleaning model. Experiments show that the performance index of principal component analysis-random sampling consistency algorithm has no significant decrease compared with random sampling consistency algorithm, but the number of iterations is reduced to 492, and the time consumed is reduced to 196ms, which is 4.45% and 5.14% of the traditional random sampling consistency algorithm, respectively. The proposed improved method alleviates the volatility of the random sampling consistency algorithm, greatly reduces the number of iterations and time-consuming of the algorithm, and has a significant cleaning effect.

Keywords: Concrete Compressive Strength; Data Cleaning; Random Sampling Consensus; Principal Component Analysis

1 前言

混凝土是建筑行业最重要的结构材料之一，其抗压强度（Concrete Compressive Strength, CCS）对建筑的荷载和安全性能起着决定性作用[1]。混凝土抗压强度预测有两大关键：数据与模型。而模型又依赖于数据，训练机器学习预测模型及构建传统经验公式预测模型的前提，都需要可靠的数据。

混凝土抗压强度数据集主要来源于物理测量，具体测量方法为：通过设计不同混凝土配比参数[2-4]，并在标准条件下等待一定时间后，进行物理测定。在此过程中，测量、计算、录入等过程均可能发生误差，导致数据集存在异常值。这些异常值不仅不能训练模型，反而会干扰模型准确性。因此，在训练模型前，需要进行异常数据清洗。

本文通过随机采样一致性算法（Random Sampling Consensus, RANSAC）结合主成分分析（Principal Component Analysis, PCA）降维方法，提出了PCA-RANSAC 算法来实现对 CSS 数据的分析与清洗。

2 混凝土抗压强度数据采集与分析

按照《混凝土强度检验评定标准》（GB / T 50107-2010），在标准条件下进行混凝土抗压强度数据采集[5]。收集约 1000 余条混凝土抗压强度数据样本，其参数名称、单位、级配最值、均值与标准差可由表 1 所见，表中最后一栏描述了各个参数在数据集中作为输入特征还是输出特征。

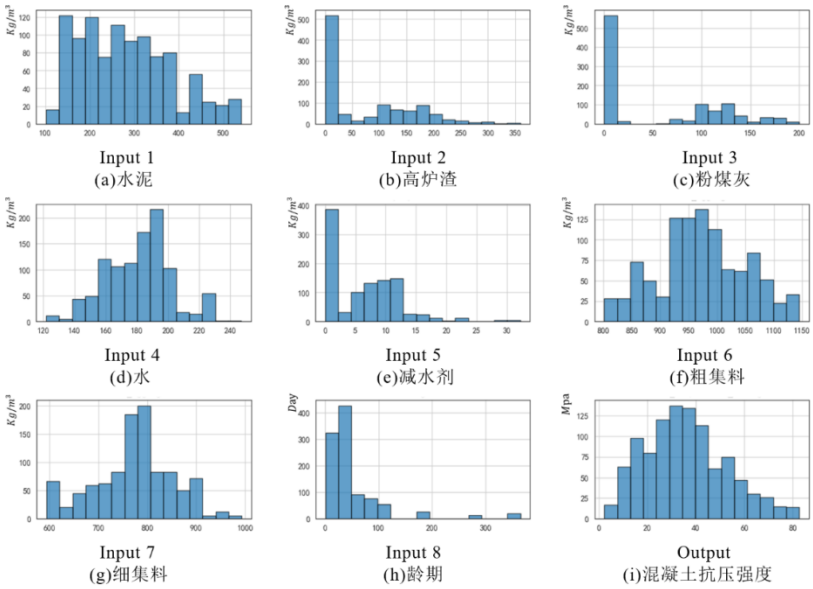


图 1 各参数统计分布

表 1 CCS 数据集参数统计表

参数名称	单位	*Max*	*Arg*	*Min*	*SD*	描述
x_1：水泥	Kg/m^3	540.0	281.2	102.0	104.5	Input 1
x_2：高炉渣	Kg/m^3	359.4	73.9	0.0	86.3	Input 2
x_3：粉煤灰	Kg/m^3	200.1	54.2	0.0	64.0	Input 3
x_4：水	Kg/m^3	247.0	181.6	121.8	21.4	Input 4
x_5：减水剂	Kg/m^3	32.2	6.2	0.0	6.0	Input 5
x_6：粗集料	Kg/m^3	1145.0	973.0	801.0	77.8	Input 6
x_7：细集料	Kg/m^3	992.6	773.6	594.0	80.2	Input 7
x_8：龄期	*Day*	365.0	45.7	1.0	63.2	Input 8
y：混凝土抗压强度	*Mpa*	82.6	35.8	2.3	16.7	Output

通过直方图分析如图 1，可以看出混凝土抗压强度数据集中各组分参数的数据分布情况，包括分布形态、峰值位置、数据集中程度等信息。以上数据特征说明数据集组分之间存在某种分布规律，通过构建预测模型，挖掘这些组分与混凝土抗压强度之间的规律，即可预测未知混凝土组分的抗压强度。

3 PCA-RANSAC 算法的 CCS 数据清洗模型

3.1 异常数据准则

异常数据准则指的是一组定量或定性的规则，用于确定数据集中哪些数据点是异常值。异常值的定义有较多种，较为公认的一种说法为 1984 年由 Barnett 提出的定义，具体为：指在数据集中，与其他数据观测的规律不一致的一个或多个样本点构成的集合。

异常数据对算法训练的影响可能包括以下几个方面：

①扰乱模型拟合：异常值可能会对模型的拟合产生显著的扰动，导致模型难以捕捉数据中的真实模式。

②降低模型的准确性：异常值可能会导致模型在测试集上的准确性下降。这是因为模型可能会过度依赖于异常值，并将其视为整个数据集的代表性样本，从而导致模型在测试集上的表现不佳。

③影响模型的稳定性：异常值可能会导致模型的稳定性下降。这是因为异常值的引入可能会导致模型参数估计不稳定，从而导致模型的表现在不同数据集上产生很大的波动。

本文所指的异常样本，为在抗压强度数据采集、录入过程中发生错误，可能干扰模型训练的数据样本，而由于数据本身的特性导致的“异常”，则需要尽可能保留。

3.2 RANSAC 算法设计

RANSAC 应用于异常数据清洗[6]，主要利用一组包含外点的数据集通过迭代的方法拟合模型参数，在阈值范围的样本数据评定为内点，在阈值以外的数据评定为外点。

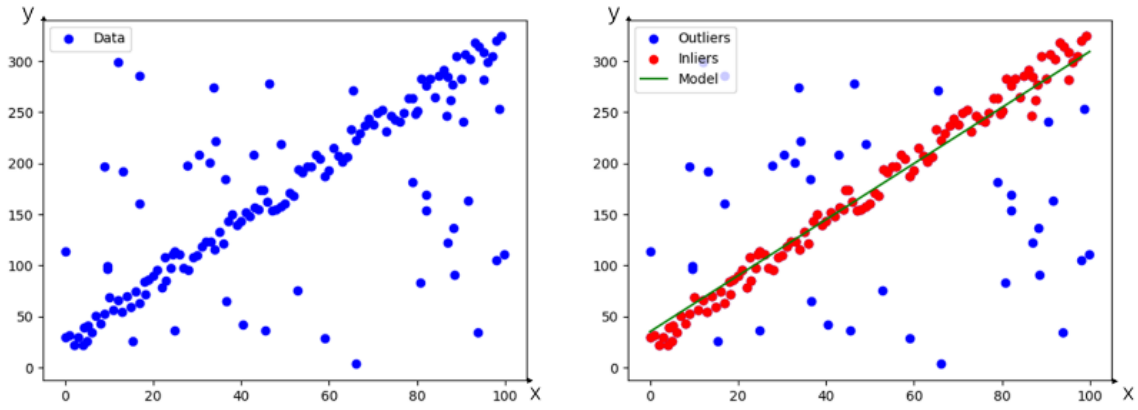
设样本数据集为 $D=\{p_1,p_2,...,p_m\}$，生成的模型为 *Model*，迭代次数为 K，模型匹配的最小样本数为 N，判断样本点是否处于模型公差带上的阈值为 T，最大内点数记为 $N_{\max}$，初始值为 0。据此进行迭代，其主要步骤如表 2 所示。

表 2 RANSAC 异常数据清洗算法

算法 1：RANSAC异常数据清洗算法	
1：	输入：样本集 $D=\{p_1,p_2,...,p_m\}$，模型 *Model*，模型参数 *param*，最小样本数据 N，最大迭代次数 K，阈值 T；
2：	步骤 1：第i次迭代，从样本集 $D=\{p_1,p_2,...,p_m\}$ 中随机选取 n 个采样点作为内点，构建内点集 $D'=\{p_1',p_2',...,p_n'\}$；
3：	步骤 2：利用内点集 D 构建模型 *Model*，并训练得到模型参数 *param*；
4：	步骤 3：计算样本集 D 除去内点集 D 的其他样本点与模型 *Model* 的误差 $\{d_1,d_2,...,d_{m-n}\}$，并比较误差与阈值 T 的关系，统计在阈值范围内的样本点总数，记为 N_i；
5：	步骤 4：判断 N_i 与 N 的关系，若 $N_i<N$，重复步骤 1-3，否则比较 N_i 与 $N_{\max}$ 关系，更新较大者为 $N_{\max}$；判断迭代次数 i 与 K 的关系，若达到 K 则停止。
6：	输出：内点与外点的划分结果

通过以上步骤，RANSAC 可较为准确的区分内点和外点，其示意图如图 2 所示（图 2 中随机设置了 150 个

数据样本，其中约 100 个内点，50 个外点）：



(a)原始数据 (b)RANSAC处理后

图 2 RANSAC 数据清洗示意图

同时，利用 RANSAC 算法剔除异常数据，其关键在于随机取样与迭代，与传统算法不同，RANSAC 使用较少的数据点估计基础模型参数来生成解决方案，因此对于数据集应该满足：正常数据多于异常数据和异常数据在数据集中随机分布的条件。而本文中混凝土抗压强度预测数据集基本满足以上要求，能很好适用该算法。

3.3 PCA-RANSAC 算法设计

（1）PCA 降维

主成分分析（Principal Component Analysis，PCA）核心思想是将高维数据映射到低维空间中，以保留数据的主要特征[7]。这个过程可以通过对数据的协方差矩阵进行特征值分解来实现。

设原始数据集维度为 n，数据个数为 m，用矩阵 $\mathbf{X}$ 表示 $n\times p$，其中第 i 项样本的特征向量可表示为公式（1）：

$$x^{(i)}=(x_1^{(i)},x_2^{(i)},...,x_n^{(i)})^T \tag{1}$$

T 代表转置。首先对数据进行标准化处理，将每个特征向量减去其均值，然后除以标准差，使得不同特征的数据具有相同的尺度。然后根据公式（2），计算数据的协方差矩阵 $\mathbf{C}$：

$$\mathbf{C}=\frac{1}{m-1}(\mathbf{X}-\bar{\mathbf{X}})^T(\mathbf{X}-\bar{\mathbf{X}}) \tag{2}$$

其中 $\bar{\mathbf{X}}$ 为 $\mathbf{X}$ 的均值向量。将 $\mathbf{C}$ 进行特征值分解，得到特征值 $\lambda_1,\lambda_2,...,\lambda_n$ 与对于的特征向量 $v_1,v_2,...,v_n$。选择前 k 项特征向量作为新的特征空间基向量，将原始数据点映射到新的特征空间中，得到降维后的数据 $\mathbf{Z}$：

$$\mathbf{Z}=\mathbf{X}\mathbf{V}_k \tag{3}$$

其中，$\mathbf{V}_k$ 由前 k 项特征向量组成，即 $\mathbf{V_k}=[v_1,v_2,..v_k]$，$\mathbf{Z}$ 的维度为 $m\times k$。

（2）模型设计步骤

由于本文数据集为多维数据，先采用 PCA 算法进行数据降维，在降维后的数据中，利用三维空间半随机采样的规则寻找 4 个采样点，并返回结果至 RANSAC 算法[8]。具体步骤如下：

①将降维的数据进行三维化，在样本点中随机选取 2 个点作为初始采样点 1 与采样点 2（对应图 3-a 中的红点）。

②找到初始采样点 1 与 2 的中间点（对应图 3-b 中的绿点），并将空间中距离该点最大的点标记为采样点 3（对应图 3-b 右下角红点），并判断三点是否共线，共线则返回步骤①，不共线则继续。

③通过采样点 1、2 与 3，构建空间三角形，并将距离该空间三角形内心（对应图 3-c 右下方绿点）最大的点记为采样点 4（对应图 3-c 上方红点）。

④记初始采样点为采样点 1-4（对应图 3-d 中红点）。

PCA-RANSAC 异常数据清洗算法由表 3 所示。

表 3 PCA-RANSAC 异常数据清洗算法

算法 2：PCA-RANSAC异常数据清洗算法	
1：	输入：样本集 $D=\{p_1,p_2,...,p_m\}$，模型 *Model*，模型参数 *param*，最小样本数据 N，最大迭代次数 K，阈值 T；
2： 3：	步骤 1：PCA 数据降维； 1.1 对样本集 D 进行标准化处理，使得不同特征的数据具有相同的尺度，并计算样本集 D 的协方差矩阵 **C**；
4：	1.2 对 **C** 进行特征值分解，得到特征值和特征向量并降序排序；
5：	1.3 选择前 3 项特征值对应的特征向量作为新的特征空间基向量；
6：	1.4 将原始数据点映射到新的特征空间中，得到降维后的数据 **Z**。
7：	步骤 2：改进的随机采样点选取规则；
8：	2.1 在降维后的云图中，随机选取 2 个采样点，记为 $p_1^{'},p_2^{'}$；
9：	2.2 计算 $p_1^{'},p_2^{'}$ 中点 p_c，记距离 p_c 最远的点为 $p_3^{'}$，判断 $p_1^{'},p_2^{'},p_3^{'}$ 三点是否共线，若是，返回 2.1 否则继续 2.3；
10：	2.3 计算 $p_1^{'},p_2^{'},p_3^{'}$ 三角形内心为 p_o，记距离 p_o 最远的点为 $p_4^{'}$；
11：	2.4 选取与 4 个采样点对应的原数据样 $D^{'}=\{p_1^{*},p_2^{*},p_3^{*},p_4^{*}\}$。
12：	步骤 3：利用内点集 D 构建模型 *Model*，并训练得到模型数 *param*；
13：	步骤 4：计算样本集 D 除去内点集 D 的其他样本点与模型 *Model* 的误差 $\{d_1,d_2,...,d_{m-4}\}$，并比较误差与阈值 T 的关系，统计在阈值范围内的样本点总数，记为 N_i；
14：	步骤 5：判断 N_i 与 N 的关系，若 $N_i<N$，重复步骤 2-4，否则比较 N_i 与 $N_{\max}$ 关系，更新较大者为 $N_{\max}$。判断迭代次数 i 与 K 的关系，若达到 K 则停止迭代。
15：	输出：内点与外点的划分结果

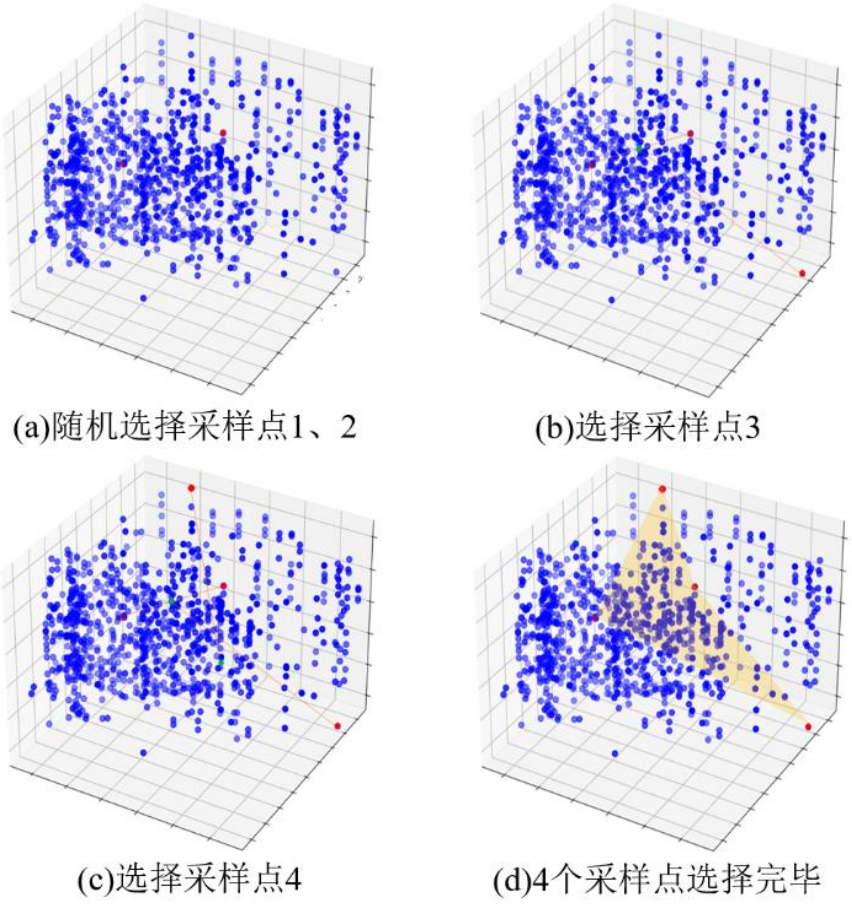


图 3 PCA-RANSAC 采样点选取步骤示意图

4 实验评估

4.1 实验设置

本文采用 RTX3060 为实验用显卡，CPU 采用 i7-12700H、内存为 16GB。使用 python3.10 解释器，使用 Scikit-Learn 作为后端框架，由于在 RANSAC 模型的变换矩阵有 8 个未知参数，而一组匹配点可得到 2 个线性方程，因此本文中选取 4 个样本点。为统计模型效果，最大迭代次数设为 inf。为避免偶然误差，本文抗压强度预测模型均运行 6 次，取平均/稳定结果作为最后结果。

4.2 混淆矩阵评价体系

为评价模型优劣，可使用预测准确率，即预测准确样本占比，但显然，对于本文的数据集，正常样本与异常样本数量差异较大，若单纯利用准确率对模型优劣进行评价，很容易训练出一个极度偏好正常数据的模型，这种评价方案有失偏颇。为此，基于本文数据集特点，利用基于混淆矩阵的评价法[9, 10]对指标进行评价，并根据四种类型指标组合，形成评价体系，下文与此类似的模型性能均将基于此进行评价。

混淆矩阵 4 种类型：

表 4 中 TP 代表真实值为正常，而通过 RANSAC 将其判断为正常（判断为内点），判断准确。FP、FN、TN 的含义与此类似。同时，

依据混淆矩阵，可定义：

表 4 混淆矩阵表

		真实值	
		正常	异常
预测值	正常	TP	FP
	异常	FN	TN

①准确率（Accuracy），即准确预测样本占比，准确率越高，模型对整体识别率越好，表示为公式（4）：

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

②精准度（Precision），即预测为正常的样本中，实际正常样本占比，主要表现模型的“纳伪”能力，表示为公式（5）：

$$P = \frac{TP}{TP + FP} \tag{5}$$

③召回率（Recall），即正常数据判定为正常的占比，主要表现模型的“弃真”能力，表示为公式（6）：

$$R = \frac{TP}{TP + FN} \tag{6}$$

④F1 分数，即权衡精准度与召回率之后的一种评价指标，表示为公式（7）：

$$F1 = 2\frac{P*R}{P + R} \tag{7}$$

4.3 对比算法

为了验证本文模型有效性，选择了与孤立森林算法（isolation Forest，iForest [11, 12]）、K-means 聚类两种数据清洗处理算法进行对比，并分别计算了在本数据集上的评价指标。

其中孤立森林算法通过利用一种特殊的二叉树（孤立树）来实现异常数据孤立，它将正常数据点尽可能地划分为同一叶节点，而异常值则很难划分到同一叶节点，通过判断叶节点数量情况，即可划分出异常数据点。

而 K-means 聚类[13, 14]是一种常见的基于聚类的异常值检测方法，其原理为通过设定聚类中心，根据数据集之间的逻辑联系对样本点划分类别，使得具有相似特征的样本点聚集为同一类，而离群点则识别为不属于任何群体的数据点，据此判定为异常点。这两种算法在数据清洗领域，均有大量应用，因此具有一定的代表意义。

4.4 公开数据集实验

由于混凝土抗压强度数据集无法知晓样本是否为异常样本，因此借用公开数据集对设计的模型进行评价。由于公开数据集 annthyroid_21feat_no-rm-alised（https://codeload.github.com/GuansongPang）包含与本文混凝土抗压强度数据集相似的数据类型、数据分布，因此选择此公开数据集作为评价数据集。数据集中包含 22 个指标，其中前 21 维为样本特征（Dim_0—Dim_20），最后 1 维为类别特征（0 或 1）。该数据集样本量为 7200，类别特征为 1 表正常数据共 6666 条，类别特征为 0 的代表异常数据共计 534 条。

各属性可由表 5 知悉。

表 5 数据集描述

描述	特征名称	描述	特征名称
“0”至“1”的连续变量	Dim_0、Dim_16、Dim_17、Dim_18、Dim_19、Dim_20	“0”或“1”的类别变量	Dim_1、Dim_2、Dim_3、Dim_4、Dim_5、Dim_6、Dim_7、Dim_8、Dim_9、Dim_10、Dim_11、Dim_12、Dim_13、Dim_14、Dim_15、Class

按照 RANSAC 算法流程[15, 16]，设置阈值为τ，随机选取 4 个样本点作为预设内点，设置迭代次数 K 初值为 0，并按照图 4 进行模型训练。

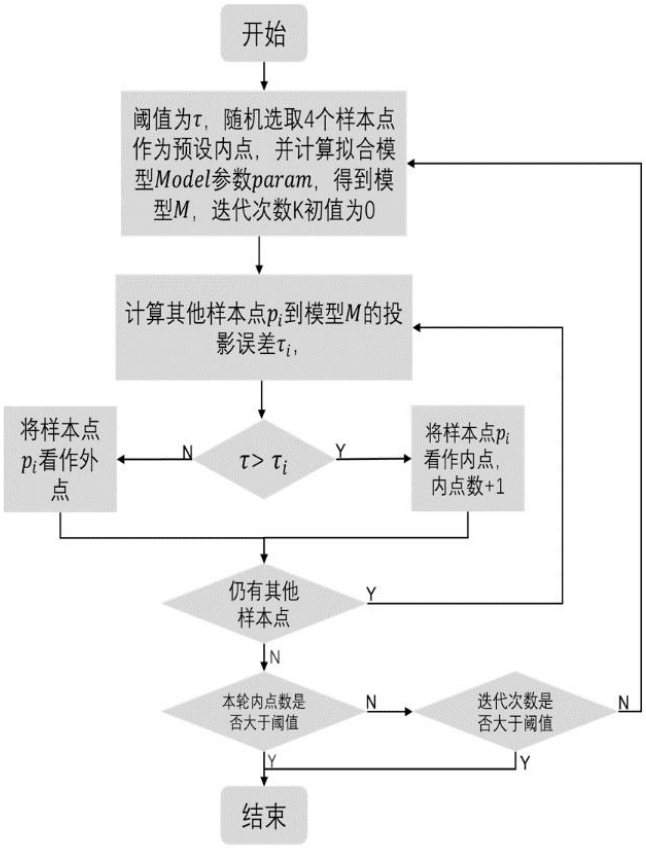


图 4 RANSAC 训练流程图

同样的，先利用 PCA 算法对数据进行降维处理，并设置阈值为 τ，按照图 5 虚线框所述对采样点进行选取，并按照图 5 对 PCA-RANSAC 算法进行训练。

通过在公开数据集 annthyroid_21-feat_norm-alised 上进行训练与测试，对 RANSAC 模型、PCA-RANSAC 模型及对比算法孤立森林算法、K-means 进行评价，值得注意的是，在改进的算法训练流程中，选取采样点 3 时，需要判断前 3 个采样点是否共线，若共线则无法形成空间三角形。

上述模型在数据集的表现如表 6（重复 6 次取均值），表中红色为最优值，蓝色为次优值。

表 6 PCA-RANSAC 模型数据清洗结果对比表

	准确率	精准度	召回率	F1
孤立森林	0.9039	0.9345	0.9287	0.9316
K-means	0.8662	0.9196	0.8914	0.9053
RANSAC	0.9358	0.9416	0.9813	0.9615
PCA-RANSAC	0.9347	0.9428	0.9686	0.9555

根据表 6 的数据，本研究所采用的 RANSAC 算法在异常数据识别准确率、精度、召回率及 F1 分数方面均优于孤立森林算法与 K-means 聚类方法。由于孤立森林算法采用树结构，在深度增加时正常样本与异常样本很难位于同一叶节点，因此其准确率相对较高。而 K-means 聚类方法易受数据集本身影响，若正常数据与异常数据差异度不高，则难以加以区分。相比之下，本研究所采用的 RANSAC 算法在召回率方面表现出卓越的性能，这说明该算法能够有效应对样本数据量不平衡对模型训练的影响。除此之外，在其他指标方面，RANSAC 算法也具有较好的性能，相较于其他模型表现更优秀。因此，针对数据清洗的问题，RANSAC 算法具有显著的优势。

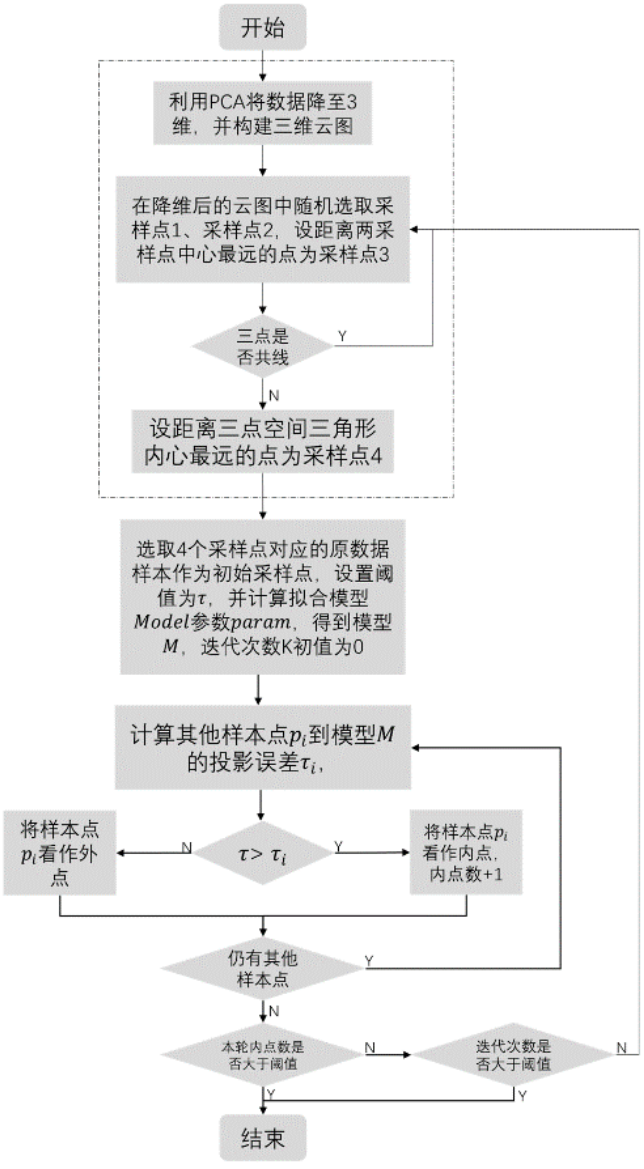


图 5 PCA-RANSAC 训练流程图

同时，PCA-RANSAC 算法对比 RANSAC 算法在准确度、召回率与 F1 分数上均有轻微下降，在精准度上有稍许提升，在统计学范畴可认为两者模型性能上保持一致。

4.5 混凝土抗压强度数据集实验

本文所采用的数据集在精密实验室测算得到，理论上异常样本数量较低，因此将内点阈值设置为 1000，

为了使得模型充分训练，得到最佳数据清洗模型，将最大迭代次数设置为 inf。同时，为尽可能保障去除的为真实异常点，重复进行异常数据清洗 6 次，并记在 6 次实验中，出现 2 次及以上的数据点为异常数据点。记录最终结果，参见表 7。

以水泥含量为横坐标，混凝土抗压强度为纵坐标绘制散点图，对上表结果进行可视化，其中红色点为该次模型判断出的异常点，蓝色点为正常点，结果如图 6 所示。

表 7 RANSAC 模型重复实验结果统计表

每次输出结果（按数据集内序号）	最终结果
114, 175, 177, 178, 179, 328, 332, 333, 357, 358, 362, 363, 368, 404, 463, 501, 514, 686, 699, 746, 763, 774	114, 175, 177, 178, 179, 182, 328, 332, 333, 353, 357, 358, 362, 363, 368, 381, 383, 404, 463, 464, 501, 514, 584, 686, 699, 746, 763, 774
114, 177, 178, 182, 328, 332, 333, 353, 357, 358, 362, 363, 368, 383, 404, 463, 464, 501, 514, 746, 763, 774	
114, 175, 177, 178, 179, 182, 332, 333, 353, 357, 358, 362, 368, 381, 383, 464, 501, 514, 584, 686, 699, 774	
114, 175, 177, 178, 179, 182, 328, 353, 358, 362, 363, 368, 381, 383, 501, 514, 584, 686, 746, 763	
114, 175, 177, 178, 179, 182, 328, 332, 333, 357, 358, 362, 363, 368, 381, 383, 404, 463, 464, 514, 584, 686, 699, 746, 763, 774	
114, 177, 178, 182, 328, 332, 362, 368, 383, 404, 463, 464, 501, 514, 584, 686, 699, 746, 763, 774	

由于 RANSAC 算法主要依靠随机采样与迭代，为验证模型效果，在以上结果的基础上，增加迭代次数与耗时统计，相关情况见表 8。

表 8 RANSAC 模型耗时与迭代次数统计表

总耗时（ms）	平均耗时（ms）	总迭代次数	平均迭代次数
1530	3809.33	4388	11067.83
2807		8350	
1459		4148	
5185		15287	
8415		23930	
3460		10304	

据上表，由于 RANSAC 对于最优模型十分依赖数据点的选择，若初始采样点选择不好，则容易陷入局部困境，使得模型迭代次数和总耗时极具增加。在上表中，第 4、5 次模型总迭代次数与耗时相对于平均迭代次数与耗时而言，均出现了陡然提升，这说明该次采样点选取效果并不理想。

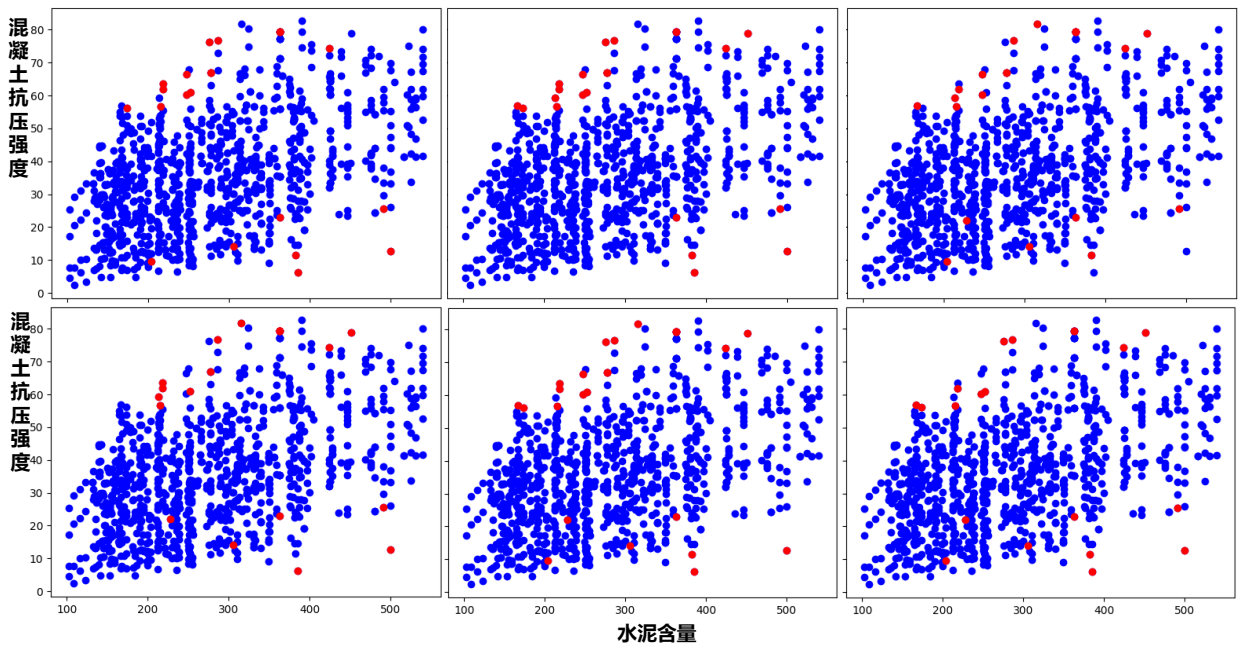


图 6 基于 RANSAC 异常数据清洗模型可视化结果

同样的，利用 PCA-RANSAC 算法在本文混凝土抗压强度数据集上进行训练，得到的数据清洗结果如图 7（以水泥含量为横坐标，混凝土抗压强度为纵坐标红色点为该次模型判断出的异常点，蓝色点为正常点）。

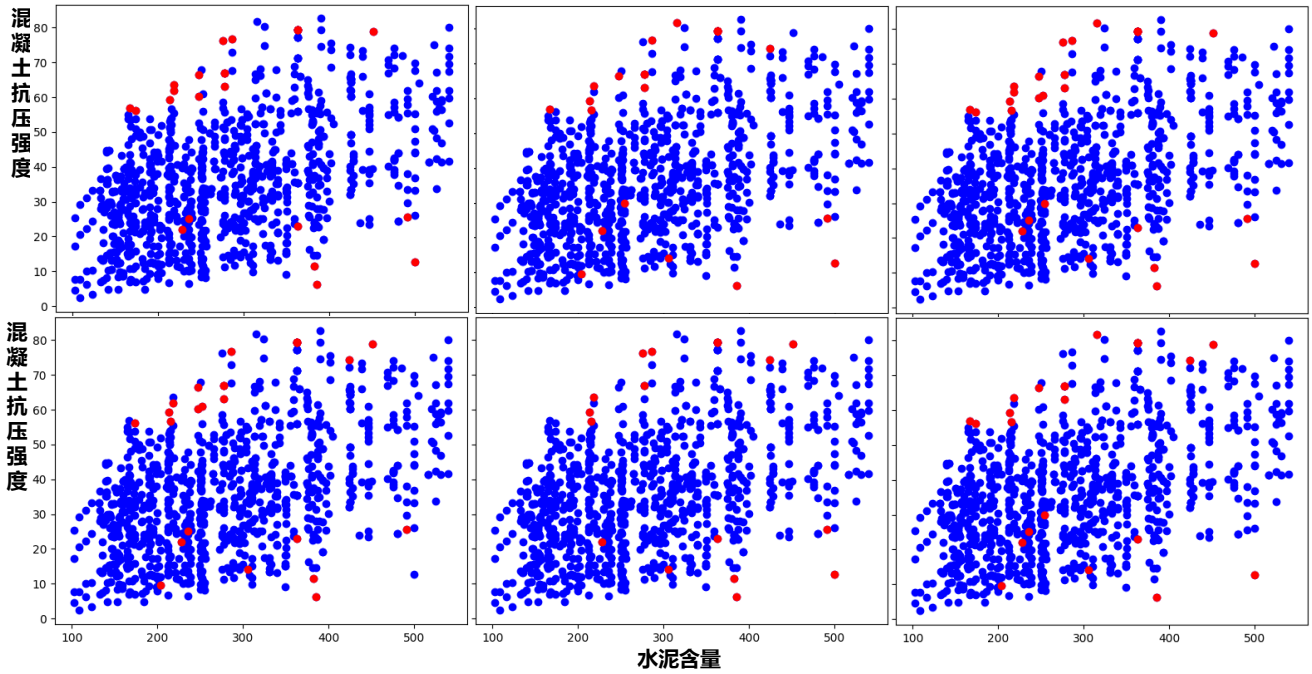


图 7 PCA-RANSAC 异常数据清洗模型可视化结果

统计出现 2 次及以上的数据点为：114, 175, 177, 178, 179, 182, 332, 333, 353, 356, 358, 362, 363, 368, 381, 383, 404, 463, 464, 501, 514, 584, 610, 620, 686, 699, 746, 763, 774。取改进算法结果与原结果交集为最终结果为：114, 175, 177, 178, 179, 182, 332, 333, 353, 358, 362, 363, 368, 381, 383, 404, 463, 464, 501, 514, 584, 686, 699, 746, 763, 774。

同时，统计 PCA-RANSAC 模型迭代次数与耗时于表 9。

表 9 PCA-RANSAC 模型耗时与迭代次数表

总耗时（ms）	平均耗时（ms）	总迭代次数	平均迭代次数
159	196.00	390	492.33
248		641	
166		394	
236		621	
226		560	
141		348	

根据表 6、表 8 与表 9 可知，相较于其他两种算法，PCA-RANSAC 算法性能均较好，相较于传统 RANSAC 算法性能基本相当。同时，由于优化了选点策略，使得在本文数据集上迭代次数与耗时均有明显下降，迭代次数减少至 492 次，耗时减少至 196ms，分别减少至传统 RANSAC 算法的 4.45%和 5.14%。并且相较于传统的 RANSAC 算法，其迭代次数波动不大，总体性能更加稳定。据此，可以判断 PCA-RANSAC 算法对于异常数据清洗有较高可信度的同时，兼具高效与稳定的优点。

5 总结

混凝土抗压强度对于建筑物的结构安全与耐久性至关重要，如何在混凝土早期快速而准确地进行抗压强度预测对实际工程具有重要意义。在收集混凝土抗压强度数据集的过程中会产生一定的异常数据，因此本文引入随机采样一致性算法的原理，并结合主成分分析，提出了一种改进的数据清洗方法。通过大量的

实验得出该方法不仅可以达到随机采样一致性算法在数据清洗上的优异性能，同时缓解了 RANSAC 算法的易波动性，极大降低了算法的迭代次数和耗时，为后续混凝土抗压强度预测模型的训练提供了有力的数据支持。

参考文献

[1] Coffetti D, Crotti E, Gazzaniga G, et al. Pathways towards sustainable concrete [J]. Cement and Concrete Research, 2022, 154: 106718.

[2] Sear L K A, Dews J, Kite B, et al. Abrams law, air and high water-to-cement ratios[J]. Construction and Building materials, 1996, 10(3): 221-226.

[3] De Larrard F. Concrete mixture proportioning: a scientific approach [M]. CRC Press, 1999.

[4] Wu X, Zhu F, Zhou M, et al. Intelligent Design of Construction Materials: A Comparative Study of AI Approaches for Predicting the Strength of Concrete with Blast Furnace Slag [J]. Materials, 2022, 15(13): 4582.

[5] Yeh I C. Modeling slump of concrete with fly ash and superplasticizer [J]. Computers and Concrete, An International Journal, 2008, 5(6): 559-572.

[6] Fischler M A, Bolles R C. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography [J]. Communications of the ACM, 1981, 24(6): 381-395.

[7] Maćkiewicz A, Ratajczak W. Principal components analysis (PCA) [J]. Computers & Geosciences, 1993, 19(3): 303-342.

[8] 苏宇, 刘海燕, 李国勇. 一种结合随机采样一致性与主成分分析的点云配准方法 [J]. 广西科技大学学报, 2022, 33(04):70-77. DOI: 10.16375/j.cnki.cn45-1395/t.2022.04.011.

[9] Luque A, Carrasco A, Martń A, et al. The impact of class imbalance in classification performance metrics based on the binary confusion matrix [J]. Pattern Recognition, 2019, 91: 216-231.

[10] 李港, 李莉, 林国义等. 硬盘故障预测模型的建立与实现 [J]. 控制工程, 2022, 29(10): 1788-1792. DOI: 10.14107/j.cnki.kzgc.CAC2020-1537.

[11] 徐昊, 王永生, 许志伟等. 基于生成对抗网络多变量风电时间序列异常值处理 [J]. 太阳能学报, 2022, 43(12): 300-311.

[12] Ding Z, Fei M. An Anomaly Detection Approach Based on Isolation Forest Algorithm for Streaming Data using Sliding Window [J]. IFAC Proceedings Volumes, 2013, 46(20).

[13] 赵耀, 虞莉娟, 苏义鑫等. 基于聚类分析和 Pearson 相关系数法的电网负荷数据清洗与去重 [J]. 船电技术, 2023, 43(06): 69-75. DOI: 10.13632/j.meee.2023.06.019.

[14] 张静, 陈燕林. 基于 K-means-CNN 耦合的采砂大数据智能清洗模型研究 [J]. 现代信息科技, 2023,7(18): 99-105. DOI: 10.19850/j.cnki.2096-4706.2023.18.020.

[15] Jongmoo Choi, Gérard G. Medioni. Starsac: Stable Random Sample Consensus For Parameter Estimation[C], Computer Vision and Pattern Recognition, 2009, 2009(1): 675-682.

[16] Wei Ruoyan, Wang Junfeng. FSASAC: Random Sample Consensus Based on Data Filter and Simulated Annealing [J], IEEE Access, 2021, 9: 164935-164948.