基于分类和回归树决策树的网络大数据集离群点动态检测算法

傅丽芳[1],陈 卓[2],敖长林[2]

（1.东北农业大学 理学院，哈尔滨 150030；2.东北农业大学 工程学院，哈尔滨 150030）

摘 要：针对大数据集中存在海量数据，当数据规模扩大到一定程度时，离散点检测处理效率受到限制的问题，提出了一种基于分类和回归树（CART）决策树的网络大数据集离群点动态检测算法。首先，划分大数据集异常数据标准，利用方差衡量数据离散程度，使用支持向量机建立异常数据样本关联规则矩阵，明确大数据集异常数据范围，并通过动态网格划分策略降低离群点检测计算量；然后，运用CART决策树方法在分支节点采取布尔检测，将待检测数据统一拟作连续数据，升序排列训练数据集，计算数据最高信息增益，剪枝决策树直到没有非叶子节点可被替换，得到离群点动态检测结果。仿真结果证明，本文算法离群点检测准确率高、检测耗时短，具备显著的计算优势，能为大数据集的可靠应用提供积极帮助。

关键词：分类和回归树决策树；大数据集；离群点检测；数据预处理；网格划分；基尼系数

中图分类号：TP393 **文献标志码**：A **文章编号**：1671-5497（2023）09-2620-06

DOI：10.13229/j.cnki.jdxbgxb.20220434

Dynamic outlier detection algorithm for network large data set based on classification and regression trees decision tree

FU Li-fang[1],CHEN Zhuo[2],AO Chang-lin[2]

（1. *College of Science*, *Northeast Agricultural University*, *Harbin* 150030, *China*; 2. *College of Engineering*, *Northeast Agricultural University*, *Harbin* 150030, *China*）

Abstract: There are massive data in big data sets, and when the data scale expands to a certain extent, the processing efficiency of discrete point detection is limited. Therefore, a dynamic outlier detection algorithm based on CART decision tree was proposed. Firstly, the abnormal data standard of large data set was divided, the data dispersion degree by variance was measured, the abnormal data sample association rule matrix by support vector machine was established, the abnormal data range of large data set was clarified, and the amount of outlier detection calculation by dynamic meshing strategy was reduced. Then, the classification and regression trees（CART）decision tree method was used to take Boolean detection at the branch nodes, unify the data to be detected as continuous data, arrange the training data set in ascending order, calculate the maximum information gain of the data, prune the decision tree until no non leaf nodes

收稿日期：2022-04-18.

基金项目：国家自然科学基金项目（71874026）.

作者简介：傅丽芳（1974-），女，副教授，博士.研究方向：农业网络舆情，大数据分析.E-mail：fulifang7895@163.com

通信作者：敖长林（1964-），男，教授，博士.研究方向：资源环境管理.E-mail：chenweiliang7895@163.com

can be replaced, and obtain the dynamic detection results of outliers. Simulation results show that the proposed algorithm has high outlier detection accuracy, short detection time, significant computational advantages, and can provide positive help for the reliable application of large data sets.

Key words: classification and regression trees (CART) decision tree; large data sets; outlier detection; data preprocessing; meshing; Gini coefficient

0　引　言

伴随信息技术的持续进步和爆发式增长,各类企业组织均累积了海量数据[1]。针对此类海量数据,如何提取出可用信息是当前信息领域的热门话题。目前,数据库系统即便能够完成数据的增加、删除、查找等操作[2],但仍难以实现对数据的深度分析。数据挖掘技术很好地弥补了上述需求,该技术融合了统计学、人工智能、机器学习等多种学科知识,可以在数据中进行宏观统计,获取数据之间的关联性和未来趋势。

通常意义而言,数据集内会具备离群点数据元素,它们可能是由于数据采集时的人为偏差或测量设备故障而生成的异常值。异常值的存在会降低数据采集质量,需要使用对应的离群点检测方法将其剔除。针对离群点检测问题,张倩倩等[3]引入孤立度板块分析样本点孤立特征,使用放大因子凸显数据点间的差异,运用增大算法提升离群点敏感性,得到离散点检测结果,但该方法没有考虑数据自身稀疏性特征,导致检测结果误差过高。江峰等[4]从入侵数据检测角度出发,把入侵行为拟作离群点,将粗糙熵描述为离群点特性,提出了粗糙熵下的离群点检测方法。该方法计算过程复杂,在时效性方面略有不足,无法在预期时段内完成检测目标。

综上,为改进以往离群点检测方法的不足,本文提出了一种基于分类和回归树(Classification and regression trees,CART)决策树的网络大数据集离群点动态检测算法。首先,深入分析数据集中异常数据的规律特征,通过网格划分预处理数据,节省了计算时间,运用CART决策树逐层归类节点属性;然后,利用剪枝操作替换非叶子节点,准确筛选离群点,保证了网络大数据集完整性和可用性;最后,进行仿真实验,从精准度、加速比和扩展性3个方面证明了本文方法的优越性,可为大数据集信息的高效甄别带来一定的参考和借鉴意义。

1　网络大数据集预处理

数据集预处理阶段,通过数据过滤方法大致筛选出具备异常特征的数据范围[5],利用动态网格划分完成数据集处理,表1为大数据集异常数据分类准则。

表1 大数据集异常数据分析标准

Table 1　Analysis standard of abnormal data of large data set

序号	类型	潜在表现	分析结果
1	数据失效	无法完成解析任务	无需过滤
		不在有效范围内	无需过滤
2	数据跳变	数据产生大幅度改变	无需过滤
		数据发生改变后随即恢复正常	无需过滤
		数据改变幅度极小	需要过滤
3	其他	数据状态稳定	需要过滤

数据过滤计算过程中,需要预先设定一个安全数据范围,在此范围内的数据直接过滤。将过滤周期设为1 min,在原有基础上,跳变分析单个数据,若数据跳变值超出预设临界值[6],则对该时段中的数据采取离散水平验证。离散水平越高,证明数据异常概率越大。在本文方法中,保存离散程度高的数据[7],直接过滤离散程度小的数据。

上述过程中,临界值的选取直接决定了过滤结果的有效性。方差是权衡随机变量或数据离散程度的标准,本文使用方差描述数据的离散程度。方差$A(X)$的计算公式为:

$$A(X)=\sum_{i=1}^{n}b_i\cdot(x_i-c)^2 \tag{1}$$

式中:n为数据总数;b_i为数据分量;x_i为第i个数据项;c为数据平均值。

为提升数据过滤可靠性,在式(1)前提下,引入支持向量机方法挖掘异常数据[8],聚类异常数据明确其所处范围。假设支持向量机中的最小距离节点为N_j,失真敏感参数为C_j^n,有效数据过滤后的数据集为D,由此得到大数据集无偏异常评估值$\mathrm{SN}(X)$为:

$$\exp \mathrm{SN}(X)=\sum_{T_d\supseteq X\wedge T_d\supseteq D}E(X,T_d) \tag{2}$$

式中:T_d为一次检测中提取数据相关性的采样时间间隔;$E(\cdot)$为数据异常函数。

利用支持向量机获得集合D内的异常数据样本,进而得到异常数据样本关联规则矩阵F,确定包含离群点的数据范围。

$$F=\begin{bmatrix}0 & y_1 & \cdots & y_n\\ y_1 & F_{11} & \cdots & F_{1n}\\ \vdots & \vdots & & \vdots\\ y_n & F_{n1} & \cdots & F_{nn}\end{bmatrix} \tag{3}$$

式中:y为关联因子;F为正定矩阵。

由于数据点仅能与其相邻的数据点存在关联,与较远的数据无耦合联系,故使用网格划分策略将大数据分割成多个子集,降低离群点检测计算量。针对数据流的实际特征,依据数据的持续性采取网格分裂、融合,储存网格内与分布有关的统计信息[9],保存网格内有可能是离群点的数据。

为处理数据流数据的概念转移问题[10],给予过往数据较小权重,减少其影响力,这样网格能更精准地展现目前数据的分布状态。假设数据权重的衰退指数为α,数据在t时段的权重为α^{t-t_o},其中,t_o为数据出现的时间,原始数据的权重为1,且数据权重会随着时间的增长而减小。将网格状态G表示为:

$$G=\left\langle H,\alpha^{t-t_o},d_H,p_H,t_{\text{latest}}\right\rangle \tag{4}$$

式中:H为k维空间内的超方体;d_H为落入H的数据点数;p_H为网格内的候选离群点集合;t_{latest}为此网格最近的更新时间。

如果网格密度达到临界值[11],对其采取分割操作,网格分裂融合就是让不同数据集被匹配至不同网格。正态分布下,网格划分解析式为:

$$L=\int_{\min j} q\frac{1}{\sqrt{2\pi\gamma_j}} \tag{5}$$

式中:L为网格分裂融合总数;γ_j为网格数据的标准偏差;$\min j$为网格内第j维上的最小值;q为网格融合指数。

2 CART决策树下网络大数据集离群点动态检测算法

为增强大数据集离群点检测精准度,提出了一种基于CART决策树的网络大数据集离群点动态检测算法。CART是决策树的一种表达形式,为一种二分递归分割算法[12]。该算法在分支节点实施布尔检测,倘若评估条件为真则归类为左分支,条件为假则归类为右分支,最终构成一棵二叉决策树。

利用CART决策树处理大数据集离群点时,因数据取值状况较多,统一将数据拟作连续数据[13],检测具体过程如下所示。

步骤1 升序排列待训练大数据集,离散化连续数据,O个样本涵盖O-1种离散策略,一般是将连续的两个样本平均值视为分割点[14],小于分割点的被归类至左节点,大于分割点的被归类至右节点。

步骤2 推算O-1种状态下的最高信息增益,与此同时,依照相对的分割点把O个样本归类至左节点和右节点。把父节点定义为R,凭借任意分割点归类左右子节点[15,16],将节点R的基尼系数$\text{Gini}(R)$表示为:

$$\text{Gini}(R)=1-\sum_{i=1}^{u}S_i^2 \tag{6}$$

式中:u为样本类型;S_i^2为R节点内样本类属第i类的概率。

若全部样本数据为相同类型,则基尼系数值等于0。

分割点把R节点归类后的基尼系数$\text{Gini}_z(R)$为[17]:

$$\text{Gini}_z(R)=q_l\text{Gini}_l+q_r\text{Gini}_r \tag{7}$$

式中:$q_l\text{Gini}_l$、$q_r\text{Gini}_r$分别为左、右节点内样本个数占父节点的比率。

由此,将信息增益ΔGini记作:

$$\Delta\text{Gini}=\text{Gini}(R)-\text{Gini}_z(R) \tag{8}$$

步骤3 如果树的深度已经达到预设的最高深度,全部的叶子节点基尼系数都为0[18,19],就停止划分子节点。

为防止产生过拟合现象,提升CART决策树的泛化能力,对决策树实施剪枝操作。剪枝指剔除非叶子节点T_i内的误差增益,使用子树的叶子节点替换非叶子节点T_i,反复执行该过程,直至没有非叶子节点可被替换[20],剪枝结束,输出离群点动态检测结果。将剪枝过程描述为:

$$\eta=\frac{Z(i)-Z(T_i)}{\left|L'(T_i)\right|-1} \tag{9}$$

$$Z(i)=rl(i)\,p'(i) \tag{10}$$

$$Z(T_i)=\sum_{j=1}^{n}rl_j(i)\,p'_j(i) \tag{11}$$

式中:$Z(i)$为叶子节点顶替第i个非叶子节点后生成的偏差;$rl(i)$为节点i的偏差概率;$p'(i)$为节点i内的样本数量占训练集合样本的百分比;$Z(T_i)$为节点i没有裁剪时子节点T_i内全部叶子节点的偏差总和;$L'(T_i)$为偏差绝对值。

3　仿真实验

为验证本文方法对大数据集离群点检测的正确性,挑选文献[3]近邻传播法、文献[4]粗糙熵法与本文方法进行实验对比。仿真平台为Matlab,模拟表2所示的4种数据量,分析不同数据量下方法的检测性能。其中,离群点检测精准度(Accuracy,AC)的计算公式为:

$$AC = p'/Q_N \tag{12}$$

式中:p'为离群点检测正确的个数;Q_N为离群点总数。

表2 实验样本数据信息

Table 2　Experimental sample data information

样本	数据量/万条	样本大小/MB
A	30	894
B	40	1056
C	50	1719
D	60	2493

3.1　精准度分析

在表2数据前提下,设定每个数据样本内均涵盖25个离群点,依次记录不同数据量下3种方法计算出的离群点个数,结果如图1所示。由图1可以看出,本文方法与其他两个方法相比,检测出的离群点数量与实际离群点数量最为相符,运算准确性得到显著提升,分析其原因是本文方法在检测前预先划分异常数据点范围,获得更多异常点信息特征,有效降低了检测误差。

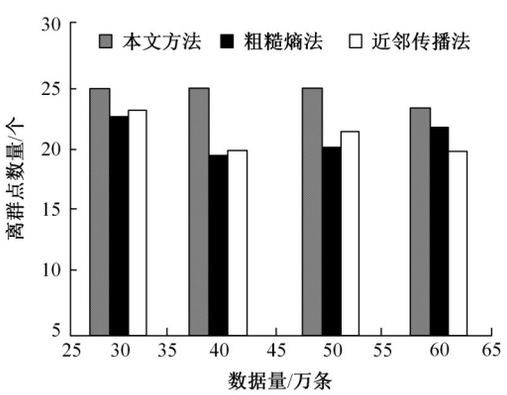

图1　离群点检测精准度对比

Fig. 1　Comparison of outlier detection accuracy

3.2　时效性分析

加速比和扩展性是权衡方法计算效率的核心指标。加速比表示处理相同工作时单节点与多节点运行时间的比率;扩展性表示方法可伴随节点个数和处理数据规模的增多所呈现出的计算性能。

在加速比实验中,设定节点个数为12,记录每次离群点检测的计算时长,分析其加速比大小,仿真结果如图2所示。由图2可以看出,随着节点数量的增多,加速比呈上升趋势,本文方法加速比始终高于其他两个方法,相同实验环境下加速比数值更大,证明本文方法加速比具备优秀的稳定性,检测效率也随之提高。

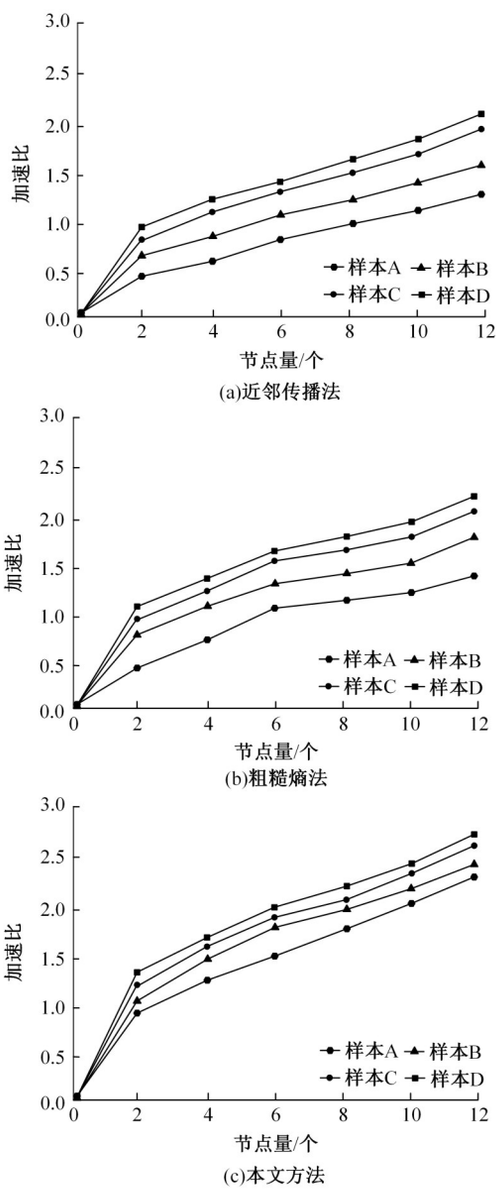

图2　三种方法加速比实验结果对比

Fig. 2　Comparison of experimental results of acceleration ratio of three methods

在扩展性实验中,把4种样本依次在2、4、6、8个节点上实施离群点检测,运行状况如图3所示。由图3可以看出:4个样本随着节点个数的增多,运行时间均呈下降趋势,且数据集规模越大,下降态势越显著。本文方法运行时间始终小于近邻传播法和粗糙熵法,证明数据规模和节点个数均增大时,本文方法依旧具备高效的数据处理能力,扩展性良好,有效减少了检测消耗时长,能够满足日常大数据离群点的处理需求。

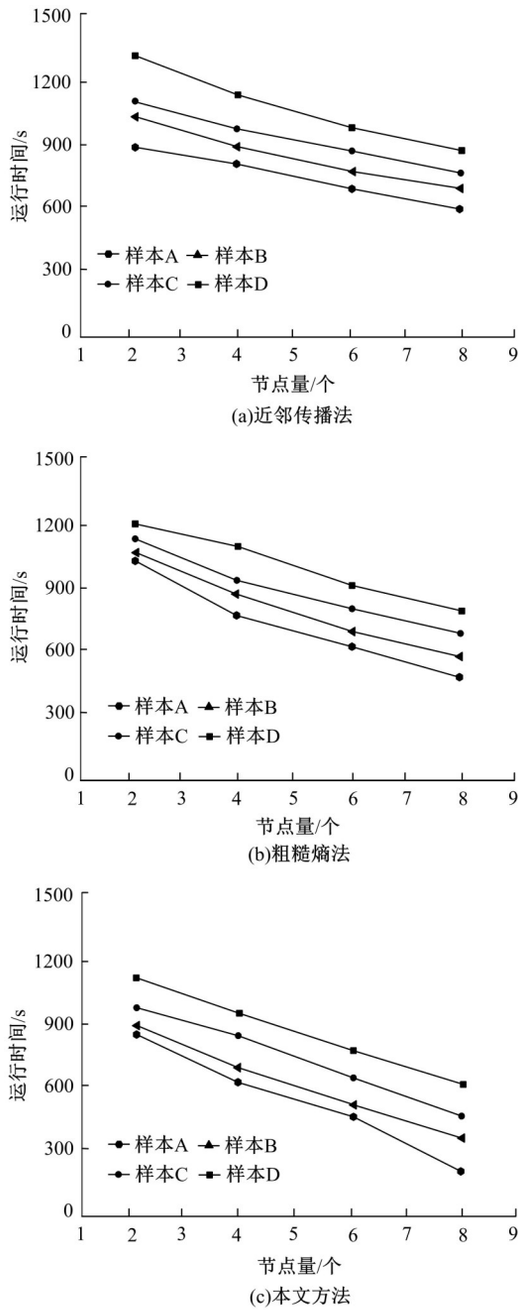

图3 三种方法扩展性实验结果对比

Fig. 3 Comparison of scalability experimental results of three methods

4 结束语

为改进传统数据离群点检测方法应用方面的缺陷,提出了一种基于CART决策树的网络大数据集离群点动态检测算法。首先,预处理大数据集中的异常数据,并采用数据空间网格划分构成若干无交集的网格单元,提升检测整体效率;然后,运用CART决策树得到高精度离群点检测输出结果,本文方法能够满足大数据时代离群点检测的需求。

参考文献:

[1] 杨晓玲, 冯山, 袁钟. 基于相对距离的反k近邻树离群点检测[J]. 电子学报, 2020, 48(5): 937-945.
Yang Xiao-ling, Feng shan, Yuan Zhong. Outlier detection based on reversed k-nearest neighborhood mst of relative distance measure[J]. Acta Electronica Sinica, 2020, 48(5): 937-945.

[2] Vafaei N, Ribeiro R A, Camarinha-Matos L M. Comparison of normalization techniques on data sets with outliers[J]. International Journal of Decision Support System Technology, 2022, 14(1): 1-17.

[3] 张倩倩, 于炯, 李梓杨, 等. 基于近邻传播的离群点检测算法[J]. 计算机应用研究, 2021, 38(6): 1662-1667.
Zhang Qian-qian, Yu Jiong, Li Zi-yang, et al. Outlier detection algorithm based on affinity propagation[J]. Application Research of Computers, 2021, 38(6): 1662-1667.

[4] 江峰, 王凯郦, 于旭, 等. 基于粗糙熵的离群点检测方法及其在无监督入侵检测中的应用[J]. 控制与决策, 2020, 35(5): 1199-1204.
Jiang Feng, Wang Kai-li, Yu Xu, et al. A rough entropy-based approach to outlier detection and its application in unsupervised intrusion detection[J]. Control and Decision, 2020, 35(5): 1199-1204.

[5] Belhadi A, Djenouri Y, Djenouri D, et al. Deep learning versus traditional solutions for group trajectory outliers[J]. IEEE Transactions on Cybernetics, 2020, 52(6): 4508-4519.

[6] 袁庆军, 王安, 王永娟, 等. 基于流形学习能量数据预处理的模板攻击优化方法[J]. 电子与信息学报, 2020, 42(8): 1853-1861.
Yuan Qing-jun, Wang An, Wang Yong-juan, et al. An improved template analysis method based on power traces preprocessing with manifold learning[J]. Journal of Electronics & Information Technology,

2020, 42(8): 1853-1861.

[7] Ghani M U, Rafi M, Tahir M A. Discriminative adaptive sets for multi-label classification[J]. IEEE Access, 2020, 8: 227579-227595.

[8] 邓泓, 刘志超, 彭莹琼, 等. 基于Fibonacci采样的数据预处理方法研究[J]. 江西师范大学学报: 自然科学版, 2021, 45(1): 60-66.
Deng Hong, Liu Zhi-chao, Peng Ying-qiong, et al. The study on data preprocessing method based on fibonacci sampling[J]. Journal of Jiangxi Normal University (Natural Sciences Edition), 2021, 45(1): 60-66.

[9] Sripriya T P, Srinivasan M R, Gallo M. Robust distance measure to detect outliers for categorical data[J]. Soft Computing, 2020, 24(18): 1-8.

[10] Li N, Zhao X W, Mu H L, et al. Research on the self-repairing model of outliers in energy data based on regional convergence[J]. Energies, 2020, 13(18): No. 4909.

[11] 刘云, 郑文凤, 张轶. 模糊残差算法对离群点数据的优化研究[J]. 小型微型计算机系统, 2021, 42(6): 1321-1326.
Liu Yun, Zheng Wen-feng, Zhang Yi. Optimization of outlier data by fuzzy residual algorithm[J]. Journal of Chinese Computer Systems, 2021, 42(6): 1321-1326.

[12] 王习特, 朱宗梅, 于雪苹, 等. 异构分布式环境中的并行离群点检测算法[J]. 湖南大学学报: 自然科学版, 2020, 47(10): 100-110.
Wang Xi-te, Zhu Zong-mei, Yu Xue-ping, et al. Parallel outlier detection algorithm in heterogeneous distributed environment[J]. Journal of Hunan University (Natural Sciences), 2020, 47(10): 100-110.

[13] Yang L, Lu Y Z, Yang S X, et al. An evolutionary game based secure clustering protocol with fuzzy trust evaluation and outlier detection for wireless sensor networks[J]. IEEE Sensors Journal, 2021, 21(12): 13935-13947.

[14] 水泽农, 张星宇, 沙朝锋. 基于最优输运和k-近邻的离群文档检测[J]. 计算机科学, 2021, 48(7): 105-111.
Shui Ze-nong, Zhang Xing-yu, Sha Chao-feng. Outlier document detection via optimal transport and k-nearest neighbor[J]. Computer Science, 2021, 48(7): 105-111.

[15] Yu K Q, Shi W, Santoro N. Designing a streaming algorithm for outlier detection in data mining—an incrementa approach[J]. Sensors, 2020, 20(5): No. 1261.

[16] Hagan R, Langston M A. Molecular subtyping and outlier detection in human disease using the paraclique algorithm[J]. Algorithms, 2021, 14(2): No. 63.

[17] 林雪. 海量不确定数据集中离群点快速检测方法仿真[J]. 计算机仿真, 2021, 38(6): 378-382.
Lin Xue. Simulation of quick detection method for outliers in massive uncertain data sets[J]. Computer Simulation, 2021, 38(6): 378-382.

[18] Mouret F, Albughdadi M, Duthoit S, et al. Outlier detection at the parcel-level in wheat and rapeseed crops using multispectral and sar time series[J]. Remote Sensing, 2021, 13(5): No. 956.

[19] 董泽, 贾昊. 基于EWT-LOF的热工过程数据异常值检测方法[J]. 仪器仪表学报, 2020, 41(2): 126-134.
Dong Ze, Jia Hao. Outlier detection method for thermal process data based on EWT-LOF[J]. Chinese Journal of Scientific Instrument, 2020, 41(2): 126-134.

[20] Riahi-Madvar M, Azirani A A, Nasersharif B, et al. A new density-based subspace selection method using mutual information for high dimensional outlier detection[J]. Knowledge-Based Systems, 2021, 216(2): No. 106733.