

近红外光谱的通用聚苯乙烯牌号在线识别方法

方 圆, 何张平, 朱世超, 梁显荣, 晋 刚*

华南理工大学聚合物新型成型装备国家工程研究中心, 广东省高分子先进制造技术及装备
重点实验室, 华南理工大学聚合物加工工程教育部重点实验室, 广东 广州 510640

摘 要 在聚合物加工过程中, 如果在同一生产线上混用不同牌号的原材料, 可能会影响产品性能, 降低产品合格率。然而采用传统方法识别相同类型不同牌号的聚合物往往耗时长且具有滞后性, 目前还缺乏一种快速实时的牌号识别方法。因此, 以 5 种不同牌号的通用聚苯乙烯(GPPS)为研究对象, 利用自主开发的安装于挤出机上的在线近红外光谱测量系统, 将近红外光谱与化学计量学、机器学习算法相结合, 实现对挤出过程中 GPPS 牌号的快速在线识别。首先利用在线近红外光谱测量系统实时采集 5 种不同牌号 GPPS 熔体的在线近红外光谱, 波长范围为 900~1 700 nm。经过谱图分析后, 利用主成分分析结合 K 均值聚类算法验证在线近红外光谱数据对于不同牌号的可行性。最后采用偏最小二乘判别分析和随机森林两种算法分别建立 GPPS 牌号识别模型并进行对比。结果表明: ①经过基线校正、最大最小归一化、7 点移动平均平滑预处理后, 在线近红外光谱在 1 207, 1 388, 1 407 和 1 429 nm 处的特征峰峰值会随着牌号的变化呈阶梯状改变, 以前 3 个主成分得分作为 K 均值聚类的输入变量得到聚类正确率为 88%, 说明了不同牌号 GPPS 在线近红外光谱数据的可行性; ②所建立的两种预测模型均能够对 GPPS 牌号有效识别, 最佳主因子数为 3 的偏最小二乘判别分析模型对验证集的分类正确率为 90.4%, 以前 5 个主成分得分作为输入变量建立的随机森林模型对验证集的分类正确率达 95.6%, 所以随机森林模型的牌号识别性能更好。因此, 在线近红外光谱测量系统结合化学计量学、机器学习算法可以实现 GPPS 牌号的快速在线识别, 为在生产线上利用近红外光谱识别同种聚合物的不同牌号提供参考。

关键词 近红外光谱; 牌号识别; 通用聚苯乙烯; 偏最小二乘判别分析; 随机森林

中图分类号: O657 **文献标识码:** A **DOI:** 10.3964/j.issn.1000-0593(2021)09-2759-05

引 言

化工企业通常根据用途不同会为同种塑料开发多种牌号, 不同牌号的塑料其性能和加工参数存在较大差异^[1]。在实际加工过程中, 同一生产线上混用不同牌号的原材料, 会影响产品性能, 降低产品合格率。目前常用的塑料牌号识别方法是测量材料的熔融指数、流变性能, 这些方法耗时长且具有滞后性。因此, 寻求一种快速、实时、准确的塑料牌号识别方法是有必要的。

近红外光谱技术(near-infrared spectroscopy, NIR)通过测量不同基团的吸收波长与强度来测量样品组成与含量^[2], 该方法测定速度快, 适用范围广, 操作简便, 在物质定性和定量分析中应用广泛^[3-5]。化学计量学的发展进一步促进了

近红外光谱技术的应用, 解决了近红外光谱谱峰重叠严重、指纹性差的问题。多篇文献报道了利用近红外光谱和化学计量学方法实现塑料分类的研究工作。Mikio 等^[6]以近红外光谱数据为基础, 利用决策树中的分类与回归树(classification and regression trees, CART)算法实现了 18 种塑料的准确分类识别。郝勇等^[7]结合近红外光谱和偏最小二乘判别分析(partial least squares-discriminant analysis, PLS-DA)实现了对 6 类食品接触塑料的精准识别。这些工作主要进行塑料种类的识别, 目前关于塑料牌号的识别还鲜有研究。在本课题组前期工作中^[8], 利用近红外光谱结合化学计量学, 实现了 3 种聚乳酸牌号的离线识别, 但是离线测量存在延时性问题。

为了实现塑料牌号的实时识别, 自主开发了在线近红外光谱测量系统, 在挤出机狭缝模具上安装近红外光谱传感

收稿日期: 2020-08-24, 修订日期: 2020-12-21

基金项目: 国家重点研发计划(2019YFB1704902), 佛山市南海区“蓝海人才计划”项目(201701210004)资助

作者简介: 方 圆, 女, 1998 年生, 华南理工大学硕士研究生 e-mail: fang_yuan1998@163.com

* 通讯作者 e-mail: pmrdd@scut.edu.cn

器,实现对通用聚苯乙烯(general purpose polystyrene, GPPS)熔体的实时测量。通过谱图分析和 K 均值聚类(K-means)算法验证了不同牌号在线光谱数据的可分性,为 GPPS 牌号的在线识别提供了依据。利用 PLS-DA 和随机森林(random forest, RF)分别建立牌号识别模型并进行对比,实现了对 GPPS 牌号的准确识别,提供了一种在生产线上利用近红外光谱在线识别 GPPS 牌号的方法。

1 实验部分

1.1 材料

实验采用的 5 种 GPPS 牌号为: 158K, 5250, 525, PG-33, GP-150。对应的熔融指数分别为 7.3, 7.4, 8.5, 10.0, 10.2 $\text{g} \cdot (10 \text{ min})^{-1}$ 。

1.2 装置

在线近红外光谱测量系统的原理如图 1 所示,卤素光源(LS-3000, 广州标旗有限公司)通过光纤和探头(QR400-7-VIS-NIR, Ocean Optics Inc, USA),将入射光照射到单螺杆挤出机(RESM-20/25, 普同实验分析仪器有限公司)狭缝模具中的熔融物料上,携带样品信息的反射光由探头(QR400-7-VIS-NIR, Ocean Optics Inc, USA)收集,并经光纤传输至近红外光谱仪(NIRQUEST 512, Ocean Optics Inc, USA),通过 USB 接口将光谱数据传输至计算机,在光谱采集软件中实时显示。

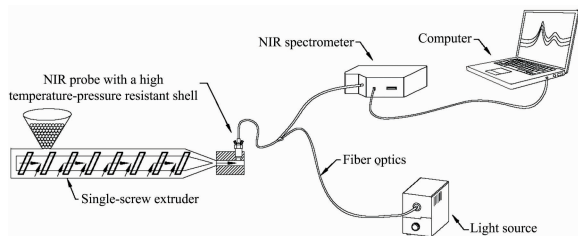


图 1 在线近红外光谱测量系统原理图

Fig. 1 Schematic diagram of in-line near-infrared spectral

1.3 方法

料筒各段温度: 180, 190, 200 和 210 $^{\circ}\text{C}$, 螺杆转速: 80 $\text{r} \cdot \text{min}^{-1}$ 。按照熔融指数由低到高的顺序, 连续挤出不同牌号的 GPPS。光谱波长范围为 900~1 700 nm, 分辨率为 3.1 nm, 积分时间设置为 500 ms, 积分次数设置为 3。共采集到 1 730 条在线近红外光谱, 每种牌号选取 100 条光谱, 共 500 条光谱作为训练集, 在训练集之外, 每种牌号选取 50 条光谱, 共 250 条光谱作为验证集。

1.4 光谱数据处理及建模

近红外光谱按测量方式分为透射光谱和漫反射光谱, 由于挤出过程中的熔融 GPPS 料层较厚, 导致其透射光的强度较弱, 而漫反射光强度高、易于分析, 因此采用近红外漫反射光谱进行分析。首先采用基线校正、最大最小归一化、7 点移动平均平滑 3 种光谱预处理方法对 1 730 条原始在线近红外光谱进行预处理, 以消除光谱的平移、漂移、无关信息和噪声。然后采用主成分分析(principal component analysis, PCA)对预处理后的光谱数据进行降维, 提取光谱数据的主要特征分量, 以简化建模过程。利用 K-means 聚类算法对降维后的光谱数据进行聚类分析, 验证不同牌号在线的近红外光谱可分性, 为 GPPS 牌号的在线识别提供依据。最后利用 PLS-DA 和 RF 分别对降维后的不同牌号在线近红外光谱数据进行分类识别, 并对比两种模型的牌号识别准确率。上述过程均利用 Python scikit-learn 机器学习库来实现。

2 结果与讨论

2.1 不同牌号在线近红外光谱的可分性分析

图 2(a)是经过预处理后的在线近红外光谱图, 每种牌号各选取了 5 条光谱, 共 25 条光谱。可以看到, GPPS 在 1 143, 1 207, 1 388, 1 407 和 1 429 nm 处存在特征峰。其中 1 143 nm 属于芳烃 C—H 伸缩振动的二级倍频, 1 207 nm 属于亚甲基 C—H 伸缩振动的二级倍频, 1 388, 1 407 和 1 429 nm 属于亚甲基 C—H 对称和反对称伸缩与弯曲振动的组合

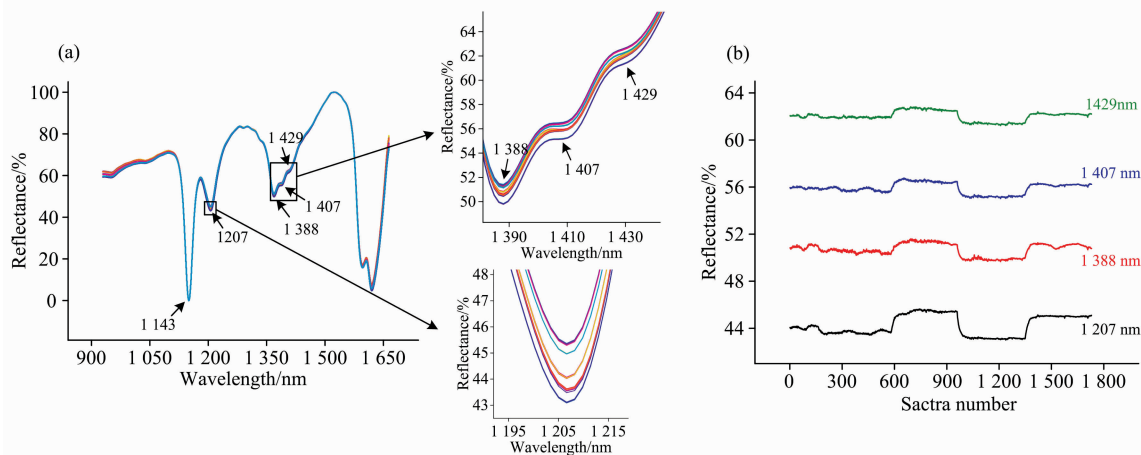


图 2 (a) 不同牌号 GPPS 的在线近红外光谱图, (b) 在线近红外光谱特征峰的峰强变化

Fig. 2 (a) In-line NIR spectra of different grades of GPPS, (b) Changes in characteristic peaks of in-line NIR spectra of different grades of GPPS

频^[9]。由图 2(a)可以发现 1 143 nm 处的特征峰峰强较弱,且特征峰几乎没有差异,但 1 207, 1 388, 1 407 和 1 429 nm 处的特征峰存在细微差异,因此将后四处的特征峰峰强在图 2 (b)中按照光谱采集顺序对预处理后的 1 730 条在线近红外光谱进行分析,其中 158K, 5250, 525, PG-33 和 GP-150 对应的光谱数约为 0~300, 350~600, 650~900, 1 000~1 300 和 1 400~1 730, 其余光谱为共混过程光谱。由图 2(b)可以发现,除前两个牌号 GPPS 的特征峰峰强变化较小外,后续特征峰峰强均随着牌号的变化出现明显阶梯状改变,初步可以判定在线近红外光谱对不同牌号 GPPS 具有一定的区分度。

训练集经 PCA 降维得到的主成分解释变量累计贡献率如图 3 所示, PC3 及之后的累计贡献率均大于 98.67%, 足以代表全部光谱特征。对降至 3 维的训练集进行聚类分析, K-means^[10] 聚类结果如图 4 所示, 错误聚类共 60 例, 参与聚类的光谱共 500 条, 因此, 聚类总正确率为 88%, 表明不同牌号在线近红外光谱的可分性, 这是利用近红外光谱在线识别 5 种牌号 GPPS 的前提和基础。

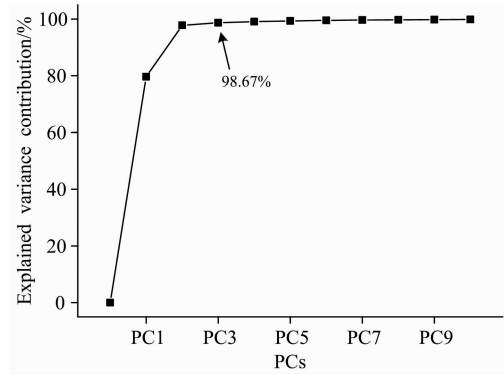


图 3 在线近红外光谱主成分累计解释变量图

Fig. 3 Explained variance contribution of principal components of in-line NIR spectra

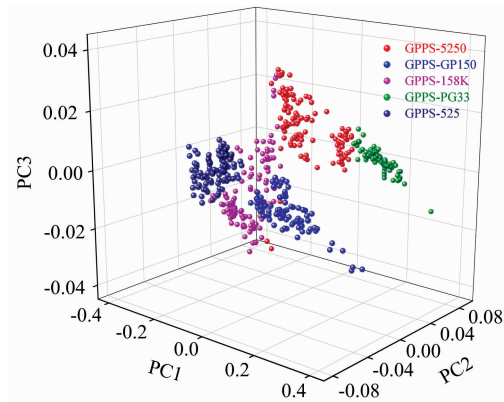


图 4 不同牌号在线近红外光谱的 K-means 聚类结果图

Fig. 4 K-means clustering result of in-line NIR spectra of different grades of GPPS

2.2 PLS-DA 算法识别结果

采用预处理后的训练集和验证集进行建模和预测。为达到最佳识别效果,通过 5 折交叉验证^[11]对主因子数寻优。

图 5 为交叉验证得到的预测残差平方和 (prediction residual error sum of squares, PRESS) 与主因子数的关系图,由图 5 可以发现, PRESS 值下降到不再发生显著变化时,对应的主因子数为 3,此时的 PRESS 值为 26.644, 决定系数 R^2 为 0.973。因此采用最佳主因子数为 3 建立 PLS-DA 模型,得到训练集和验证集的分类正确率分别为 92.0% 和 90.4%。验证集的分类结果如图 6(a)所示,详细结果见表 1。

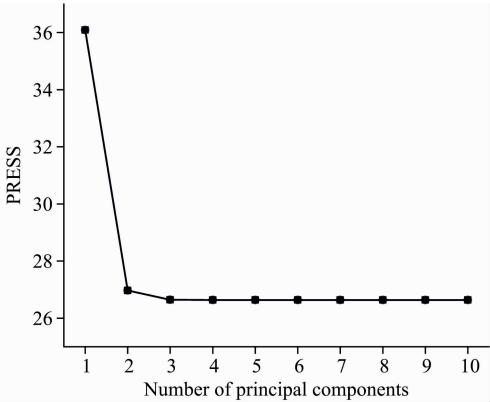


图 5 PRESS 值与主因子数的关系图

Fig. 5 Relationship between PRESS and the number of principal components

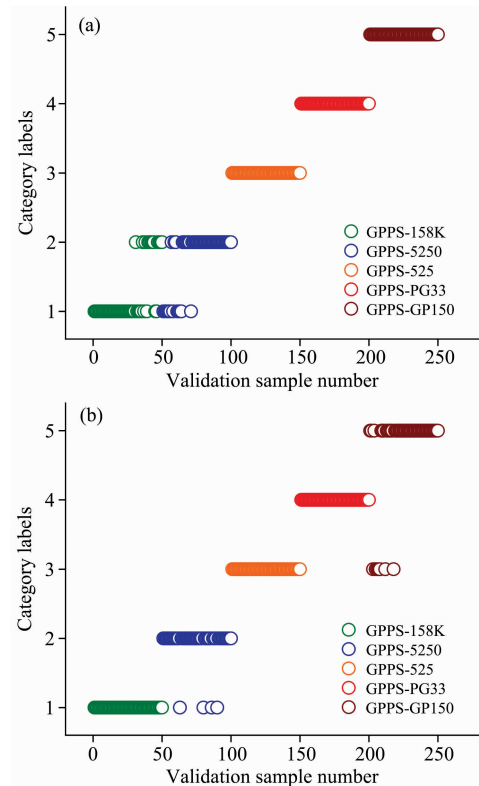


图 6 GPPS 牌号识别模型的验证集判别结果

(a): PLS-DA 模型; (b): RF 模型

Fig. 6 Identification results for different grades of GPPS in validation set

(a): PLS-DA model; (b): RF model

表 1 基于 PLS-DA 判定的不同牌号 GPPS 识别结果

Table 1 Identification results of different grades of GPPS based on the PLS-DA algorithm

分类变量	牌号	验证集个数	识别正确个数	识别正确率/%
1	158K	50	38	76
2	5250	50	38	76
3	525	50	50	100
4	PG-33	50	50	100
5	GP-150	50	50	100

2.3 RF 算法识别结果

RF 算法是通过在 CART^[12] 中引入集成学习装袋方法 (Bagging)^[13] 进行多次随机抽样而构建的^[14], 可以有效地避免模型过拟合。采用 PCA 将预处理后的训练集和验证集降至 5 维, 在训练集样本中, 每次随机选取 100 个样本作为子模型, 共建立 500 个子模型来训练 RF 模型, 并对 RF 的基学

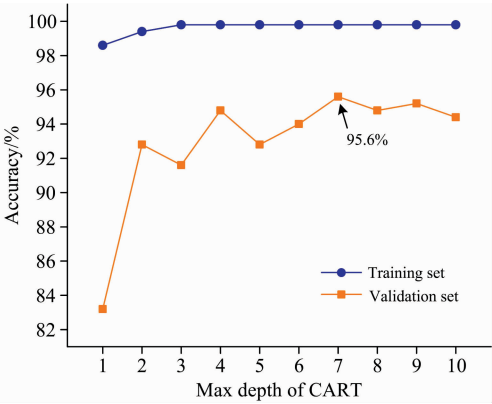


图 7 CART 最大深度与不同牌号 GPPS 的分类正确率关系图
Fig. 7 Relationship between maximum depth of CART and classification accuracy of different grades of GPPS

习器 CART 的最大深度进行寻优, 以达到最佳的识别效果。图 7 为 CART 最大深度与训练集、验证集的分类正确率关系图, 当 CART 最大深度为 7 时, 对应的验证集分类正确率最高, 此时训练集的分类正确率为 99.8%, 验证集的分类正确率为 95.6%。验证集的识别结果如图 6(b)所示, 详细结果见表 2。

表 2 基于随机森林判定的不同牌号 GPPS 识别结果

Table 2 Identification results of different grades of GPPS based on the RF algorithm

分类变量	牌号	验证集个数	识别正确个数	识别正确率/%
1	158K	50	50	100
2	5250	50	46	92
3	525	50	50	100
4	PG-33	50	50	100
5	GP-150	50	43	86

与 PLS-DA 模型相比, RF 模型的牌号识别正确率更高。这是因为: ①光谱与牌号的对应关系并非完全线性, 而 PLS-DA 是一种基于线性回归的判别算法^[15], 不具备 RF 算法的非线性数据处理能力; ②集成学习方法可以提高单一弱分类器的准确率, 在复杂样本中表现更加优异^[16]。

3 结 论

利用自主开发的在线近红外光谱测量系统实时采集了 5 种不同牌号 GPPS 熔体的近红外光谱数据, 通过谱图分析和 K-means 聚类分析方法验证了不同牌号 GPPS 在线近红外光谱数据的可分性, 建立的 PLS-DA 和 RF 模型均实现了对不同牌号在线近红外光谱的准确识别, 其中 RF 模型的识别准确率更高。因此, 近红外光谱是一种在线测量牌号的有效手段。

References

[1] CHEN Le-yi(陈乐怡). Synthetic Resins and Plastic Grades Manual(合成树脂及塑料牌号手册). Beijing: China Petrochemical Press(北京: 中国石化出版社), 2003.

[2] CHU Xiao-li(褚小立). Molecular Spectroscopy Analytical Technology Combined With Chemometrics and Its Application(化学计量学方法与分子光谱分析技术). Beijing: Chemical Industry Press(北京: 化学工业出版社), 2011. 10+259.

[3] Mclauchlin A R, Ghita O, Gahkani A. Polymer Testing, 2014, 38(18): 46.

[4] LI Wen-huan, JIN Shang-zhong, CHEN Ling-ling, et al(李文环, 金尚忠, 陈玲玲, 等). China Plastics Indusrty(塑料工业), 2016, 44(12): 124.

[5] Rani M, Marchesi C, Federici S, et al. Materials, 2019, 12(17): 2740.

[6] Kaihara M, Satoh M. Analytical Sciences, 2007, 23(7): 921.

[7] HAO Yong, WEN Qin-hua, RAO Min, et al(郝 勇, 温钦华, 饶 敏, 等). Food and Machinery(食品与机械), 2018, 34(4): 124.

[8] ZHU Shi-chao, YOU Jian, JIN Gang, et al(朱世超, 游 剑, 晋 刚, 等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2018, 38(10): 71.

[9] WU Xiao-yu, LI Jia, YAO Lin-ping, et al. Journal of Cleaner Production, 2020, 246: 118732.

[10] Camastra F, Verri A. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27(5): 801.

[11] SUN Cheng, WANG Wei, LIU Fang-tian, et al(孙 成, 王 卫, 刘方田, 等). Research of Environmental Sciences(环境科学研究), 2019, 32(9): 1500.

- [12] Zacharis N Z. International Journal of Intelligent Systems and Applications, 2018, 10(3): 1.
- [13] Yariyan P, Janizadeh S, Phong T V, et al. Water Resources Management, 2020, 34: 3037.
- [14] Breiman L. Machine Learning, 2001, 45(1): 5.
- [15] Ballabio D, Consonni V. Analytical Methods, 2013, 5(16): 3790.
- [16] XU Hao-ran, XU Bo, XU Ke-wen(徐浩然, 许波, 徐可文). Computer Engineering and Applications(计算机工程与应用), 2020, 56(12): 19.

In-Line Identification of Different Grades of GPPS Based on Near-Infrared Spectroscopy

FANG Yuan, HE Zhang-ping, ZHU Shi-chao, LIANG Xian-rong, JIN Gang*

National Engineering Research Center of Novel Equipment for Polymer Processing, Guangdong Provincial Key Laboratory of Technique and Equipment for Macromolecular Advanced Manufacturing, Key Laboratory of Polymer Processing Engineering of Ministry of Education, South China University of Technology, Guangzhou 510640, China

Abstract Misusing the wrong grades polymer during the polymer processing in the same production line may lead to poorer product performance and a lower qualification ratio. The traditional methods identifying the grades from same kind of polymer are usually time consuming and hysteretic. There has not yet been discovered a fast and real time method for grade identification. In this work, 5 different grades of GPPS were the research object. An in-line near-infrared spectral measurement system installed on the extruder was developed. Near-infrared spectroscopy was combined with chemometrics and machine learning algorithms. The different grades of GPPS could be fast and in-line identified during the extrusion process. First, the in-line near-infrared spectra of GPPS melts of 5 different grades were collected in real time by the developed system with a spectral range of 900~1 700 nm. After spectrum analysis, a K-means clustering algorithm in combination with PCA was performed to verify the distinguishability of in-line near-infrared spectra for different grades. Last, PLS-DA and RF algorithm were used to establish the grade identification models respectively, and the identification ability of these two models was compared. The results show that: ①After baseline correction, maximum and minimum normalization, and 7-point moving average smoothing, the characteristic peak values at 1 207, 1 388, 1 407, 1 429 nm of the in-line near-infrared spectra change in a step-like manner with the change of grades. With the first three principal components scores as input variables, the clustering accuracy by K-means can reach 88%. It shows the distinguishability of the in-line near-infrared spectral data of different grades of GPPS; ②The two prediction models established by PLS-DA and RF can both effectively identify the grades of GPPS. The classification accuracy on the validation set of the PLS-DA model with the optimal principal components of 3 can reach 90.4%. The classification accuracy on the validation set of the RF model with the first five principal components as input variables can reach 95.6%. The RF model shows better grade identification performance than that of the PLS-DA model. Therefore, combined with chemometrics and machine learning algorithms, the in-line near-infrared spectral measurement system can realize the rapid and in-line identification of GPPS grades. It provides a reference for the in-line identification of different grades of the same kind of polymer by near-infrared spectroscopy in a production line.

Keywords Near-infrared spectroscopy; Grade identification; General purpose polystyrene; Partial least square-discriminant analysis; Random forest

(Received Aug. 24, 2020; accepted Dec. 21, 2020)

* Corresponding author