文章编号：2095-2163(2020)05-0093-05 中图分类号：TP391 文献标志码：A

基于 loess 回归加权的单细胞 RNA-seq 数据预处理算法

高美加
（哈尔滨工业大学 计算机科学与技术学院，哈尔滨 150040）

摘 要：单细胞 RNA 测序数据提供了研究细胞异质性和在生物条件下差异表达基因的机会，其中一些在细胞中表达量有显著变化的高变异基因对单细胞测序数据的下游分析有着关键的作用。本文提出一种基于 LOESS 回归加权的单细胞 RNA-Seq 数据预处理算法，处理基因在细胞中的表达量数据，使高变化基因在分析过程中作用加强，达到基因软筛选与数据降噪的目的。进一步，选取 6 组单细胞 RNA-seq 数据对算法进行测试，首先对生成的基因表达矩阵进行预处理，然后分析预处理对后续分析（可视化、聚类、差异表达分析）的影响，实验结果表明该算法有效提升了下游分析准确度，显示出良好应用价值。
关键词：单细胞；RNA 测序；数据预处理

Single cell RNA-seq data preprocessing algorithm based on LOESS regression weighting

GAO Meijia
（College of Computer Science and Technology, Harbin Institute of Technology, Harbin 150040, China）

【Abstract】 Single-cell RNA-seq data provides us with the opportunity to study cell heterogeneity and differentially expressed genes under biological conditions. Some highly variable genes play a key role in the downstream analysis of single-cell sequencing data. This paper proposes a single-cell RNA-Seq data preprocessing algorithm based on LOESS regression weighting to process gene expression data in cells, so that high-variation genes are strengthened in the entire analysis process to achieve gene soft screening and data noise reduction . Further, I selected 6 single-cell RNA-seq datasets to test the algorithm, first preprocessed the gene expression matrix generated, and then analyzed the impact of pretreatment on subsequent analysis (visualization, clustering, differential expression analysis), experimental results shows that the algorithm effectively improves the accuracy of downstream analysis and shows good application value.
【Key words】 Single cell; RNA-seq; data preprocessing

0 引 言

相比于传统的细胞测序方法，单细胞 RNA-seq 的测序数据提供了研究细胞异质性和基因差异表达的机会，但是单细胞 RNA-seq 通常表现出比来自大量细胞群的 RNA-seq 数据更高水平的噪声和更多的零值。scRNA-seq 数据的计算分析包括质量控制、定位、定量、标准化、聚类几个步骤，用于鉴定差异表达的基因。上游的步骤可能对结果产生实质性的影响。scRNA-seq 的大多数分析，如基因差异表达分析、细胞类型特异性基因的鉴定、分化轨迹的重建等，都依赖于基因表达测量的准确性。目前，对单细胞 RNA-seq 得到的矩阵的预处理方法主要是对矩阵进行插值，以此减轻过多零值对后续的影响。此方法利用单细胞基因表达数据的结构，通过利用相关细胞或者基因表达之间的相似性来校正基因的表达量[1]。例如：sclmpute 是利用混合模型来定位可能的缺失值，之后对其进行插补；MAGIC 和 SAVER 是对矩阵去噪，生成一个新的矩阵，以上都是通过线性来对矩阵进行去噪[1]。另外，也有一些使用神经网络的方法来进行插补的算法，使用自编码器可以通过无监督的方式，最小化重建数据和原始数据之间的误差，来进行非线性的差值，同时也可以进行有效的数据压缩，例如：DCA 算法。

另外，在单细胞 RNA-seq 数据的降噪方法中常用的有基因筛选和降维。基因筛选即筛选出在细胞中表达量变化大的基因，这样可以去除低变化高表达量基因对后续分析的影响[2]。在 RNA-seq 数据分析中常用的降维方式有 PCA、KPCA 和 t-SNE 等。

常用的基因筛选算法有 Seurat 包里的 disp、vst、mvp 等。但是，虽然一些管家基因的表达信息对于细胞的分类并不能起到什么关键的作用[3]，降低这些基因的影响，可能会对后续分析（细胞聚类等）有一些提升。在单细胞表达矩阵的预处理过程中，通常会先回归拟合基因在细胞中表达量的标准差与平

作者简介：高美加（1997-），女，硕士研究生，主要研究方向：计算机科学与技术、生物信息。
收稿日期：2020-02-06

均值的变化曲线来对基因进行筛选,但是这样会损失一部分信息,从而影响后续的分析质量。

基于以上问题,本文提出一种基于 Loess 回归加权的单细胞转录组数据预处理算法,通过 Loess 回归曲线定量计算基因表达偏移水平,并基于偏移水平构造基因加权系数,达到基因软筛选与数据降噪的目的。本文选择 6 组单细胞 RNA-seq 数据从可视化和聚类两方面对算法预处理效果进行测试,实验证明该方法可以有效降低低质量基因对分析过程的影响,提升下游分析的精准水平,显示出较好应用价值。

1 预处理方法研究

1.1 回归加权

通过量化基因在每个细胞里表达的高变异度,对表达量矩阵进行加权来降低变化度的基因对后续分析的影响。在此使用局部加权回归(LOESS)拟合基因在细胞中的表达量的标准差与平均值的变化曲线,使用实际的标准差和预测的标准差之间的差值作为每个基因的权重,然后生成新的表达矩阵,如式(1)~式(4)所示。

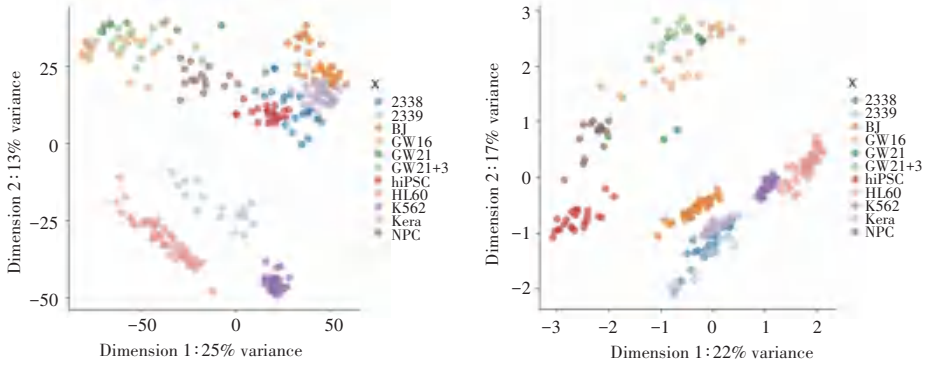
$$Model = Loess(sd_i \sim mean_i), \tag{1}$$

$$weight_i = sd_i - Predict(mean_i), \tag{2}$$

$$weight_i = log\frac{(weight_i - weight_{min})}{(weighty_{max} - weight_{min})}, \tag{3}$$

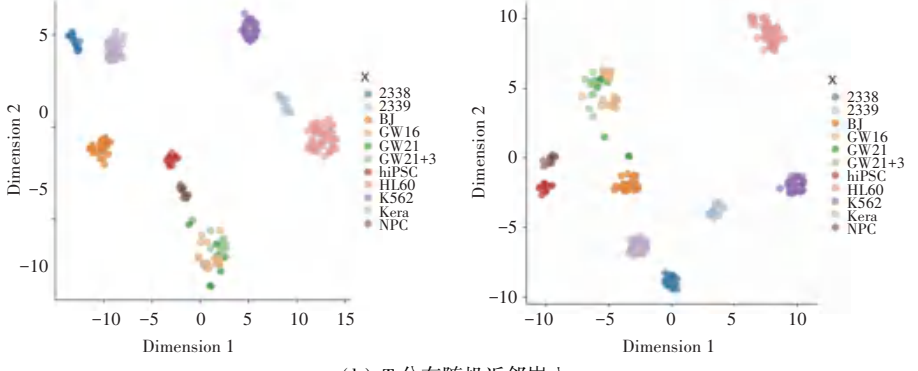
$$x'_{ij} = x_{ij} \times weight_i. \tag{4}$$

其中,$mean_i sd_i$ 为基因 i 的表达的平均值和标准差,x_{ij} 为矩阵中的元素,x'_{ij} 为新生成的表达值。以 Pollen 数据集为例,Pollen 数据集经变换后,如图 1 所示。在图 1(a)中,可以看出,经 PCA 降维后,预处理后数据集的可视化效果要好一些,有几类细胞在图中被有效分离开,图 1(b)是经过 TSNE 降维后的效果,各个簇也更聚集一些。



(a) 主成分分析

(a) PCA



(b) T 分布随机近邻嵌入

(b) TSNE

图 1 Pollen 数据集

Fig. 1 Pollen dataset

1.2 标准化

由于技术原因,单细胞 RNA-seq 数据中基因表达显示出明显的细胞差异,可能是由于生物学和技术上的双重原因造成的。在此使用了 Hafemeister 等人提出的一个标准化方法来降低测序深度对基因表达造成的影响,公式(5)和公式(6)如下。

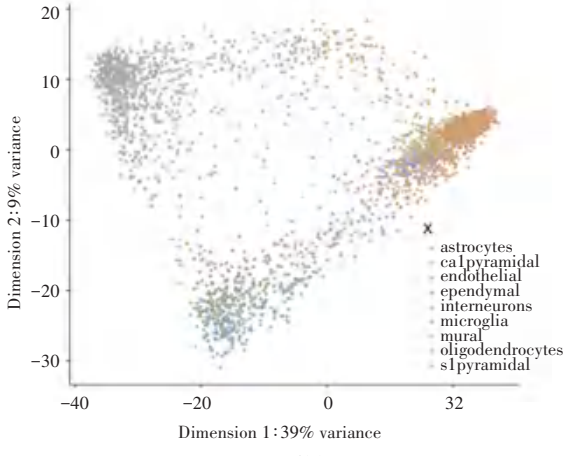
$$\log(x_i \cdot) = \beta_0 + \beta_1 \log_{10} m, \tag{5}$$

$$\mu_{ij} = \exp(\beta_{0_i} + \beta_{1_i} \log_{10} m_j. \tag{6}$$

其中：m 为细胞 j 中基因的总的表达量，分别对每个基因在细胞中的表达量和细胞总的表达量做线性回归，计算出每个基因在细胞中期望的表达值，根据新的矩阵使用上述的方法做预处理。图 2 中，以 Zeisel 数据集为例，图 2(a)为未处理过的数据，图 2(b)为使用 LOESS 加权处理过的新的矩阵，图 2(c)为标准化后加权处理形成矩阵的可视化效果。可以看出，在图 2(a)中形成了 3 个比较大的簇，在图 2(b)和(c)中都有其它的簇分裂出来，可视化效果较好。

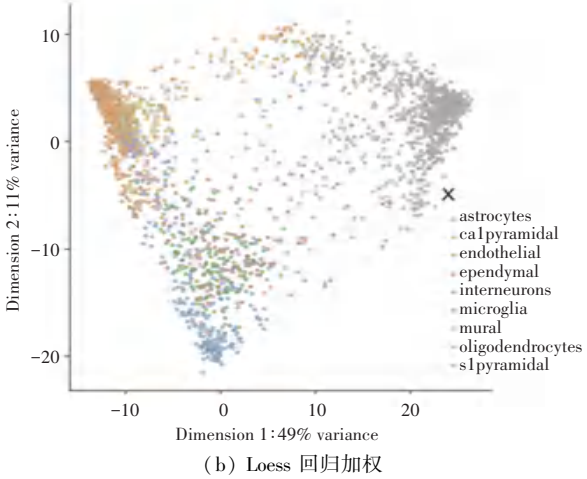
2　实验结果

分别从可视化效果和无监督聚类效果两方面来对预处理效果进行评价。在这里选取了 6 个数据集：pollens，Biase，Yan，Goolams，Deng，Zeisel。其中 Zeisel 数据集中的标签为 SC3 算法得出的标签。数据来源：https://hemberg-lab.github.io/scRNA.seq.datasets/。本实验使用 R 语言中的 scater 包进行大部分的单细胞 RNA-seq 数据分析。



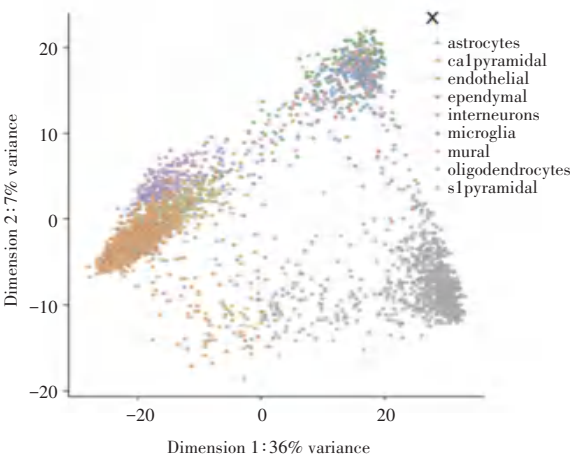
(a) 原数据

(a) Raw data



(b) Loess 回归加权

(b) Loess regression weighting



(c) 标准化处理后数据

(c) Data after standardization

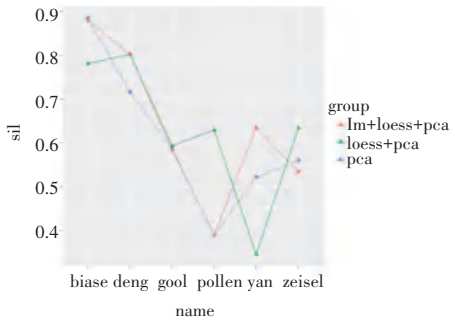
图 2　Zeisel 数据集

Fig. 2　Zeisel dataset

2.1　对可视化效果影响

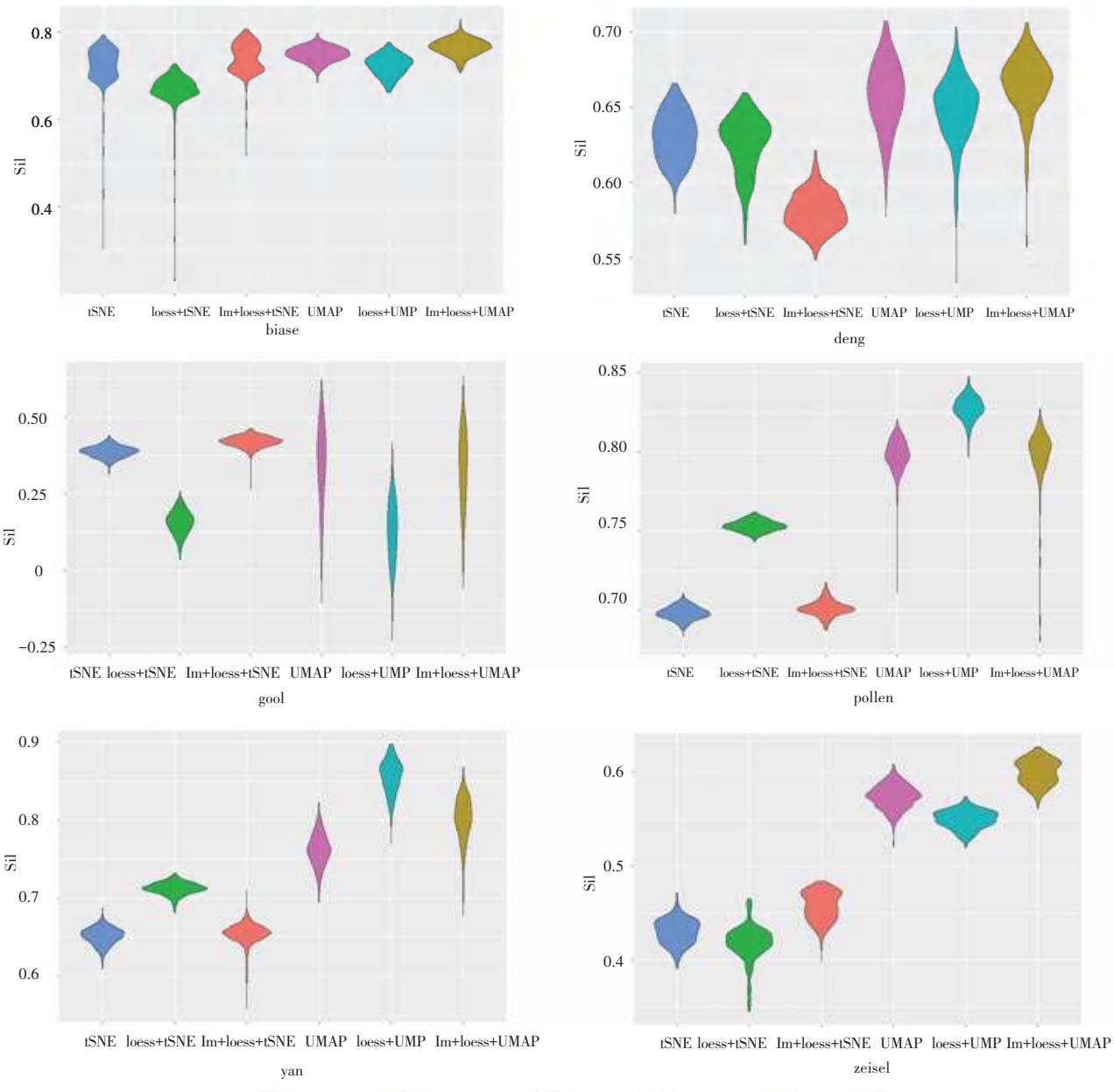
选取了在单细胞 RNA-seq 分析中比较常用的 3 种可视化方法来进行对比实验，这 3 种方法分别是：PCA、TSNE、UMAP。其中 TSNE 和 UMAP 由于实现过程中有随机性，所以重复了 500 次实验进行对比。使用轮廓系数(Silhouette Coefficient)来对图中同种细胞的聚集程度和不同种细胞的离散程度进行量化。轮廓系数就是针对样本空间中的一个特定样本，计算它与所在聚类其它样本的平均距离 a，以及该样本与距离最近的另一个聚类中所有样本的平均距离 b，该样本的轮廓系数为 $(b-a)/\max(a, b)$，将样本空间中所有样本的轮廓系数取算数平均值，作为聚类划分的性能指标 s。在前两个维度里，类间距离越大，类内距离越小，就认为这个可视化效果是好的。

可视化实验结果如图 3 所示(每张图中前 3 个为 TSNE 可视化的轮廓系数对比，后 3 个为 UMAP 可视化后轮廓系数对比；这三列分别为原始矩阵、标准化后又加权的矩阵和只回归加权的可视化轮廓系数对比。图 3(b)从左到右，从上到下分别为：biase、deng、gool、pollen、yan、zeisel 数据集)。通过比较发现，对于 PCA 来说，在 biase、deng 和 yan 数据集中标准化之后再进行预处理的效果较好，在 pollen 和 zeisel 数据集中直接进行预处理效果较好。对于 tSNE 和 UMAP 算法来说，biase、gool、zeisel 数据标准化之后再进行预处理效果较好，在 pollen、yan、deng 这 3 个数据集中直接进行预处理效果更好。从实验结果来看，基因表达矩阵经算法处理过之后，经过降维，前两维中数据同类样本更集中，不同类样本之间也更加分散，它对后续的可视化效果是有一定提升的。



（a）主成分分析

（a）PCA



（b）T 分布随机近邻嵌入+ 统一流形逼近与投影

（b）tSNE+UMAP

图 3 可视化效果对比

Fig. 3 Visual effect comparison

2.2　对无监督聚类的影响

选取 3 个常用的单细胞测序数据的聚类算法：SC3、SIMLR、和 Seurat。使用 *F*1 - *score* 来对聚类结果进行评价。*F*1 - *score* 具体定义为公式(9)：

通过公式(7)和(8)计算每个类别下的 *precision* 和 *recall*：

$$precision_k = \frac{TP}{TP + FP}, \tag{7}$$

$$recall_k = \frac{TP}{TP + FN}, \tag{8}$$

$$f_{1_k} = \frac{2 \times precision_k \times recall_k}{precision_k + recall_k}. \tag{9}$$

其中，*TP*（True Positive）预测答案正确；*FP*（False Positive）错将其他类预测为本类；*FN*（False Negative）本类标签预测为其他类标。最后通过公式(10)计算各个类别下的 *F*1 - *score* 的平均值：

$$score = (\frac{1}{n}\sum f_{1_k})^2. \tag{10}$$

*F*1 - *score* 是精确率和召回率的调和平均数，最大为 1，最小为 0。

聚类分析结果如表 1 所示，其中 lm+loess 为经标准化后的回归加权方法。经过比较发现两种预处理方法对这三种聚类方法的实验结果都有一定的提升作用，其中 LM+LOESS+SC3、SC3、LOESS+SIMLR 在多数数据集中表现都比较好，说明回归加权的方法是对后续的无监督聚类分析有一定的提升作用。

表 1　聚类结果比较

Tab. 1　Comparison of clustering results

Method	pollen	biase	gool	deng	yan
SC3	0.962 835 1	**1.000 000 0**	0.773 636 4	0.758 702 1	0.727 272 7
LOESS+SC3	**0.962 842 6**	0.914 847 2	0.666 974 2	0.744 048 8	**0.740 514 1**
LM+LOESS+SC3	0.962 632 0	**1.000 000 0**	0.773 636 4	0.770 214 8	0.727 272 7
Seurat	0.869 367 4	0.731 993 3	0.711 538 5	0.655 350 6	0.698 042 9
LOESS+Seurat	0.869 367 4	0.731 993 3	**0.926 880 2**	0.584 643 7	0.722 790 7
LM+LOESS+Seurat	0.869 367 4	0.731 993 3	**0.926 880 2**	0.663 053 6	0.722 790 7
SIMLR	0.784 462 9	0.982 817 9	0.491 541 4	0.850 008 3	0.683 896 6
LOESS+SIMLR	0.947 943 6	0.969 837 6	0.613 817 5	**0.917 950 0**	0.667 961 2
LM+LOESS+SIMLR	0.847 168 0	0.982 817 9	0.601 795 8	0.770 569 3	0.657 068 1

3　结束语

本文提出的基于 Loess 回归加权单细胞 RNA-seq数据的预处理算法。可以看出，在一些数据集中，预处理之后的可视化和无监督聚类过程都有一定的提升作用，数据经过 PCA 或者 t-SNE 降维后，经处理后的数据同类细胞间往往表现的更加聚集，不同类之间更加分散，这同样会加强后续的聚类效果，使聚类算法表现更好。gool丨、deng、yan 等人的数据集经过预处理后，聚类结果准确度明显有了很大的提升。但是此算法也有一定的局限性，预处理之后的数据产生的值并不符合矩阵中元素为基因在细胞中的表达量这一定义，不利于差异表达基因等下游分析的进行，还有待进行一些分析与改进。

参考文献

[1] ERASLAN G, SIMON L M, MIRCEA M, et al. Single-cell RNA-seq denoising using a deep count autoencoder[J]. Nat Commun 10, 390 (2019).

[2] LI W V, LI J J. An accurate and robust imputation method scImpute for single-cell RNA-seq data[J]. Nat. Commun. 9, 997 (2018).

[3] van DIJK D. MAGIC: a diffusion-based imputation method reveals genegene interactions in single-cell RNA-sequencing data [J]. bioRxiv (2017).

(上接第 92 页)

[4] 贾丽姣，邱勇. 基于数值模拟的激光超声裂纹检测仿真技术[J]. 中国科技信息，2019(18)：89-91.

[5] 刘辉，郑宾，王召巴，等. 激光超声透射波表征表面缺陷深度的仿真研究[J]. 中北大学学报(自然科学版)，2017，38(2)：119-123+1.

[6] 赵曦明，战宇，王雨，等. 激光超声方法检测材料缺陷的数值模拟[J]. 科学导报，2016，(3)：242-243.

[7] 王明宇，周跃进，郭冲. 激光超声检测表面裂纹深度的数值模拟[J]. 激光技术，2017，41(2)：178-181.

[8] QING Y, CHUNXIA H. Application of Nondestructive Testing in Composite Materials[J]. Engineering & Test, 2009, 49(2): 24-28.

[9] XU B Q, LIU H, XU G. Mixed stress-displacement finite element method for laser-generated ultrasound [J]. Laser technology, 2014, 2: 230-235.

[10] 刘长福，牛晓光，李中伟，等. 基于 ANSYS 的超声纵/横波传播仿真计算[J]. 无损检测，2011，33(6)：15-18，26.

[11] 卢旭，陈智军，黄鑫，等. 基于 COMSOL 的声表面波标签仿真[J]. 压电与声光，2012，34(4)：494-497.

[12] 艾春安，韩兆林，李剑，等. 基于有限元方法的超声波仿真研究[J]. 电声技术，2015，39(8)：39-41+47.

[13] 张建炎. 激光超声非接触检测高温金属表面缺陷研究[D]. 江苏：南京航空航天大学，2016.