Software Engineering and Applications 软件工程与应用, 2020, 9(1), 82-92
Published Online February 2020 in Hans. http://www.hanspub.org/journal/sea
https://doi.org/10.12677/sea.2020.91010

Hans汉斯

Stability of RNN Classified Gene Sequence on Variant Maps

Tao Li, Jeffrey Zheng*

School of Software, Yunnan University, Kunming Yunnan
Email: 1977675165@qq.com, *conjugatelogic@yahoo.com

Received: Feb. 5th, 2020; accepted: Feb. 20th, 2020; published: Feb. 27th, 2020

Abstract

The recurrent neural network (RNN) has an excellent role to analyze the data characteristics of integer sequences. Using the principle of feature classification, the selected gene sequences are pre-classified, and subsequent processes are performed on the classified sequences. In this paper, the pre-classified gene sequences are segmented and shifted. For the sequence operation after classification, the gene sequence data sets with different shift-length values are obtained, and the different data sets are respectively used as the inputs of the RNN classifier. The sequences of different detection data sets are replaced to obtain the final visualized sequence data set. The variant maps are provided to show a series of visualization results of the variant probability statistics. The stability of the RNN classifier is analyzed through comparison and analysis of the variant maps and other diagrams. Multiple substitution relationships in the replacement operation change lengths of the shift operations to provide a variety of visualization and comprehensive cross-comparisons to support the analysis and in-depth exploration of the stability problems of the RNN classifier.

Keywords

RNN Classifier, Stability, Shift, Sequence Operation, Visualization

基于变值图示判定RNN基因序列分类器的稳定性

李　桃，郑智捷*

云南大学软件学院，云南　昆明
Email: 1977675165@qq.com, *conjugatelogic@yahoo.com

收稿日期：2020年2月5日；录用日期：2020年2月20日；发布日期：2020年2月27日

*通讯作者。

文章引用: 李桃, 郑智捷. 基于变值图示判定 RNN 基因序列分类器的稳定性[J]. 软件工程与应用, 2020, 9(1): 82-92.
DOI: 10.12677/sea.2020.91010

摘　要

循环神经网络(RNN)在分析整数序列的数据特征中具有优异的作用。利用其特征分类的原理，本文将预分类和分类后的基因序列进行包括移位等的序列操作，按照文中框架下模块操作得到一系列的变值概率统计图示可视化结果，通过变值图示以及其他的图示比较与分析，对RNN分类器的稳定性进行分析。在替换操作中多种类的替换关系和移位操作中长度的变化，提供丰富的可视化结果，综合交叉比较结果，有利于对RNN分类器稳定性问题进行分析和深入探索。

关键词

RNN分类器，稳定性，移位，序列操作，可视化

1. 引言

从 1990 年代起，循环神经网络(Recurrent Neural Network, RNN) [1]在处理整数序列中展现出强大的分析和处理功能。通过大规格集成电路等相关计算工具的发展，目前已经有了长足的进步。近年来，针对不同应用，出现了各种基于 RNN 的变体结构，并被应用在各个特殊的应用领域中[2]，循环神经网络已经成为在深度学习领域中一类非常重要的模型。

在现实生活中满足序列性质的数据是普遍存在的，这类数据也助推了 RNN 在实际应用中获得不断地拓展，例如：时间序列处理，股票行情预测[3]，金融数据预测[4]，雷达临近预报[5]；语音序列识别[6]等。对 RNN 的应用研究偏重于的是实际应用方面。本文将基因序列分类作为 RNN 应用的代表，探究 RNN 分类器的稳定性问题，利用变值理论[7] [8] [9] [10]体系框架，达到展示 RNN 分类器所生成的分类序列的稳定性特征，从可视化的角度，展现其移位后不变的特性。

通过可视化方法将 RNN 稳定性抽象概念具体形象化，展示 RNN 分类器分类结果数据集中的数据分布情况以达到对 RNN 分类器稳定性探索与研究的目的。

2. 体系结构

本文将预分类的基因序列进行分段、移位。针对预分类后的序列操作，得到在不同移位长度值条件下的基因序列数据集合，将不同数据集分别作为 RNN 分类器的输入，得到对应的检测数据集。对不同检测数据集中的序列进行替换，得到最终可视化的序列数据集。然后通过统计计算模块操作提供可视化的模块展示。

文中使用的体系结构主要分为以下几个模块：预分类序列操作模块、RNN 分类模块、可视化序列操作模块、统计计算模块、可视化模块。

该体系结构下，数据处理的流程如图 1。

3. 核心处理模块

在文中所使用的框架下不同的处理模块会有不同的操作。

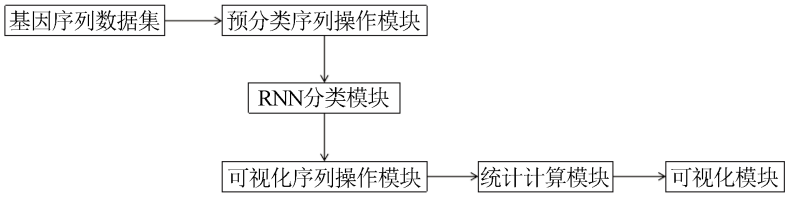

Figure 1. Flow Chart of Data processing
图 1. 数据处理流程图

3.1. 预分类序列操作模块

该模块主要包括序列分段操作和序列移位操作，参阅“图 2”。

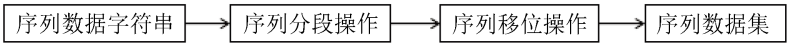

Figure 2. Operation diagram of pre-classification sequence
图 2. 预分类序列操作

首先是序列分段操作，将整个基因序列数据作为字符串，然后按照分段长度(dl)进行分段处理，将长度小于分段长度的序列舍去，使得基因序列字符串转变成为由固定分段长度的序列组成的数据集。

其次是序列移位操作，将序列数据集中的单条序列字符串分别向左平移，平移单位为移位长度值(ml)个字符，最后不同移位长度的移位操作可获得不同的序列数据集。而移位操作中的长度值范围为：1 分段长度值。

例如，序列数据字符串为：GCTGGTCCGCAGCAACACGACCAGGTTGACGTACCGAT，其序列总长度为 38，若将分段长度(dl)固定为 9，则此时可得到数目为 4 段的序列数据集：GCTGGTCCG，CAGCAACAC，GACCAGGTT，GACGTACCG。而移位操作中若是移动长度(ml)为 2，则将序列数据集中的每条数据序列向左平移 2 个序列字符，此时以上数据集的移位操作结果就为：TGGTCCGGC，GCAACACCA，CCAGGTTGA，CGTACCGGA。

3.2. RNN 分类模块

该模块下，将测试数据集中的数据序列进行移动操作后作为 RNN 分类器的输入，进行分类操作后，得到相应的结果数据集作为输出，如“图 3”。

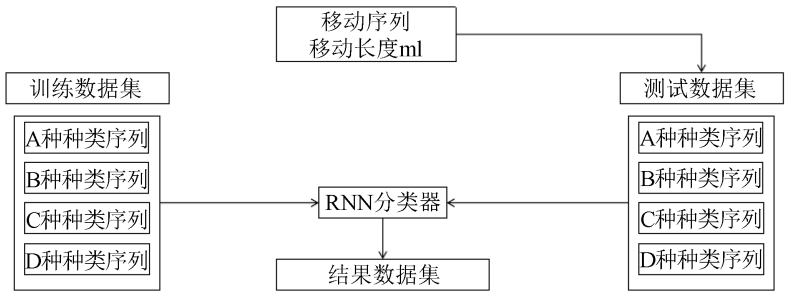

Figure 3. Flow Chart of Classifier classification
图 3. 分类器分类流程图

3.3. 可视化序列操作模块

可视化序列操作主要是两次序列替换操作。

任意一条 DNA 序列根据四种碱基的不同性质，可以被唯一描述为 3 种独立的嘌呤(R)和嘧啶(Y)的分布、氨基(M)和羧基(K)的分布、强氢键(S)和弱氢键(W)的分布，本文将三种分布作为替换关系，如下：

1) RY：嘌呤(R) = A、G；嘧啶(Y) = C、T；

2) MK：氨基(M) = A、C；羧基(K) = G、T；

3) SW：强氢键(S) = G、C；弱氢键(W) = A、T；

根据以上对应的替换关系，一条基因序列可映射成三条不同的序列，此为序列的第一次的替换操作，其主要针对的是序列中的元素。

例如：序列 GTCCACTGGCATGGT 可替换成三条独立的序列：1) RYYYRYYRRYRYRRY；2) KKMMMMKKKMMKKKK；3) SWSSWSWSSSWWSSW。

任意一条序列经过了第一次替换操作后其中的位置关系有四种，如下：

1) RY：‘RR’，‘RY’，‘YR’，‘YY’；

2) MK：‘KK’，‘KM’，‘MK’，‘MM’；

3) SW：‘SS’，‘SW’，‘WS’，‘WW’；

而该序列操作中的第二次替换是在所有位置关系选出特定的位置关系字符串作为替换时的判断依据，将整条数据序列中与之相等的子字符串替换为“1”，否则就替换为“0”，最终将整条序列替换成只包含“0”和“1”的序列。

综上可知，任意一条经过第一次替换的数据序列按照不同的位置替换关系一共可替换成 4 条“01”序列，作为统计计算模块的输入。

例如，若经过第一次替换后的序列：SWSSWSWSSSWWSSW，则第二次替换结果为：① ‘SW’：100101000100001；② ‘SS’：00100001100010；③ ‘WS’：01001010000100；④ ‘WW’：00000000001000。

3.4. 统计计算模块

统计计算模块主要是统计替换后的“01”序列中“1”字符的总数量，然后计算其测度，用“1”的统计总数除以序列的长度。

同一条序列因为不同的序列替换关系会有相对应不同的统计值集合。

3.5. 可视化模块

该模块中主要包括序列差异、分类结果统计、分类结果数据集可视化操作，以下主要介绍针对不同的可视化操作的数据操作。

3.5.1. 序列差异

序列移位操作前后的差异性可通过编辑距离[11]计算两序列相似度[12]来量化。其中编辑距离是指两个字串之间，由一个转成另一个所需的最少编辑操作次数。编辑操作包括将一个字符替换成另一个字符，插入一个字符，删除一个字符。一般来说，编辑距离越小，两个串的相似度越大。

编辑距离用数学语言描述有：

$$ldist_{a,b}(i,j)=\begin{cases}\max(i,j),\min(i,j)=0\\ \min\begin{cases}ldist_{a,b}(i-1,j)+1\\ ldist_{a,b}(i,j-1)+1\\ ldist_{a,b}(i-1,j-1)+1\end{cases},otherwise\end{cases}$$

其中有：$ldist_{a,b}(i,j)$ 指的是 a 字符串中前 i 个字符和 b 字符串中前 j 个字符之间的编辑距离。

其中计算相似度的公式如下：

$$r = \frac{\left(sum - ldist_{a,b}\right)}{sum}$$

以及有：

$$sum = m + n$$

其中 m 和 n 分别为两个字符串的长度，而 $ldist_{a,b}$ 表示字符串 a 和字符串 b 之间的编辑距离。

计算可得序列进行不同长度的移位操作后序列与不进行移位操作序列之间的相似度集合，该集合就作为序列差异可视化的输入数据。

3.5.2. 分类结果统计

将进行不同长度值移动操作后所得的测试数据集经过 RNN 分类器分类，然后对结果数据集进行统计操作，即将结果数据集中数据按照种类为单位进行数据统计。

结果可得到一系列不同长度值移动操作下分类结果的种类数量的统计值，将其作为分类结果统计可视化的数据源。

3.5.3. 分类结果数据集统计

此模块中的可视化数据来源于统计计算模块中的数据集，因为可以选择不同的元素替换关系以及不同的位置替换关系，所以可得不同的数据集。其后选择不同的数据值组合作为可视化 2 维图示的坐标值也就可以得到相应不同的可视化结果，从而获得不同的观察分析角度。

4. 数据简介

文中所涉及的数据为四类细菌的基因序列：① 肠沙门氏菌(*Salmonella enterica*)；② 脓肿分枝杆菌(*Mycobacteroides abscessus*)；③ 桑特氏三角菌(*Terriglobus saanensis*)；④ 苏云金芽孢杆菌(*Bacillus thuringiensis*)。

所有数据全部来源于美国生物基因数据库 NCBI 网站。

本文中将分段长度值固定为 100，移位长度值范围为 1~99。

5. 结果展示与分析

5.1. 结果展示

本文结果主要是结果可视化模块中的内容。

5.1.1. 分类结果数据集统计

将不进行移动操作的数据序列与进行移动操作后的数据序列之间的相似度按照以上公式计算出来，然后计算整个数据集的差异性平均值，最后将其所得平均值可视化，结果如“图 4”。

图中其横坐标为序列移动操作中的移动长度值，其为连续性变化数值，范围就为 1~99，纵坐标为序列移动前后两序列字符串的相似度计算值。

观察可视化图示结果发现，移动长度数值范围在 41~61 时，两序列字符串差异性最大，同时对应的相似度也最低。

5.1.2. 分类结果统计可视化

将分类结果统计结果可视化后，可得如“图 5”。

统计结果是按照分类操作中序列数据集的种类为单位可视化，本文中所涉及的数据种类数量为 4。

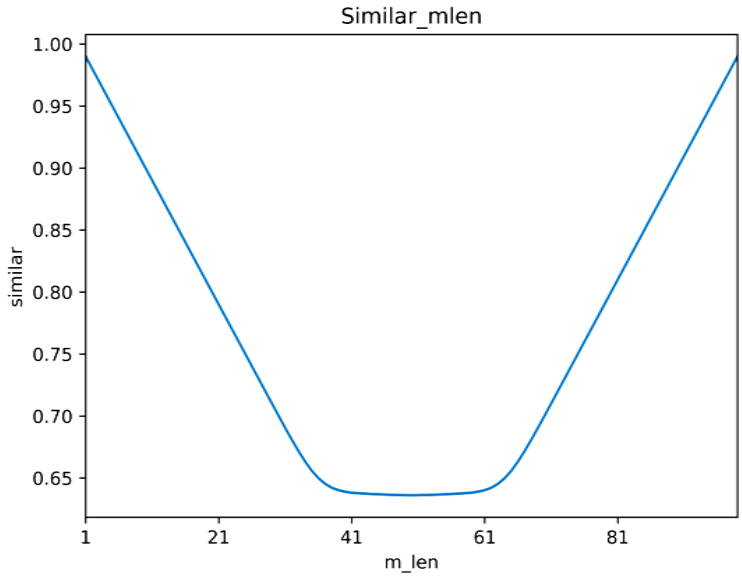

Figure 4. Flow Chart of Classifier classification
图 4. 序列差异值图示

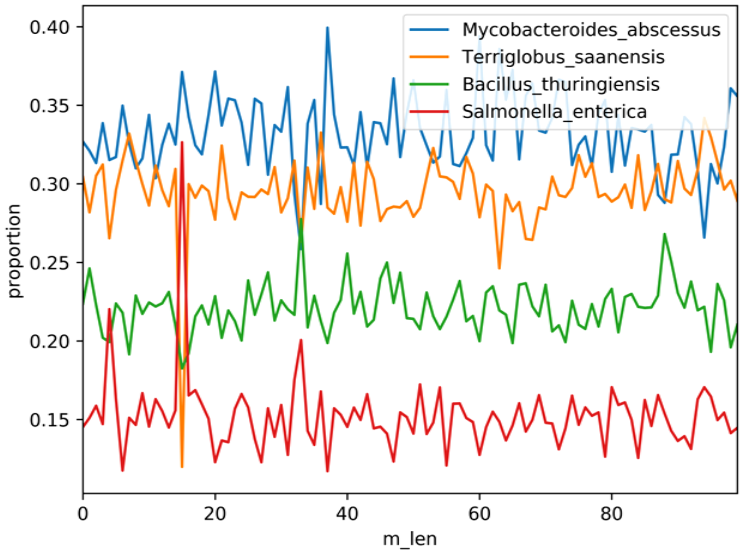

Figure 5. Proportion diagram
图 5. 分类结果数据集比例图

横坐标为不同的连续序列移动操作的移动长度值，值所取范围是 0~99，纵坐标为其中某一类数据所占结果数据集的比例。

从结果图示中，可知在不同的移动长度数值的移动操作条件下，所对应的不同结果数据集中四种不同类别的数据集数量变化并不大。

5.1.3. 分类结果数据集可视化

分类结果数据集统计结果可视化后可以得到一批不同的图示结果，在其中挑选出有特定序列差异值的数据集可视化图示进行展示如图 7 至图 11，对应的移动操作的移动长度分别为 0，1，21，41，61，81。

图示结果为二维频次直方图，其中横纵两坐标轴的值分别是数据集经过两次替换操作后的相应统计值，而值的选取可以是多种不同的组合，由于第一次替换关系总共有三种，而每一种替换关系里位置替换关系总共有四种，所以可以获得大量的可视化组合结果图示。图示中颜色的分布信息表示数据分布情况，颜色由‘红黄青蓝’组成，其中蓝色表示数据含量最少的投影，红色表示数据含量最多的投影。

本文中以下可视化结果展示的是：第一次替换关系是‘SW’，第二次位置替换关系是‘SW’和‘WW’，具体结果如下图系列。

其中单张图示中包括(a)、(b)、(c)、(d)四个部分，每个部分分别代表不同的分类结果，从上至下，从左至右顺序依次是：① 肠沙门氏菌(*Salmonella enterica*)；② 脓肿分枝杆菌(*Mycobacteroides abscessus*)；③ 桑特氏三角菌(*Terriglobus saanensis*)；④ 苏云金芽孢杆菌(*Bacillus thuringiensis*)。

不进行移动操作时，按‘SW’与‘WW’位置关系统计值可得如图“图 6”：

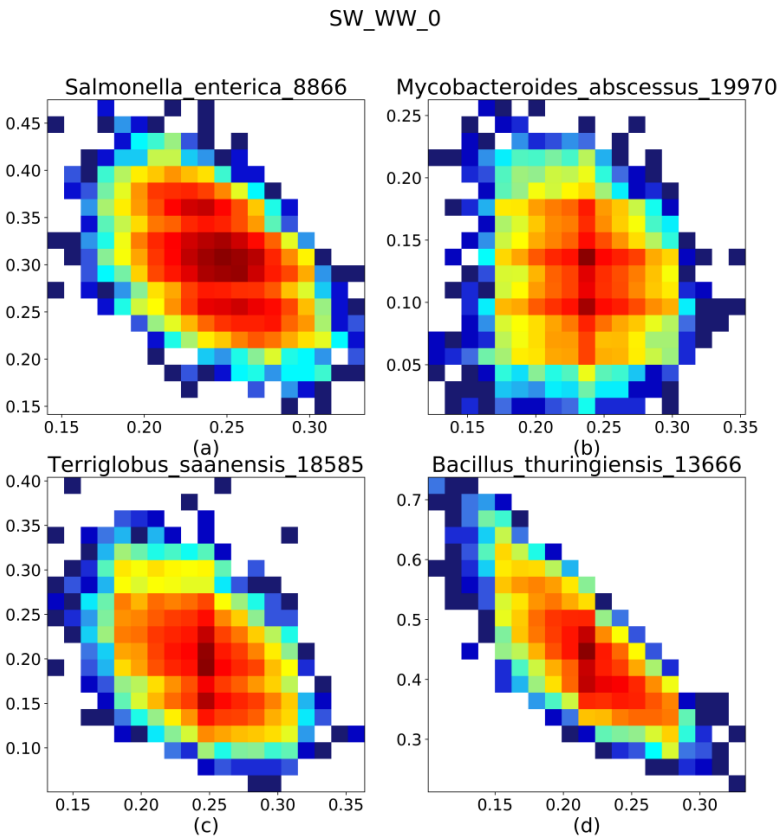

Figure 6. Result of 0 Length-Shift
图 6. 移位长度 0 结果图

进行移位操作，且移位操作中移动长度(mlen)为 1，21，41，61，81 时，按‘SW’与‘WW’位置关系统计值可如下：

1) 移动长度(mlen)为 1 可得如图“图 7”；
2) 移动长度(mlen)为 21 可得如图“图 8”；
3) 移动长度(mlen)为 41 可得如图“图 9”；
4) 移动长度(mlen)为 61 可得如图“图 10”；
5) 移动长度(mlen)为 81 可得如图“图 11”；

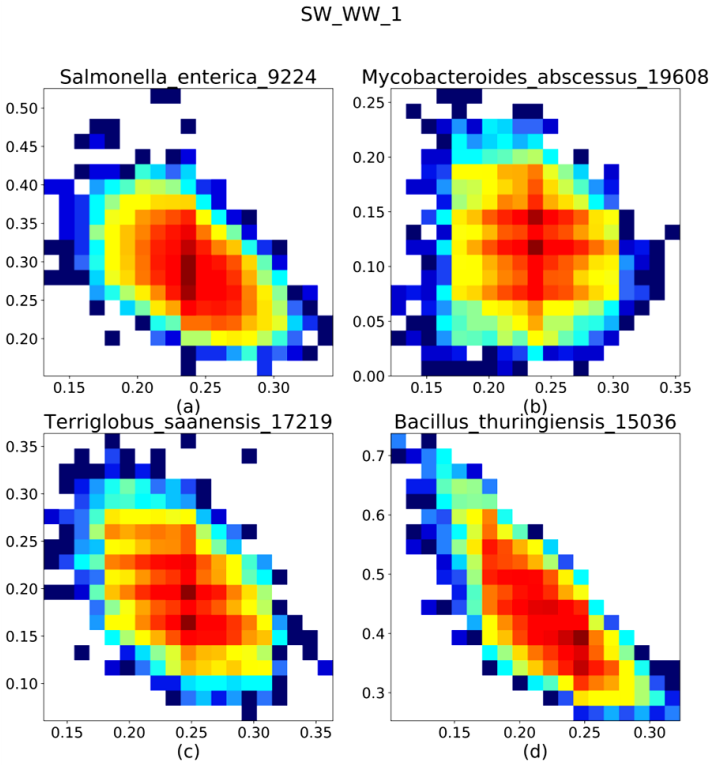

Figure 7. Result of 1 Length-Shift
图 7. 移位长度 1 结果图

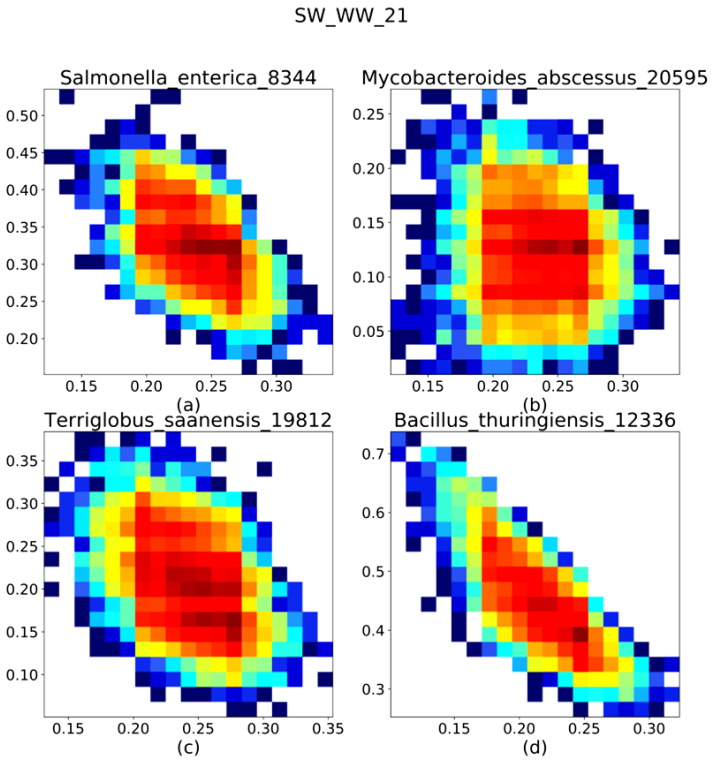

Figure 8. Result of 21 Length-Shift
图 8. 移位长度 21 结果图

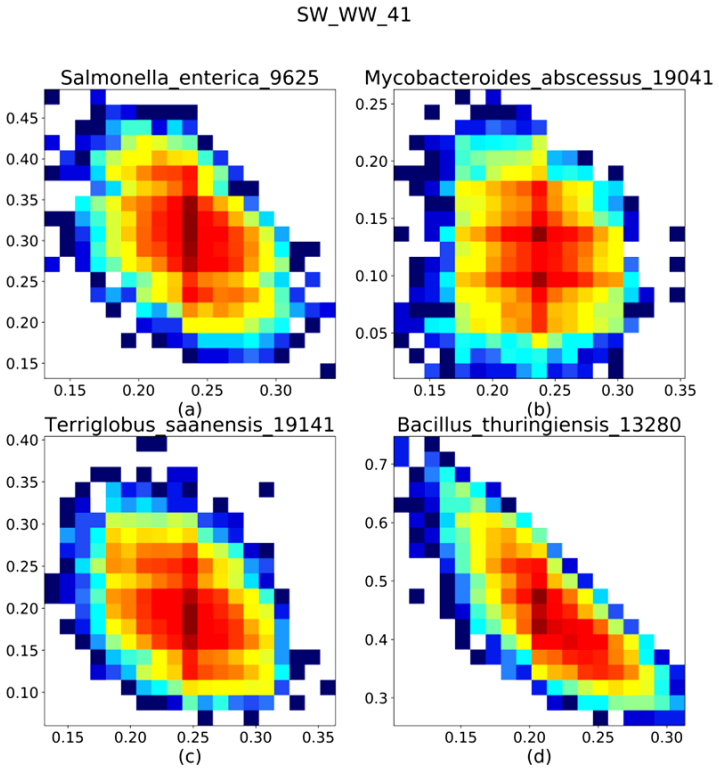

Figure 9. Result of 41 Length-Shift
图 9. 移位长度 41 结果图

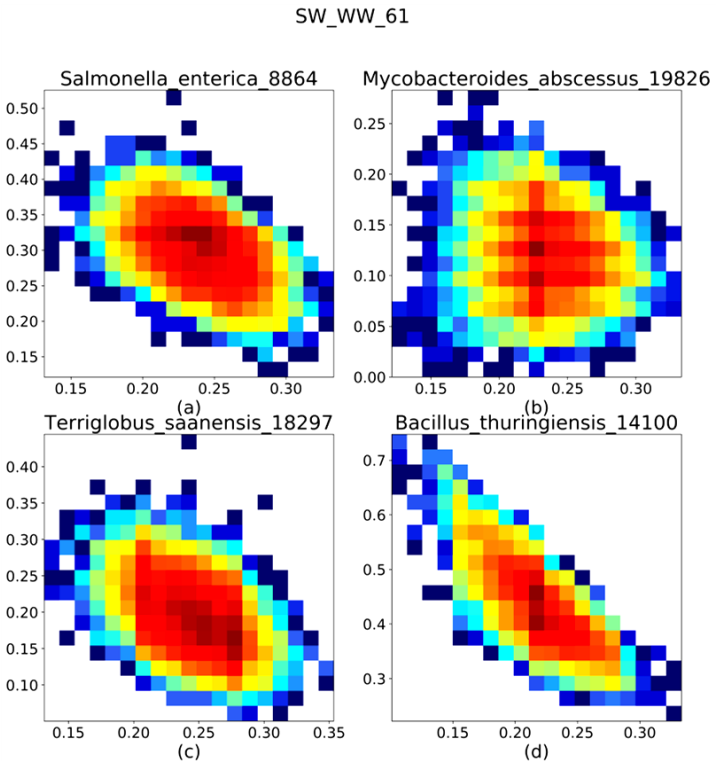

Figure 10. Result of 61 Length-Shift
图 10. 移位长度 61 结果图

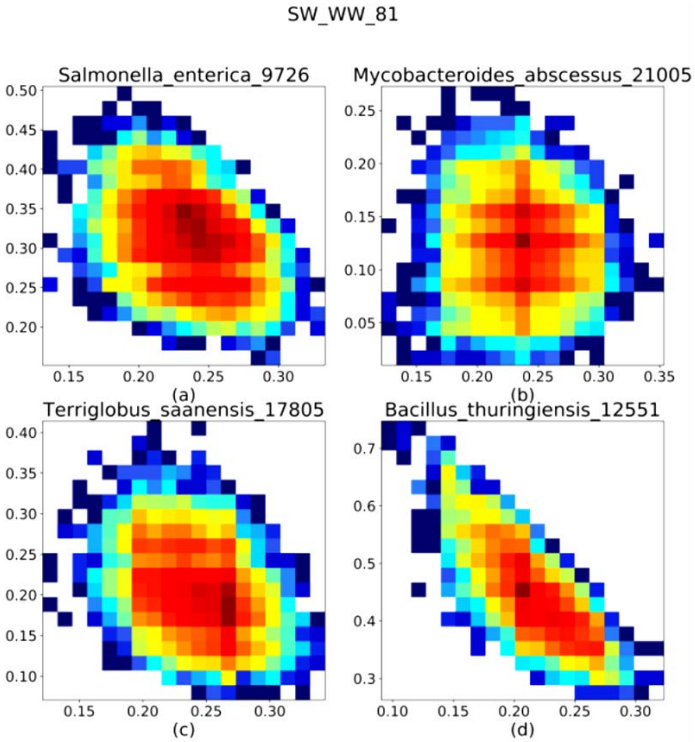

Figure 11. Result of 81 Length-Shift
图 11. 移位长度 81 结果图

观察图示可发现，分类结果数据集的变化并不大，虽然数据分布情况仍有变动，但是数据集数据整体分布情况变化不大。

综上，将图 6 作为对比项，较之图 7 至图 11 可视结果中的差异数据统计占比如“表 1”：

Table 1. System resulting data of standard experiment
表 1. 标准试验系统结果数据

mlen = 1 (图 7)	mlen = 21 (图 8)	mlen = 41 (图 9)	mlen = 61 (图 10)	mlen = 81 (图 11)
12.42%	14.04%	15.30%	14.21%	14.43%

5.2. 结果分析

综上所有的结果图示，虽然在移动长度为 41~61 的范围内进行移动操作后的序列与不进行移动操作的序列两者之间的相似度最低即差异性最大，但在该区间范围内分类结果数据集的每个种类分类的结果统计量变化不大，且单一种类的结果数据集的数据分类情况整体稳定，具有固有的分布特征。虽然本文处只展示了‘SW’和‘WW’一种组合的图示结果，但是其他组合结果图示在同样的操作流程下也是可得的，并且所有图示所反映结果信息的也是一致的。

RNN 分类器在序列元素位置变化的情况下，分类操作时仍然是稳定的，序列数据中元素的位置改变在一定范围内对 RNN 分类器的影响是十分有限的。

6. 总结

本文利用变值测量模型对 RNN 分类器的稳定性进行了初步的分析与探索，其处理的流程主要包括：

序列操作模块、RNN 分类器分类模块、统计计算模块、结果可视化模块。其中多种组合的图示结果提供了较为全面的观察与分析角度，从数据分布特征对 RNN 分类器的分类稳定性研究提供了一种可行的处理流程和方法，为观察 RNN 分类器结果的稳定性提供了系列可视的比较结果。

基于目前的研究结果，整个框架仍有可拓展的部分，虽然可视化结果中展示出了 RNN 分类器的稳定，但是其中仍有细微的差异，可将差异作为独立的可视化单元进行进一步的分析与探索。

致　谢

感谢云南大学软件学院，感谢云南省软件工程重点实验室提供良好的工作环境。感谢云南省科技计划项目(KC1810123)对该项目提供的资金支持。

参考文献

[1] Salehinejad, H., Sankar, S., Barfett, J., Colak, E. and Valaee, S. (2018) Recent Advances in Recurrent Neural Networks. https://arxiv.org/pdf/1801.01078.pdf

[2] 杨丽, 吴雨茜, 王俊丽, 刘义理. 循环神经网络研究综述[J]. 计算机应用, 2018, 38(S2): 1-6

[3] Aungiers, J. (2018) Tim Series Prediction Using LSTM Deep Neural Networks. https://www.altumintelligence.com/articles/a/Time-Series-Prediction-Using-LSTM-Deep-Neural-Networks

[4] 黄有为, 高燕. 基于循环神经网络的金融数据预测系统[J]. 软件导刊, 2019, 18(1): 28-33, 226.

[5] 韩丰, 龙明盛, 李月安, 等. 循环神经网络在雷达临近预报中的应用[J]. 应用气象学报, 2019, 30(1): 61-69.

[6] 唐美丽, 胡琼, 马廷淮. 基于循环神经网络的语音识别研究[J]. 现代电子技术, 2019, 42(14): 152-156.

[7] 郑智捷, 郑昊航. 变值测量结构及其可视化统计分布[J]. 光子学报, 2011, 40(9): 1397-1404.

[8] Zheng, J. (2011) Conditional Probability Statistical Distributions in Variant Measurement Simulations. Acta Photonica Sinica, 40, 1662-1666. https://doi.org/10.3788/gzxb20114011.1662

[9] Zheng, J. (2019) Novel Pseudorandom Number Generation Using Variant Logic Framework. Variant Construction from Theoretical Foundation to Applications, Springer, Singapore, 289-295. https://doi.org/10.1007/978-981-13-2282-2_18 .

[10] Zheng, J. (2019) Variant Logic Construction under Permutation and Complementary Operations on Binary Logic. In: Zheng, J., Ed., Variant Construction from Theoretical Foundation to Applications, Springer, Singapore, 3-21. https://doi.org/10.1007/978-981-13-2282-2_1

[11] Levenshtein, V. (1966) Binary Codes Capable of Correcting Deletions, Insertions, and Reversals. Soviet Physics Doklady, 10, 707-710.

[12] 赵作鹏, 尹志民, 王潜平, 等. 一种改进的编辑距离算法及其在数据处理中的应用[J]. 计算机应用, 2009, 29(2): 424-426.