Histone Modification Data Compression for Fast Classification of Gene Expression using Temporal Convolution Network

Imam Mustafa Kamal[1]
Imamkamal52@gmail.com

Natanael Yabes Wirawan[1]
yabes.wirawan@gmail.com

Nur Ahmad Wahid[1]
nurahmadwhd@gmail.com

[1]Big Data Department,
Business Service Computing Laboratory
Pusan National University

1. Problems

Histones were one of the fundamentals parts to regulate genes. When modified, it can change the spatial arrangement of the DNA which allows or restricts the binding of different proteins to the DNA and finally leads to different form of gene regulation [1].

Histone modifications are epigenetic which means that any changes to the histone are reversible. DNA mutations however, are not epigenetic since the effect of such modifications is irreversible. This leads to the importance of histone modifications role in developing drugs for cancer treatment [1].

DeepChrome is the first deep-learning implementation for predicting gene expression from histone modification [2]. The idea is that given a matrix of histone modification data represented as an image, predict the gene expression using Convolutional Neural Network (CNN). This method outperforms SVM and Random Forest altogether for 56 prediction tasks of REMC data [3].

However DeepChrome search space is rather large since it predicts histone modifications data into X^N i.e. 100^5, where X represents the number of row for each gene ID, and N represents the number of histone modifications which is equals to five.

Therefore, we propose a method to convert the histone modification data into high dimensional features such that each gene has 500 features instead of 5 features with 100 rows. Afterward, using the generated one-dimensional data, we would like to incorporate Temporal Convolution Network (TCN) to predict the gene expressions for each gene.

Our method outperforms the state-of-the-art method i.e. Convolution Neural Network (CNN) in terms of Area under Curve (AUC) score. Moreover, our method has linear search space i.e. only 1^{500}.

2. Related Works

The related works of this paper is organized as follows: first we would like to explain about the DeepChrome algorithm. Then we would like to explain about the dimensionality reduction as a way to reduce the search space of input layer. Lastly, we would like to explain about Temporal Convolution Network and why it is suitable for this case study.

2.1. DeepChrome

DeepChrome utilizes a deep convolutional neural network model to predict gene expression given histone modifications data i.e $N = 5$ [2]. The network was able to learn both the combinatorial interactions and the classifier in unified manner. Moreover, it also introduces a visualization technique to extract interactions to make the model interpretable. The architecture of DeepChrome can be summarized in **Figure 1**.

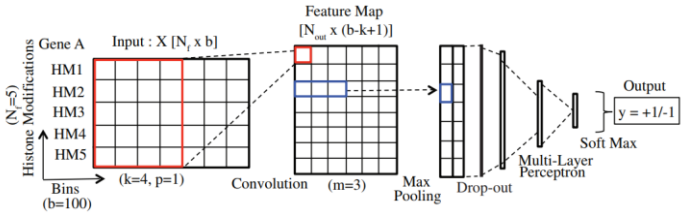

Figure 1 DeepChrome CNN Architecture [6]

2.2.Dimensionality Reduction

Dimensionality Reduction is a technique to reduce overlap features by obtaining a set of principle variables [4]. It divided into two key components i.e. **1) feature selection** to find a subset of the original set of variables or features, to get a smaller subset which involves filtering, wrapping, and embedding the variables and **2) feature extraction** to reduce the data in high dimensional space to a lower dimension space i.e. space with lesser number of dimension. Our approach is using the combination of feature selection and feature extraction i.e. by embedding histone modification data for each 100 rows of gene ID into single row of data, and thus converts the data into higher dimensional space.

2.3. Temporal Convolution Network

Temporal Convolution Network (TCN) is a new family of Convolutional Neural Network (CNN) architecture where the convolutions in the architecture are causal and can take sequence of any length and map it to an output sequence of the same length [5].

TCN leverages dilated causal convolution and residual block which enable an exponentially large receptive field and allows layers to learn modifications to the identity mapping rather than the entire transformation respectively [6].

Moreover, a TCN can process a long input sequence as a whole thus it has low memory requirement for training. It also can change its receptive fields in multiple ways e.g. stacking more dilated convolutional layers (which were used in this work), using larger dilation factor, or increasing the filter size, thus afford better control against model's memory size. TCN also known to have more stable gradients compared with other recurrent architectures. Lastly, it was able to cope variable inputs with arbitrary length by sliding the 1D convolutional kernels, thus can be adopted as drop-in replacements for RNN for any sequential data of any arbitrary length. **Figure 2** summarized the architecture of TCN.

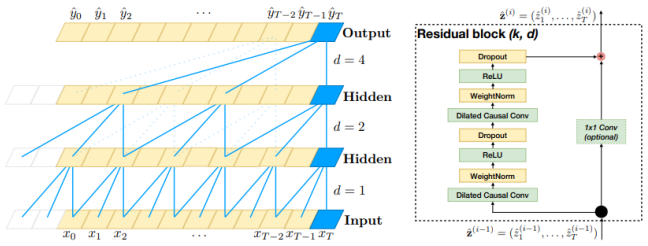

Figure 2 Architecture of Temporal Convolutional Network (TCN) [6]

3. Proposed Idea

The initial idea is begun with the research [2]. By harnessing the deep learning approach, this research provides better result compared with previous research. However, the existing research represents the data as two-dimensional data as shown in figure 1. In this study, we proposed to comprise the data into merely one-dimensional data with regarding minimize the processing time in doing classification task.

Table 1 The multi-dimensional data of gene expression using DeepChrome

Row	GeneID	HM1	HM2	HM3	HM4	HM5	Label
1	1	0	7	0	5	2	1 (on)
2	1	0	0	1	8	0	1 (on)
…	…	…	…	…	…	…	…
100	1	1	2	0	0	0	1 (on)
1	2	0	1	4	0	0	0 (off)
2	2	0	1	8	1	3	0 (off)
…	…	…	…	…	…	…	…
100	2	4	0	1	1	0	0 (off)

Table 2 The one-dimensional data of gene expression in our approach

Row	GeneID	HM1	HM2	HM3	HM4	HM5	Label
1	1	1	20	1	7	2	1 (on)
1	2	20	1	4	0	0	0 (off)

Furthermore, we incorporate Temporal Convolutional Network (TCN) [9] to perform the classification task, and finally, compare the result with the state-of the-art-method. **Figure 3** summarized the TCN architecture of our method.

4. Evaluation

Generally, binary prediction tasks should be resulted in [0, 1] value to predict whether a set of features is belong to a certain class identified by those [0, 1] values. However, typical classification models give more than just binary label i.e. float values which can be used to measure the confidence whether a set of features tends to belong into 0 or 1 class. Since it is important to calculate confidence of the prediction result, thus the evaluation metric for gene expression prediction tasks from histone modification dataset is based on Area under Curve (AUC) score from Receiver Operating Characteristic (ROC) curve [7] under the following equation:

$$A = \int_{\infty}^{-\infty} TPR(T)FPR'(T)\,dt$$

Where:

A : Area under the ROC curve
$TPR(T)$: True positive rate given threshold T
$FPR(T)$: False positive rate given threshold T

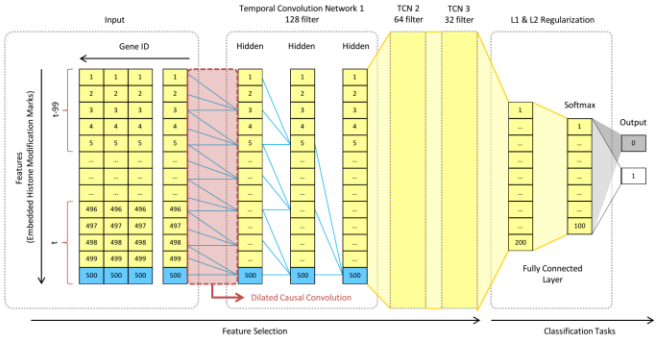

Figure 3 Our Temporal Convolutional Network Architecture

Furthermore, we also calculate evaluate the experiment in terms of accuracy, precision, recall, f-score, and specificity against REMC dataset [2] under the following equations.

To compute accuracy:

$$accuracy = \frac{TP + TN}{TP + FN + FP + TN}$$

To compute precision:

$$precision = \frac{TP}{TP + FP}$$

To compute recall:

$$recall = \frac{TP}{TP + FN}$$

To compute f-score:

$$f - score = 2 \times \frac{recall \times precision}{recall + precision}$$

And finally, to compute specificity:

$$specificity = \frac{TP}{TP + FN}$$

The experiment was carried using Tensorflow and Keras as our deep learning framework with Intel Core i7 4790K CPU, 32 GB RAM, and NVDIA GeForce GTX 1080Ti GPU setup.

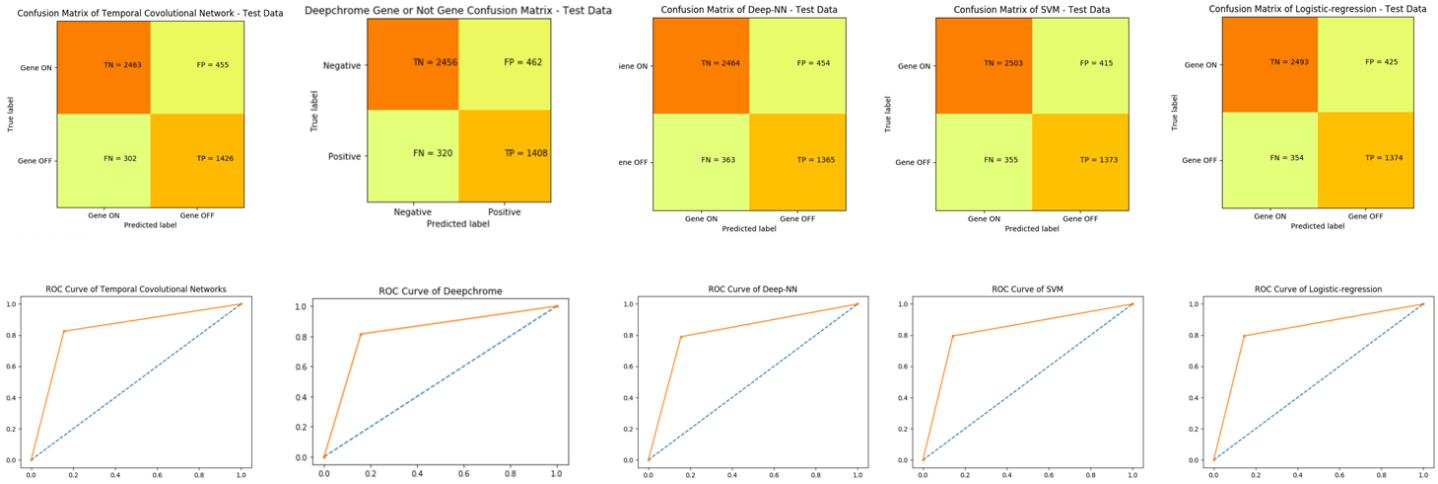

Figure 4 Confusion Matrix and ROC Curve for all experiments against various machine learning models

5. Experiment Results

Although AUC is sufficient to measure the performance for each machine learning model in this problem, we also conduct experiment under accuracy, precision, recall, f-score, and specificity.

The TCN outperforms the state-of-the-art-method in terms of AUC, accuracy, recall, and f-score, while Deep NN fails miserably in all metrics because in tends to be over-fitting. Interestingly enough, the SVM outperforms TCN in terms of precision and specificity metrics.

The confusion matrix, AUC, accuracy, f-score, precision, recall, and specificity score can be summarized as seen in **Table 3**, **Figure 4** and **Figure 5** respectively.

Table 3 Experiment results for all machine learning models

Methods	AUC	Accuracy	Precision	Recall	F-score	Specificity
TCN	0.835	0.837064141	0.758107	0.825231	0.790247	0.8440713
DeepChrome	0.828	0.831683168	0.752941	0.814815	0.782657	0.8416724
DeepNN	0.817	0.824149806	0.750412	0.789931	0.769665	0.8444140
SVM	0.826	0.834266035	0.767897	0.794560	0.781001	0.8577793
Logistic Regression	0.825	0.832328885	0.763758	0.795139	0.779132	0.8543523

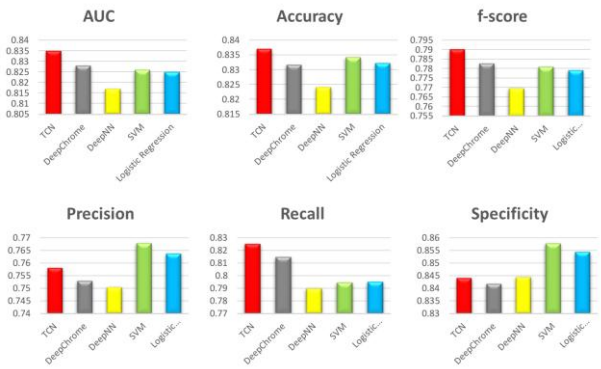

Figure 5 Experiment results against all machine learning models evaluated under AUC, accuracy, f-score, precision, recall, and specificity metrics

6. Conclusion

Our method outperforms the state-of-the-art method in terms of AUC, accuracy, recall, and f-score. In the future we would like to use different set of data to measure the robustness of the proposed method.

7. References

[1]. Bannister, A. J. and Kouzarides, T. (2011). Regulation of chromatin by histone modifications. Cell Research, 21(3), 381– 395.

[2]. Singh, R., Lanchantin, J., Robins, G., & Qi, Y. (2016). DeepChrome: deep-learning for predicting gene expression from histone modifications. Bioinformatics, 32(17), i639-i648.

[3]. Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z.,Wang, J., Ziller, M. J., et al. (2015). Integrative analysis of 111 reference human epigenomes. Nature, 518(7539), 317–330.

[4]. Roweis, S. T., & Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. science, 290(5500), 2323-2326.

[5]. Razavian, N., & Sontag, D. (2015). Temporal convolutional neural networks for diagnosis from lab tests. arXiv preprint arXiv:1511.07938.

[6]. Bai, S., Kolter, J. Z., & Koltun, V. (2018). An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. arXiv preprint arXiv:1803.01271.

[7]. Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology, 143(1), 29-36.