A WaveNet based Ion Channel prediction Method

Yukun Teng¹, Deheng Chen¹, Yichun Zhou ², Nathaniel M. Vegh ³ and Ren Zhang ⁴

¹China University of Mining & Technology, Beijing, 100000, China
¹Wuhan University, Wuhan, Hubei, 442000, China
²New York University, New York, 10041 NY212, USA
³Pangea International Academy, Chicago, 60601, USA
⁴Sichuan University, Chengdu, Sichuan, 610000, China
⁵Corresponding author’s e-mail: 2017301040065@whu.edu.cn

Abstract. There are many diseases associated with the dysfunction of ion channels. Ion channels are a class of proteins embedded in cell membranes that serve principle physiological functions such as memorization, learning and pain signaling. In order to help scientists learn more about ion channels and how they cause diseases, a team in Liverpool University published a public dataset about Ion channels. This paper presents a Wavenet based model to predict the number of Ion channels open at each time point. Wavenet is a type of deep neural network designed for one-dimensional signals. We used the Kalman filter to denoise the original signals and compared WaveNet with classical machine learning methods like SVM and Naïve Bayes to show the superior performance of the WaveNet model.

1. Introduction

1.1. Summary and problem review

Ion channels are membrane proteins that provide passage for ions to cross the membrane along their electrochemical gradient. In 1999, Yoshihisa Kurachi proposed an article about Ion channels which shows that Ion channels are vital in the physiology of cells. The team discussed the structure and function of ion channels and illustrated that a rapidly growing number of animal and plant diseases are derived or resultant from these structures and channels. In recent years, researchers have turned to machine learning methods to predict or cluster the information gathered from Ion channels. In the work [7, 8], the authors have offered conclusions for some machine learning approaches as regards Ion channel prediction. For instance, the naïve bayes classifier can be used to train a system on electrical signals. It is an interesting case, since it shows the feasibility of using biological signals to predict the type of Ion channels. In the paper [15], LF Yuan and other authors did experiments on radial networks to predict the types of ion channels. In this paper, we propose a WaveNet [9] based intelligent algorithm to predict the number of Ion channels open at each time point. Deep learning (DL) is the part of a broader family of machine learning based methods based on artificial neural networks with representation learning. Machine Learning can be supervised, semi-supervised or unsupervised. It has a lot of applications like image recognition [3], email spam filtering [4] and biomedical deep learning [5]. To compare the difference in performance between our algorithm and classic machine learning algorithms, algorithms like support vector machine (SVM), logistic regression and naïve bayes are
tested in the experimental phase. For the classical machine learning algorithms, some data mining and feature engineering are done to enhance the final results. Whereas Wavenet can learn deep patterns from electrical signals so that little feature engineering is needed.

1.2. Related Work
Ion channels prediction is based on electrical signals in cells. There are two ways to solve this problem in machine learning. It is possible to use classical machine learning algorithms like the support vector machine (SVM). However, data mining and feature engineering are required to generate information from historical electrical data, such as using lag features or rolling features. The other method is using deep neural network such as WaveNet. WaveNet, which was proposed by Google, is a network used for generating raw audio waveforms. The model is probabilistic and autoregressive which can generate the predictive distribution for each audio sample. In this task, we used this model on the Ion channel signals since it is a kind of waveform. Besides, Kalman filters are applied to preprocess signals and remove noisy data [16]. In control theory, Kalman filtering, also known as linear quadratic estimation (LQE), uses a series of measurements of time, containing statistical noise, and produces an estimation of unknown signals. In this task, Kalman filtering is used to remove the noise existing in electrical signals, which enhances the final result.

2. Our Contribution

2.1. Article Structure
This paper proposes a WaveNet based algorithm to predict Ion channel numbers at each time point. The dataset released from Liverpool University contains millions of electrical sample points. In this paper, to deal with such a massive amount of data, the Kalman filtering algorithm is used to remove the drift noise from the original data. The first step is important otherwise the dirty data would ruin our models. The WaveNet model consists of WaveNet blocks which can learn the distribution from waveform to make better predictions. WaveNet is able to accurately model different voices with the accent and tone of the input correlating with the output. In this paper, it is utilized to learn a different kind of signal, which is those from the Ion channels. To demonstrate the improved performance of our Wavenet model, we also implemented other classical models like support vector machine, logistic regression, and Naïve bayes. The final experiments showed the WaveNet based model outperforms all the other models on the F1 score.

The remainder of this paper is organized as follows. The second section mainly discusses data descriptions and data filtering methods. Section III presents the WaveNet based deep learning model and the structure of this model. In section IV, the experiments based on metric of accuracy and F1 score are conducted. Finally, In Section V, we draw a conclusion for this paper.

2.2. Feature Engineering
The Ion Channel dataset consists of two sets: the training set and test set. The training set has two million electrical sample points. The test set contains 500,000 sample points. Figure 1 shows the waveform of signals and the related ground-truth data.

![Figure 1. The waveform of electrical signals](image-url)
Figure 2. The ground truth of open channels

The first graph shows the wave form of electrical signals and the second graph shows the ground truth of open channels. From the two figures, it is easy to find out that there are some relationships between ground truth and electrical signals. To improve prediction accuracy, it is essential to do data cleaning. Kalman filtering methods are carried out to process the training set and test set. For classical machine learning algorithms, feature engineering is needed, which is covered in the Table 1.

| Features            | Feature Description            |
|---------------------|--------------------------------|
| Gradient features   | Gradient of signals            |
| Rolling features    | Rolling features of signals    |
| Percentage features | Percentage signals in a period |
| Ewm features        | Ewm of signals                 |
| Extreme features    | Max, min features of signals   |
| High pass features  | High pass signals              |
| Low pass features   | Low pass signals               |
| Shifted features    | Lagging and leading features of signals |

These features above like gradient features, rolling features and percentage features are used for classical methods.

3. Wavenet Based Ion Channels Prediction
In this section, we will describe the model’s structure and parameters. Figure 3 is an illustration of the model’s architecture.

Figure 3. The WaveNet Structure
3.1. WaveNet Structure Characteristic
The WaveNet consists of several WaveNet blocks followed by fully connected layers. The block contains Dilated convolutional layers that can gain information in a short time period or long time period. Fully connected layers are used as the multi-class classifier for this task. Details are presented in Table 2.

We used grid search to fine tune the parameters of our models in a certain range. For practical purposes, it is a useful strategy to improve the final result. Fine tuning the parameters improved the F1 score from 0.938 to 0.940. In particular, grid search is acting on some key parameters like n_estimators (number of estimators), num_leaves (number of leaves) and boosting_type (gbdt boosting type) etc. By setting the learning rate to a small value, the learning curve becomes more smooth, and num_leaves decides the basic decision tree complexity. In general, a more complicated tree has a better learning ability. For this task, the leaves number is set to 280. More details can be found in Table 3. Parameters not shown in Table 3 are set to default values.

| WaveNet | Filter Number | Dilation Rate |
|---------|---------------|---------------|
| Block1  | 16            | 12            |
| Block2  | 32            | 8             |
| Block3  | 64            | 4             |
| Block4  | 128           | 1             |
| Dense6  | Unit Number   | -             |

The table above shows the parameters and structures of WaveNet. It is easy to find out the model contains four blocks, each one has a different filter number. As the layers go deeper, the filter number increases to learn deeper features. The dilation rate stands for the sample range, a higher value means a bigger scope. Normally, the former layers have a bigger scope to capture information from electrical signals. Finally, fully connected layers are used as the classifier. It has 11 units since the maximum number of Ion channels is 11. In this layer, it is can be activated with a softmax function and the unit with the maximum probability is recognized as the prediction result. Since it is a classification task, the loss is set to categorical cross entropy.

3.2. Experiment On Our Model
In this section, the details of training the deep neural network on Liverpool Ion Switching dataset are shown. The model is trained on a workstation with a NVIDIA RTX 2080 Ti graphics card and Intel i7 processors. The GPU can accelerate the training of the WaveNet model. To train a WaveNet from scratch, it takes about 2 hours to finish training on our workstation.

To show the performance of different methods, the metrics like micro F1 and accuracy are used. In table 3, it shows the micro F1 and accuracy of the three classical models and WaveNet.

| Models         | Micro F1 | Accuracy |
|----------------|----------|----------|
| Logistic Regression | 0.862    | 0.926    |
| SVM            | 0.927    | 0.952    |
| Naïve bayes    | 0.902    | 0.938    |
| WaveNet       | 0.940    | 0.970    |
As it is shown in table 3, it is easy to find out that WaveNet has the best micro F1 score and accuracy, which means the proposed model outperforms all other classical methods without doing feature engineering.

The Figure 3 shows the metrics the F1 score and accuracy in different epoch. It is clear that the model converge very quickly, it takes about 20 epochs for model to get a good score.

![Figure 4. F1 score and accuracy for different epoch](image)

4. Conclusion
In this paper, we present a WaveNet-based algorithm to predict the number of ion channels open at each time point. It can learn the relationships between historical electrical signals without feature engineering. In section I, we mainly introduced ion channels’ role in disease and some related work in this area. In section II, the wave form of signals, the Kalman filtering and the related feature engineering is covered. In section III, we discussed the structure and parameters of our WaveNet model. The performances of all four models are listed in table 3, and figure 3 shows the F1 score and accuracy per epoch. Due to the similarity of audio data and our data about ion channel, wavenet's high performance remains.

Acknowledgement
The author Yukun Teng & Deheng Chen thanks Yichun Zhou for his interesting ideas on WaveNet. It is also nice for Nathaniel M. Vegh to do feature engineering and data cleaning to optimize the final score of the WaveNet and other classical models. Finally, DD offered help in experiments and training models.

References
[1] Ashcroft, F. M. (1999). Ion channels and disease. Academic press.
[2] Kurachi, Y., & North, A. (2004). Ion channels: their structure, function and control—an overview. The Journal of Physiology, 554(2), 245-247.
[3] Wu, M., & Chen, L.. (2015). Image recognition based on deep learning. Chinese Automation Congress. IEEE.
[4] Al-Jarrah, O., Khater, I., & Al-Duwairi, B. (2012). Identifying potentially useful email header features for email spam filtering. In The sixth international conference on digital society (ICDS) (Vol. 30, p. 140).
[5] Ramos-Pollán, R., Guevara-López, M. Á., & Oliveira, E. (2012). A software framework for building biomedical machine learning classifiers through grid computing resources. Journal of medical systems, 36(4), 2245-2257.
[6] Lin, H., & Chen, W. (2015). Briefing in application of machine learning methods in Ion channel prediction. The Scientific World Journal, 2015.

[7] Lin, H., & Ding, H. (2011). Predicting ion channels and their types by the dipeptide mode of pseudo amino acid composition. Journal of theoretical biology, 269(1), 64-69.

[8] Murphy, K. P. (2006). Naive bayes classifiers. University of British Columbia, 18, 60.

[9] Oord, A. V. D., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., ... & Kavukcuoglu, K. (2016). Wavenet: A generative model for raw audio. arXiv preprint arXiv:1609.03499.

[10] Suykens, J. A., & Vandewalle, J. (1999). Least squares support vector machine classifiers. Neural processing letters, 9(3), 293-300.

[11] Tolles, J., & Meurer, W. J. (2016). Logistic regression: relating patient characteristics to outcomes. Jama, 316(5), 533-534.

[12] https://www.kaggle.com/c/liverpool-ion-switching/overview

[13] Welch, G., & Bishop, G. (1995). An introduction to the Kalman filter.

[14] Chen, X. W., & Jeong, J. C. (2007, December). Enhanced recursive feature elimination. In Sixth International Conference on Machine Learning and Applications (ICMLA 2007) (pp. 429-435). IEEE.

[15] Yuan, Lu-Feng, et al. "Prediction of the types of ion channel-targeted conotoxins based on radial basis function network." Toxicology in Vitro 27.2 (2013): 852-856.

[16] Welch, Greg, and Gary Bishop. "An introduction to the Kalman filter." (1995): 41-95.