Neuron classification with a data-driven workflow

Longfei Li1*

1College of Computer Engineering and Science, Shanghai University, Shanghai, Shanghai, 200000, China
*Corresponding author’s e-mail: lilongfei@shu.edu.cn

Abstract. The shape feature of a neuron reveals the behavior of a neuron cell. Usually, a set of features of neuron anatomy can define the morphological characteristics of a neuron for comparison or analysis of a group of neuron cells. At present, the best set of neuron morphological description characteristics cannot be determined in the field of brain science research. Based on a data-driven strategy, this paper establishes a workflow from neuron feature processing to morphological classification.

1. Introduction
The problem of neuron classification is still inconclusive in more than 100 years of controversy. Generally speaking, the classification of neurons is mainly based on a set of common features in a group of neurons [1]. But even for GABA neurons in the cerebral cortex, the consensus category information and the best characteristics that define them are still unknown [2]. As the amount of morphological data increases, the difficulty of classification gradually increases. At present, it is particularly important to establish standardized neuron category information and feature categories. Both the European Human Brain Project and the American BRAIN Initiative have established the task of cell classification as the primary task [3], [4]. In the era of big data where the number of neurons is rapidly increasing, the higher expert knowledge requirements of artificial classification methods are no longer suitable for solving this classification problem. In contrast, the automated neuron classification process with wide applicability and high data throughput is obviously more advantageous. Based on the perspective of neuron morphological characteristics, this paper constructs a data-driven workflow for feature extraction and machine learning classification, then evaluates the accuracy of applying this process to systematically classify 2000 neurons from NeuroMorpho.org [5].

2. Method

2.1. Definition
For a neuron classification data set $D$ with $n$ samples, each sample $d$ can be described as a combination of observations of $k$ morphological variables, then through the feature selection step, $m$ selected feature variables can be regarded as the neuron sample Effective description.

$$d \rightarrow (f_1, \ldots, f_k) \rightarrow (f_1, \ldots, f_m)$$ (1)

For each sample, the class label $l$ can be output by the classifier $C$.

$$C: d_1, \ldots, d_n \rightarrow l_1, \ldots, l_k$$ (2)

In the supervised classification task, the feature combination and category information of the neuron are known, and the classifier $C$ establishes the corresponding relationship between the feature
combination and the category information. According to this corresponding relationship, the predicted category of each test neuron can be output based on the characteristic information of each sample. In the unsupervised classification method, the algorithm directly divides neurons into interpretable $k$ clusters based on the characteristic information in the neuron data set.

2.2. Data preparation

Pyramidal neurons have a complex morphology and are named because they all shaped like cones. They are main projection neurons of brain, thus it has been widely selected in various studies. We obtained a total of 2001 pyramidal neurons from hippocampus and neocortex in rat and mouse from NeuroMorpho.org for classification task. Some of these samples are shown in Figure 1, the blue part of a neuron represents its apical dendrites, and the purple part represents the basal dendrites. It can be seen from Figure 1 that these neurons are difficult to distinguish visually.

![Figure 1. Some samples of neuron dataset.](image)

The number of neurons in each category is shown in Table 1, the four types of neurons have close sample numbers.

|            | Mouse hippocampus | Rat hippocampus | Mouse neocortex | Rat neocortex |
|------------|-------------------|----------------|-----------------|---------------|
| Count      | 458               | 483            | 549             | 511           |

As a feature measurement tool, L-Measure [6] is widely used in neuroscience research. It can calculate more than 40 morphological feature values in neurons, including cell body area and branch number, etc. Among these features, the maximum, minimum, standard deviation of local feature such as length can also be measured. The numerical calculation of neuron features performed in this experiment was obtained using version 5.3 of the L-Measure tool. Table 2 shows the 5 feature informations of all L-measure features. For the global features like $N_{\text{stems}}$, the feature value is used directly. However, for the features of locality like Length, only the average value of every compartment is selected as a morphological parameter.

| Description          |                          |
|----------------------|--------------------------|
| $N_{\text{stems}}$   | Number of stems attached to the soma |
| $N_{\text{bifs}}$    | Number of bifurcations in the neuron |
| $N_{\text{branch}}$  | Number of branches in the neuron |
| Volume               | Volume of a compartment in neuron |
| Length               | Length of a compartment in neuron |
For the outliers or missing values obtained after processing by the L-measure tool, the average value of this characteristic attribute is used for replacement. Finally, each feature value is normalized between 0 and 1.

2.3. Classification algorithms
In this paper, the following supervised classification algorithms are selected for neuron classification, Linear discriminant analysis (LDA) [7], random forest classifier [8], decision tree [9], Support Vector Classifier (SVC) with linear kernels (SVC-linear) and radial basis functions kernels (SVC-RBF) [10], Gaussian Naive Bayes (GNB) [11]. We use the above algorithm to carry out 10 times 3-fold classification experiments, and the training and test set division of each experiment is carried out randomly. In each trial, the training set was used to fit the classifier, and the classification results were output on the test set, and the average value of 10 classification assessment was counted as the final effect evaluation. The evaluation indicators like Precision, Recall, Accuracy, and F1-score are selected as the comprehensive evaluation of the classification effect. For the prediction results of each sample, according to the combination of the original category label and the predicted label of the sample, the number of four classification results can be divided into four sets: True Positive (TP), False Positive (FP), True Negative (TN), False Negative (FN), based on these results, the calculation of the 4 indicators is as follows:

\[
\text{Precision} = \frac{TP}{TP + FP} \tag{3}
\]
\[
\text{Recall} = \frac{TP}{TP + FN} \tag{4}
\]
\[
\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}
\]
\[
F1 \text{ score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{6}
\]

3. Results
For 2001 neuron data to be classified, the classification results of the six classification algorithms are shown in Figure 2, each bar represents a classification indicator. The Random forest algorithm has achieved the best results in various indicators.
The specific classification indicators are shown in Table 3. From the classification indicators, it can be seen that the algorithm with the best classification effect on the data set is Random forest, which has achieved a F1-score of 0.924. This is derived from the advantages of ensemble learning and the integration of multiple models to make predictions, the neuron feature space is divided more reasonable. The LDA algorithm and SVC models with two different type of kernel functions have achieved similar results (F1-score: 0.849, 0.849, 0.770). The worst classification effect is predicted by the Naive Bayes algorithm, which only achieves accuracy of 0.547, precision of 0.562, recall of 0.541 and F1-score of 0.526. This may be because the Bayesian model assumes that features are independent of each other, but the correlation between neuron characteristics is relatively high. For example, the number of branches and the number of bifurcation points, the total area and the total length are often related in a neuron tree.

| Table 3. Detailed classified metrics. |
|--------------------------------------|
|                                      |
| Accuracy | Precision | Recall | F1-score |
| LDA      | 0.851     | 0.860  | 0.847    | 0.849    |
| SVC-RBF  | 0.854     | 0.856  | 0.849    | 0.849    |
| SVC-Linear | 0.777   | 0.791  | 0.771    | 0.770    |
| GNB      | 0.547     | 0.562  | 0.541    | 0.526    |
| Decision tree | 0.878 | 0.877  | 0.876    | 0.876    |
| Random forest | 0.925  | 0.925  | 0.924    | 0.924    |

4. Conclusion
As an important research basis for exploring neuron morphological characteristics, neuron classification task is still very challenging in the research community. The current continuous increase in the amount of data has given more space for the application of data-driven machine learning algorithms. In this work, machine learning algorithms are applied to neuron feature data sets for classification experiments. The experimental results show that machine learning methods can effectively help solve the problem of neuron classification. Among these algorithms, the random forest algorithm shows good task adaptability. In fact, its decision-making process is also the process of defining neuron category characteristics. On the one hand, through a powerful model, the results of classification experiments help researchers find accurate and objective descriptions of neuron morphology. On the other hand, the precise morphological definition also exposes the morphological regularity of neurons and accelerates the analysis process of neuron behavior.

In the complete neuroscience workflow, neuron classification is an intermediate step that connects the reconstruction of neural cells and the generation of virtual neurons using relevant feature parameters. Since most of the current classification work still focuses on the morphological characteristics of neurons, a broader and more accurate feature description method can undoubtedly accelerate the development of classification research. At present, morphological description methods based on branch topology are gradually emerging. Future classification methods should consider introducing these description methods to improve the specificity of neuron description and help machine learning methods to better understand neuron data. Besides, with the increase of feature descriptions, the strategic steps of feature selection like PCA or mRMR [12] (Max-Relevance and Min-Redundancy) will also have important value in future generation methods.

As the data standards of various research laboratories are gradually unified, feature measurement models and tools are gradually developing, the task of neuron classification will undoubtedly be completely solved, and the ultimate research goal of simulating the human brain in the field of brain science will eventually be realized.
References

[1] R. Armañanzas and G. A. Ascoli, “Towards the automatic classification of neurons,” Trends in Neurosciences, vol. 38, no. 5, pp. 307–318, May 2015, doi: 10.1016/j.tins.2015.02.004.

[2] J. DeFelipe et al., “New insights into the classification and nomenclature of cortical GABAergic interneurons,” Nat Rev Neurosci, vol. 14, no. 3, pp. 202–216, Mar. 2013, doi: 10.1038/nrn3444.

[3] T. R. Insel, S. C. Landis, and F. S. Collins, “Research priorities. The NIH BRAIN Initiative,” Science, vol. 340, no. 6133, pp. 687–688, May 2013, doi: 10.1126/science.1239276.

[4] H. Markram, “The human brain project,” Sci Am, vol. 306, no. 6, pp. 50–55, Jun. 2012, doi: 10.1038/scientificamerican0612-50.

[5] G. A. Ascoli, D. E. Donohue, and M. Halavi, “NeuroMorpho.Org: A Central Resource for Neuronal Morphologies,” J. Neurosci., vol. 27, no. 35, pp. 9247–9251, Aug. 2007, doi: 10.1523/JNEUROSCI.2055-07.2007.

[6] R. Scorcioni, S. Polavaram, and G. A. Ascoli, “L-Measure: a web-accessible tool for the analysis, comparison and search of digital reconstructions of neuronal morphologies,” Nature Protocols, vol. 3, no. 5, pp. 866–876, May 2008, doi: 10.1038/nprot.2008.51.

[7] Y. Aliyari Ghassabeh, F. Rudzicz, and H. A. Moghaddam, “Fast incremental LDA feature extraction,” Pattern Recognition, vol. 48, no. 6, pp. 1999–2012, Jun. 2015, doi: 10.1016/j.patcog.2014.12.012.

[8] Tin Kam Ho, “The random subspace method for constructing decision forests,” IEEE Trans. Pattern Anal. Machine Intell., vol. 20, no. 8, pp. 832–844, Aug. 1998, doi: 10.1109/34.709601.

[9] G. V. Kass, “An Exploratory Technique for Investigating Large Quantities of Categorical Data,” Applied Statistics, vol. 29, no. 2, p. 119, 1980, doi: 10.2307/2986296.

[10] Y. Lee, Y. Lin, and G. Wahba, “Multicategory Support Vector Machines: Theory and Application to the Classification of Microarray Data and Satellite Radiance Data,” Journal of the American Statistical Association, vol. 99, no. 465, pp. 67–81, Mar. 2004, doi: 10.1198/016214504000000098.

[11] C. K. I. Williams and D. Barber, “Bayesian classification with Gaussian processes,” IEEE Trans. Pattern Anal. Machine Intell., vol. 20, no. 12, pp. 1342–1351, Dec. 1998, doi: 10.1109/34.735807.

[12] Hanchuan Peng, Fuhui Long, and C. Ding, “Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy,” IEEE Trans. Pattern Anal. Machine Intell., vol. 27, no. 8, pp. 1226–1238, Aug. 2005, doi: 10.1109/TPAMI.2005.159.