Mining Big Neuron Morphological Data

Maryam Aghili,Maria Presa, Jingan Qu

Abstract

In this research we explored a part of a raw dataset of 3D reconstructed images of neurons. We applied different feature selection and dimensionality reduction algorithm to increase the baseline accuracy of different classifiers. Multiple combination of pre-processing methods, single and multi label classifiers and different subset of instance space were tested on this dataset aiming to increase the classification accuracy. The result of different single label classifiers and multi label classifiers compared during this study and best algorithm with its performance is defined as the benchmark.

1 Introduction

Neuromorphology is a branch of science that studies neural systems form, shape, geometry, and structure. Scientists have proven that there is a relationship between form and shape of the neurons and their functionality and properties. Many experiments have been done to identify the different categories of the neurons based on their guise and function. [1] .It is believed that neurons morphology differs based on the different species, disparate regions in living body, cells function and evolutionary stages [2]. knowing the functionality of a neuron is an essential tool for diagnosis the root of neural networks disease for exploring the main cause of many disorders like emotional disorders, learning disabilities and also malfunction of some parts in the injured brain [3]. Due to the importance of the neuromorphology along with the burdensome task of manual classification, many attempts has been started to exploit computer power for automatic neurons classification. In this research, we tried to use multiple techniques of data pre-processing, feature selection, dimensionality reduction and classification to tackle the problem mentioned above.

2 Related Work

Neuromoph is a repository of almost all identified neurons samples from different laboratory around the world. This public dataset provides researchers of the different field rather than biology with the opportunity of accessing a valuable data. The original data, which is a collection of 3D reconstruction of neurons image with corresponding metadata are available via a website. The 3D nature of neurons image and difficulty of the pattern finding in the cluttered data expose a new challenge for the researchers. In recent years, many researchers focus on finding the patterns in this huge raw dataset and exploring new way for automatic classification of 3D neurons images. Da Fontoura in 2010, applied principal component analysis and canonical analysis to decrease the 20 features to a number that is possible to visualize in 2D spaces. PCA is a statistical method that transform a large set of features to a smaller collection in a way that remaining attributes represent the original variance, in other words, PCA gives the most important features of the entire set. [4]. By this transformation that allow visualization of the data, they discover some new facts about the data set including that while neurons only have a small proportion of its surrounding environment, only two principle component can represent half of the data variability and fifty percent of the total variance in original data. They also reported that only one region detected in morphological space has density peak. Their result shows that, cells with similar types, region and species tend to form cluster together and by applying PCA these clusters become more substantial. [5] following their research, Polavaram et al used L-measure to extract more than 100 features from the data and then they used an unsupervised clustering algorithm and principle component analysis to cluster the dataset. Their results show that there are some morphological differences between specific cell types and animal species which is not independent of the origin laboratory[6] Guerra et al moved toward supervised classification instead of unsupervised clustering in order to reap the benefits of prior knowledge in the field to distinguish neocortical pyramidal cells from interneurons. They compared their result of Decision Tree, Naive Base, Multilayer Perceptron, Logistic regression and KNN. The outcomes of hierarchical clustering algorithm show the superiority of supervised classification in morphological task. Additionally, they applied some dimensionality reduction techniques of principle component analysis and feature subset selection. They made a limited set of highest priority features and removed the groups of variables which lessened the accuracy. Their database consist of 128 pyramidal cells and 199 interneurons from mouse neocortex, sixty-four variable extracted and

Apical Dendrite used as the ground truth. [7] while many researchers have been working on this dataset, still there are lots of open challenge in this area which motivated us to leverage the neoromorphology scope by probing the variety of techniques of automatic classification on highly inconsistent datasets.

3 Methods

First 17123 3D reconstructed images of neurons extracted from the NeuroMorph website. using Lmeasure toolbox [8], 190 features extracted including bifurcation features, number of branches, width, depth and neuron compartment level. From the meta data corresponding to neurons images,6 label variables including Gender, Primary Brain Region, Second Brain Region, Primary Cell Class, Secondary Cell Class and Tertiary Cell selected as the main classes that should be predicted based on the image of a neuron. Remaining 190 features supposed to be used as labeled data.There were more than 190 features but as they had too many missing values they identified as not useful data at the first step and deleted from the dataset. Afterwards some pre-processing steps had done to extract the most useful data and exclude the redundant data which not only were not useful but in some cases also decreased the performance of the classifiers. Following in this section detailed information about pre-processing methods and the result of different classifiers on the data is provided.

3.1 Data Preparation

3.1.1 Marking Missing Values

In the label variables, there was a lot of classes named "Unknown" and "Notreported". It is reasonable to treat these kind of class values with a unique name to have an integrate schema, so all those values replaced with a unified name. in order to prevent loss of data we keep those samples that have some missing attributes.

3.1.2 Classes Removal

Considering there are a lot of classes in label variables with very low rate of occurrence, Dealing with these kinds of classes should be done before classification, otherwise, there won't be enough information from data to predict these kinds of classes. Here, a threshold of 20 is set up, which

means, the classes whose sample size is below 20, will be removed. After that, there are 12,288 samples left.

3.1.3 Making Up Missing Value

After some classes such as "Unknown", "Notreported", are marked as missing values, there are quite a few missing values in the data file, like Figure 1, about one fourth of record containing missing value. Ignoring all the records with missing value will lose a lot of information. So another strategy is used: Making up the missing values with predictions by the other variables (including the label variables). By the way, all the missing values to be made up should be put in the training data, otherwise, it should effect the accuracy of the estimation in the next section.

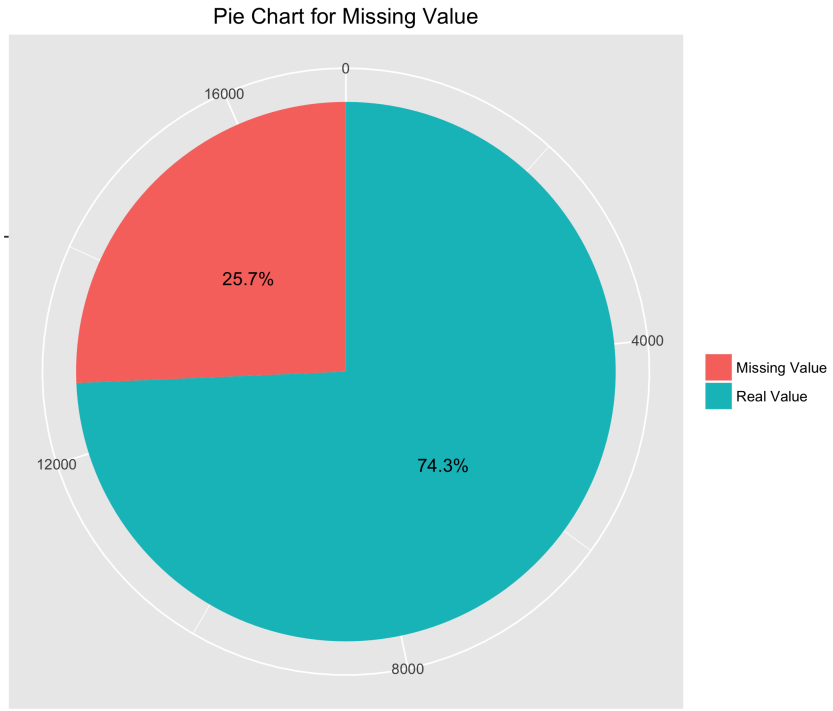

Figure 1: Pie Chart for Missing Values

Delgado et al in their study [9] have proven that Random Forest is the best classifier in most of the case so we applied this algorithm for making up missing values by prediction.

3.1.4 Features Removal

exploring the feature set, we discovered that a few of the attributes are constant in all the samples and there is no gain in keeping them so they were omitted and only 177 features left for further exploration. (there were 190 features originally), which means 13 features were removed.

3.2 Improvements

In this project, seven state of the art algorithms including KNN, logistic regression, decision tree, AdaBoost, Naive Bayes, Neural Network, Random Forest were used to build up the model for classification. Based on these seven algorithms, the label variable relationship and ensemble method are introduced.

3.2.1 Label Variables Relationship

There are 6 label variables to be predicted (classification), here, the chi-square test is used to test whether there is a significant relationship between any two label variables. table 1 are part of the results of chi-square tests, in which *** means the p-value is smaller than .001, indicating that there is a significant relationship between every two label variables. It suggests that when predicting one label variable, the information from the other label variables can be use for the prediction. The our strategy is that, when When one model to predict one label variable is being building, the other models to predict the other label variables are also included in this model.

label	Gender	PrimaryBrainRegion	SecondBrainRegion
Gender	0 ***	0 ***	0 ***
PrimaryBrainRegion	0 ***	0 ***	0 ***
SecondBrainRegion	0 ***	0 ***	0 ***
PrimaryCellClass	0 ***	0 ***	0 ***
SecondaryCellClass	0 ***	0 ***	0 ***
TertiaryCellClass	0 ***	0 ***	0 ***

Table 1: Part of the Results of Chi-Square Test

3.2.2 Ensemble Method

Taking one label variable Tertiary Cell Class as an example, logistic regression is used to draw a learning curve shown in Figure 2, from Figure 2, one

important information can be seen easily seen that this model is underfitting the data set. Meanwhile it is indicated that the training error and the test error are close when the training data points are around 3,000. So the training data is divided into 3 folds, and there are around 3,000 records in each fold. For each fold, KNN, logistic regression, decision tree, AdaBoost, Naive Bayes, Neural Network, Random Forest are applied separately to build 7 models to predict one label variable. Next, weighted ensemble method is used to combine these 7x3 models into one final model. By the way, the weights that is used in weighted ensemble method are the accuracy of each algorithm.

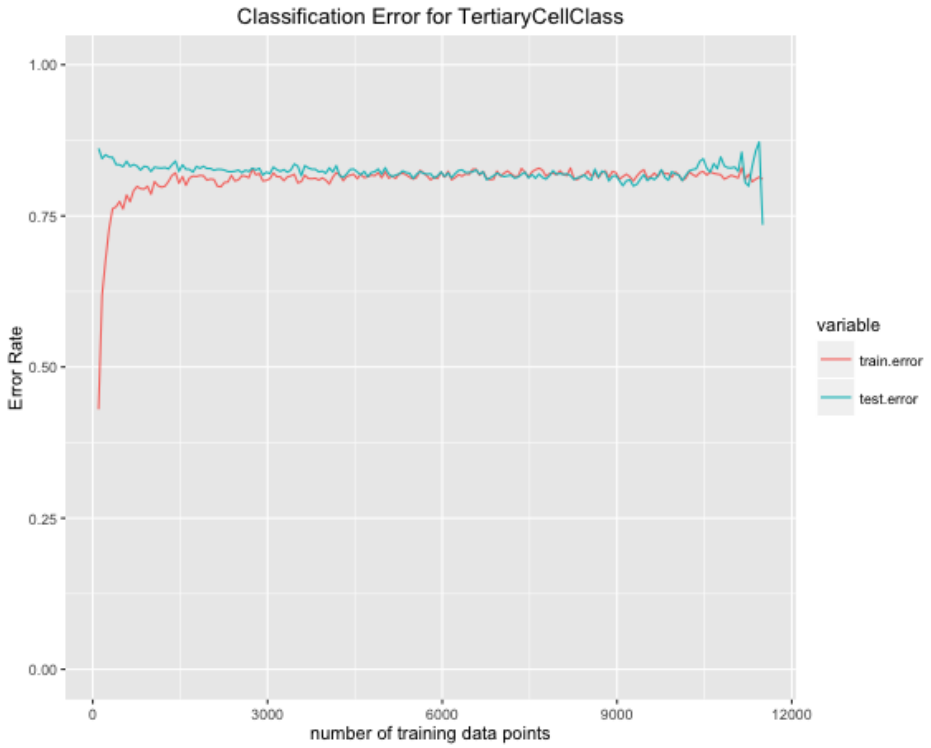

Figure 2: Learning Curve of Label Variable TertiaryCellClass

4 Experimental Results

4.1 Baseline Accuracy

Before the data preparation, KNN, logistic regression, decision tree, AdaBoost, Naive Bayes, Neural Network, Random Forest are applied separately to the original data set to do classification, and 3 folds cross validation is used the evaluate the accuracy. These results shown as Figure 3 are used as baseline accuracies to compare with the ones after the improvements.

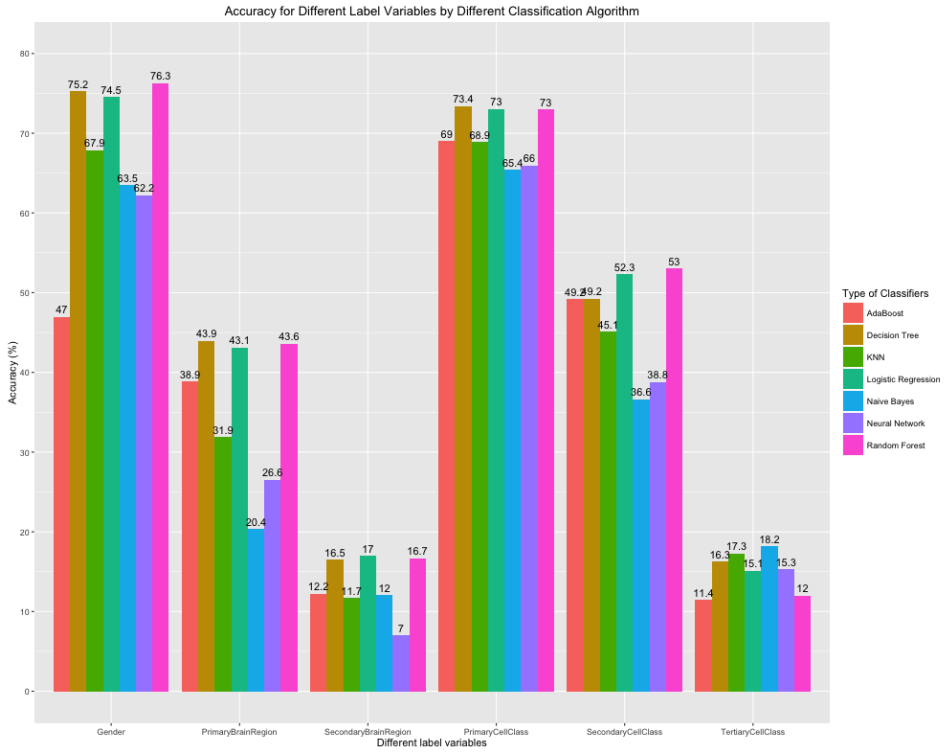

Figure 3: Baseline Accuracy for Label Variables by Different Algorithm

4.2 Accuracy Improvement

The data set is divided into two parts, two thirds are used as training data and one third are used as test data. A model is built as in the method part and then this model is used to predict the test data. Repeat this precess for each label variable. The results are as shown in Figure 4

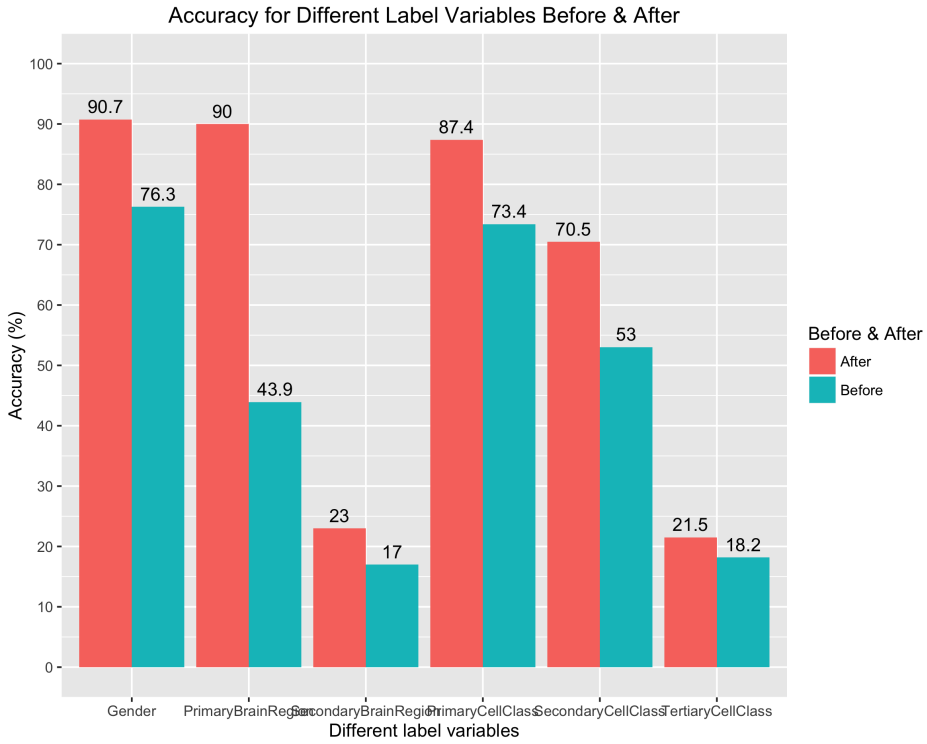

Figure 4: Accuracy for Different Label Variables Before & After

5 Discussion of Results

In Figure 4, it is indicated that the accuracy of PrimaryBrainRegion was improved the most.

There are some comparison between the accuracy of single and multi label classifiers. while it was supposed that combination of the label set should improve the accuracy of the classifiers, the results did not support that. The noisy data is the main reason of this contradiction and it is planned to study the data more in depth in order to extract noisy data and considering only valuable part to detect pattern on it.

References

[1] G. M. Villegas, "Electron microscopic study of the vertebrate retina," *The Journal of general physiology*, vol. 43, no. 6, pp. 15–43, 1960.

[2] R. Armañanzas and G. A. Ascoli, "Towards the automatic classification of neurons," *Trends in neurosciences*, vol. 38, no. 5, pp. 307–318, 2015.

[3] G. S. Benson, J. McConnell, L. I. Lipshultz, J. N. Corriere Jr, and J. Wood, "Neuromorphology and neuropharmacology of the human penis: an in vitro study," *Journal of Clinical investigation*, vol. 65, no. 2, p. 506, 1980.

[4] K. Person, "On lines and planes of closest fit to system of points in space. philiosophical magazine, 2, 559-572," 1901.

[5] L. d. F. Costa, K. Zawadzki, M. Miazaki, M. P. Viana, S. N. Taraskin *et al.*, "Unveiling the neuromorphological space," *Front Comput Neurosci*, vol. 4, p. 150, 2010.

[6] S. Polavaram, T. A. Gillette, R. Parekh, and G. A. Ascoli, "Statistical analysis and data mining of digital reconstructions of dendritic morphologies." *Frontiers in neuroanatomy*, vol. 8, 2014.

[7] L. Guerra, L. M. McGarry, V. Robles, C. Bielza, P. Larranaga, and R. Yuste, "Comparison between supervised and unsupervised classifications of neuronal cell types: a case study," *Developmental neurobiology*, vol. 71, no. 1, pp. 71–82, 2011.

[8] 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, 2008.

[9] M. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim, "Do we need hundreds of classifiers to solve real world classification problems?" *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 3133–3181, 2014.