# INFS7203 Cell Communication Functionality for Genes of E. coli

Tong Z. Author
School of Information Technology and Electrical Engineering
The University of Queensland, Qld., 4072, Australia

### Introduction

*This project aims to design and apply multiple data mining techniques to solve real-world problems.*

*The proposal will focus on the general design of the project on different steps, including pre-processing, model training and validation etc. It will also discuss the potential problems to be faced during the implementation of classification methods. Eventually, it gives estimated time management of the entire project.*

## 1 Data Pre-processing

Outlier detection should be applied primarily since the outlier data can significantly affect the statistical results of the training. [1] Considering the high-dimensional data provided, compute the mean and standard deviation of the data firstly, then perform outlier detection to remove the outlier data from the data set.

Imputation should be applied to the data as it shows that both the numerical and nominal data have missing values. For each feature, there could be more than one missing value. Hence, it might not be wise to simply choose the most common feature value as a replacement since it will lead to imbalanced results that a large proportion of instances will be assigned with the same feature value. Instead, when multiple nominal values are missing, a good strategy might be that missing values can be replaced according to the weight of each value type among all the instances.

If multiple missing values exist within numerical data, simply using the average feature value could possibly alter the overall distribution of values. Since it's hard to consider all possible data distributions, assumptions or testing about the data distribution can be helpful. [2] Imputation can be more precise when the distribution is known in prior. For instance, if the data follows the Gaussian distribution, multiple missing values can be assigned with values that follow the same distribution status to reduce the side effect of imputation.


*Fig 1. Weighted Imputation Diagram*

According to observations, the features have rather wide ranges of values. Hence, normalization methods like Min-Max and Z-Score can be applied in pre-processing. [3] For data that follows a Gaussian distribution, Z-Score might have a better performance.

However, the distribution of data and other details are unknown. To choose the optimal method among the optional pre-processing methods mentioned above, cross-validation can be used to get representative results. By dividing the data into k folders, each trial uses the k-1 folders data as the training data that applies the optional pre-processing methods above. The rest of 1 folder will be regarded as the validation set. Average results of k times training and validation, choose the pre-processing method with higher accuracy.

## 2 Classification Models

To apply decision tree classification to the data, purity measures like the Gini index, Gain ratio, Information gain can help select the optimal splitting feature. Using cross-validation, do calculations within the training set to obtain the splitting features that provide the highest Information gain and Gain ratio or the lowest Gini index. (For most of the continuous data, bi-partition can be used to construct splitting value.) Then, validate the chosen features in the validation data set and calculate the accuracy. After k times calculations, comparing the results and make the decision on which features to apply primarily. Eventually, construct the complete decision tree with the features selected.

Random Forest can help with building multiple decision trees instead of only one. K features are randomly selected out of the total d features and get grouped up. For each feature group, a decision tree will be constructed in the same way as building a classic decision tree, which has been mentioned above. The output of all classifiers will be combined

by majority voting to complete the predictions. In this way, the diversity between individual decision trees can be kept to a great extent. In terms of hyperparameter tuning, the value of k can be optimized by CV training with values within an interval. For instance, k within [a, b] where a > d/10 and b < d/2.

Moreover, the instance-based classification method k-NN can construct a table that records the distance between each data point. The distance measurement can be done with Minkowski distance. Find k nearest neighbours of each data, and predict the label of the data using the majority voting. To determine the optimal value of k, cross-validation on different k is a useful method. The results of each k will be averaged, after comparing the averaged evaluation results, choose the k value from the trial which has the best performance.

Naïve Bayes is another good classification technique. For nominal data, if p (x | y) = 0 exists, Laplacian correction can be applied primarily. Due to numerous continuous values in the data set, estimate the mean and standard deviation of the training data. For instance, assume that the data follow a univariate Gaussian distribution, then calculate the p using the mean and standard deviation.

After when all classifications are finished, ensemble learning can make great use of the results. Because these classification methods all have their own advantages and also the drawbacks. They might also have a different assumption on the distribution of data. By combining the outputs using methods like majority voting, perhaps the prediction can achieve a better generalization.

## 3 Model Evaluation

One of the advantages of using cross-validation is that even if the models have different structures and complexity, their CV accuracy still can be compared. Using CV on the classification method means that there will be k evaluation results for each of them. These evaluations will be combined together to generate the overall performance. [4]

To evaluate the output of each model, directly using accuracy (the rate of correct predictions) is generally not a good way due to the imbalanced property of most datasets. Precision emphasis having correct predictions, while recall focuses on the proportion of instances that were retrieved. They both have their own characteristic.

Nevertheless, the F1 metric, which is the harmonic mean of precision and recall, could make a comprehensive evaluation of the result by combining two metrics.

## 4 Implementation Timeline

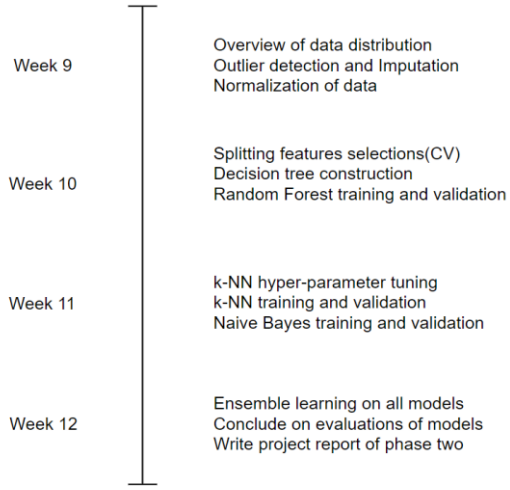


*Fig 2. Timeline of Model Implementations*

## References

[1] Siddharth M., Oghenekaro O., Mark P., (2020). Machine Learning for Subsurface Characterization. https://doi.org/10.1016/B978-0-12-817736-5.00001-6.

[2] Kropko, J., Goodrich, B., Gelman, A., & Hill, J. (2014). Multiple Imputation for Continuous and Categorical Data: Comparing Joint Multivariate Normal and Conditional Approaches. Political Analysis, 22(4), 497–519. http://www.jstor.org/stable/24573085

[3] Jo, J.-M. (2019). Effectiveness of Normalization Pre-Processing of Big Data to the Machine Learning Performance,14(3),547–552. https://doi.org/10.13067/JKIECS.2019.14.3.547

[4] Tabe-Bordbar, S., Emad, A., Zhao, S.D. et al. A closer look at cross-validation for assessing the accuracy of gene regulatory networks and models. Sci Rep 8, 6620 (2018). https://doi.org/10.1038/s41598-018-24937-4