Graph Based Link Prediction between Human Phenotypes and Genes

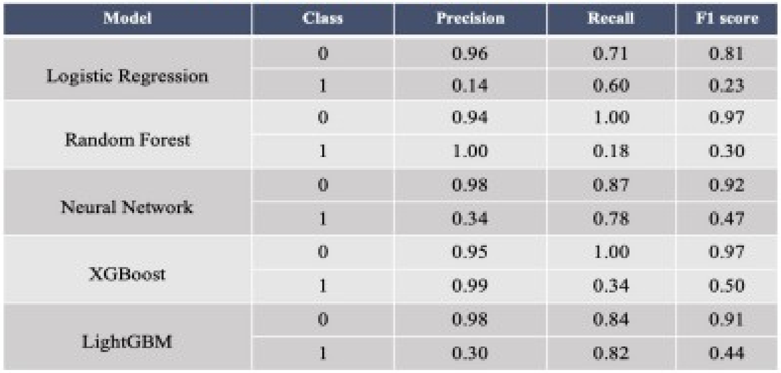
Abstract

The learning of genotype-phenotype associations and history of human disease by doing detailed and precise analysis of phenotypic abnormalities can be defined as deep phenotyping. To understand and detect this interaction between phenotype and genotype is a fundamental step when translating precision medicine to clinical practice. The recent advances in the field of machine learning is efficient to predict these interactions between abnormal human phenotypes and genes.

Comparing Performances of different models

To evaluate the predictive performance of these five models, Logistic Regression, Random Forest, Neural Network, XGBoost & LightGBM on human phenotype-gene dataset we use all the many different metrics including AUROC, AUCPR, Micro, Macro & Weighted Average precision, recall & F1 score, etc. To appropriately evaluate the imbalance nature of the dataset we calculate important metrics for each class instance in our case we just have two classes 0 and 1.

Table 1 *Class-wise Evaluation*



Model	Class	Precision	Recall	F1 score
Logistic Regression	0	0.96	0.71	0.81
	1	0.14	0.60	0.23
Random Forest	0	0.94	1.00	0.97
	1	1.00	0.18	0.30
Neural Network	0	0.98	0.87	0.92
	1	0.34	0.78	0.47
XGBoost	0	0.95	1.00	0.97
	1	0.99	0.34	0.50
LightGBM	0	0.98	0.84	0.91
	1	0.30	0.82	0.44

As you can see from Table 1, it is much easier to identify which model is doing a great job in identifying each class. Based on these metrics we can see that XGBoost & Random Forest performs well in identifying positive samples who are positives i.e., When these predict a link between nodes, they are correct 99% & 100% of the time respectively. Contrastingly, LightGBM beats all other methods in identifying correct actual positive – in other words, it correctly predicts 82% of all the links between these nodes. From Table 2, we can confer that LightGBM is better than all other models in terms of AUROC & AUCPR.

Conclusion

In this study, we presented an approach to predict links between human phenotype & genes using heterogeneous knowledge resources i.e., orphanet. The most important part of this study is to represent data into a graph and then finding a way to represent this graph into an appropriate feature set which will allow us to use it for down streaming tasks like a prediction. In essence, we provided a way to get the embedding vectors by using an algorithm called node2vec and then using these embeddings to build five different machine learning models. We evaluated and compared the performances using different quantitative metrics including AUROC, AUCPR, Micro, Macro & Weighted Precision, Recall, and F1 score. Some of these metrics were calculated for each class instance to better understand the situation for imbalanced class, in our case positive samples. Based on these metrics we found very interesting results. If we want to just focus on positive samples meaning the measure of the link that we correctly identify having associations of all the actual associations in the graph (we refer to it as Precision), then we may either use XGBoost or Random Forest algorithm. On the other hand, if we just want to focus on accurately identifying positives from True Positives i.e., actual links in the graph (Recall) then use LightGBM.

Adaptive Machine Learning Algorithm and Analytics of Big Genomic Data for Gene Prediction

Abstract

Artificial intelligence helps in tracking and preventing diseases. For instance, machine learning algorithms can analyze big genomic data and predict genes, which helps researchers and scientists to gain deep insights about protein-coding genes in viruses that cause certain diseases. To elaborate, prediction of protein-coding genes from the genome of organisms is important to the synthesis of protein and the understating of the regulatory function of the non-coding region. Over the past few years, researchers have developed methods for finding protein-coding genes. Notwithstanding, the recent data explosion in genomics accentuates the need for efficient gene prediction algorithms. This book chapter presents an adaptive naive Bayes-based machine learning (NBML) algorithm to deploy over a cluster of the Apache Spark framework for efficient prediction of genes in the genome of eukaryotic organisms. To evaluate the NBML

algorithm on its discovery of the protein-coding genes from the human genome chromosome GRCh37, a confusion matrix was constructed and its results show that NBML led to high specificity, precision and accuracy of 94.01%, 95.04% and 96.02%, respectively. Moreover, the algorithm can be effective for transfer knowledge in new genomic datasets.