ANALYSIS FOR DISEASE GENE ASSOCIATION USING MACHINE LEARNING

P.PAVANI, P.KUSUMA, K.ANUSH VARMA, Students, Dept.of.CSE, Raghu Engineering College, Dakamarri (V), Bheemunipatnam, Visakhapatnam Dist.Pincode: 531162.
Mr.Y V NAGESH MEESALA(Ph.D), Assistant Professor, Dept.of.CSE, Raghu Engineering College, Dakamarri (V), Bheemunipatnam, Visakhapatnam Dist. Pincode: 531162.

ABSTRACT: In this study, we propose and analyze some novel computational methods for the identification of genes associated with diseases. Some advance topological and biological features that are overlooked currently are introducing for identifying candidate genes. We evaluate different computational methods on disease-gene association data from DisGeNET based on TP rate, FP rate, precision, recall, F-measure, and ROC curve evaluation parameters. The results reveal that various computational methods with advanced feature set outperform previous state-of-the-art techniques by achieving precision up to 93.8%, recall up to 93.1%, and F- measure up to 92.9%. Significantly, we apply our methods to study three major disease types: Group, Disease and Phenotype. Simulation results sho44w that the proposed Extreme Gradient Boosting Algorithm (XGBoost) gives more accurate results as compared to previously published approaches.

INTRODUCTION

To recognize the basis of disease, it is essential to determine its underlying genes. Understanding the association between underlying genes and genetic disease is a fundamental problem regarding human health. Identification and association of genes with the disease require time consumingand expensive experimentations of a great number of potential candidate genes. Therefore, the alternative inexpensive and rapid computational methods have been proposed that can identify the candidate gene associated with a disease. Most of these methods use phenotypic similarities due to the fact that genes causing same or similar diseases have less variation in their sequence or network properties of protein-protein interactions based on-premises that genes lie closer in protein interaction network that causes the similar or same disease. However, these methods use only basic network properties or topological features and gene sequence information or biological features as a prior knowledge for identification of gene-disease association,which restricts the identification process to a single gene-disease association. A gene is the basic physical and functional unit of heredity that is responsible for different biological processes in an organism. The mutation in a single gene sequence may mutate a biological process and leads to a certain disease. The genes in the human body are not isolated they interact with one another, therefore, the mutation in a single gene may affect its interacting gene which may also play a part in the mutation of different biological processes and cause different diseases.

Correctly predicting new gene-disease associations has long been an important goal in computational biology. One very successful strategy has been the so-called guilt-by-association (GBA) approach, in which new candidate genes are found through their association with genes already known to be involved in the condition studied. This association can in practice be derived from many different types of data. Goh et al construct a network where genes are connected if they are associated with the same disease, whereas Tian et al. combine protein interactions, genetic interactions, and gene expression correlation, and Ulitsky and Shamir combine interactions from published networks and yeast two-hybrid experiments.Therefore,consideration of biological mechanisms and based on these mechanisms discovering the relationship between the diseases and genes is a serious challenge in modern biology and medicine. Understanding the association between casual genes and their genetic disease is a fundamental problem regarding human health. Technology is involved in the detection and monitoring of various human diseases such as Parkinson. Also, the Internet of Medical Things (IoMT) is in focus for addressing human health. Different experimental methods have been proposed to associate genes with a disease but these methods are expensive in terms of cost and time.

PROPOSED SYSTEM

In existing system, Machine Learning models used are Random Forest, Light Gradient Boosting Machine (LGBM) and Support Vector Machines are used for the detection of Disease Gene Association.

DISADVANTAGES:

- Low accuracy.
- Difficult to handle
- Low reliability.
- Time Consuming.

In proposed system, we implement supervised machine learning algorithm Extreme Gradient Boosting (XGBoost Classifier), for detection of the Disease Gene Association. Our Proposed model outperforms existing methods and give best results.

ADVANTAGES:

- High accuracy.
- Time Saving.
- High reliability.
- Low complexities.

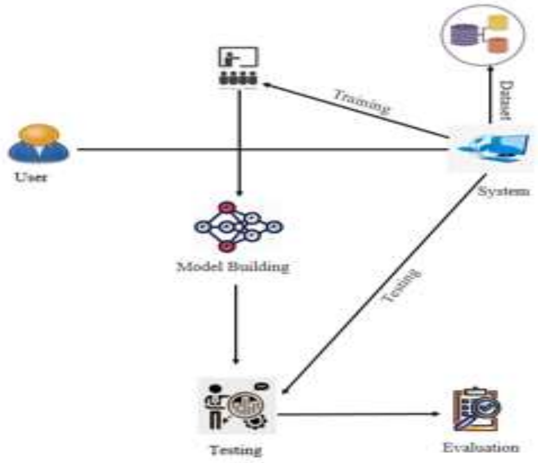
LITERATURE SURVEY

[1] C. K. Saket Navlakha, “The power of protein interaction networks for associating genes with diseases,” Bioinformatics, vol. 26, no. 8, p. 1057–1063, 2010.

- Understanding the association between genetic diseases and their causal genes is an important problem concerning human health. With the recent influx of high-throughput data describing interactions between gene products, scientists have been provided a new avenue through which these associations can be inferred. Despite the recent interest in this problem, however, there is little understanding of the relative benefits and drawbacks underlying the proposed techniques.

[2] O. M. E. R. T. S. R. S. Oron Vanunu, “Associating Genes and Protein Complexes with Disease via Network Propagation,” PLoS Computational Biology, vol. 6, no. 1, pp. 1-9, 2010.

- A fundamental challenge in human health is the identification of disease-causing genes. Recently, several studies have tackled this challenge via a network-based approach, motivated by the observation that genes causing the same or similar diseases tend to lie close to one another in a network of protein-protein or functional interactions. However, most of these approaches use only local network information in the inference process and are restricted to inferring single gene associations. Here, we provide a global, network-based method for prioritizing disease genes and inferring protein complex associations, which we call PRINCE. The method is based on formulating constraints on the prioritization function that relate to its smoothness over the network and usage of prior information. We exploit this function to predict not only genes but also protein complex associations with a disease of interest. We test our method on gene-disease association data, evaluating both the prioritization achieved and the protein complexes inferred.



ALGORITHMS

XGBoost:

XGBoost is an algorithm that has recently been dominating applied machine learning and Kaggle competitions for structured or tabular data. XGBoost is an implementation of gradient boosted decision trees designed for speed and performance.

XGBoost is a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework. In prediction problems involving unstructured data (images, text, etc.) artificial neural networks tend to outperform all other algorithms or frameworks. However, when it comes to small-to-medium structured/tabular data, decision tree based algorithms are considered best-in-class right now.

Bagging: Now imagine instead of a single interviewer, now there is an interview panel where each interviewer has a vote. Bagging or bootstrap aggregating involves combining inputs from all interviewers for the final decision through a democratic voting process.

XGBoost and Gradient Boosting Machines (GBMs) are both ensemble tree methods that apply the principle of boosting weak learners (CARTs generally) using the gradient descent architecture. However, XGBoost improves upon the base GBM framework through systems optimization and algorithmic enhancements.

Random Forest:

First, Random Forest algorithm is a supervised classification algorithm. We can see it from its name, which is to create a forest by some way and make it random. There is a direct relationship between the number of trees in the forest and the results it can get: the larger the number of trees, the more accurate the result. But one thing to note is that creating the forest is not the same as constructing the decision with information gain or gain index approach.

The author gives four advantages to illustrate why we use Random Forest algorithm. The one mentioned repeatedly by the author is that it can be used for both classification and regression tasks. Overfitting is one critical problem that may make the results worse, but for Random Forest algorithm, if there are enough trees in the forest, the classifier won’t overfit the model. The third advantage is the classifier of Random Forest can handle missing values, and the last advantage is that the Random Forest classifier can be modeled for categorical values.

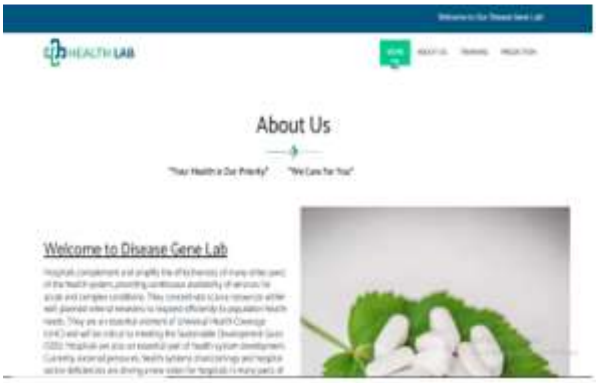
There are two stages in Random Forest algorithm, one is random forest creation, the other is to make a prediction from the random forest classifier created in the first stage.

STEPS:

1. Randomly select “K” features from total “m” features where k << m
2. Among the “K” features, calculate the node “d” using the best split point
3. Split the node into daughter nodes using the best split

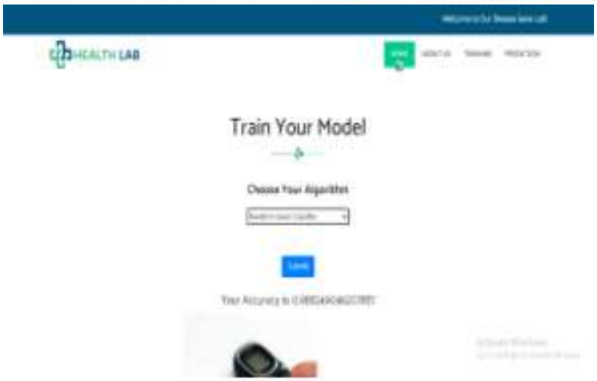
4. Repeat the a to c steps until “l” number of nodes has been reached
5. Build forest by repeating steps a to d for “n” number times to create “n” number of trees.

SAMPLE RESULTS



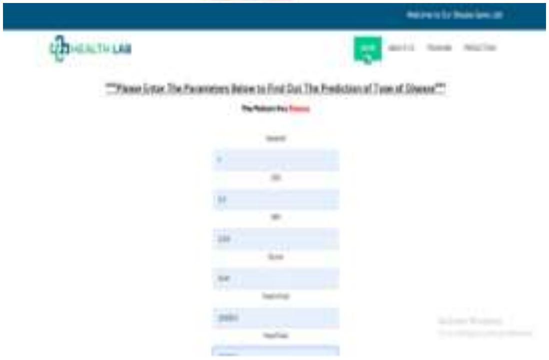
About Us

Welcome to Disease Gene Lab



Train Your Model

Choose Your Algorithm



Please Enter The Parameters Below to Find Out The Prediction of Type of Disease

CONCLUSION

In this project, we have successfully created a ML models to predict the type of Disease Gene. This is developed in a Jupyter Notebook. We noticed that out of XGBoost Classifier, Random Forest Classifier, Light GBM, K-Nearest Neighbors and Support Vector Classifier, XGBoost Classifier performs well with better accuracy.

REFERENCES

C. K. Saket Navlakha, “The power of protein interaction networks for associating genes with diseases,” Bioinformatics, vol. 26, no. 8, p. 1057–1063, 2010.
O. M. E. R. T. S. R. S. Oron Vanunu, “Associating Genes and Protein Complexes with Disease via Network Propagation,” PLoS Computational Biology, vol. 6, no. 1, pp. 1-9, 2010.

M. S. Mabrouk, “A Study of the Potential of EIIP Mapping Method in Exon Prediction Using the Frequency Domain Techniques,” American Journal of Biomedical Engineering, vol. 2, no. 2, pp. 17-22, 2012.

G. M. M. T. P. T. A. P. C. Y. J. F. F. R. S. V. R. B. T. D. M. P.-Y. K. C. A. M. F. P. R. Rahul C Deo, “Prioritizing causal disease genes using unbiased genomic features,” Genome Biology, vol. 15, no. 12, pp. 1-19, 2014.

D. M. Z.-H. D. Adarsh Jose, “A gene selection method for classifying cancer samples using 1D discrete wavelet transform,” International Journal of Computational Biology and Drug Design, vol. 2, no. 4, pp. 398-411, 2009..

R. J. M. Q. Z. S. L. Xuebing Wu, “Network-based global inference of human disease genes,” Molecular Systems Biology, vol. 4, pp. 1- 11, 2008.

S. B. T. M. M. H. S. Yu Qian, “Identifying disease associated genes by network propagation,” BMC Systems Biology, vol. 8, no. 1, pp. 1- 7, 2014.

W. A. U. I. B. X. W. M. S. L. Y. Z. L. J. Z. C. Aisha Sikandar, “Decision Tree Based Approaches for Detecting Protein Complex in Protein Protein Interaction Network (PPI) via Link and Sequence Analysis,” IEEE Access, vol. 6, pp. 22108-22120, 2018.