A STUDY OF CLASSIFICATION BASED MACHINE LEARNING METHODS FOR DISEASE GENE PREDICTION

K.Jayanthi and C.Mahesh
Department of Information Technology, Veltech University, Avadi, India

S.Sithsabesan
Department of Mathematics, Veltech Multitech Dr.RR and Dr. SR Engineering College, Avadi, India

ABSTRACT

One of the most challenging problems in biomedical is prediction of genes associated to a disease. Now a day’s different types of methods have been proposed to exploit existing data sources for solving this problem. This work to study various existing methods related to Machine Learning for disease gene prediction based on classification.

Keywords: Classification/Supervised Learning Techniques, disease gene prediction.

http://www.iaeme.com/IJMET/issues.asp?JType=IJMET&VType=8&IType=10

1. INTRODUCTION

Disease gene prediction, the process of identifying the most related candidate disease genes, is a predominant issue in biomedical research and many research studies have been done for this[3,4,11,12]. Reorganization of disease – correlated genes also leads to more constructive researchers about therapies for genetic diseases and progressively approaches a future of personalized medicine[5,6,7]. Early analysis and linkage analysis were used to identify disease genes, in which susceptible genes are scrutinized and thus it is very expensive for doing many experiments in various labs. So that, Ranking of genes method used for candidate the genes are launched (i.e. Genes are ranked by their relevance to a disease of interest). The gene which is having high rank is further investigated to find out associated biomedical confirmation [1]. Finally the goal of gene ranking also referred as prioritization method is used to predict disease associated genes. Formerly annotation based methods were proposed to address this problem. Later for providing high throughput technologies data of interaction between genes has emerged quickly and almost covered for genomes for that network based methods for the issue is becoming prominent. By using machine learning techniques formulation can be done...
by a classification for disease gene prediction. Various numbers of supervised learning techniques and various types of gene annotation data have been used to solve the disease gene prediction problem. Machine learning which aims to develop computer algorithms that improve with experience, Machine learning methods have been applied to a broad range of areas within genetics and genomics. Machine Learning is perhaps most useful for the interpretation of large genomic data sets and has been used to annotate a wide variety of genomic sequence elements.[1] Machine learning techniques are used to solve various important problems in biomedical, Such as genomic annotation. Pattern recognition, Classification of micro array data inference of gene regulatory networks, prediction of drug – target and discovery of gene- gene interaction in disease data[1]. In particular, they have been applied to identifying disease – associated genes. First the problem is formulated as a classification(Supervised Learning)

Problem, where the task is to learn classifiers from training data, and then the learned classifier is used to predict whether or not a test, candidate gene is a disease gene.

CLASSIFICATION BASED METHOD FOR DISEASE GENE PREDICTION

2. LITERATURE SURVEY
For the prediction of disease associated genes are generally approached by three main categories
1. Functional annotation Based.
3. Network based.

2.1. FUNCTIONAL ANNOTATION METHOD
It prioritized candidate genes by calculating the degrees of similarity of each candidate genes to a set of known disease genes based on profiles which were built from many functional annotation data sources. This method focuses on the integration of different biological data set to find more accurate similarity. So, these approaches are limited in that functional annotation data sources have not covered whole human genome yet[8,9,10].
2.2. MACHINE LEARNING METHOD
For the disease gene prediction machine learning addresses a classification. Where a classifier is learned from training data. After that the experience based classifier is used to predict whether or not a test/candidate gene is a disease gene based on the learning samples of classification methods are comprised of positive and negative training samples such as decision Tree, K-nearest neighbour, native Bayesian classifier, artificial neural network techniques and random forest. For providing high-throughput technologies data of interaction between genes has emerged quickly and almost covered for genomes for that network based methods for the issue is becoming prominent. By using machine learning techniques formulation can be done by a classification for disease gene prediction. Various numbers of supervised learning techniques and various type gene annotation data have been used to solve the disease gene prediction problem[1][2].

2.3. NETWORK BASED METHOD
Network-based methods are usually relied on a disease module principle that functionally similar genes are associated with similar phenotypes or diseases. Among them, methods solely based on gene/protein networks only exploit that principle by structural modules in the gene/protein networks. Meanwhile, others based on integration of these networks with a disease similarity network better exploit the principle and consequently result in higher prediction performance[2].

3. CLASSIFICATION
In the field of machine learning classification is the basic task. It is the reorganization of the category labels of instances that are normally described by a set of attributes in a dataset.

3.1. NEED OF CLASSIFICATION
To accurately predict class labels of instances whose values of attributes are known, but the labels of classes are unknown.

3.2. TYPES OF CLASSIFICATION
Classification can be described in machine learning as Binary, Multi class, Multi labelled, and Hierarchical. The classification task with disease gene prediction problem is to assign a disease label to a particular instance.

3.3. BINARY CLASSIFICATION
Using Binary classification, Identification of disease associated genes, where training set is comprised of positive and negative samples[1]. In which positive samples are constructed from known disease genes, whereas negative samples are the remaining which are not known to be associated with disease genes. This is one of the limitations of the binary classification based solutions, since the negative training set should be actual non-training genes. However construction of this set is nearly impossible in biomedical researches.

3.4. BINARY CLASSIFICATION USED IN DISEASE GENE PREDICTION
Various algorithms used in data mining classification can also specified in binary classification to solve the problem in Disease gene prediction.

**Decision Tree Algorithm:**
Using distinctive sequence features of known disease proteins compared to all human protein these binary class classification techniques has been used[1].
**K-Nearest Neighbour Algorithm (K-NN):**
It is used to find the growth of interaction data between proteins for the disease prediction.

**Instance based classifier:**
It is used based on topological properties of protein on a human protein interaction network.

**Naive Bayesian Classifier:**
To identify human disease genes by integrating multiple types of genomic, phenotype and interactomic data. Mainly used on different genomic datasets to identify human mitochondrial disease genes.

**Support Vector Machine:**
It is used for the problem identifying on both interaction and sequence data of protein. It provides better performance while compare to K- Nearest Neighbour Algorithm.

**Artificial Neural Network: (ANN)**
It is used to identify disease genes for four complex diseases (i.e Cancer, Type 1 Diabetes, Type 2 Diabetes, and Ageing) Using eight topological features calculated from protein interaction Network.

Moreover, unlike the methods mentioned above where the classifier was trained on all disease genes associated to all diseases, learning model was trained for each disease ontology term. Binary class classification techniques have extensively been used for disease gene prediction problem using many kinds of annotated genes.

### 3.5. MULTI CLASS CLASSIFICATION

Classification which deals with high dimensional data is referred as multi class classification. Classifying multiple disease genes profiles. This classification method focuses on detecting small sets of genes in which the relative comparisons of their expression values leads to class discrimination.

For an m-class problem, the classification rule typically depends on a small number of n-gene sets which provide transport decision boundaries and allow for potential biological interpretations.

### 4. CONCLUSION

In this paper, we have studied about addressing one of the biomedical challenging problem related to disease gene prediction, solutions can be analyzed by various existing research papers and Machine learning methods, this paper provides different kinds of methods and also describes about classification and it its types, such as Binary classification and Multiclass classification. Machine learning – based approaches usually treat disease gene prediction as a Binary – Class classification problem rather than Multiclass Classification. Another updated approach to disease gene prediction is to use unary/one class classification techniques, in which the classifier is learned from only positive samples (i.e Known disease genes). Our future work will include the comparison and analysis of how both Binary classification and Multiclass Classification differentiated and prediction of disease can be identified.
REFERENCES


