ML problems(5 July)

05 July 2016 15:56

A brief summary of different papers we are looking at. Here the problem statement, methods used for building models including feature selection, preprocessing and validation techniques used in the various papers have been mentioned. Further type of data set each study uses has also been listed.

Machine Learning in Genomic - It in general describes usage of ML to predict the cell variables rather than directly phenotype and use the cell variables instead later.
Problems-
Computational Model of Splicing – Uses input features extracted from genome near regions where splicing occurs and then predict the frequency with which exons are kept or excluded from mRNA. To train the model RNA-seq is used which counts the read determining relative abundance of isoform.

Computational Model for Protein and DNA binding – The sequences are taken and corresponding protein binding strength is our interest. We get a score for sequence and protein bonding. Microarray and sequencing For training we use this above score. Chip-Seq and microarray have been used for training data.

Machine Learning applications in cancer prognosis and prediction – It deals with different applications of ML in cancer prognosis. Deals with studies in susceptibility, recurrence and survival. Uses gene expression as well as clinical data. Different types of classifiers applied ranging from ANNs, SVM, RF etc. Also focused on Semi Supervised approaches.

Classification of Breast Cancer Subtypes by combining Gene Expression and DNA Methylation Data – Data extracted from TCGA. Builds a gene expression, DNA methylation and a combined model for prediction of subtypes. Also compares with the control model PAM50(original list of genes used for segregation of subtype specific to breat cancer). RF was used for building classification model. Uses recursive-elimination and bootstrapping for feature elimination(not clear in paper). Also uses some MiniDecreaseGiniIndex(specific to RF). Model validation using bootstrapping and area under ROC. Also does a feature list validation by intersecting with known breast cancer gene

MiRNA and Gene Expression based Cancer Classification using Self-Learning and Co-Training Approaches – Uses semi-supervised learning to classify samples using miRNA/expression profiles. Self Learning and Co-training approaches are used.

Gene-Expression-Based Cancer Subtypes Prediction Through Feature Selection and Transductive SVM – Uses microarray data of 4 tumor samples. First feature selection was performed(suitable genes were selected) . Consistency based feature selection was performed followed by signal-to-noise-ratio to get appropriate genes. Then these feature sets fed to TSVM. Note this is a semi-supervised learning approach.

Classification models for clear cell renal carcinoma stage progression, based on tumor RNAseq expression trained supervised machine learning algorithms – This study aims to develop a model to distinguish between different cancer stages based on gene expression profiles. Data taken from TCGA.All ML steps including data preprocessing, feature selection, generating classification models and independent testing performed using WEKA. It was basically a correlation based feature selection. It was then fed to models such as RF, NB, SMO, J48. For performance evaluation - accuracy, sensitivity or recall, specificity, Matthews Correlation Coefficient (MCC), F-value and auROC. 10 fold validation and independent testing used.

Pathway Based classification of subtypes - Main aim is to find pathways or gene sets as basis for differentiation of subtypes rather than relying on single genes as there is little overlap among them in different studies. Creates a two level feature vector with genes at leaves and gene sets at higher level. Different models purposed but I end up chosing pathways based on GSEA and then use certain genes expression from pathways to act as feature vectors. All above done for training data. These fed to SVM which in the end tells us the class.