Diagnosis of Mesothelioma's Disease Based on Classification

Sahba Akhavan[1], Shaojun Zhang[1], Delaram Ghoreishi[2]

[1]Department of Statistics, [2]Department of Physics, University of Florida

Introduction

Malignant mesotheliomas (MM) is very aggressive tumors of the pleura. MM is a rare disease but depending on region could be considered as one of the major public health problems.[1]

MM Diagnosis usually appears with symptoms like shortness of breath, pain in the chest and painful or persistent coughing, none of which immediately alert the doctor to a diagnosis of mesothelioma.[1]

There is no doubt that evaluations of data taken from patients and decisions of experts are the most important factors in diagnosis. However, sometimes different classification techniques are needed to diagnose the disease.[1]

Classification is often a very important part of process in many different fields like medicine. In this study possibility of using different classification methods is investigated for MM's disease diagnosis and the results are compared to Multilayer Neural Network (MLNN), Learning Vector Quantization neural network (LVQ) and Probabilistic Neural Network (PNN).The methods used are Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), classification trees, random forest (RF), Support Vector Machines (SVM) and k-nearest neighbors (KNN).

Data Set

In the dataset, all samples have 35 features that are considered more effective than other factors by doctor's guidance. These features are; age, gender, city, asbestos exposure, type of MM, duration of asbestos exposure, diagnosis method, keep side, cytology, duration of symptoms, dyspnoea, ache on chest, weakness, habit of cigarette, performance status, white Blood, cell count (WBC), hemoglobin (HGB), platelet count (PLT), sedimentation, blood lactic dehydrogenise (LDH), Alkaline phosphatise (ALP), total protein, albumin, glucose, pleural lactic dehydrogenise, pleural protein, pleural albumin, pleural glucose, dead or not, pleural effusion, pleural thickness on tomography, pleural level of acidity (pH), C-reactive protein (CRP), class of diagnosis.[1]

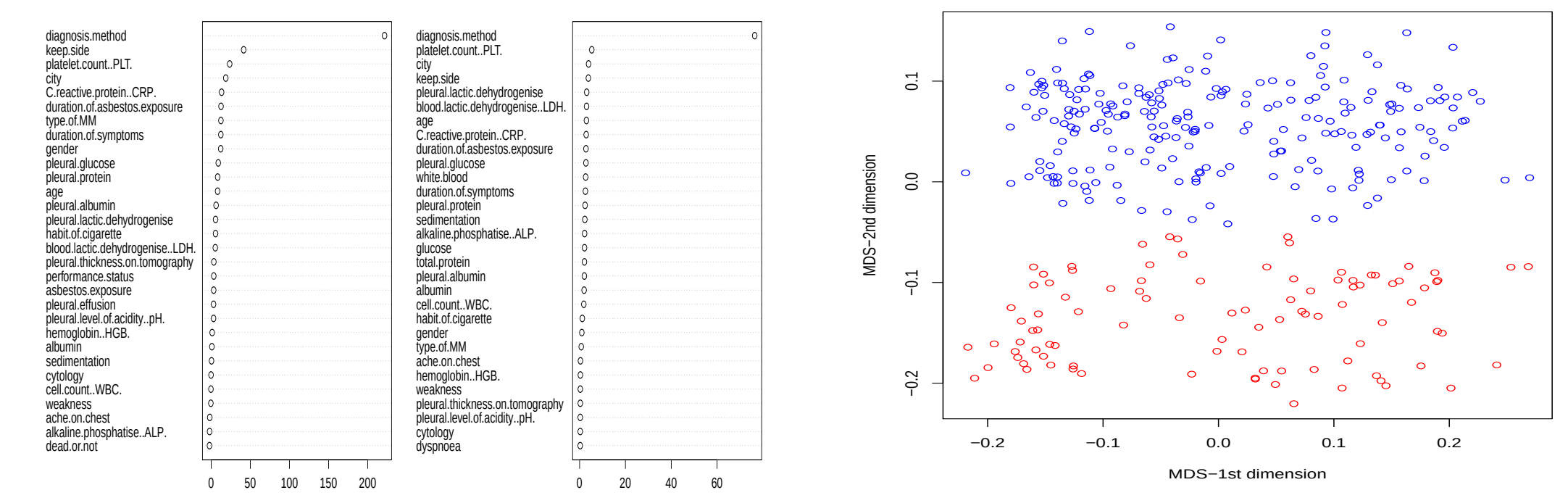
Challenge

The major difficulty associated with classifying this dataset is having many categorical variables which violate the assumptions needed for some of the classification methods like LDA and QDA.

Results

Different classification methods has been tested in order to classify class of diagnosis (1 if the patient is healthy, 2 if the patient is diagnosed to have Mesothelioma) based on all other features.

Among all different methods, random forest gave the perfect result, which led us to a more thorough interpretation of the data set. It is apparent from the variable importance plot that diagnosis.method is far more important than the other variables. The MDS plot has given a clean separation between the two class of data set.



(a) Variable Importance Plot for RF. Notice the importance of Diag.Method.

(b) MDS plot shows a clear separation between the two classes.

Figure 1 : VIP and MDS for the complete data set.

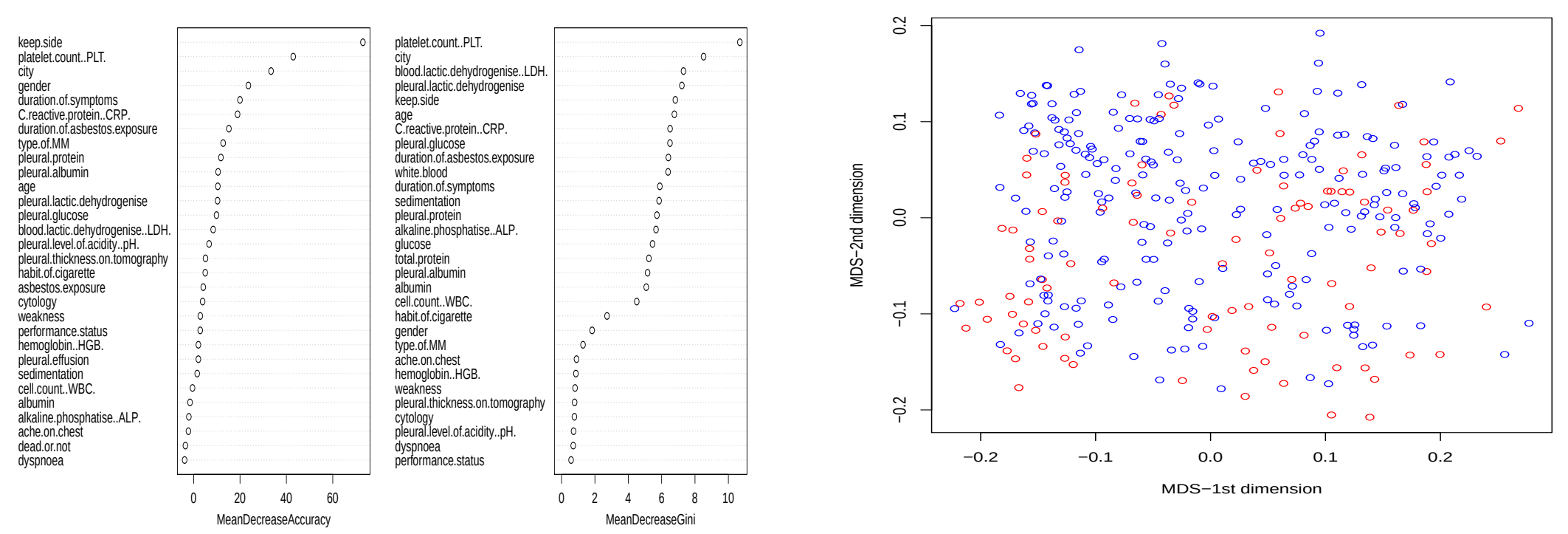
More detailed examination of the data set led us to realize that diagnosis.method is exactly the same as the class.of.diagnosis.

Notice

Diagnosis Method has been omitted from the data set because of the complete resemblance it had with the class of diagnosis

Results on the New Data Set

The variable importance of RF and MDS plots for the new data set have been shown in Figure 2. From these plots it is obvious that classification of this data set is not overall an easy task. From variable importance plot it can be observed that keep.side, platelet.count.PLT, city and gender are important variables in terms of accuracy and platelet.count.PLT, city,blood.lactic.dehydrogenise.LDH,pleural.lactic.dehydrogenise are important in terms of gini index.



(a) Variable Importance Plot for RF.

(b) MDS plot, class types are not easily separabale.

Figure 2 : VIP and MDS for the complete data set.

Classification Methods

LDA-QDA

Assumption used in LDA and QDA methods is that features have multivariate normal distribution in each class. Having many categorical variables in our data set, this assumption does not hold. For this reason, classification with LDA and QDA using all variables, would give us a singular within-class covariance matrix. We had to exclude some of the categorical variables to be able to perform LDA and we were not able to perform QDA.

Classification Trees

Trees are more capable in predicting the results from a mixture of numerical and categorical data types. Trees split the numerical variables in a binary mode based on some threshold value. The same applies to ordinal variables. Three fold cross validated tree has been tested on the data set.

Classification Methods

KNN

knn is a non-parametric classification method that is very easy to interpret and implement. However, it suffers from the curse of dimensionality. Here k is chosen to be 9 and the same three-fold cross-validation is applied.

SVM

SVM enables us to enlarge the feature space in order to accommodate a non-linear boundary between the classes. The radius basis function kernel is used and leave-one-out cross-validation is implemented.

Random Forest

Random forests produce an ensemble of trees so it is safe to assume that it will produce a more accurate result than classification trees.

Table below shows the error rate of each method. Note that for the neural network methods the variable diagnosis.method has not been excluded.

Method	Error Rate
LDA	0.25
Tree	0.28
KNN	0.33
SVM	0.30
RF	0.18
PNN	0.04
MLNN	0.06
LVQ	0.09

Table 1 : Error rates of different classification methods.

References

[1] O. Er, A. C. Tanrikulu, A. Abakay, and F. Temurtas. An approach based on probabilistic neural network for diagnosis of Mesothelioma's disease. *Comput. Electr. Eng.*, 38(1):75–81, jan 2012.

[2] George Michailidis. Applied multivariate analysis. *J. Econom.*, 3(3):320, 1975.