Classification of Mesothelioma's Disease

Shaojun Zhang
Department of Statistics
University of Florida

Sahba Akhavan Niaki
Department of Statistics
University of Florida

Delaram Ghoreishi
Department of Physics
University of Florida

Abstract

Malignant mesothelioma (MM) is an aggressive progress tumor that results from mesotel cells and pleura usually incurs. The two important causes, in MM etiologies are known as asbestos and erionite, both mineral fibers. Correctly classifying the type of the disease has a significant meaning in medicine. However, this data set was made public on the UC Irvine Machine Learning Repository recently and not much analysis has been done. In this project, several statistical methods will be applied to classify the type of the disease. The results of these methods will be analyzed and compared.

PROJECT PLAN

A. Data

The data set contains 324 MM patient data each with 34 categorical and quantitative features. The 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) and class of diagnosis. A separate dataset as test data is not available.

B. Methods

Different possible classification methods will be applied to this data set to determine the class of diagnosis for each patient. The methods include LDA, QDA, classification trees, random forest, K-nearest neighbors, SVM and logistic regression. The performance of these methods will be evaluated based on cross-validation.

Three different artificial neural networks (ANNs) structures, probabilistic neural network (PNN), multilayer neural network (MLNN) and learning vector quantization (LVQ) neural network, have been applied for classification of this dataset, with accuracies of 96.3%, 94.41% and 91.14% respectively [1]. Therefore, we will also compare the results from our methods to those from ANNs.

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," *Computers Electrical Engineering*, vol. 38, no. 1, pp. 75 – 81, 2012, special issue on New Trends in Signal Processing and Biomedical Engineering.

[2] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning: With Applications in R*. Springer Publishing Company, Incorporated, 2014.