A Comparison of Classification Methods to Diagnose Vertebral Column Disorder

Sowjanya Ratho, Guenter Tusch, PhD
Medical and Bioinformatics Graduate Program, School of Computing and Information Systems,
Grand Valley State University, Allendale, MI, USA

Summary

Spinal disorders are extremely common in two third of adults. In this project, we are analyzing dataset on biomechanical features of Vertebral Column to classify patient in to normal or abnormal. To improve the accuracy of prediction, it is also important to know the correlation between variables and their impact on classification. Principle comment analysis, correlation matrix and feature selection libraries are used for feature selection, and we carried analysis with 6 variables, these variables explain about 95% of data. Hence, further we try to fit logistic regression, Random forest and SVM algorithms for classification on new data.

Background

Spinal is a complex network of vertebras, muscles, nerves and discs. Most of the causes linked to muscle strain, injury or overuse or it's a symptom of the specific condition of the spine like herniated disc, degenerative disc, spondylolisthesis.

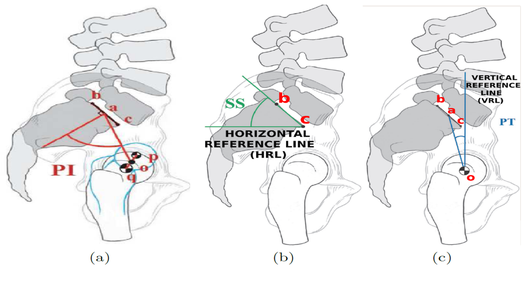
Radiographs, which were at first frequently used to diagnose spine problems now turn out to be less appealing since Magnetic Reverberation Imaging (MRI) was found where it can portray the spinal cord and other soft tissues issues better. Although MRI is a good diagnostic system, examining MRI pictures requires significant experience. Considering the fact, the mistake in the MRI picture investigation can prompt the wrong treatment. Use of Computer-Aided Diagnosis (CAD) framework can help the radiologist recognize problems in the medical field to support their decision by pattern recognition and machine learning.

This dataset has 310 observations which grouped into two classes. Each record comprises of 12 attributes and 1 class(Label).

Pelvic incidence = pelvic tilt + sacral slope

a line form center of the S1 endplate to the center of the femoral head and second line is drawn perpendicular to a line drawn along the S1 endplate. (Figure-1)

The angle between these two lines is the pelvic incidence



Material and Methods

Data mining typically utilized for classification grouping and applied in prediction, it's one of the key tasks. Classification techniques come under supervised learning in this it maps the data into predefined goals. It builds a classifier based upon a defined class with specific attributes to describe the objects or one attribute to describe the group of objects. Then classifier predicts the label(class) of new inputs based on standards of other attributes in the dataset.

Nature of dataset: All the attributes are numerical and continuous. Dataset comprises of patients categorized into one of two classes: Normal or Abnormal .Data looks clean after performing the Anderson-Darling normality test.

Correlation Matrix: We have used the corrplot library in R, to see the correlation and preliminary visualizations of our data to explore more about features and to get an idea about the variables and relations amongst them. Our dataset has some highly-correlated variables that contribute to classifying data correctly. (Figure 2 a)

Principal Components Analysis (PCA): PCA is applied to extract key variables from a large set of variables. With 12 variables, there will be more than 200 three-dimensional scatterplots to be studied. To interpret data in a more meaningful way, it is important to decrease the quantity of variables to less, interpretable linear combinations of the data (Figure 2 b)

Logistic Regression Model: Performed logistic regression using a function called glm in the R, this function used for the binary regression model, it is the best method when a dependent variable is categorical with binary classification

Random Forests Model: In R "randomForest" package has the function called randomForest () which is used to create and analyze random forests. This algorithm takes the concept of the decision tree and generates many trees. (Figure 4)

Support Vector Machine (SVM): SVM is a strong algorithm for classification and pattern recognization and this model is better when there is no clear boundary between classifiers because it uses a technique called kernel trick to transform data. Depending upon these data transformations SVM model finds an optimal boundary between possible outcome it also effective with high dimensional data. It have a great ability to classify multiple classes cases, however, by nature, this algorithm is a binary classifier.

Acknowledgements:

I would like to express my deepest appreciation to Dr. Henrique da Mota for making the data available in UCI machine learning repository. R and Analytical bloggers for sharing their knowledge

User Interface and Evaluation

Machine Learning (ML)can be effectively applied to medicinal industry to find hidden patterns, new information. This project examined existing condition of classification methods in medical data mining that we can use in medical and diagnostic industry. When we compared all 6 models k-Nearest Neighbor gave the highest accuracy in prediction (96.77%), (Figure 3) followed by the logistic regression (95.13%) model. Feature selection and PCA analysis were helpful to identify important variables. Choosing only principle components for classification analysis proven to be effective to be able to improve accuracy in predicting target variables. Decision trees or random forest are helpful in interpretability of data and they produce good visualization to understand results. Best classification algorithm can be applied in medical industry to support the physician to make a good decision without hesitation. They can eliminate medical errors, improve quality and it saves time.

In addition to the project, I would like to apply k-NN, ANN algorithms, because k-NN seems to be more accurate and effective than the methods which I have used. A global algorithm must be developed in the future so that it can be applicable to the numerous data types.

References:

Rocha Neto, A. R. & Barreto, G. A. (2009). 'On the Application of Ensembles of Classifiers to the Diagnosis of Pathologies of the Vertebral Column: A Comparative Analysis', IEEE Latin America Transactions, 7(4):487-496.

S. K. Reddy, S. R. Kodali and J. L. Gundabathina, "Classification of Vertebral Column using Naive Bayes Technique," International Journal of Computer Application, pp. 38-42, 2012.

Y. Unal and E. Kocer, "Diagnosis of Pathology on the Vertebral Column with Backpropagation and Naive Bayes Classifier," Technological Advances in Electrical, Electronics and Computer Engineering, pp. 278- 281, 2013.

S. Ansari, N. Naveed, F. Sajjad and I. Shafi, "Diagnosis of Vertebral Column Disorders Using Machine Learning Classifiers," Information Science and Application, 2013.

Corrilation Matix

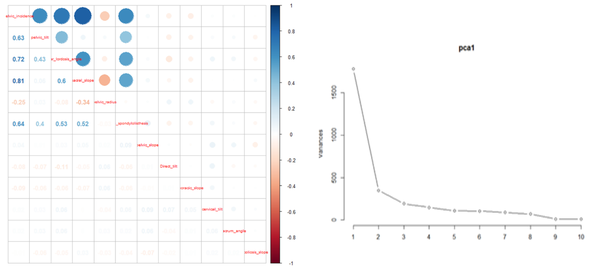


Figure 2: a Correlation matrix b: PCA

ACCURACY OF THREE CLASSIFIERS ON TWO VERTEBRAL COLUMN DATASETS (WITHOUT GENETIC ALGORITHM AND BAGGING)

Validation	Vertebral Column 3 Classes			Vertebral Column 2 Classes		
	Accuracy			Accuracy		
	Naive Bayes	Neural Network	k-Nearest Neighbor	Naive Bayes	Neural Network	k-Nearest Neighbor
Cross Validation	81.94%	84.52%	81.94%	77.42%	84.19%	85.16%
90% - 10%	80.65%	83.87%	90.32%	80.65%	83.87%	87.10%
80% - 20%	82.26%	80.65%	77.42%	77.42%	83.87%	79.03%
70% - 30%	81.72%	83.87%	77.42%	77.42%	82.80%	77.42%
60% - 40%	83.06%	87.90%	78.23%	79.84%	86.29%	79.84%

TABLE II
ACCURACY OF THREE CLASSIFIERS ON TWO VERTEBRAL COLUMN DATASETS (WITH GENETIC ALGORITHM AND BAGGING)

Validation	Vertebral Column 3 Classes			Vertebral Column 2 Classes		
	Accuracy			Accuracy		
	Naive Bayes	Neural Network	k-Nearest Neighbor	Naive Bayes	Neural Network	k-Nearest Neighbor
Cross Validation	86.13%	88.06%	89.03%	82.90%	87.74%	88.71%
90% - 10%	90.32%	90.32%	96.77%	83.87%	90.32%	96.77%
80% - 20%	90.32%	88.71%	90.32%	83.87%	90.32%	91.94%
70% - 30%	87.10%	88.17%	87.10%	87.10%	89.25%	90.32%
60% - 40%	87.10%	88.71%	87.90%	86.29%	88.71%	88.71%

Figure 3: Previous Research Summary

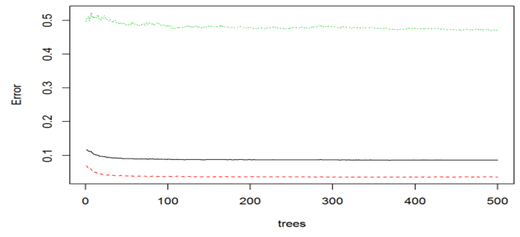


Figure 3: Error rate with number of trees

Over View of Variable Separation

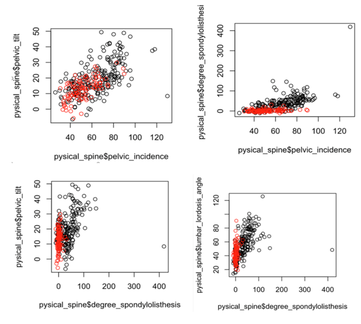


Figure 4: See the data pattern