SAS® GLOBAL FORUM 2016

IMAGINE. CREATE. INNOVATE.

Application of Gradient Boosting through SAS® Enterprise Miner™ to Classify Human Activities

#SASGF

Application of Gradient Boosting through SAS® Enterprise Miner™ to Classify Human Activities

Minh Pham, Mostakim Tanjil, Mary Ruppert-Stroescu

Oklahoma State University

ABSTRACT

Using smart clothing with wearable medical sensors integrated to keep track of human health is now attracting many researchers. However, body movement caused by daily human activities inserts artificial noise into physiological data signals, which affects the output of a health monitoring/alert system. To overcome this problem, recognizing human activities, determining relationship between activities and physiological signals, and removing noise from the collected signals are essential steps. This paper focuses on the first step, which is human activity recognition.

Our research shows that no other study used SAS® for classifying human activities. For this study, two data sets were collected from an open repository. Both data sets have 561 input variables and one nominal target variable with four levels. Principal component analysis along with other variable reduction and selection techniques were applied to reduce dimensionality in the input space. Several modeling techniques with different optimization parameters were used to classify human activity. The gradient boosting model was selected as the best model based on a test misclassification rate of 0.1233. That is, 87.67% of total events were classified correctly.

DATA EXPLANATION

- We used a dataset collected from the Center for Machine Learning and Intelligent Systems at University of California, Irvine, which was made available in the open repository in July 2015. The data were collected with sensors embedded in a smartphone which was attached on the waist when subjects performed certain activities.
- Independent variables included the 3-dimension acceleration and 3-dimension angular rate that were collected with a sampling rate of 50Hz.
- Dependent variable included human activities such as standing, sitting, lying, walking, walking downstairs and walking upstairs

DATA PREPARATION

New independent variables

- Those 6 raw signals were put through a filter to remove some noise. Then 561 features, which were calculated from both time and frequency domains, were extracted by using sliding window method with 2.56 seconds window width and 1.28 seconds step size.
- Each window can be treated as an observation and there were totally 7,767 observations. Those 561 features were considered independent variables

New dependent variable

- Four new levels were created: 1 refers to *Moving* (including walking, walking downstairs and walking upstairs) , 4 refers to *Sitting*, 5 refers to *Standing*, and 6 refers to *Lying*

Data transformation

- Out of 561 variables, 249 variables needed to be transformed because of their high skewness and kurtosis values. 12 variables were transformed using Maximize Normality Power, 4 variables using Exponential, 223 variables using Log with base 10, and 10 variables using Square Root

Principal Component Analysis

- By using this dimensional reduction technique, from original 561 variables, we selected 15 PCs, which explained 75.56% of total variation in the input space, and then 1these 5 PCs were fed into subsequent modelling steps.

MODELING

- **Decision Tree** models with different splitting rule criteria – Probability of Chi-square, Gini and Entropy, different number of branches and different depth were built. In all cases, Bonferroni adjustments to the p-values were done before the tree split is chosen.
- **Neural Network** models with different network architecture (Generalized Linear Model, Multilayer Perceptron, Ordinal Radial with equal and unequal width, Normalized Radial with equal width and height, and Normalized Radial with unequal width and height) with hidden units 3 and 4 were built. For all the neural network models, default optimization technique was applied.
- **Gradient boosting** models with training proportion 50 and 60 were used. For splitting rule, only Square Loss Function was used. As Huber M-Regression Loss function is more appropriate for interval target, it was not used. The base learner was built with maximum branch 2, and maximum depth 2 and 3. For all the gradient boosting models, number of iteration and shrinkage parameter were set to default, which are 50 and 0.1 respectively.

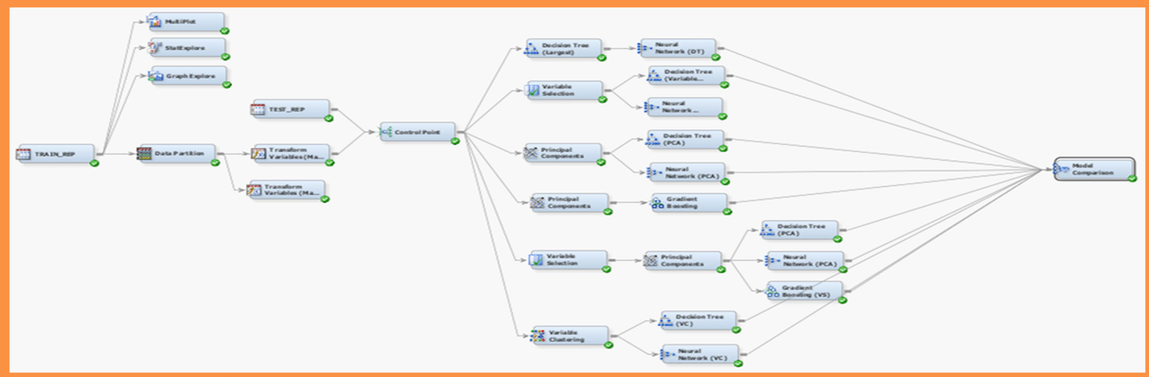

Fig 1. Partial process flow diagram

Application of Gradient Boosting through SAS® Enterprise Miner™ to Classify Human Activities

Minh Pham, Mostakim Tanjil, Mary Ruppert-Stroescu

Oklahoma State University

RESULTS

	Test Misclassification Rate	Test ROC Index	Test Gini Coefficient
Gradient Boosting (PCA)	0.1233	0.993	0.987
Neural Network (PCA)	0.1262	0.997	0.993
Decision Tree (PCA)	0.1632	0.927	0.855
Decision Tree (Variable Selection)	0.2565	0.878	0.756
Neural Network (Variable Selection + PCA)	0.2951	0.944	0.889
Gradient Boosting (Variable Selection)	0.3254	0.837	0.673
Decision Tree (Variable Selection + PCA)	0.3289	0.799	0.597
Neural Network (Decision Tree)	0.5088	0.5	0
Neural Network (Variable Clustering)	0.5088	0.5	0
Decision Tree (Variable Clustering)	0.8242	0.5	0

Fig 2. Comparison of Fit Statistics of Top 10 Models

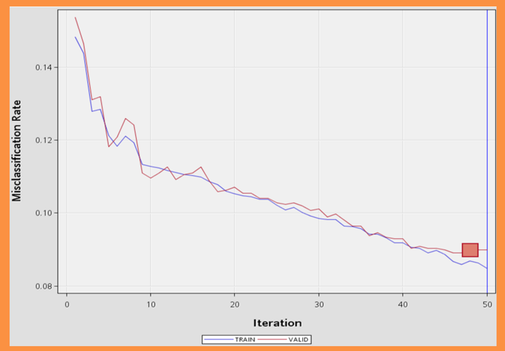

Fig 3. Subseries Plot of Gradient Boosting Model

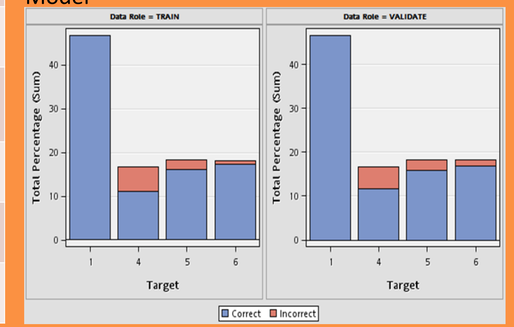

Fig 4. Classification Chart of Train and Validation Data

- The Gradient Boosting model was iterated 50 times in the series to use in the final model
- Classification chart shows that Gradient Boosting correctly classified '1 – Moving' with 100% precision in both training and validation data (Fig. 4). On the other hand, false positive rate in the validation data for '4 – Sitting', '5 – Standing' and '6 – Lying' are 4.97%, 2.53% and 1.3% respectively

CONCLUSIONS

In order to assure the validity of physiological data collected from textile-based sensors, attention must be given to the noise portion of the signals. With the goal of understanding how a person's movement can influence biomedical signal quality gathered from textile-based sensors embedded in a garment, we employed the Human Activity Recognition principle. In order to classify a person's movement, we applied recently developed SAS algorithms to create a model that has 87.67% accuracy when classifying the four different activities of moving, sitting, standing, and lying down.

Future study will include integrating a gyroscope and accelerometer into the prototype smart medical garment that we developed and to do human wear tests. Data will be collected and processed using the same model as this study. Then the classification algorithm using SAS® Enterprise Miner will be employed to score newly collected data to validate the efficiency of the model.

REFERENCES

Reyes-Ortiz, J.L., Oneto, L., Samà, A., Parra, X. and Anguita, D., 2016. "Transition-aware human activity recognition using smartphones." *Neurocomputing*, 171:754-767.

Tapia, E.M., Intille, S.S. and Larson, K., 2004. "Activity recognition in the home using simple and ubiquitous sensors" *Springer Berlin Heidelberg,* 158-175.

Hong, Y.J., Kim, I.J., Ahn, S.C. and Kim, H.G., 2010. "Mobile health monitoring system based on activity recognition using accelerometer." *Simulation Modelling Practice and Theory*, 18(4):446-455.

Aggarwal, J.K. and Ryoo, M.S., 2011. "Human activity analysis: A review." *ACM Computing Surveys (CSUR)*, 43(3):16.

Maldonado, M., Dean, J., Czika, W. and Haller, S., 2014. "Leveraging Ensemble Models in SAS® Enterprise Miner™." *Proceedings of the SAS Global Forum 2014 Conference*. Available at https://support.sas.com/resources/papers/proceedings14/SAS133-2014.pdf

SAS® GLOBAL FORUM 2016

IMAGINE. CREATE. INNOVATE.

LAS VEGAS | APRIL 18-21

#SASGF