Paper 11801-2016

Application of Gradient Boosting through SAS® Enterprise Miner™ 12.3 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.

INTRODUCTION

Recently, research about smart garments has become a key area of interest. There are currently a few products available on the commercial market, such as Heathwatch, Adidas Micoach, and Hexoskin, which can be used in medical or sport applications. Wearable electrodes are embedded in those smart garments in order to monitor continuous physiological signals while maintaining wear comfort. The signals collected normally include ECG (Electrocardiogram) and respiration, from which features such as heart rate and respiration rate are extracted. Then they are used for detection of some heart-related disorders, or just simply to alert when heart rate is abnormal. Still, there are several drawbacks to those products. To ensure comfort, the ECG electrodes in such products are designed so that they do not stick on human skin, and are not tightly pressing the skin. The ability of electrodes to conduct the signal from the body to the health monitoring system is very sensitive to body movement, and variation in the space between the electrode and the skin creates artificial noise within the signal. Therefore, the artificial noise in these signals might hide important features for analysis or could easily lead to a false alarm. Classifying movement is an essential component to the process of determining the relationship of textile electrode signal quality to the subject's physical movements, and provides an important foundation for the development of viable medical-quality textile-based sensing systems.

Currently we are designing a smart medical garment embedded with wearable textile electrodes. Its function is to monitor physiological signals including ECG, respiration, air flow and Spo2 (saturation of oxygen in the blood). We have developed a prototype that successfully acquires all signals. However, if the wearer is moving, such as walking or doing postural transitions, we found that the noises caused by the motion artifact were added to the ECG signal, which made analysis and detection tasks harder. This is why Human Activity Recognition (HAR) was integrated: to increase signal quality and reduce the false alarm rate. For example, the movement-related signals collected during an experiment can be used for adaptive filtering to remove noise from physiological signals; or based on activity recognition output, we may use decision fusion to help generate accurate conclusions based on the specific goal of the monitoring system, such as alerts, medical intervention, or other actions.

HAR in daily life has been researched for the past decade since systems with human-computer interaction was developed, such as the smart home or mobile health monitoring systems, etc. Traditionally, HAR research has been widely conducted by using computer vision, where cameras have been used to record user's images, however, in this case user privacy risks being violated. Recently, with the development of MEMS (micro electro-mechanical systems), a wearable Inertial Measurement Unit (IMU), which includes gyroscope and accelerometer, has emerged as a key sensor for human activity

recognition. Pattern recognition, or machine learning, is a popular and useful approach of HAR researchers. However, based on our literature review, there is almost no method that satisfies generalization characteristics by providing a model that can apply properly to different people with different styles of activity or behavior.

In this paper, we applied the data mining approach to HAR based on a dataset collected from the Center for Machine Learning and Intelligent Systems at University of California, Irvine. The dataset was made available in the open repository in July 2015. The experiment was conducted on 30 people with age ranging from 19 to 48 years. This is a supervised classification problem and we employ the software SAS® Enterprise Miner 12.3 to solve it.

DATASET EXPLANATION

The data were collected with sensors including a 3-axis gyroscope and 3-axis accelerometer embedded in a smartphone. Thirty people were asked to wear the smartphone on the waist and to perform six activities including standing, sitting, lying, walking, walking downstairs and walking upstairs.

Our independent variables included the 3-dimension acceleration and 3-dimension angular rate that were collected with sampling rate of 50Hz. Those 6 raw signals were put through a filter to remove some noise, then 561 features were extracted by using sliding window method with 2.56 seconds window width, and 1.28 seconds step size. They were calculated from both time and frequency domains. For example, some features were extracted by calculating mean value, standard deviation, median, maximum, minimum values, etc. in each dimension of acceleration or angular rate. The explanation of features is fully mentioned in the paper [9]. Each window can be treated as an observation and all those features were considered independent variables of the dataset used in the SAS® Enterprise Miner 12.3.

The dependent variables included the human activities that were video-recorded and labeled manually. Initially, the target variable had 12 levels, including 6 activities mentioned above and 6 postural transitions which are stand-to-sit, sit-to-stand, sit-to-lie, lie-to-sit, stand-to-lie, and lie-to-stand. However, in our project, we used only 4 levels, in which 3 levels were standing, sitting, and lying, and all the other activities and transitions were transformed into 1 level, considered as a moving state. Following are the 4 levels that our target variable has:

- 1 refers to ‘Moving’
- 4 refers to ‘Sitting’
- 5 refers to ‘Standing’
- 6 refers to ‘Lying’

DATA PREPARATION

The data came in separate datasets; dataset-1 and dataset-2 were used for this study. Dataset-1 has 7,767 observations and dataset-2 has 3,162 observations. For honest assessment, dataset-1 was split into 70% train and 30% validation, which resulted in 5,433 observations in train and 2,334 observations in validation. Dataset-2 was used for test. Both datasets have 561 interval input variables and one nominal target variable. None of the dataset has missing values.

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. Transformation resulted in substantial reduction of both skewness and kurtosis values. However, transformation was applied only to the training partition.

The dataset has a very high number of input variables. To reduce redundancy and irrelevancy from the input variables as well as to address multicollinearity issues, both unsupervised variable reduction and supervised variable selection techniques were applied. Variables selected via unsupervised PC Analysis and Variable Clustering, and supervised Variable Selection, Decision Tree and variables selected by Variable Selection node transformed into PCs were fed into subsequent model building nodes.

PRINCIPAL COMPONENT ANALYSIS

The primary purpose of unsupervised variable reduction technique – PC Analysis is to reduce the redundancy of information contained in the input variable. PCs are linear combination of original input variables where "the first PC is the linear combination with maximal variance; the second PC is the linear combination with maximal variance in a direction orthogonal to the first principal component and so on [10]." The 561 input numeric variables of this study were converted into 561 uncorrelated PCs which could be written as

- $PC1 = W_{1,1} * F1 + W_{1,2} * F2 + \ldots\ldots\ldots + W_{1,561} * F561$
- $PC2 = W_{2,1} * F1 + W_{2,2} * F2 + \ldots\ldots\ldots + W_{2,561} * F561$
- ……..
- $PC561 = W_{561,1} * F1 + W_{561,2} * F2 + \ldots\ldots\ldots + W_{561,561} * F561$

Where W's are weights of respective original input variables.

Usually the following general guide lines are considered to decide how many PCs to retain:

- To account for a specified percentage of total variation.
- Eigen values greater than 1.
- In the Scree plot of Eigen values or Log of Eigen values, where there is a natural break or elbow.

In this study, it is decided to retain 15 PCs for subsequent modelling steps. 15 PCs explained 75.56% of total variation in the input space (

Table 1).

Eigenvalues of Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	284.473060	239.254804	0.5071	0.5071
2	45.218256	28.185574	0.0806	0.5877
3	17.032682	4.410514	0.0304	0.6180
4	12.622168	2.911345	0.0225	0.6405
5	9.710823	1.110295	0.0173	0.6579
6	8.600528	1.372809	0.0153	0.6732
7	7.227719	0.238830	0.0129	0.6861
8	6.988889	0.706485	0.0125	0.6985
9	6.282404	1.277940	0.0112	0.7097
10	5.004464	0.321942	0.0089	0.7186
11	4.682522	0.253661	0.0083	0.7270
12	4.428861	0.194986	0.0079	0.7349
13	4.233875	0.431929	0.0075	0.7424
14	3.801945	0.181006	0.0068	0.7492
15	3.620940	0.199701	0.0065	0.7557

Table 1. Eigen Values of Correlation Matrix and Total Percentage of Variation Explained by PCs

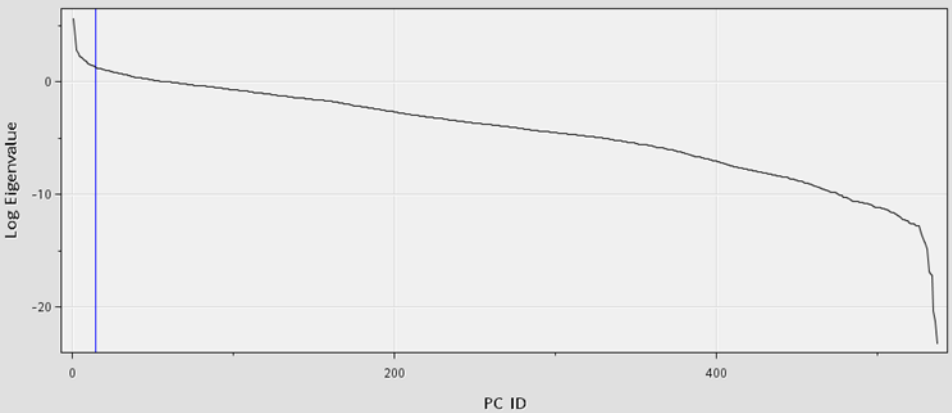

Figure 1. Scree Plot of Log of Eigen Values

Scree plot shows that there is no obvious natural break in the log of Eigen values (

Figure 1). For 15 PCs, the Eigen value was 3.62. However, increasing the number of PCs did not increase subsequent models' efficiency substantially. Therefore, lower number of PCs were preferred to avoid dimensional complexity.

MODEL BUILDING

Variables selected by Decision Tree, Variable Selection, Variable Clustering, PC and Variable Selection passed through PC were used to build different classification models. Several modelling techniques with different properties were applied for model optimization and to build the best possible model (Fig. 2).

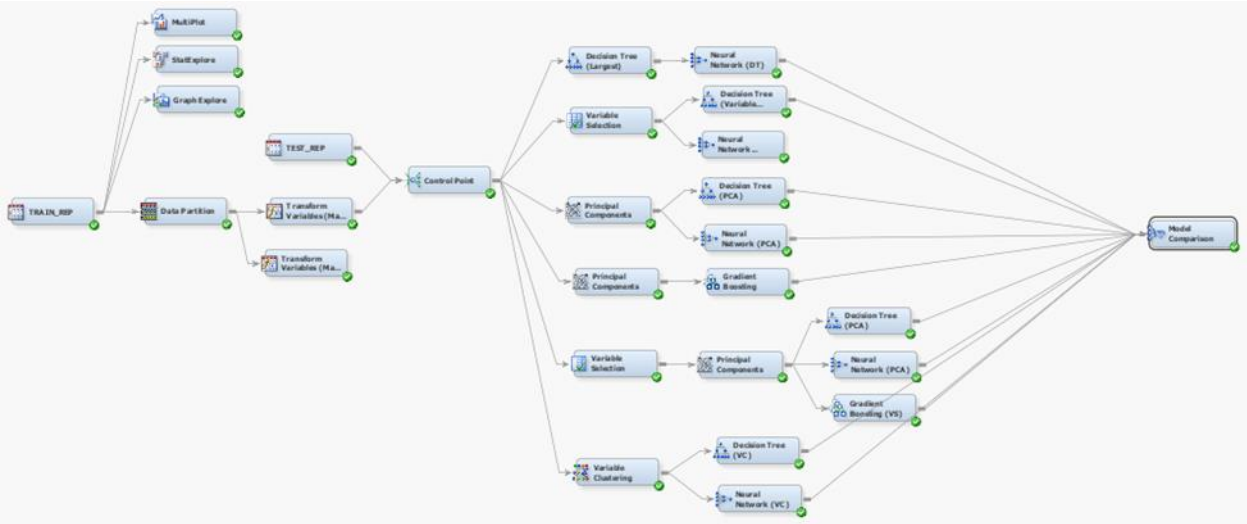

Figure 2. Partial Process Flow Diagram

- 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.
- To build Gradient boosting models, 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.

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

Table 2. Comparison of Fit Statistics of Top 10 Models

All the models were iterated several times and their results were compared using the model comparison node of SAS® Enterprise Miner. As the target variable is nominal with four levels, test misclassification

rate was considered as the primary criteria to select the best model. Misclassification rate is the percent of outcomes predicted incorrectly. It is computed by the following formula:

Misclassification rate = $\frac{False\ Positive + False\ Negative}{Total\ Observations}$

The lower the value of misclassification rate is, the better the model is. Although the test misclassification rate was the primary criteria to select the best model, other fit statistics were also compared (Table 2).

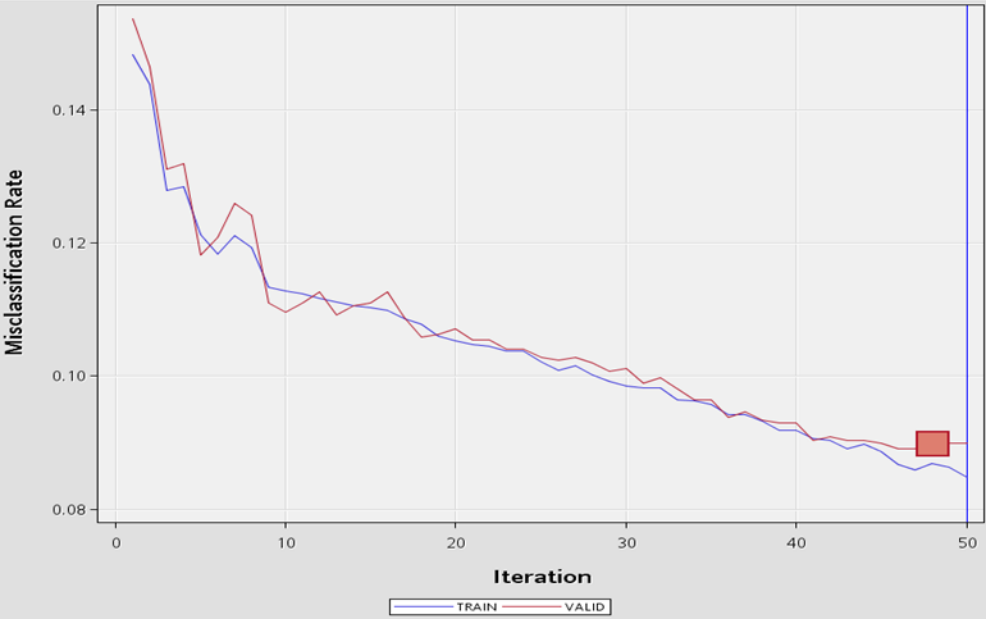

Figure 3. Subseries Plot of Gradient Boosting Model

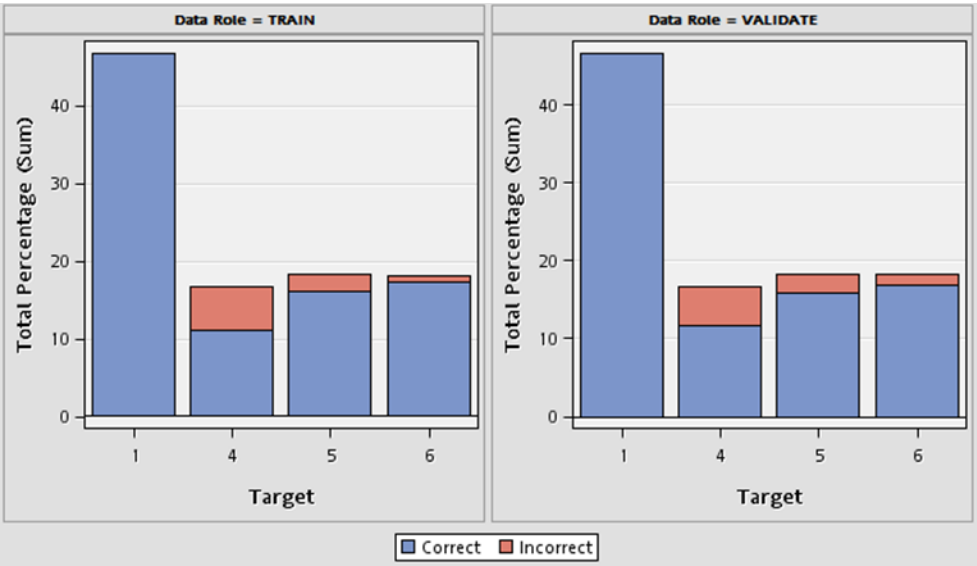

Figure 4. Classification Chart of Train and Validation Data

Comparison of fit statistics of top 10 models shows that Gradient Boosting Passed through PC (15 PCs) outperformed other models and was selected as the best model by the model comparison node. The best

model had training proportion 60, number of iteration 50, shrinkage parameter 0.1 and Square Loss Function with the base learner of maximum branch 2 and maximum depth 2. Both test ROC Index and Gini coefficient show that Neural Network passed through PC (15 PCs) performed better than Gradient Boosting model, nevertheless, values of Gradient Boosting are comparable to that of Neural Network.

Despite having easy interpretability and ability to handle missing values, decision trees sometimes are unstable in the presence of different training datasets. On the other hand, Gradient Boosting is a form of ensemble model which is a sequential iteration of several decision trees. It starts with developing a decision tree as base learner. Then the second base learner is built to fit the pseudo residuals of the previous base learner. It readjusts weights for misclassified observations at each step and misclassified observations are given higher weight than correctly classified observations. This readjustment is done until it reaches a loss function; in this study it is the Square Loss Function.

The Gradient Boosting model was iterated for 50 times in the series to use in the final model, however, for validation data, misclassification rate flattened out and did not drop any more after 48 iteration (Figure 3). Hence, 50 iteration helped the model to achieve the lowest misclassification value.

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.

CONCLUSION AND FUTURE SCOPE

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.

After the actual motion data and physiological data are collected, the association mining technique will be employed to find out rules between human activities, which are already classified based on our current study, and physiological characteristics, or between human activities and artificial noise. Then, adaptive filtering will be applied to remove noise out of the human physiological signals to get a reliable dataset. Resulting data will be used in future phases of our research and will aid with decision-making. In the specific case of the medical smart garment we are developing, these decisions could relate to medical alerts, intervention, or other caregiving activities.

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

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Minh Pham
Oklahoma State University
Stillwater, OK 74074
Email: minh.pham@okstate.edu
Work Phone: 415-306-4055

Minh Pham is a doctoral student majoring in Electrical and Computer Engineering at Oklahoma State University. He completed his master's in Management Information System in 2012 from Spears School of Business of Oklahoma State University. He gained SAS® and OSU Data Mining Certificate, SAS® certified Predictive Modeler using SAS® Enterprise Miner 6.1., SAS® 9 certified Base Programmer in 2012.all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc.

Mostakim Tanjil
Oklahoma State University
Stillwater, OK 74074
Email: imran.tanjil@okstate.edu
Work Phone: 313-603-1678

Mostakim Tanjil is a master's student in Design, Housing and Merchandising at College of Human Sciences of Oklahoma State University (OSU). He is also pursuing Graduate Certificate in Business Datamining and SAS® and OSU Predictive Analytics Certification from Spears School of Business, OSU. He works as an analyst (graduate assistant) for Center for Health Systems Innovation of OSU. He has a Bachelor of Science in Textile Engineering. Before joining the graduate program, he worked five and a half years for two textile manufacturing industries in Bangladesh as a Senior Engineer. He holds BASE SAS® 9, SAS Certified Statistical Business Analyst Using SAS® 9: Regression and Modeling and SAS® Certified Predictive Modeler Using SAS® Enterprise Miner™ 13 credentials. He presented a paper in AATCC International Conference 2015 and a poster in Analytics Conference 2015.

Mary Ruppert-Stroescu
Oklahoma State University
Stillwater, OK 74074
Email: mary.ruppert-stroescu@okstate.edu

Mary Ruppert-Stroescu, Ph.D. is an Assistant Professor with the Design, Housing, and Merchandising Department in the College of Human Sciences at Oklahoma State University. She is the director of the Center for Wearable Electronic Sensing Systems and Technologies (CWESST) and has developed patent-pending technologies related to textile-based sensing systems.