Predicting Human Activity Sensor Data Using an Auto Neural Model with Stepwise Logistic Regression Inputs

Sudarshan Vennelakanti, Oklahoma State University

Abstract

Due to advances in medical care and rise in living standards, life expectancy, on an average increased to 79 years in US. This resulted in the increase of aging population and increase in demand for development of technologies to aid elderly people to live independently and safely. It is possible to achieve this challenge through Ambient-assisted living (AAL) technologies that help elderly people to live more independently. Much research has been done on Human Activity Recognition (HAR) in the last decade and this research work can be used in development of assistive technologies and HAR is expected to be future technology for e-health systems.

In this research, I discuss the need to predict human activity accurately by building various models in SAS® Enterprise Miner™ 14.1, on a free public domain dataset, which contains 165,633 observations and 19 attributes. Variables used in this research, represent the metrics of accelerometers mounted on waist, left thigh, right arm and right ankle of 4 individuals performing five different activities recorded over a period of eight hours. The target variable predicts human activity such as sitting, sitting down, standing, standing up and walking. Upon comparing different models, the winner is a Auto Neural whose input is taken from Stepwise Logistic Regression, which is used for variable selection. The model has an accuracy of 98.73% and sensitivity of 98.42%

Introduction

The advances in medical care and rise in living standards have increased life expectancy to 79 years in US, in 2015. As per US Census Bureau," In 2050, the population aged 65 and over is projected to be 83.7 million, almost double its estimated population of 43.1 million in 2012 "[3]. Due to these facts, the demand for development of technologies to aid elderly people has increased more than ever. But developing new technologies that enable elderly and the chronically ill to lead a more independent and safer life has become a difficult task. It is possible to achieve this arduous task through Ambient-assisted living (AAL) technologies that help elderly people to live more independently.

Research on HAR has been active for more than a decade and HAR is considered to be future technology for e-health systems. There are prospects of developing assistive technologies to support care of the older adults using research work on HAR. E-Health systems like AAL can be developed by using patients' routine data provided by activity recognition. The most common methods used to recognize human activity are image processing and usage of wearable sensors. Though image processing doesn't require patient to wear any equipment, it has some major problems such as requiring installation of cameras and good light. In addition, its operations are restricted to indoor environments and users have privacy concerns. The use of wearable sensors has addressed all these problems, but requires wearing of equipment by the user for long durations[1]. In this paper, I discuss various models built using SAS® Enterprise Miner™ 14.1, on a free public domain dataset containing 165,633 observations and 19 attributes and compare each model with another . Data used in this discussion, represents the metrics of accelerometers mounted on waist, left thigh, right arm, and right ankle of 4 subjects performing five different activities. These activities, such as sitting-down, standing-up, standing, walking, and sitting, recorded over a period of eight hours of four healthy subjects is used for our analysis.

Data

Data was collected from accelerometers mounted on waist, left thigh, right arm, and right ankle of 4 subjects while they were performing five different activities. This data was collected over a period of 8 hours of activity, each subject performing activities for 2 hours. Each activity was performed separately. This data contains 165633 observations and 19 attributes. The list of all variables is given below.

Variable	**Description**
User	Name of the subject
Gender	Man and Woman; Man is represented as 1 and Woman as 0
Age	Age of the subject
Height	Height of the subject
Weight	Weight of the subject
BMI	Body Mass Index of the subject
x1	X-axis value of the 1st accelerometer, mounted on waist
y1	Y-axis value of the 1st accelerometer, mounted on waist
z1	Z-axis value of the 1st accelerometer, mounted on waist
x2	X-axis value of the 2nd accelerometer, mounted on the left thigh

y2	Y-axis value of the 2nd accelerometer, mounted on the left thigh
z2	Z-axis value of the 2nd accelerometer, mounted on the left thigh
x3	X-axis value of the 3rd accelerometer, mounted on the right ankle
y3	Y-axis value of the 3rd accelerometer, mounted on the right ankle
z3	Z-axis value of the 3rd accelerometer, mounted on the right ankle
x4	X-axis value of the 4th accelerometer, mounted on the right upper-arm
y4	Y-axis value of the 4th accelerometer, mounted on the right upper-arm
z4	Z-axis value of the 4th accelerometer, mounted on the right upper-arm
Class(Target)	sitting = 1, sitting down = 2, standing = 3, standing up = 4, walking = 5

Table 1: Variable Description

Data Preparation

Gender has two values Man and Woman. For the convenience of model building, Man is coded as '1' and Woman as '0'. The target variable has five levels: sitting, sitting down, standing, standing up and walking. These levels of target variable class are coded as follows: sitting as '1', sitting down as '2', standing as '3', standing up as '4' and walking as '5'. The dataset has no missing values. Variables Gender and Class are nominal and rest of the variables are interval.

Descriptive Analytics

User and Class are nominal and rest of the variables in the dataset are interval variables. Debora has the highest number of observations and Jose Carlos has the lowest number of observations. The characteristics of subjects such as gender, age, height, weight and body mass index are shown in the below figure. X-Y-Z values from all accelerometers are of interval type and their distributions in the dataset are shown below. Z4 is slight left-skewed. X2-Y2-Z2 are platykurtotic and rest of the accelerometer variables are almost normally distributed.

user	gender	age	how_tall_in_meters	weight	body_mass_index	N Obs
debora	0	46	1.62	75	28.6	51577
jose_carlos	1	75	1.67	67	24	13161
katia	0	28	1.58	55	22	49797
wallace	1	31	1.71	83	28.4	51098

Figure 1: Characteristics of the subjects

class	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	50631	30.57	50631	30.57
2	11827	7.14	62458	37.71
3	47370	28.60	109828	66.31
4	12415	7.50	122243	73.80
5	43390	26.20	165633	100.00

Figure 2: Frequencies of different activities

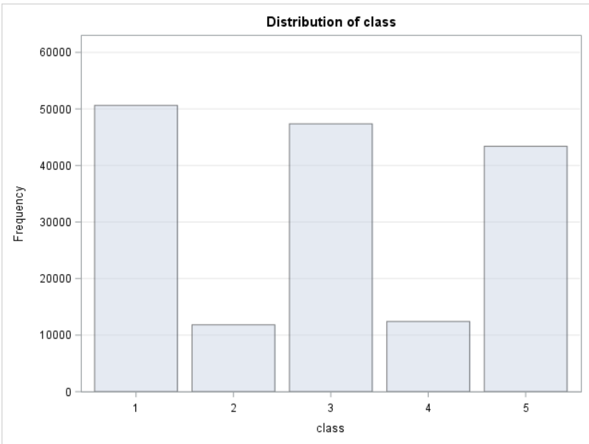

Figure 3: Histogram showing frequency distribution of different activities in data

Based on the accelerometer readings, 30.57% of the observations in the dataset are recorded as sitting, 7.14% as sitting down, 28.60% as standing, 7.50% as standing up and 26.20 as walking. X-Y-Z values from all accelerometers are of interval type and their distributions in the dataset are shown below. Z4 is slight left-skewed. X2-Y2-Z2 are platykurtotic and rest of the accelerometer variables are almost normally distributed.

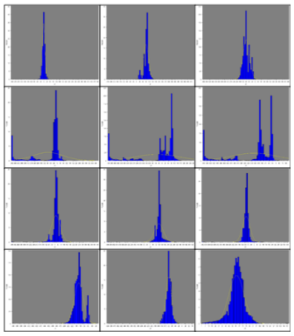

Figure 4: Histograms of Accelerometer variables

Data Partition

Data is partitioned using Simple Random Method. 60% of the data is used for training and 40% for validation..

Predictive Modeling

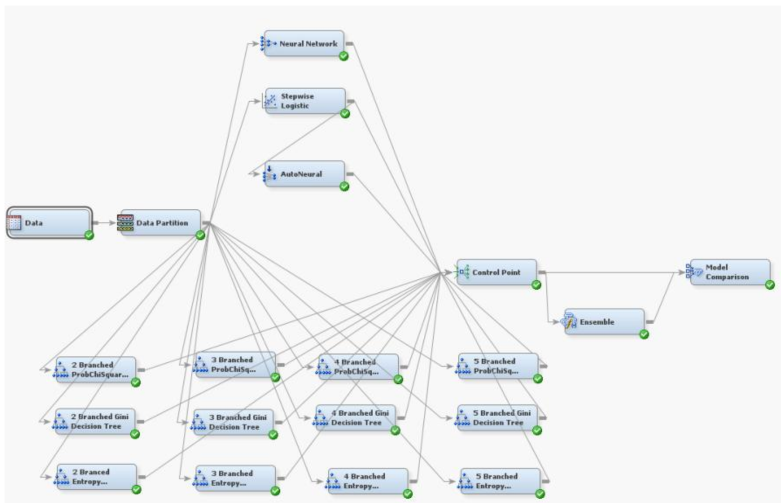

Figure 5: Model Diagram

As shown in Figure 3, Neural Network, Stepwise Logistic Regression, AutoNeural with input from Stepwise Logistic Regression, Decision Tree(Two, Three, Four and Five Branched Decision tree models across all nominal target criteria such as Prob-Chisq, Gini and Entropy) and Ensemble models are built to predict the target variable, 'class'.

Neural Network is built with four hidden layers, Maximum iterations set to 100 and no preliminary training. A Stepwise Logistic Regression is modeled by choosing logistic for regression type, stepwise regression for selection model, validation misclassification for selection criterion and all other default values. The output of stepwise logistic regression is fed to AutoNeural so that only a few selected important variables are given as input, to increase the model performance. AutoNeural is modeled with number of hidden units set to 1, tolerance value set to low, and logistic chosen for activation function. Logistic function is chosen as Activation function, as it is better than other functions at predicting categorical variables. Two, three, four and five branched Decision Tree models are built across all nominal target criteria such as Prob-

Chisq, Gini and Entropy.

A control point node is used to collect output from all models and assign input to Ensemble node and Model Comparison node. Model Comparison node compares models with each other and selects the best of all.

As the objective of this paper is to predict the activity performed by the subject accurately, based on accelerometer readings, Validation Misclassification rate is chosen as the Assessment Measure to compare the performance of different models. After the comparison of models, it is found that AutoNeural has the lowest validation misclassification rate of 0.0126, followed by Ensemble and Five Branched Entropy Decision Tree with misclassification rates of 0.018 and 0.026. Training and Validation Misclassification rates of all models are shown below.

Selected Model	Model Description	Selection Criterion: Valid: Misclassification Rate	Train: Misclassification Rate
Y	AutoNeural	0.012648	0.011431
	Ensemble	0.018248	0.015818
	5 Branched Entropy Decision Tree	0.026127	0.020054
	5 Branched Gini Decision Tree	0.027576	0.021071
	5 Branched ProbChiSq Decision Tree	0.029025	0.022731
	4 Branched Gini Decision Tree	0.029946	0.025075
	4 Branched Entropy Decision Tree	0.030278	0.025941
	4 Branched ProbChiSq Decision Tree	0.032632	0.030097
	3 Branched Gini Decision Tree	0.049869	0.047575
	3 Branched Entropy Decision Tree	0.051032	0.048853
	3 Branched ProbChiSq Decision Tree	0.071514	0.070256
	Neural Network	0.089717	0.091729
	2 Branched Gini Decision Tree	0.123255	0.124079
	2 Branced Entropy Decision Tree	0.135541	0.134544
	2 Branched ProbChiSquare Decision Tree	0.159661	0.159026
	Stepwise Logistic	0.167328	0.169752

Figure 6: Training and Validation Misclassification Rates

Approximately, 98.3% of observations in the validation data are predicted correctly by AutoNeural model, 98.2% by Ensemble model and 97.4% by Five Branched Entropy Decision Tree. One of the reasons for better performance of AutoNeural compared to Neural Network is the input from Stepwise Logistic Regression.

The top three models with lowest misclassification rates are discussed below with their fit statistics and classifications. It is also shown how the number of wrong classifications is calculated in validation data.

Results of AutoNeural Model

Fit Statistics	Statistics Label ▲	Train	Validation
AIC	Akaike's Information Criterion	9019.51	.
AVERR	Average Error Function	0.016123	0.018621
ASE	Average Squared Error	0.003684	0.004034
DFE	Degrees of Freedom for Error	397016	.
DIV	Divisor for ASE	496900	331265
ERR	Error Function	8011.51	6168.346
FPE	Final Prediction Error	0.003693	.
MAX	Maximum Absolute Error	1	1
MSE	Mean Squared Error	0.003689	0.004034
MISC	Misclassification Rate	0.011431	0.012648
DFM	Model Degrees of Freedom	504	.
_NW_	Number of Estimated Weights	504	.
WRONG	Number of Wrong Classifications	1136	838
RASE	Root Average Squared Error	0.060697	0.063512
RFPE	Root Final Prediction Error	0.060774	.
RMSE	Root Mean Squared Error	0.060736	0.063512
SBC	Schwarz's Bayesian Criterion	14509.58	.
SUMW	Sum of Case Weights Times Freq	496900	331265
NOBS	Sum of Frequencies	99380	66253
SSE	Sum of Squared Errors	1830.645	1336.251
DFT	Total Degrees of Freedom	397520	.

Figure 7: Fit Statistics of AutoNeural

Number of wrong classifications = Number of observations × Misclassification rate. Therefore there are 1136(99380×0.011431) wrong classifications in training data and 838(66253×0.012648) in validation data.

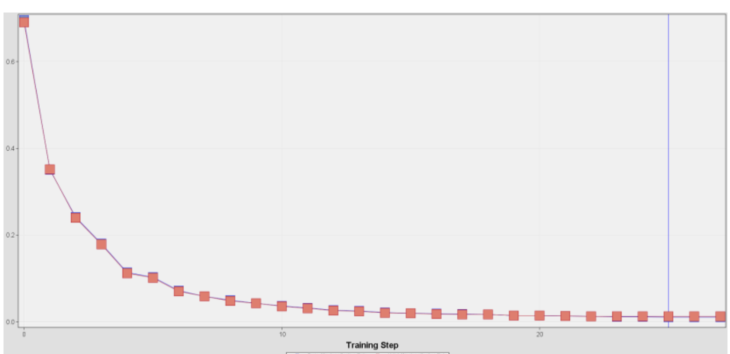

Figure 8: Iteration Plot

As seen in Figure 8, the model is optimized at 25th iteration and has a total of 25 neurons.

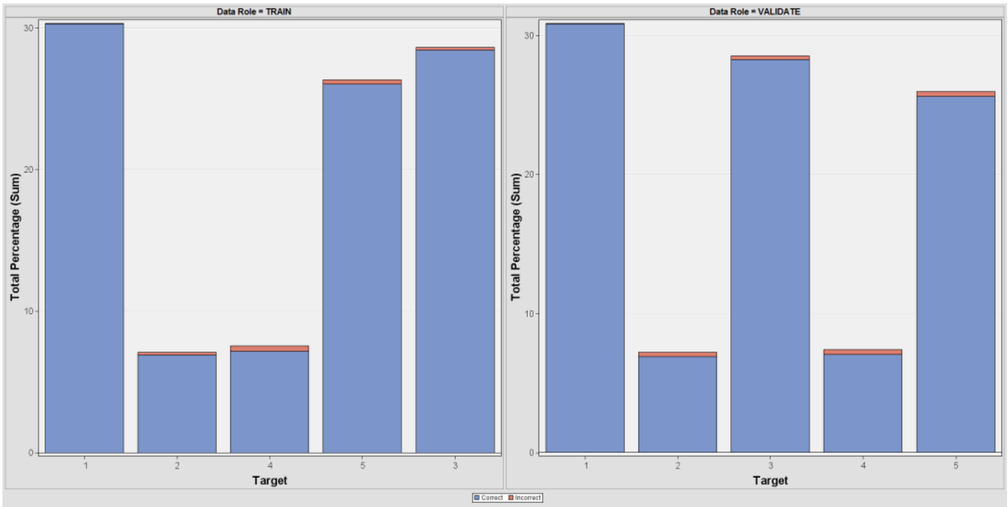

Figure 9: Classification Chart of Auto neural

There are a few incorrect predictions in 2, 3, 4, 5 levels of the target variable as per chart shown above.

Misclassification rate of Ensemble model is 0.016 in training data and 0.018 invalidation data. The number of wrong classifications is 1572 in training data and 1209 in validation data. There are 66253 observations in validation data and the misclassification rate is 0.018248. Multiplying both (66253×0.018248 = 1208.98) gives the number of wrong classifications(≈1209) in validation data.

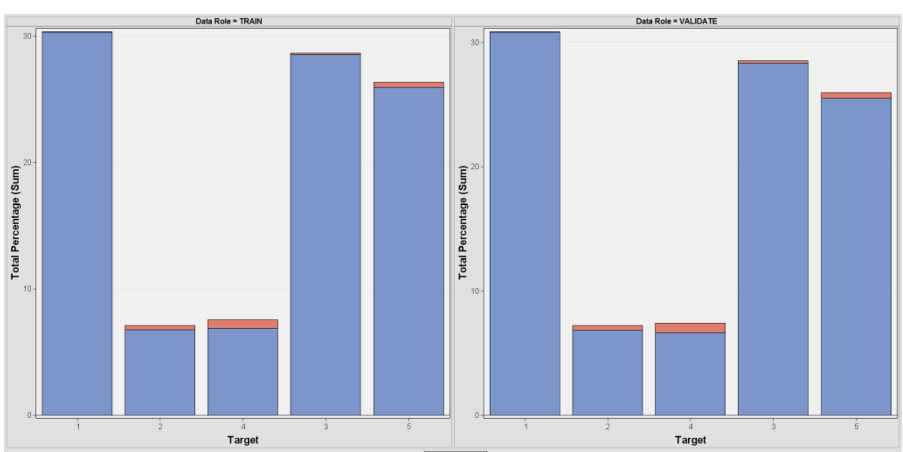

Figure 10: Classification chart of Ensemble model

Misclassification rate of the 5-Branched Entropy Decision Tree model is 0.02 in training data and 0.026 in validation data. The number of wrong classifications is 1993(99380×0.020054) in training data and 1731(66253×0.026127) in validation data.

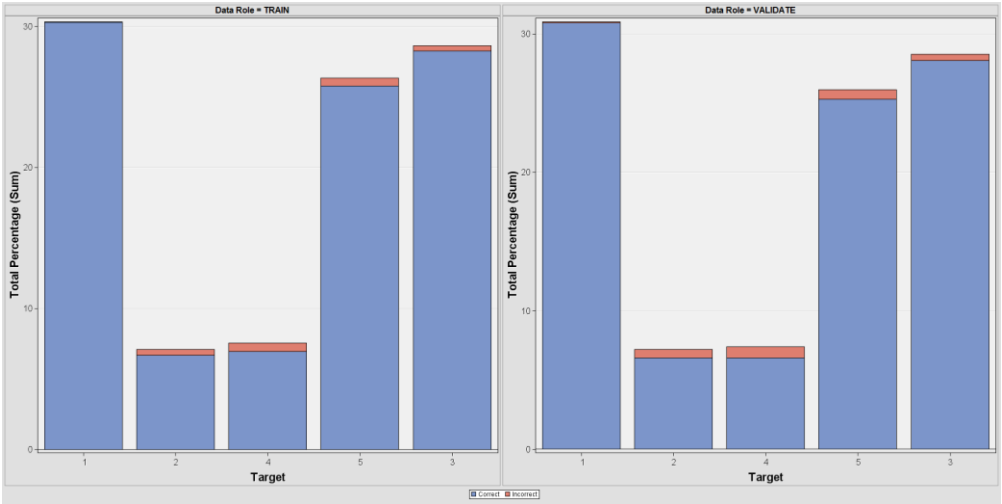

Figure 11: Classification chart of 5-Branched Entropy Decision tree model

Conclusion

Human Activity Recognition has a very wide applications ranging from security-related applications and logistics support to location-based services. Predicting HAR sensor data accurately can help in the development of AAL technologies that help elderly people to live more independently and safely. In this paper, we designed different models to predict human activity recognition and Auto Neural whose input taken from Stepwise Logistic Regression proved to be the best model.

References

1. Ugulino, W.; Cardador, D.; Vega, K.; Velloso, E.; Milidiu, R.; Fuks, H. Wearable Computing: Accelerometers' Data Classification of Body Postures and Movements. Proceedings of 21st Brazilian Symposium on Artificial Intelligence. Advances in Artificial Intelligence - SBIA 2012. In: Lecture Notes in Computer Science. , pp. 52-61. Curitiba, PR: Springer Berlin / Heidelberg, 2012. ISBN 978-3-642-34458-9. DOI: 10.1007/978-3-642-34459-6_6.
   Available at: [Web Link]
2. Yu-Jin, H., Ig-Jae, K., Sang Chul, A., Hyoung-Gon, K.: Activity Recognition Using Wearable Sensors for Elder Care. In: Proceedings of Second International Conference on Future Generation Communication and Networking, FGCN 2008, vol. 2, pp. 302–305 (2008), doi:10.1109/FGCN.2008.165
3. An Aging Nation: The Older Population in the United States Population Estimates and Projections Current Population Reports
   Issued May 2014 P25-1140
   By Jennifer M. Ortman, Victoria A. Velkoff, and Howard Hogan