Instructor: Hrant Davtyan
Course: Business Analytics
Date: 15/05/20

Embry Tech

Final Project

Group Members:
Aspram Grigoryan
Mariam Hakobyan
Svetlana Asatryan
Mariana Ghazaryan

Abstract

Activity recognition using technology is not a new topic for researchers. Among recent literature, more common methods used in this endeavor have ranged from mobile devices, such as smartphones and wearables, as well as more unique methods such as eye movement analysis. Automatic recognition of activities of daily living (ADL) is an essential component in understanding energy balance, quality of life, and other areas of health and well-being. In this paper, we elaborate on human activity recognition tracked by the shoe insoles that have several sensors on them. The project's objective is to find out the most significant features, through transformation and elimination of the non-important features, and build a model with high accuracy to track the human movement. The dataset is provided by an Armenian start-up company, Embry Tech, that has collected the data among several people. The methods used in the model building process are Correlation Analysis, Binary Logistic Regression, CART Decision Tree, and Random Forest. The findings illustrate a better predictive power and accuracy for Random Forest.

**Embry Tech is a hardware & software solution that allows 100% consistency with weight tracking, several other physical activity tracking through connected shoes and its novel technology. The product is a premium insole, which uses a patent pending technology to automatically track body weight trends and physical activity. By using a wide range of sensors, such as pressure sensors, accelerometer sensors, they make predictions about the physical conditions and activity of the person. The main goal of the project is helping people track their daily activities to stay informed about their health conditions and work on making progress.*

Introduction

Daily living activity recognition is one of the most important research topics that can provide valuable information for various applications such as health monitoring, security, intelligent assistance in daily life, fall detection, and surveillance.

Wearable sensors, such as accelerometers, gyroscopes, and magnetometers and several other ubiquitous computing applications, are used to translate human motion into signal patterns for activity recognition. Thanks to the advanced technologies, sensors can be placed in various portable devices such as phones, watches, gadgets, shoes, shoe insoles, and nonportable objects like cars, furniture, and even walls. Even though these devices overcome most of the shortcomings of the external approach, they still have some. One of the disadvantages is that a single wearable cannot cover the entire body and, therefore, may fail to accurately track the body movement. By concentrating on the shoe insole tracking system, we are planning to come up with a model that given the potential drawbacks of tracking a single body part will suggest a competitive accuracy.

Our work was focused on recognizing simple human activities of moving versus not moving. Recognizing complex activities remains a challenging and active area of research especially when the activity is mixed, e.g. walking that includes stairs up movement.We focused on accurate detection of the human activities based on a predefined activity model. Therefore, we tried to build a conceptual model first, and then implement the model by building a suitable system. For that purpose we experimented with different classification algorithms to be able to find a more accurate classifier.

Related Work

Approaches to activity recognition can be divided into two categories: obtrusive and unobtrusive. Obtrusive activity recognition usually includes computer vision or wearable sensors for detecting human activities from the sequence of the user's motion. A considerable number of studies were proposed in the literature for HAR using wearable sensors. Accelerometric signals are very common to all HAR applications. Several studies used that information alone, while more often, accelerometers were combined with gyroscopes and magnetometers. Recent researches have brought the idea of embedding sensors in shoes or insoles for human gait monitoring in academic society. *Noshadi et al*. proposed a smart shoe called "Hermes" for monitoring the person's walking behavior and using an instability assessment model to generate quantitative value with episodes of activity identified as necessary. *Bamberg et al.* developed a wireless wearable system called "GaitShoe" for gait data collection outside the confines of the traditional motion laboratory. It includes accelerometers, gyroscopes, force sensors, bidirectional bend sensors, pressure sensors, and other electric field height sensors. There are several others to mention. The problem is that those systems place the insole-like units inside the shoe, and the micro-controllers or IMU are strapped and mounted outside the shoe, making the system inconvenient to set up and discomfort to use in daily life. For this purpose, another technology has been developed: an in-shoe monitoring system based on a textile fabric sensor array. This system measures the spatial and temporal plantar pressure distributions.

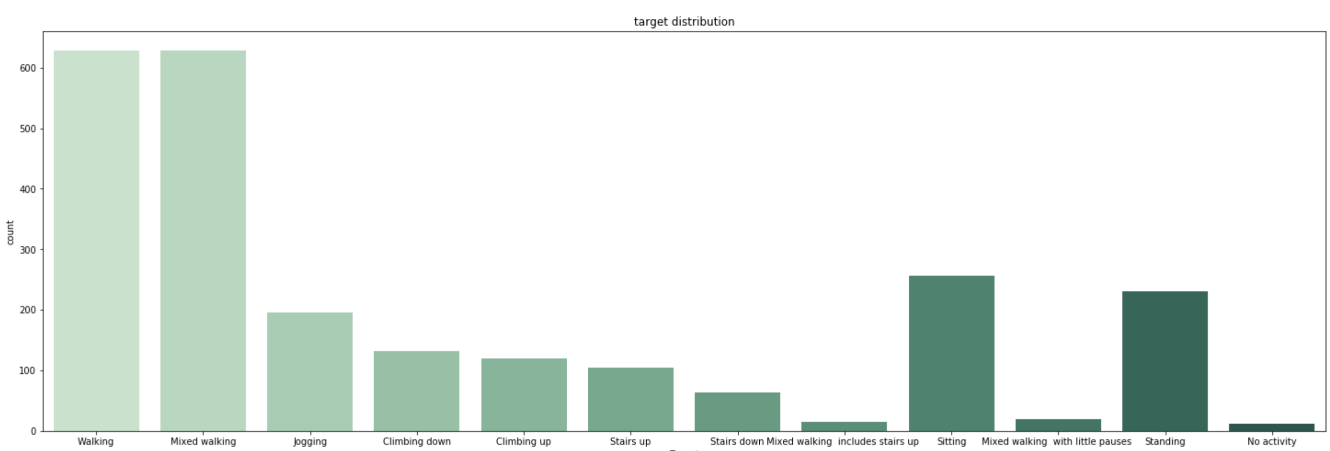
Once the sensors acquire the data, a kind of processing is usually applied to extract a set of informative features. In general, most extracted features belong to the time-domain, such as mean, standard deviation, minimum value, maximum value, range, and the frequency-domain (e.g., mean and median frequency, spectral entropy, signal power, entropy). We can find several variables in literature, such as time-frequency domain variables or the cepstral features proposed by *San-Segundo et al*.

Regarding the dimensions of the obtained feature sets, three different approaches were presented in the literature. The first one is no dimensionality reduction, where the whole set of variables was used for the recognition phase. The second approach achieves dimensionality reduction by transforming the original set of variables into lower dimensionality ones. Last but not least, the most common method is principal component analysis (PCA). Finally, several other methods were used to reduce the number of variables without any transformation: Minimum Redundancy Maximum Relevance, Information Gain, recursive feature elimination, or evolutionary algorithms.

Because the goal of a HAR application is to identify the performed human activity, in the final step, a proper learning algorithm should be applied. The vast majority of the studies were based on supervised learning algorithms, like decision trees, random forest, support vector machine, multilayer perceptron, as well as neural networks.

Dataset and Features

8 participants with different gender, age, height, and weight were recruited to perform 12 types of activities: walking, mixed walking, jogging, climbing down, climbing up, stairs up, stairs down, mixed walking that included stairs up movement, mixed walking with little pauses, standing, sitting and no activity at all. These activities correspond to the most common activities in people's daily life and are useful for many applications for Embry Tech. Then the data was divided into fixed ten second length windows.



The dataset had 6 core features - aggregated accelaration and psressure measures for tyhree different locations each. The dataset also had some interesting statistical features computed based on 10 sec. Intervals. Some commonly used

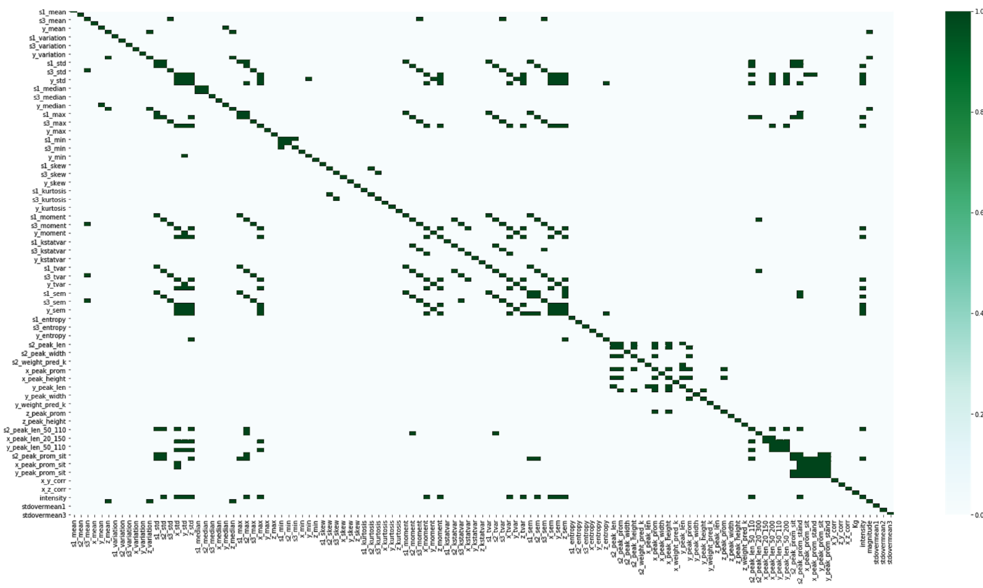
The dataset consisted of pressure and acceleration features that were captured by the sensors inside the insoles. The data were captured in three different locations inside the shoe insole resulting in effective vector data. For the purposes of feature engineering , first physical features were created. These were derived based on the physical interpretations of human motion. Specifically, two measures that are known to be significant predictors in HAR were computed.

$MI(t) = \sqrt{a_x(t)^2 + a_y(t)^2 + a_z(t)^2}$ - *Movement Intensity (MI). This feature is independent of the orientation of the sensing device, and measures the instantaneous intensity of human movements at index t.*

$SMA = \frac{1}{T}\left(\sum_{t=1}^{T} |a_x(t)| + \sum_{t=1}^{T} |a_y(t)| + \sum_{t=1}^{T} |a_z(t)| \right)$ - *Normalized Signal Magnitude Area (SMA). The acceleration magnitude summed over three axes within each window normalized by the window length. This feature has been used in previous studies and is regarded as an indirect estimation of energy expenditure.*

Secondly, From this mechanical motion measurements different statistical parameters were computed. Through feature transformation a total of 122 features were gathered. Some commonly used features such as mean, variance, correlation and entropy were used. A few statistical measures were added such as mean over st.deviation of pressure. Some of them have been intensively investigated in previous studies and proved to be useful for activity recognition . For example, variance has been proved to achieve consistently high accuracy to differentiate activities such as walking, jogging, and hopping. On the other hand, Correlation between each pair of sensor axes helps differentiate activities that involve movement in a single dimension such as walking and running from the ones that involve a movement in multi-dimension such as stair climbing.

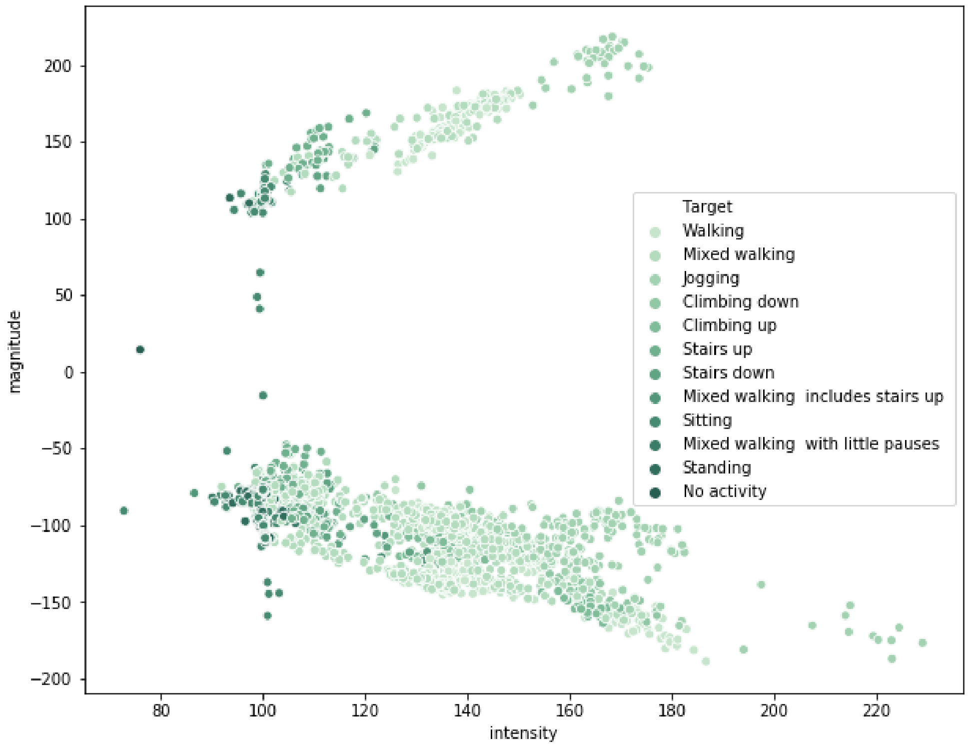
The problem, however, was that some of the features were irrelevant and redundant, showing high correlation with other features. These features were not adding new information to improve the classification accuracy, as well as the high dimensionality was confusing the classifiers rather than helping them. The image below shows the pairwise correlation of features. Dark greens represent correlation of more than 0.75.



To achieve the best classification performance the dimensionality was reduced. A correlation-based feature selection method was used to select a subset of features. All the variables that had a correlation of more than 0.75 were excluded from the set. Only complementary features were kept. This is important not only in terms of more accurate prediction, but also it may result in less computational costs for the company.

Methodological Approach

Accurate activity recognition is challenging because human activity is complex and highly diverse. Literature survey performed in this area has revealed data mining algorithms are employed for classification of activities. Hybrid mining techniques, Naive Bayes with SVM and C4.5 with Neural Network are proved to be efficient in classifying the accelerometers and pressure data. In this paper, we have used three different machine learning algorithms to classify the target variable. For that purpose we have splitted the target into two categories that define whether the human wearing the insole is moving or not. The categories of sitting, standing and no activity were combined to represent no movement and the rest to represent some movement, e.g. walking, mixed walking. As we can see from the scatterplot below, no-movement activities that are colored in darker green seem to have clustered in the beginning of these two dot streams.



As the data set suffered from high-dimensionality it was divided into training and testing sets for validating the model accuracy on test data as well. The splitting was done manually to avoid random splits that would include the same person both in train and test.

We have used multinomial logistic regression algorithms that might be useful not only for prediction but also for identifying important features and eliminating unimportant ones. This approach may remarkably reduce the dimension and maintain the most discriminative information. Dummies for persons recruited for the experiment were used to identify whether there is a person specific component in target activities.

Secondly, a CART decision tree and Random forest were used. RF is known for achieving high accuracy in classification. It's robustness in classifying large data sets is promising.

Experimental studies are conducted to study the activity recognition capability of the models, the results are compared with popular supervised classification techniques. It is observed that the proposed RF model outperformed the other classification techniques in comparative study.

Experiments and results

Preliminary logistic regression showed a big number of insignificant features. Using backward elimination technique and closely following the predictive power of the model we removed some of the variables.

We have optimized all three models through hyperparameter optimization and finding the best parameters for all three models.

Logistic Regression Train:

	precision	recall	f1-score	support
0	0.99	0.99	0.99	1729
1	0.96	0.97	0.96	462
accuracy			0.98	2191
macro avg	0.97	0.98	0.98	2191
weighted avg	0.99	0.98	0.98	2191

Logistic Regression Test:

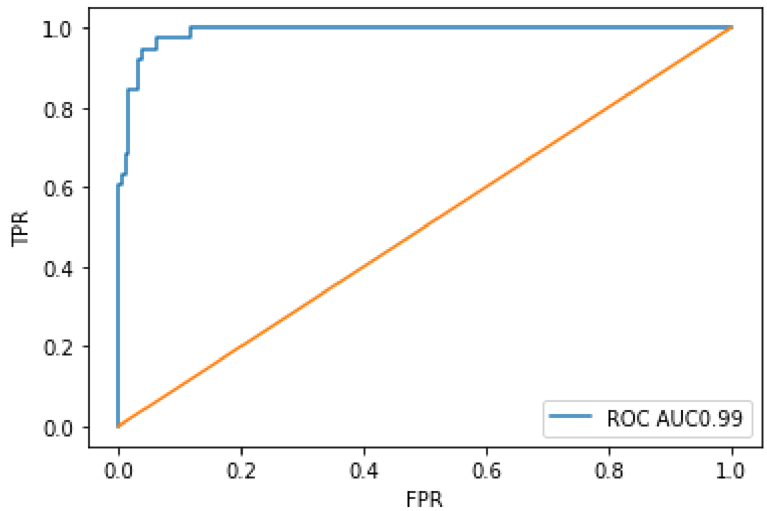
	precision	recall	f1-score	support
0	0.98	0.96	0.97	179
1	0.81	0.92	0.86	38
accuracy			0.95	217
macro avg	0.90	0.94	0.92	217
weighted avg	0.95	0.95	0.95	217

Going beyond the accuracy of the model we see that for the test data Logistic Regression had a precision of 0.81. for positive cases (not moving). Precision is defined as the number of true positives divided by the number of true positives plus the number of false positives. False positives are cases the model incorrectly labels as positive that are actually negative, or in our example, individuals the model classifies as not moving(1) while they are moving (0). While recall expresses the ability to find all relevant instances in a dataset, precision expresses the proportion of the data points our model says was relevant actually were relevant.

As a more accurate model was our primary objective we have relied mostly on a more complex model of Random Forest that gave a good-enough result in terms of its accuracy to predict the occurrence of an event on unseen data. However, an issue arises because as the accuracy increases so does the model interpretability.

```
Mean 5-fold ROC AUC score for Tuned DT 0.96
Mean 5-fold ROC AUC score for Tuned RF 0.98
Mean 5-fold ROC AUC score for Tuned Logit 0.97
```

A ROC curve was constructed for the winning model by plotting the true positive rate (TPR) against the false positive rate (FPR).



As we have a probabilistic classifier, it gives a probability or score that reflects the degree to which an instance belongs to one class rather than another, we created a curve by varying the threshold for the score. Our classifier gave a curve closer to the top-left corner indicating a better performance. As a baseline, a random classifier is expected to give points lying along the diagonal (FPR = TPR). As our model also has AUC near to 1 it means it has good measure of separability.

Conclusion

Through feature engineering, transformation and important feature extraction <u>we reduced the dimensionality of the dataset that helped to construct a more accurate model and less computational costs for the company.</u> We found that for the binary classification task with high dimensional data Random Forest outperformed Logistic Regression and Decision Tree.

References:

- A review on applications of activity recognition systems with regard to performance and evaluation - Suneth Ranasinghe, Fadi Al Machot, Heinrich C Mayr, 2016. (2020). *International Journal of Distributed Sensor Networks*, [online] p. Available at: https://journals.sagepub.com/doi/full/10.1177/1550147716665520 [Accessed 15 May 2020].
- Bose R, Helal A. Observing Walking Behavior of Humans using Distributed Phenomenon Detection and Tracking Mechanisms. Proceedings of the International Workshop on Practical Applications of Sensor Networks; 2008. [Google Scholar]
- Gu T, Wu Z, Tao X, Pung HK, Lu J. epSICAR: An Emerging Patterns based approach to sequential, interleaved and Concurrent Activity Recognition. Proceedings of the IEEE International Conference on Pervasive Computing and Communications; 2009. [Google Scholar]
- Deep Learning for Sensor-based Human Activity Recognition: Overview, Challenges and Opportunities https://arxiv.org/pdf/2001.07416.pdf