Alex Teboul
Paper Review 1

DSC 540: Advanced Machine Learning
Professor: Casey Bennett

Paper: Physical activity classification in free-living conditions using smartphone accelerometer data and exploration of predicted results
Authors: Kangjae Leea, Mei-Po Kwanb

Review

Summary

Researchers Kangjee Lee and Mei-Po Kwan at the University of Illinois at Urbana-Champaign explored the application of the Random Forest (RF) and Gradient Boosting (GB) algorithms towards physical activity classification in free-living conditions using smartphone accelerometer data. They argued that the ability to accurately identify the types of physical activity individuals are engaged in based on smartphone accelerometer data is important for researchers studying the relationship between physical activity and health outcomes. The prevalence of smartphones in the U.S., where this research was conducted, provides a low-cost, easily accessible population for which this technique could be applied. In this study, free-living is defined as outside of laboratory conditions.

Overall, the authors claim that their RF and Gradient Boosting models successfully classified free-living physical activities and that this technique is valuable. They concede that transportation, the means of carrying and using one's smartphone, and the type of smartphone accelerometer can lead to poor classification using this technique. I argue that their data collection process and presentation of results is problematic and potentially misleading. That said, this research should be followed up with a more robust study. Any future research should provide greater clarity regarding the methods used to evaluate the performance of the classification models on free-living conditions and means by which accelerometer data is labelled by physical activity type.

Data

The researchers used two separate datasets in their study - WISDM for training their models and data they collected on two subjects for their free-living study. The first dataset, the publicly available WISDM accelerometer dataset, is composed of roughly 1 million recorded x, y, and z accelerations. These measurements were taken for 36 subjects using Android smartphones that were in the subject's front pockets. All 1 million data points have associated labels for the type of physical activity the subjects were engaged in - either walking, jogging, upstairs, downstairs, sitting, or standing. To process this time-series accelerometer data a low-pass filter was used to reduce noise and a moving window approach was taken on every 200 samples to determine examples with features. With 50% overlap, this approach means that every 1000 points returned 9 examples with calculated features. Overall, they used 59 features including magnitude of total acceleration, min-max mean, and mean dominant frequency.

The second dataset used was collected by the researchers themselves in order to apply the classification algorithms to smartphone accelerometer data in a free-living context. Overall they had one used subject's data which had about 49000 samples. The means by which this data was collected and how subjects were chosen to be included in the study are somewhat ambiguous and potentially problematic. They recorded accelerometer data using the same processing techniques as in the WISDM dataset, but this time subjects went about daily life with their smartphones recording accelerations and GPS locations being logged for each recorded instance. The authors were very unclear about how they labelled their data in terms of jogging, walking, sitting, and standing physical activity types. They ended up calculating accuracies and confusion matrices for this dataset, but it seems they also came up with the labels visually based on the GPS data associated with records. This could create problems in terms of the replicability of this study and call into question the researchers claims of successful classification. It is important to note that for this study, all upstairs/downstairs

Machine Learning Methods

To review, the WISDM dataset is a publicly available smartphone accelerometer dataset with physical activity labels that was created from in-lab experiments on subjects who were asked to perform 6 different physical activities. The researchers used this dataset to train their Random Forest and Gradient Boosting models for predicting physical activity type. Random Forest is a machine learning algorithm that takes multiple decision trees and predicts class labels based on the votes of all the trees taken as an ensemble. Subsets of features are chosen for each tree and best split is applied based on these subsets. Gradient boosting utilizes weak classifies to learn over multiple iterations with the use of regression tree models. Loss function minimization is the means by which a

stronger model is built. The authors don't describe the risk of overfitting, but these techniques do present some challenges if regularization isn't properly applied. The authors used R to train the RF ('randomForest') and GB ('xgboost'), models on the WISDM dataset. They presented confusion matrices and accuracies for the model performance on this dataset and combined upstairs/downstairs classes with walking. They then took these trained models and applied them to the dataset they collected on the single *valid* subject's accelerometer data. From here they presented the same confusion matrices that gave very positive results.

Concerns

The way the authors of this study present their findings makes the results seem very compelling, but I have a few concerns. The authors claim that their RF and GB models achieved up to 99% accuracy in the WISDM dataset and over 95% in their free-living conditions. I have 3 main points of concern with these results - which seem a bit too good to be true. Specifically, the authors use of a single subject for evaluating their models, the means by which physical activity labels were generated for the free-living context, and the exclusion of *bad data* from their free-living performance evaluation.

First, the use of a single subject in this trial is problematic. It is not problematic because of the quantity of data collected for use in modeling or evaluation. Rather it is problematic because they had actually included other subjects but selectively removed the results of these free-living trials because of poor performance of their models and messy accelerometer data. It's hard to claim a method of classifying physical activity using smartphones can achieve 95% accuracy when 2 out of every 3 study participants had to be dropped because of the way a smartphone was carried or that the accelerometer was different. To better support their findings, the authors could have recruited another participant, or at least detailed more clearly the results in the failed subject cases.

Second, the researchers are not clear on how they generated their *Actual Class* physical activity type labels for the data they collected. While they claim to have +95% accuracies and perfect sitting vs. standing classification, it appears they are labelling these cases themselves. Based on GPS location they seem to be making determinations about what that actual class of physical activity is. With respect to sitting and standing, I would also like an explanation of how the accelerometer is picking up on that, because there's no acceleration in either case, I imagine the orientation of the phone could mess with this classification.

Third, if the researchers are also saying that using the phone, being indoors, and being in a moving vehicle throw off the results, why isn't this reflected in the evaluation performance metrics they present? These seem awfully high given all the caveats the authors present in terms of cases where their classification techniques breakdown. It's not a true free-living study if the contexts in which classification fails are dismissed.

Conclusion

The authors aimed to describe an approach to physical activity classification in free-living conditions for physical activity research using smartphone data. They claimed incredible success with this method by presenting high classification accuracies for Random Forest and Gradient Boosting models. That said, there are problematic choices made in the manner of presenting these findings. First, they used a single subject to collect data on and removed from the study the two others which gave poor results. Second, they likely labelled the accelerometer data they collected using GPS data points which could be inaccurate. I say 'likely' because they don't actually explain how they labelled those data points exactly to evaluate performance. Finally, the selective exclusion of particular data points that were improperly classified such as when a smartphone was in a jacket pocket instead of a pants pocket, individual was indoors or using transportation, or a smartphone was in use, was problematic. I think this research has potential, I would just like to see a more robust research paper exploring the topic and addressing the concerns outlined above.