Bag-of-Words Method Applied to Accelerometer Measurements for the Purpose of Classification and Energy Estimation

Kevin M. Amaral
PhD. Student
Computer Science
University of Massachusetts Boston

Ping Chen, PhD.
Associate Professor
Computer Science and Engineering
University of Massachusetts Boston

Scott Crouter, PhD.
Assistant Professor
Exercise Physiology
University of Tennessee-Knoxville

Wei Ding, PhD.
Associate Professor
Computer Science
University of Massachusetts Boston

April 2017

1 Abstract

Accelerometer measurements are the prime type of sensor information most think of when seeking to measure physical activity. On the market, there are many fitness measuring devices which aim to track calories burned and steps counted through the use of accelerometers. These measurements, though good enough for the average consumer, are noisy and unreliable in terms of the precision of measurement needed in a scientific setting. The contribution of this paper is an innovative and highly accurate regression method which uses an intermediary two-stage classification step to better direct the regression of energy expenditure values from accelerometer counts.

We show that through an additional unsupervised layer of intermediate feature construction, we can leverage latent patterns within accelerometer counts to provide better grounds for activity classification than expert-constructed time-series features. For this, our approach utilizes a mathematical model originating in natural language processing, the bag-of-words model, that has in the past years been appearing in diverse disciplines outside of the natural language processing field such as image processing. Further emphasizing the natural language connection to stochastics, we use a gaussian mixture model to learn the dictionary upon which the bag-of-words model is built. Moreover, we show that with the addition of these features, we're able to improve regression root mean-

squared error of energy expenditure by approximately 1.4 units over existing state-of-the-art methods.

2 Introduction

2.1 Background and Related Work

In 2005, Crouter et. al. introduce the two-regression model which alternates between a quadratic regression model and a linear regression model based on the coefficients of variations of each bout. [3] This novel approach broke the overall problem objective into two key parts: first, separating instances by their variability into two groupings based on their coefficients of variation; second, applying to each grouping a regression model which is more appropriate for instances of that variability.

In 2012, Trost et. al. was able to improve physical activity classification accuracy as well as low root mean squared-error (RMSE) in energy expenditure estimation with an Artificial Neural Network (ANN) model. [7]

In that same year, Mu et. al., revisited the two-regression model of Crouter et. al. and extended it to a number of regression models, one per each activity type. [5] The data used in this study including each activity bout therein was structured rather variably, which made it analogous to a free-living data collection. This method utilized distance metric learning methods to learn the underlying block structure of variable-length activity bouts.

In 2014, Staudenmayer et. al. expanded on the field with another ANN model which they applied to their own dataset. [6] However, their classification procedure was targeting learned activity types, as opposed to expert-defined types. They produced these types through clustering based on their signal activity levels.

In 2015, Montoye et. al. did an analysis of accelerometer placement for the purpose of energy expenditure estimation and found in their results that the thigh-mounted accelerometer produced the most accurate measurements of all considered mount-points. [4]

Bastion et. al. published an evaluation of cutting-edge methods outside of the rigid laboratory setting and confirmed the activity classification community's suspicions that existing methods would not perform well in the free-living setting. [1]

3 Methods

For our experiments, we utilized a subset of the dataset used in [5] whose activities most-closely resembled those of [7]. In total, one hundred and eighty-four (184) child participants' data were used. For each of these participants, one bout of lying resting for up to thirty (30) minutes with a median time of seventeen (17) minutes. All other activities were performed for up to twelve (12) minutes with a median time of four (4) minutes.

Sedentary	Lying Rest
	Playing Computer Games
	Reading
Light Household and Games	Light Cleaning
	Sweeping
	Workout Video
Moderate-Vigorous Household and Sports	Wall Ball
	Playing Catch
Walk	Brisk Track Walking
	Slow Track Walking
	Walking Course
Run	Track Running

Table 1: Activity Classes and the types of activities performed within them.

In Table 1, we list the types of activities which have been included in our experiment. In the left column, we have the activity classes of which each activity bout in the dataset only corresponds to one. On the right, are the more specific activity types that each class consists of.

In Table 2, the number of bouts associated with each activity class is listed, as well as the total number of intervals in each class.

	Sed.	**LHH**	**MtV**	**Walk**	**Run**
Bouts	259	116	79	150	23
Intervals	16475	2505	1570	3775	485

Table 2: Dataset Summary

3.1 Classification Model

Artificial neural networks are the state of the art method for activity classification. Trost et. al. used in their work a feed-forward neural network with a single hidden layer to predict MET values directly [7]. Staudenmayer et. al. also use an artificial neural network as their model [6], in the first step to predict the physical activity type and then afterwards separately to predict the MET values. Our method improves on the ideas of the two-regression framework by expanding the number of regression models to one per each activity type and leverages the model from Staudenmayer's and Trost's works as our framework's final classification component.

Our classification model is a three-stage framework which consists of a clustering phase over the activity windows, a bag-of-words construction phase for each unique activity bout, and a neural network classification phase over the new bag of words features.

The Bag-of-Words structure within the framework brings with it its representation power from the field of Natural Language Processing, and integrating it into our activity classification and energy expenditure estimation framework bridges the gap between these two disciplines.

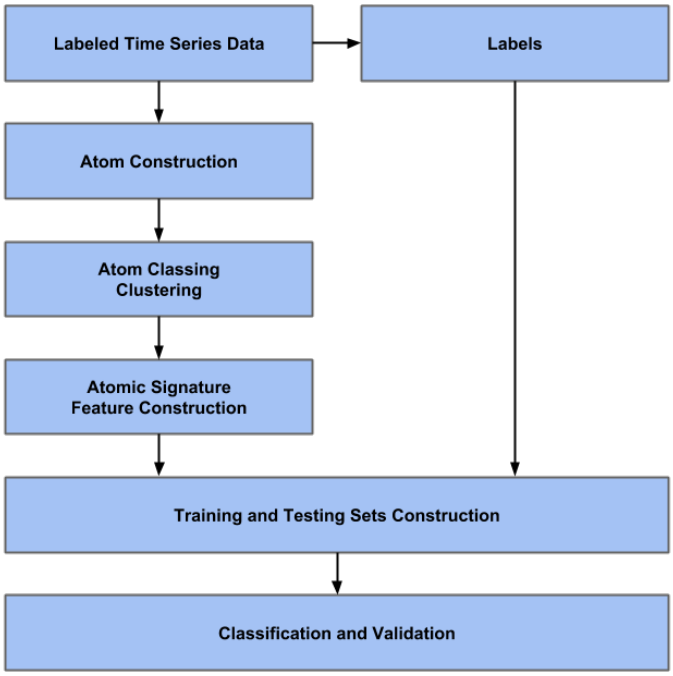

Figure 1: Classification Framework Diagram

3.1.1 Time Series Features - Atom Construction

The standard feature construction seen in the state of the art for physical activity time series has been to use percentile features. These features represent the time series signal by their moments, more specifically their 10th, 25th, 50th, 75th, and 90th percentiles. [6] [7]. We will start with these features and include lag-k autocorrelation features in our initial feature construction phase.

For the instance construction, we favor segmentation of the signal into 12-second blocks. This allows us to best manage the inconsistent bout lengths in our dataset. In the work of Staudenmayer et. al., their activity bouts were all of a fixed-length of 10 minutes. Trost et. al.'s dataset considered for fixed-length bouts of 2 minutes, however they opted to increase their data resolution by further partitioning their bouts into 10, 15, 20, 30, and 60 seconds for their experiments. In line with Trost et. al.'s work, our narrower windows allow us

to tightly capture the volatile patterns of child physical activity, whereas wider windows would be too long.

In either case, those works have a data setting which we do not have: fixed-length activity bouts. In later subsections of Section 3, we show how we're able to overcome this challenge within our dataset through our intermediate feature construction phase to produce a fixed-length data instances.

3.1.2 Clustering Phase - Atom Classing

In Staudenmayer et. al., they cluster activity instances based on their signals to produce their activity classes. [6] However, we do not use the clustering phase alone to determine physical activity classes. We reject the idea that classifying the signal alone will give us the true activity class. Each performance of an activity differs and any one moment spent idle or performing the activity in a non-standard way will greatly affect the signal as a whole.

Again, we consider brief windows of 12-seconds to be characteristic of atomic micro-performances within an activity. This is in line with previous work in Mu et. al. [5] in which we considered the block structure of the timeseries signal. By contrast, however, we are considering much smaller blocks which are expected to lose their homogeneity with the rest of the signal, as opposed to the 1-minute windows used in that paper. One minute of an activity may look like any other minute of the same activity but as we choose this finer resolution, each 12-second window will be more distinguished from other block units in the same activity. By considering these very brief local acts, we can better classify the activity as a whole. Henceforth, we will call these micro-performances "atoms" as they represent our smallest considered unit of activity.

These atoms' types must be learned latently. While we have some high-level idea of what types of atoms we should be able to find within the accelerometer measurements, such as jumping, taking a step, climbing a stair, etc., it is not clear which atoms best describe the space of accelerometer counts over the types of activities we're considering. It's expected that the truly descriptive atom classes are abstract spatial patterns which we can not define empirically.

As such, we seek to identify these atom classes through clustering. The model we've chosen to represent our clusters is a Gaussian Mixture model. In statistics, mixture models best model a distribution for which there are distinct subdistributions which constitute the whole probability space [2]. In a Gaussian mixture, we assume that each and every subpopulation in our sample distribution is modelled at least approximately by a multivariate Gaussian.

Using variational Bayesian methods to approximate the Guassian mixture over our sample, we learn the most-likely set of subpopulations or clusters that our 12-second windows fall into. The associated distributions of these subpopulations are each associated with a unique atom class.

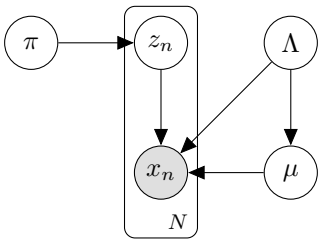

Figure 2: Graphical representation of the Bayesian mixture of Gaussians model

3.1.3 Bag of Words Phase - Latent Feature Construction

A bag is a mathematical structure likened to that of the mathematical set except that it allows duplicate elements. It has a higher representation power than the set in that each element in the bag has an associated count or frequency. When used as a collection of words, it represents word frequency within a sentence or document in natural language processing and has found use in other fields of machine learning. We seek to apply the bag-of-words model to our context in the following way.

With each 12-second window of an activity bout assigned to a unique cluster, we now have a basis for constructing an activity signature for the entire bout. If we take each atom class to be a word, each bout can be seen as a sentence composed of words. Intuitively, the idea follows from the concept of the atoms in and of themselves, which are short meaningful chunks of the whole activity. Looking at them independent from the activity doesn't give us any indication of what activity is being performed. This is the same relationship between words and sentences.

For the bag-of-words construction, we must select a dictionary. The dictionary defines which words may appear in the bag structure. We must also determine if we are going to use word counts or word frequency. Not all activities are of the same length in the same way that not all sentences are of the same length. As such, we will use a bag-of-words model over term frequency.

The Guassian mixture acts as our dictionary where each cluster is a word. For any one activity bout, the frequency associated with each word in its bag-of-words is the ratio of how many atoms in the bout belong to that word's associated cluster. In the bag's representation, we include words with zero frequency. This along with the fixed-length dictionary allows us to build fixed-length vector representations of the bag-of-words model.

This allows us to leverage technologies we weren't able to before. Before this step, we had many activity bouts of various lengths; each instance in our data would have variably many observations and would not reside in a fixed-dimensional vector space. As a result, we were not able to use classification models which depended on those conditions directly.

However, now we have produced exactly those conditions. At the end of this phase, we have converted our data to a fixed-dimensional vector space with

exactly one bag-of-word vector associated with each bout.

3.1.4 Classification on Latent Features Phase

We apply the same feed-forward neural network model used in previous works to the newly constructed bag-of-word features.

The neural network is applied to the bag-of-word vector with each of its components being nodes in the input layer. For the hidden layer, we use an affine layer of 25 nodes. This layer includes a bias vector $\vec{b}$ whose components b_i function as the bias terms for each node in the layer. The layer also includes a weight matrix W whose components $w_{i,j}$ function as the weights associated with the edges in the neural network. For the hidden layer's activation function, the hyperbolic tangent function was used.

For the output layer, we use softmax to give us a categorical output value from the neural network. This output is our model's class prediction.

The model was trained using the MATLAB Neural Network Toolbox which uses the Levenberg-Marquardt algorithm for training neural networks.

3.2 Regression Model

In our regression stage, we learn a least-squares linear regression model. As inputs to the model, we use the atoms with their original time-series features, as well as the activity class prediction of the bout it belongs to from the Classification Stage. This is a carry-over from previous work in which we justified that including the class prediction of an activity increases the accuracy of estimation of energy expenditure. We also include as input the Bag-of-Words features associated with the bout each atom belongs to. The reasoning behind this is that Bag-of-Words features carry higher-level knowledge of the moments surrounding each atom which thereby improves the energy expenditure estimation for each atom in the bout.

As output from this model, we get a low-error estimation of MET. We can then aggregate the MET predictions over all atoms in a bout to get the energy expenditure of the bout as a whole.

4 Results

As can be seen in Table 4, during the classification stage we manage to get competitive accuracies with our competing methods. However, the major success in our model comes from our acceptable classification misses: run-class atoms are never classified as sedentary-class atoms. Table 3 shows that relatively few misclassifications happen between distant classes when they happen at all. This implies that the only misclassifications that occur are happening at boundary or outlier cases.

As for our regression results, we provide the standard least-squares linear regression model on the data by itself as a baseline. This emphasizes the effect

	Sed.	**LHH**	**MtV**	**Walk**	**Run**
Sed.	**16455**	20	0	0	0
LHH	90	**2085**	310	20	0
MtV	20	240	**1290**	20	0
Walk	25	0	25	**3685**	40
Run	0	0	0	135	**350**

Table 3: Confusion Matrix as atoms for Classification Phase

	Sed.	**LHH**	**MtV**	**Walk**	**Run**
Sed.	**99.88**	0.12	0.00	0.00	0.00
LHH	3.59	**83.23**	12.38	0.80	0.00
MtV	1.27	15.29	**82.17**	1.27	0.00
Walk	0.66	0.00	0.66	**97.62**	1.06
Run	0.00	0.00	0.00	27.84	**72.16**

Table 4: Confusion Matrix as percentages for Classification Phase

of each aspect of our regression stage. Table 5 shows that each additional element to the model has a significant impact on overall RMSE. In fact, simply by including the activity class prediction, we beat the state of the art's regression model in RMSE by nearly 0.5 units. Including our Bag-of-Word features increased accuracy by another 0.1 units.

From Table 6, we can see that our model has an in-class RMSE advantage over all other methods. We do not simply beat the methods overall, but in regressing any type of activity, our model performs with the least error. Especially significant is the 1.1 unit decrease in RMSE from the ANN in the Run-class and the 1.2 unit decrease in RMSE from the ANN in the Moderate-to-Vigorous-class. These activity classes contain the highest error rates for all models as they are the most difficult to estimate. Our model improves greatly on these difficult classes.

	RMSE
Linear Regression on Raw Features	2.3690
Linear Regression w/ Class Prediction	0.9548
Linear Regression w/ Class Prediction and BoW Features	0.8502
Artificial Neural Network	1.4402

Table 5: Root Mean Squared Error for MET estimates of each Regression Model

	Sed.	LHH	MtV	Walk	Run
LR-RF	2.0105	2.7549	3.3990	2.6670	4.1593
LR+CP	0.2284	1.4094	1.8695	1.4612	2.5591
LR+CP+BoWF	0.1798	1.3477	1.5964	1.3494	2.0146
ANN	0.3999	2.2715	2.7789	2.1308	3.6695

Table 6: RMSE for MET estimates of each Regression Model by Activity Class

5 Conclusion

We presented a Classification-Regression framework for predicting activity classes and estimating energy expenditure from time-series data collected from hip mounted accelerometers. Our approach of utilizing an unsupervised intermediate feature construction layer has been shown to generate meaningful and useful knowledge that contributes to high classification accuracy and lower regression error. Integrating the Bag-of-Words model into our representation was shown to have a significant impact on our results over those of other methods. We further show that our results exceed that of the state-of-the-art method.

References

[1] Thomas Bastian, Aurélia Maire, Julien Dugas, Abbas Ataya, Clément Villars, Florence Gris, Emilie Perrin, Yanis Caritu, Maeva Doron, Stéphane Blanc, Pierre Jallon, and Chantal Simon. Automatic identification of physical activity types and sedentary behaviors from triaxial accelerometer: laboratory-based calibrations are not enough. *Journal of Applied Physiology*, 118(6):716–722, 2015.

[2] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[3] Scott E. Crouter, Kurt G. Clowers, and David R. Bassett. A novel method for using accelerometer data to predict energy expenditure. *Journal of Applied Physiology*, 100(4):1324–1331, 2006.

[4] AH Montoye, Lanay M Mudd, Subir Biswas, and Karin A Pfeiffer. Energy expenditure prediction using raw accelerometer data in simulated free-living. *Medicine & Science in Sports & Exercise*, 47(8):1735–1746, 2015.

[5] Y. Mu, H. Z. Lo, W. Ding, K. Amaral, and S. E. Crouter. Bipart: Learning block structure for activity detection. *IEEE Transactions on Knowledge and Data Engineering*, 26(10):2397–2409, Oct 2014.

[6] J. Staudenmayer, D. Pober, S. Crouter, D. Bassett, and P. Freedson. An artificial neural network to estimate physical activity energy expenditure

and identify physical activity type from an accelerometer. *J. Appl. Physiol.*, 107(4):1300–1307, Oct 2009.

[7] S. G. Trost, W. K. Wong, K. A. Pfeiffer, and Y. Zheng. Artificial neural networks to predict activity type and energy expenditure in youth. *Med Sci Sports Exerc*, 44(9):1801–1809, Sep 2012.