Engineering Proceedings

MDPI

Proceedings

Low-Cost Environmental and Motion Sensor Data for Complex Activity Recognition: Proof of Concept †

Rok Novak[1,2,*], David Kocman [1], Johanna Amalia Robinson [1,2], Tjaša Kanduč [1], Denis Sarigiannis [3,4,5], Sašo Džeroski [2,6] and Milena Horvat [1,2]

[1] Department of Environmental Sciences, Jožef Stefan Institute, 1000 Ljubljana, Slovenia; david.kocman@ijs.si (D.K.); johanna.robinson@ijs.si (J.A.R.); tjasa.kanduc@ijs.si (T.K.); milena.horvat@ijs.si (M.H.)

[2] Jožef Stefan International Postgraduate School, 1000 Ljubljana, Slovenia; saso.dzeroski@ijs.si

[3] Environmental Engineering Laboratory, Department of Chemical Engineering, Aristotle University of Thessaloniki, Thessaloniki, Greece; denis@eng.auth.gr

[4] HERACLES Research Centre on the Exposome and Health, Center for Interdisciplinary Research and Innovation, Thessaloniki, Greece

[5] University School of Advanced Study IUSS, Pavia, Italy

[6] Department of Knowledge Technologies, Jožef Stefan Institute, 1000 Ljubljana, Slovenia

* Correspondence: rok.novak@jsi.si

† Presented at the 7th Electronic Conference on Sensors and Applications, 15–30 November 2020; Available online: https://ecsa-7.sciforum.net/.

Published: 15 November 2020

Abstract: Merging new sensing technologies with machine learning methods can be used as a tool to recognize complex activities. A wearable PM sensor in combination with a motion tracker was provided to 97 individuals for 7 days in two seasons. These data sets were used in three different models, based on three classification of activity algorithms: IBk, J48 and RandomForest, which showed for hourly (minute) an accuracy of 31.0 (23.1)%, 28.6 (22.0)% and 35.7 (23.0)%, respectively. Most misclassified instances concerned vaguely defined activities. Low accuracy can also be explained with the differences in time scales. The accuracy could be improved by more clearly defining the activities and collecting per-minute data.

Keywords: activity recognition; classification; machine learning; particulate matter

1. Introduction

Exposure to particulate matter (PM) and the intake dose can be heavily dependent on a specific activity an individual is performing [1,2]. By aggregating data per activity, instead of per time interval, the user is provided with another view to better discern where steps should be taken to reduce possible harm, caused by increased PM exposure or intake dose. Although activity recognition software is widely used in many commercial and research devices, confined to recognizing simpler activities, such as walking, running or other sports activities [3,4]. Recognizing complex activities still proves to be quite challenging [5]. Devices (in general) use integrated movement sensors, such as accelerometers and gyroscopes, for activity recognition. These sensors are also present in smartphones, allowing them to perform activity recognition, e.g., counting steps. Adding environmental sensors to the input dataset could potentially improve the accuracy of recognition of complex activities. Measuring the concentration of PM, the temperature and relative humidity in the vicinity of an individual could give valuable insight into their activity. Elevated levels of PM have been found for complex activities, such as cooking, cleaning and smoking [6–8], and combining these data points with ambient temperature, heart rate, and movement could allow the

algorithms to distinguish between these activities, e.g., high PM and high temperature for cooking, high PM and low heart rate for smoking, etc.

Machine learning classification algorithms can be used for activity recognition, and with powerful algorithms, such as Random Forest, the percent of accurately labeled instances is in certain cases >99% [3,4]. A training dataset which provides quality data ("quality" can be differently defined on a case by case basis) can sufficiently train the model to provide high accuracy from correctly labeling data points. This can sometimes mean that the model needs data with high temporal resolution or clearly defined activity labels, clearly delimited sets of activities, etc.

2. Methodlogy

2.1. Data Collection

Data used in the study was collected from 97 participants in the ICARUS campaign [9]. Most participants were involved in the winter (February to March 2019) and summer (April to June 2019) season of the campaign for approximately 7 days each and equipped with two sensor devices:

1. A Garmin Vivosmart 3 Smart Activity Tracker (SAT) [10], which was strapped on each participant's wrist for the entire duration of the data collection period. Temporal resolution for the data was one minute. The data used from the SAT was primarily the average minute heart rate and the number of steps and distance per minute, which indicated movement.
2. A Portable PM measuring device (PPM), which was developed for the ICARUS project by IoTech Telecommunications [11], using a Plantower [12] pms5003 sensor, based on the laser light scattering principle. The device provided minute resolution data for three size classes of PM (1 μm, 2.5 μm, 10 μm), temperature, relative humidity and speed.

All participants had to fill out a Time Activity Diary (TAD), where information about their activities was provided for each hour. They were given 7 blank daily TADs, where they were able to fill in circles for each activity they performed for every hour of the day. These files were collected and digitalized. Information about all indoor and outdoor activities was used.

2.2. Data Overview

The minimum, 1st quartile, median, mean, 3rd quartile, and maximum values for all numeric variables in the final dataset are presented in Table 1.

After cleaning the data, all values are within expected limits. The PM values were fixed at 180 μg/m^3 as the highest possible value, otherwise the mean, median and quartile values are as expected. Values for speed are quite low, due to the fact that all values above 20 km/h were removed, as there are no activities included in this research where speed could be above 20 km/h.

Table 1. Basic statistics for all numeric variables in the dataset.

	Median	***Mean***	***Max***	***Min***	***1st Q***	***3rd Q***
PM_1 [μg/m^3]	9.0	15.2	180	0.0	5.0	17.0
$PM_{2.5}$ [μg/m^3]	12.0	21.2	180	0.0	7.0	24.0
PM_{10} [μg/m^3]	13.0	23.7	180	0.0	7.0	26.0
Temperature [°C]	24.1	24.0	35.2	5.8	22.8	25.3
Relative humidity [%]	32.7	33.0	80.7	6.7	28	37.9
Speed [km/h]	0.52	1.21	20.0	0	0	1.65
Avg. Heart rate [bpm]	71.0	74.1	205	34	62	83
Steps [nr.]	0	5.40	276	0	0	0

For a more thorough overview of the dataset, the average values, were calculated for each activity separately and plotted in Figure 1.

Running has the highest value of speed, heart rate, steps and MET, and the lowest for temperature. sports.OUT and sports.IN also stand out in all of these values, while also having low

average PM concentrations. Importantly, sports.IN has also a higher average temperature and relative humidity than sports.OUT.

Highest PM values are observed for smoking, followed by cooking and cleaning, and lowest for sleep. Sleep also has the lowest speed, heart rate, number of steps and MET, all of which is expected. It doesn't stand out in regard to temperature and humidity.

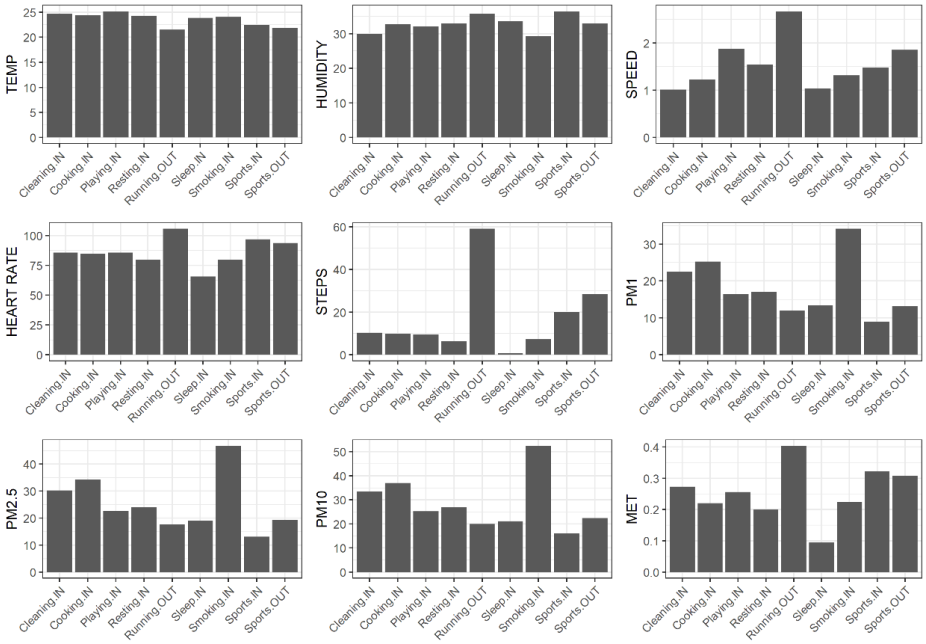

Figure 1. Average values for all variables and activities.

2.3. Classifiers Used

Three classification algorithms were chosen, based on best practices and recommendations. The classifiers used are listed in Table 1 along with a short description of each. All of these algorithms are included in WEKA 3.8.3 [13], which was used for the analysis. After the data was imported to WEKA and before it was analyzed, it was normalized by rescaling all attributes to the range of 0 to 1 as the distribution was not Gaussian.

Table 1. Classifiers used in this research with a short description.

Classifier	Description
IBk [14]	Instance Based learner, otherwise known as the k-nearest neighbor (kNN) classifier; selects value of k based on internal cross-validation.
J48 [15]	J48 is a Java implementation of the C4.5 decision tree algorithm developed in 1993 by Ross Quinlan [16]. It can be used for classification and allows a high number of attributes. Deemed as "machine learning workhorse", ranked nr. 1 in the *Top 10 Algorithms in Data Mining* [17].
RandomForest [18]	Constructs a forest of decision trees in a randomized manner. Developed by Leo Breiman in 2001 [19].

3. Results and Discussion

3.1. Comparing Classifiers

There are several measures of predictive performance of classifiers, such as the overall classification accuracy and the K coefficient, and the (per-class and average) true positive (TP) and false positive (FP) rates, and precision, among others. Table 4 shows a comparison of the listed metrics for all the classifiers used in this research.

10-fold cross-validation was used the evaluation methodology.

Table 2. Summary of results for all models.

Classifier	*Correctly Classified*	*Kappa*	*TP*	*FP*	*Precision*	*ROC Area*	*PRC Area*
IBk	32.7%	0.2424	0.327	0.084	0.363	0.621	0.220
J48	39.5%	0.3195	0.395	0.076	0.407	0.767	0.370
RandomForest	43.1%	0.3601	0.431	0.071	0.432	0.807	0.444

As evident in Table 4, the RandomForest method mostly preforms better in this specific task than IBk and J48. It correctly classifies 10.4 percentage points of instances more than IBk and 3.6 percentage points more than J48. Its Kappa coefficient is also better than IBk and J48. FP rates are lower and TP rates are higher for RandomForrest. All metrics show that in terms of accurately predicting an activity the models can be ranked: RandomForest > J48 > IBk.

3.2. IBk

The IBk classifier correctly classified 2939 (32.7%) of instances, with a K (Kappa Coefficient) of 0.2424. True positive (TP) rates are >0.4 for two activities: the highest (0.7) for sleeping, the next best for resting (0.5). Most misclassified instances of sleep were labeled as resting. This is expected, as the two activities share several similar characteristics, such as low heart rate, no movement and low levels of PM.

A relevant observation is that sleeping typically has very clearly defined time intervals (at night), a low heart rate and no movement. Sleeping is also one of the few activities that every participant indicated and is consequently very homogeneously distributed. On top of this, it is the only activity that is performed consecutively for several hours, without interruptions, which in turn means that there are very few instances where there are distorted minute values present inside an hour. An example of such distorted values would be that a person only runs for 20 min, but indicates that running was the main activity in that hour. Only 1/3 of the data would really confirm this fact, the other 40 min are other activities, which distort the final result. On the other hand, this is not common for sleeping, as most people sleep in one single block of time.

Resting is also somewhat characterized with longer consecutive time intervals without interruptions. It also has the most misclassifications and highest False positive (FP) rate, which is due to the fact that resting is the second most frequent activity chosen by participants (after sleeping) in the whole study and in turn should overlap with most activities very frequently (the "default" activity being resting). It is also vaguely defined and open to interpretation, which can prompt participants to include a whole swath of activities under this term, e.g., reading a book, playing board or computer games, watching television, chatting with friends, taking a leisurely walk, napping, having a dinner party, etc. All of these activities can differ in many aspects, such as heart rate, movement, speed or PM concentrations, which would make accurate predictions more difficult.

Besides sleeping and resting, TP rates are >0.25 for all activities, with the example of smoking (0.030). An interesting observation is that running also had quite a small False positive (FP) rate of 0.009, mostly being misclassified as sports outdoors. This could also be a consequence of activities being mislabeled by the participants (confusing sports outdoors and running when labeling activities).

3.3. J48

Results show that the model learned by J48 correctly classified approximately 22.0% of instances, with a K value of 0.1221. One noticeable difference of TP rate is evident with cooking, where it was 0.187 with IBk and 0.320 with J48, otherwise the TP values do not differ much between the different models. Similar patterns are obvious with all other measures of accuracy.

Running has the lowest number of misclassified instances, where most of the latter are labeled as being indoor or outdoor sports activities. This is expected, as these activities share a distinctive

pattern of an elevated heart rate and more movement. In this case, resting faired a bit worse than with IBk as there are fewer correctly classified instances.

3.4. RandomForest

The results from the model based on RandomForest, showed the highest accuracy (23.6%) and lowest errors. Although the TP and FP values are somewhat higher, they don't differ much from IBk and J48, with sleeping and resting being again on top with 0.790 and 0.408 TP rates, respectively. A similar pattern as in the previous classifiers was observed in the confusion matrix, where running had the fewest misclassified instances, mostly being sports outdoors. Again, very few activities were misclassified as sleeping, the only outlier being resting with 74 misclassifications.

4. Conclusions

All the used classifiers had accuracy above 30%, with RandomForest being the most accurate (43.1%). As the labeled data consisted of hourly labeled activities, this gives it less resolution and more errors (some activities don't last an hour, and most don't last exactly a set number of full hours). A future improvement would be to label data by minute, not by hour. This would match the desired output of per-minute predictions and allow finer granularity.

All of the models had the most misclassified instances from resting activity. This could be the result of the vague definition of resting in comparison to sleeping, running and most other activities. On the other hand, sleeping or smoking are quite well-defined activities, where there is little room for subjectivity. A prospect for future studies would be to take the most ambiguous or subjective activities and break them down into more defined activities, as specified above. Although, this would mean more challenges for collecting data, it could provide more detailed and accurate final results.

Combining the data points used in this research with environmental stressors, measured with portable low-cost sensors, could provide detailed results of exposure and intake dose. Further research is needed to test and validate these approaches.

As low-cost sensors become more widely used and individuals are able to gain access to more information about their living environment, it is crucial for researchers to provide adequate tools to assess and improve accuracy of activity classification. A promising step forward would be to reduce the input of individuals and increase the role of machine learning. This research shows a novel approach of using classification methods with data from low-cost portable environmental and activity sensors, to recognize specific activities without direct human input.

Author Contributions: R.N. and D.K. conceptualized the idea, collected the data with J.A.R. and T.K., analyzed, validated and visualized the data, and prepared the original draft. S.D. reviewed and edited the content. D.S. coordinated and led the design efforts for the ICARUS project. M.H. headed the project on a local level, ensured funding and contributed to the final review and editing. All authors approved the content of the manuscript.

Funding: This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 690105 and the Young researchers program funded by the Slovenian Research Agency.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ferro, A.R.; Kopperud, R.J.; Hildemann, L.M. Elevated personal exposure to particulate matter from human activities in a residence. *J. Expo. Sci. Environ. Epidemiol.* **2004**, *14*, S34–S40, doi:10.1038/sj.jea.7500356.
2. Patel, S.; Sankhyan, S.; Boedicker, E.K.; DeCarlo, P.F.; Farmer, D.K.; Goldstein, A.H.; Katz, E.F.; Nazaroff, W.W.; Tian, Y.; Vanhanen, J.; et al. Indoor Particulate Matter during HOMEChem: Concentrations, Size Distributions, and Exposures. *Environ. Sci. Technol.* **2020**, *54*, 7107–7116, doi:10.1021/acs.est.0c00740.
3. Barna, A.; Masum, A.K.M.; Hossain, M.E.; Bahadur, E.H.; Alam, M.S. A study on Human Activity Recognition Using Gyroscope, Accelerometer, Temperature and Humidity data. In Proceedings of the 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE), 2019; pp. 1–6.

4. Shelke, S.; Aksanli, B. Static and Dynamic Activity Detection with Ambient Sensors in Smart Spaces. *Sensors* **2019**, *19*, 804, doi:10.3390/s19040804.
5. Dernbach, S.; Das, B.; Krishnan, N.C.; Thomas, B.L.; Cook, D.J. Simple and Complex Activity Recognition through Smart Phones. In Proceedings of the 2012 Eighth International Conference on Intelligent Environments, 2012; pp. 214–221.
6. Holm, S.M.; Balmes, J.; Gillette, D.; Hartin, K.; Seto, E.; Lindeman, D.; Polanco, D.; Fong, E. Cooking behaviors are related to household particulate matter exposure in children with asthma in the urban East Bay Area of Northern California. *PLoS ONE* **2018**, *13*, e0197199, doi:10.1371/journal.pone.0197199.
7. Wan, M.-P.; Wu, C.-L.; To, G.-N.; Chan, T.-C.; Chao, C. Ultrafine particles, and PM 2.5 generated from cooking in homes. *Atmos. Environ.* **2011**, *45*, 6141–6148, doi:10.1016/j.atmosenv.2011.08.036.
8. Corsi, R.L.; Siegel, J.A.; Chiang, C. Particle resuspension during the use of vacuum cleaners on residential carpet. *J. Occup. Environ. Hyg.* **2008**, *5*, 232–238, doi:10.1080/15459620801901165.
9. 'ICARUS2020.eu', *icarus*. Available online: https://icarus2020.eu/ (accessed on 12 October 2018).
10. Garmin and G. L. or Its Subsidiaries, 'Vívosmart 3|Activity Tracking', *Garmin*, Nov. 16, 2018. Available online: https://buy.garmin.com/en-US/US/p/567813 (accessed on 16 November 2018).
11. IoTECH TELECOMMUNICATIONS. Available online: https://iotech.gr/ (accessed on 13 May 2019).
12. Plantower PMS5003 Datasheet.pdf. Available online: http://www.aqmd.gov/docs/default-source/aq-spec/resources-page/plantower-pms5003-manual_v2-3.pdf (accessed on 15 June 2019).
13. Frank, E.; Hall, M.A.; Witten, I.H. *The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques"*, 4th ed.; Morgan Kaufmann: 2016.
14. IBk. Available online: https://weka.sourceforge.io/doc.dev/weka/classifiers/lazy/IBk.html (accessed on 9 June 2020).
15. J48. Available online: https://weka.sourceforge.io/doc.stable-3-8/weka/classifiers/trees/J48.html (accessed on 9 June 2020).
16. Salzberg, S.L. C4.5: Programs for Machine Learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1993. *Mach. Learn.* **1994**, *16*, 235–240, doi:10.1007/BF00993309.
17. Wu, X.; Kumar, V.; Ross Quinlan, J.; Ghosh, J.; Yang, Q.; Motoda, H.; McLachlan, G.J.; Ng, A.; Liu, B.; Yu, P.S.; et al. Top 10 algorithms in data mining. *Knowl. Inf. Syst.* **2008**, *14*, 1–37, doi:10.1007/s10115-007-0114-2.
18. RandomForest. Available online: https://weka.sourceforge.io/doc.stable-3-8/weka/classifiers/trees/RandomForest.html (accessed on 9 June 2020).
19. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32, doi:10.1023/A:1010933404324.