Interactive visualization towards facilitating physical activity researchers in training machine learning algorithm

1st Author Name
Affiliation
City, Country
e-mail address

2nd Author Name
Affiliation
City, Country
e-mail address

3rd Author Name
Affiliation
City, Country
e-mail address

ABSTRACT

Physical activity recognition using accelerometer data from ubiquitous computing devices is a useful tool to access physical behavior in real-time and designing just-in-time health behavior intervention. Physical activity researchers are domain experts in the sensor data generation process, but likely do not possess an adequate understanding of machine learning techniques. Tools to aid these domain experts in developing machine learning algorithms is necessary. Without the availability of such tools, debugging algorithms and evaluating models becomes less efficient and more prone to error. Our research aims to explore interactive visualizations as a way to support debugging and training of physical activity classification algorithms by physical activity researchers. We have developed Intuit, an interactive visualization tool, designed to explain misclassifications during training to facilitate end-users with an exploration of data and building of a sound mental model of the algorithm. We present Intuit's interactive visualization technology and the design process.

Author Keywords

Physical activity recognition; accelerometer; machine learning; interactive visualization; debugging; end-user mental model

ACM Classification Keywords

Human-centered computing → Ubiquitous and mobile computing; *Ubiquitous and mobile computing design and evaluation methods*

INTRODUCTION

Accelerometer-based sensors have been extensively used to objectively monitor physical activity [1]. Physical behaviors have important health implications [2-3], as a result, the need for assessment of these behaviors in real-time within population studies is growing.

Physical activity researchers are the domain experts in this field and should steer the process of building robust and accurate physical activity recognition machine learning (ML) algorithms. However, most likely they lack a thorough understanding of ML and technical skills to efficiently and correctly build such algorithms.

We aim to aid such end-users by exploring interactive visualization. In particular, we designed and developed a web-based interactive visualizer that is designed to explain the misclassifications of the ML algorithm based on the raw data, that they are most familiar with, which could help them acquire clearer mental model so that they can make informed decisions on how to resolve the issue.

RELATED WORK

The user's mental model of the learning system must appropriately map onto the behavior of the actual system to facilitate an iterative-feedback loop between the system and the user. The phrase interactive machine learning was popularized by Fails et al. while describing how a train-feedback-correct cycle allowed users to correct the mistakes of an image segmentation system [4]. Since then, researchers have explored this cycle to enable better model selection by end-users [5], and to elicit labels for the most important instances [6]. ModelTracker [7] is an interactive visualization designed to support direct error examination and debugging in ML, as it subsumes information contained within numerous summary statistics while displaying example-level performance. Recent work has also focused on building intelligible ML algorithms so that end-users can understand, validate, edit and trust a learned model [8]. Explanatory debugging [9] incorporates both machine-learned knowledge representation and user-driven learning simultaneously. Kapoor et al. proposed ManiMatrix that allows users to directly interact with confusion matrix and to view the implications of incremental changes to the matrix via a real-time interactive cycle of re-classification and visualization [10]. A confusion matrix is a specific table layout that allows visualization of supervised learning. Each column of the matrix represents the instances in a predicted class while each row represents the instances in an actual class. To our knowledge, most studies have focused on visually explaining the learning algorithm, but not enough focus has been given to build intuitive

interactive visualizations of the actual data and features and evaluate its significance in training of the system. Recent works have focused on visually explaining additive classifiers[11], presenting a graphical view of confusion matrices to help users understand the relative merits of various classifiers[12].

Model agnostic explanation of classification prediction is available, but they might not be accessible to users, who are domain experts in a field that would benefit from ML, but they might lack required understanding of ML and technical skill to use such tools in their domain.

By shifting the ML focus into context that is familiar to the users, our goal to facilitate the on-boarding of the users to troubleshoot such algorithms. In this paper, we present Intuit's design process based on expert interviews. We also present our proposed technique for interactive visualization for supporting physical activity research train ML algorithms.

To our knowledge, little research exists on how to support physical activity researchers in troubleshooting ML algorithms during training. *Intuit* builds upon the success of previous interactive visualizers as we analyze the need of a new user group: physical activity researchers.

INTUIT'S DESIGN

In this section, we will give a brief overview of two physical activity researchers, who volunteered to be interviewed. To help with the design process, we wanted to understand the needs of our end-users and how they envisioned the system would be to help them train an ML algorithm.

Participant 1 – Novice in ML

Our interview with Participant-1 (P1) was motivated by identifying the needs of a physical activity researcher, who is a novice in ML, while they learn how to train ML algorithm. We approached P1 because he is a user of products that use sensors to capture movements or behaviors (such as fit bands) and is also a researcher whose job is to generate data that later will be used by other people to create ML algorithms.

Participant description

P1 is a user of products that use sensors to capture movements or behaviors (such as fit bands) and is also a researcher whose job is to generate data that later will be used by him and his colleagues to create ML algorithms. Asked about his knowledge in the topic, P1 stated to be familiar with ML and understanding its fundamentals. His interest in the subject is influenced by his current job, and he has been doing machine learning courses to know how to prepare better datasets for future algorithms. He is aware that many times the raw data obtained from 'real life' does not match the algorithm resulting in unexpected behavior.

Insight from the interview

When asked about visualizations about ML or visualization tools for it, P1 indicated that he had not seen any about the ML algorithms. However, for him, being able to see how the decisions are made would be more useful than actually knowing more about the algorithm. This approach is represented in the sketch that we asked him to do about a possible ML visualization, where for him the important element was the decision flow in order to see where the error in the algorithm happened.

P1's feedback about our initial sketches was really useful: he remarked that the visualization needed to have a very good hierarchy and very good labels to avoid confusing the user, especially in the items names; and highlighted that the visualization needed to allow comparisons of peaks between the different charts at the same time. Overall, he stated that our idea would save a lot of time and would allow two types of use (i) tagging the instances that were incorrectly classified, and (ii) fixing the algorithm by experts by analyzing the tagged instances

Participant 2 – Some experience in ML

Our interview with Participant-2 (P2) was motivated by identifying the needs of physical activity researcher, who has some experience in ML, while they are training ML algorithms. We approached P2 to understand how physical activity researchers, who are ML users, generate / re-use / correct or change algorithms in order to get more accurate and expected results. We wanted to learn what might be missing that is important to that group.

Participant description

P2, a Ph.D. student researcher, focuses on applied machine learning techniques to build physical activity classification algorithms. She has two years of experience using ML in her work.

Insight from the interview

Asked about how she chooses or implements ML, she explained to us that the election of one algorithm instead of another depends mainly on the type of results that the researcher wants: best precision or best accuracy algorithms usually require different nests of features, while algorithms in mobile phones or similar devices need to be fast and, therefore, simple.

Asked if she uses visualization tools in her work, she remarked that sometimes visualizing data is not an easy task, especially when managing big data. Interviewee-2 stated that in fact most of the times there are not tools to visualize those results. Right now, she studies which feature show more correlation with the results by using R and visualizations, but she agrees that a tool for the raw data and the feature would be a good instrument to have in order to view what is happening in the algorithm and improving her knowledge of this process.

Her feedback about our initial sketches was really useful to identify strengths, weaknesses, and opportunities in our

idea. P2 thought that our visual exploration was a little simplistic and that we gave too much importance in the features, while the process of selecting them tends to be done at the very beginning of the process and not at the end when our visualization is supposed to be used. She did not see the functionality in the frequency feature histograms for this very same reason.

However, she stated that a visualization that highlighted outliers and would help her summarizing and filtering data would be a great implementation.

Task analysis

Table below shows the distilled task that needed to be done based on the findings of our interviews.

Domain	Analytic task	Technical task
Summarize the performance of machine learning algorithm		Visualize confusion/adjacency matrix with true vs. predicted labels for multiple classes
Search and identify misclassified and correctly classified instances	Filter	Use confusion matrix to select cells that belong to correctly or incorrectly classified instances
Compare the raw data of the misclassified instances with the correctly classified instances to discover any errors they could fix to improve performance	Correlate find anomalies	Visualize raw time-series data of correctly classified and misclassified instances juxtaposed together to check for noisy labels.

Final design decisions

The final version of our design consists of two main sections. In the first section (Figure 1), we provide a high-level explanation of machine learning algorithms. This explanation serves as an introduction and brief background for interpreting the visualizations and data. The second part of the design consists of the main features of the application developed through an adjacency matrix, an item list, and line charts. We first discuss our design choices from a user interaction perspective and then discuss how our design decisions follow visualization principles.

Through our interviews, we realized that one of the things users want is the ability to make comparisons at different levels. This is important for sense-making and discovery. Therefore, while our visualization is designed as a guided step-by-step process (where each of the visualizations only appears once the users interact with the visualization), users can also see everything as a single main view. We did this to better assist users in comparing different levels of information. For instance, if they discover something through the line's charts about a specific activity and need to compare it with another activity in the matrix, they can easily see the matrix as it is in the same page. We also wanted to complement our visualization with a tutorial tour. However, since some of the features of the page are first non-existent and only appeared after the user demanded them, we tried to provide as much information as possible at each step, to better assist users through the process.

We now discuss the three main visualizations. Our first visualization is a confusion matrix. We chose the confusion/adjacency matrix (Figure 2) to show the prediction results of the machine learning algorithm. As previously mentioned, rows represent the classes of actual activities and the columns represent the classes of activities that have been predicted by the ML algorithm. The value in each cell is equal to the number of all samples with a specific actual class and predicted class of activity. Therefore, the number inside the cells that are on the diagonal - for which the actual class is equal to the predicted class - represent the number of samples which have been correctly predicted for a specific class. Figure 2 illustrates the visualization of the confusion matrix.

We chose a matrix for two main reasons. First, because confusion matrices are concepts that most users with some basic understanding of ML are already familiar with, and second because they are easily scalable for a large number of classes. For instances, alternative designs such as tree structure would have been hard to manage given we have more than 20 classes. The users can click on any cell of their interest on the adjacency matrix to get the details on demand. For representing numbers in each cell, we initially

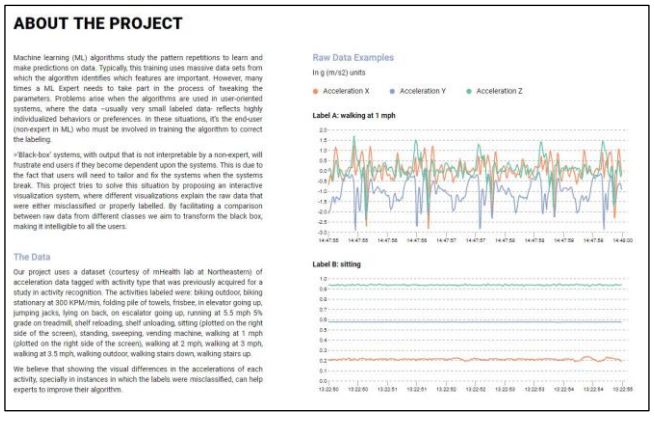

Figure 1 Top part of the interactive visualizer showing the high-level explanation of data and algorithm

Figure 3 Confusion matrix summarizing the classifications

thought of creating other versions of the visualization. For example, one of our ideas was the use of circles with different sizes at the intersection of each column and row (bigger sizes of circles for higher cell numbers). However,

we realized that size cannot be easily scalable for large numbers in the cells. Therefore, we chose to use color intensity instead of size to represent numbers. We thought that colors are a good choice for helping users to detect anomalies. To better assist users in detecting anomalies, we decided to choose the diverging color map (consisting of two colors of blue for correctly classified samples and red for incorrectly classified samples). Our choice of diverging color map is guided by the fact that we have quantitative data and are interested in extreme values. We used the aforementioned colors for two main reasons. According to Madden et al. [14], the color blue is usually associated with the attribute of "high quality". So, we chose this color for the samples that are correctly classified. On the other hand, as red is a stimulating color we used it to grab users' attention to cells that have been incorrectly classified samples. Secondly, we wanted two colors that support color-blind design. Aside from using colors, we also included the actual number of instances within each cell so that users can quickly search and identify any misclassification conditions they want to work on. Colors are therefore an intentional redundancy in choice in our design.

Based on the cell that the user chooses from the confusion matrix the users pick, they are provided with the list of samples represented by the cell. Those samples could either be a list of misclassified or correctly classified instances for which a list of attributes such as participant ID, start and end time, true activity and predicted activity are shown. This list is sorted based prediction confidence (as most ML models output probability score for each prediction) of the algorithm. For the list of items, we chose a simple table to list all of those instances. Users can scroll and see them. We think that while tables are very simple, they can depict the list of instances and their properties very well. The interactions that are available is the selection of each row. Figure 3 illustrates our visualization for the list of items.

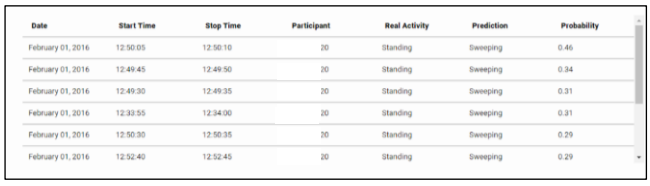

Figure 3 Item list showing instances corresponding to the selected cell from confusion matrix

After selecting a specific row, users see three-line charts showing the visualization of raw data are. We chose line charts to plot raw accelerometer data, as they are apt to show trends as well as patterns in the time-series data. The line chart of the misclassified instance is in the middle while the line chart for instance of the actual class and

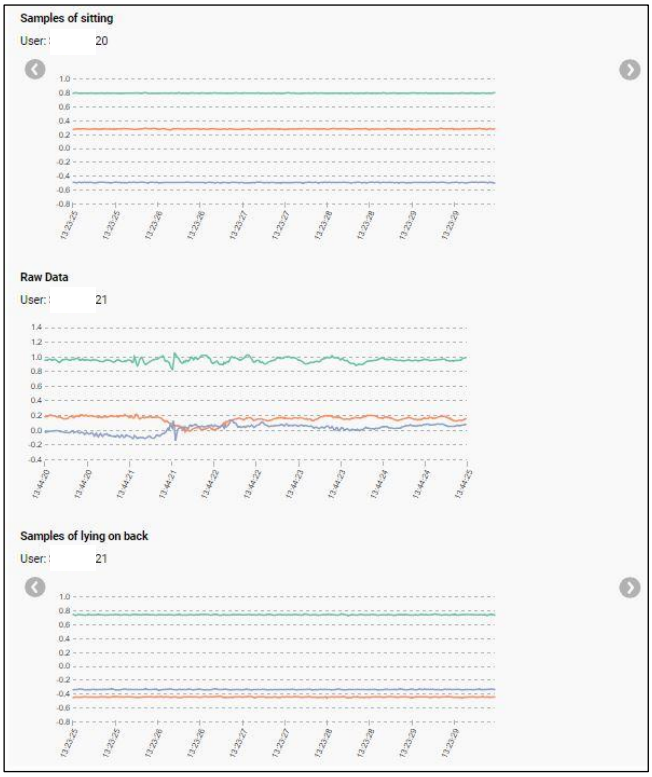

Figure 4 Visualization of raw data for true class (top), misclassified class (middle), predicted class (bottom)

predicted class are on the top and bottom, respectively. The goal is to facilitate an easier and quick comparison of raw data. We have also added a left and right button for the true/predicted instance so that the users can compare the misclassified instance with more than one correctly identified instance from both the classes. We have also used three different colors for each accelerometer direction (x, y, and z). We used the categorical color map, with three distinct hues. Figure 4 illustrates our visualization of the line charts.

DISCUSSION

We showed the web-based interactive visualizer prototype that we developed to P1 and P2 as a part of participatory design process. The goal was to understand if it addressed some of their needs and to gather additional feedback to guide us improve future iterations of the system.

Despite a relatively short time of exposure (~5 minutes) with our interactive visualization prototype, both the users found the interactive visualizations easy to use. P1 commented that the use of confusion matrix to show the performance of multiclass machine learning algorithms was apt given he is familiar with it. The users were able to understand the significance of color choices to represent correctly labelled cells and incorrectly labelled cells. The users also thought that the layout and the listing of information about the instances was intuitive and helpful, especially while troubleshooting prediction errors in machine learning. P2 mentioned that she found the

visualization to compare raw data for different classes helpful to understand the effect of raw data on the algorithm.

P2 suggested that it would be helpful in debugging if there is a way to display the multidimensional features in isolation or combination that could potentially shed more insight into their richness and relevance.

We also realized that one of the issues with our design was that users might have issues working with a large number of raw-data sample instances. At the current state of our visualization, the user has to manually peruse to individual instances during the decision process, which can potentially increase the mental burden and decision fatigue. We think that the design could be improved by either grouping samples or filtering them.

These preliminary findings based on our user-study in a convenient sample (n=2) is not conclusive yet encouraging. The impressions from the two users is in accord with our initial hypothesis that a simple intuitive interactive visualization of the raw data from the misclassified instances might help in the training and debugging process of ML algorithms.

CONCLUSION AND FUTURE WORK

An interactive visualizer prototype was designed and developed to aid physical activity domain experts, with negligible to low ML expertise) while they train and debug ML algorithms. We interviewed two potential end-users of our system to understand their needs, expectations, and understanding to drive our design process. Both participants had positive impressions while testing the web-based prototype that was developed.

Future work would include the inclusion of interactive visualizations of features (as suggested by P2) to aid better mental model building of the end-users regarding the ML algorithm.

This would further enable us to run user-studies to test our hypotheses of whether going through the troubleshooting process using our proposed visualization could actually help users with building a sound mental model of the algorithm, and that they could train machine learning algorithms with better performance faster than users that did not use our visualization.

REFERENCES

[1] Bonomi, A.G., Physical activity recognition using a wearable accelerometer, in Sensing Emotions: The impact of context on experience measurements, J. Westerink, M. Krans, and M. Ouwerkerk, Editors. 2011, Springer Netherlands: Dordrecht. p. 41-51

[2] Lee, I.M., E.J. Shiroma, F. Lobelo, P. Puska, S.N. Blair, and P.T. Katzmarzyk, Effect of physical inactivity on major non-communicable diseases worldwide: an analysis of burden of disease and life expectancy. Lancet (London, England), 2012. 380(9838): p. 219-229.

[3] Celis-Morales, C.A., D.M. Lyall, P. Welsh, J. Anderson, L. Steell, Y. Guo, R. Maldonado, D.F. Mackay, J.P. Pell, N. Sattar, and J.M.R. Gill, Association between active commuting and incident cardiovascular disease, cancer, and mortality: prospective cohort study. BMJ, 2017. **357**: p. j1456

[4] Fails, J. A., Olsen, D. R., and Jr. Interactive machine learning. In Proceedings of the 8th International Conference on Intelligent User Interfaces (2003), 39–45

[5] Amershi, S., Fogarty, J., Kapoor, A., and Tan, D. Examining multiple potential models in end-user interactive concept learning. In Proceedings of the ACM Conference on Human Factors in Computing Systems (2010),1357–1360

[6] Cakmak, M., Chao, C., and Thomaz, A. L. Designing interactions for robot active learners. IEEE Transactions on Autonomous Mental Development 2, 2 (2010), 108–118

[7] Amershi, S.,Chickering, M.,Drucker, S.,Lee, B.,Simard, P.,Suh, J. ModelTracker: Redesigning performance analysis tools for machine learning. ACM HFCS 2015

[8] Caruana, R.,Lou, Y.,Gehrke, J., Koch, P., Sturm, S.,Elhadad, N. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. KDD 2015

[9] Kulesza, T.,Stumpf, S.,Burnett, M.,Weng-Keen, Wong,Riche, Y.,Moore, T.,Oberst, I., Shinsel, A., McIntosh, K.Explanatory debugging: Supporting end-user debugging of machine learned programs VL/HCC 2010

[10] Kapoor, A.,Lee, B.,Tan, D., Eric Horvitz, E. Interactive optimization for steering machine classification,Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, April 10-15, 2010, Atlanta, Georgia, USA

[11] Poulin, B., Eisner, R., Szafron, D., Lu, P., Greiner, R., Wishart, D.S., Fyshe, A., Pearcy, B., MacDonell, C., Anvik, J. Visual explanation of evidence in additive classifiers, Proceedings of the 18th conference on Innovative applications of artificial intelligence, p.1822-1829, July 16-20, 2006, Boston, Massachusetts

[12] Talbot, J., Lee, B., Kapoor, A., Tan, D.S. EnsembleMatrix: interactive visualization to support machine learning with multiple classifiers, Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, April 04-09, 2009, Boston, MA, USA

[13] Thomas J. Madden, Kelly Hewett, Martin S. Roth (2000) Managing Images in Different Cultures: A Cross-National Study of Color Meanings and Preferences. Journal of International Marketing: Winter 2000, Vol. 8, No. 4, pp. 90-107.