Vol.2019-UBI-63 No.5
Vol.2019-ASD-15 No.5
2019/8/26

Improving Annotation for Activity Recognition with Active learning and Gamfication

Nattaya Mairittha[1,a)] Tittaya Mairittha[1,b)] Sozo Inoue[1,c)]

Abstract: The "quality" and quantity of annotations can have a significant impact on the performance of activity recognition systems. Earlier, we explored a method to exploit uncertainty based active learning to calculate gamification points and employed those points to motivate users through smartphone notifications to collect accurate labels for activity recognition systems using smartphone sensors [17]. We conducted the experiments with three conditional methods and evaluated activity recognition performances of each method with one supervised learning classifier. It is better to investigate several machine learning algorithms for evaluating the data collected and show that our method is extremely efficient. Here, we evaluate activity recognition results with several algorithms and show that our proposed method has improvements in among data quality (the performance of several classifications), data quantity (the number of data collected), and user engagement (click-through rate for push notifications) that reflect our method could improve annotation for activity recognition systems.

1. Introduction

Smartphone-based activity recognition systems aimed at physical activities recognition such as walking or running, based on mobile sensor data. The sensor data may be recorded directly on the subject such as by carrying smartphones that have accelerometers and gyroscopes [1]. Understanding what users are doing in the physical world allows the smartphone app to be smarter about how to interact with them. A central challenge in smartphone-base activity recognition by self-labeling is data annotation studies in order to assess the labels describing the current activity while this activity is still on-going or recent to ensure that the dataset is labeled correctly. The quality of annotations can have a significant impact on the performance of the activity recognition systems. Hence, it is inevitable to rely on the users and to keep them motivated to provide labels. However, collecting accurate labels (*annotation*) comes with a hefty price tag, in terms of human effort. Either to have the data labeled by third-party observers or self-labeling both are costly, time-consuming, tedious, incorrect segments, and they have the risk of missing some of the activity labels.

To address the problem, we aim to leverage a gamification strategy [4] for activity data collection by employing game elements such as point systems [14] to reward and give such feedback to the users. The goal is to keep the users motivated continuously to provide activity labels. We introduce a new method using uncertainty based active learning [20] to evaluate the score of user's activity data collection performance and use that score as gamification points. The score is evaluated by approximating the unlabeled examples according to the current model uncertainty in its prediction of the corresponding activity labels using the entropy method. In this context, every activity that the users annotated will be evaluated to score for each activity class. Therefore, the users are getting gamification points of every single class as feedback that motivates activity data collection from several classes, not just one —these gamification points based on their activity data collection performance. To evaluate our proposed method, we began this work with the research questions: can we improve data quality, data quantity, and user engagement for activity data collection using the proposed method compared to the other methods? (See the detail of each method in Table 2). To answer this question, we conducted the controlled experiments and gathered 1,236 activity labels with mobile sensors data from 11 participants. We then reviewed the collected data and evaluated the methods using activity recognition processes [1] and Click-through rate (CTR) analysis [13]. We explored the answer to our research question by showing the proposed method is better than the others on data quality (i.e., the performance of classi cation several machine learning models), data quantity (i.e., the number of data collected), and user engagement (i.e., click-through rate for push notifications). In summary, the contribution of this paper is that **gamification points (the score of data annotation quality is measured by an uncertainty based active**

[1] Graduate School of Engineering, Kyushu Institute of Technology, 1-1 Sensui-cho, Tobata-ku, Kitakyushu-shi, Fukuoka, 804-8550, JAPAN
[a)] nattafahh@gmail.com
[b)] callmefons@gmail.com
[c)] sozo@brain.kyutech.ac.jp

ⓒ 2019 Information Processing Society of Japan

learning approach) are used to motivate the users for activity data collection.

2. Background and Related Work

There are two main areas of background work relevant to our current research. We will first explore activity recognition that focuses on uncertainty based active learning and we then explore existing studies of activity recognition and gamification.

2.1 Activity recognition and Uncertainty based active learning

Active learning has evolved into a popular paradigm for utilizing user's feedback to improve the accuracy of learning algorithms. Activity recognition using smartphone sensors has abundant unlabeled data instances that make active learning as an ideal solution by selecting the most informative sample among unlabeled data [11, 5] (i.e., the data sample in which current classifier is least confident, and querying the label of that point from the user). One of the most straightforward example to query the user for activity recognition that is to ask what the user was doing at a certain timestamp; however, all users do not have the ability to precisely recall an event at a certain timestamp. Uncertainty based active learning for activity recognition has been investigated by very few researchers [15, 21]. The well-known entropy is a popular uncertainty measurement widely used in previous studies on active learning [10] in order to calculate the informativeness of activity data instances. While other pieces of literature as mentioned above presented to use active learning to alleviate the labeling effort and ground truth data collection in activity recognition pipeline by relabeling the data instances, our research, in contrast, we aim to exploit such a strategy to evaluate activity data collection performance for gamification points.

2.2 Activity recognition and Gamification

The authors of [4] define gamification as *"the use of game design elements in non-game contexts"* to improve user experience and user engagement. Nevertheless, conceptualizing gamification from the authors of [9] indicates that gamification provides positive effects, however, the effects are greatly dependent on the context in which the gamification is being implemented, as well as on the users using it. Common gamification elements include points, badges, and leaderboards. Points can make sense of progression that motivates continued effort while leaderboards provide a social status element, and badges are a visual representation of some achievement used to encourage and recognize specific behaviors. Gamification points are the simplest way to reward a user for completing an action or a series of actions. In our study, we will utilize gamification points for activity data collection. Hence, if gamification points are allowed to be part of the data annotation process – to be a powerful `motivator` about the data collection, if we will – the goal is to keep them motivated to provide labels. Gamification in the context of mobile applications has been explored in recent years [23, 8, 7] while investigation of gamification aimed to support smartphone-based activity recognition directly is scarce. The authors of [23] showed the potential for gamification by using points as gamification elements in an experience sampling method (ESM) study on smartphones and notifications, describing both its effect on response quantity and quality. While other literature [8] explored opportunities and challenges that exist when using mobile sensors as input for game elements to engage people at events such as university orientation, we will present differences both on methods and applications as well as a contrast to their purpose.

3. Methods

In this section, we will give an overview of the proposal, as shown in Figure 1. We first detail the proposed gamification points using uncertainty based active learning method. We then describe the traditional method by using the accuracy of activity recognition model as gamification points.

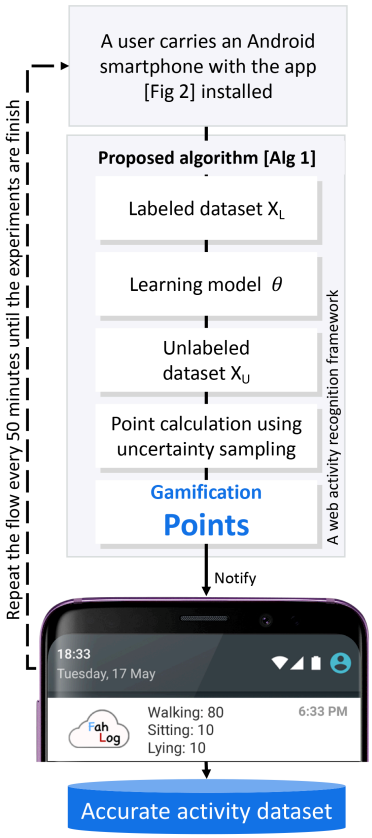

Figure 1 An overview of the proposed method

3.1 Proposed gamification points using active learning

Algorithm 1 shows the processing flow of the proposed method and Table 1 summarizes the additional mathematical expressions introduced in this section. To use an uncertainty based active learning approach to measure activity data collection performance, we assume that the uncertain

ⓒ 2019 Information Processing Society of Japan

score of each data sample was calculated by entropy measurement can be gamification points for each activity class. Thus, the idea behind this motivates activity data collection from several classes, not only one. In our method, we follow the standard pool-based active learning setting as described by the author of [20], and then integrate the proposed idea into the algorithm.

Algorithm 1 follows five steps:

(1) Let X_U be the set of unlabeled activity data instances and X_L be the set of labeled activity data instances.

(2) Use the labeled activity data set X_L to train the activity classifier using Random Forest [2].

(3) Predict the activity class of X_U using the activity classifier from (2) and map the predicted activity class is returned.

(4) Calculate X_U were mapped with activity classes as gamification points using entropy measurement:

$$H(x) = -\sum_{y \in Y} P_\theta(y|x) log P(x|y) \qquad (1)$$

(5) Average entropy score of each activity class and return as gamification points.

Algorithm 1 Proposed method

INPUT:
$X_U = \{(x^0, x^1, ..., x^i)\}$
$X_L = \{(x_1^0, x_1^1, ..., x_1^i, y_1), (x_2^0, x_2^1, ..., x_2^i, y_2)\}$, $y_i \in \{C_1, C_2, C_3, ...C_n\}$
OUTPUT: $O = \{(C_1{=}80), (C_2{=}10), ..., (C_i{=}10)\}$, the $H(\cdot)$ scores on average of each C to be used as gamification points for each C

$D = \{\}$
$\theta = \text{train}(X_L)$
for every $x_i \in X_U$ **do**
  using Eqn 1 to calculate entropy of X_U according to model θ
  $d =$ map the predicted class for $H(x_i)$ by $P_\theta(y|x_i)$ is returned
  $D = D + d$
average scores of $H(\cdot)$ of D for each C and assign into O

Table 1 Nomenclature reference

Symbol	Summary
X_L	set of initial labeled instances (x,y)
X_U	a pool of labeled instances
C	set of activity classes
x, y	input data instance and corresponding label
$(x^1, x^2, ..., x^i)$	the sequence of feature vectors
θ	model
$H(\cdot)$	entropy
O	set of $H(\cdot)$ scores on average of each activity class to be used as gamification points for each activity class

3.2 Traditional gamification points using the accuracy

Accuracy [6] is one traditional metric for evaluating classification models; therefore, it can measure the performance of activity data collection as well. Thus, in this context, we will use the accuracy as gamification points for a baseline method to compare with our proposed method. To use the accuracy as gamification points, we train a machine learning algorithm with the collected smartphone sensor data and activity labels of each user. We then evaluate accuracy of a classifier for gamification points. However, such a strategy is often not feasible in reality due to a problem of accuracy paradox [22] (i.e., when a model may have a high level of accuracy but be too crude to be useful). For example, if the incidence of 'Walking' is dominant, being found in 99% of cases, then predicting that every case is 'Walking' will have an accuracy of 99%. Thereby, we superinten to compute gamification points reasonably for users to avoid the accuracy paradox problem: where a dataset is unbalanced; the overall accuracy is not representative of the true performance of a classifier. The following formula is used to calculate the accuracy as traditional gamification points:

$$gamification\ points = accuracy^1/c \qquad (2)$$

by bending the curve using accuracy to the power of one over classes, the traditional gamification points will be weighted by the inverse of classes, where **c** is the number of activity classes by the user. Additionally, it is as well to motivate activity data collection for more classes.

3.3 Experimental setup

An overview of our experimental design is explained in detail in Table 2. The participants were required to carry Android phones in their pants pockets, install the app on the phones that we extend from [16] by including notifications, to select and record their daily life activities from the list of predefined labels (depicted in Figure 2), get notifications of gamification points, and submit data to our server. Each participant performs the experiments for 6 days (12 hours from 8 AM to 8 PM). We do and repeat our processes as we described in the method section every 50 minutes.

To compare our proposed method with others, we created notifications on smartphones that displays 3 different versions. Each notifications version only differs in the user interface and algorithm for calculating gamification points. Each participant will receive all three conditions of the notifications, each of which will show 2 days. We randomly display the conditions for participants to ensure that they are not affected by the day of experiments for each term. We also request the users click the push notifications sent to assure that the users have seen the notifications. Then, we collected log events when the user clicks on notifications; these log events are such valuable to get insight into message delivery and user engagement.

4. Experimental evaluation

In this section, we will evaluate the proposed method using a standard activity recognition process by comparing its performance with other methods (Table 2). Then, we will explore the notification logs collected and examine how users have engaged with notifications in each method.

ⓒ 2019 Information Processing Society of Japan

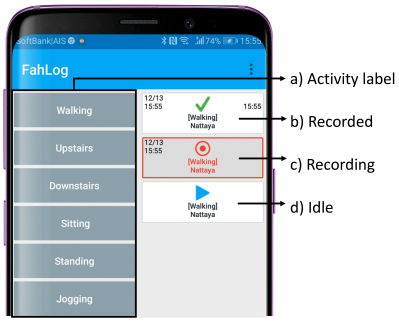

Figure 2 An android app for collecting sensor and labels

Table 2 Experimental design

Method	Conditional detail
Proposed	Receive notifications of the proposed points (Alg 1)
Traditional	Receive notifications of traditional gamification points (Equ 2)
Without	Receive notifications with messages "What are you doing?", but without gamification points

4.1 Activity recognition with smartphone sensors

Since we propose a standard activity recognition chain and a supervised learning approach for evaluations, we first preprocess the dataset collected and then evaluate it.

4.1.1 Data description

The dataset was collected between May 2019, from 11 subjects within an age bracket of 21-26 years, performing one of 12 regular activities (as shown in the left column of Table 3) while carrying an Android smartphone (Wiko Tommy3 Plus) in the pants pockets that recorded the movement data (accelerometers in smartphones). The total number of labels is 1,236.

Table 3 Number of activities for each activity class

Activity class	labels	Activity class	labels
Walking	410	Running	3
Sitting	370	Standing	213
Downstairs	61	Upstairs	50
Lying	40	In vehicle	32
Cycling	25	On train	15
Phone	6	Carrying	11
Total = 1,236 labels			

4.1.2 Data preprocessing

We put together the dataset by including 3-axis accelerometer sensor data and the activity labels on the smartphones without clock and time synchronization because the sensor and the labeling system are both in the same device.

We used sliding windows of 1 minute with no overlapping. For each axis, average, standard deviation, maximum value and minimum value were extracted as features. An example of feature extraction is shown in Table 4. Before data proceeding, we excluded missing values. As a result, we obtained multivariate data of 21,132 samples with 12 variables for feature vectors.

Figure 7 shows the activity labels distribution of the data samples in our dataset. It is worth noting that the distribution is highly skewed, where some classes appear more frequently than others. Since imbalanced dataset can negatively influence the generalization and reliability of supervised learning algorithms [12], we employed the SMOTE algorithm: Synthetic Minority Over-sampling Technique as presented in [3] (an oversampling technique that creates new synthetic data samples in the minority classes, varying the features values of the existing data points based on their k nearest neighbors in the feature space) in order to balance our dataset. We create a balanced dataset by matching the number of samples in the majority class with resampling from the minority class. By upsampling the data separately for the training set and test set, by splitting this will give us a balanced training set and a balanced test set.

Table 4 An example of feature extraction

Feature	value
$mean_x$	num -0.82054216867469876 ...
max_x	num -0.622 ...
min_x	num -1.207 ...
sd_x	num 0.085123482931909022 ...
$mean_y$	num 1.3659708029197057 ...
max_y	num 5.468 ...
min_y	num -4.118 ...
sd_y	num 1.1740472194146572 ...
$mean_z$	num 9.819719626168224 ...
max_z	num 9.909 ...
min_z	num 9.742 ...
sd_z	num 0.894526836753883 ...

4.1.3 Evaluation method

In this section, we present the effectiveness of the proposed method when we give gamification points using active learning through smartphone notifications. The experiment was designed to test the performance of our classifier for a user-dependent scenario. In this case, the classifiers were trained and tested for each individual with her/his own data, and average accuracy and was computed. We show that several machine algorithms have improvements in the classification performance. We also present the proposed method has improvements in the number of data collected compared to the traditional method. To evaluate the proposed method using a technique of supervised learning algorithm for multiclass classification. We trained each participant separately using several standard machine learning classifiers, including Linear discriminant analysis (LDA), k-nearest neighbors (KNN), Decision tree (CART), Naive Bayes (NB), Support-vector machine (SVM), and Random Forest (RF).

To test the model's ability we used stratified k-fold cross-validation. The folds are made by preserving the percentage of samples for each class to ensure each fold is a good representative of the whole. To account for label imbalance, the model performance was presented using the weighted average of precision, recall, F1-score of each class for the multiclass task. So the average is weighted by the support, which is the number of samples with a given label.

4.2 Measuring user engagement

User engagement refers to the quality of the user experience with notifications on a smartphone. In this section, we

ⓒ 2019 Information Processing Society of Japan

propose Click-through rate (CTR) [13] to assess users' depth of engagement with each notifications version displayed. To measure CTR, we use the click logs collected; the CTR fomula is defined as follows:

$$CTR = (\frac{Total\ measured\ clicks}{Total\ measured\ notification\ impressions}) \times 100 \quad (3)$$

where **'Total measured clicks'** are the total amount of clicks on notifications and **'Total measured notification impressions'** are number of times notification was sent on smartphones (which were counted by Google Analytics for Firebase service [18]).

5. Results

Following the evaluation approach discussed above, we report our results of the validation together with a discussion of such results. We show the proposed method has improvements in data quality (the classification performance) compare to the traditional method. The average classification performance of all models results are shown in Figure 3. We also present the proposed method has improvements in data quantity (the number of data collected) compare to the traditional method. Figure 7 shows the number of collected activity labels for both methods.

5.1 Quality of collected activity data

Figure 3 shows F1-score, precision, and recall performance results of all machine learning models were improved with our proposed method compared to the traditional method and without method. Comparing the proposed to traditional. The F1-score was improved from 0.6378 to 0.7733 **(+0.1355)**. The precision was improved from 0.6634 to 0.7987 **(+0.1353)**. The recall of improved from 0.6488 to 0.7721 **(+0.1233)**. Comparing the proposed to without. The F1-score was improved from 0.6240 to 0.7733 **(+0.1493)**. The precision was improved from 0.6500 to 0.7987 **(+0.1487)**. The recall of improved from 0.6381 to 0.7721 **(+0.134)**.

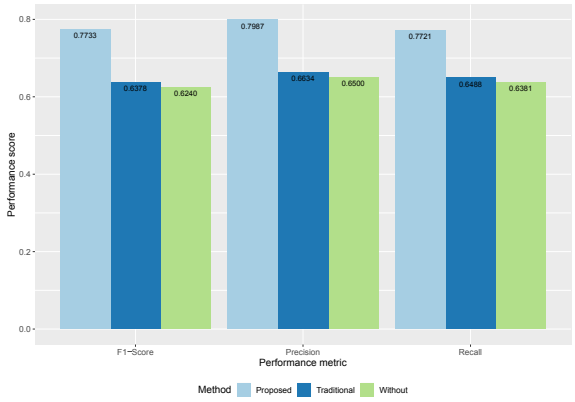

Figure 3 The average classification performance of all models for each method

Figure 4 shows F1-score performance results of all machine learning models were improved with our proposed method compared to the traditional method and without method. Comparing the proposed to traditional. The F1-score of CART was improved from 0.674 to 0.790 **(+0.116)** The F1-score of KNN was improved from 0.697 to 0.811 **(+0.114)**. The F1-score of LDA was improved from 0.654 to 0.785 **(+0.131)**. The F1-score of NB was improved from 0.435 to 0.624 **(+0.189)**. The F1-score of RF was improved from 0.714 to 0.833 **(+0.119)**. The F1-score of SVM was improved from 0.654 to 0.797 **(+0.143)**. Comparing the proposed to without. The F1-score of CART was improved from 0.656 to 0.790 **(+0.134)** The F1-score of KNN was improved from 0.686 to 0.811 **(+0.125)**. The F1-score of LDA was improved from 0.607 to 0.785 **(+0.178)**. The F1-score of NB was improved from 0.469 to 0.624 **(+0.155)**. The F1-score of RF was improved from 0.703 to 0.833 **(+0.13)**. The F1-score of SVM was improved from 0.623 to 0.797 **(+0.174)**.

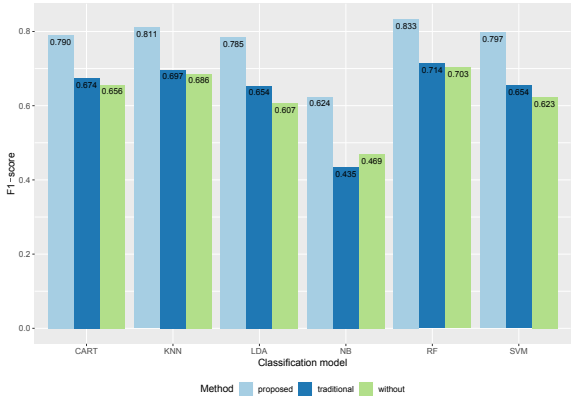

Figure 4 The F1-score performance results of several machine learning models

Figure 5 shows precision performance results of all machine learning models were improved with our proposed method compared to the traditional method and without method. Comparing the proposed to traditional. The precision of CART was improved from 0.703 to 0.821 **(+0.118)**. The precision of KNN was improved from 0.696 to 0.808 **(+0.112)**. The precision of LDA was improved from 0.644 to 0.776 **(+0.132)**. The precision of NB was improved from 0.607 to 0.792 **(+0.185)**. The precision of RF was improved from 0.720 to 0.833 **(+0.113)**. The precision of SVM was improved from 0.611 to 0.762 **(+0.151)**. Comparing the proposed to without. The precision of CART was improved from 0.683 to 0.821 **(+0.138)**. The precision of KNN was improved from 0.679 to 0.808 **(+0.129)**. The precision of LDA was improved from 0.602 to 0.776 **(+0.174)**. The precision of NB was improved from 0.622 to 0.792 **(+0.170)**. The precision of RF was improved from 0.703 to 0.833 **(+0.13)**. The precision of SVM was improved from 0.610 to 0.762 **(+0.152)**.

Figure 6 shows recall performance results of all machine learning models were improved with our proposed method

ⓒ 2019 Information Processing Society of Japan

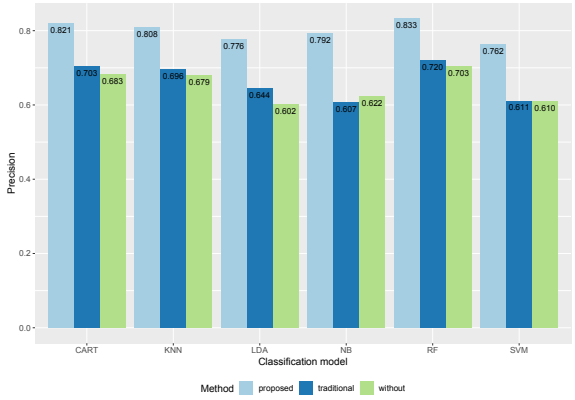

Figure 5 The precision performance results of several machine learning models

compared to the traditional method. Comparing the proposed to traditional. The recall of CART was improved from 0.658 to 0.769 **(+0.111)**. The recall of KNN was improved from 0.705 to 0.818 **(+0.113)**. The recall of LDA was improved from 0.684 to 0.803 **(+0.119)**. The recall of NB was improved from 0.400 to 0.557 **(+0.157)**. The recall of RF was improved from 0.715 to 0.838 **(+0.123)**. The recall of SVM was improved from 0.730 to 0.847 **(+0.117)**. Comparing the proposed to without. The recall of CART was improved from 0.639 to 0.769 **(+0.13)**. The recall of KNN was improved from 0.702 to 0.818 **(+0.116)**. The recall of LDA was improved from 0.637 to 0.803 **(+0.166)**. The recall of NB was improved from 0.450 to 0.557 **(+0.107)**. The recall of RF was improved from 0.710 to 0.838 **(+0.128)**. The recall of SVM was improved from 0.690 to 0.847 **(+0.157)**.

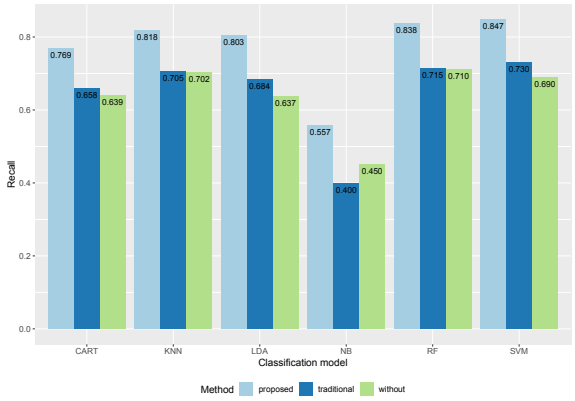

Figure 6 The recall performance results of several machine learning models

5.2 Quantity of collected activity data

As we can see from Figure 7, the number of collected activity labels was increased with our proposed method. The number of activity labels was increased from 409 to 498 **(+89)** compared to traditional method. The number of activity labels was increased from 329 to 498 **(+169)** compared to without method.

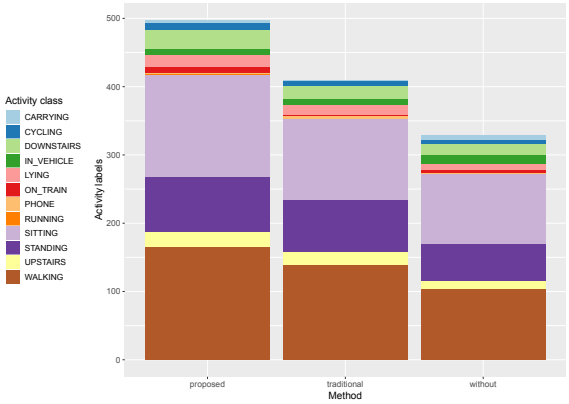

Figure 7 The number of activity labels for each method

Table 5 The number of activity labels of each activity class for each method

Activity class	Proposed	Traditional	Without
Walking	166	140	104
Sitting	150	118	102
Standing	81	77	55
Downstairs	27	18	16
Upstairs	21	18	11
Lying	18	14	8
Cycling	11	7	7
In vehicle	9	10	13
On train	8	2	5
Phone	2	3	1
Carrying	4	1	6
Running	1	1	1
Total	**498**	**409 (+89)**	**329 (+169)**

5.3 User engagement

As we can see from Figure 8, the percentage of CTR was increased with our proposed method. The percentage of CTR was increased from 78.3% to 80.4% **(+2.1%)** compared to traditional method. The number of activity labels was increased from 14.3% to 80.4% **(+66.1%)** compared to without method.

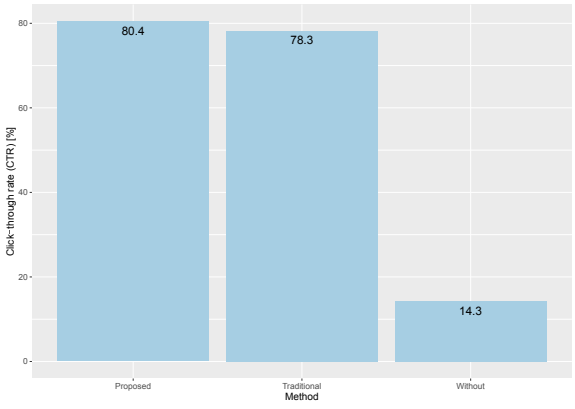

Figure 8 CTR for each method

6. Discussion and Future directions

By evaluating with the dataset and comparing with the traditional method, the results reflect that our proposed method has improvements in data quality for all machine learning models evaluated, data quantity, and user engagement that indicate improvements in activity data collection.

ⓒ 2019 Information Processing Society of Japan

RF achieves the highest F1-score at 83.3%, and also has the most improvements by 81.9%. RF achieves highest the precision at 83.3%, NB has the most improvements by 18.5%. SVM achieves highest the recall at 83.3%, LDA has the most improvements by 16.6%. While this study enabled us to improve activity data collection effectively, there are some limitations that we would like to point out and reference in the future.

In this work, we did not explore task interruptions in mobile notification systems (i.e., the duration of the gamification points sent to the user was designed without considering interruptibility and task performance when interrupted). According to existing work, the onslaught of interruptions from notifications has caused many people to choose to disable (or not enable) notifications for particular applications [19]. Hence, activity data collection can be interrupted by a poorly timed mobile notification as well. In future work, it is important to focus on the proper time for interruption of notification when notifying gamification points to users. Examining opportune moments for interruptions might produce better results.

We were also not able to ascertain who was not clicking on notifications because they did not see it or they dismissed it. We do know the total amount of clicks on notifications and number of times notification was sent on smartphones, but with our log access we did not handle when a user dismisses a notification sent. We believe that capturing this event will valuable to help us get insight into message delivery and user engagement.

There are relevant models of user engagement that were not tracked, such as user segmentation (e.g., daily active users, their demographics, their age), time in the app, etc. It would be better to scrutinize several approaches to measure user engagement, not just CTR. For instance, the time a user spends on the mobile app is useful to take a more in-depth look at user engagement. There may be a group of users that have not been spending much time in the app for one reason or another, so by measuring time spent, we can ask questions of them to find out what is causing this. We then can motivate the user to spend more time on data collection that another way to optimize activity data collection. Collecting more user engagement metrics are left to future work.

In the future, we are also interested in analyzing the relation between classification performance, a number of collected activity labels, CTR analysis as well as activity classes to show whether and how strongly pairs of variables are related. For example, *does CTR affect the classification performance?*, *do gamification points affect the number of activity labels and classes?* Answering these questions, it would also be helpful to understand user motivations and support activity data collection further.

7. Conclusion

We have proposed a method to use an uncertainty based active learning approach to measure the efficiency of collect activity data as gamification points for optimizing activity data collection in smartphone-based activity recognition. The proposed method was validated with mobile sensors and 1,236 activity labels that we collected from 11 participants. By evaluating with the dataset, the results show our proposed method had improvements in the performance of all supervised learning classifications, improved the average F1-score and precision by 0.14 at maximum and recall by 0.13 at maximum. Also, the proposed method had improvements in the number of activity labels by 169 at maximum compared to the other methods. Furthermore, through CTR analysis, our proposed method had a minor improvement in the rate of notification responses compared to the traditional method that reflects better user engagement.

References

[1] Ling Bao and Stephen S Intille. Activity recognition from user-annotated acceleration data. In *International conference on pervasive computing*, pages 1–17. Springer, 2004.

[2] Leo Breiman. *Classification and regression trees*. Routledge, 2017.

[3] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[4] Sebastian Deterding, Miguel Sicart, Lennart Nacke, Kenton O'Hara, and Dan Dixon. Gamification. using game-design elements in non-gaming contexts. In *CHI'11 extended abstracts on human factors in computing systems*, pages 2425–2428. ACM, 2011.

[5] Tom Diethe, Niall Twomey, and Peter Flach. Bayesian active transfer learning in smart homes. In *ICML Active Learning Workshop*, volume 2015, 2015.

[6] T Fawcett. An introduction to roc analysis pattern recognition letter. 2006.

[7] Zachary Fitz-Walter, Dian Tjondronegoro, and Peta Wyeth. Orientation passport: using gamification to engage university students. In *Proceedings of the 23rd Australian computer-human interaction conference*, pages 122–125. ACM, 2011.

[8] Zachary Fitz-Walter and Dian W Tjondronegoro. Exploring the opportunities and challenges of using mobile sensing for gamification and achievements. In *UbiComp 11: Proceedings of the 2011 ACM Conference on Ubiquitous Computing*, pages 1–5. ACM Press, 2011.

[9] Juho Hamari, Jonna Koivisto, Harri Sarsa, et al. Does gamification work?-a literature review of empirical studies on gamification. In *HICSS*, volume 14, pages 3025–3034, 2014.

[10] Yu-chen Ho, Ching-hu Lu, I-han Chen, Shih-shinh Huang, Ching-yao Wang, Li-chen Fu, et al. Active-learning assisted self-reconfigurable activity recognition in a dynamic environment. In *Proceedings of the*

ⓒ 2019 Information Processing Society of Japan

2009 IEEE international conference on Robotics and Automation, pages 1567–1572. IEEE Press, 2009.

[11] HM Sajjad Hossain, Md Abdullah Al Hafiz Khan, and Nirmalya Roy. Active learning enabled activity recognition. *Pervasive and Mobile Computing*, 38:312–330, 2017.

[12] Bartosz Krawczyk. Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, 5(4):221–232, 2016.

[13] Janette Lehmann, Mounia Lalmas, Elad Yom-Tov, and Georges Dupret. Models of user engagement. In *International Conference on User Modeling, Adaptation, and Personalization*, pages 164–175. Springer, 2012.

[14] Zakkoyya H Lewis, Maria C Swartz, and Elizabeth J Lyons. What's the point?: a review of reward systems implemented in gamification interventions. *Games for health journal*, 5(2):93–99, 2016.

[15] Rong Liu, Ting Chen, and Lu Huang. Research on human activity recognition based on active learning. In *2010 International Conference on Machine Learning and Cybernetics*, volume 1, pages 285–290. IEEE, 2010.

[16] Nattaya Mairittha, Tittaya Mairittha, and Sozo Inoue. A mobile app for nursing activity recognition. In *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*, pages 400–403. ACM, 2018.

[17] Nattaya Mairittha, Tittaya Mairittha, and Sozo Inoue. Optimizing activity data collection with gamification points using uncertainty based active learning. In *Proceedings of the 2018 ACM International Joint Conference and 2019 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*. ACM, 2019. (to appear).

[18] Laurence Moroney. Google analytics for firebase. In *The Definitive Guide to Firebase*, pages 251–270. Springer, 2017.

[19] Martin Pielot, Karen Church, and Rodrigo De Oliveira. An in-situ study of mobile phone notifications. In *Proceedings of the 16th international conference on Human-computer interaction with mobile devices & services*, pages 233–242. ACM, 2014.

[20] Burr Settles. Active learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 6(1):1–114, 2012.

[21] Maja Stikic, Kristof Van Laerhoven, and Bernt Schiele. Exploring semi-supervised and active learning for activity recognition. In *2008 12th IEEE International Symposium on Wearable Computers*, pages 81–88. IEEE, 2008.

[22] Francisco J Valverde-Albacete and Carmen Peláez-Moreno. 100% classification accuracy considered harmful: The normalized information transfer factor explains the accuracy paradox. *PloS one*, 9(1):e84217, 2014.

[23] Niels Van Berkel, Jorge Goncalves, Simo Hosio, and Vassilis Kostakos. Gamification of mobile experience sampling improves data quality and quantity. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(3):107, 2017.

ⓒ 2019 Information Processing Society of Japan