Sensors and Materials, Vol. 32, No. 9 (2020) 2999–3017
MYU Tokyo

S & M 2318

Personalizing Activity Recognition Models by Selecting Compatible Classifiers with a Little Help from the User

Trang Thuy Vu and Kaori Fujinami*

Department of Computer and Information Sciences, Tokyo University of Agriculture and Technology,
2-24-16 Naka-cho, Koganei, Tokyo 184-8588, Japan

(Received May 2, 2020; accepted August 11, 2020)

Keywords: human activity recognition, machine learning, wearable sensors, personalization

In daily life, people perform activities every moment differently from one another. Thus, it is necessary to develop a robust system that can recognize human activities and cope with their individual differences. In this article, we propose a new method of individualizing a classifier by choosing the most suitable one based on the estimation of compatibility with a set of classifiers, which we call compatibility-based classifier personalization (CbCP). To make CbCP effective and reduce the burden on the user, the number of activities that a user needs to perform to provide data should be as small as possible. We propose two methods of ranking activities that are as effective in estimating the compatibility as using all activities: difference-based and correlation-based approaches. Additionally, we evaluated four methods of handling a case when more than two classifiers have the same level of compatibility, i.e., multi-compatible classifier handling, random choice, average compatibility reference, and ensemble classification with and without weighting. An offline experiment was carried out using two public datasets, i.e., Physical Activity Monitoring for Aging People 2 (PAMAP2) and Daily Life Activities (DaLiAc), to understand the characteristics of these methods. The results showed that the correlation-based method for activity ranking and the average compatibility reference for multi-compatible classifier handling are the best combination in terms of classification performance, the burden on the user, and computational complexity.

1. Introduction

The noninvasive monitoring of human activities using mobile and wearable devices is gaining considerable attention in various application domains such as fitness,[1] sports,[2] healthcare,[3] and work performance management[4] owing to the enhanced computational and processing capabilities of these devices. In general, machine learning and deep learning technologies are used to identify an activity label of a particular time period,[5] in which a recognition model is trained in advance using a dataset obtained from a certain number of people. A single "recognizer" or "classifier" is often built for all prospective users, which is commonly known as a user-independent[6] or one-fits-all (OFA) classifier.[7] The generalizability of the person-independent classifier often poses an issue regarding real-world

*Corresponding author: e-mail: fujinami@cc.tuat.ac.jp
https://doi.org/10.18494/SAM.2020.2917

ISSN 0914-4935
https://myukk.org/

use because people have individual characteristics of movement and physical properties such as age and gender. The recognition performance improves when a larger number of people provide their data because of the increasing degree of heterogeneity.[1–10] Therefore, a large number of people are required to make the recognition system robust for new users; however, it is quite challenging to build a human activity recognition system from a large amount of data with sufficient heterogeneity.

The other end of the classifier performance enhancement technique is to adjust the recognition system to individual users. This is called a user-dependent or personalized classifier approach. Personalization techniques have already been practically applied in web-based systems such as search and recommendation systems, in which the provided contents are adjusted to individual users.[11] A straightforward approach is to ask the user to collect a training dataset by himself/herself at the beginning of using the system; however, although the effectiveness of this approach is well known[9,12–14] and a user-friendly user interface may support the user with annotating collected data,[14] building a classifier from scratch is burdensome for users, especially in cases of activities of people with diseases, such as Parkinson's disease, infrequently occurring activities of vulnerable people, such as falls of children and elderly people, and activities that are difficult to achieve, such as running at the speed of athletes.

Model adaptation techniques have been proposed to accelerate the personalization process. A personalized activity recognition system can be made by adjusting the weights in fusing multiple classifiers without the user's intervention, which is considered to be a hyperparameter adaptation approach.[15] The unsupervised adjustment of the thresholds of decision trees to the user also fits this category.[16] These methods are challenging because the user's intervention is not assumed. In our previous work, a classifier personalization method was proposed to choose one classifier from a classifier pool based on the compatibility with the target user, which we called compatibility-based classifier personalization (CbCP).[17] Here, the term "compatibility" represents the capacity of using a classifier trained without the target user as if it were trained with his/her data. We assume that there is a compatible classifier for each user because typical ways of performing activities exist in a group of people in general. Although a promising result was obtained in a preliminary experiment,[17] a critical issue is that the compatibility metric is calculated from data obtained from all types of activities. This means that a new user needs to perform all activities at the beginning of using the system, which can be quite burdensome for the user. Therefore, in this article, we propose a method of selecting effective activities from an existing dataset, aiming at only listing a set of activities that have the same capability of identifying a compatible classifier as that when selecting all activities.

We consider that CbCP is complementary to active learning.[18] In active learning, a learning algorithm itself specifies unlabeled data for learning and a human annotator provides labels as answers. Thus, the recognition system can gradually adapt to the user by starting with a "semi-finished" or base classifier through the use of the device.[19] In Ref. 20, a framework that accelerates active-learning-based personalization by choosing a semi-finished classifier based on the compatibility with data given by the user was proposed. In the framework, other components that support remembering the label anytime when the user is available and

motivating the user to perform labeling were provided. Therefore, by incorporating effective activity selection into the CbCP framework, the user's burden would be significantly reduced.

The remainder of this article is organized as follows. In Sect. 2, the notion of CbCP and the extension of identifying a compatible classifier with a set of activities are presented. Also, a method of ranking effective activities is proposed. Furthermore, experimental settings including the description of datasets are presented. Section 3 shows the results and discussion, which is followed by a conclusion in Sect. 4.

2. Methods

In this section, we describe CbCP and the experimental methodology to evaluate the idea of CbCP as well as its functional components.

2.1 Basic idea of CbCP

CbCP chooses the most compatible classifier based on information from the user at the beginning of the system's operation [Fig. 1(a)], rather than using a single common classifier provided for all users [Fig. 1(b)]. The metric of compatibility can be any metric that shows classification performance characteristics according to the design of the recognition system such as accuracy and F-measure (F1-score). The same features are used for classification and for calculating compatibility. The classifiers whose compatibility metrics are evaluated for selection are called *candidate classifiers* or simply *candidates*. The candidate classifiers can be formed in many ways, such as by taking any possible combination of people who provide training data and making groups from all the data as heterogeneous as possible to match as many users as possible. By contrast, in a traditional method, only one classifier is built from all collected data and shared with all the users, which is often called OFA classifier formation.

Let us assume that there are N candidate classifiers with the names $C_{i\in[1,N]}$ and that the compatibility between a new user and a candidate classifier C_i using the data of an entire set of target activity $\boldsymbol{A}$ is represented as $M_{A,i}$. The classifier to be used for the user is represented as

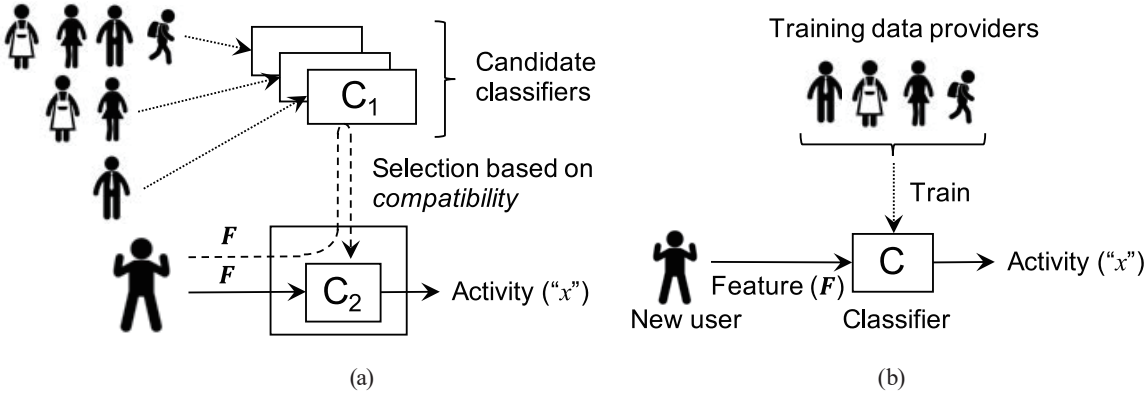


Fig. 1. Schematic diagrams of (a) CbCP-based classification and (b) OFA-based classification.

$C_{k\in K_A}$, where $\boldsymbol{K}_A$, a set of classifier indices, is defined by Eq. (1). $\boldsymbol{K}_A$ may contain the indices of more than two classifiers that have the same compatibility with the user's data. Therefore, any element in the set can be chosen as the classifier to be used in such a case. Note that this principle is extended in the next section.

$$\boldsymbol{K}_A = \left\{ \underset{i\in[1,N]}{\arg\max}\, M_{A,i} \right\} \tag{1}$$

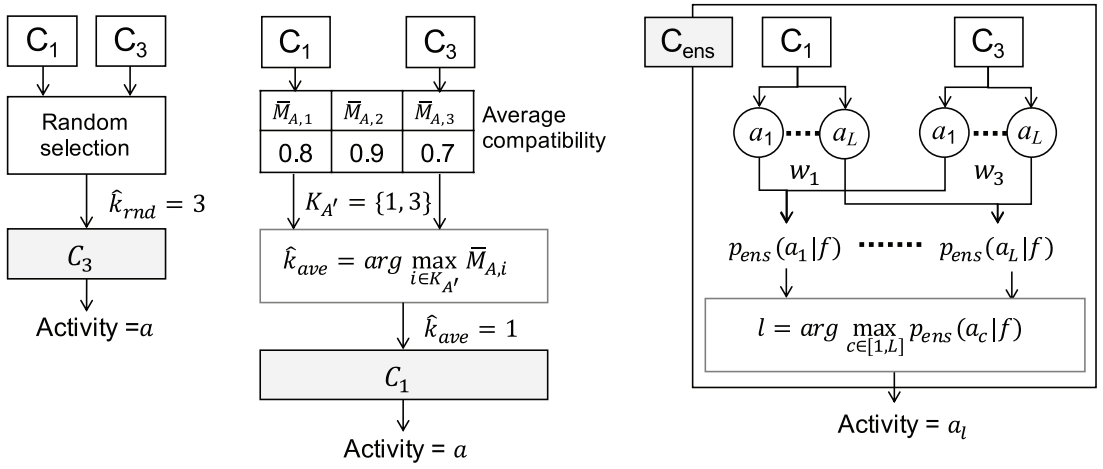
The notion of CbCP can be applied to hierarchical classifier formation, which deals with a microscopic view of compatibility, i.e., per group of activities. A hierarchical classification consists of more than two layers of classifiers. The top layer has one classifier, while the lower layers have more than two classifiers that classify more concrete activities with increasing layer depth. The hierarchical approach is expected to improve the overall classification performance because the compatibility becomes more concrete for a particular group of activities. In our previous work,[17,21] the effectiveness of CbCP over OFA was examined, in which the data of all supported classes, i.e., activities, were used for calculating the compatibility. The result showed that the CbCP approach outperformed the OFA approach in both flat and hierarchical methods.

2.2 CbCP using a subset of target activities

In the above basic idea, the compatibility metric is calculated using the data of all activities. This indicates that a new user is requested to perform all the activities to collect data for this purpose. In the case of activity recognition with a large number of activities, the burden on a user would be large. Thus, the activities the user is asked to do should be limited, which we assume to be determined in advance in a manner presented in Sect. 2.3. In this section, the formulation of the classifier personalization based on the selection of a candidate classifier with limited activities is presented.

The candidate classifiers (C_i) are trained with data of all activities in set $\boldsymbol{A}$. The compatibility metric ($M_{A',i}$) can be calculated using the data of subset $\boldsymbol{A'}$ of an entire activity set $\boldsymbol{A}$. Referring to Eq. (1), the set with the most compatible classifiers calculated from activity subset $\boldsymbol{A'}$ is represented by $\boldsymbol{K}_{A'}$. Unlike the case in which an entire activity set $\boldsymbol{A}$ can be used in the compatibility calculation process, it does not mean that either one of the classifiers $C_{j\in K_{A'}}$ can be used because the calculated compatibility $M_{A',i}$ is not identical to $M_{A,i}$. Thus, an actual classification performance may vary depending on the classifier(s) finally used. This requires appropriate handling methods in the case that multiple candidates have the greatest compatibility, which we call the multi-compatible classifier handling method, and we propose three approaches: (1) random choice, (2) average compatibility reference, and (3) ensemble classification. Figure 2 illustrates these approaches, which assume that two candidate classifiers, C_1 and C_3, have the same compatibility regarding the subset of activities $\boldsymbol{A'}$, i.e., $M_{A',1} = M_{A',3}$.

The random choice approach is very straightforward: one candidate classifier ($C_{\hat{k}_{rnd}}$) is selected randomly at the beginning or at any time during use. $\hat{k}_{rnd}$ given by Eq. (2) is the index



(a) (b) (c)

Fig. 2. Flows of multi-compatible classifier handling methods with an assumption that two classifiers (C_1 and C_3) have the same values of compatibility: (a) random choice, (b) average compatibility reference, and (c) ensemble classification.

of the chosen classifier and *random* is a function that returns an index in set $\boldsymbol{K}_{A'}$. An actual classification is carried out using $C_{\hat{k}_{rnd}}$. In Fig. 2(a), C_3 is chosen for use in the classification.

$$\hat{k}_{rnd} = random(\boldsymbol{K}_{A'}) \tag{2}$$

The average compatibility reference is a deterministic approach defined by Eq. (3), in which $\hat{k}_{ave}$ and $\bar{M}_{A,i}$ are the index of the chosen classifier and the average compatibility of classifiers C_i, respectively. The average compatibility is assumed to be the compatibility for a general population, not that for a particular person. The average compatibilities are calculated using an existing dataset, in which the data obtained from an individual person are used for constructing the candidate classifiers. In Fig. 2(b), let us assume that four persons (P_1, P_2, P_3, and P_4) provided their data to train three classifiers, C_1, C_2, and C_3. For example, the data from P_1 are used to train C_1, while C_2 is trained using the data from P_2 and P_3. Each classifier is tested with the data of each person, and the resultant compatibilities are averaged. In the example in Fig. 2, $\boldsymbol{K}_{A'}$ consists of the indices of classifiers 1 and 3. Thus, by comparing the average compatibilities $\bar{M}_{A,1}$ and $\bar{M}_{A,3}$, i.e., 0.8 and 0.7, respectively, the index of the chosen candidate $\hat{k}_{ave}$ is 1, and thus C_1 is used for this user. Note that, in practice, the compatibility of a classifier trained by data including those of the person to be tested is excluded in the averaging process because the condition is nonrealistic. In practice, the data of a new user are not included in the training data. Thus, the compatibility obtained in such a way needs to be eliminated. Although there can be more than two $\hat{k}_{ave}$ even in this case, any element in the set can be used for the same reason as in Sect. 2.1, and the classification is carried out using one of the $C_{\hat{k}_{ave}}$.

$$\hat{k}_{ave} = \arg\max_{i \in K_{A'}} \bar{M}_{A,i} \tag{3}$$

The ensemble classification approach utilizes all candidates in $\boldsymbol{K}_{A'}$. Ensemble classification in this case involves calculating average subsequent probabilities over the candidates and finding the activity that has the maximum posterior probability, which is often called soft voting in an ensemble classification paradigm.[22] Let the posterior probability of class a_c for a given feature vector $\boldsymbol{f}$ calculated by classifier C_s be represented by $p_s(a_c \mid \boldsymbol{f})$ and $\boldsymbol{w}$ be the weight vector. The posterior probability of ensemble classifier ($p_{ens}(a_c \mid \boldsymbol{f})$) is obtained as Eq. (4), in which w_s represents the normalized weight assigned to classifier C_s. The class that has the largest posterior probability (a_l) is chosen as the output of the classifier as shown in Eq. (5). Regarding the weighting, we propose two approaches: unweighted and weighted approaches. In the unweighted approach, the outputs of classifiers are just averaged, so it can be regarded as an equal-weighted approach. By contrast, we use the average compatibilities ($\bar{M}_{A,i \in K_{A'}}$) via normalization as weights in the weighted approach, in which the outputs of classifiers with larger weights are more likely to be reflected in the final decision. Figure 2(c) shows the structure of the ensemble classifier (C_{ens}) using two classifiers C_1 and C_3 and the activity recognition process.

$$p_{ens}(a_c \mid \boldsymbol{f}) = \sum_{s \in K_{A'}} w_s \cdot p_s(a_c \mid \boldsymbol{f}) \tag{4}$$

$$l = \arg\max_{c \in [1,L]} p_{ens}(a_c \mid \boldsymbol{f}) \tag{5}$$

2.3 Estimating effectiveness of individual activities

The compatibility of a classifier is obtained by using the data of all activities, which means that a new user is requested to perform the activities at the very beginning of the system. When the number of activities is large, it is burdensome for the user. To address this issue, the number of activities should be reduced. In other words, a subset of activities, which represents the compatibility equivalent to that obtained using all activities, should be found. Such limited activities are regarded as "effective". We propose two approaches to estimate the effectiveness of an individual activity: difference-based and correlation-based approaches.

2.3.1 Difference-based approach

A metric of the effectiveness of an activity is represented by the gap between the maximum compatibility obtained using all activities ($\hat{M}_A$) and the estimated compatibility using a single activity ($\hat{M}_{a \in A}$), as shown in Eq. (6), in which $\hat{M}_A$ is represented by Eq. (7). $\hat{M}_a$ represents the overall compatibility assuming that the candidate(s) having the highest compatibility obtained using activity a is utilized to classify the data of all activities. First, a set of classifier indices with the highest compatibility with respect to activity a is identified. This means that $\boldsymbol{K}_{A'}$ in Sect. 2.2 is obtained, in which $\boldsymbol{A'}$ consists of only one activity a. Then, by following the multi-

compatible classifier handling methods, $\hat{k}_{rnd}$ and $\hat{k}_{ave}$ are obtained for $M_{a,\hat{k}_{rnd}}$ and $M_{a,\hat{k}_{ave}}$, which correspond to $\hat{M}_a$ of the random choice and average compatibility reference methods, respectively. Moreover, $\hat{M}_a$ for the ensemble classification approach is obtained as a result of classifying the data of all activities using the ensemble classifier (C_{ens}) consisting of all the classifiers in $\boldsymbol{K}_{A'}$. An ideal case is that the difference (δ_a) is zero, meaning that the classifier estimated with the data of a particular activity a has equivalent classification performance to that chosen using the data of all activities.

$$\delta_a = \hat{M}_A - \hat{M}_a \tag{6}$$

$$\hat{M}_A = \max_{i\in[1,N]} M_{A,i} \tag{7}$$

The hyperparameters in machine learning models are often determined automatically by testing possible combinations as well as empirically. In automatic hyperparameter tuning, a technique called cross-validation is often utilized. We perform leave-one-person-out cross-validation (LOPO-CV) to specify the most effective activity. Figure 3 shows this process. Let us assume that an entire dataset consists of the data obtained from P persons. The candidate classifiers are trained by the data from $P-1$ persons (from P_2 to P_P in the first column of Fig. 3, for example), while the data from one particular person (P_1 in this case) is used to calculate the compatibility metric M and the associated δ_a for activity a. This process is repeated P times by changing the target person and the average $\overline{\delta_a}$ is obtained. The average values are calculated for all activities in the activity set ($\boldsymbol{A}$). A smaller $\overline{\delta_a}$ value indicates that the corresponding activity is more effective.

2.3.2 Correlation-based approach

The second approach is to use the correlation between the compatibility using all activities ($M_{A,i}$) and that using a particular activity a ($M_{a,i}$). The idea behind this approach is that an activity

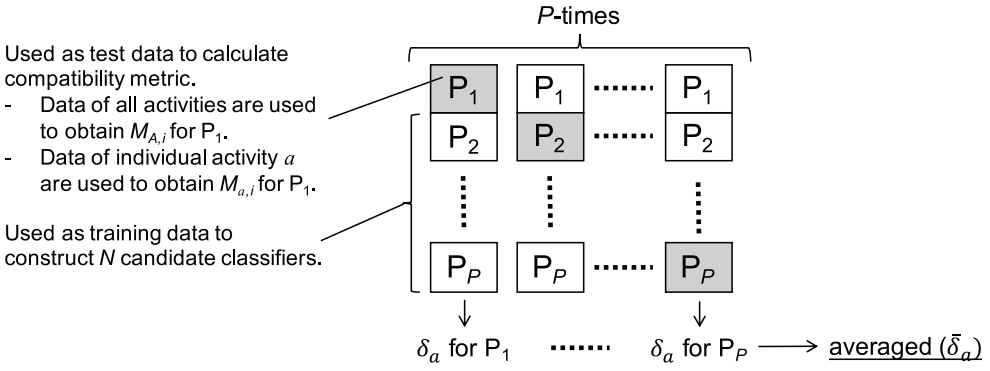


Fig. 3. LOPO-CV scheme for calculating the metric of effectiveness of particular activity (δ_a).

that has a higher correlation with the compatibility using all activities should be more likely to represent the global characteristics of all target activities. We use the Pearson correlation coefficient to represent the correlation. r_a is calculated from all combinations of P persons and candidate classifiers (C_i), as illustrated in Fig. 4. Given that there are N candidate classifiers and P persons in the collected dataset, up to $N \times P$ compatibility metrics are obtained. As described in Sect. 2.2, the compatibility between the data from a person and a classifier trained by the data containing his/her data was excluded when calculating the value. The effectiveness increases with r_a. Unlike the difference-based approach, the correlation-based approach does not need to specify the classifiers to be used. In other words, the effectiveness metric (r_a) is directly calculated from $M_{A,i}$ and $M_{a,i}$ and does not depend on the multi-compatible classifier handling method, making the calculation process simpler than the difference-based approach.

2.4 Experiment

The objectives of the experiment are to evaluate (1) the effectiveness of CbCP using limited activities for specifying a compatible classifier(s), (2) the effectiveness of the methods of ranking activities, and (3) the effectiveness of the methods of handling multiple candidates that have the same compatibility.

2.4.1 Methodology

The first objective is addressed by confirming that the classification performance using a compatible classifier(s) chosen by limited activities is better than that obtained by both an OFA classifier and CbCP using all activities. The classification performance for CbCP is calculated by changing the size of the effective activity subset $\boldsymbol{A'}$. Thus, prior to the calculation, the effectiveness of individual activities is evaluated in the ways proposed in Sect. 2.3. The activity subset $\boldsymbol{A'}$ is extended in order from the most effective one. Thus, given that L activities are subject to recognition, the size of $\boldsymbol{A'}$ varies from one to L. The size L is a special case in which all activities are used, i.e., $\boldsymbol{A'} = \boldsymbol{A}$, and the most burdensome for the user. If the classification performance using a reduced activity subset is higher than that with the OFA classifier, CbCP

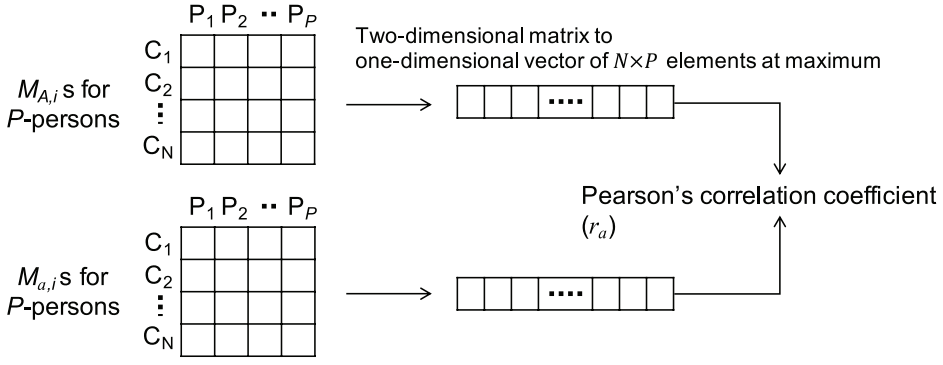


Fig. 4. Scheme for calculating Pearson's correlation coefficient (r_a) between the compatibility obtained using the data of all activities ($M_{A,i}$) and that obtained using the data of a particular activity a ($M_{a,i}$).

will be proved to be a feasible solution for obtaining a good classification result, where the user's involvement is needed but limited.

Regarding the second objective, the difference-based effective activity estimation (Sect. 2.3.1) and correlation-based estimation (Sect. 2.3.2) are compared with respect to the size of the activity subset that shows comparable classification performance to the OFA classifier obtained in the experiment for the first objective and to the case of all activities ($\boldsymbol{A}$).

The third objective is verified by comparing four methods of handling multiple classifiers presented in Sect. 2.2, i.e., random choice (RND), average compatibility reference (AVE), weighted ensemble classification (ENS_W), and unweighted ensemble classification (ENS_UW).

To realistically evaluate the classification performance, the data from a person used for a test are not used in training candidate classifiers or in finding effective activities. The candidate classifiers are generated by combining data from persons who are not subject to the test. Suppose that the data from Q persons can be utilized as the training dataset, then the number of candidate classifiers is $\sum_{i=1}^{Q} {}_{Q}C_i$. The special case with $i = Q$ represents the classifier being trained by the data from all (Q) persons, which is equivalent to the case with OFA classification.

Note that the F-measure is used as the evaluation criterion throughout the experiment, which is the harmonic mean of recall and precision. Recall is the ratio of the number of true positives, i.e., correctly classified cases, to the total number of positive cases, while precision is the ratio of the number of true positives to the total number of cases classified as positive. We implement an offline experiment system using the Application Programming Interface (API) of the Weka machine-learning toolkit,[23] in which a Random Forest classifier is used as a classification model. The number of estimators in Random Forest is set to 100.

2.4.2 Dataset and dataset-specific settings

To investigate the applicability of the proposed methods, we use two public datasets: Physical Activity Monitoring for Aging People 2 (PAMAP2)[24] and Daily Life Activities (DaLiAc).[25] The PAMAP2 dataset contains data of 18 different physical activities performed by nine persons who wear three inertial measurement units (IMUs) and a cardiac rhythm monitor. To calculate the compatibility metric, the data to be used should contain the same activities. Therefore, we choose seven persons who have 10 common activities, which include a wide variety of body movements and postures, as summarized in Table 1. The IMUs consist

Table 1
Numbers and names of activities in the two datasets used in the experiment.

Dataset	# of activities	Activities (abbreviation)
PAMAP2	10	ascending stairs (AS), ironing (IR), standing (ST), cycling (CY), lying (LY), vacuum cleaning (VC), descending stairs (DS), Nordic walking (NW), walking (WK), sitting (SI)
DaLiAc	13	sitting (SI), lying (LY), standing (ST), washing dishes (WS), vacuum cleaning (VC), sweeping (SW), walking (WK), ascending stairs (AS), descending stairs (DS), treadmill running (TR), bicycling on ergometer w/ 50 W (CYL), bicycling on ergometer w/ 100 W (CYH), rope jumping (RJ)

of a three-axis accelerometer and a three-axis gyroscope with a sampling frequency of 100 Hz, which were attached to the wrist, chest, and dominant side's ankle, and a heart rate monitor with a sampling frequency of up to 9 Hz. Although the sampling frequencies of the inertial and heart rate sensors are different, data were recorded in one file synchronously, with the label "NaN" for the nonsensing period of time of the heart rate sensor. Therefore, such periods of time are linearly complemented using two adjacent measured values before feature calculation. The features are calculated in a window of 512 samples (= 5.12 s) overlapping by 50% in accordance with existing work on activity and context recognition.[24,26,27] Nine features from the time and frequency domains, i.e., mean, median, standard deviation, peak, absolute integral, peak frequency, power ratio of the frequency bands 0–2.75 and 0–5 Hz, energy, and spectral entropy, are calculated for the *x*-, *y*-, and *z*-axes of the three accelerometers on the body. Additionally, three Pearson correlation coefficients are included in the times series data. The data from the heart rate sensor attached to the chest are used to calculate the mean and normalized mean, resulting in 83 features in total.

The evaluation is carried out in the LOPO-CV scheme, in which the data of six persons are utilized to train candidate classifiers, and the effectiveness of activities is evaluated using the data of the six persons. This means that Q in Sect. 2.3.1 is six. The data of one remaining person are used for the test. This process is iterated by changing the test person seven times and an average F-measure is obtained. The number of candidate classifiers is 63 ($= \sum_{i=1}^{6} {}_6C_i$) for each test person. The numbers of candidate classifiers, training persons, and test persons, as well as the scheme of the test, are shown in Table 2.

The DaLiAc dataset consists of inertial sensor data captured from 19 persons performing the 13 daily activities shown in Table 1. Four IMUs (three-axis accelerometers and three-axis gyroscopes) are attached to the right hip, chest, right wrist, and left ankle. The sampling rate is 204.8 Hz. The features are calculated in both the time and frequency domains in accordance with Ref. 25. Four types of time domain features, i.e., minimum, maximum, and mean amplitudes and the variance of amplitudes, are utilized. As frequency domain features, the spectral centroid and bandwidth are used. The six features are calculated for each axis of one sensor node. Additionally, energy is calculated for the sensor types, i.e., the accelerometer and gyroscope, of a sensor node. The total number of features is 152. A window consisting of 1024 samples (= 5 s) is slid with 50% overlap. Among 19 persons, we specify 13 persons whose data contain at least 10 feature vectors per activity.

Unlike the case with PAMAP2, we split a group of 13 persons into a training group of six persons and a test group of seven persons. Therefore, the number of candidate classifiers is 63. These numbers are shown in Table 2. The average F-measure of seven persons is calculated. The rationale behind this decision is to keep the number of candidate classifiers small; the number of candidate classifiers in the case of Q = 12 reaches 4095, which would require a huge

Table 2
Numbers and test schemes in the experiment.

Dataset	# of candidates	# of training persons	# of test persons	Test scheme
PAMAP2	63	6	7	Leave-one-person-out CV
DaLiAc	63	6	7	Split training and test groups

amount of time for training and evaluation. The formation of an effective candidate classifier is required to reduce the number of classifiers, which will be a target of future work.

3. Results and Discussion

3.1 Effectiveness of CbCP using limited activities in identifying compatible classifier

3.1.1 Effective activities for estimating overall maximum compatibility

Figure 5 shows the metrics indicating the effectiveness of individual activities for estimating the overall maximum compatibility ($\hat{M}_A$) obtained by the (a) difference-based and (b) correlation-based approaches in the PAMAP2 dataset. The values in Fig. 5(a) represent $\overline{\delta}_a$ as defined in Sect. 2.3.1, which is the difference between the maximum compatibility obtained using all activities ($\hat{M}_A$) and the estimated one using activity a ($\hat{M}_a$). The values are the averages of seven persons in the training dataset, and each bar represents the method of handling multiple compatible classifiers. A lower value indicates a more effective activity. For RND, Nordic walking was the most effective activity (0.039), followed by lying (0.049) and descending stairs (0.059). This means that the classifier chosen with the data of Nordic walking based on RND is inferior to that using all activity data by 0.039 (3.9%) in terms of classifying the test data. Nordic walking is also the most effective activity in AVE, with a value of 0.030, followed by lying (0.034) and descending stairs (0.046). In the case of the ensemble methods, lying is the most effective activity, followed by Nordic walking and descending stairs. The three most effective activities are common to all multi-compatible classifier handling methods. Additionally, the least effective activity, i.e., sitting, is also common to the different methods. Note that, as described in Sect. 2.3.1, the difference-based approach calculates the F-measure using the compatible classifier(s) found by a single activity. Thus, the average difference ($\overline{\delta}_a$) is obtained by the handling method.

The value in Fig. 5(b) is Pearson's correlation coefficient (r_a) between the compatibility using all activities ($M_{A,i}$) and that using a particular activity a ($M_{a,i}$) as defined in Sect. 2.3.2.

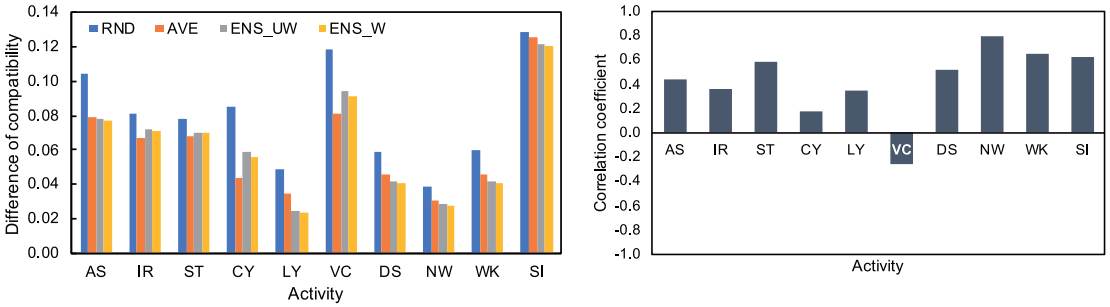


Fig. 5. (Color online) Effectiveness of individual activity in PAMAP2 dataset: (a) difference-based and (b) correlation-based approaches. Note that, in the difference-based method, a smaller value indicates a more effective activity, while in the correlation-based method, a larger value indicates a greater effectiveness.

A higher value indicates that the compatibility using the particular activity is more strongly correlated with that using all activities and thus more preferable. Since the correlation-based approach does not depend on the handling method, the bar shows the average r_a of six persons. From the figure, we can confirm that Nordic walking is the most effective activity (0.803), followed by walking (0.648) and sitting (0.629), and vacuum cleaning is the least effective (−0.259).

Figure 6 shows the effectiveness metrics per activity in the DaLiAc dataset. The way of reading the figure is the same as that of Fig. 5. Generally, treadmill running shows effectiveness in both the difference- and correlation-based approaches, i.e., it is the fourth (0.058), first (0.026), third (0.023), and second (0.023) most effective activity in RND, AVE, ENS_UW, and ENS_W, respectively, as well as the third (0.503) in the correlation-based approach. Rope jumping is also effective in AVE (0.028), ENS_UW (0.023), and ENS_W (0.023), but ineffective in the correlation-based approach (0.072).

Since the F-measure is calculated on the basis of the classification of the data containing all activities, the value depends on the dataset consisting of different activities. Thus, it is natural that the order of effective activities varies, which means that the effectiveness of activities must be evaluated for each dataset. The comparison between the difference-based and correlation-based approaches is presented in Sect. 3.2.

3.1.2 Classification performance by changing size of effective activity subset in identifying compatible classifier

The effectiveness of CbCP with limited activities over OFA-based classification is evaluated with regard to the F-measure by extending the activity subset $\boldsymbol{A'}$ in order of the effectiveness of activity. The F-measures corresponding to the PAMAP2 and DaLiAc datasets are summarized in Figs. 7 and 8, respectively. In each figure, (a) presents the result of the difference-based effectiveness estimation, while that of the correlation-based estimation is presented in (b). The five lines in each figure present the four types of handling method in the case that there are more than two elements in $\boldsymbol{K}_{A'}$ in addition to OFA. An F-measure of 1 indicates the performance in which the most effective activity was used to identify a compatible classifier(s), while the rightmost values (10 and 13 for PAMAP2 and DaLiAc, respectively) are

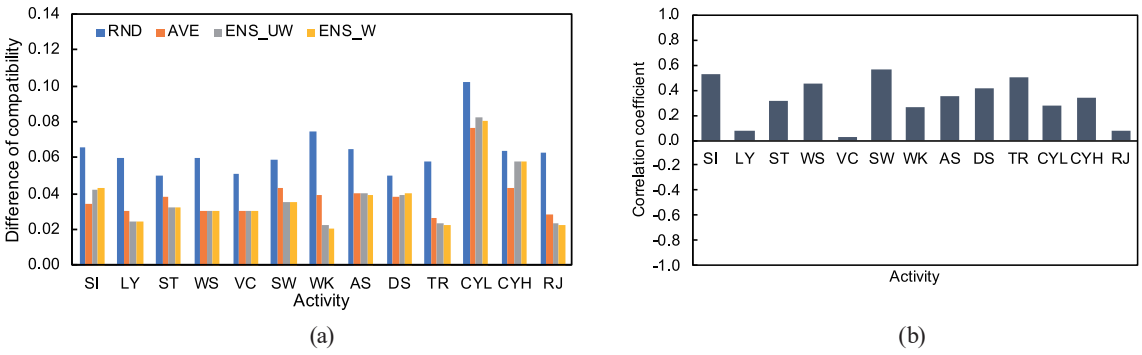
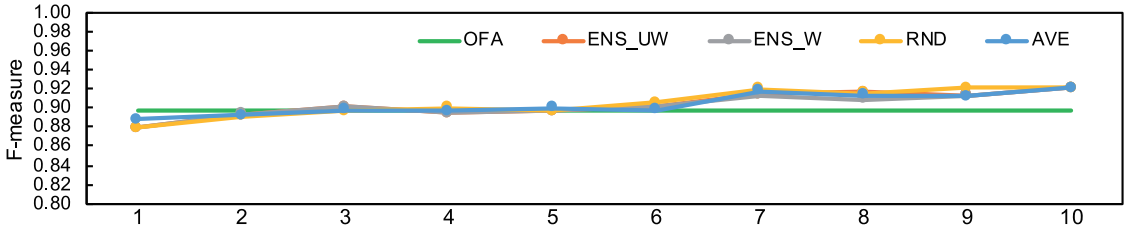


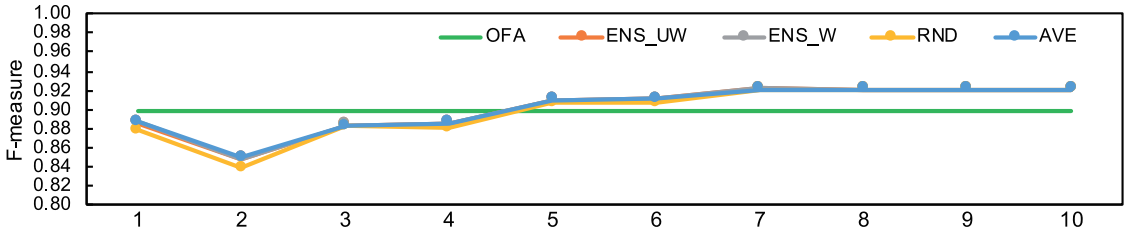
Fig. 6. (Color online) Effectiveness of individual activity in DaLiAc dataset: (a) difference-based and (b) correlation-based approaches. Note that in the difference-based method, a smaller value indicates a more effective activity, while in the correlation-based method, a larger value indicates a greater effectiveness.



	1	2	3	4	5	6	7	8	9	10
OFA	0.898	0.898	0.898	0.898	0.898	0.898	0.898	0.898	0.898	0.898
ENS_UW	0.879	0.893	0.901	0.894	0.897	0.903	0.914	0.916	0.913	0.921
ENS_W	0.879	0.893	0.901	0.894	0.896	0.903	0.913	0.908	0.913	0.921
RND	0.879	0.892	0.897	0.900	0.897	0.905	0.920	0.916	0.921	0.921
AVE	0.888	0.892	0.898	0.896	0.900	0.898	0.917	0.914	0.913	0.921

Number of activities to be asked to the user

(a)



	1	2	3	4	5	6	7	8	9	10
OFA	0.898	0.898	0.898	0.898	0.898	0.898	0.898	0.898	0.898	0.898
ENS_UW	0.886	0.848	0.883	0.885	0.910	0.910	0.921	0.921	0.921	0.921
ENS_W	0.886	0.848	0.883	0.885	0.910	0.911	0.921	0.921	0.921	0.921
RND	0.878	0.838	0.883	0.880	0.906	0.906	0.920	0.921	0.921	0.921
AVE	0.887	0.850	0.883	0.885	0.910	0.911	0.921	0.921	0.921	0.921

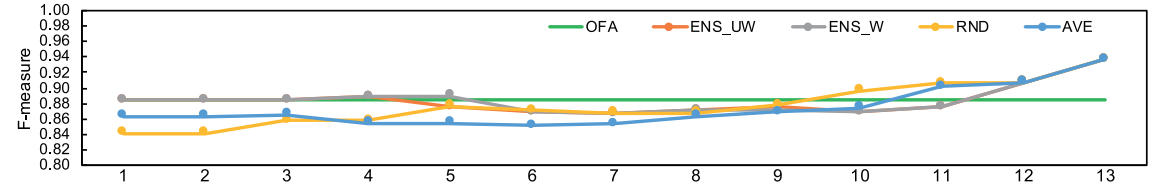
Number of activities to be asked to the user

(b)

Fig. 7. (Color online) Relationship between number of activities and F-measure in PAMAP2 dataset: (a) difference-based and (b) correlation-based approaches.

the performances in which the data of all activities are used. Note that the number of activities is specific to CbCP, and thus the OFA-based approach is not related to the number. However, for comparison, the line for the OFA-based approach, which is distinguished from the others by a line without a marker, is shown in the figures.

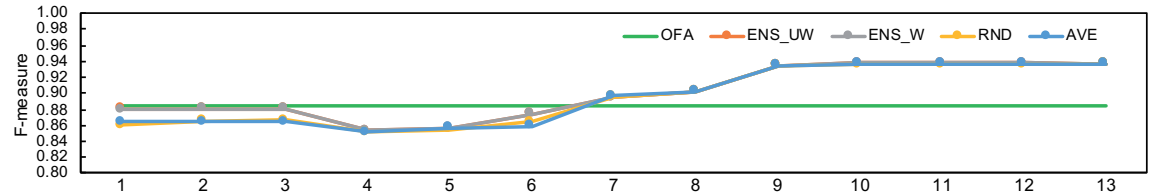
As described in Sect. 2.4.2, the evaluation on PAMAP2 was carried out with the LOPO-CV scheme. Thus, the effectiveness of individual activities varies among the test persons. Table 3 shows the median rank of effectiveness of the test persons. The rank indicates the order of adding to the activity subset A'. Here, DIFF and CORR represent the difference-based and correlation-based activity effectiveness estimation methods, respectively. In the case of DIFF, the four types of multi-compatible classifier handling methods used in conjunction with the difference-based method are presented individually. In the case of DaLiAc, the persons in the entire dataset were split into the training and test data groups. Therefore, the effectiveness of individual activities is common within the multi-compatible classifier handling methods, as summarized in Table 4 by referring to Fig. 6.



	1	2	3	4	5	6	7	8	9	10	11	12	13
OFA	0.884	0.884	0.884	0.884	0.884	0.884	0.884	0.884	0.884	0.884	0.884	0.884	0.884
ENS_UW	0.884	0.884	0.884	0.889	0.875	0.868	0.867	0.871	0.876	0.868	0.875	0.907	0.937
ENS_W	0.884	0.884	0.884	0.889	0.889	0.868	0.867	0.871	0.871	0.868	0.875	0.907	0.937
RND	0.841	0.841	0.857	0.857	0.876	0.871	0.867	0.867	0.878	0.896	0.906	0.907	0.937
AVE	0.863	0.863	0.865	0.855	0.855	0.851	0.853	0.863	0.868	0.874	0.902	0.907	0.937

Number of activities to be asked to the user

(a)



	1	2	3	4	5	6	7	8	9	10	11	12	13
OFA	0.884	0.884	0.884	0.884	0.884	0.884	0.884	0.884	0.884	0.884	0.884	0.884	0.884
ENS_UW	0.880	0.881	0.881	0.853	0.856	0.874	0.895	0.902	0.934	0.938	0.938	0.938	0.937
ENS_W	0.880	0.881	0.881	0.854	0.856	0.874	0.895	0.902	0.934	0.938	0.938	0.938	0.937
RND	0.860	0.865	0.866	0.852	0.855	0.864	0.896	0.902	0.934	0.936	0.936	0.936	0.937
AVE	0.865	0.865	0.865	0.851	0.856	0.859	0.897	0.902	0.934	0.937	0.937	0.937	0.937

Number of activities to be asked to the user

(b)

Fig. 8. (Color online) Relationship between number of activities and F-measure in DaLiAc dataset: (a) difference-based approach and (b) correlation-based approach.

Table 3
Median ranks of effectiveness representing the orders of adding to the activity subset A' in the PAMAP2 dataset.

Method	AS	IR	ST	CY	LY	VC	DS	NW	WK	SI
DIFF+ENS_UW	7	6	7	6	1	9	3	2	4	10
DIFF+ENS_W	7	6	7	5	1	8	3	2	4	10
DIFF+RND	8	6	6	6	2	9	3	1	4	10
DIFF+AVE	8	6	7	4	3	8	5	1	4	10
CORR	6	7	4	9	7	10	5	1	2	3

Table 4
Median ranks of effectiveness representing the orders of adding to the activity subset A' in the DaLiAc dataset.

Method	SI	LY	ST	WS	VC	SW	WK	AS	DS	TR	CYL	CYH	RJ
DIFF+ENS_UW	11	4	7	6	5	8	1	10	9	3	13	12	2
DIFF+ENS_W	11	4	7	5	6	7	1	9	10	2	13	12	3
DIFF+RND	11	7	2	6	3	5	12	10	1	4	13	9	8
DIFF+AVE	6	5	7	3	4	12	9	10	8	1	13	11	2
CORR	2	11	8	4	13	1	10	6	5	3	9	7	12

As shown in the figures, the performance generally increases with the number of activities. The rightmost values are the performances in which all the data are used to find candidate classifiers ($\hat{M}_A$), which are regarded as ground truth or target values. In the case of the PAMAP2 dataset, the value is 0.921, which is much higher than that of OFA (0.898). This means that CbCP is more effective than the traditional approach if a user provides data of all activities at the beginning of the system use. With increasing size of the activity subset, the

performance exceeds that of OFA. This is considered to be a break-even point (BEP) of CbCP. For example, in Fig. 7(a), the BEPs of ENS_UW and ENS_W are observed in the case with three activities, and the F-measure is 0.901. In other words, three activities are required for a higher performance than OFA. According to Table 3, they are lying, Nordic walking, and descending stairs, although the orders in the ranking represent the medians for the test subjects and may be slightly different in an actual calculation. In Fig. 7(b), the performances in the case with seven activities in ENS_UW, ENS_W, and AVE are equivalent to those of the case with all activities, i.e., 0.921. The user's burden of performing activities can be reduced by three activities to obtain the full benefit of CbCP, which could be vacuum cleaning, cycling, and lying or ironing.

A similar tendency can be found in the case of the DaLiAc dataset (Fig. 8). The BEPs of ENS_UW and ENS_W are found in the case of four activities (0.889), in which the activity subsets comprise walking, treadmill running, rope jumping, and lying as shown in Table 4. Even using 12 activities, there are still gaps compared with the case with all 13 activities, although the F-measures themselves are much higher than that obtained by the OFA classifier. By contrast, in the correlation-based effective activity estimation method, the BEPs correspond to seven activities, and the gap between the case of all activities (0.937) and the case of using nine activities is 0.003. This means that it is not necessary to perform vacuum cleaning, rope jumping, lying, and walking.

3.2 Methods of calculating effectiveness metric in estimating compatibility

As described in Sect. 3.1, the difference-based and correlation-based approaches resulted in different orders of effectiveness of individual activities asked of the user. By looking at Figs. 7 and 8, we find that the correlation-based approach tends to reach the BEP earlier than the difference-based approach in both the PAMAP2 and DaLiAc datasets, and also reaches a comparable value to the all-activity cases. For example, in the case of PAMAP2, the selection of a compatible classifier(s) using seven activities showed an F-measure of 0.920 or 0.921 in the correlation-based approach [Fig. 7(b)], while no subset of activities, which was comparable to all-activity cases in the difference-based approach, existed [Fig. 7(a)]. Thus, we consider that the correlation-based approach provides better results than the difference-based approach.

Note that the proposed method deals with ranking the effectiveness of individual activities, which means that the effectiveness does not necessarily represent that of a subset as a whole. Therefore, a subset evaluation method needs to be investigated to identify the best subset activity. We can apply the feature (or attribute) subset evaluation techniques in machine learning.

3.3 Handling methods of multi-compatible classifiers

In Sect. 2.2, four types of methods that handle the issue of multi-compatible classifiers were introduced, i.e., RND, AVE, ENS_UW, and ENS_W, which occurs when more than two classifiers are found to have the same compatibility. The average numbers of compatible classifiers under all experimental conditions were 2.0 and 9.1 for the PAMAP2 and DaLiAc

datasets, respectively. Additionally, the average number of compatible classifiers per size of the activity subsets, i.e., the number of activities required to find a compatible classifier, as well as per activity effectiveness estimation method, is shown in Fig. 9. As shown in the figure, the number of compatible classifiers decreases as the number of activities asked of the user increases. We consider that this is because the diversity of the data increases with the number of activities and that of the test data themselves increases, preventing the F-measure from taking the same value.

By taking into account the discussion in Sect. 3.1.2 that the number of activities that go beyond BEP appears in the latter half of the number of activities, the number of compatible classifiers is one or two. For example, the average values in PAMAP2 are 1.1 and 1.3 for the difference-based and correlation-based effectiveness estimation methods using seven activities, while those of DaLiAc are 1.0 and 1.0 using 12 and nine activities, respectively. Thus, the impact of the multi-compatible classifier handling method seems to be limited in the two datasets.

Nevertheless, we discuss the characteristics of the methods for their future use in other datasets. As shown in Figs. 7 and 8, RND often outperforms other methods with a medium to a large number of activities in the difference-based approach, such as seven and nine in PAMAP2 and 10 and 11 in DaLiAc; however, the value of RND is an average of the results of individual classifications under a condition that only one classifier is used at a time. In other words, it is an expected value of randomly chosen classifiers. Thus, the result could be lower than the average value in some cases. By contrast, the other three methods are deterministic and showed almost the same F-measures when the number of activities was larger than the BEP. By considering the principle of ensemble classification, the computational complexity depends on the number of classifiers. If two classifiers are used, the computational complexity is doubled, and the fusion of the outputs of the two classifiers is an extra process compared with a single-classifier approach. AVE utilizes only one classifier at a time. Thus, we recommend the use of AVE, which has a low computational complexity. By combining AVE with the correlation-based approach, the number of activities can be reduced while keeping the classification performance comparable to the all-activity case.

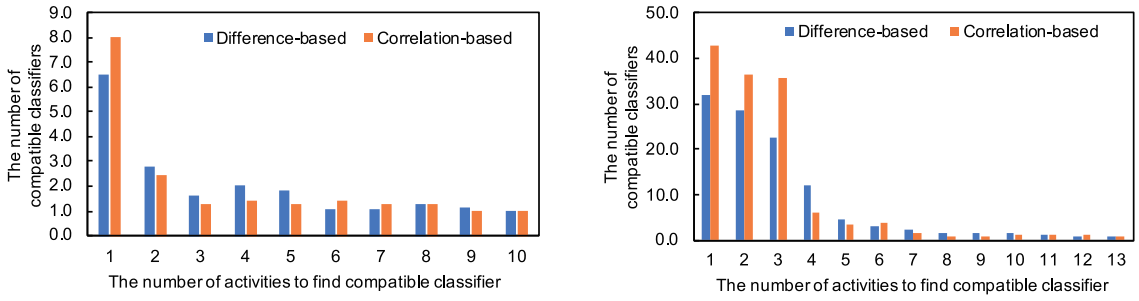


Fig. 9. (Color online) Average numbers of compatible classifiers found by using the data of k-activities: (a) PAMAP2 and (b) DaLiAc datasets.

4. Conclusion

In this article, we proposed an activity recognition system that finds a classifier(s) for each user in a set of pretrained ones (candidate classifiers). The idea behind this approach is that there should exist a suitable classifier for each user, which we call a compatible classifier. The process of finding such a classifier is called CbCP. The compatibility can be best calculated using all activities supported in the recognition system; however, asking every user to perform all activities is burdensome for him/her. Thus, we investigated the difference-based and correlation-based approaches to estimating the effectiveness of activities to identify a subset of activities that are comparable to the case where all target activities are used. However, the classifier selection process may find more than two classifiers that have the same compatibility. Thus, we attempt to resolve this multi-compatible classifier issue by proposing four approaches: random choice, average compatibility reference, and ensemble classification with and without weighting.

Offline experiments were carried out to evaluate the proposed methods using two public datasets: PAMAP2 and DaLiAc. We compared the classification performance, i.e., F-measure, with that obtained by a traditional single classifier (OFA classifier). Also, the performance upon changing the number of activities to find a suitable classifier(s) was compared. The findings throughout the experiment are summarized as follows:

- CbCP outperforms the OFA approach. For example, the maximum F-measures for CbCP and OFA in the PAMAP2 dataset are 0.921 and 0.898, respectively.
- The correlation-based approach reaches a comparable level to an all-activity case faster than the difference-based approach. For example, nine activities are required in the correlation-based approach, while all activities need to be used in the difference-based approach in the DaLiAc dataset.
- The number of compatible classifiers found in the classifier selection process is found to be less than two on average. This indicates that the impact of the number of compatible classifiers on the different multi-compatible classifier handling methods is limited.

By considering the computational complexity, we can conclude that the combination of correlation-based activity effectiveness estimation and the average compatibility reference for multi-compatible classifier handling should be used.

As future work, an efficient subset evaluation method needs to be investigated to find the best subset of activities. Furthermore, an effective candidate classifier generation method needs to be investigated to reduce the number of classifiers required to calculate the compatibility. In addition, compatible classifiers should be efficiently found in a large number of candidate classifiers without evaluating all candidates. Addressing these two issues would improve the processing speed when a user first uses the system.

Acknowledgments

This work was supported by the Japan Society for the Promotion of Science (JSPS) (Grant No. 18H03228).

References

1 A. Subasi, M. Radhwan, R. Kurdi, and K. Khateeb: Proc. 2018 15th Learning and Technology Conf. (IEEE, 2018) 29–34. https://doi.org/10.1109/LT.2018.8368507
2 M. Jones, C. Walker, Z. Anderson, and L. Thatcher: Proc. 2016 ACM Int. Joint Conf. Pervasive and Ubiquitous Computing: Adjunct. (ACM, 2016) 856–860. https://doi.org/10.1145/2968219.2968535
3 M. Gjoreski, H. Gjoreski, M. Luštrek, and G. Matjaž: Sensors **16** (2016) 6. https://doi.org/10.3390/s16060800
4 M. Zhang, S. Chen, X. Zhao, and Z. Yang: Sensors **18** (2018) 8. https://doi.org/10.3390/s18082667
5 D. Thakur and S. Biswas: J. Ambient Intell. Human Comput. (2020). https://doi.org/10.1007/s12652-020-01899-y
6 P. Siirtola and J. Röning: Int. J. Interact. Multimedia Artif. Intell. **1** (2012) 38. https://doi.org/10.9781/ijimai.2012.155
7 J. W. Kamminga, H. C. Bisby, D. V. Le, N. Meratnia, and P. J. M. Havinga: Proc. 2017 ACM Int. Joint Conf. Pervasive and Ubiquitous Computing and Proc. 2017 ACM Int. Symp. Wearable Computers. (ACM, 2017) 597–606. https://doi.org/10.1145/3123024.3124407
8 M. A. Awan, Z. Guangbin, H.-C. Kim, and S.-D. Kim: Int. J. Ad Hoc Ubiquitous Comput. **20** (2015) 3. https://doi.org/10.1504/IJAHUC.2015.073170
9 J. W. Lockhart, T. T. Pulickal, and G. M. Weiss: Proc. 2012 ACM Int. Conf. Ubiquitous Computing (ACM, 2012) 1054–1058. https://doi.org/10.1145/2370216.2370441
10 N. Ogawa, K. Kaji, and N. Kawaguchi: Proc. Int. Workshop on Frontiers in Activity Recognition using Pervasive Sensing (ACM, 2011) 48–51.
11 M. Gao, K. Liu, and Z. Wu: Inf. Syst. Front. **12** (2010) 607. https://doi.org/10.1007/s10796-009-9199-3
12 B. Ao, Y. Wang, H. Liu, D. Li, L. Song, and J. Li: Sensors **18** (2018) 11. https://doi.org/10.3390/s18113604
13 G. M. Weiss and J. W. Lockhart: Proc. Workshops at 26th AAAI Conf. Artificial Intelligence (AAAI, 2012) 98–104. https://www.aaai.org/ocs/index.php/WS/AAAIW12/paper/view/5203/5564
14 J. B. Gomes, S. Krishnaswamy, M. M. Gaber, P. A. C. Sousa, and E. Menasalvas: Proc. 2012 IEEE 13th Int. Conf. Mobile Data Management (IEEE, 2012) 316–319. https://doi.org/10.1109/MDM.2012.33
15 A. Reiss and D. Stricker: Proc. 2013 Int. Symp. Wearable Computers (ACM, 2013) 25–28. https://doi.org/10.1145/2493988.2494349
16 Z. Zhao, Y. Chen, J. Liu, Z. Shen, and M. Liu: Proc. 22nd Int. Joint Conf. Artificial Intelligence (AAAI, 2011). https://www.aaai.org/ocs/index.php/IJCAI/IJCAI11/paper/view/2983/3731
17 T. T. Vu and K. Fujinami: Proc. 2019 Joint 8th Int. Conf. Informatics, Electronics & Vision and 2019 3rd Int. Conf. Imaging, Vision & Pattern Recognition (IEEE, 2019) 97–102. https://doi.org/10.1109/ICIEV.2019.8858540
18 B. Settles: University of Wisconsin-Madison Department of Computer Sciences Technical Report (2009). https://minds.wisconsin.edu/bitstream/handle/1793/60660/TR1648.pdf
19 H. M. S. Hossain, A. A. H. Khan, and N. Roy: Pervasive Mobile Comput. **38** (2017) 2. https://doi.org/10.1016/j.pmcj.2016.08.017
20 K. Fujinami, T. T. Vu, and K. Sato: Proc. 17th Int. Conf. Pervasive Intelligence and Computing (IEEE, 2019) 885–888. https://doi.org/10.1109/DASC/PiCom/CBDCom/CyberSciTech.2019.00162
21 T. T. Vu and K. Fujinami: Proc. 2019 IEEE 8th Global Conf. Consumer Electronics (IEEE, 2019) 544–545. https://doi.org/10.1109/GCCE46687.2019.9014644
22 A. N. Hassan and A. El-Hag: Energies **13** (2020) 7. https://doi.org/10.3390/en13071735
23 University of Waikato Machine Learning Group: https://www.cs.waikato.ac.nz/ml/weka/ (accessed April 2020).
24 A. Reiss and D. Stricker: Proc. 5th Int. Conf. Pervasive Technologies Related to Assistive Environments (ACM, 2012) Article No. 40. https://doi.org/10.1145/2413097.2413148
25 H. Leuteheuser, D. Schuldhaus, and B. M. Eskofier: PLoS One **8** (2013) e75196. https://doi.org/10.1371/journal.pone.0075196
26 K. Fujinami: Information **7** (2016) Article No. 21. https://doi.org/10.3390/info7020021
27 S. Pirttikangas, K. Fujinami, and T. Nakajima: Ubiquitous Computing Systems, Eds. H. Y. Youn, M. Kim, and H. Morikawa (Springer, Berlin, 2006) pp. 516–527. https://doi.org/10.1007/11890348_39

About the Authors

Trang Thuy Vu received her B.S. degree from Hanoi University of Science and Technology (HUST), Vietnam, in 2009 and her M.S. degree from Tokyo University of Agriculture and Technology (TUAT), Japan, in 2012. Since 2018, she has been a Ph.D. candidate at TUAT. Her research interests are in machine learning, activity recognition, human–computer interaction, and ubiquitous computing. (trangit2@gmail.com).

Kaori Fujinami received his B.S. and M.S. degrees in electrical engineering and his Ph.D. degree in computer science from Waseda University, Japan, in 1993, 1995, and 2005, respectively. From 2005 to 2006, he was a visiting lecturer at Waseda University. From 2007 to 2017, he was an associate professor in the Department of Computer and Information Sciences at TUAT. Since 2018, he has been a professor at TUAT. His research interests are in machine learning, activity recognition, human–computer interaction, and ubiquitous computing. (fujinami@cc.tuat.ac.jp).