Research Article

Activity Graph Feature Selection for Activity Pattern Classification

Kisung Park, Yongkoo Han, and Young-Koo Lee

Department of Computer Engineering, Kyung Hee University, Seocheon-Dong, Giheung-Gu, Yongin 449-701, Republic of Korea

Correspondences should be addressed to Young-Koo Lee; yklee@khu.ac.kr

Received 2 January 2014; Accepted 26 March 2014; Published 15 April 2014

Academic Editor: Ruay-Shiung Chang

Copyright © 2014 Kisung Park et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Sensor-based activity recognition is attracting growing attention in many applications. Several studies have been performed to analyze activity patterns from an activity database gathered by activity recognition. Activity pattern classification is a technique that predicts class labels of people such as individual identification, nationalities, and jobs. For this classification problem, it is important to mine discriminative features reflecting the intrinsic patterns of each individual. In this paper, we propose a framework that can classify activity patterns effectively. We extensively analyze activity models from a classification viewpoint. Based on the analysis, we represent activities as activity graphs by combining every combination of daily activity sequences in meaningful periods. Frequent patterns over these activity graphs can be used as discriminative features, since they reflect people’s intrinsic lifestyles. Experiments show that the proposed method achieves high classification accuracy compared with existing graph classification techniques.

1. Introduction

The advances in sensor technology make activity recognition possible. Activity recognition is a technique that automatically recognizes human activities by analyzing sensor data [1–6]. Recently, several studies [7–10] have been performed to analyze activity patterns from an activity database gathered by activity recognition. Activity pattern classification is a technique that predicts class labels of people based on activity patterns. The class labels can be not only individual identification but also meaningful groups such as nationalities, genders, jobs, and hobbies. Therefore, activity pattern classification has many applications which vary the class labels accordingly. For example, recommender systems can recommend similar items to users with the same hobbies.

Accurate classification of activity patterns requires a deep understanding of activity patterns. Activity patterns are styles in which people perform their activities and they reflect people’s lifestyles. People have both intrinsic and common activity patterns. The intrinsic patterns play a key role in distinguishing each individual from the others. They can appear in individuals’ specific activities in different frequencies [11, 12], orders, days, and periods [13]. In order to find these intrinsic activity patterns, we need to explore activity patterns by adjusting the frequencies in all combinations of days. For example, certain people perform their hobbies on weekends, but others may perform the same hobby once a month.

However, it is hard to explore frequent activity graphs in all combinations of days. This naive approach requires a very long running time and produces redundant frequent patterns. The lifestyle of people is itself a solution to solve this problem. People usually repeat similar activity patterns in specific periods such as daily, weekly, monthly, and yearly. Therefore, we only need to explore activity patterns from every combination of days in specific periods. By avoiding the exploration of meaningless combinations, we can reduce the search space in order to find discriminative features. From a classification viewpoint, these features are as effective as the features from all the combinations of days.

The other important point is to determine pattern types for activity pattern classification. The types are determined depending on activity data models. Various activity data modeling studies have been proposed to represent activity data collected from sensors such as statistics, sequence, and graph-based modeling. Statistics-based activity models [7, 8] calculate the average frequency or duration of each activity.
Sequence-based activity models [9] represent activities as daily sequences based on the occurrence time. Graph-based activity models [10] generate activity graphs by combining activity sequences in various periods such as daily, weekly, monthly, and yearly. Nodes and edges of the activity graphs represent activities and the occurrence order between two activities, respectively. Among these modeling techniques, the graph-based activity model is suitable for the activity pattern classification problem. This model can maintain sufficient information to mine the intrinsic activity patterns such as frequencies and occurrence orders and various meaningful periods of activities.

In this paper, we propose a novel feature selection framework for classifying activity patterns effectively. The proposed framework generates activity graphs by combining all combinations of daily activity sequences in each meaningful period. By performing frequent subgraph mining over these graphs, we can find all discriminative frequent activity patterns efficiently, which can reflect individuals’ intrinsic lifestyles. In order to remove redundant frequent patterns, we select highly discriminative features by adopting topological similarity based feature selection (TSFS) [14]. Since topologically similar graphs involve similar activity patterns, we can effectively remove redundancy. Through experiments, we show that the proposed framework can achieve a high performance in classifying activity patterns.

The remainder of this paper is organized as follows. We briefly introduce the existing activity modeling and graph classification studies as related work in Section 2. We analyze the effectiveness of various types of activity patterns in Section 3. In Section 4, we define the activity pattern classification problem and propose our feature selection framework in detail. Section 5 presents the experimental results of the proposed framework, and Section 6 concludes this paper.

2. Related Work

Activity recognition has gained a lot of interest in recent years due to its potential and usefulness in context-aware computing such as smart homes [3–6] and aged care monitoring [7, 8]. The purpose of activity recognition is to infer people’s behaviors from low level data acquired through sensors in a given setting, from which other critical decisions are made. There are two approaches to acquire human activities using sensor systems. First, sensors are attached on the body and the signal readings are interpreted [1, 2]. This approach can recognize low level activities such as “sitting,” “running,” and “walking.” Second, sensors are deployed to objects and devices in the environment and the sensor readings are interpreted [3–6]. This approach can recognize high level activities such as “eating,” “sleeping,” “showering,” and “leaving home.” The low level activities are used for short-term activity monitoring such as the elderly falling down. The high level activities are used for long-term activity pattern monitoring such as healthcare.

Mining techniques generally require appropriate data models to find informative patterns to improve effectiveness or efficiency. Various studies have been proposed to represent activity data collected from sensors such as statistics, sequence, and graph-based modeling. Statistics-based activity models [7, 8] calculate the average frequency and duration of each activity. Large deviations from the average time or number are considered abnormal activity patterns. Sequence-based activity models [9] represent activities as daily sequences based on the occurrence time. Sequential pattern mining techniques can be applied to activity sequences to mine informative patterns. The graph-based activity model [10] generates activity graphs by combining the daily activity sequences in every monitoring period. The activity graphs can maintain various activity related information through multilables of nodes and edges such as frequencies, time, durations, and locations of activities. The main advantage of this activity model is that we can analyze activity patterns in various frequencies and periods.

Graph classification studies [14–18] have been proposed to classify graph modeled data such as chemical compounds, social networks, and XML documents. The techniques represent graphs as feature vectors with values indicating the presence or absence of graph structural features, and a discriminative power of each feature is estimated by feature evaluation criteria including G-tests and information gain (IG). The graphs are then classified by using a machine learning classifier.

The existing techniques mostly adopt frequent subgraphs as graph structural features for classification. Many efficient frequent subgraph mining algorithms have been proposed such as FSG [11] and gSpan [13]. These algorithms enable us to extract frequent subgraph features in practical time. TSFS [14], M$^3$T [15], and LEAP [16] use frequent subgraph features. COM [17] has shown that cofrequent patterns can have high discriminative powers. Structure feature selection [18] has shown that frequent subgraphs have different discriminative powers according to their spatial distribution in a graph database. However, they cannot achieve the high classification performance for activity graphs since they only consider the frequency of the features.

3. Analyzing Activity Patterns in Various Activity Data Models

Individual lifestyles are hidden in one’s frequently performed activity patterns. It has also been shown in literature [19] that frequent patterns are highly discriminative in various classification problems. Therefore, we adopt frequent patterns as features for this activity pattern classification problem.

Frequent patterns involve different information depending on data models. The representative models for activity data are statistics-, sequence-, and graph-based models. In these models, frequent activity patterns include the frequency of activities, frequent activity sequences, and frequent subgraphs, sequentially. We analyze a discriminative power of each type of frequent activity pattern.

The frequency of activities does not involve enough information for activity pattern classification, because it cannot express the occurrence order among activities. People
can have similar frequencies but different activity orders. Therefore, the occurrence order is very valuable information. For example, Figure 2 shows parts of activity sequences in two consecutive days. The frequencies of “sleeping,” “eating,” and “toileting” are 4, 1, and 1 in the activity sequences. From this information, we cannot perceive that the person is suffering from insomnia. This kind of distinct patterns can be helpful in distinguishing people.

Frequent activity sequences involve the occurrence order of activities. However, the orders are valid only in their own sequences, because we have fractions of sequences. Occurrence order relationships among activity sequences, especially when a sequence shares common segments, can provide a more precise occurrence order among activities of different sequences. For example, we can interpret the activity sequences as two different meanings in Figure 1, that is, “eating or toileting” or “eating and toileting” in the middle of sleeping.

Frequent activity subgraphs represent the occurrence order of activities together in multiple activity sequences from various periods. From these graphs, we can get the occurrence order of activities at a similar time in different periods, which is useful knowledge for activity pattern classification. For example, we can accurately interpret an individual’s activity patterns as “eating or toileting” or “eating and toileting” in the middle of sleeping.

Frequent activity subgraphs represent the occurrence order of activities together in multiple activity sequences from various periods. From these graphs, we can get the occurrence order of activities at a similar time in different periods, which is useful knowledge for activity pattern classification. For example, we can accurately interpret an individual’s activity patterns as “eating or toileting” or “eating and toileting” in the middle of sleeping.

4. Classifying Activity Patterns Based on Activity Graphs

The proposed method uses activity database accumulated with unit activities recognized from various sensors. In this paper, we assume unit activities are recognized exactly. We present the proposed feature selection framework for activity pattern classification. In Section 4.1, we define an activity pattern classification problem. We analyze discriminative activity graph features in Section 4.2 and suggest a method that mines the discriminative features efficiently in Section 4.3. In this paper, our scope is limited to knowledge discovery process in Figure 2.

4.1. Notation. In this section, we present notations related to the activity graph approach [8] and the formal definition of the activity classification problem.

Definition 1 (unit activity). A unit activity, \( a_i \), is an activity performed in a certain continuous time. \( a_i \) represents a unit activity, \( a_i \), performed in time \( t \). Unit activities are recognized from sensor data and become nodes in an activity graph.

Definition 2 (activity sequence). An activity sequence, \( s = d_1^t \rightarrow d_2^t \rightarrow \cdots \rightarrow d_m^t \), is a sequence of unit activities, where \( t_i < t_{i+1} \). Any unit activity, \( a_i \), can appear multiple times in different \( t \). The sequence of activities in a day is regarded as a daily activity sequence, \( s_d = d_1^t \rightarrow d_2^t \rightarrow \cdots \rightarrow d_m^t \).

Definition 3 (modeling period). A modeling period, \( p (p \in N) \), is a time unit used to generate activity graphs. \( p \) is generally set to a meaningful number of days such as a week \((p = 7)\), a month \((p = 30)\), or a year \((p = 365)\).

Definition 4 (combination days). Given the modeling period, \( p \), the combination days of \( p \) are all possible combinations of days in \( p \), that is, \( c_p \), \( 1 \leq i \leq p \). We denote them as \( c_p = d_1^i \cdot d_2^i \cdots d_1^m (1 \leq |d_i| \leq p) \). For example, the combination days are \( d_1^1 \cdot d_1^2 \cdots d_1^p \). When \( |c_p| = 2 \).

Definition 5 (activity graph). An activity graph, \( G = (V, E, L, \Sigma) \), is a graph that consists of a set of activity nodes, \( V \), and a set of edges, \( E \), where an edge, \( e \in E \), represents the order between two activity nodes in \( V \). \( L \) is a set of node and edge labels and \( \Sigma \) is a function assigning labels to nodes and edges.

Activity sequences are generated every day. In order to represent activities of more than one day concisely, we combine corresponding activity sequences and generate activity graphs using multiple sequence alignment (MSA) [18]. The number of combined sequences is determined depending on the modeling period, \( p \), and combination days, \( c \).

Figure 3 is an example of an activity graph generated using MSA, when \( p \) is 3. MSA first combines the activity sequences \((s_1 \text{ and } s_2)\) that share the greatest number of common activity nodes. The common activity nodes are represented as a single node and increase the edge label by one. In the same way, activity sequence \( s_1 \) is combined with \( s_2 \) and \( s_3 \).

Definition 6 (activity pattern classification). Given the set of activity sequences, \( S = \{s_1, s_2, \ldots, s_n\} \), for \( m \) people, and a set of class labels, \( \{y_1, y_2, \ldots, y_m\} \), an activity
Sequences and perform frequent subgraph mining on the graphs by combining all combinations of daily activity days, and periods. Therefore, we should generate activity expressed by activity patterns in different frequencies, orders, graph features are mined efficiently.

In this paper, we focus on mining discriminative features for activity pattern classification. In order to mine these features efficiently, we represent activity data as activity graphs by combining every combination of daily activity sequences in each meaningful period and find frequent subgraphs, \( F = \{ f_1, f_2, \ldots, f_n \} \), from the activity graphs. Among \( F = \{ f_1, f_2, \ldots, f_n \} \), we finally select highly discriminative features, \( f^* = \text{argmax} \ q(f) \), to remove redundancy, where \( q(\cdot) \) is a feature evaluation function.

**Definition 7** (frequent activity pattern). A frequent activity pattern, \( f_i \), is a subgraph of \( f \) that occurs no less than the minimum support. A support of subgraph \( f_i \), denoted by support \( (f) \), is calculated as \( |f \subseteq G|/|GD| \), where \( |f \subseteq G| \) is the number of graphs containing subgraph \( f \) and \( |GD| \) is the total number of activity graphs in an activity graph DB.

**Definition 8** (discriminative feature). Given the set of activity graphs, \( G_{cp} = \{ g_1, g_2, \ldots, g_n \} \), with \( cp \) and \( P = \{ p_1, p_2, \ldots, p_m \} \); a discriminative feature is a frequent subgraph that can be mined only in a specific set of activity graphs, \( G_{cp} \).

In this paper, we focus on mining discriminative features for activity pattern classification. In order to mine these features efficiently, we represent activity data as activity graphs by combining every combination of daily activity sequences in each meaningful period and find frequent subgraphs, \( F = \{ f_1, f_2, \ldots, f_n \} \), from the activity graphs. Among \( F = \{ f_1, f_2, \ldots, f_n \} \), we finally select highly discriminative features, \( f^* = \text{argmax} \ q(f) \), to remove redundancy, where \( q(\cdot) \) is a feature evaluation function.

**Theorem 9.** For each \( p \), given the set of activity graphs, \( G_{cp} = \{ g_1, g_2, \ldots, g_n \} \), with \( cp \), \( G_p \) can have discriminative frequent subgraphs, \( F_p = \{ f_1, f_2, \ldots, f_m \} \).

**Proof.** Suppose that we generate all possible combinations within a modeling period; each combination of an activity sequence is unique. Moreover, daily activity sequences are different in a specific period. Therefore, we generate different graphs with each combination of days of \( cp \) and each graph set, \( G_i \), with \( cp \) can have discriminative frequent subgraphs, \( F_p = \{ f_1, f_2, \ldots, f_m \} \).

**Theorem 10.** Given the set of activity graphs, \( G_p = \{ g_1, g_2, \ldots, g_n \} \), with a set of modeling periods, \( P = \{ p_1, p_2, \ldots, p_m \} \), each graph set, \( G_i \), with \( p_i \) can have discriminative frequent subgraphs, \( F_p = \{ f_1, f_2, \ldots, f_m \} \).

**Proof.** We prove it by contradiction. Suppose that a set of frequent subgraphs, \( F_p = \{ f_1, f_2, \ldots, f_m \} \), is mined in \( G_p = \{ g_1, g_2, \ldots, g_n \} \)
\[ \{g_1, g_2, \ldots, g_n\} \] by varying \( p_i \). Assume that, for all \( i \) and \( j \) (\( i < j \)), two conditions, \( F_i - F_j = \emptyset \) and \( F_i - F_j = \emptyset \), are always established. We show the case, not establishing these conditions. Suppose a certain frequent subgraph, \( f \), is in \( G_j \). It is possible that \( f \) does not appear in \( G_i \), since we have \( i < j \). Therefore, we have \( F_i \notin F_j \). This also means that the support of \( f \) in \( G_j \) is no less than the support of \( f \) in \( G_i \). Therefore, \( f \in F_i \land f \notin F_j \) is also possible when we mine frequent subgraphs with the same minimum support in \( G_i \) and \( G_j \). Since \( F_i - F_j = \emptyset \) and \( F_i - F_j = \emptyset \) are not always established, \( F_i \) can have frequent subgraphs, not containing \( F_j \), and vice versa. \( \Box \)

We present an example explaining Claim 1. Figure 4 shows two sets of frequent subgraphs, \( F_2 \) and \( F_{30} \), mined in \( G_2 \) and \( G_{30} \), which denote weekly and monthly patterns, respectively. Suppose that activity patterns \( f_3 \) and \( f_4 \) are for two months. \( f_3 \) is the activity pattern performed the first week in each of the two months. The support of \( f_3 \) is 0.25 (= 2/8) in \( G_2 \) but 1 (= 2/2) in \( G_{30} \). As shown in this case, the supports of frequent subgraphs increase in larger modeling periods. Conversely, suppose that \( f_4 \) is the activity pattern performed every week for a month and then performed for two weeks in the second month. For two months, the support of \( f_4 \) is 0.75 in \( G_2 \) but 0.5 in \( G_{30} \). \( f_4 \) appears to have two different activity patterns in \( G_{30} \) because the activity graph involves the frequency of activities in a modeling period.

**Observation 1.** People repeat similar daily activity sequences in specific periods according to their own lifestyles; that is, \( s_i \equiv s_{i+p} \).

**Claim 1.** Given all the activity sequences for \( n \) days, \( S = \{s_1, s_2, \ldots, s_n\} \), and the set of modeling periods, \( P = \{p_1, p_2, \ldots, p_m\} \), \( F_n \equiv \{F_{p_1} \cup F_{p_2} \cup \cdots \cup F_{p_m}\} \), where \( F_{p_i} \) is a set of frequent subgraphs mined in \( c_{p_i} \).

**Proof.** Based on Observation 1, graph set, \( G_{k_p} \), with \( k \) multiples of \( p \) is similar to the graphs in \( G_{c_{p_i}} \). If \( k = 2 \), \( F_p \equiv F_{2p} \) is established. Suppose \( k \) is a limitation, \( F_p \equiv F_{2p} \equiv F_{3p} \equiv \cdots \equiv F_{kp} \) is established. When we apply the above property to the set of modeling periods, \( P = \{p_1, p_2, \ldots, p_m\} \), \( \{F_{p_1} \cup F_{p_2} \cup \cdots \cup F_{p_m}\} \equiv \{F_{k_p}, F_{k_p}, \ldots \cup F_{k_p}\} \) is established. Almost all of the frequent subgraphs in \( F_{p_i} \) appear in \( k \) multiples of the set of periods. Therefore \( F_n \equiv \{F_{k_p}, \cup F_{k_p}, \ldots \cup F_{k_p}\} \equiv \{F_{p_1} \cup F_{p_2} \cup \cdots \cup F_{p_m}\} \) is satisfied, and \( F_n \equiv \{F_{p_1} \cup F_{p_2} \cup \cdots \cup F_{p_m}\} \) is established. \( \Box \)

As we observed above, people repeat similar activity patterns in very specific periods; that is, \( s_i \equiv s_{i+p} \). Therefore, Claim 1 is convincing. We find discriminative features by \( F_1 \cup F_2 \cup \cdots \cup F_m \) rather than \( F_n \).

### 4.3. Mining Discriminative Features

Based on Claim 1, we mine frequent subgraphs in the activity graphs generated in every \( c_{p_i} \), with \( P = \{p_1, p_2, \ldots, p_m\} \). A set of all activity sequences for \( n \) days, \( S = \{s_1, s_2, \ldots, s_n\} \), is divided into subsets of activity sequences according to each period, \( p_i (i \leq m) \), such as \( S_1 = \{s_1, s_2, \ldots, s_p\}, S_2 = \{s_{p+1}, s_{p+2}, \ldots, s_{2p}\}, \ldots \), and \( S_k = \{s_{(k-1)p+1}, s_{(k-1)p+2}, \ldots, s_n\} \). In each subset of activity sequences, we construct the sets of activity sequences for every \( c_{p_i} \). Activity graphs are generated by combining activity sequences in every \( c_{p_i} \). We then mine frequent subgraphs in the activity graphs.

Figure 5 shows the processing steps of mining frequent subgraphs when \( p = 3 \). \( S = \{s_1, s_2, \ldots, s_n\} \) is divided into \( S_1 = \{s_1, s_2, s_3\}, S_2 = \{s_4, s_5, s_6\}, \ldots \), and \( S_m = \{s_{2n-2}, s_{2n-1}, s_{2n}\} \). In each \( S_i \), activity graphs are generated by combining activity sequences in all possible combination days, \( c_i \). For example, when \( |c_1| = 2 \), activity graphs \( g_1, g_2, \) and \( g_3 \) are generated from combination days \( s_1, s_2, s_3, s_4, s_5, s_6, s_{2n-2}, s_{2n-1}, \) and \( s_{2n} \). The sets of frequent subgraphs, \( F_1, F_2, \) and \( F_3 \), are mined from each set of activity graphs.

Though we generate activity graphs with all possible combinations within the periods, \( p^2 \) graphs are generated for the period, \( p \). Graphs generated in a smaller period are generated again in a larger period. Duplicate generation is a particularly severe problem for combinations of a small number of sequences, since these combinations are involved in most of the larger periods. In order to avoid generating duplicate graphs, we generate graphs from the largest period to the smallest period. In each period, we generate graphs beginning with those that have the largest number of sequences to the smallest number of ones in a combination. In this way, we can avoid generating graphs that were already generated in a larger period. We can efficiently remove a lot of duplicated graphs.

Figure 6 shows an example in which duplicate graphs are generated with all possible combinations in a set of modeling periods, \( P = \{3, 5\} \). When we generate activity graphs with all possible combinations in modeling periods 3 and 5, the combinations of sequences, \( s_1, s_2, \) and \( s_3 \), are duplicated. We generate graphs from the combination days of period 5, \( c_5 \), to period 3, \( c_3 \). In this way, we can avoid generating graphs that were already generated in \( c_5 \).

After generating activity graphs, frequent subgraph mining is performed on the activity graphs to extract frequent activity patterns. A number of duplicate frequent subgraphs occur, since activity graphs are generated from the same activity sequences. Many of them have similar graph structures to each other. These redundant patterns degrade the performance in both the accuracy and running time of the classification.

We select highly discriminative features by removing the redundant frequent subgraphs. A number of feature selection methods [14–20] have been proposed. Among them, we
adopt the TSFS approach [14]. This approach has proved that topologically similar graphs have similar discriminative powers and a method has been proposed that selects discriminative subgraphs by clustering frequent subgraphs based on their similarity. TSFS is very suitable for an activity pattern classification problem since similar activity graphs involve similar activity patterns.

We generate frequent subgraph clusters, \( C_1, C_2, \ldots, C_m \), by clustering the set of frequent subgraphs, \( F = \{ f_1, f_2, \ldots, f_n \} \). For the clustering, we can use any clustering algorithm such as \( k \)-means and any graph similarity measure such as graph edit distance [21] or maximum common subgraph [22, 23]. The highest discriminative frequent subgraph is selected in each cluster. The discriminative power of each frequent subgraph can be estimated by feature evaluation functions such as information gain, mutual information, and \( \chi^2 \) statistic.

Figure 7 shows a processing step that selects highly discriminative features from a set of frequent subgraphs, \( F = \{ f_1, f_2, \ldots, f_n \} \). For example, the frequent subgraphs, \( f_1, f_2, f_7, \) and \( f_{10} \), are topologically similar to each other. These activity patterns also have similar semantics, “sleeplessness.” Therefore, they are clustered into the same cluster, \( c_1 \). The set of highly discriminative features, \( F^* \), is selected by selecting only one highly discriminative feature in each cluster.

We then convert each activity graph, \( G_p = \{ g_1, g_2, \ldots, g_n \} \), into a set of feature vectors. Equation (1) is a feature-graph matrix, \( M_{i,j} \), that indicates whether each feature, \( f_j \), is present or absent in \( g_i \). We build a classification model by training machine learning classifiers with the converted activity graphs and class labels, \( y_i \in \{ y_1, y_2, \ldots, y_m \} \). Consider

\[
M_{i,j} = \begin{cases} 
1 & \text{if } f_j \subseteq G_i, \\
0 & \text{otherwise}.
\end{cases}
\]  

Algorithm 1 is the pseudocode for mining highly discriminative features and training an activity pattern classifier.

Algorithm 1 takes the set of activity sequences, \( S \), the minimum support, \( \text{min_sup} \), and the set of modeling periods, \( P \), as input and returns a set of highly discriminative features, \( F^* \), and the activity pattern classification model, \( H \). It first mines frequent subgraphs and stores them in \( F \) in each period (line 2–4). A set of activity pattern clusters, \( C \), is generated by clustering \( F \) (line 5–9). Finally, each graph set, \( G_i \), is represented as feature vectors by generating the feature-graph matrix, \( M_i \), and the activity pattern classification model, \( H \), is built (line 10–11).

5. Experiments

In this section, we experimentally evaluate the effectiveness of our feature selection framework on a real dataset.

5.1. Experimental Setup. Due to the lack of activity datasets using sensor-based activity recognition, we gathered a real activity dataset for the experiments. The dataset was gathered by eight Kyung Hee University students for a year. Users inputted their activities manually on the activity gathering
Algorithm 1: Activity pattern classifier construction ($S, P, \text{sup}$).

Input:
$S = \{s_1, s_2, \ldots, s_n\}$: A set of Activity sequences
$P = \{p_1, p_2, \ldots, p_m\}$: A set of modeling periods sorted in ascending order
min$_{\text{sup}}$: A minimum support threshold
t$_s$: A topological similarity

Output:
$F^*$: A set of optimal activity patterns
$H$: An activity pattern classification model

1. $F \leftarrow \emptyset$
2. for $p \in P$ do
3. Divide $S$ by modeling period $p$
4. for $i = 1, i \leq p$ do
5. generate graph $g$ with $d_f$
6. $G_i \leftarrow G_i \cup \{g\}$
7. for $i = 1, i \leq p_m$ do
8. $F \leftarrow F \cup \text{fpmine}(G_i, \text{min}_{\text{sup}})$ // Perform frequent subgraph mining
9. $C = \{c_1, c_2, \ldots, c_l\} \leftarrow \text{clustering}(F, t_s)$
10. for each $c_i \in C$ do
11. choose the highest discriminative feature, $f^*$, in $c_i$
12. $F^* \leftarrow F^* \cup \{f^*\}$
13. Generate feature-graph matrix, $M$, by using $F^*$
14. Train model $H$ using $F^*$ and $M$
15. return $F^*$, $H$

5.2 Effectiveness Evaluation. The first experiment aims to show the effectiveness of graph features. We compare classification accuracy among features mined from statistics-

(3) discriminative features in each period: comparison of classification accuracy between proposed method and naive method extracting features in a single period,

(4) effectiveness of TSFS: comparison of classification accuracy among feature selection algorithms, top-$k$, MMRFS [19], and TSFS,

(5) ineffectiveness of conventional graph classification: comparison of classification accuracy between the proposed method and conventional graph classification algorithms such as the model-based search tree ($\text{MbT}$) [15] and maximal marginal relevance feature selection (MMRFS) [19].
sequence-, and graph-based activity data modeling. We select the top 60 discriminative frequent activity patterns as features in each modeling technique. We mine the frequent activity patterns with various minimum supports such as 10, 15, and 20. We set the modeling period to 7 for generating activity graphs.

In Figure 9, the graph-based activity modeling outperforms other modeling techniques over all the minimum supports. As we explained in Section 3, graphs involve sufficient information such as the order of activity occurrences and relationships among activity sequences. This experiment demonstrates that graph-based activity modeling is suitable for an activity pattern classification problem.

The second experiment aims to select the best minimum support threshold for mining frequent activity subgraphs. We vary the minimum support when the topological similarity, \( t_s \), is set to 70%.

In Figure 10, the highest classification accuracy is 72% when the minimum support is 10%. Frequent subgraphs with a small minimum support appear in most of the graphs having any class labels. Frequent subgraphs with a large minimum support appear only in a few graphs. Therefore, frequent subgraphs with too small or too large of a minimum support cannot have high discriminative powers. We chose to set the minimum support threshold to 10%.

The third experiment aims to show that discriminative features are involved in each period. We compare classification accuracy between a fixed period and various periods.

We set the fixed period to 7 and the various periods to \( P_1 = \{1, 2\} \), \( P_2 = \{1, 2, 5\} \), \( P_3 = \{2, 5, 7\} \) and \( P_4 = \{1, 2, 5, 7\} \).

In Figure 11, the results show that frequent subgraphs mined in activity graphs with various periods are highly discriminative according to Theorems 9 and 10. In all the period settings, the proposed method achieves better classification accuracy compared with the fixed period, except for \( P_1 \). The proposed method especially achieves the highest performance when the periods are set to \( P_4 = \{1, 2, 5, 7\} \).

The fourth experiment aims to show the effectiveness of TSFS. We compare the classification accuracy among feature selection algorithms, Top-k, MMRFS, and TSFS. We use

### Table 1: Description of dataset.

|                          |           |            |
|--------------------------|-----------|------------|
| Number of total activities | 35678     |            |
| Number of total users    | 17        |            |
| Type of activities       | 19        |            |

### Table 2: Target activities of dataset.

| Activities | Total number | Average duration (minute) |
|------------|-------------|--------------------------|
| Wakeup     | 1975        | 24.3                     |
| Wash room  | 5491        | 9.9                      |
| Bathing    | 1788        | 20.2                     |
| Dress up   | 1923        | 33.4                     |
| Breakfast  | 1468        | 16.25                    |
| Traveling  | 4677        | 21.8                     |
| Lab work   | 2912        | 167.2                    |
| Meetings   | 349         | 85                       |
| Classes    | 371         | 103                      |
| Lunch      | 1463        | 29.7                     |
| Relaxing   | 2996        | 51.2                     |
| Exercise   | 245         | 31.3                     |
| Exam       | 15          | 87.5                     |
| Dinner     | 1819        | 30.9                     |
| Sleeping   | 2362        | 292                      |
| Cleaning   | 691         | 33.63                    |
| Laundry    | 188         | 72.9                     |
| Cooking    | 1438        | 19.4                     |
| Sickness   | 52          | —                        |
| Personal   | 2697        | 111.37                   |
| Outing     | 403         | 245                      |
| Home visit | 181         | —                        |
| Seminar    | 174         | 130                      |
the same number of features for all the feature selection algorithms. We vary the topological similarities by 65%, 70%, 75%, and 80%.

In Figure 12, the results show that TSFS achieves the highest classification accuracy compared with Top-k and MMRFS over all of the topological similarities. Particularly, TSFS achieves the highest accuracy when the topological similarity is set to 70%. A loose topological similarity loses discriminative features because many features belong to one cluster. A tight topological similarity incurs redundant features because too many features as clusters are generated. From this experiment, we can observe that topologically similar activity graphs involve similar activity patterns.

The fifth experiment aims to show the ineffectiveness of conventional graph classification algorithms for the activity pattern classification problem. We compare the classification accuracy with representative graph classification algorithms, MbT and MMRFS.

In Figure 13, the proposed method achieves a classification accuracy that is about 5% higher than the existing algorithms. The existing algorithms do not consider frequent patterns from various periods, since they are developed for other graph data such as compounds, web page link networks, and social networks. Moreover, the existing algorithms cannot remove redundant frequent activity patterns well, since they do not consider the topological similarity for feature selection.

6. Conclusion

Classification is an important technique for analyzing activity data. We have defined the activity pattern classification problem and proposed an effective feature selection framework for classifying activity patterns. We have shown that a graph model is effective for activity pattern classification because the activity graphs reflect individuals’ specific activities in different frequencies, orders, days, and periods. By analyzing the lifestyles of people that repeat similar activity patterns in specific periods, we have proposed an effective and efficient feature selection technique. We select frequent activity patterns in all combinations of daily sequences in meaningful periods.

Experimental results have shown that the proposed method achieves (1) suitable model for activity pattern classification, (2) better discriminative power extracting combinational features than that extracting features in a single period, and (3) higher accuracy than that of existing graph classification methods. In addition, we discussed the optimal
thresholds of minimum support and topological similarity for activity pattern classification.

Conflict of Interests
The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgment
This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (No. 2012R1A2A2A01047478).

References
[1] L. Bao and S. S. Intille, "Activity recognition from user-annotated acceleration data," in Pervasive Computing, vol. 3001 of Lecture Notes in Computer Science, pp. 1–17, 2004.
[2] J. R. Kwapisz, G. M. Weiss, and S. A. Moore, "Activity recognition using cell phone accelerometers," ACM SIGKDD Explorations Newsletter, vol. 12, no. 2, 2010.
[3] E. Munguia Tapia, S. S. Intille, and K. Larson, "Activity recognition in the home setting using simple and ubiquitous sensors," in Pervasive Computing, vol. 3001 of Lecture Notes in Computer Science, pp. 158–175, 2004.
[4] T. Van Kasteren, A. Noulas, G. Englebienne, and B. Kröse, "Accurate activity recognition in a home setting," in Proceedings of the 10th International Conference on Ubiquitous Computing (UbiComp '08), pp. 1–9, Seoul, Korea, September 2008.
[5] R. J. Robles and T.-H. Kim, "Applications, systems and methods in smart home technology: a review," International Journal of Advanced Science and Technology, vol. 15, 2010.
[6] J. Xiao and R. Boutaba, "The design and implementation of an energy-smart home in Korea," Journal of Computing Science and Engineering, vol. 7, no. 3, pp. 204–210, 2013.
[7] G. Virone, M. Alwan, S. Dalal et al., "Behavioral patterns of older adults in assisted living," IEEE Transactions on Information Technology in Biomedicine, vol. 12, no. 3, pp. 387–398, 2008.
[8] Y. Han, M. Han, S. Lee, A. M. Sarkar, and Y. K. Lee, "A framework for supervising lifestyle diseases using long-term activity monitoring," Sensors, vol. 12, no. 5, 2012.
[9] P. Rashidi and D. J. Cook, "Mining sensor streams for discovering human activity patterns over time," in Proceedings of the 10th IEEE International Conference on Data Mining (ICDM ’10), pp. 431–440, Sydney, Australia, December 2010.
[10] Y. Han, K. Park, and Y. K. Lee, "Graph model based activity pattern mining for healthcare," Journal of KIIEE: Databases, vol. 38, no. 5, 2011.
[11] M. Kuramochi and G. Karypis, "Frequent subgraph discovery," in Proceedings of the 1st IEEE International Conference on Data Mining (ICDM ’01), pp. 313–320, San Jose, Calif, USA, December 2001.
[12] X. Yan and J. Han, "gSpan: graph-based substructure pattern mining," in Proceedings of the 2nd IEEE International Conference on Data Mining (ICDM ’02), pp. 721–724, Maebashi, Japan, December 2002.
[13] J. Han, G. Dong, and Y. Yin, "Efficient mining of partial periodic patterns in time series database," in Proceedings of the 15th International Conference on Data Engineering (ICDE ’99), pp. 106–115, Sydney, Australia, March 1999.