Robust Classification of Step Data of Exercise

Haoqi Fu
Department of Electronic Engineering
Shanghai Jiao Tong University
Shanghai, China
fuliye725@sjtu.edu.cn

Keqin Shi
Department of Electronic Engineering
Shanghai Jiao Tong University
Shanghai, China
keqinshi@sjtu.edu.cn

Weiqiang Sun
Department of Electronic Engineering
Shanghai Jiao Tong University
Shanghai, China
sunwq@sjtu.edu.cn

***Abstract*—The step data of exercise(SDE) is a type of daily life data that is generated by humans and recorded by mobile devices. Collection of such data has become more and more popular, as it provides users with information regarding his/her level of activity, and a stimulus for persisting in an active lifestyle. When SDE becomes more widely available, it also provides health professionals with a new tool to interact with their clients. Most existing methods for classifying exercise modes rely on training models that are manually labeled, or micro-pattern analysis using manually selected lengths and micro-pattern quantities. However, SDE datasets are highly personalized, inaccurate and the cost of these methods is high. So we propose a new method to classify the SDE by utilizing its inherent characteristics. With the understanding that SDE reflects daily lifestyle, is personalized and stable over at least a period of time, this paper takes the understanding of users' exercise preference, obtained from dataset of respective users, as prior knowledge. Then, instead of using raw step data, we quantize the raw data into different levels of exercise and takes into account the moment of exercise occurrence. Results show that our method provides an easy to interpret and robust characterization of user's exercise habit.**

***Keywords*-SDE, smart device, classification, prior knowledge, exercise habit**

I. Introduction

The wave of exercise has spread across the whole nation. More and more scientific evidence indicates the positive impacts of exercise on health. People associate exercise with health and pay more and more attention to their own level of exercise [1]. Recording one's daily activity with smart devices provides users with the information of long term activity level and a stimulus for persisting in an active lifestyle. This leads to the proliferation of activity tracking software and devices [2] [3]. These software and tools record exercise data, such as daily steps, cadence, and types of exercise, which is called step data of exercise(SDE). The wide availability of SDE also provides a tool for health professionals to interact with their customer in a new way[4-7].

Proper ways to understand the SDE is critical for any application of the data. The difficulty lies in the fact that the data may not be consistent in different software or devices. Researches indicate that step counts collected by different platforms may possess a discrepancy of 20% or even more. This means that handling the step records as precise numerical data points may lead to over-stated and unrobust user behavior characterization. To obtain reliable and robust results, inaccuracy should be taken into account and the data should be treated as a rough characterization of user's lifestyle.

Given the fact that the exercise behavior, as an important component in one's lifestyle, is typically stable over time, data generated alongside such behavior should also be stable. The application of such characteristic will greatly facilitate the handling of SDE toward an accurate and reliable description of one's exercise behavior. In this paper, we extract the 'stable part' of one's exercise behavior from the statistical distribution of his/her SDE dataset, and use that as prior knowledge to understand the individual samples in the dataset. With the knowledge, we divide the individual SDE samples into segments. The time of exercise, as well as the total exercise duration, is then calculated for each time segment. A distance function based on time and duration is used to classify the samples. Comparing with existing solutions that often treats SDE datasets as generic time series and use standard methods to do classification, the proposed method in this paper has low computational complexity, is robust over the variability in step quantity, and, most importantly, delivers easy to interpret results. The contributions of this paper include the following:

- We analyzed a large SDE dataset and obtained the characteristics of SDE. We then argue that three quantities related to exercise, instead of the raw step counts, should be used in classifying the SDE samples.
- We propose a new distance function for the SDE. This distance function captures the main features of the exercise and reduces the impact of noise introduced by high-dimensional data. It is simple and efficient.
- We propose a personalized segmentation of SDE samples based on prior knowledge. To the best of our knowledge, we are the first to use this method in the study of exercise data.
- Our approach shows much higher accuracy than the baseline clustering algorithm on the SDE dataset. The results show that our algorithm can capture the exercise

modes with different exercise habits.

II. Related Work

Although there are few studies on SDE, we can get inspiration from the analysis of other activity data. Activity analysis studies follow three general directions. The first approach constructs a model of some preselected activities, and establishes the fitness of this model through methods such as Bayesian Learning [9] and Hidden Markov Models [10]. The obtained models can serve to predict people's house activities, to group the users based on their activity routines, or to identify common activity routines [9]. Model-based methods are commonly applied to datasets of location and exercise sensors. To obtain sound results in their models, researchers study incorporate domain expert knowledge (and perhaps manually annotate the dataset). This requires substantial effort and constrains the quality of the analysis to the extent of the expert's knowledge ahead of the quality of the dataset. On a highly personalized SDE dataset, the implementation of an annotated dataset requires a high level of expert knowledge and significant cost, which is difficult to achieve.

As an alternative, studies from the second approach extract features from frequently occurring patterns, and then construct classifiers based on these features. Subsequently, for classification, studies either apply state-of-the-art supervised learning techniques such as Support Vector Machines, Decision Trees [11] or incorporate custom data structures (like graph-based clustering [12], and routine-tree [13].

The third method breaks the conventional thinking and combine pattern recognition with other research areas. The original time series is matrix-decomposed to form two matrices, and clustered separately. The clustering results of the two matrices are cross-fused to obtain the final cluster [14]. Although a new clustering method is proposed, the characteristics of the exercise time series itself are not taken into account, and the dynamic time warping distance (DTW) [15] is still selected for clustering.

These methods can be well applied in daily life data, but when applied to SDE, the accuracy is going to be severely affected by inaccuracy and variability of step count data. Existing researches indicate that step counts collected by different platforms may possess a discrepancy of 20% or more. This means that handling the step records as precise numerical data points may lead to over-stated and unrobust user behavior characterization.

Therefore, we need to study and analyze the SDE itself in order to design a robust classification method on the basis of understanding the characteristics of SDE.

III. A First Look at SDE Samples

The exercise data analyzed in this paper are a kind of step data recorded by wearable devices or smartphones, which reflect the exercise habits of people. Fig. 1 shows two SDE

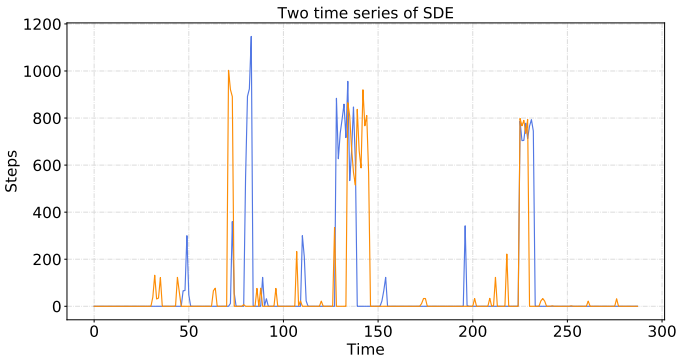

Figure 1. Sample:Two time series of SDE.

samples of a particular user. In the figure, the x-axis is time of the day whose unit is five minutes, and the y-axis is steps. The blue and orange polylines represent two different time series consisting of steps. These values are influenced by sensors of smart devices, which may not accurately reflect the actual number of steps. The steps recorded by different devices may also have deviations. We do not know whether these two time series are recorded by the same device. The value of each point in the time series seems to be a number that is completely randomly distributed between 0 and an upper threshold. From these random numbers, we can't seem to find any regularity.

However, from the two samples, we can see that: 1. Both samples indicate long lasting exercise during the day; 2. The position of the exercise occurs, although there are some offsets on the time axis, are not far apart. 3. The duration of the corresponding exercise is similar.

Our statistical analysis on a large SDE dataset reveals the fact that, even though the step counts in the samples may vary dramatically against time and on different platforms, the duration and time of exercise from different samples can be very reliable. This suggests that instead of using raw step counts as the input to the classification method, we should use the duration and time of exercise, which can be derived from the raw data, to handle SDE samples. Our analysis also suggests that the number of exercise, which can range from 0 to 4 for most people, is also reliable over time. In short, one's daily exercise behavior can be better characterized by: 1. number of exercise, 2. occurrence time of exercise, 3. duration of exercise.

At present, most of the researches on SDE time series do not consider the characteristics of exercise itself. They directly input steps into the algorithm, using Euclidean Distance or Dynamic Time Wrapping Distance to calculate the distance between the series, trying to discover the exercise modes. However, sensors from different smart devices affect the number of steps and in addition, the step number itself is also very volatile. For example, the number of steps generated in two 20-minute exercise may differ by more

than a thousand steps, but should be recognized as the same exercise practice. When a point-to-point Euclidean Distance is used to calculate the distance between two time series, the distance between two similar exercise will be large, even if they only differ by half an hour. Using the Euclidean Distance will result in the algorithm not being able to distinguish between more similar series and completely dissimilar series. A further study of the recognition of exercise modes uses Dynamic Time Warping Distances(DTW) [14] to measure the similarity between series. DTW was first used in the field of speech recognition. This distance algorithm can well describe the similarity of shapes and has a certain degree of tolerance for the expansion and translation of time series in shape. In contrast to Euclidean Distance, DTW considers local misalignments and reports the optimal warping path between the given two series. The DTW Distance between the time series data Q and P can be calculated as:

$$DTW(Q,P) = \min_{W} \left(\sum_{k=1}^{K} d\left(w_k\right) \right) \tag{1}$$

However, due to the characteristics of the DTW itself, when the distance between SDE is calculated using the DTW Distance, the duration of the exercise is lost. DTW focuses on 'similarity', and in speech recognition, similar waveforms represent the same character. The difference between 'similarity' and 'congruence' is that 'similarity' ignores the actual length and only calculates the degree of conformity after scaling. If the DTW Distance is adopted, the distance between the two time series, which generated by exercising one hour and exercising ten minutes, may be 0. That is, the characteristics of the distance algorithm itself cause the length of the exercise time to be ignored. The DTW algorithm also does not describe the distance between exercise time series very well.

Therefore, we hope to design a new method for calculating the distance between SDE time series according to the essential characteristics of the exercise itself, and then further distinguish the user's exercise modes.

IV. Methods

One exercise activity can be naturally characterized by its duration and time of occurrence. Therefore, instead of using raw step counts, we derive the two values from the raw data, and use the two values, together with the number of exercise, as the input for classification. In this section, we introduce the detailed steps of our method.

The overall flow chart is shown in Fig. 2. The main steps taken by our approach are as follows:

- Design a filter to remove the random noise and irregular activities of raw data. The filtered data can reflect the duration and the time of exercise. We name it time series of exercise.
- Superimpose the time series of exercise over the past hundreds of days for a given user and obtain a probability distribution of exercise throughout a day.
- Consider the probability distribution of exercise as prior knowledge and dividing series of exercise activity into segments. Then the duration of exercise is extracted as a feature for each segment. So the time series of exercise can be convert to feature vector consisting of feature points. The feature vector will be used to measure the similarity between SDE series.
- Based on the prior knowledge, quantify the duration of exericse in each segment into three levels. Traversing all combination of these levels, we initialize all classifiers and each classifier represents one exercise mode.
- Euclidean Distance calculated between samples and classifiers will become the basis of classification. The classification is completed after abnormal series and empty classes are removed.

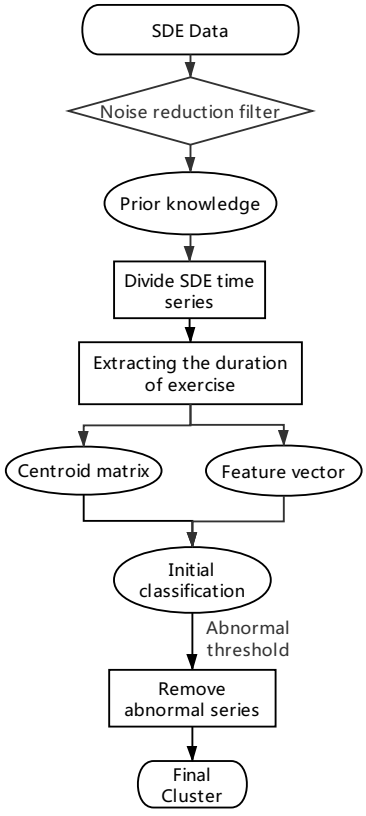

Figure 2. The data flow in our approach.

A. Processing of raw data

Step count data collected with smart phone and other wearable devices reflect users' daily activities. Aside with habitual movement such as commuting to/from work and planned exercise, random and irregular activities will also

be reflected in the data. When we try to characterize the user's exercise habit, it is important that data generated by random and irregular activities are effectively removed.

To realize this, we designed a noise reduction filter, as shown in the following algorithm table. In the filter, only activities that has duration (in time) and intensity (in step counts) greater than some threshold value are deemed as valid exercise events. At the same time, we propose to use the duration of exercise instead of the original steps to analyze the exercise modes. In the same situation, the difference in the number of exercise steps is far greater than the exercise time, so using duration (in time) to analyze the exercise modes can significantly reduce the impact of noise. The filter produces a binary time series, with 1 for a valid 'exercise state', and 0 for a 'non-active state'. It is clear that after filtering, the moment and duration of one exercise event will be retained, while the short-bursts of activities and irregular movements are removed. It is also important to note that the intensity of exercise in terms of step count is temporally left aside. Fig. 3 shows an instance of noise reduction filter. The upper bar chart indicates the original SDE time series and the middle bar chart is the SDE time series after filtering out the noise steps. The bottom curve represents the exercise state and we name it time series of exercise. Comparing raw SDE time series with the time series of exercise, we can find that the brief or light movement is treated as non-active state. The time series of exercise successfully remove the random and irregular part of raw data.

Algorithm 1 Noise reduction filter

Input: input parameters step series 'Step', window width 'W', walk threshold 'T', length of Step 'len'
initialization: walk state series 'Walk' = [0] * len(Step)
Output: Walk

for $i = 0; i < len - W; ++i$ **do**
  if $\sum_{k=i}^{i+W} Step[k] > T$ **then**
    $Walk[k] : Walk[k+W] = 1$
return $Walk$

We tested the filter on a SDE dataset from a university in China. The filter removed 12% of the points and reduced the step noise by 28% after the operation for all data of all users.

B. New distance of SDE time series

According to the analysis in section slowromancapiii@ , we can find that there are still errors if we use the Euclidean Distance or DTW Distance to measure the similarity of series processed in A. Both distance algorithms have their own drawbacks. Euclidean Distance is too sensitive to the moment-shift of exercise, and DTW Distance does not consider the duration of exercise. We hope to extract the

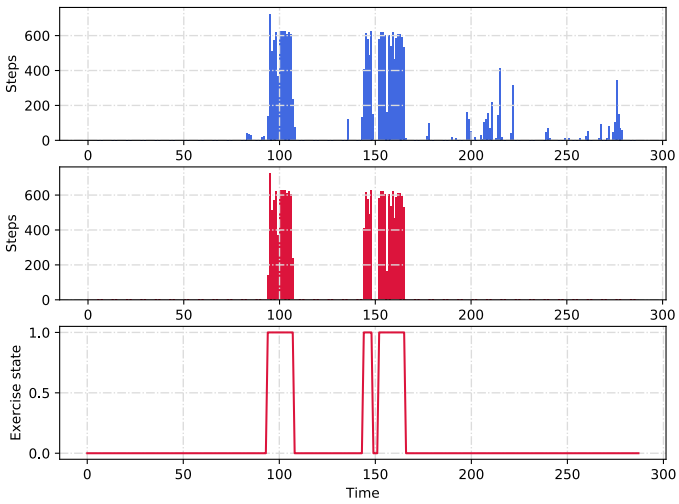

Figure 3. Effect of the filter.

essential features of SDE time series. Therefore, we design a new method of distance calculation.

The new distance algorithm should meet the following requirements: 1. There must be a certain degree of tolerance for the moment-shift of the exercise. 2. It is necessary to consider the duration of the exercise. 3. The number of exercise should also be included. Based on these considerations, we propose to segment the time series of exercise, which can have a more appropriate tolerance for the moment-shift of the exercise, and also consider the number of exercise. Further, we utilize the total duration of exercise over each time segment which takes the duration of the exercise into account. What is very innovative is that in order to achieve a more appropriate segmentation, we calculate the user's historical data to obtain prior knowledge to divide the SDE time series.

Based on the appropriate segmentation and duration extraction, we obtain feature vectors that reflect the characteristics of the user's exercise. By calculating the Euclidean Distance between the feature vectors, we can obtain the similarity between the original SDE series. Detailed operations are as follows:

- **Acquisition of prior knowledge**
  With abundant historical data, we can extract the prior knowledge from them and the prior knowledge can be used to enhance the accuracy and interpretability of classification. By superimposing the time series obtained from A for a given user, we can get a long-term distribution of the user's exercise throughout a day, as shown in Fig. 4. The x-axis of Fig. 4 is time of the day whose unit is five minutes, with the y-axis being the probability. This figure depicts the probability that a user has exercise throughout a day. We can see the moment when the user has a regular exercise.

Fig. 4 reflects that the probability of exercise at different moments is quite different. But in the macroscopic view, the probability distribution shows obvious exercise preference, which is intuitively consistent with the regularity of human life. The exercise probability distribution from different users can be very different, because each user has unique exercise preference and habits, making SDE data highly personalized.

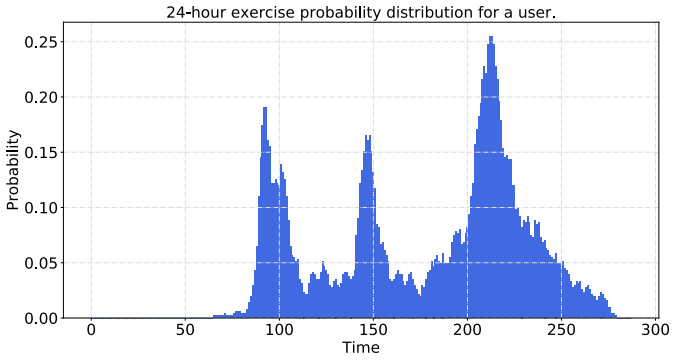

Figure 4. Example: 24-hour exercise probability distribution for a user.

This probability distribution provides a rough sketch of user's exercise preference. With the sketch, we can see the regularity and randomness in user's exercise habits, during a particular window of interest. And we can find that a slight change in the user's exercise habits or sudden occurrence of certain special circumstances, such as equipment problems, can lead to burrs and low-level peaks in the probability distribution of exercise, which are not long-term habits of users. We hope to obtain a stable exercise preference of the user and filter out the burr parts caused by a few special cases. Therefore, this paper uses the classical Hodrick and Prescott filter in economics to smooth the exercise probability distribution. This filter separates long-term trends and short-term fluctuations, helping us to extract the probability distribution of long-term, stable exercise preference. The filter decomposes a given time series object $Y = (y_1, \ldots, y_m)$ into a summation $Y_t = T_t + C_t$ such that the objective function is minimized over $(T_1, T_2, \ldots, T_m)$, where T_t represents the trend component (the desired output), and C_t represents the cyclical component. Increasing the smoothing parameter (λ) results in smoother trend components at a cost of more information loss.

$$Y_t = \sum_{t=1}^{m} C_t^2 + \lambda \sum_{t=2}^{m-1} \left((T_{t+1} - T_t) - (T_t - T_{t-1})\right)^2 \tag{2}$$

The H-P filter is used to smooth the probability distribution to obtain a stable trend part of the probability distribution, as shown in Fig. 5.

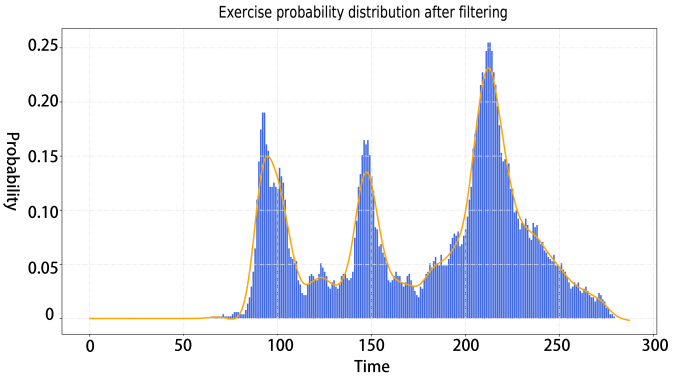

Figure 5. Exercise probability distribution after filtering.

In the figure, the x-axis is time of a day, and the y-axis is the probability of exercise occurring. The yellow curve is the probability distribution after H-P filtering. The filter helps us filter out unnecessary noise information and get a longer-term, more stable probability distribution.

Since the distribution is obtained from data over a relatively long period of time, it can be very stable, so it provides us with important prior knowledge to handle user's step data further.

- **Segmentation of SDE time series**

From Fig. 5 we can see that the user's athletic activity is very different at different time of the day. For example, in the middle of the night, there is no exercise; but there is a greater possibility of exercise at noon or off work. Therefore, we hope to divide the time of day into several time periods for segmentation research. At the begining,we divided the time of day into three parts, the moring, the noon and the night. However, we found that some users' continuous exercise is cut off when all user' exercise time series are divided by a uniform way. For example, the day is divided into three segment of eight hours (0 to 8 o'clock, 8 o'clock to 16 o'clock, 16 to 24 o'clock), but the user is used to exercising from 7 o'clock to 9 o'clock, and the series is cut off from 8 o'clock , resulted in the separation of consecutive exercise.

Based on the observation of a large number of samples, we find that the unified time segmentation method is less effective for SDE time series. Therefore, we propose a personalized segmentation method for SDE time series. With the probability distribution in Fig. 5, we can easily see that the user exercise regularly in the mornings, at noons and during the evenings. It is thus natural to conclude that if we separate the time of a day into 3 sections, the user's exercise activities, which might happen 0 to 3 times for a particular day, will fall into the respective sections. Provided that different

users may have totally different distributions, it is important that we perform personalized time separation and divide the time at which the peak of probability occurs and part of its neighborhood into a segment, such that user's exercise activity is less likely be truncated. Based on prior knowledge, we will know the most appropriate location of the division and store the information in moment vector p. We divide all the series according to the value in the moment vector p, which is personalized. For example, there are three main peaks in the Fig. 5 and the moment vector p generated based on prior knowledge is: $[123, 170]$ ($[10:15, 14:10]$) , so we should divide all series of this user at 10:15 and 14:10 in time axis.

Comparing the method of personalized segmentation with the method of taking eight hours as a segment, the probability that the former cuts off countinuous exercise decreased by 84.3% , which is a very significant optimization.

- **Calculation of new distance**
  The operation of personally dividing time series takes into account the two characteristics of the number of exercise and the time when they occured. Then we also included the duration of the exericises into the distance calculation. We extract the duration of exercise in each segement to form a feature vector, which captures the main features of exercise directly or indirectly, and is the core of our distance calculation.

  The Euclidean Distance is adopted to calculate the distance between the feature vectors to characterize the similarity between original SDE time series. This method is concise and efficient. It captures the essential features of SDE and reduces the noise interference.

C. Classification of exercise modes

The users' exercise modes do not have an accurate number, and there is no absolute 'right' mode and 'wrong' mode. Different users may have distinctly different exercise preference. The method of specifying the number of classifications in advance, based on statistical information, may result in poor interpretability of the classification, which is not conducive to the semantic understanding of the analysis results. Therefore, this paper proposes a framework for constructing individual exercise modes recognition methods based on prior knowledge, instead of specifying the optimal number of modes in advance.

We obtained the prior knowledge from the user's probability distribution of exercise analyzed in Subsection B. To refine the more concise modes, we quantify the duration of exercise. The distribution of the total duration of exercise in every time segment is calculated, and then the elbow rule is used to find the optimal number as the level of quantization. In this paper, the exercise time in each segment is quantized into three levels, and every quantified level values of each segement is stored in the level matrix L. For example, based on the vector p, the matrix L of the user in Fig. 5 should be:

$$\begin{bmatrix} 6 & 10 & 10 \\ 12 & 20 & 20 \\ 18 & 30 & 30 \end{bmatrix}$$

Each column of a matrix represent each segment, and the lists of matrices represent the values of each quantization level. The actual duration is equal to the value multiplied by five, because each point are five minutes of data. Traverse Cartesian product of all columns in matrix L to construct initial exercise behavior classifiers. Calculate the distance between feature vectors and the centroids of each classifier, where Euclidean Distance is used. The feature vector will be classified into a exercise mode classifier corresponding to the smallest distance.

In each classifier, sort the distances between each feature vector and the centroid. Remove the feature vector with the largest distance from the classifier in turn, and calculate the distance sum of all vectors left to the centroid after each vector is removed. Based on this operation, we may obtain a curve between the total distance within the classifier and the vector removed. According to the elbow rule, the curve's inflection point is found. The distance at the inflection point of the curve is set as a threshold value, and vectors whose distance are greater than the threshold value are moved to the abnormal classifier to obtain the final classification result. Different from the unified setting of the abnormal threshold in the general classification algorithm, our algorithm personally calculates the corresponding abnormal threshold for each classifier, which can improve the accuracy of identifying the abnormal mode.

The classification method in this paper fully considers the characteristics of SDE time series. It is concise and efficient, can accurately identify different modes of exercise behavior, and the analysis results are highly interpretable. In addition, the method has better scalability, and the clustering precision can be conveniently adjusted, with good usability.

V. Experiments

A. Datasets

- **Synthesized Dataset.** The dataset is a manually generated virtual dataset. We selected seven typical SDE time series of one user as the core series of the seven basic exercise modes. A certain degree of random fluctuations are generated from the three aspects of duration of exercise, the steps at each time point and the moment at which the exercise occurs, generating some virtual time series in the same exercise mode and tagging them. For step series data $Step = (S_1, S_2, \ldots, S_m)$, we define methods for generating the same type of data $Step' = (S'_1, S'_2, \ldots, S'_m)$ as follows:

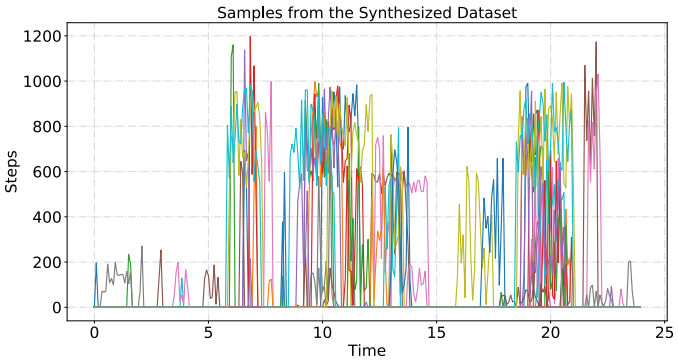

Figure 6. Samples from the Synthesized Dataset.

Moment-shift:

$$S_i' = \mathcal{M}(S_i) = S_{i+t(\bmod m)}, i = 1, 2, \ldots m$$

Duration-shift:

$$S_i' = \mathcal{D}(S_i) = \beta S_i, i = 1, 2, \ldots m, \text{ where } 0.9 < \beta < 1.1$$

Steps-shift:

$$S_i' = \mathcal{N}(S_i) = \max\{S_i + s, 0\}$$

,where

$$s \sim \mathcal{N}\left(0, \sigma^2\right), \sigma = 0.1 * \max\{S_i\}$$

Fig. 6 presents a superposition of partial time series.The x-axis of Fig. 6 is twenty-four hours a day, and the y-axis is the number of steps.

- **JDJK-walk Dataset.** This dataset comes from a university in China and there are a total of 418 users' time series of steps in the dataset. The earliest occurrence date of the SDE time series in the dataset is August 9, 2017. The longest consecutive days of data is 510 days. The dataset has a total of 26,632 pieces of data. Each piece of data contains the anonymized id, the date, the steps recorded every five minutes and the total steps of a day. Steps are recorded every five minutes, so the dataset remains a high level of accuracy. Based on it, we can analyze each user's exercise preference and exercise modes.

B. Evaluation

Overall comparison: We compare our method with some well-known baseline algorithms (namely, K-means, k-modoids, and agglomerative hierarchical clustering) and the method in Decomposing Activities of Daily Living to Discover Routine Clusters[14](DADL). We employed both Euclidean Distance and DTW Distance for these methods.

- **Synthesized Dataset Result.** On the tagged dataset, we apply the Euclidean Distance and the DTW Distance respectively to a variety of baseline algorithms to calculate the accuracy of the classification. From TABLE I, we can see that the baseline algorithm mostly performs poorly because it does not consider the characteristics of the SDE time series themselves. The result at Euclidean Distance is particularly poor because the Euclidean Distance calculates the point-to-point distance between two time series, and a large distance is generated when there are hundreds of steps of fluctuation, while the exercise mode has not changed in fact. At the same time, the adoption of the Euclidean Distance makes the algorithm very sensitive to the moment-shift of the exercise on time axis. The exericse of the same duration, offset by 30 minutes on the time axis, will result in a very large distance. When the DTW Distance is used for calculation, the time-position information of the exercise and the duration of the exercise are ignored to some extent. The difference in the steps of each time point is amplified, and the classification accuracy is low.
From the results, we can find that our method grasps the essence of SDE time series, and thus, we can better capture the user's exercise modes, and obtain satisfactory results.

Table I
THE ACCURACY SCORES FOR THE SYNTHESIZED DATASET

Methods	**Distance**	**Accuracy**
K-means	Euclidean Distance	48.0%
K-medoids	Euclidean Distance	32.0%
	DTW Distance	54.3%
Hierarchical	Euclidean Distance	44.5%
	DTW Distance	51.2%
DADL		62.3%
Our method		89.5%

- **JDJK-walk Result.** The SDE time series of a user is classified by our method as an example. The centroids of initial classifier are shown in Fig. 7. In Fig. 7, the x-axis is 24 hours in a day, and the y-axis is the duration of exercise. It can be seen that the time series is divided into three segments according to the user's exercise preference.

TABLE slowromancapii@ is the initial classification of the user's SDE time series. From this table, we can analyze the typical exercise modes of the user. The user has more than one-third of the dates without continuous exercise for more than 15 minutes (M9), which means all movements are low-duration and sudden movements in these days; the second and third mode are M0 and M1, indicating that this user prefers low-to-medium-intensity exercise, and he/she likes exercising more in the evening. From the proportion of each mode and the characteristics of each mode, we can conclude that this is a user with a regular life who prefers a moderate duration of exercise, and has a tendency to exercise in the evening. Fig. 8 shows the proportion of different

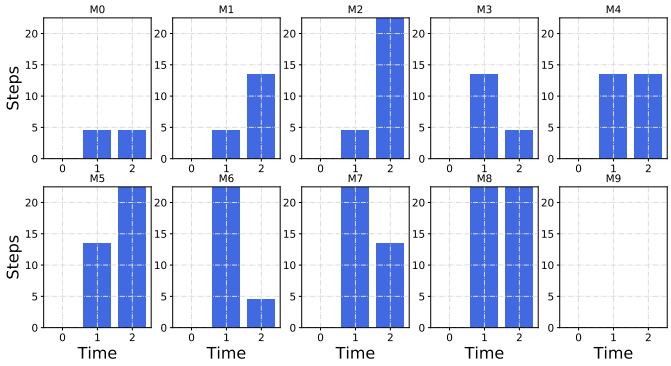

Figure 7. The centroids of initial classifiers for a user in the JDJK-Walk dataset.

Table II
NUMBER OF SERIES IN EACH CLASSIFIER

Mode	**M0**	**M1**	**M2**	**M3**	**M4**
Number	147	113	12	8	1
Proportion	33%	25%	2.7%	1.8%	0.2%
Mode	**M5**	**M6**	**M7**	**M8**	**M9**
Number	3	5	0	0	159
Proportion	0.7%	1.1%	0%	0%	35.5%

modes obtained by our algorithm in different time lengths. The line for each color represents a exercise mode of the user, with the x-axis being length of the dataset(days) and y-axis being the proportion of the mode. It can be seen that the state probability after 200 days tends to be stable, and the result converges to a certain value, that is, the method can find a more stable exercise mode of the user. At the same time, from this figure we can find changes of the user's exercise habits: as time goes on, the user gradually reduces the night's exercise.

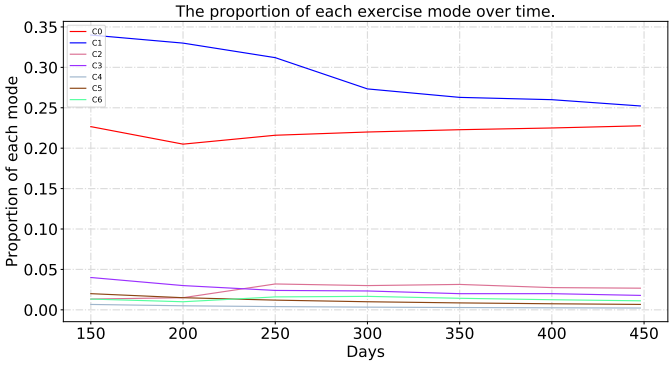

Figure 8. The proportion of each exercise mode over time.

We remove the time series in the abnormal classifier, and superimpose the time series before and after the removal operation. The time series can reflect the duration and position of the user's exercise, so their superimposed graphics express the information of the dataset. As shown in Fig. 9, the blue portion is the superimposed pattern before the removal operation, and the red dashed line is the superimposed pattern after the removal operation. We can see that the shape of the graphics after the superposition is basically unchanged ,which means the information loss of the dataset is very small.

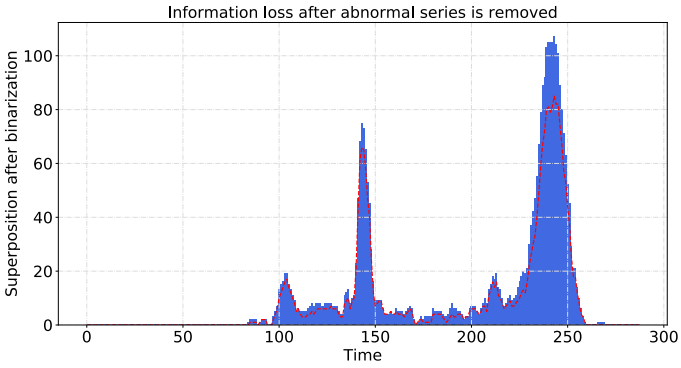

Figure 9. Information loss after abnormal series is removed.

Finally, we classify the data of all the users who have more than 30 days data in the dataset. Rank the proportion of all exercise modes and calculate the number of modes that make up the first 80% . This dataset comes from the teachers who have stable working arrangements during the week. Regular work and rest limit the number of exercise modes. From the Fig. 10, we can see that most users have 0-4 permanent exercise modes, and this result is in the line with the habits of people.

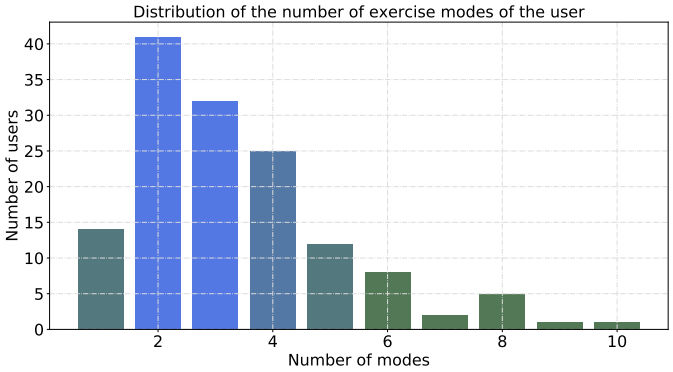

Figure 10. Distribution of the number of exercise modes of the users.

VI. CONCLUSIONS

We propose a simple and personalized classification method for SDE. It takes into account the time and duration, both of which are the essential features of exercise. Based on the two features, we propose new distance functions during the classification, greatly reducing the computational

complexity and minimizing the effect of possible data inaccuracy in SDE time series. Experiment results show that the proposed method significantly outperforms existing SDE classification methods. Furthermore, our method produces results with great interpretability. It may provide health professionals with a robust tool to understand and interact with their clients.

The results obtained from this research is highly dependent on the fine granular dataset itself. Our next step is to use the insight obtained through this research to analyze data with coarser granularity, for instance, data series collected with wider, or even non-uniform intervals.

ACKNOWLEDGMENT

We acknowledge the support of the National Natural Science Foundation of China under grant 61433009 and we thank the Yiqizou Company (http://yiqizou.com) for providing the dataset.

REFERENCES

[1] Tien-Chin Tan. 'The Transformation of Chinaś National Fitness Policy: From a Major Sports Country to a World Sports Power.' The International Journal of the History of Sport 32.8(2015):1071-1084.

[2] S. Becker, T. Miron-Shatz, N. Schumacher, et al., 'mHealth 2.0: Experiences, possibilities, and perspectives,' JMIR mHealth uHealth, vol. 2, no. 2, pp. 148–159, 2014.

[3] M. J. Deering, E. Siminerio, and S. Weinstein, 'Issue brief: Patient-generated health data and health IT,' Office Nat.Coordinator Health Inf. Technol., Washington, DC, USA, 2013.

[4] Ryu, Borim. 'Impact of an Electronic Health Record-Integrated Personal Health Record on Patient Participation in Health Care: Development and Randomized Controlled Trial of MyHealthKeeper,' Journal of Medical Internet Research 19.12(2017):e401.

[5] Aral, Sinan , and C. Nicolaides . 'Exercise contagion in a global social network,' Nature Communications 8(2017):14753.

[6] Smarr, L. 2012. 'Quantifying your body: A how–to guide from a systems biology perspective,' Biotechnology Journal 7(8):980–991.

[7] Tollmar, K., Bentley, F., and Viedma, C. 'Mobile health mashups: Making sense of multiple streams of well- being and contextual data for presentation on a mobile device,' In Pervasive Computing Technologies for Healthcare (PervasiveHealth), 2012 6th International Conference on, 65–72. IEEE.

[8] Hodrick, R. J., and Prescott, E. C. 1997. 'Postwar us business cycles: an empirical investigation,' Journal of Money, credit, and Banking 1–16.

[9] Zheng, J., and Ni, L. M. 2012. 'An unsupervised framework for sensing individual and cluster behavior patterns from human mobile data,' In Proceedings of the 2012 ACM Conference on Ubiquitous Computing, 153–162. ACM.

[10] Cook, D. J.,'Learning setting-generalized activity models for smart spaces,'IEEE intelligent systems 2010(99):1.

[11] Patel, D.; Hsu, W.; and Lee, M. L. 2012. 'Integrating frequent pattern mining from multiple data domains for classification,',In Data Engineering (ICDE), 2012 IEEE 28th International Conference on, 1001–1012. IEEE.

[12] Vahdatpour, A., Amini, N., and Sarrafzadeh, M.2009,'Toward unsupervised activity discovery using multi- dimensional motif detection in time series,' In IJCAI, vol- ume 9, 1261–1266.

[13] Ali, R., ElHelw, M., Atallah, L., Lo, B., and Yang, G.-Z.2008.'Pattern mining for routine behaviour discovery in pervasive healthcare environments,' In Information Technology and Applications in Biomedicine, 2008. ITAB 2008. International Conference on, 241–244. IEEE.

[14] Yürüten, Onur, J. Zhang , and P. Pu . 'Decomposing Activities of Daily Living to Discover Routine Clusters. ' Aaai Conference on Artificial Intelligence 2014.

[15] Berndt, D. J., and Clifford, J. 1994. 'Using dynamic time warping to find patterns in time series,' In KDD workshop, volume 10, 359–370. Seattle, WA.

[16] Codella J , Partovian C , Chang H Y , et al. 'Data quality challenges for person-generated health and wellness data,'. IBM Journal of Research and Development, 2018, 62(1):3:1-3:8.

[17] Benkabou, Seif Eddine , K. Benabdeslem , and B. Canitia . 'Unsupervised outlier detection for time series by entropy and dynamic time warping.' Knowledge and Information Systems 2017.

[18] Kerdprasop, Kittisak , N. Kerdprasop , and P. Sattayatham . 'Weighted K-Means for Density-Biased Clustering,' 2005.