A Demonstration of 4D Digital Avatar Infrastructure for Access of Complete Patient Information

Project acronym: MyHealthAvatar

Deliverable No. 8.3
Data Analysis

Grant agreement no: 600929

Dissemination Level		
PU	Public	**X**
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

COVER AND CONTROL PAGE OF DOCUMENT	
Project Acronym:	MyHealthAvatar
Project Full Name:	A Demonstration of 4D Digital Avatar Infrastructure for Access of Complete Patient Information
Deliverable No.:	D8.3
Document name:	Data Analysis Toolbox
Nature (R, P, D, O)[1]	R
Dissemination Level (PU, PP, RE, CO)[2]	RE
Version:	1
Actual Submission Date:	DD/MM/YYYY
Editor: Institution: E-Mail:	Dr. Xujiong Ye University of Lincoln (LIN) xye@lincoln.ac.uk

ABSTRACT:

This deliverable focuses on the work of building a visual data analysis suite to support data analysis. This includes noisy data processing, data aggregation and interpretation from different sources (e.g. wearable sensors, mobile phones, etc). A number of advanced data mining technologies have been developed to extract useful information that is valuable to each individual.

We have implemented the prototype for estimating/predicting people's overall daily active scoring from low-level data/attributes (i.e. walking steps, travel duration, distance, etc). This provides user a good summary/ indication of general active status. The framework can be used to answer medical questionnaires where categorical output is required. As an example, the current data analysis suite also provides functionalities for visual assessment by assessing individual's night driving capability using mobile 's move data.

[1] **R**=Report, **P**=Prototype, **D**=Demonstrator, **O**=Other

[2] **PU**=Public, **PP**=Restricted to other programme participants (including the Commission Services), **RE**=Restricted to a group specified by the consortium (including the Commission Services), **CO**=Confidential, only for members of the consortium (including the Commission Services)

KEYWORD LIST:
Daily active state, genetic programming, health data analysis, hidden Markov model, supervised learning, visual assessment, wireless sensor data, mobile app data

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 600929.

The author is solely responsible for its content, it does not represent the opinion of the European Community and the Community is not responsible for any use that might be made of data appearing therein.

MODIFICATION CONTROL			
Version	**Date**	**Status**	**Author**
1.0	18/12/2015	Draft	Dr Ji Ni
2.0	20/01/2016	Final	Dr Xujiong Ye

List of contributors

- University of Lincoln

Contents

1 Executive Summary

This document presents in details MHA data analysis suite for predicting individual's daily active states from sequential life-logging data captured using wearable sensors, mobile phones, etc.

Multiobjective Genetic Programming hidden Markov model (MOGPHMM) is used as a supervised machine learning technique to assess and provide scoring for daily active states. According to both medical reference and real data characteristics, we generate a large amount of synthetic training data to extrapolate possible imbalances or missing data embedded in the read data, from which MOGP has trained classifiers to group high-dimensional sequential data (i.e. walking steps, travel duration, distance, etc) into a number of one-dimensional discrete states in order to establish an effective emission matrix for HMM stage. We demonstrate that our method can accurately and robustly predict individual's active states from sequential life-logging data.

The framework can also be used to address medical questionnaires where categorical output is required. As an example, the current data analysis suite also provides functionalities for visual assessment by automatically assessing individual's night driving capability using mobile 's move data.

This document is organized as follows: Section 3 introduces the methodologies how the HMM corporates with the classifier hybrid scheme, including the details of methods and data representation. We also introduce how this general framework is applied to night-driving prediction for VF-14 questionnaires [4]. Finally, we conclude that MOGPHMM effectively predict daily active states for health care purpose with better generality in supervised learning from life-logging data.

2 Introduction

2.1 Sequential life-logging data analysis

Life-logging data has drawn great attention to monitoring people's daily activity for health, fitness and a wide range of other purposes [1]. With the development of mobile devices and wireless communication, varieties of wearable devices has been developed and life-logging data increases exponentially with respect to data dimensions. To process and analyse these data for general health purpose remains an active research area. For instance, daily active level is one of the most important assessment for health and fitness for daily life. NHS has a public guide of minimum activity for general health, while there are also a range of research that indicate some other measurements, e.g., steps, distance, for maintaining a healthy daily active states.

The conventional standard normally considers some statistical thresholds as minimum requirement and was used in the commercial wearable products for health and fitness monitoring, e.g., Fitbit, Withings. However, further processing on these data beyond simple statistical analysis has rarely been seen in those products. For instance, Fitbit only summarises, for example, step or distance, for each person weekly or monthly. So far, there is not a trivial way to use all the data provided to output a general active state for each day.

In this document, we are analysing life-logging data in a sequential manner captured using wearable sensors, mobile phones, and employing hidden Markov model (HMM) as a proved machine learning technique to predict individual's daily active states and to provide visual assessment. Considering there are K active states from inactive to extremely active for each day, where K is predetermined, we can train personalised HMM model for each individual and predict his/her active states (e.g. K states) from all the collected data. Quantification of people's daily activity in terms of K-state is more intuitive and the method can be extended to other applications in healthcare (e.g., medical surveys, medical questionnaires)

HMM can be used for supervised, semi-supervised and unsupervised learning tasks with corresponding algorithms [2]. In the context of machine learning however, supervised learning is generally an easier task than and is better theoretically justified to the other two cases [3]. In this document, we have employed supervised learning scheme to achieve best possible results. Commonly methods in supervised HMM employ expectation maximisation (EM) for training the emission matrix while determining the prior π and transition matrix directly by accounting the sequential tags (supervision) in the training data. To train the emission matrix, a Gaussian mixture model (GMM) is often used for continuous feature input or Bernoulli for categorical, followed by EM to optimise

parameters in the mixture model employed. However, the results are highly dependent on the model assumed, as well as on the initialisation due to the local optima EM can yield. To address this problem, in this document, we use a multiobjective Genetic Programming (MOGP) hidden Markov model as a multiclass classifier to transform the original high-dimensional continuous feature space (e.g. step, duration, distance from wireless sensors or mobile phone) into a new one-dimensional discrete class space using multiclass classification to construct a compact HMM for daily active state prediction.

This report is constructed as follows: the next section will introduce the methodologies how the HMM corporates with the classifier hybrid scheme, including the details of methods and data representation. We also introduce how this general framework is applied to night-driving prediction for VF-14 questionnaires [4]. Finally, we conclude that MOGPHMM effectively predict daily active states for health care purpose with better generality in supervised learning from life-logging data.

3 Methodology

3.1 General Framework

A conventional supervised learning HMM framework as in Figure 2.1. After having been labelled and assigned tags, the real data is partitioned into training and test sets for HMM stage. A Gaussian mixture model is commonly used as a parametric optimisation via expectation maximisation. The trained HMM is then used to perform a Viterbi decoding on test data for prediction assessment.

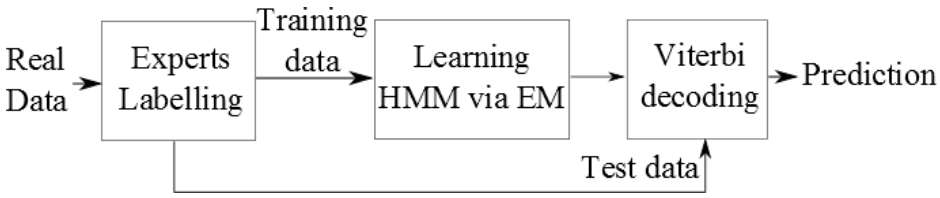


Figure 2.1. General framework of supervised HMM

Different from the conventional framework, in this report, we have proposed our new method optimised for daily activity state prediction task shown in figure 2.2. The main advantage of the new framework has avoided the pre-settings of training HMM via expectation maximisation. Instead, it maps the original high-dimensional continuous feature space into a one-dimensional finite discrete class space and constructs a compact HMM to solve the problem.

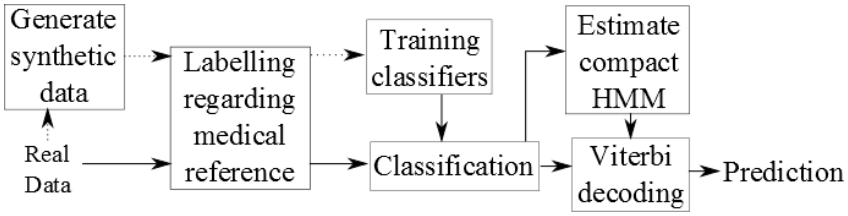


Figure 2.2. Framework of daily active states prediction using supervised classifier-HMM.

In our framework, we firstly extract some statistical characteristics from real data, which will be introduced in Section 4.1.1, to generate a large amount of synthetic data that covers all possible states. The synthetic data can be used to extrapolate possible imbalances or missing data embedded in the read data. We then label the synthetic and real data regarding medical reference, although we further consider variant noises for real data for assessment purposes. The labelled synthetic data will be then used for training multi-class classifiers that transform (classify) real data from continuous real space to a finite and discrete class space. We then use labels of part of the real data as prior knowledge,

together with the classes to empirically estimate a compact HMM, which will be used later for a Viterbi decoding on the real data to provide outputs for the prediction.

3.2 Multiobjective Genetic Programming

3.2.1 GP and MOGP

Genetic programming (GP) is a non-parametric evolutionary optimisation algorithm that uses tree-based syntax to represent models [5], as opposed to parametric optimisation algorithms like genetic algorithm (GA) to optimise a length of coefficients. Unlike other parametric optimisation methods, a key advantage of GP is that it does not need predetermined model structures and the tree-based syntax will commonly provide richer model candidates for the searching. We employ GP to generate discriminative classifiers that map the multidimensional continuous observation vectors to finite and discrete classes as HMM input.

The multiobjective method aims to simultaneously optimise multiple tasks that are usually competitive. In this document, we employ MOGP to simultaneously minimises empirical 0/1 loss and the node count as syntactic complexity measure to evolve a Pareto-front that presents the trade-off between empirical error and model complexity. Minimizing tree size not only control bloat but also constrain a form of upper bound of model complexity [6]. MOGP is essentially a practical scheme to generate models with small expected risk in the statistical learning perspective.

3.2.2 Details of MOGP

Genetic programming is a population based algorithm which is constructed by a group of candidate models. Offspring (new) models are generated by crossover and mutation process to inherit part of the structural characteristics, termed gene, from their parent models. Pareto-comparison is then employed to rank and sort all models and only the elitists are preserved. A rank based algorithm is employed for selecting parent models for breeding [7]. The evolutionary process terminates when certain criterion is met, for instance, target training error < 0.001 or reaching the computing limit. Figure 2.5 illustrate the brief evolutionary process.

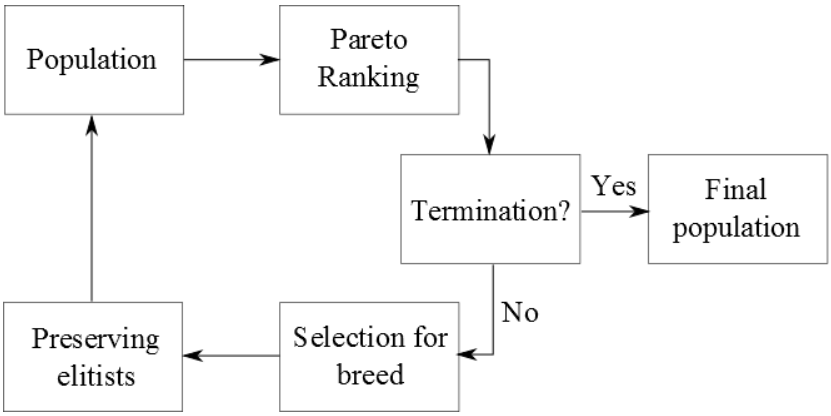


Figure 2.5 Framework of GP algorithm

In out algorithm, we have employed the steady-state evolutionary strategy to pursue a steadily-converged Pareto-front, which is called Pareto converging genetic programming (PCGP) [8]. This methods select two models in the population, perform crossover and mutation, and generate two offspring models. The newly generated models are appended to the population for re-ranking, after which the bottom-ranked two models are discarded. This strategy is reported to have superior performance among a range of other strategies [9].

The GP parameter is summarised in Table 2.2. We have run up to 80 000 tree evaluations, each of which obtains the fitness vector of newly generated model. Or, the evolutionary process terminates when 0/1 loss is zero.

Table 2.2 GP parameters

Population size	100
Initialisation	Ramped [18]; 30 repetitions
Termination criterion	80 000 evaluation or 0/1 loss = 0
Crossover Mutation	Point crossover [18] Point mutation [18];Tree depth = 4
Node Type	Unary minus Addition, Subtraction Multiplication Analytic Quotient [19]

3.2.3 Multiclass classification

To classify the observed input into L classes, we have adopted a form that is similar to one-vs-all scheme [10]. The main issue of multiclass classification is the ambiguous decision for the same pattern [10]. Taking advantage of the following HMM stage, the classification errors (classes vs. ground truth) can be characterized by the emission matrix, and further refined by the HMM process.

A simple but consistent strategy is to train m independent classifiers f_i ($i \in m$) from the synthetic training data that provides equal priors for each class, either using MOGP or SVM; where m = 5 in our case. These classifiers are sorted in ascending order with respect to their training errors to obtain $f_1, f_2 \ldots f_m$, where f_1 has minimum training error and f_m the maximum. These classifiers are used to label the real data, such that every current classifier always incurs the least risk/error for the later ones. An extra class for patterns without a positive class assignment is introduced. Thus, there are m+1 classes against m true states and an m+1 by m emission matrix is constructed instead of m continuous PDFs.

3.3 Hidden Markov Model

A hidden Markov model $\Theta(\pi, A, B)$ is a probabilistic description of a series of observations X, where there are K variant hidden states $z \in Z_K$ and M variant observations $x \in X_M$ for the case observation being discrete and finite. π is a K dimensional vector that represents the prior of each hidden state z_i, defined $\pi_K(i) = p(z_i),\ i \in K$. The K by K matrix A represents the transition probability of hidden states from z_j to z_i and is a conditional probability defined by $A_{K\times K}(i,j) = p(z_i|z_j), i \in K,\ j \in K$. The emission matrix B is a set of hidden-state conditional probabilities of observations. Each element in the M by K matrix is defined by $B_{M\times K}(i,j) = p(x_i|z_j),\ i \in M,\ j \in K$. In a more generalized case where observation x is a n-dimensional continuous vector, as in our case, each column of the emission matrix is generalized to a continuous probability density function (PDF) for state-conditional probabilities, defined by $B_K(x,k) = p(x|z_k), x \in \mathcal{R}^n, k \in K$. Conventional supervised learning for HMM assumes a Gaussian mixture model (GMM) with prefixed number of parameters to perform the density estimation via expectation maximization (EM) and locate a local optima for the emission matrix B, whereby the performance can vary from case to case.

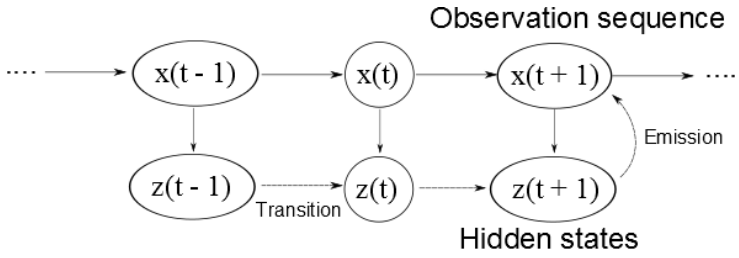


Figure 2.3 hidden Markov model

After having trained the emission matrix B, π and A are directly identified by the tags in the training data. A Viterbi decoding method is then used to predict hidden states, which is to find a *Viterbi path* (a sequence of hidden states) $z^* = max_z P(X, z|\theta)$ that maximises the joint probability of observation and hidden states sequences given the model θ. In other words, the output hidden states sequence z^* is the most likely to be true among all the other possibilities.

To briefly introduce Viterbi decoding, we have the simple implementation that

$$p(z_{0,k}) = \pi_k \cdot B(x_0|z_{0,k})$$

and

$$p(z_{t,k}) = \max_k(p(z_{t-1,k}) \cdot A(z_{t,k}|z_{t-1,k}) \cdot B(x_t|z_{t,k})).$$

The solid line in Figure 2.4 represents the maximal likely routine for state $z_{t,k}$, dot line the non-optimal routine.

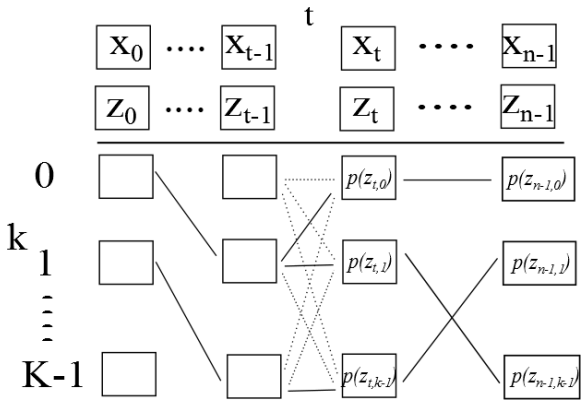


Figure 2.4 Viterbi decoding

In the final column, we could determine which state to be chosen among k for all states $z_{n-1,k}$ by choosing the maximum. Then, the back-tracking procedure of the solid line will reconstruct the optima *Viterbi path*.

Using Viterbi to translate the observational data to the hidden daily active states requires an explicit HMM as we previously mentioned. We have employed an empirical estimation using the true tags and the transferred classes to construct the compact HMM as we have employed classification methods to map 3-dimensional real input into M = K+1 classes, hence the HMM is more compact and easy to be empirically estimated.

4 Life-logging data demos

4.1 Daily active state prediction

4.1.1 Datasets collections and synthetic training data

The real datasets were collected by "moves-app" using the accelerometer and GPS in the cell phone. The real dataset is constructed by a sequence of vectors, each of which consists of individual's daily accomplishment of distance, duration and number of steps from physical activities. For instance, walking or running but excluding distance from transportation. There are ten people who contribute their daily activity data for this sstudy. The number of patterns of each dataset (per person) ranges from 118 (days) to 401 which covers around 4 months to more than 1 year.

As the behavioural characteristics vary from person to person, some people live in an inactive life from which highly active pattern is hardly observed. To perform a supervised learning via MOGP requires all states explicitly existed in the training data for learning purpose. Hence, we generate synthetic data regarding one's behavioural characteristics to construct a training dataset that contains all possible states with equal prior. The synthetic data can also be used to extrapolate possible imbalances or missing data embedded in the read data. Since the elements of each input vector – steps, duration, distance are highly correlates, we extract speed S and step frequency F as following,

$$S = distance/duration$$

,and

$$F = steps/duration.$$

These are two main characteristics for generating the synthetic data. We compute the mean and standard deviation of both quantities for each person over all his/her real data. Then we use a uniform random generator to generate a random duration dr_i. The steps and distance will be generated by

$$step_i = dr_i \cdot \mathcal{N}(\mu_F, \sigma_F^2),$$

$$distance_i = dr_i \cdot \mathcal{N}(\mu_S, \sigma_S^2)$$

where $\mathcal{N}$ is Gaussian distribution with its mean μ, variance σ^2. The synthetic data will be further tagged with its active states, details of which to be introduced in next section. We finalise the training set with 200 data pattern per active state which summed up to 1000 training data for each person.

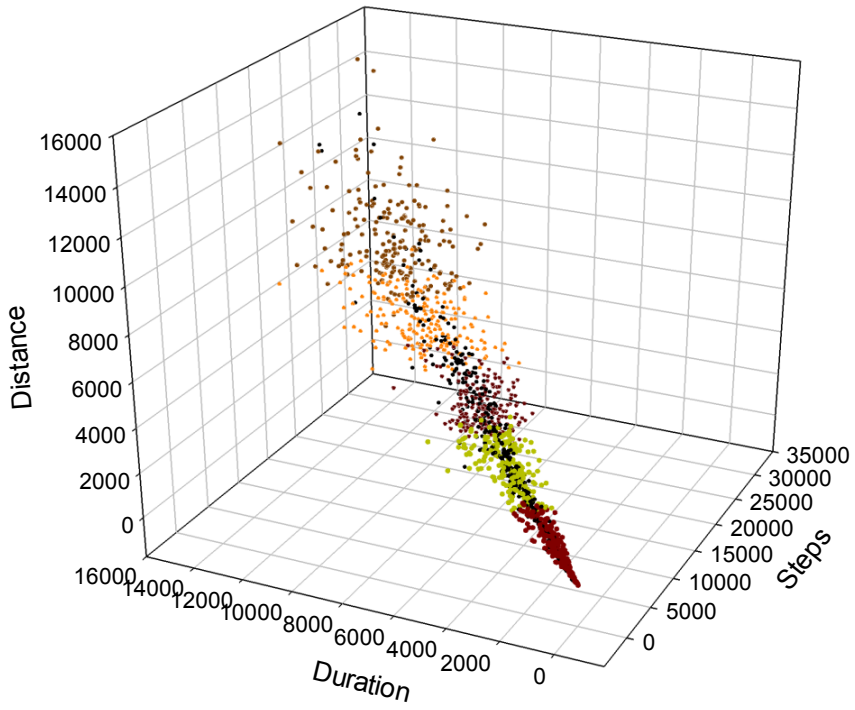


Figure 2.6. A sample of synthetic and real data distribution

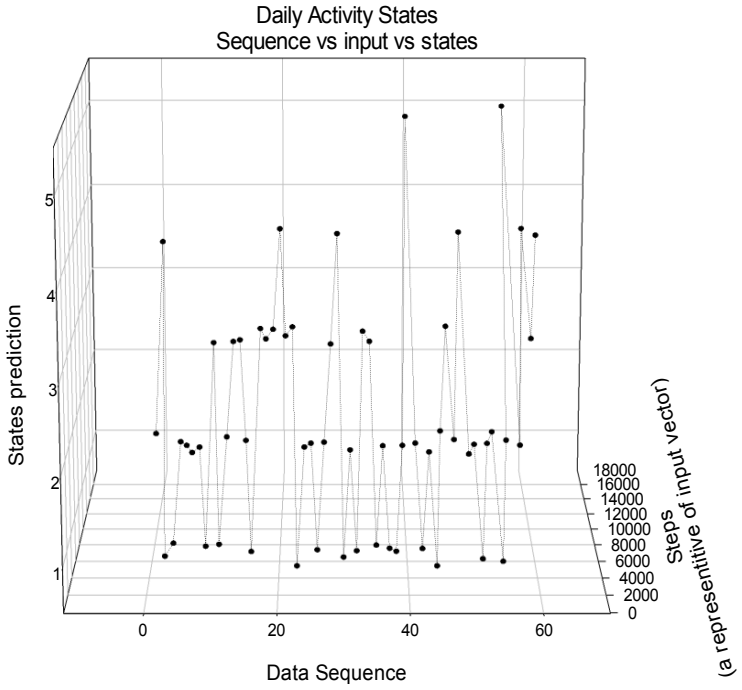
As shown in Figure 2.6, the real data are the black pots surrounded by the coloured plots. The five-coloured plots are synthetic data with equal weight of each active state, which is presented by a different colour. This figure illustrates that our synthetic data provides a practical simulation to the real data and reasonable for training purpose.

4.1.2 Tagging Data

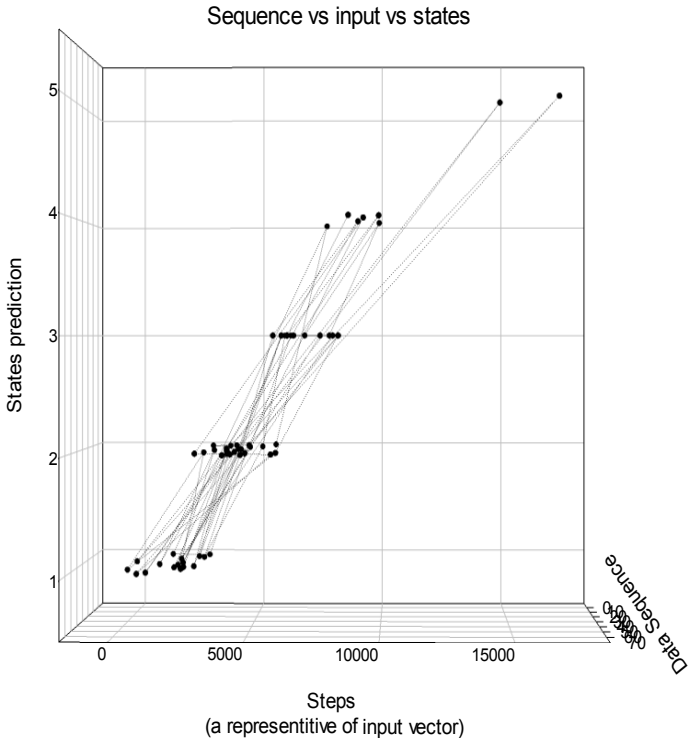
Daily activity level has been widely studied. The typical conclusion or health guidance is using some threshold to justify if one is active or not each day, for instance, 10,000 steps [12], alternatively 8 km walking distance [13], or 30 minutes averagely per day [14]. We thus adopt 10,000 steps, 8km, and 60 minutes [15] as the median to justify the active states level. In our case, we consider five different active states. To tagging a datum with active states, we have employed a linear model that accounts the contribution from each of these three quantities.

4.1.3 Case Study

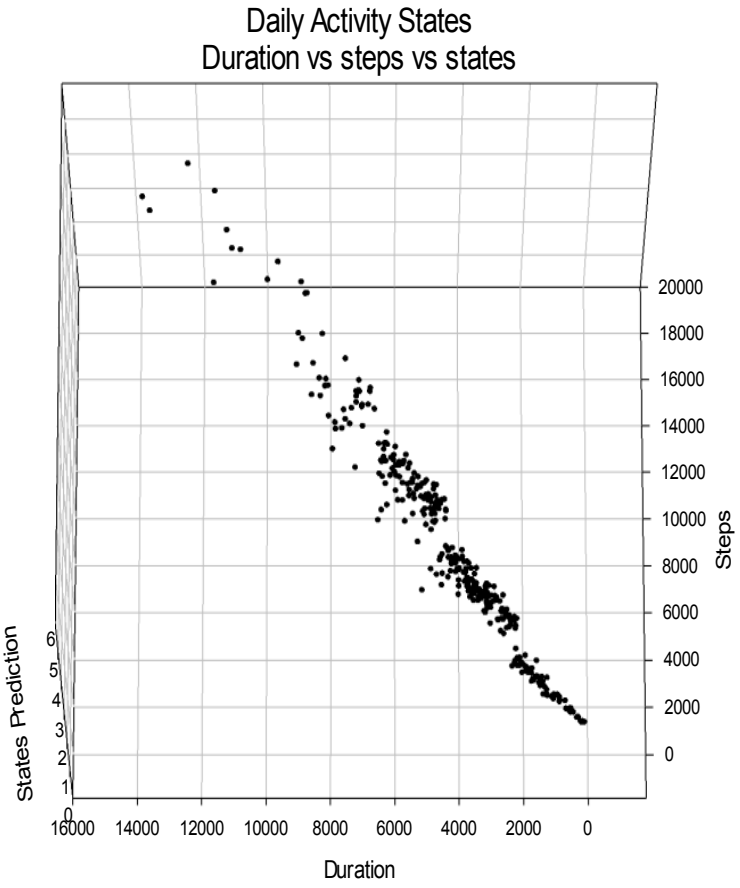
Figure 2.7 and 2.8 show two examples of daily activity states from person 1 and person 3.



(a) daily activity states *vs* data sequence and steps

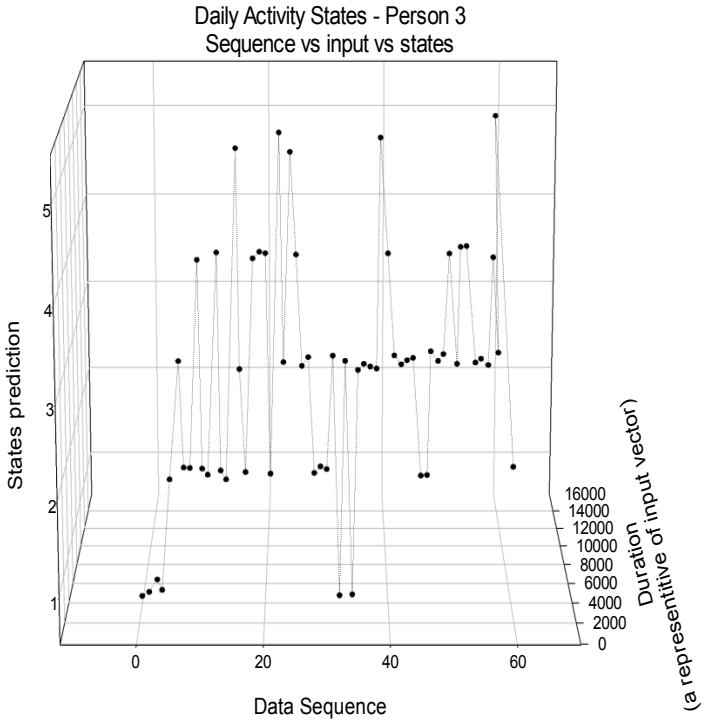


(b) daily activity states *vs* steps and data sequence

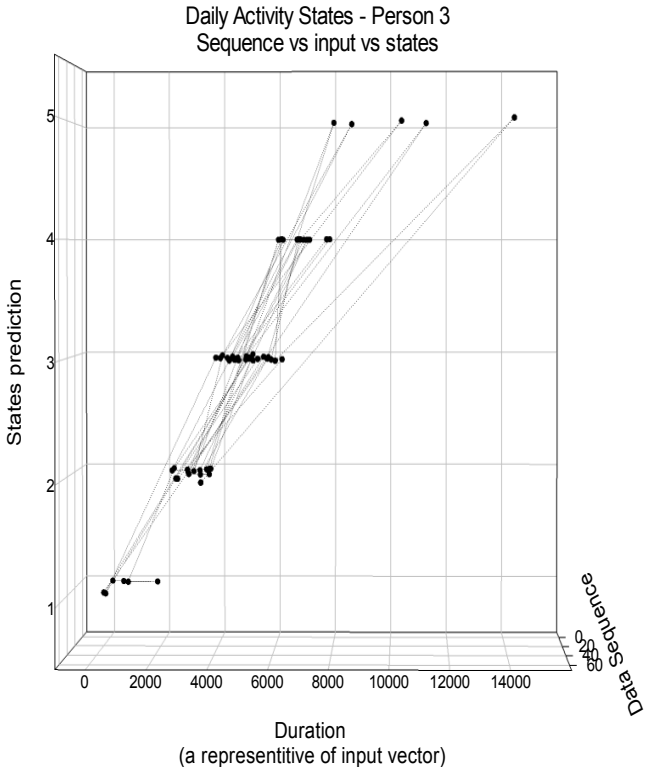


(c) daily activity states *vs* duration and steps (view from top)

Figure 2.6 an example of daily activity states using the first 60-days data from person 1



(a) daily activity states *vs* data sequence and duration



(a) daily activity states *vs* duration and data sequence

Figure 2.7 an example of daily activity states using the first 60-days data from person 3

4.2 Visual assessment- Night driving

The data analysis toolbox can be used to automatically predict discrete states from sequential observations. In addition to the daily active state prediction, the other potential applications include medical questionnaires or medical survey where categorical output is required. As an example, in this section, we present how this framework is used in predicting night-driving evaluation in VF-14 questionnaire [4]. To adopt this framework in night-driving prediction, we only need to slightly redefine data as discussed in the following section.

4.2.1 Tagging Data

From the data available in our platform, we use those labelled with “transport”, of which “duration” and “distance” are used as components of input vector. The output is one of the four states as in VF-14 questionnaire. Since there is no reference for determining an active or inactive driving states, we have considered all data available to have an average daily driving duration of 1546.36 seconds and distance of 33392.68 meters, which are considered as our empirical standard rather than those standard from medical reference in daily active states. We have considered “night” as from “18:00:00” to “23:59:59”, although the “start” and “end” time can be changed to the hour to present any duration in one day with minimum scale of one hour. The duration and distance standard of the six-hour “night” time is assumed one quarter of the daily standard in 24-hour time. The standard of “night” can be further improved by characterising average completion of duration and distance in each hour in 24 hours a day and have more accurate empirical standard in “night” time.

4.2.2 Case Study

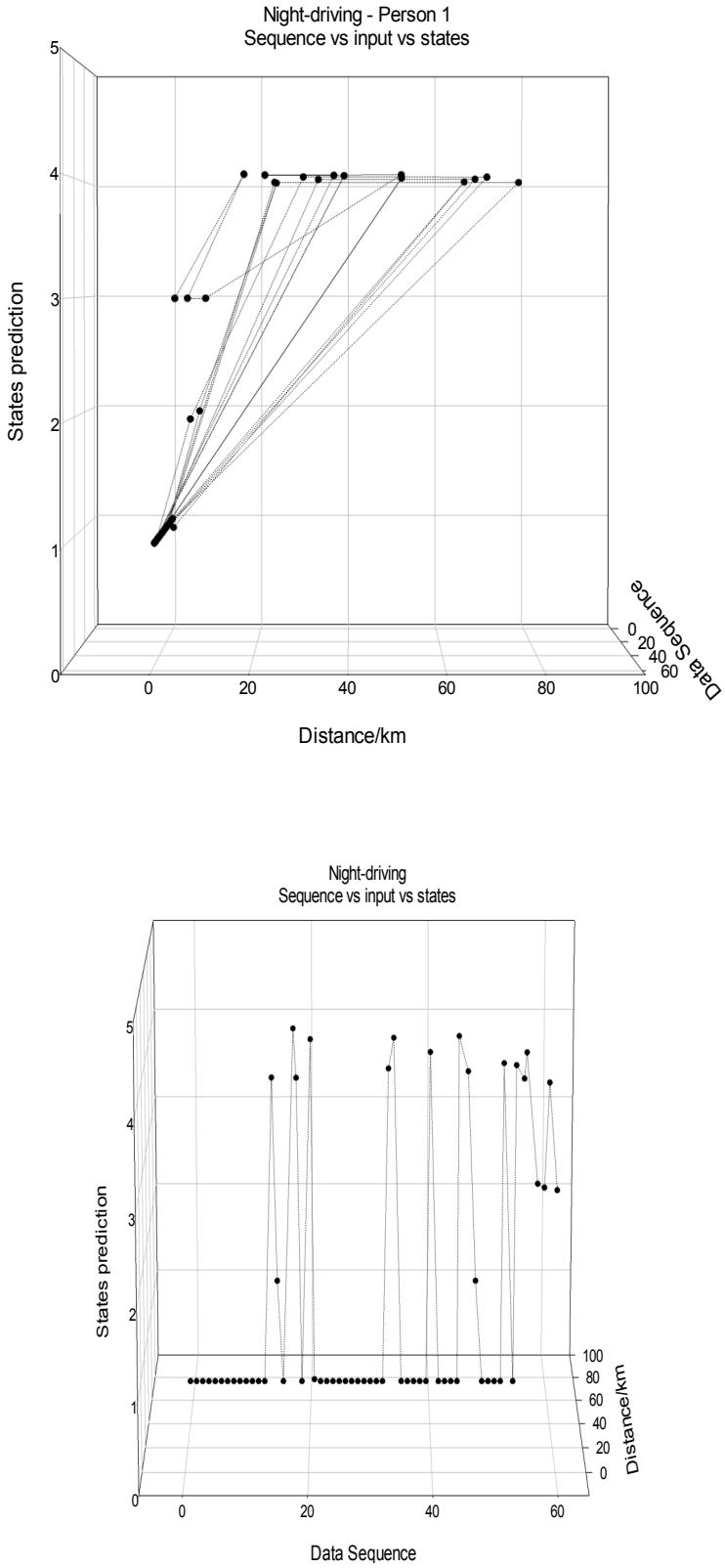


Figure 2.7 an example of night driving states using the data from person 1

5 Conclusion

In this deliverable, we have implemented the prototype for estimating/predicting people's overall daily active scoring from low-level data/attributes (i.e. walking steps, travel duration, distance, etc). We use MOGPHMM as a supervised classification-HMM method to predict daily active states from sequential life-logging data. This provides user a good summary/ indication of general active status. The framework can also be used to answer medical questionnaires where categorical output is required, As an example, the current data analysis suite also provides functionalities for visual assessment by assessing individual's night driving capability using mobile 's move data. We conclude that MOGPHMM can effectively predict daily active states for health care purpose with generality and potential for other use in supervised learning from life-logging data.

6 References

[1] A. Garratt, L. Schmidt, A. Mackintosh, and R. Fitzpatrick, "Quality of life measurement: bibliographic study of patient assessed health outcome measures," British Medical Journal, vol. 324, p. 1417, 2002.

[2] C. M. Bishop, *Pattern Recognition and Machine Learning*: Springer, 2007.

[3] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. New York: John Wiley, 2001.

[4] C. B. Terwee, M. N. Gerding, F. W. Dekker, M. F. Prummel, and W. M. Wiersinga, "Development of a disease specific quality of life questionnaire for patients with Graves' ophthalmopathy: the GO-QOL," British Journal of Ophthalmology, vol. 82, pp. 773-779, 1998.

[5] R. Poli, W. B. Langdon, N. F. McPhee, and J. R. Koza. (2008). *A Field Guide to Genetic Programming*. Available: http://www.gp-field-guide.org.uk/

[6] J. Ni and P. Rockett, "Tikhonov regularization as a complexity measure in multiobjective genetic programming," IEEE Transactions on Evolutionary Computation, vol. 19, pp. 157-166, 2015.

[7] D. E. Goldberg, Genetic Algorithms in Search, Optimization and Machine Learning. Reeding, Massachusetts: Addison-Wesley, 1989.

[8] R. Kumar and P. Rockett, "Improved sampling of the Pareto-front in multiobjective genetic optimizations by steady-state evolution: a Pareto converging genetic algorithm," *Evolutionary Computation,* vol. 10, pp. 283-314, 2002.

[9] Y. Zhang and P. Rockett, "A Comparison of three evolutionary strategies for multiobjective genetic programming," *Artificial Intelligence Review,* vol. 27, pp. 149-163, 2007.

[10] C. M. Bishop, *Pattern Recognition and Machine Learning*: Springer, 2007.

[11] V. N. Vapnik, *The Nature of Statistical Learning Theory*, 2nd ed. New York, 2000.

[12] B. C. K. Choi, A. W. P. Pak, J. C. L. Choi, and E. C. L. Choi, "Daily step goal of 10,000 steps: a literature review," Clinical and Investigative Medicine, vol. 30, pp. 146-151, 2007.

[13] W. W. K. Hoeger, L. Bond, L. Ransdell, J. M. Shimon, and S. Merugu, "One-mile step count at walking and running speeds," ACSM's Health & Fitness Journal, vol. 12, pp. 14-9, 2008.

[14] NHS. *Physical activity guidelines for adults*. Available: http://www.nhs.uk/livewell/fitness/pages/physical-activity-guidelines-for-adults.aspx

[15] G. C. Le-Masurier, C. L. Sidman, and C. B. Corbin, "Accumulating 10,000 steps: does this meet current physical activity guidelines?," Research quarterly for exercise and sport, vol. 74, pp. 389-394, 2003.