Assignment 3 Report

You Hu[2631052,*yhu*310], Wenchen Lai[2643117,*wli*310], and Tian Chen[2632807,*tcn*740]

Vrije University Amsterdam, ML4QS Group 100

1 Introduction

Lifelogging app and sensors on mobile devices have grown more and more popular these days. They enable individuals to digitally monitor and record their daily lives in a very complete and detailed way[7]. So, in other words, analyzing and interpreting the data of this size can be overwhelming but also very rewarding[1]. These gathered data can provide valuable information about users daily life for various purposes, including reproducing and predicting their whole action track during one specific period of time[2].

The data used in this research, which we obtained from NTCIR-Lifelog website, consists lifelogs of three anonymous individuals collected by several mobile sensors and Autographer wearable cameras in 100 days. Every day, 1000 to 1500 pictures of the participants were taken while they were wearing mobile sensors, which recorded their locations and activities.

However, we found that in this data set, not time with picture was labeled by locations and activities. The record of three participants locations and activities can be more continuous. Therefore it is possible to complete their action track by the prediction of their locations and activities with their pictures. Unfortunately, the amount of different locations in the data is too high while samples of every location are too few. Thus, we will only focus on their activities.

2 Data exploration and Feature engineering

2.1 Data description

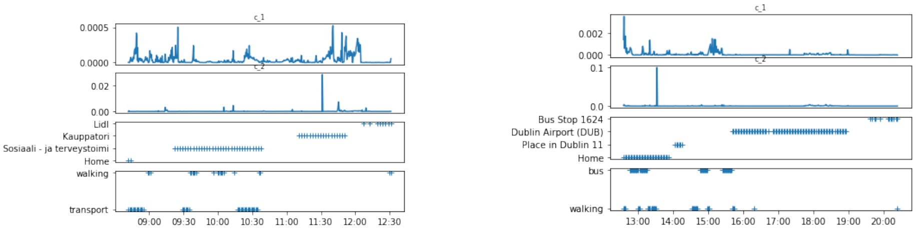
The NTCIR Lifelog dataset consists of about 30 days data from 3 lifeloggers. The dataset contains three parts:

- Images: The lifeloggers take 1000-1500 photos per day using Autographer wearable cameras to record their daily life.
- Visual Concepts(a txt file): This data is automatically extracted by CAFFE visual concepts detector[5]. It calculates the possibilities of 1000 daily objects (such as pizza, school bus, radio, etc...) appearing in each image.
- Semantic Content(an XML file): It contains 130 semantic locations(e.g., University, Dublin Airport, Lidl, Home) and semantic activities(e.g., walking, transport, running). They are captured and recognized by Moves App on Mobile phone. The device collects data every minute.

We choose one minute to be the time step size of our dataset. If there are more than one image in one minute, we average 1000 image features. We separate data of 3 users into 3 Dataframes and Table.1 presents our data format. We also plot two days data of user3 with ‘c_1’ and ‘c_2’ as Fig.1 shows.

date_time	c_1	...	c_1000	location	activity
2015-02-23 07:04:00	0.000018	...	0.008851	Home	walking
2015-02-23 07:04:01	0.000000	...	0.000167	Home	-

Table 1: Aggregated Data



(a) 2015-06-15

(b) 2015-08-03

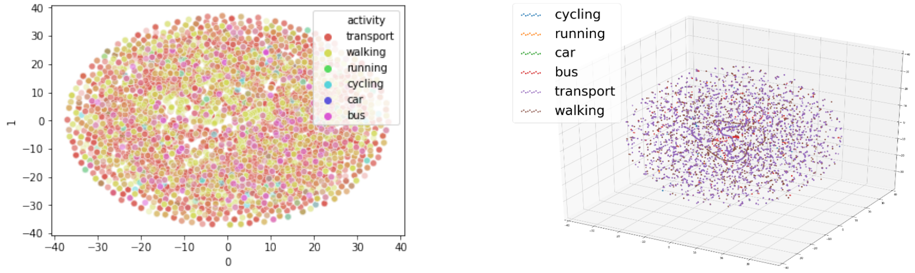
Fig. 1: user3 data visualization

2.2 t-SNE

As we have 1000 CAFFE concepts as features of images which describe the environment around user, it’s naturally to see whether we can separate each location/activity from others in feature space. However, the feature dimension is too high. We hope to reduce the dimension and map those data point from high dimension to 2 or 3 dimension. Therefore we can visualize the result and find whether different categories can be separated.

Here we use T-distributed Stochastic Neighbor Embedding (t-SNE) which is known as a technique for visualizing high-dimensional data by giving each data point a location in a 2D or 3D map. The advantage of t-SNE is that the method is good at handle non-linear relation as it transforms distance between points to the probability of similarity[6].

Below shows the result of dimension reduction(with parameters: perplexity 20,learning rate 50 and number of iteration 2000) As it is shown, there is no significant pattern which can separate groups of activities. Different type of activities are mixed in both 2 dimension and 3 dimension. However, we can not judge whether different activities are not able to be classified. Besides, the main



(a) Data points on 2 dimensional map

(b) Data points on 3 dimensional map

Fig. 2: Dimension reduction visualization

purpose of t-SNE reduction is to visualize the data set rather than processing data for model training because we can directly utilize non-linear model to handle row data. Therefore,we still performed PCA dimension reduction to handle the problem of linear correlation between features.

2.3 PCA

Because there are 1000 features in our dataset, we try to find the principle components and reduce the dimension of our dataset. We apply principle components analysis(PCA) to all our 1000 attributes of images. We need to determine the number of components first and ensure the data after PCA still contains most of the information. We modify the source code of Chapter 2 in our textbook[3]. Fig.3 shows the explained variance by the different principal components. We can observe that the explained variance declines after about 15 components. Thus, we choose 15 as the number of principle components and apply PCA. Finally, we plot two days data of user3 with ‘pca_1’ and ‘pca_2’ as Fig.4 shows.

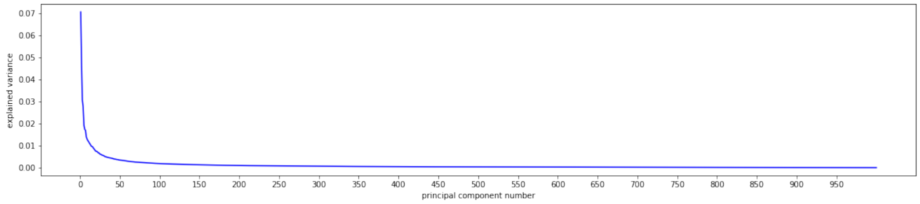
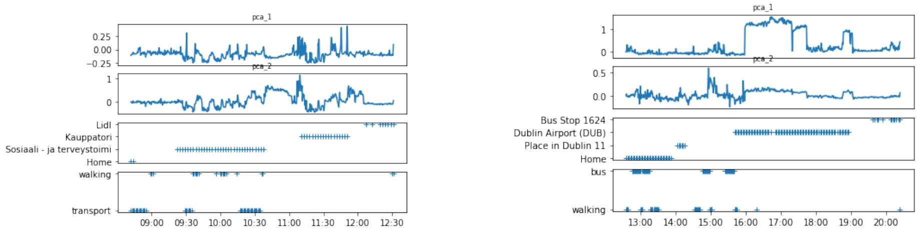


Fig. 3: Explained variance by principal components ranked on importance

2.4 Clustering

In order to have a better insight of the data, we cluster 15 PCA features using k_means method and agglomerative approach [4]. To find the best number of



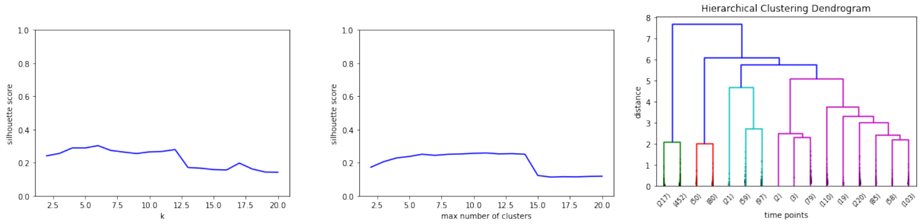
(a) 2015-06-15

(b) 2015-08-03

Fig. 4: user3 data visualization after PCA

clusters(k), first we run the algorithm with different value of k(from 2 to 20) and measure the quality of clustering by silhouette scores. The results are shown in Fig.5. From Fig5a, we can find that k = 6 results in highest silhouette score which is 0.3027. Thus, we run k means with k = 6. Table.2 describe that Cluster 1 covers more than 50% of the instances with 'walking' and 'car' labels. 'Transport' and 'bus' instances mainly occurs in Cluster 1 and Cluster 4. Most of the instances covered by Cluster 3 are the instances with 'bus' label.

Clustering using k_means approach does not provide a really clear and good result, so we also perform hierarchical clustering(agglomerative approach). When we create the clusters, the highest silhouette score is 0.2586 for k=11 (as Fig.5b shows) which is lower than that of k_means approach. Fig.5c presents the hierarchical clustering dendrogram with k=11. We will see whether clustering has effects on activities classification in Section 3.



(a) silhouette scores of k_means approach

(b) silhouette scores of agglomerative clustering

(c) Hierarchical Clustering Dendrogram(k=11)

Fig. 5: Clustering

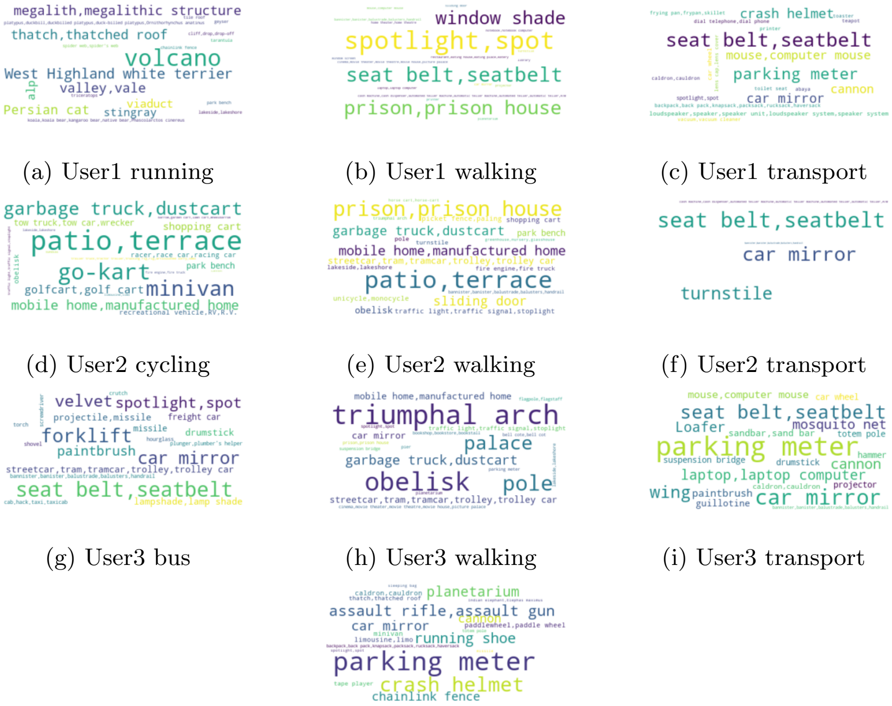
2.5 Word cloud

The PCA dimension reduction has its limit that we will lose the meaning of features which make model harder to be explained and to adjust feature engineering. Therefore, we select the union of 10 highest features for each user and

Activity	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
transport	40.29%	14.31%	4.50%	35.38%	0.00%	5.52%
walking	69.39%	5.19%	7.91%	10.12%	2.72%	4.67%
car	50.00%	16.67%	0.00%	16.67%	0.00%	16.67%
bus	29.56%	1.03%	19.79%	43.96%	0.26%	5.40%

Table 2: Distribution of activities over clusters for k_means(k=6)

each activity as input features and visualize each of them with word cloud. The word cloud is made of the top features' names and the shapes represent the weights which are average of this feature for one user and one activity.



(a) User1 running
(b) User1 walking
(c) User1 transport
(d) User2 cycling
(e) User2 walking
(f) User2 transport
(g) User3 bus
(h) User3 walking
(i) User3 transport
(j) User3 car

Fig. 6: Word clouds for each of users and each of activities

From the word clouds pictures, we can compare them horizontally(different activities for same user) or vertically(same activity in multiple users). Here, as we hope to fill the missing activity, we focus more on horizontal comparison. We can observe that there is difference between activities. Take user3 as example, in Fig. 6g, we can find "forklift" in there with large font size which represent its weight, while it does show in Fig. 6h, 6i and 6j. Therefore, we can use "forklift" to distinguish "bus" from other activity for user3. Beside of this single feature, we can find other features that are unique for one activity of one user.However, we can also observe that there are many features common to show in multiple categories. It might be not helpful for the classification.

3 Modelling and evaluation

3.1 Random forest model

For this paper, we chose Randomforest as classifier for multi-classification . Random forest(RF) is an ensemble learning model based on decision tree. It uses bagging theory which means random forest use multiple decision trees to train with different random data from whole dataset. The trained model $\hat{f}$ can be formally described as below:

$$\hat{f} = \frac{1}{B} \sum_{b=1}^{B} f_b \left(x'\right) \tag{1}$$

It can be represented as the average of prediction of all single decision tree f_b by B times. The Random forest model has many benefit for our task. First, it is based on tree model, which can handle non-linear relation and good at multi-classification task;second, it is an ensemble model which can reduce noise that common in single tree model.

3.2 F1 score

To evaluate our model, we use F_1 score as metrics because our task is to classify multiple activities. Therefore, for each category we take both precision and recall into consideration.

$$F_1 = \left(\frac{recall^{-1} + precision^{-1}}{2}\right)^{-1} = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{2}$$

In our case, we calculated both overall F_1 score and F_1 score for each single category.

3.3 Parameter tuning

For Random Forest model, there are multiple parameters that should be considered. In our case, we mainly focus on three parameters: max depth of tree, max input features and number of trees/estimators. Besides, to ensure the robustness, we performed 5-fold cross validation. The way we used to find the best parameter setting is gridsearch. The max depth of tree is from 2,5,10 and 20, the higher the model has better performance on training set and higher chance to get overfitting. The max input features(probation) are from 0.1,0.5 and 0.7, which represents the proportion of features are used in tree model. The number of trees is from 100,200 and 300, the higher the better while it would cost more time and get no improvement when larger than some extent.

3.4 Results

The whole data set is divided into training set and testing set with random extraction which means samples in adjacent time might be in different sets because our final goal is to fill up the missing log. The training set contains 70% of whole dataset and rest are in testing set. The results are show in Table 3, the scores in overall and for each activity, and the parameters of the best model are shown for different feature settings. We extract 15 PCA features, 3 clustering features and 62 top content features in Section 2. Here, initial data set contains 1000 CAFFE concept; ‘Init + PCA + cluster’ dataset contains 1000 basic features, 15 PCA features and 3 clustering features; only 15 PCA features are covered by the dataset with PCA features; ‘top content features’ dataset only contains 62 top content features; and ‘top content features’ dataset consists of 62 top content features and 3 clustering features.

	Overall		F1 scores of each activity				Best parameter setting		
Data set	Train	Test	Bus	Car	Transport	Walking	Max depth	Max features	N_esimators
Initial	0.995	0.800	0.8	0.	0.787	0.810	10	0.1	300
Init+PCA+cluster	1.0	**0.804**	0.824	0.	0.779	0.813	20	0.1	200
Only PCA features	1.0	0.715	0.644	0.	0.705	0.756	20	0.1	300
Top content features	1.0	0.783	0.786	0.	0.758	0.798	20	0.1	200
Top content+cluster	0.988	0.766	0.772	0.	0.730	0.788	10	0.1	100

Table 3: Results of parameter tuning

From this table we can see that the initial feature plus PCA features and cluster features achieves the highest score on testing set with 0.804 F_1 score. However, we can see that the margin between its performance and that of initial data set is not significant. And note that the depth of tree is bigger that of initial set while the number of estimators is smaller, we think the PCA and cluster features have little help for model performance as both two feature settings have more model complexity in one dimension and less complexity in another dimension. We reckon that the performance for both of them with same parameter setting will be close and the model can ”learn” the latent information extracted by PCA. Besides, all of the best models used only 10% of all input features. Therefore we think for each activity there are few unique concepts related to them and the performance of ”Top content features” proves it: only few concepts can help to identify the activity.

As to the performance on each activity, the confusion matrix is shown in Fig. 7 (label 0-3 are ”bus”,”car”,”transport” and ”walking”). We can see that for ”car”, the performance is 0 as in training set this kind of activity is not covered which means the model can not classify it. Beside it, the performance on other activities are good. And it has lower F_1 score on transport, we think it is because the definition of transport is fuzzy.

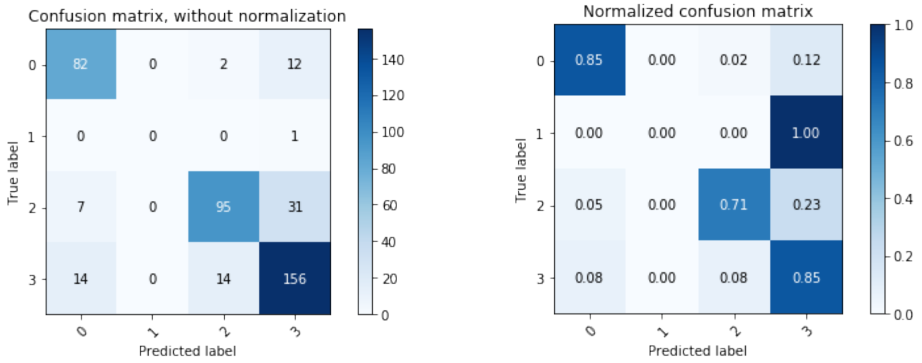


Fig. 7: Confusion matrix for initial+PCA+cluster set

4 Discussion

As we talked about the reason we don't use location information, the few number of sample will lead to more situation like "Car" in activity: the whole related data are covered on only one side. Another issue is that we don't have negative samples, therefore we can't "stop" identifying. For model training, there are two ways of sampling:random division or split by date. Since we hope to impute the missing activity we chose the first way. If we only use wearable device for identification, then we should apply the second way of sampling and the performance were about 0.67 in overall. For parameter setting, the smallest number of max features is 0.1, for initial set it can be 100 while for others the numbers are about 2-3. We think it is not fair for them. On the other hand, there is still a space for initial set to reduce the number of input features since according to the default setting of model, the number of feature will be square root of total which is about 33 and much lower than 0.1 of total. Due to the time limit, we didn't explore it deeper.

5 Conclusion

According to the performance of models with different feature and parameter settings, the task that classifying different activities can be solved with high accuracy. We think based on this method, we can identify the untracked activities though the concepts of images taken by wearable devices. And then, when we have more identified activities as training data, we can outline the whole day of users by those images.

References

1. Cathal, G., Alan, F.S., Aiden, R.D.: Lifelogging: Personal big data. Foundations and Trends in Information Retrieval **8**(1), 1–107 (2014)
2. Cathal, G., Hideo, J., Frank, H., Liting, Z., Rami, A.: Ntcir lifelog: the first test collection for lifelog research. In: SIGIR'16 - the 39th International ACM SIGIR conference on Research and Development in Information Retrieval. pp. 17–21 (2017)
3. Hoogendoorn, M., Funk, B.: Basics of Sensory Data, pp. 15–24. Springer International Publishing, Cham (2018). https://doi.org/10.1007/978-3-319-66308-1_2, https://doi.org/10.1007/978-3-319-66308-1_2
4. Hoogendoorn, M., Funk, B.: Clustering, pp. 73–100. Springer International Publishing, Cham (2018). https://doi.org/10.1007/978-3-319-66308-1_5, https://doi.org/10.1007/978-3-319-66308-1_5
5. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: Convolutional architecture for fast feature embedding. In: Proceedings of the 22nd ACM international conference on Multimedia. pp. 675–678. ACM (2014)
6. Maaten, L.v.d., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(Nov), 2579–2605 (2008)
7. Martin, D., Rob, K.: outlines of a world coming into existence: Pervasive computing and the ethics of forgetting. Environment and Planning B: Planning and Design **34**, 431–445 (2007)