StudentLife

COMP9417 project

Zhu qinshi z

Yang Keyu z5177443

Semester 3 2019

Introduction

With rapid development of information technology, information explosion has become difficult issues in various industries at this stage, which means massive data can not be counted and analyzed by human in a short time， especially in educational system. Manager may have difficulty in obtaining effective information about student life.

In this case,  it is essential to handle large amount of data and information. In order to predict health and behaviour, prediction methods like KNN, decision tree arises to makes use of many features to complete the prediction. There are two psychology-related phenomena manager can utilize to evaluate their performance. These prediction methods aims at making use of many smart phone features to speculate current state of students. Prediction and most of students' behaviour are based on existing user groups data.

 DATA

The dataset we used is divided into **input section** and **output section.**

For input part, the first one is **Physical Activity Inferences** which reflects the status of students as time goes on. And there are four types of student’s status which respectively are Stationary, Walking, Running and Unknown. And after calculating the amount of these four types of all the students, we found that the majority part is the stationary status.

The second one is **Audio dataset** which indicates the participant's physical audio inferences in 10 weeks, and the audio inferences also divided into four types, Silence, Voice, Noise and Unknown. The majority section of students’ phone audio is Silence,and then followed by Noise proportion。

**Conversation dataset** includes each communication’s start time and end time, and we could use it to calculate the length of whole conversation of one student.

**GPS Location** give us the statistic of students location which is reflected by latitude, longitude, altitude these three fields. And the network and provider and network type are another important fields which can be expressed by one-hot encoding.

The next one is **wifi-location**. And it provides us the location of mac address. And **Wifi-location** calculate each participant's on-campus rough location and we could infer the GPS coordinates of each building[1]. For example, in timestamp 1364357009 and the content in location field is in[kemeny ,then we could know the location of the students.

The last three datasets respectively are **Light, Phone lock, Phone charge**, which has the same data format as the Conversation dataset and reflects the time duration of phone information of students.

For **output section,** there are two scores we have to predict which are **flourish score** and **pana scores.** Flourish score is a single psychological well-being score and its dataset provide us the student’s life feeling in 8 aspects[2]. And in pana datasets, it has two different score metrics which are positive affect score and negative affect score, and the mean value of positive score is 33.3 and negative is 17.4. This dataset indicates the feeling of each students in past week.According the reference, the flourishing scale is associated with conversation duration and number of co-locations. Similarly, the pana scores is related to conversation, activity, and co-locations too.

DATA GROUPING

We have used the combinations method which aims to all the features from the input file,

And in feature importance criteria of our paper, we will discuss about how we group the data explicitly.

Method

Clear presentation of methods applied

Decision tree

Decision tree is a machine learning method which is a tree structure, and each internal node of the decision tree represents the judgment on attribute, different judgment corresponds to different conditions,and each branch represents the output of one judgment result, and finally each leaf node represents only one classification result.

Decision tree is a very common classification method, but it can only be established by supervised learning. A bunch of samples and their labels are given to us in supervised learning, and each sample has many attributes and one known classification result. We create this decision tree through learning  these samples by the labels, and then this tree can give the correct classification of new data.But Decision tree is easy to have the overfitting problem if there are too many leaf nodes or the maximum depth is high.

KNN

The core idea of the KNN algorithm is that if most of the k nearest samples in the feature space belongs to one certain category, then the sample also belongs to this category and has the same features which the majority class has in this category. In the decision-making of classification, KNN method only determines the category of the samples which can  be classified according to the category of the nearest one or several samples. In other words, the KNN method is only related to a small number of adjacent samples. This is because the KNN method mainly depends on the limited adjacent samples nearby, rather than on the method of identifying the class domain to determine the category. It is better when the sample set to be divided with more overlapping or overlapping class domain compared to other methods .

SVM

Support vector machine (SVM) is a kind of generalized linear classifier which classifies data according to the supervised learning method. the decision boundary of svm is the maximum margin hyperplane for learning samples

Clear presentation of pre-processing and feature extraction (if applicable)

**Pre-processing and feature extraction:** Several significant features have been introduced in [3], which is attached in the given dataset. Some features are easy to extract while others are not. Here are the examples about pre-processing and extraction of some most important features:

(1) **conversation**. The feature conversation is divided into two parts that are duration and frequency. The sensing data of conversation has two start and end timestamps which is easy to implement. The gap between the end and start timestamps indicate the time of a single conversation. The amount of all conversation represents the duration and the length of all rows in the csv file shows the frequency. Our design is to divide conversation data into different duration of one day to reflect the student’s communication information. And the duration is respectively day (9am - 6pm), evening (6pm - 12pm), and night (12pm - 9am). We also split the timestamps into these three sections in other features.

(2) **activity duration:** the method of calculating activity duration is similar as that of conversation frequency. The accelerometer records the state of motion of each phone every 2-3 seconds whether is sedentary or moving. As the probability of 2 or 3 seconds is close, we count all records as one time if the state is not sedentary (type 0).

(3) **co-location:** the sensing data of Bluetooth is used to calculate the number of co-locations. In order to remove the phones passed by, we should record how many times the certain MAC address has appeared in the sensing data and set a threshold to eliminate these distributions. As for this project, the threshold is set to 6. After that, we calculate the total times of the MAC addresses have appeared.

(4) **sleep duration:** instead of using wearable devices, [4] introduce completely unobtrusive way to predict sleep duration. As the result of their paper, 4 features are used including stationary, silence phone-lock and phone-charge and their ecoefficiencies are 0.5445, 0.3484, 0.0512 and 0.0469 respectively. At last, the feature will be classified into different time period depends on the requirement and the time zone should be set to America/Ne.

 (5)panas score: the questions of panas can be divided into two parts, the first one is positive score and another is negative score,and we calculate the total score of each students for these two choices. Because the value of students is not complete and some students features value are missing. For the fairness and calculation , we compare the situation which use mean value or zero.to replace these values.

(6)flourish score: similar as the panas score, the flourish score are also missing some values, so we also compare the performance then using these two values.

Result

We use three different classification models to predict two psychology-related phenomena which are flourishing scale and panas score. By using different combinations of data to fit model and setting optimal parameters, we get three models which have the highest accuracy on the test data set. Here is the accuracy table of three models

|  |  |  |  |
| --- | --- | --- | --- |
|  | Flourish score | Positive panas score | Negative panas score |
| Decision tree | 0.8260869565217391 | 0.8260869565217391 | 0.717391304347826 |
| KNN | 0.8478260869565217 | 0.8478260869565217 | 0.7608695652173914 |
| SVC | 0.8043478260869565 | 0.7608695652173914 | 0.6739130434782609 |

Table1: highest accuracy of three models

**Metrics**

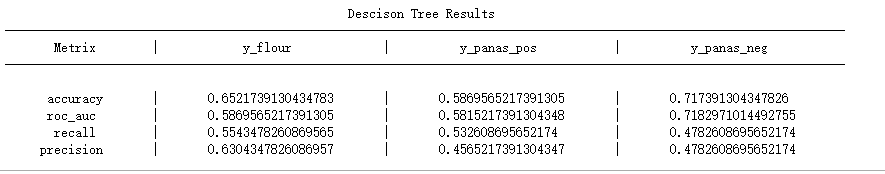
We tried four metrics to represent the performance of the model. By setting the ‘score parameter, it is easy to get the corresponding score of the model. **accuracy**: The proportion of the correct number of samples to the total sample.

**roc\_auc**: roc\_auc score, area under the ROC curve

**precision**: samples predicted correctly as a percentage of all predicted positive samples

**recall**: positive samples predicted correctly as a percentage of all predicted correctly samples

Here are metric score features we get when using model to fit all features.

 Figure1: decision Tree metrics

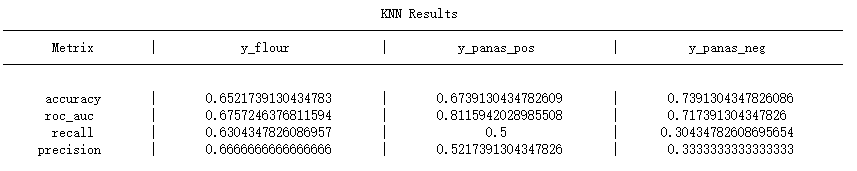


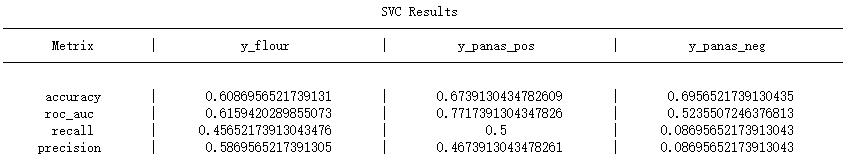
                                                       Figure2:  KNN metrics

                           Figure 3: SVC metrics

From features we can see, a model with a high accuracy may have a low recall and precision score. Because a model may have a low ability to predict correct sample. In order to make the model perform better overall, we chose acc and roc\_auc to reflect the performance of a model.

From table1, among the three methods, the decision tree model has the highest accuracy on all three labels. But except for svc, which does not perform well when predicting the flourishing scale, all other models have nearly 70% accuracy when predicting three labels.

1. **Decision tree:**

Decision tree classifier has three main parameters the first one is max\_depth which limit the depth of tree models and the second one is min\_sample\_split which control the minimum number of samples required for internal node subdivision．And the third one is min\_samples\_leaf.

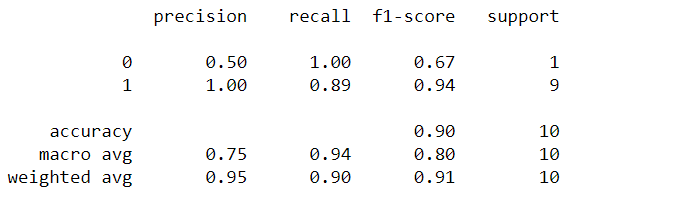
1.flourish score

The using features are x\_con\_duration\_day, x\_con\_freq\_day, x\_activity\_day,

x\_traveled\_day, x\_indoor\_night respectively represent the total conversation time in one day, conversation frequency, activity time and

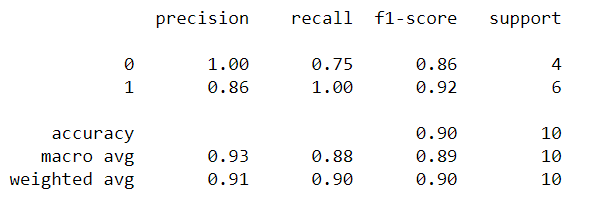
indoor time at night.

And we could obtain the max\_depth is 4 and min\_sample\_split is 2 when we use the grid search and cross validation to prevent overfitting and find the optimal solution for decision tree, which is 0.8260869565217391. The graph below is the score of using the optimal parameters.



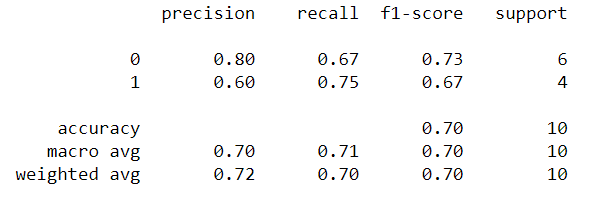
2.for panas positive

Parameters: max\_depth=8, min\_samples\_split=10



3.for panas negative

And if we add the 'min\_samples\_leaf' parameters, the accuracy get decreased which is 0.717391304347826. And the best parameters is {'max\_depth': 2, 'min\_samples\_leaf': 2, 'min\_samples\_split': 4}



1. **KNN**

Knn classifier has 2 important parameters: n\_neighbors and weights. The first one indicates how many neighbor points is used and the second one decides how to calculate distance of which the default setting is Euclidean distance.

In this project, Euclidean distance is not suitable because the features are using different measurements. The distance will be dominated by the features of larger numbers, so we should do further preprocessing before we train the model.

**By applying grid search method, we can easily get the model of optimal accuracy. For all the experiment in this paper, we use 80% for training and 20% for testing. As for grid search, we use 5-fold cross validation to prevent overfitting. The best accuracy is 0.695 and the corresponding parameters are {'n\_neighbors': 16, 'weights': 'uniform'}. The accuracy is not satisfied.**

**After this we tried to find a better model by adding or reducing features.**

**It would take huge amount of time to compute when using combinations, but we have a lot. As is described above, the Knn is sensitive to number size. If we use ‘distance’to train model, the best result is 0.783 and the features used are traveled distance and traveled distance during night. The higher accuracy may not suggest a better model and as can be expected, the traveled distance is relatively large numbers.**

**After applying min-max scaler, the result is better. The average accuracy return by the grid search is 0.870 and the 6 feature used is x\_con\_duration\_nig, x\_con\_freq\_eve, x\_co\_location, x\_traveled\_day, x\_indoor\_day, x\_indoor\_nig**

1. **SVC**

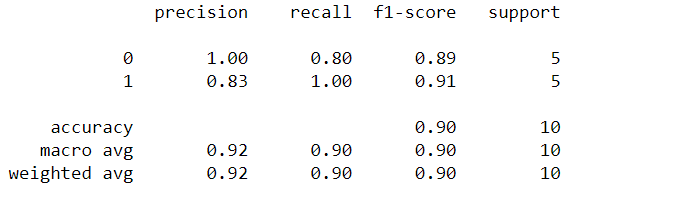
There are three important parameters which respectively is kernel,gamma and C.

kernel: we choose the default value ‘rbf’ gamma: Kernel coefficient for ‘rbf’. C: the penalty coefficient C of the objective function, using the balanced classification interval margin and the wrong sample, the default C is 1.0. We expand the value of C (1, 10, 100, 1000) to seek the optimal solution.After getting lists value of a hyper parameters, we pass these lists to gridSearchCV model.

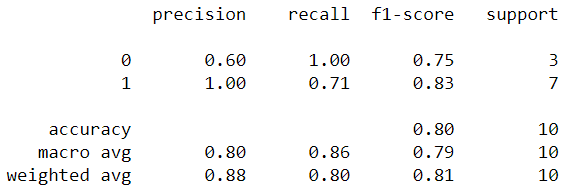
1. For flourish score

The first five features we test are x\_con\_duration\_day, x\_activity\_nig, x\_traveled\_day, x\_traveled\_eve and x\_traveled\_nig

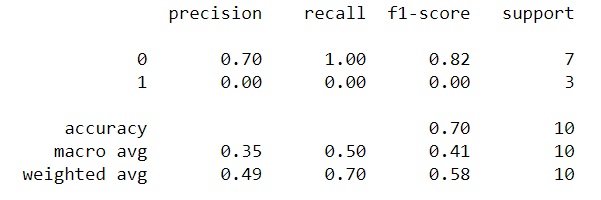
The highest accuracy is 0.8043478260869565 and the optimal parameter is {'C': 1000, 'kernel': 'linear'}. The graph below is the score of using the optimal parameters.

 '

1. For pana positive score

The features we use are by x\_con\_freq\_day, x\_con\_freq\_nig,x\_activity\_nig, x\_traveled\_eve ,x\_traveled\_nig, x\_indoor\_day from the reference, and then we can obtain the score is 0.7608695652173914.

3.for panas negative score, we can obtain the score is 0.6739130434782609



**Features**

There are total 8 features we generate from the dataset. They are physical activity inferences, audio dataset, conversation dataset, GPS location, wifi location, light, phone lock and phone charge. Different features have different effects on flourish score, panas positive score and panas negative score. In order to find use which feature to train model can get best performance on test data, we use a combination algorithm to get the full combination of these features and use the generated training set to train the model to find the most accurate model. Finally, we get that the GPS location, wifi location and physical activity features influence the flourish score most. The conversation and GPS location feature influence the panas positive score most. The conversation and wifi location features influence the panas negative score most.

reference

[1]<https://studentlife.cs.dartmouth.edu/dataset.html>

[2]

Diener, E., Wirtz, D., Tov, W., Kim-Prieto, C., Choi, D., Oishi, S., & Biswas-Diener, R. (2009). New measures of well-being: Flourishing and positive and negative feelings. Social Indicators Research, 39, 247-266.

[3]. Wang, Rui ; Chen, Fanglin ; Chen, Zhenyu ; Li, Tianxing ; Harari, Gabriella ; Tignor, Stefanie ; Zhou, Xia ; Ben-Zeev, Dror ; Campbell, Andrew. [StudentLife: assessing mental health, academic performance and behavioral trends of college students using smartphones](https://primoa.library.unsw.edu.au/primo-explore/fulldisplay?docid=TN_acm2632054&context=PC&vid=UNSWS&lang=en_US&tab=default_tab&query=any,contains,StudentLife:%20Assessing%20Mental%20Health,%20Academic%20Performance%20and%20Behavioral%20Trends%20of%20College%20Students%20using%20Smartphones). Proceedings of the 2014 ACM International Joint Conference on pervasive and ubiquitous computing, 13 September 2014, pp.3-14

[4]. Z. Chen, M. Lin, F. Chen, N. D. Lane, G. Cardone, R. Wang,

T. Li, Y. Chen, T. Choudhury, and A. T. Campbell. Unobtrusive

sleep monitoring using smartphones. In Proc. of

PervasiveHealth, 2013.