Strava Data Analysis Report II

Yiliang Zhang 17/11/2016

In this period of time, what I try to do can be summarized in 4 points:

1. Try to scrape data in time series for the same athlete. Then use these data to figure out the change of some statistics in time series.

2. Use API to get more details for every activity, try to find the covariances of some statistics (time, distance and so on).

3. Considering that the data from API contains athletes' gender, try to use pattern recognition methods to find a way to predict an athlete's gender via other statistics of an activity.

4. If there are valid data for club members, then comparing the change of data in their activities might help to conclude whether there is a conspicuous "peer effect" in the athlete groups.

1. SETBACKS OF STUDY IN POINT ONE & FOUR

When I try to complete the first point above, I face two big problems:

1. The activity data from the same athlete are in everyone's " Training Logs" (which contains one's every activity in the time line) in strava, but most of the athletes set their training logs as private so I cannot get their data. In fact, I found that no more than 5% athletes in strava set their training logs public. Thus, it is quite hard to get sufficient data to study the change of one athlete's activity data in a statistical level.

2. Data in training logs can only be taken in login status. But I am not able to find the way to let my computer scraping while maintaining the login status. Normally in programming, I only need to request to the webserver attaching request header containing login user id, password and URL information. But strava’s webpage needs a dynamic access token(this token is not the same as access_token needed by strava API) for every request. This token is produced unknownly(which means I don’t know the rule of producing that token) inside the web browser.

One thing that I need to mention is that in the previous task, the python code I wrote for data scraping is NOT able to log in. But the distance data, time data and activity type in every activity do not necessarily require a login status, which is totally different from this task.

In these two problems I think the second one is much more difficlut to solve. After all, if I search sufficient number of athletes, there will be sufficient training logs there that are available and the first problem can be solved.

Then I tried to find if the strava API and the python library in *https://github.com/hozn/stravalib* can solve the second problem. However, I am not able to input an athlete id and get all his activity data; instead I need to input an activity id and get the data of that specific activity. Thus if I want to get all the activity data from one specific athlete, I need to input all activities' ids, get all activities' information and filter the activities whose athlete is the one I want -- an impossible way.

Until now I could not be able to solve the second problem . I have already spent entirely several days trying to figure out how to write the code to login but ended up with a setback. I communicated with several of my friends who major in Computer science in Tsinghua University. Unfortunately they could not make it as well. Therefore I suppose temporarily I am not able to finish point 1 and point 4 although I have a strong curiosity to study that.

2. RESULTS IN STUDY OF POINT 2 & 3

In the following contents, I will mainly discuss the result I got after studying point 2 and point 3. In this part, I am mainly going to The python library in *https://github.com/hozn/stravalib* is not actually able to get detail information about one specific activity. The corresponding function is not workable. Thus via the API, I write my own R code to get the data. What I am interested in is the relationship between one activity's:

1. athlete gender

2.distance

3.time

4.max speed

5.activity type

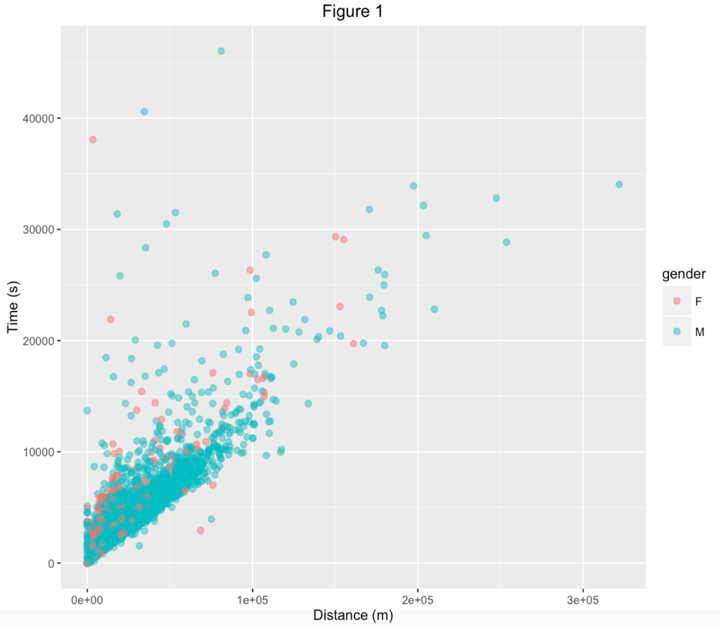
The reason why I choose these five aspects is that I believe these 5 aspects can roughly cover all the feature of one activity.

2.1 Some Basic Statistics of Data

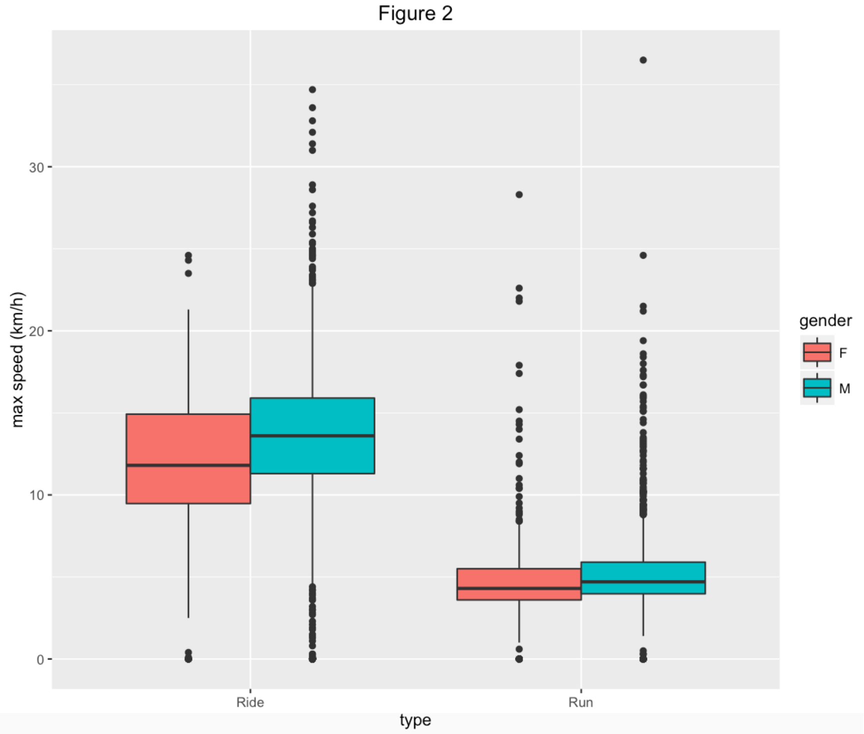
Via the API, I download 4265 valid data for study and here is a brief summary of the row data:

	female	male	ride	run
Row data number	679	3923	2924	1341
Average riding time/s	4700	4719	4718	
Average riding distance/m	25701	29774	29388	
Average riding max speed/(km/h)	12.08	13.47	13.34	
Average running time/s	2908	2809		2833
Average running distance/m	6769	7743		7507
Average running max speed/(km/h)	4.96	5.30		5.20

Figure 1 contains the distribution of data in time and space dimension, I mark every point's color with athlete's gender. There is no surprise that male athletes' average activity speed is higher than female's, as the red points are comparatively upper than green points.



The boxplot in figure 2 shows a distribution of activity type and max speed of female and male:



2.2 Covariance between differnt data dimension

I separate the data with gender and type into 4 groups and get the matrix of correlation coefficient respectively. And here's is the result:

female ride	distance	time	max speed
distance	1	0.924	0.326
time	0.924	1	0.187
max speed	0.326	0.187	1

female run	distance	time	max speed
distance	1	0.541	0.315
time	0.541	1	0.111
max speed	0.315	0.111	1

male ride	distance	time	max speed
distance	1	0.879	0.418
time	0.879	1	0.320
max speed	0.418	0.320	1

male run	distance	time	max speed
distance	1	0.815	0.245
time	0.815	1	0.132
max speed	0.245	0.132	1

From the table above we can easily find that for both female and male, riding activity's time, distance and max speed are more related to each other than running. One possible explanation for this is that riding speed does not have such disparity between athletes as in the running activity. After all, riding is strongly dependent on bikes, and there is not so much difference among bikes than the difference of muscle strength among athletes that decide the speed of running.

By showing the relation between variables, correlation coefficients might indicate that if one dimension of data has some conspicuous behavior, whether or not other dimensions will also show similar behavior will depends on the correlation between these two dimensions. That’s why in the previous task, I found that both data in time and distance dimension have a “cluster” behavior.

2.3 gender prediction

In this section, four methods of pattern recognition is being used to see if there is an appropriate way to distinguish the gender of an athlete only by one of his or her ride or run activity, which contains the activity type, time, distance as well as the max speed.

(1) Fisher linear discriminant analysis

Firstly, I use Fisher linear discriminant analysis to see if there is an linear way to discriminate the gender in the high dimension space.

I separate the total 4265 data into training set(2000) and test set(2265). Via R code my result in the test set is in the following table:

true test	male	female
male	1938	320
female	7	0

And the discriminant error rates are:

$$\begin{cases} \varepsilon_m = 0.2\% \\ \varepsilon_f = 100\% \\ \varepsilon_{all} = 7.7\% \end{cases}$$

Obviously, the Linear discriminant does not fit our data, female data can’t be distinguished by linear model.

(2) Support Vector Machine (SVM)

Next, I use the SVM to try to use non-linear discriminants for separation. I respectively chose the kernel function of the machine as linear, polynomial and sigmoid and find that when choosing polynomial of degree 2, the error rate is comparatively lowest and results here:

true test	male	female
male	1241	123
female	704	197

Discriminant error rates are:

$$\begin{cases} \varepsilon_m = 36.1\% \\ \varepsilon_f = 38.4\% \\ \varepsilon_{all} = 36.5\% \end{cases}$$

As we can see, our data is more likely to be fitted with non-linear model.

(3) Back Propagation neural network (BPNN)

Third, I used the Back Propagation neural network (BPNN) for discrimination. I built several kinds of network structure and find that the simple hidden layer network has comparatively the best result. After careful select, I choose the structure (5-1-2) as my final network type (5 input nodes, 1 node in hidden layer, 2 output nodes). One thing that I need to mention is that I expand the type and gender dimension separately into 2 dimensions, that is every gender data corresponds with a 2-dimension vector: “M” ～ (0,1) and “F” ～ (1,0). Same transformation is made in activity type data. This transformation will make the analysis easier:

true test	male	female
male	1200	135
female	745	185

Discriminant error rates are:

$$\begin{cases} \varepsilon_m = 38.3\% \\ \varepsilon_f = 42.2\% \\ \varepsilon_{all} = 38.9\% \end{cases}$$

Usually, neural networks have a good behavior in pattern recognition, but in our case, this method does not seem to have an ideal result. The reason might be that in our data, the female part and male part of data have much superposition with each other, which is obviously impossible to be distinguished by linear discriminant model just like the Fisher method. Thus, considering the BP propagation is also operated in linear method (the key work is to decide the weight of each line between two nodes in two adjacent layers), this result may not be a disappointing result as a linear method.

(4) Logistic regression

All these three methods do not give an ideal discriminant result here (all error rates are larger than 30%). Finally, I choose logistic regression model for discriminant and receive a better result thereafter:

true test	male	female
male	1735	92
female	210	228

Discriminant error rates are:

$$\begin{cases} \varepsilon_m = \ 10.1\% \\ \varepsilon_f = 28.7\% \\ \varepsilon_{all} = \ 13.3\% \end{cases}$$

The discriminant error is much more smaller than previous, because the logistic model fits the situation of “Yes/No” problem (our case is “M” and “F”) quite well.

3. Summary and Discussion

In this report I mainly analyze data downloaded from strava API in three aspects: first is to see the basic statistics of the data if there is any interesting phenomenon; second is to study the covariance of different dimensions; third is to predict athlete gender via other his activity data. What impresses me most is the third part as when I use different method for discriminant, the results have such a disparity.

Throughout my study in strava data, I find that data from sports-based social platform is amazing. It not only contains single activity and single athlete data, but also athlete groups and friends network data, which means that I can not only study the behavior of sports data but also social networks and other social phenomenon. For study in “target behavior”, strava has the potential to make the study easier as it provides athlete with a record on which athletes can set their target in time or distance in advance before activity. But it is disappointing that current I am not able to scrape data in login status, thus this information cannot be fetched.

Unlike the previous task, this task is not processed smoothly. In fact I spend a large amount of time attempting to solve the problems described in first part of this report, but still I am not able to make it. If I am able to get one athlete's all activity from strava, I will be able to see the behavior of his activities' statistics in time series. I have already get the top 200 athlete numbers in a marathon held in New York. I would also be able to compare the activity's statistics of professional athlete with normal exercisers. Besides, I find that strava API provides me with members' ids in every clubs, so if I have the data in time series, I will be able to study the "peer effect" in athlete clubs by studying the change of their activity data and the similarity of their activity with their club members.