Social Media User Modeling

Yuanfei Pan

Abstract

In this study, we designed an item-based recommendation system to predict users' favorable POIs (i.e. Point of Interest), built SVMs and Bayes Networks for predicting users' gender information. In POI prediction task, we have 35.12% prediction accuracy, and in gender prediction task the accuracy is 76.39%.

1 Introduction

Social media has become the most popular way that people contact each other. As people can publicly post anything that they want to share on social media, they generate heterogeneous information on those media platforms. This kind of user information is especially useful for business decisions such as targeted advertisement. The term "User Modeling" describes the process of building up a conceptual understanding of the user. The main goal of user modeling is customization and adaptation of systems to the user's specific needs. Since user behavioral data on social media is heterogeneous, it's still challenging to effectively leverage the heterogeneous information for user modeling.

In this study, we are provided with a social media dataset including three kind of different information: geographical information, users' personal information and natural language texts (tweets posted by users). We are going to complete the following two tasks: (i) Interested Location Prediction, given users' some historical location visits and other provided information, predict what locations a user may be interested to visit in the future. (ii) User Profiling, given users' other information except profiles; predict each user's profile information (gender information).

We interpret task (i) as a recommendation task and build up an item-based system to perform recommendation. For task (ii), we trained two kinds of different binary classifiers, namely SVMs and Bayes Nets, to perform user profiling. The performance of both systems is satisfying. Basically, the prediction accuracy of task (i) is 35.12%, and prediction accuracy of task (ii) is 76.12%.

2 Dataset

The dataset is retrieved from a social media platform. For each user, we have data on his previously visited POIs, anonymous tags, his posted tweets and other users he followed. The POI information is a tuple of unique ID of POI, primary category, secondary category, precise category, latitude, longitude and name of POI.

An overlook of dataset is shown in Table 1. A significant characteristic of this dataset is that it is missing lots of data. For example, we have 3122 users' check-in data, while having only 946 users' tweets data, which mean 2176 users' tweet data are missing. Due to missing data, any system works on this dataset must be designed to make prediction on partial information.

Data Type	Users	Records
Check-in	3122	319,095
Tag	2013	8,667
Social	3094	3,094
Profile	3121	3,121
Tweets	946	1,640,923

Table 1: An overlook on whole dataset

3 Proposed Approach

3.1 Task I

An item-based recommendation system was built for task one. POIs are seen as items, and in the recommending process, the most similar items will be recommended to users. With the given training dataset, a user database is built to store the known information for recommending. When making recommendation to one user, program search the database for most similar items. To measure the similarity between items, a series of numeric features of items and a corresponding distance metric are specially designed.

3.1.1 Feature Extraction

Vectors of numeric features are created for representing items. We fully used all given POI information except the name of POI as item features, which includes longitude, latitude and three levels of category of POI. An example of feature vectors is shown in Table 2.

3.1.2 Similarity Measurement

Our similarity measurement is based on two basic assumptions: (1) Users will not go to a distant POI that is far away from his range of activity. (2) Users prefer similar classes of POIs. Therefore, closer POIs of similar classes score higher in our similarity measurement. The similarity between two items is defined as below:

$$similarity = distance^{-1}$$

$$distance = \gamma^{\alpha} \cdot d_e$$

$d_e$is Euclid distance between two POIs, and γ^{α} measures the similarity of category classification. γ is a scaling factor; it must be in range (0, 1). α is the similarity of POI category. If the primary category is matched, $\alpha = 1$; if secondary category is matched, $\alpha = 2$; if precise category is matched, $\alpha = 3$; If no category is matched, $\alpha = 0$. Distance can be seen as a monotonous increasing function of d_e or a monotonous decreasing function of α.

3.2 Task II

Users' gender prediction is a binary classification task. Tag data and tweet data are utilized to perform classification. We first build a SVM, which only takes tag data into consideration, and then we try the Bayes Network, which integrates both tag and tweet data.

Longitude	Latitude	Primary Category	Secondary Category	Precise Category
31.3043	120.6835	19	38	NaN
39.9077	116.3488	169	248	82
39.9433	116.0982	19	12	13

Table 2: An example of feature vectors

3.2.1 SVM

Each user has a series of tags. Statistic on the given dataset shows that tags and sex are correlated. So, we use tag as a feature to classify user into male and female. There are 163 different tags in total, so there is a 163 bits vector for each user. According to the occurrence of one tag, the corresponding bit is set to zero or one. As the vector is sparse and high-dimensional, it is difficult for SVM to work well. So we calculated a 50-dimensional "tag-embedding" by multiplying sparse feature vector with a randomly assigned embedding matrix. Then feed the 50-dimensional vectors into SVM to make the classification. Four types of kernel functions have been tested: linear, polynomial, radius-based and sigmoid function.

3.2.2 Bayes Network

A Bayes Network is built to predict user's gender. We first built a simple network with only two nodes, gender and tag, as shown in Figure 1. Then we take natural language texts (tweets) into consideration, and built another Bayes Network with two nodes, which is gender and tweet topic. By applying LSI topic model on tweets posted by users, we got vectors of topics for each user. User's gender information can be reflected by topic vectors. Finally, we integrated two networks and form a new network with three nodes.

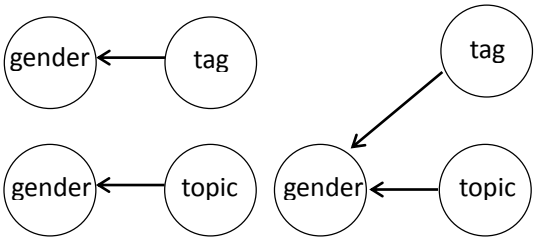

Figure 1: Graphical illustration of three kind of Bayes Networks.

4 Analysis and Discussion

4.1 Task I

The recommend system for task one is evaluated by three indexes.

$$P_i@K = \frac{H_i}{K} \quad (1)$$

$P_i@K$ measures the accuracy of prediction. K is the number of POIs to predict, H_i is the number of correctly predicted POIs. Instead of matching unique ID of POIs, we match categories of POIs, which means a POI is considered to be correctly predicted if its three categories are strictly matched with those of reference data.

$$R_i@K = \frac{H_i}{V_i} \quad (2)$$

$R_i@K$ measures the utilization rate of given information. V_i is the number of visited POIs of a user, which is known to our recommendation system.

$$F_i@K = (P_i@K^{-1} + R_i@K^{-1})^{-1} \quad (3)$$

$F_i@K$ is an integrated index. It is the harmonic average of $R_i@K$ and $P_i@K$.

Our System currently scores 4.12% in F@K, 35.14% in P@K, and 4.67% in R@K. Current system has a relatively high performance in prediction accuracy. There are 280,267 different POIs in this dataset. If we guess at random, the expected accuracy would be far less than 10^{-5}, that means 35% accuracy is good.

However, the utilization rate is low. There is no significant improvement in accuracy when more information about visited POIs is given. As shown in Figure 2, prediction accuracy converges quickly when more information is provided. The way current system utilize these information is inefficient. As we simply search for top K most similar POIs of all given visited POIs, further information underlying these given POIs is not exploited at all.

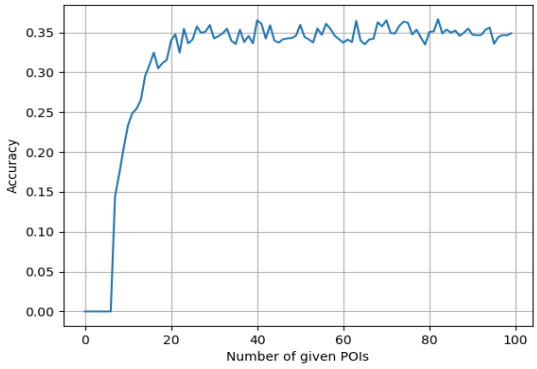

Figure 2: Prediction Accuracy vs. Given Information.

4.2 Task II

We tested two kind of system, SVM and Bayes Networks in this section. The performance of SVMs is shown in Table 3. SVM with linear kernel outperform three other kind of SVM. As the feature vectors we feed into SVM are high-dimensional, there is no good to project these vectors onto even higher dimension; so linear kernel here is the most suitable one.

The performance of Bayes Networks is shown in Table 4. It is clear that integrating both tag and tweet data indeed boost the accuracy. But we have not tuned the parameter of LSI model; there is still possibility of improvement in performance.

SVM Kernel	**Accuracy**
Linear	68.12%
2nd order Polynomial	64.01%
Radius Based Function (RBF)	66.91%
Sigmoid Function (tanh)	50.00%

Table 3. Comparison between kernels

Bayes Network	**Accuracy**
tag only	75.85%
tweet topic only	71.43%
tag and topic	76.39%

Table 4. Performance of Bayes Network

5 Future

The two required tasks have been completed. For task one, the utilization rate of information still need to be improved, we will later consider collaborative filtering and try different similarity metric to improve it. For task two, the later work will be tuning LSI parameters and test LDA model instead of LSI.

References

Cao, J.X., Dong, Y., Yang, P.W., Zhou, T., Liu, B. (2016), *POI Recommendation Based on Meta-Path in LBSN*, Chinese Journal of Computers, **39**(4), pp.675-684.

Wei, H.H., Zhang, F.Z., Yuan, N.J., Cao, C., Fu, H., Xie, X., Rui, Y., Ma, W.Y. (2017), *Beyond the Words: Predicting User Personality from Heterogeneous Information*, WSDM 2017, February 06-10, 2017, Cambridge, United Kingdom

Wang, J.J., Li, S.S., Huang, L. (2014), *User gender classification in Chinese microblog*, Journal of Chinese Information Processing, **28**(6), pp.150-155.

Zou, Y.G., Wang, J., Liu, Z.H., Xia, Y. (2012), *Point of interest recommendation method based on similarity between items*, Application Research of Computers, **29**(1), pp.116-126