Restaurant Recommendation for Facebook Users

Qiaosha Han
Computer Science
Stanford University
qiaoshah@stanford.edu

Vivian Lin
Computer Science
Stanford University
vlin17@stanford.edu

Wenqing Dai
Computer Science
Stanford University
wenqingd@stanford.edu

Abstract

In the past decades, people have gained a wide range of options as the availability of information expands. To help them make decisions, recommendation systems play an important role in all kinds of aspects, e.g. news, books, movies and so on. In this project, we built a restaurant recommendation system by incorporating the power of social networks and local business review sites. To make accurate predictions and provide efficient recommendations, we combined the data from Facebook and Yelp, tested various machine learning algorithms, evaluated the results on real world dataset and made detailed analysis on the experiment results.

1 Introduction

When searching for restaurants information and making decisions on where to eat, people rely on the review sites. But it is possible that the highly rated ones do not align with individual's tastes. How to provide targeted and good recommendation, this problem directly inspires the recommendation system, which is one of the typical challenges in the field of machine learning.

Recommendation system became a hot topic around mid 1990's [1]. During early years, people used context-based approaches where an item similar to the items highly rated by a user is provided as a recommendation for this user. One limitation of this approach is that the result would most likely be over specialized. Also, users with less previous rating data could not be provided with a well-tailored recommendation. Therefore, diversity of the recommendation system is desirable. Another widely used method, the collaborative filtering, has its own limitations too. For example, if the restaurant is reviewed by only a small number of users, it will rarely show up in the recommendation result even if it receives very high ratings from the reviewers. Moreover, efficient prediction from small amount data is very difficult in collaborative approach due to the sparsity of the data. In our work, we treat each user place pair as a data point and classify it to 2 classes, indicating whether this user will be interested in the place and concluding whether the place is a good recommendation candidate for this user.

2 Dataset

2.1 Data Collection

The data included in this work comes from two sources: Facebook and Yelp. We collected user profiles, place information, and user-place relation from Facebook website.

In order to get as much data as we could, we recruited people by sending out a written instruction to acknowledge people about our work and a step-by-step guidance on how to provide us the Facebook access tokens so that we had the permission to collect user information. Then, we gathered additional information about places from Yelp.

The place profiles were extracted after combining the data from the two sites. We have two challenges here. First, how to identify that two places from different sites are the same one? Secondly, how to combine the data of different types, with conflicting or/and duplicate information? In our work, if the same telephone number were in the extracted place information both from Facebook and from Yelp, we believed that these two places referred to the same restaurant. When combining the two pieces of information from different sites, we parsed the XML file from Facebook and JSON file from Yelp, used a great number of heuristic rules to deal with the conflicting and duplicate information, and generated the final place profiles.

In total, the dataset in this work contains 3652 users, 4692 places. Figure 1 shows the Checkin / Like frequencies distributions for places. From the figure, we can see that the user-place relation information is very sparse, and very few restaurants are checkin-ed/ liked by a large number of users.

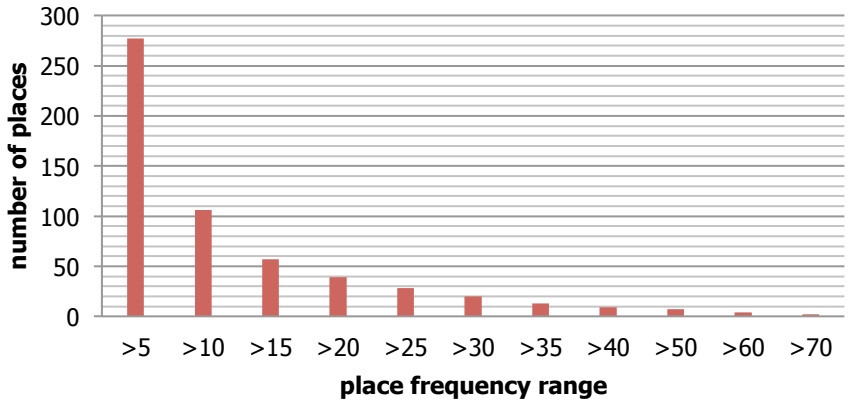

Figure 1. Checkin/Like frequencies distributions

2.2 Features

Based on the information provided from Facebook and Yelp, we had a total of 133 features:

A. **User profile features**: For each user, 33 features that described user characteristics were extracted from Facebook. These features contained information such as hometown, current location, interests, activities, and spoken languages.
B. **Place features**: This set of features described the restaurant characteristics. The profile for a place was the combination of information collected from Facebook and Yelp. For each place, 82 features were extracted from the profile, which captured the information about average rating, locations, payment options, price range, popularity, restaurant categories and so on.
C. **User and place combined features**: These set of attributes described the relations between a user and a place. For each pair, we extracted 12 features

capturing similarity between a user and a place. We considered the factors like the location similarity, the background similarity.

2.3 Sample Generation

In order to maintain the balanced number between positive and negative samples, for a single user, for each place the user liked or checkin-ed, we generated a positive sample. Then we sampled the same number of disliked places from the list of all uninterested restaurants for this user and generated corresponding user place pair as negative samples. Given the fact that the liked/ checkin-ed information was very sparse, we set the place frequency threshold and only considered certain range of frequently liked places when generating positive and negative data points.

3 Experiments and Results

In this work, we conducted experiments on different classification models and investigated feature selection process to gain a better understanding of the problem, the data, and the properties of the model we want. To evaluate the result, we used 10-fold cross validation and precision metrics.

3.1 Different Classification Algorithms

To provide more accurate predictions, we trained different models on the dataset and compared the results. In the experiment, we used the place frequency range as a parameter and tested 8 binary-class classifiers (KNN, K*, Decision Table, Random Forest, SMO, SVM, Logistic Regression, Naive Bayes).

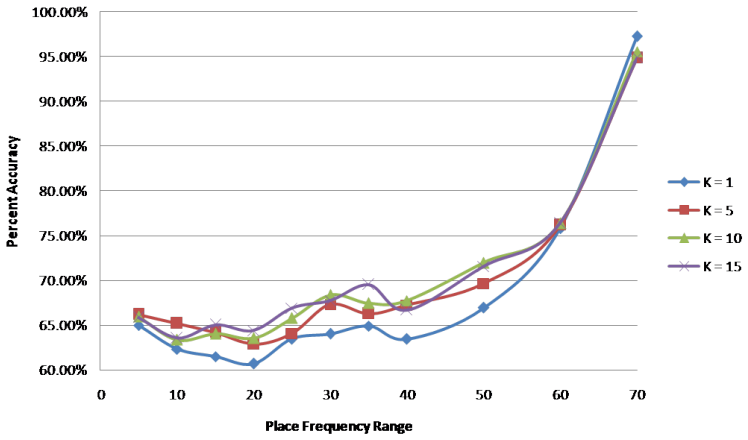

Figure 2. KNN results

For K-nearest neighbors classifier, different values of k were tested. For each k value, we obtained percent accuracies for various place frequency range (>5, >10, >15, >20, >25, >30, >35, >40, >50, > 60, >70). As shown in Figure 2, in general, the accuracy is higher when k = 15, and the accuracy is pretty low when k = 1. This is reasonable because k = 1 does the prediction based on one nearest neighbor only, which might underfit the data.

In addition to KNN, we also tried K*[2], which is similar to KNN but uses an entropy-based distance function to find the nearest neighbors. Also, for the random forest algorithm, we set #Trees = 10, #Features = total # of Features/#Trees; for SMO [3], we use linear kernel; for SVM, we use radial basis kernel. The overall result is shown in Table. 1.

Table 1. Percent Accuracy for different algorithms with different restaurant minimal frequencies

Freq	KNN (k=15)	K*	Decision Table	Random Forest	SMO	SVM	Logistic regression	NB
>5	65.85%	61.09%	68.28%	64.40%	65.47%	49.67%	62.14%	60.31%
>10	63.55%	60.28%	64.69%	58.54%	61.43%	49.70%	56.65%	57.88%
>15	65.05%	59.88%	60.84%	57.12%	62.41%	50.12%	53.65%	54.77%
>20	64.39%	60.33%	62.36%	58.26%	64.89%	49.90%	56.96%	56.08%
>25	66.90%	63.89%	65.96%	59.89%	63.75%	50.05%	55.18%	57.44%
>30	67.72%	66.73%	66.08%	59.80%	64.44%	49.71%	56.28%	56.10%
>35	69.52%	64.05%	65.08%	57.86%	65%	49.60%	52.78%	54.68%
>40	66.70%	64.61%	62.93%	56.23%	62.51%	50.58%	57.17%	53.09%
>50	71.58%	68.99%	71.19%	63.82%	69.25%	50.52%	50.26%	61.89%
>60	76.38%	74.35%	73.99%	68.82%	68.63%	50.37%	59.04%	60.15%
>70	94.86%	94.86%	93.15%	93.15%	93.84%	52.05%	75.34%	78.42%

From our experiments, we can draw the conclusions that percent accuracy is generally higher if place frequency threshold is higher even though there is less number of samples existing in the dataset. This result is reasonable since more users interested in a place then the model would have better understanding of who would like the place. On the other hand, place frequency > 5 is better than >10 ~ > 40 even though the dataset for >5 is more sparse. This happened because dataset for >5 has more samples to train the model.

When running the experiment, we also noticed that SVM is faster but SMO is more accurate. In fact, percent accuracy from SVM is worse than all other algorithms. One possible way that will help to improve the SVM result would be normalizing the feature values before training.

3.2 Feature Selection

We conducted feature selection on three datasets, i.e. low place frequency threshold (10) one, medium frequency threshold (30) one, and high frequency threshold (60) one in order to evaluate the influences. Table 2 shows the result using Decision Table [4]. We can see that the same precision is achieved using 30 features as using all 133 features, and that further feature selection can still improve the precision, both of which indicate the existence of redundant and non-indicative features in the original dataset.

Table 2. Feature selection result for decision table with place frequency threshold 10, 30, and 60

Freq\Feature#	100	70	50	30	20	10
10	64.53%	64.53%	64.76%	64.69%	65.09%	65.56%
30	66.08%	66.08%	66.14%	62.32%	62.09%	63.32%
60	73.62%	73.62%	73.62%	75.83%	71.22%	73.43%

Then we conducted the feature selection for different algorithms and plot the result. We can tell that feature selection has various influences for KNN, DT, LR. But the precision is quite stable before we select less than 50 features, which agrees with our conclusion about the existence of non-indicative features.

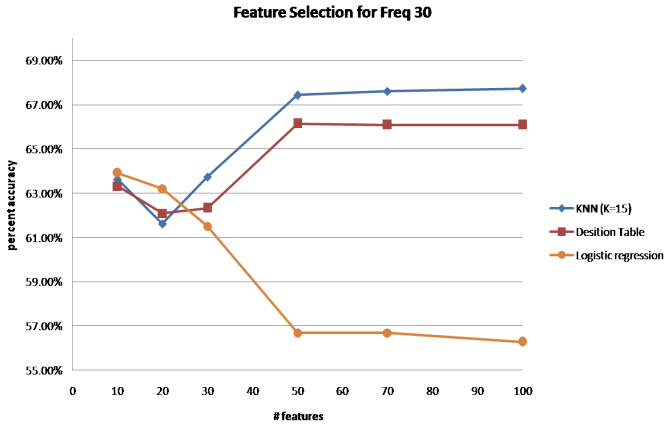

Figure 3 Result of feature selection

After testing on different feature sets, the top 8 features selected are:

- Location (state, city) of the place matches the current location of the user
- Location (country, state) of the place matches the user's hometown
- Type of the place (e.g. Club, Restaurant/Cafe, Bar, Food/Beverages)
- The number of users who ever liked this place on Facebook
- Average rating of the place from Yelp
- Number of reviews in last month from Yelp

4 Conclusions and Future Work

We conducted machine learning based recommendation methods in this work, we conclude that overall, the machine learning model has a better performance on the dataset where places have higher frequencies and the data is less sparse, and that there exists some indicative features to predict the relation between a user and a place. In the future, we can devote more time to feature design. In addition, we should pick the good models and experiment with different parameter settings to achieve better results. Finally, a larger data set would certainly help on a more accurate model and better feature selection.

Acknowledgments

We would like to thank Professor Andrew Ng for his guidance and advice on this project. We also want to thank Andrew Mass for his advice and help on data collection.

References

[1] Resnick, P., N. Iakovou, M. Sushak, P. Bergstrom, and J. Riedl. GroupLens: An open architecture for collaborative filtering of netnews. In Proceedings of the ACM conference on Computer Supported Cooperative Work, p175-186, 1994.

[2] John G. Cleary, Leonard E. Trigg: K*: An Instance-based Learner Using an Entropic Distance Measure. In: 12th International Conference on Machine Learning, 108-114, 1995.

[3] J. Platt: Machines using Sequential Minimal Optimization. In B. Schoelkopf and C. Burges and A. Smola, editors, Advances in Kernel Methods - Support Vector Learning, 1998.

[4] Ron Kohavi: The Power of Decision Tables. In: 8th European Conference on Machine Learning, 174-189, 1995.