Yelp Star is not Everything

Sau-Chin Chen

Introduction:

Yelp has been collecting the users' review for the businesses in the major cities across North America and Europe since October 2004. Based on 5 data set summarized by Data Science Capstone, this project will explore and analyze why a kind of business has reputation in a city. Data set *business* collects the information from 61184 businesses across 10 cities. Data set *checkin* stores 45166 durations and frequencies a user stay in every business. Data set *review* sumarizes 1569264 reviews for a registed business. Data set *tip* summarizes the 495107 recommendations for a business. Data set *user* collects the information from 366715 registed users.

Yelp's recommandation engine has been designed to filter the active and available review articles for a business. A couple of attributes have been summarized from the data accumulated in the past decade. The data set *business* contains 38 variables including 7 variables as the sets of minor attributes. If these attributes constituent the evaluation model how the yelp users evaluate a local businesses, will we figure, for a category of business, a global standard across cities or many local standards in each city? In use of Bayesian network, we set up the criterion to answer this question: When the Bayesian network showed that the city is not the child node of any variable, there is no a unified model for all the businesses of a category across cities. Ohterwise, there is a unified model to tell people how to evaluate a kind of business based on yelp stars or a specific attribute.

Yelp stars is the users' average responeses and is assumed as the index of overall quality for a business. If there is a global evluation model for a kind of business, the stars should be the last considered variable after considered the other atrributes. Based on this working hypothesis, we ploted the explorary analsis on the fast foods and the Chinese restaurants across ten cities. We fcosued on the two categories because they have thousands of data in some cities and the difference between the two categories. Most fast foods are the chained resturants across countries, and many Chinese resturants are managed in a location. The explorary analysis indicated that the yelp stars is the last variable in the network for the fast foods but not for the Chinese resturants. Finally we picked up the fast foods and the Chinese resturants in Phoenix and Charlotte and built the local evaluation models.

Method

Preprocessing Data

Although each bussines in *business* data set has the information of location (full_address, city, state), these variables are a mass because of users' typo. We used the geographical information (longitude, latitude) to identify the location of each business. The result is that we added the new variable 'Loc' to *business* data set. Table 1 lists the number of businesses in every city.

Table 1. Number of businesses in ten cities

Charlotte	Edinburgh	Karlsruhe	Las Vegas	Madison	Montreal	Phoenix	Pittsburgh	Urbana-Champaign	Waterloo
5151	3115	948	16490	2309	3921	25231	3041	627	351

We calcuated how many categories are collected in *business* data set and how many businesses are labeled in every category. There are 2,383 fast foods and 1496 Chinese resturants for the explorary analysis. Before run the explorary analysis, we processed the attribute vectors in the following 3 steps: (1) Delete the attributes conied in less than 10% of businesses; (2) Transfer the rest of attribute vectors to numerics; (3) Tag attribute names with simple characters(A1 to A37). In the exploratory analysis, the target data sets included 37 attributes and the five variables, stars(S), locations(L), open hours(H), review counts(R), and, number of neighborhoods(N).

Briefing Bayesian network

Bayesian network is a probabilistic graphical model that illustrates a set of random variables and their conditional dependencies via a directed acyclic graph (DAG). Take the example in the wikipedia of Bayesian network, 'Grass wet' could depend on 'Rain' and 'Sprinker', and 'Sprinkler' could depend on 'Rain'. The data of every variable have to be the same class, such as Boolean, characters, or numeric. An algorithm of Bayesian network will conduct the conditional dependence tables of every set of variables and decide the highly fitted graphical model for the variables. The arrows in the graph represent that conditional dependence of two variable. In the example of wikipedia, the data of 'Grass wet' depend on 'Rain' and 'Spinkler' and the data of 'Spinkler' depend on 'Rain'. The grapical model is illustrated in Figure 1.

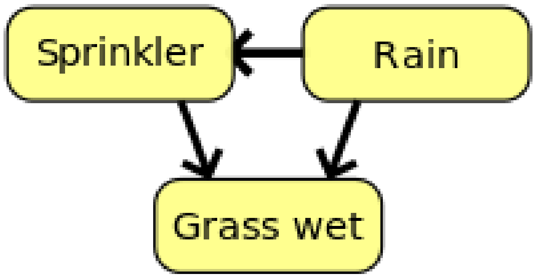

Figure 1. Graphical models of 'Grass wet', 'Rain', and 'Sprinker'

Explorary Analysis

We used hill-climbing(HC) as the algorithm to build the Bayesian network of variables for the data of fast food and Chinese resurants across ten cites. The grapical models were drew by the R package *bnlearn.* The targeted variables are stars(S) and Locations(L). If L is on the top of the network and S is at the bottom of the network, we would obtain a graphical model matched our hypothesis for a category of business.

Build Evaluation Models

To have a validation test on the evaluation model for every case, we adpoted the method in Multiple Quantitative Trait Analysis Using Bayesian Networks. In addition to Locations(L), the rest of four variables and 37 attributes entered the establishment of evaluation models. Based on the explorary analysis of the data from ten cities, we decide build the evluation models for the fast food in Phoenix and Charlotte and for the Chinese resturants in Phoenix and Charlotte.
The final evaluation model for each set of data is the averaged result of the 100 models conducted in 10 rounds of cross-validation processes. Every final evaluation model will pick up the a set of attributes as the top and bottom nodes. The evaluation of the final models depends on the predictive correlations and post correlations of the four variables: stars(S), open hours(H), review counts(R), and, number of neighborhoods(N). If these variables had lower poster correlations than the predicitve correlations, and the averaged post correlations of attributes reach one, the final model will represent the evluation process for a kind of business in that city.

Result

Explorary: Fast Food and Chinese Resturants

The graphical model of the fast foods across the 10 cities (Figure 2A) showed the locations(L) at the top and the stars(S) at the bottom. This meant that fast foods in each city has a specific set of variables to influence the stars. For the fast foods over the countres, the stars is a recommended dimenstion to understand the difference between fast food resturants. The graphical model of the Chinese resturants acrss the 10 cities(Figure 2B) showed the other attributes prior to the locations or stars. People evalute the Chinese resturants on a specific attributes rather than a score, such as a resturant has outdoor seats or TV in the resturant.

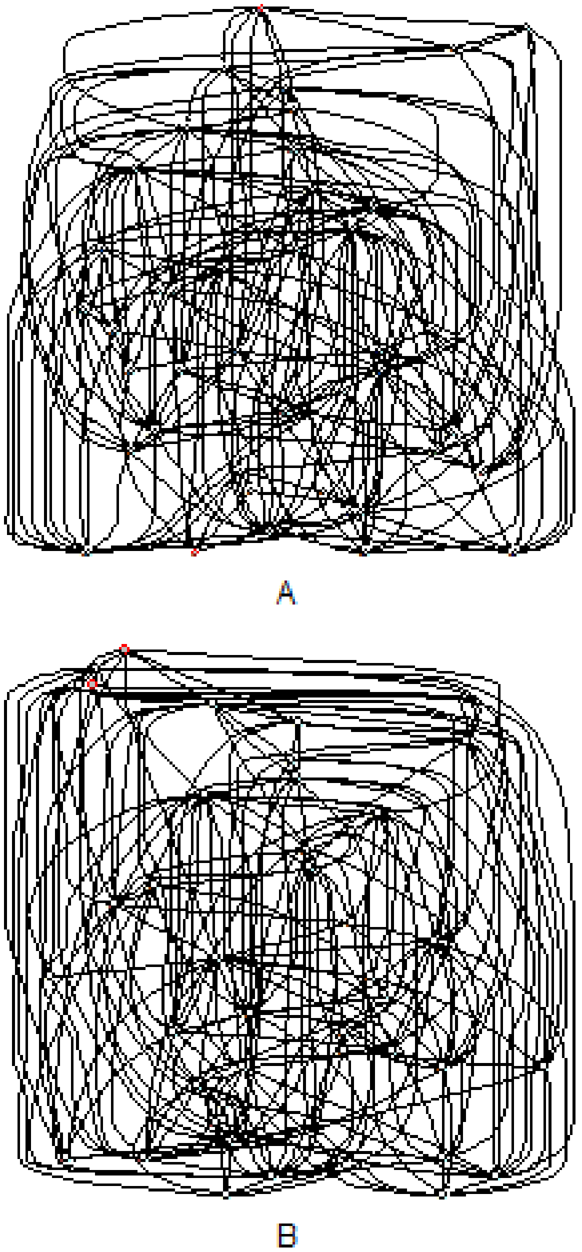

Figure 2. Explorary model for (A) Fast Food (B) Chinese Resturants.

Evaluation Models for Local Businesses

Becuase some attributes had to dependend on the other attributes, we set alpha = 0 for every cross validation process. Every evaluation model shows a variety of evaluation process for the fast foods and the Chinese resturants in Phoenix and in Charlotte. Because the area of every graphical models cover half of a page, readers are able to access the figures based on these links: (A)Phoenix, Fast Food, (B)(FastFood002.png), (C)Phoenix, Chinese resturants, (D)Charlotte, Chinese resturants. Table 2 lists the start and end nodes of each model.

Table 2. Start and End nodes of four final evaluation model

Case	Start	End
Phoenix: Fast Food	Attire	validated valet intimate classy touristy
Charlotte: Fast Food	Delivery	stars Noise.Level Wheelchair.Accessible valet intimate classy divey touristy trendy
Phoenix: Chinese	Outdoor.Seating	Drive.Thru touristy
Charlotte: Chinese	Number of neighborhoods Accepts.Credit.Cards	Delivery Takes.Reservations lunch dinner breakfast brunch intimate classy hipster touristy upscale

Each final evaluation model showes us that the yelp users in Phoenix and in Charlotte start their evaluations from a specific and local attribute. Then the yelp users make ther final recommendations based on the other attributes. In addition to the number of neighborhoods, no variables out of the attributes represent the starting point to evaluate a business in any case. The correlations of every variable and attributes were computed before and after training the final evaluation models. The results listed in Table 3 matched our purpose.

Table 3. Predictive and post correlations of variables and avergae attributes.

Category	City	Correlation	S	H	R	N	A
Fast Food	Phoenix	predict	0.296	0.256	0.35		0.803
		post	0.109	0.237	0.299		1
	Charlotte	predict	0.204	0.214	0.069	-0.004	0.779
		post	0.139	-0.03	0.031	0.033	1
Chinese Resturant	Phoenix	predict	0.182	-0.1	0.338		0.781
		post	-0.076	-0.036	0.011		1
	Charlotte	predict	0.059	0.166	0.353	-0.022	0.775
		post	0.012	0.039	0.215	-0.021	1

Discussion

About the evluation process in Yelp users' minds, we learned the following facts in this projecet:
1. Yelp stars is the final criterion for the businesses that have managed the chained stores across many cities, such as the fast food. For many local businesses, yelp stars is the intermediate factor to access the cities.

2. In every city, a category of businesses has the specific attributes being the starting point to be chosen as the potential shop. The local residents and the tourists could give the businesses the final recommendations attributes by attributes.

These facts will be benefitial to these readers:
1. **Customers:** People will know how to pick up a shop matched the personal flavor based on where they live or where they will visit.
2. **Business owners:** For who will open a shop in a city, they could understand the attributes that are mostly considered by the customers. For who is running a shop, they could review the advantages and disadvantages of the current management strategies.
3. **Yelp software engineers:** The enginners could create the apps for the local services. The apps connect Yelp's recommandation engine and filter the review articles and tips that mention the key words to evluate a category of businesses in a city. To make the apps dominate the least evaluation criterions, the engineers are able to adjust the variables and attributes at the server according to the last analysis of Bayesian network.

Acknowledge

Readers could check the full size graphical models in our rough report.