Proceedings of the 35th Chinese Control Conference
July 27-29, 2016, Chengdu, China

Reasonable Price Recommendation on Airbnb Using Multi-Scale Clustering

Yang Li[1], Quan Pan[1], Tao Yang[1 2], Lantian Guo[1]

1. School of Automation, Chinese Academy of Sciences, Northwestern Polytechnical University, Xin Shanxi 710072, P. R. China
E-mail: yangtao107@nwpu.edu.cn

2. State Key Lab for Manufacturing Systems Engineering, Xi'an Jiaotong University, Xin Shanxi 710049, P. R. China

Abstract: Reasonable house price prediction is a meaningful task, and the house clustering is an important process in the prediction. In this paper, we propose the method of Multi-Scale Affinity Propagation(MSAP) aggregating the house appropriately by the landmark and the facility. Then in each cluster, using Linear Regression model with Normal Noise(LRNN) predicts the reasonable price, which is verified by the increasing number of the renting reviews. Experiments show that the precision of the reasonable price prediction improved greatly via the method of MSAP.

Key Words: Multi-Scale Affinity Propagation(MSAP), cluster,Linear Regression model with Normal Noise(LRNN), price prediction

1 Introduction

Most of the people use the Internet to schedule their life, such as finding a point of interest, looking for a nice restaurant, renting a good hotel and even letting out their own houses. The web site like Airbnb is just one of those, which is an excellent platform that can give people the opportunity of sharing their occupied living space, and help the tenant find a good lodge anywhere. The transaction of this kind has formed the shared economy [8]. Lee et al. [6] studied that the feature of house price had close relation with sales, which always bother the host. So an attractive reasonable price is the key point for both the host and tenant.

Since the appearence of this shared economy form, there have been lots of researches about it, especially about the case of airbnb. Ikkala et al. [5] researched the host renting experience qualitatively and gave suggests on how to gain the reputation and trust from the guests. Benjamin et al. [3] studied that non-balck hosts in New York City charged approximately 12% more than black hosts for equivalent rental. Choi et al. [1] used the panel regression model investigating the impacts of Airbnb on the hotel revenue, and Lee et al. [6] analysed the features that were significantly associated with house sales, and so on. Seldom of those studies care about the reasonable price making, which will be presented in this paper.

Clustering could aggregate the house with same rental status together. In a city, the fact is the renting houses are always around the landmarks, so the clusters of the landmarks almost are the clusters of the renting house. Apart from the location, the house could also be clustered by their own characteristics, taking the facility as the example. if we aggregate the house using the geographic infomation first, and then the house rental status, which will help us study house renting price in a micro way. In this work, the proposed method of MSAP can cluster the house in different scale and make the prediction more specify.

A reasonable price is always attractive to the tenant, which will increase the house renting, so after aggregating the house, we need to analysis the reasonable price in the mirco way, and this will avoid the general advice in the city view which always mislead the host. So this paper uses the linear regression model with normal noise(LRNN) verifying the phenomenon that the greatly increasing of the renting rate of which price follows the LRNN predicted.

This work is structured as follows. Section 2 represents the multi-scale cluster by using the MSAP. Section 3 is the Linear regression model with normal noise, then followed the experiments result. Section 5 draw the conclusion about this work. Possible future work is outlined in section 6.

2 multi-scale clustering

The city landmarks which include the attractions, tour sights, museums, theatres and so on are the city symbols and always attract people from all of the world. The distance to the landmark and the popularity of the landmark usually are the latent factors to the house price. Here are the assumptions needed:

Assumption 1. The distance to the landmark is closer, and the average house renting price is higher.

Assumption 2. The popularity of the landmark takes positive effects on the house renting price.

Assumption 3. In a commiunity, the coverage of a kind of facility is reversely proportionaly to its contribution to the house value.

Each house can be regarded as a sample point n_h of the set H in R^2, and the landmark is the sample point n_m of another set M in R^2. The nearest landmark n_{mj} to the house n_{hj} could be defined as follow:

$$\forall n_{mk} \in M \quad |n_{hj} - n_{mj}| \leq |n_{hj} - n_{mk}| \tag{1}$$

The $|n_{hj} - n_{mj}|$ could use Euler distance instead.

House i can be represented by the facility embedding f_i, which is formed by the facilities that the house is equipped. We select nearly thirty facilities in each house, which include the swimming pool, kitchen, air conditioner and so on. According to Assumption 3, the corresponding number in the embedding will be in low place if the facility coverage is

This work is supported by National Natural Science Foundation (NNSF) of China under Grant 00000000.

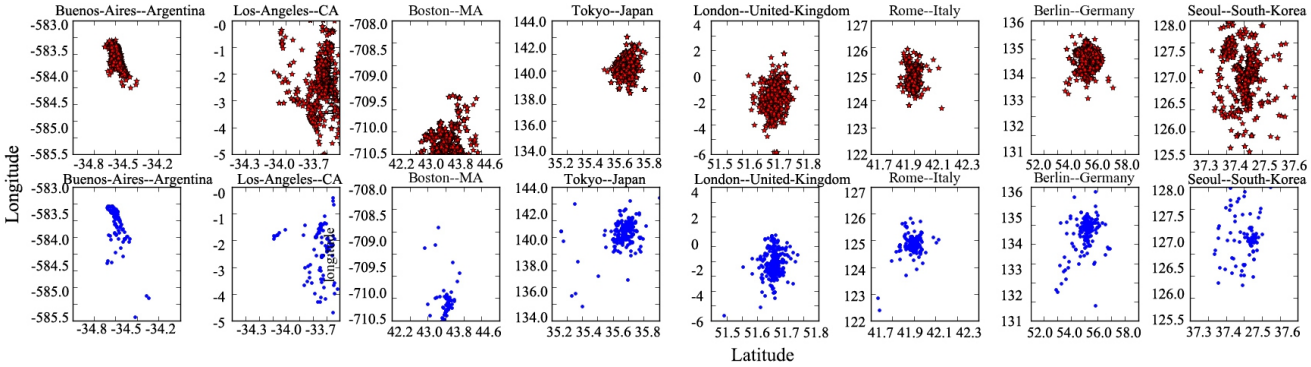

Fig. 1: Figures above are the scatter of the landmarks in different countries, and bottom ones are the scatters of the houses in the same country

high, and have $f_i = f_{i1} \cup f_{i2}$, where f_{i1} and f_{i2} are the subset of the facility embedding, then use the Huffman coding,

$$H_{ij} = H(f_{ij}) \quad j \in \{1, 2\} \tag{2}$$

and in this way, we reduce the dimension of the facility embedding from thirty to two.

From Fig. 1, we can know that the landmarks and houses almost have the same distribution in the eight most popular cities from different continents in the world[1]. So according to Assumption 1, we regard the landmark centers as the house centers when doing multi-scale clustering using MSAP.

Affinity Propagation(AP)[4], which was proposed by Dueck et al. in 2007 mainly to detect patterns, process sensory signals and so on. In this work, we improve this method and use it in multiple way distinguishing the houses in a finer way, which we call Multi-Scale Affinity Propagation(MSAP), this method is combined by two stages, in which the first stage is Landmark Clustering(LC).

When doing landmark clustering, the similarity $s_m(mi, mk) = -\|n_{mi} - n_{mk}\|^2$ indicates how well the landmark with index mk is suited to be the exemplar for landmark mi, According to Assumption 2, if the landmark is popular, it will have a large viscosity. so we add the viscosity coefficient $q_{mj} = \frac{r_{mj}}{\sum_{mj=1}^{M} r_{mj+1}}$ to the negative Euclidean distance:

$$s_m(mi, mk) = -\frac{\|n_{mi} - n_{mk}\|^2}{1 + e^{2q_{mk}}} \tag{3}$$

The responsibility $r_m(mi, mk)$ means likelihood of landmark mk attracting mi into its cluster, and the availability $a_m(mi, mk)$ measures the probability of the landmark mi choosing mk as the exemplar, here are the rules for the responsibility and availability:

$$r_m(mi, mk) = s_m(mi, mk) - max\{a(mi, mj) + s(mi, mj)\} \quad (j \in \{1, 2, .., k-1, k+1, ..., N\}) \tag{4}$$

$$a_m(mi, mk) = min\{0, r(mk, mk) + \sum_j \{max(0, r(mj, mk))\}\} \quad (j \in \{1, 2, .., k-1, k+1, ..., N\}) \tag{5}$$

[1]Buenos Aires/Los Angeles/Boston/Tokyo/London/Rome/Berlin/Seoul

After finishing the landmark clustering, shows in Fig. 2, the effects of distance and the landmark popularity on the houses that around or in the same circle may be similar, which indicates we need to do the house clustering in price-zone c, furthermore.

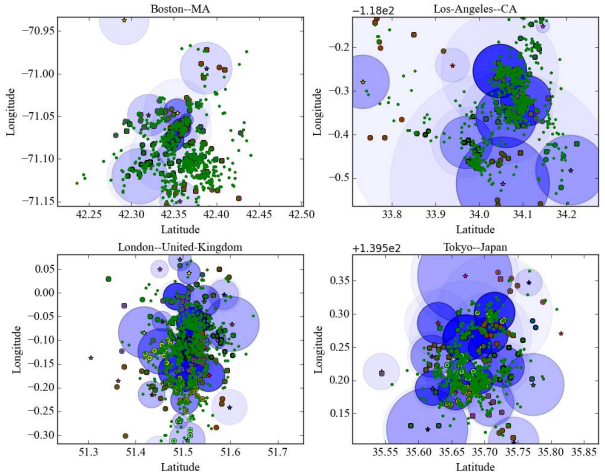

Fig. 2: The circle is the landmark influence area, and the green dot is the house

In the house clustering, the similarity s_h^c is measured by the house infrastructure f_i, so $s_h^c(hi, hk) = -\|f_{hi}^c - f_{hk}^c\|^2$, the responsibility r_h^c and availability a_h^c are same as the landmark clustering.

3 Linear Regression with normal noise

From described above, we use the factors of facility, distance to the nearest landmark, and nearest landmark popularity to indicate the house price which could be described by the multivariable linear regression model. But because of the uncertainty in reality, we need to combine the unsureness which could be simulated by the noise ΔP that will be explained in section 4 when doing the price prediction.

$$P = A \begin{bmatrix} f \\ d \\ q \end{bmatrix} + distribution(\Delta P) \tag{6}$$

In our model, we use the facility embedding f_i encoded by (2) representing house's facility status, d is the distance to

the nearest landmark which is calculated by the Euler Distance, and q is the viscosity coefficient of the landmark.

4 Experiments

4.1 Dataset

The data set was crawling from the Airbnb[2] which is a website for people to list, find, and rent lodging, the training set is the houses of which review marks are in [4.5, 5.0] for every aspects. We also crawled four famous cities' renting information about nearly four thousand items which include the location, price, facilities, reviews and so on. The basic information shows in Table I. Boston and London are familiar in house renting, especially in renting rate($\bar{R} = \sum_i^n (R_i)/n$, The average number of reviews, the bigger, the more famous) and facility variety($\bar{H} = \sum_i^n (H_{i1}H_{i2}/n)$, the bigger, the better equipped of the facilities), and with Los Angeles which is superior in facility equipment, the three cities almost are in the same level, while Tokyo has the most cost-efficient house price(The average house price$\bar{P}$ references from the site of Airbnb).

Table 1: Basic Information of the dataset

	Boston	Los Angeles	London	Tokyo
Average Review Number	165.30	175.55	161.93	155.61
Average Price	157	165	134	78
Facility variety	1391721	8675933	1430334	738492

4.2 Experiment Results

When doing the landmark clustering(LC), we aggregate the landmarks which was crawling from tripadvisor[3] into serval clusters C_P based on the negative Euler distance with viscosity coefficient. After dividing the city into several price-zones C_P which represented by the center landmark of the cluster and showed in circles on the figure, we need to subdivide the houses into the nearest one. Then use the same way in MSAP making the house that in C_P into several clusters based on the facility status.

We use the measure of Silhouette[2] which can be used for assessing how well the houses are clustered.

$$Sil_i = \frac{Bh_i - Ah_i}{max(Ah_i, Bh_i)} \tag{7}$$

Where Ah_i denote the average dissimilarity(distance) between house i and all other houses in the cluster to which i belongs. As to any other house cluster C_H in the dataset, $Dh(i, C_H)$ is the average dissimilarity(distance) of i to all houses of C_H and Bh_i is the smallest one of these.

From Tabel II, MSAP has positive effects on house clustering based on the geographic coordinates and facility variety, we need to aggregate the house which has the same facilities nearby, but it is hard to tune the weights of these two factors when using the methods of kmeans, DBSCAN and Appropagation Affinite, there are always such scenes that two houses have almost the same facilities but be far away from each other, or are in the neighborhood with different facilities totally, which lead to the low value of silhouette.

[2]www.airbnb.com

[3]www.tripadvisor.com

Table 2: Values of silhouette

Methods	Boston	Los Angeles	London	Tokyo
Kmeans	-0.073	-0.13	-0.059	0.070
DBSCAN	0.098	0.025	0.0015	-0.072
AP	-0.49	-0.66	-0.46	-0.76
MSAP	0.61	0.27	0.075	0.17

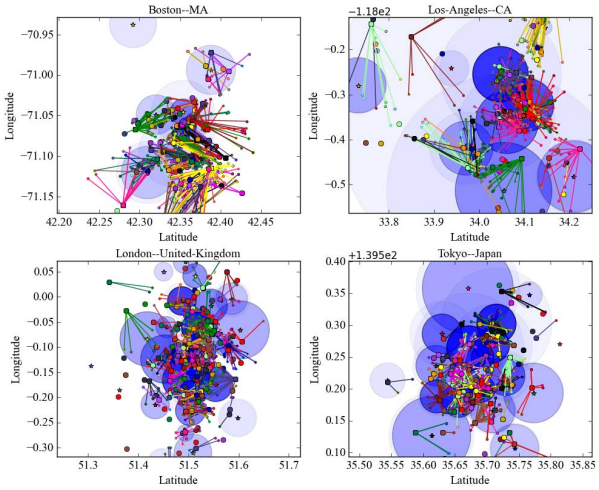

Fig. 3: The dots in different color are houses and the blue circles are the price zones. Each cluster is connected by the dots in same color. Because of using the facility difference as the cluster distance, so there may have cluster centers outer the price zone

Removing the geography information when doing the house clustering, we use the Huffman coding in (2) reducing the embedding dimension from thirty to two, which makes the house facility status more distinct. Fig. 4 shows the results of house clusters in all price-zone. From Fig. 3 and Fig. 4, we can find that the house density in Boston and London are bigger, while the house types in Los Angeles are more rich.

We analyze the price gap ΔP distribution between reasonable price and the mean price in each city, the result shows in Fig. 5. From the figure, we can see that price gaps follows the normal distribution, but has different shapes in different cities. Tokyo are sharper than other three cities, which means Boston, Los Angeles, and London have the variety price range, while the house price in Tokyo is more unifying. So in (6), we used the normal distribution as the distribution(price noise) of ΔP when doing the linear regression, which has great effects on the reasonable price recommendation. To analysis effectiveness, we use the review data that of 8th, Oct as the base line, and the 31th, Oct as the test data. The results are showing in Table III, the whole increase of the renting rates that show in the first row is low in each country, but the house prices following the LRNN model have a great improvement. So we can use LRNN method as the tool in reasonable price prediction.

Table 3: Renting rate improvment

Method	Boston	Los Angeles	London	Tokyo
ALL	13.27%	34.98%	40.85%	33.86%
LRNN	102.16%	327.51%	178.75%	288.69%

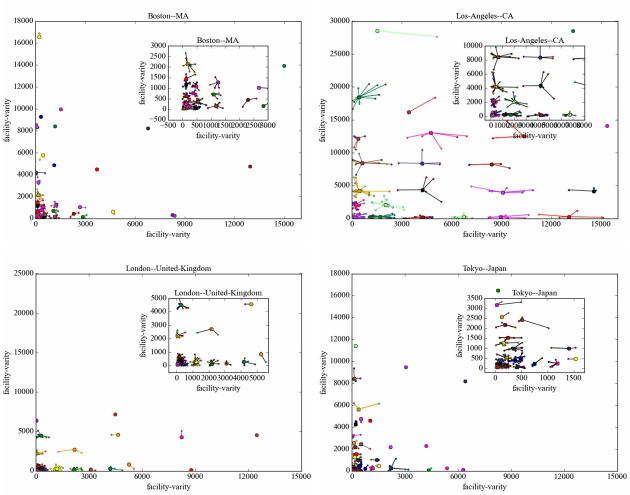

Fig. 4: The dots in different colors are houses. The clusters that in same color are connected by facility difference without geograpy information

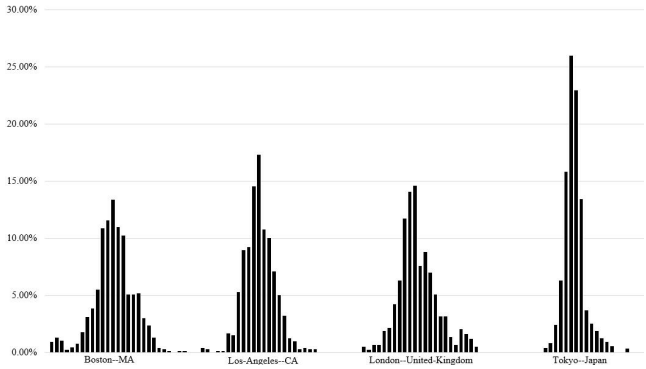

Fig. 5: Price gap distributions are different in each city, but all of them follow the normal distribution

Table 4: Precision of reasonable price Prediction in different level

Method	Boston	Los Angeles	London	Tokyo
WC	8.79%	10.28%	9.63%	18.73%
LC	11.80%	15.13%	19.96%	21.42%
MSAP	47.56%	19.59%	25.52%	55.98%

The final results are in table IV, from which we can know that after doing the multi-scale clustering, the precision of the LRNN prediction is improved continuously with the clustering stage(WC means without cluster, LC means only doing the Landmark clustering, MSAP means doing the multi-scale clustering). This phenomenon could be explained by the information entropy theory[7], the more precision of the information is, the lower uncertainty it has.

5 Conclusion

When doing the reasonable price recommendation, MSAP can cluster the house effectly and aggregate the house into different price-zone, which gives the reasonable price uniquely. The gap distribution between the house price and the city mean price reflects the diverse price in each city, which could help the method of LRNN provides an accessible way for the reasonble price prediction via the solid clustering.

6 Future Works

When doing the house price prediction, we only use the information about the distance from the landmarks, and the popularity of the nearest landmarks, so the semantic information could be used which will help improve the precision of the price recommendation. In this work, we only use simply method LR as the base model, which could be improved by some other methods, such as Gaussian Process, Ridge Regression and so on.

References

[1] K. H. Choi, J. H. Jung, S. Y. Ryu, S. Do Kim, and S. M. Yoon, "The relationship between airbnb and the hotel revenue: In the case of korea," *Indian Journal of Science and Technology*, vol. 8, no. 26, 2015.

[2] S. Dudoit and J. Fridlyand, "A prediction-based resampling method for estimating the number of clusters in a dataset." *Genome Biology*, vol. 3, no. 7, pp. 1–21, 2002.

[3] B. G. Edelman and M. Luca, "Digital discrimination: The case of airbnb.com," *Social Science Electronic Publishing*, 2014.

[4] B. J. Frey and D. Dueck, "Clustering by passing messages between data points," *Science*, vol. 315, pp. 972–976, 2007.

[5] T. Ikkala and A. Lampinen, "Defining the price of hospitality: Networked hospitality exchange via airbnb," in *Proceedings of the Companion Publication of the 17th ACM Conference on Computer Supported Cooperative Work; Social Computing*, ser. CSCW Companion '14, 2014, pp. 173–176.

[6] D. Lee, W. Hyun, J. Ryu, W. J. Lee, W. Rhee, and B. Suh, "An analysis of social features associated with room sales of airbnb," in *Proceedings of the 18th ACM Conference Companion on Computer Supported Cooperative Work & Social Computing*. ACM, 2015, pp. 219–222.

[7] Shannon and E. C., "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948.

[8] L. Zekanovic-Korona and J. Grzunov, "Evaluation of shared digital economy adoption: Case of airbnb," in *Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2014 37th International Convention on*, 2014, pp. 1574–1579.