Winter 12-8-2014

Segmenting Taipei's Real Estate Data – A Cluster Analysis

Sheng-Chi Chen

Chien-Hung Liu

SEGMENTING TAIPEI'S REAL ESTATE DATA – A CLUSTER ANALYSIS

Sheng-Chi Chen, National Chengchi University, Taiwan, 102356503@nccu.edu.tw
Chien-hung Liu, National Chengchi University, Taiwan, claude.liu@gmail.com

ABSTRACT

Data mining has been widely used for knowledge discovery from large amount of data. In this paper, clustering analysis is applied to Taiwan government open data platform (DATA.GOV.TW), segmenting the real estate data so as to understand the real estate market structure in Taiwan. This paper use design science research methodology (DSRM) as research method. The result provides valuable insights into market structures in Taipei City that has limited addressed in past research and contributes to real estate agencies and practitioners an insight-seeking approach that they can follow to generate values from data.

Keywords: Cluster analysis, data mining, design science research, real estate transaction.

INTRODUCTION

The advanced of information technology today allow enterprises to collect data that were impossible in the past and most enterprises today start to realize the importance of leveraging existing data on hands for competitive advantages. It holds true for real estate agencies in Taiwan. Taiwan's open data platform accelerates the adoption of these needs for data.

By nature, real estate is a type of product that is a long-lasting durable goods while also has a nature of investment. These characteristics make it very different from general merchandise. Traditionally, real estate industry values any information related to real estate products such as current status of buyer or seller (i.e. reason to buy or sell for a particular real estate), ownership and any transactional related information. Basically real estate agents make profits by broking sellers and buyers for real estate transaction. Equipping with critical information, real estate agents can act as a powerful intermediary between sellers and buyers and can expedite the process of transaction.

However, buyers or sellers of real estate do not hold the same information as their real estate agents do. The advantage for real estate agents are mainly due to information asymmetry. Thus, there are many disputes happened among buyers and sellers of house and real estate agents during the transaction process. This information asymmetry and intransparancy not only reduce trust but also increase the rising house prices. This conflict indeed reflects the importance of openness and transparency in the real estate transaction information requested by general public.

Taiwanese government expects to blasts real estate speculation through policy measures such as luxury tax, and real estate price disclosure. For example, to reduce information intransparency and asymmetry of real estate prices, Real Estate Value Laws was promulgated in 2011 and the general public is able to access timely and trustable information from governments now. The recent real estate policy reform in Taiwan along with the trend of open government data in the world have opened up many opportunities to researchers and industry practitioners for leveraging open government data to create more business value. This research aims to apply a data mining approach to analyze real estate market data of Taiwan, especially applying cluster analysis for better understanding of the real estate market structure in Taiwan. The findings from this paper can provide more insights about real estate market structures in Taipei City. This paper contributes to real estate agencies and practitioners an insight-seeking approach that they can follow to generate values from data.

LITERATURE REVIEW

Real estate transactions

Traditionally, real estate marketplace in Taiwan lacks of providing transparent, sufficient and real time information to interested stakeholders such buyer or seller of real estate. As a result, real estate agencies often manipulate market information and hold uneven advantage over customers (buyers or sellers). Inevitably, more and more disputes happen among real estate agencies and customers during and after transactions. As times go by, the calls from general public for real estate reform has intensified. Consequently, the Real Estate Broking Management Act was enacted by congress in 1999 to order to raise the professionalism of the real estate agency industry and safeguard consumer interest in Taiwan.

The ideal functions that real estate agents can provide to buyers and sellers are to create an efficient marketplace, reduce information search cost, and add moral hazard cost before and after transaction. Many scholars address about the relationship between real estate regulatory framework design and its market efficiency [8]. By nature, real estate products tend to be non-standardized and thus prices are determined on a case-by-case negotiation basis. In addition, the access to transaction data is often difficult, inaccurate or not real time. Today, majority of transactions in Taiwan are finished through real estate agents [1]. What bothers real estate buyers is that they neither could get accountable information from real estate agents nor could they acquired accurate information elsewhere. Information in transparency and asymmetry intensify the rising of real estate prices. Potential buyers of real estate suffer from high Misery Index.

Therefore, Taiwan government started to pay attentions to this information in transparency and asymmetry issues. For example, Taiwan's Legislative Yuan has passed the review of "Real Estate Value Law", which requires real estate buyers, real-estate agents and land administration agents register the actual transaction prices of properties within 30 days of deals being closed or face a fine. This Act was enhanced in August 1, 2012. Government and general public expect this Act would help real-estate transactions become transparent and a sound trading environment [10].

Data mining

Data mining has been grown quickly as an important issue in the application area of database. The objective of data mining is to discover knowledge hidden in a large scale of data. Data mining help analyzes large amount of data, through automatic or semi-automatic approach, builds effective models and rules [2]. Some scholar considers data mining as a process of searching and analyzing data to find the useful information hidden in the data [3]. Other scholars refer data mining to knowledge discovery from database, data warehouse, or other forms of large data storage. It extracts meaningful knowledge, including patterns, relationships or changes. From technical perspective, it refers to different forms and approaches of extracting information and knowledge from large volumes of data, which may include data visualization, machine learning and statistical techniques [4].

From business perspective, data mining is expected to extract potential, hidden and useful knowledge, pattern or trends from large volumes of daily transaction data. Today many governments start to promote open data policies in hopes that business and general public may use their innovations and capabilities to identify meaningful and valuable information. However, limited empirical researches are found in literatures about applying data mining techniques onto real estate transaction data.

Cluster analysis

Cluster analysis is a statistical classification technique used to reveals patterns, relationships, and structures in large volumes of data in which data are divided, based on similarity into different groups such that data in a cluster are homogeneous while heterogeneous between groups. Cluster analysis can identify classification rules in a seemingly messy data. Cluster analysis is useful for market structure analysis: identifying groups of similar products according to competitive measures of similarity [11].

The advantage of using cluster analysis is that users do not need to have fully understanding about target being analyzed, and in that sense is purely data driven. In other words, anonymous data set can be split into groups without users' understanding about data but purely rely on data. However, the disadvantage of this is that users cannot predict what kinds of clusters will be produced and consequently require users interpret the results by themselves [12].

Among the cluster techniques, K-means is a popular algorithm for cluster analysis in data mining, which was used by James MacQueenin 1967 [7]. K-means clustering starts with randomly selecting N observations as initial centroids for K clusters, which is also named "Center of Mass". Then K-means algorithm assigns each of N data points to its closest cluster centroids, new clusters will be computed and produced. And then this iterative process start over again and will not finish until K-means algorithm finds a N clusters with minimized variance and a maximized variance among N clusters. The advantage of K-means is quick and easy. However, K-means method is not appropriate when size of data set is too big or its density is too diverse. In addition, K-means clustering requires users to specify numbers of clusters to be developed. For example, when a user specifies N groups, then K-means algorithm would randomly select K data points as initial centroids, and this iterative process will not end until all N clusters reached to the conditions as mentioned earlier. In other words, K-means is an iterative process searching for center of mass for each cluster.

Ward's method, also called Ward's minimum variance method, is another popular algorithm for cluster analysis, especially applied in hierarchical cluster analysis, which was proposed by Joe H. Ward Jr. in 1963 [5]. Initially, Wards method treats every single data points as a clusters, and then at each step find the pair of clusters are merged according to the variance within clusters after merge (within cluster variance). The key difference between K-means algorithm and Ward's method is that the former requires users to specify K-value, the number of groups while Ward's method automatically specify number of groups based on the minimized objective function, which is the minimized total within-in variance.

RESEARCH METHOD

Peffer et al. (2008) developed the Design Science Research Methodology (DSRM), which constitutes a process model based on repeated examination, includes six steps: problem identification and motivation, defining the objectives of a solution, design and development of the artifact, demonstration of the artifact, evaluation of the artifact, and communication of the research.

This study makes use of the DSRM methodology developed by Peffer et al. to define the project objectives from a business perspective and then decide the data mining problem to be solved. Data mining processes include data sampling, feature selection, analysis and process, attribute change, model building and evaluation. Demonstration and Evaluation steps mainly validate the outputs from data miming. The accuracy can be measured through comparing the learned pattern in training set

with that of test set. Finally, the learned and validated knowledge will be applied in business as planned. The impact on business is collected and communicated, which completes the cycle of data mining.

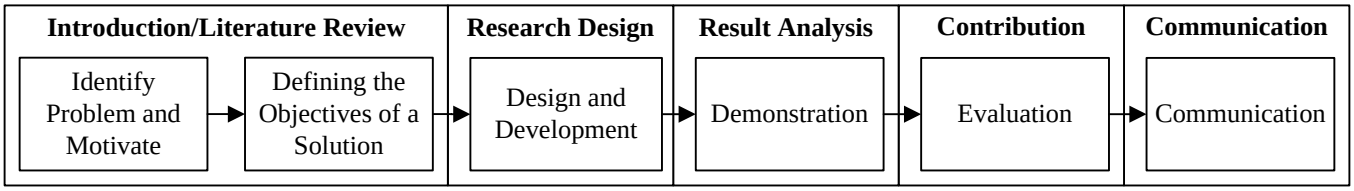

Figure 1. Research Process

Identification problem and motivate, and defining the objective of a solution, has described in the introduction section. Real estate transaction data have owned by real estate agencies and has not been opened to general public until August 1, 2012 requested by government. The objective of data mining is to identify hidden valuable information from large amount of transaction data. This research aims to understand real estate market structure in Taipei City. Cluster analysis is applied to identify groups of similar products according to competitive measures of similarity from government open data. The result will provide insights into market structures in Taipei City.

The data used for this research is acquired from Taiwan's government open data platform (DATA.GOV.TW) which contains more than 1,890 data sets with wide and diverse subjects of database such as transportation and distribution, real estate transaction, price and consumption, environment monitor data, and etc. Owing to the regional genetic characteristics of real estate transaction data, it is believed that the analysis of total cities will be hard to interpret. Thus, the research only uses the transaction data of Taipei City as the data mining target. There are 1,415 records available for data mining analysis.

RESEARCH DESIGN

Raw data is not mainly designed for analysis but for operations so it is not always easy to identify the relationship directly from raw data. Therefore, to make data mining a more insightful tool, this research takes more efforts on pre-processing step, including variable selection, data cleaning, data transformation, which will be explained in the following sections [6].

The dataset used includes 25 variables but not all of them are used in this research. To select the candidate variables for cluster analysis, this research first exclude the irrelevant and redundant variables. The candidate variables for analysis includes OBJECT OF TRADE, TOTAL FLOOR AREA TRANSFERRED (Square Meters), Type of Building, Completion Data of Building, TOTAL PRICE, UNIT PRIC of BUILDING (Per Square Meter). The candidate variable will be reduced based on the level of importance after each cluster analysis is conducted.

Data cleaning removes incomplete, inaccurate, irrelevant record, or irregular outliers from database and include only necessary data. In the data set, object of transaction contains land house, and car park. Land use zoning includes industrial land, farming land, cemetery and residential land. This research only keeps house as object of transaction and exclude land and car park in our analysis as they irrelevant to the objective of this research.

To make raw more readable and easier for analysis, the raw data is "cleaned" in the data cleaning step so that outliners are removed. Several variables are transformed to become new variables. For example, the date of building being completed is shown as "1020609" in data set. We first extract 102 from raw data to represent Taiwan Year and subtract it from current year to create "AGE of BUILDING". This year is 103 in Taiwan year so the age is "1" in this case. The unit of "Floor Area Transferred" is converted from "Square Meter" to "Ping", a unit of the size of buildings in Japan, Korea and Taiwan. One ping is equivalent to 3.3058 square meters. This transformation is more suitable for communications in Taiwan. An extended variable is thus created to serve this purpose. After the data preparation step, there are 346 records available for data mining analysis.

The aim of this research is to identify the grouping relationship of the real estate product and price and thus apply cluster analysis to actual price registered data of Taipei City. After data preparation step, the processed data set is different from the original ones. The first step of cluster analysis is to determine the objective and variables to be included in the analysis. This research follows a two-stage cluster analysis. First, Ward's method is used to determine the number of clusters. Second, Ward, Average and Centroid are three algorithms used to build the model for cluster analysis. In addition, based on the number of cluster suggested by Ward's method. This research also takes a further step to fine-tune the number of clusters. In addition to the use of two-stage approach, this research bases on importance level variables to select variables. Many candidate models are created, compared and consequently final candidate model that is meaningful and interpretable is identified and chosen as final model. This searching process is iterative and often time consuming.

RESULT ANALYSIS

This research applies cluster analysis techniques in analyzing pre-processed data of Taipei City. The variables included in the first data mining model includes object of transaction, FLOOR AREAs transferred, BUILDING TYPE, AGE OF BUILDING, TOTAL PRICE, UNIT PRICE (unit: ping). At each model building process, variables are reduced to develop new models. After many modeling, the final model show better meaningful results. Three variables are included in this final model: AGE of BUILDING, FLOOR AREA, and TOTAL PRICE. All of them have high level of importance, reflecting the these are most important factors considered by buyers or sellers of the house. The results of cluster analysis for Taipei City is shown as follows.

The characteristics of cluster analysis of Taipei City are shown in Table 1. Age of Building, Floor Area (unit: Ping) and Total Price came out as the most important criteria. The final model results of cluster analysis for Taipei City suggest 5 clusters. Cluster 4 contains more records than that of the rest, which has a total of 154 records, weighing around 45% of total sample size. It shows lower price with middle range of housing space is popular despite the house is aged. Cluster 1 contains least records- only 5 records are included, equal to 1% of total sample size. The total price of aged young and area small is close to that aged old and area medium, referred to Cluster 3 and 4, which means it could be the same cluster in the process of cluster analysis.

Table 1. Cluster analysis of Taipei City

Segment ID	Age of Building	Floor Area (ping)	Total Price
1	34.60	79.16	57,953,240
2	15.85	52.71	41,226,875
5	31.09	37.87	24,721,804
4	37.41	25.13	11,691,419
3	12.30	15.72	10,596,030

CONTRIBUTION

It is important to illustrate what data mining discovers from data in a way that everyone can possibly understand, and efficiently interpret implications behind data and help end users to make better judgment and timely decisions. In the evaluation stage of Data mining, data miners should prove (1) results from common sense; (2) possible results and make efforts to interpret; (3) results that are not clear and hard to evaluate or leave to end users for further explorations.

In the process of analysis, this study attempts to include different variables to explore the possible cluster structure existing within transaction data, identify cluster relationship and finally interpret findings. Cluster analysis is an unsupervised learning algorithm that requires careful variable selections and multiple trials in order to identify a cluster relationship that makes sense.

In the experiment design, this study consolidates the real estate open data of Taipei City, filters out seven variables as candidates for cluster analysis in order to run cluster analysis for Taipei City. Meanwhile, in variable selection process, each model adjusts its input variables for cluster analysis according to level of importance derived from data.

We believe this study has demonstrated the good use of the real estate open data provided by Taiwanese Government in the area of data mining, especially the cluster analysis, and provided the market structure results that are insightful.
The contribution this study makes to practice is threefold as following:

- Offered a holistic view of market structure insights to house buyers, sellers and real estate agents, breaking the barriers that real estate agencies hold up information on their side.
- Provided industry with a business direction and data mining approach to leverage its own data. Real estate companies can compare their cluster analysis results with that of this study so that they can quickly understand how they perform, identify sales gaps, and discover new opportunities at different segments within market structure.
- Set a new model for utilizing government open platform for data mining the real estate open data and provided insights into how to leverage big data opportunity created by government for business value.

CONCLUSION

Data mining is a rapid growing area that produces many research reports, new systems or prototyping development. For example, many new applications based on early data mining researches and a wide array of algorithms have gradually unlock the value of data residing in database. The diversification of data mining approaches, including machine learning and statistical researches, multiply the efficacy of advanced knowledge discovery. Cluster analysis is widely used across different fields; however, there is no universal agreed criteria justifying the results produced by cluster analysis. Thus, selecting a proper criteria to evaluate cluster analysis is critical. However, using cluster analysis for market structure analysis in Taiwan is received limited attentions in real estate research.

This research acquires the real estate actual price registered data from the government open data platform, applying cluster analysis to explore the data of Taipei City and includes experience from real estate agency into consideration. The results show

that AGE OF BUILDING, FLOOR AREA (unit: Ping) and TOTAL PRICE reflect different nature and meanings in different geographic areas. For example, there is a unique cluster identified in this research, which have average small floor area with very high unit price and are located in the prime area within Taipei City.

Including opinions from experienced professional from real estate industry helps define the evaluation criteria as well as result interpretation. Future researchers may consider include more market information to better adjust evaluation criteria so that the research results will be more value added. In addition, it is suggested to use different data mining techniques to analyze real estate transaction data (total market data), which will enhance efficacy of government policy, create more values for businesses as well as general public, and eventually reach a win-win situations for all stakeholders.

While much has been learned in the present research, much remain to be done. Currently, Taiwan is ranked 16th most densely populated country in the world. The population of Taipei city represents 11% of total population and contributes more than 27% of total value of transactions, which shows the importance of capital. Together with the New Taipei City, they make up the Taipei Metropolitan Area, which has a total population of 6.53 million, 28% of total population or represents 43% of total value of real estate transacted. Therefore, it is worth further including New Taipei City into future research. Furthermore, while this paper focus on using cluster analysis, a unsupervised classification approach, to group a set of real estate transactions (objects) and find whether there is some relationship between these transactions (objects), it is would be interesting for further researcher to apply supervised classification methods to open real estate data (new objects) to predict which group a new transactions (new object) belongs to.

REFERENCES

[1] Benjamin, J.D., Jud, G.D., & Sirmans, G.S. (2000) 'What do we know about real estate brokerage?' *Journal of Real Estate Research*, Vol. 20, No. 1, pp. 5-30.

[2] Berry, M.J.A. & Linoff, G. (1997) *Data Mining: For Marketing, Sales, and Customer Support*, Wiley Computer Publishing.

[3] Fayyad, U., Piatetsky-shapiro, G., & Smyth, P. (1996) 'The KDD Process for extracting useful knowledge from volumes of data', *Communication of the ACM*, Vol. 39, No. 11, pp. 27-34.

[4] Han, J. & Kamber, M. (2001) *Data Mining: Concepts and Techniques*, Morgan Kaufmann Publishers.

[5] Jain, A.K.(2010) 'Data clustering: 50 years beyond k-means', *Pattern Recognition Letters*, Vol. 31, No. 8, pp. 651-666.

[6] Kotsiantis, S., Kanellopoulos, D., & Pintelas, P. (2006) 'Data preprocessing for supervised leaning', *International Journal of Computer Science*, Vol. 1, No. 2, pp. 111-117.

[7] MacQueen, J.B. (1967) 'Some methods for classification and analysis of multivariate observations', *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, University of California Press, Vol. 1, pp. 281-297.

[8] Miceli, T.J. (1988) 'Information costs and the organization of the real estate brokerage industry in the u.s. and great britain', *AREUEA Journal*, Vol. 16, No. 2, pp. 173-188.

[9] Peffers K., Tuunanen T., Rothenberger M. A., & Chatterjee S. (2007) 'A design science research methodology for information systems research', *Journal of Management Information Systems*, Vol. 24, No. 3, pp. 45-77.

[10] Real Estate Broking Management Act, Laow and Regulations Retrieving System, *Ministery of Interior, Taiwan*, available at http://glrs.moi.gov.tw/EngLawContent.aspx?Type=E&id=210&KeyWord=%E4%B8%8D%E5%8B%95%E7%94%A2. (accessed 20 September 2014).

[11] Shmueli, G., Patel, N. R., & Bruce, P. (2007) *Data Mining for Business Intelligence*, John Wiley & Sons Inc.

[12] Ward, H.J. Jr. (1963) 'Hierarchical grouping to optimize an objective function', *Journal of the American Statistical Association*, Vol. 58, No. 301, pp. 236-244.