THE APPLICATION OF DATA MINING TECHNOLOGY IN REAL ESTATE MARKET PREDICTION

Xian Guang LI, Qi Ming LI
Department of Construction and Real Estate, South East Univ,,Nanjing, China.

Abstract: This paper introduces the application of data mining technology in real estate market and develops an application flow both in theory and practical example. Firstly, The definition， application methods were introduced and using flow of data mining in real estate was analyzed. Then Nanjing real estate market was taken as an example to illustrate how to use neural network------one of the data mining technologies to analyze and forecast real estate market. Finally, some problems about data mining technology was bring forward and prospect was discussed. The results show that data mining produces high prediction accuracy in real estate data analyses and market prediction.
Keywords: data mining; real estate market; prediction

1 Introduction

1.1 Problems in real estate data analyses

With the rapid development of Real estate industry, it has become the important industry in China which has great influence on national economy. Many problems also emerged during the process, such as exorbitance investment in real estate, the rapid rise of house price, the high vacancy ratio on, which require government and relative department to institute reasonable industry policies to channel off the industry development. And the investors and developers always rack their brains to bear the palm and get profit in the vehement competitive environment, which normally work inefficient since they are confused by such lots of data. The house consumers usually can't get any help from the real estate data as they can't find a proper tool to analyze it.

All the problems to be solved depend on the correct analyses of real estate data. Fortunately, there are lots of statistic data in the industry, such as total dimension of investment, floor space started and finished in every year, space of house pre-selling by region, the average house price in different region and so on. The important thing after we get the data is to find a efficient data analyzing tool to transform the data into knowledge and information, otherwise it will just be a "data tomb"（Zhang ,2006). Although there are already some data analyzing tools and soft, but they usually are efficient for decision in real estate industry since people know little about the development mechanism of it.

1.2 Data mining technology and its application in real estate industry

The ever increasing quantity of data in every competing environment such as real estate industry present both an opportunity to extract useful information and a challenge to process the massive volume of data effectively. During the past few decades, the pattern recognition and machine-learning communities have greatly expanded their areas of application and the kind information to be extracted. The database community joined the endeavor in early 1990 and a new multi-disciplinary field began, which we now call data mining. It is the process of extracting valid, previously unknown, comprehensible and actionable information from large database and using it to make decisions, the combination of modern artificial intelligence and statistics（David and Heikki,2003) The methods of data mining include neural network, heredity arithmetic, decision-tree arithmetic, rough set arithmetic, fuzzy sets theory and so on. The main task of different data mining activity can be divided into four kinds: relevancy analysis, clustering analysis, time-sequence model and prediction, deviation analysis. Data mining can help us understand the running mechanism of the object, discover the future trend and make relative prediction. All the mined information will be very useful for our decision.

The application of data mining technology became more widely in real estate industry. It can help

governing bodies and enterprises to get some useful information from the collected data. For example, government can understand the status and development of the industry and institute some reasonable policies, and real estate developers can find commercial chance and customers to plan for some project as well as carry into execution. The typical uses of data mining technology in real estate industry are showed in table 1:

Table 1 The application types of data mining technology in real estate market prediction

	Application types	Application purpose
The application of data mining technology in real estate market prediction	Real estate period	Research on the real estate fluctuation period, find out its rules and characteristic, and ascertain the influence factors and relationship between them.
	Market trend	Analyze the relationship between market requirement and GDP, individual governable income，space of land development，total investment for real estate development and so on，get the requirement model through statistic regression and neural network simulation.
	Property supplying types	Integrate the total residents in special region and its distribution, land using status, government layout and traffic infrastructures to understand different properties requirement through clustering and hierarchy analysis.
	Client management	Analyze the multidimensional relevancy and time-sequence model of customer, find their consuming habits and help salesman to find some way to affect their customers.

2 The features of application of data mining

Before we analyze the application of data mining for real estate market, we'd better understand the features of it, especially its features compared with traditional statistic method. Both statistics and data mining are concerned with drawing inferences from data. The aim of the inference may be understood as the patterns of correlation among data values and to predict the future data values. Classical statistics developed an approach that involves specifying a model for the probability of the data and make inferences in the form of probability statement. Data mining methods have developed in many cases been developed for problems in real estate market that that don't easily fit into the framework of classical statistics（Jonathan Hoskin.1997). We state main features of application of data mining in real estate industry as follows:

Complex models: real estate market involves complex interactions between feature variables, such as national economy, individual income and population, and with no simple relationships being apparent in the data. It is difficult to formulate a comprehensible statistic model for different city or different country. Data mining techniques such as neural networks and rule-based classifiers have the capacity to model complex relationship and should have better precision in analyzing problems.

Problems with large database: data mining often deals with problems with large data sets. This is also the consequence of complexity of problems with large amounts of data, which is needed to deduce right inferences. For problems like analyzing real estate market, the data is very abundant and large for data mining.

Controlled prediction error: prediction error is often estimated by cross-validation, a technique known to statistics but used much more widely in data mining. Data mining methods often seek to minimize the loss function expressed in terms of prediction error, including the choice and elimination of variables and their values in the model.

3 The prediction flow with data mining technology for real estate market

The data mining is an interactive and iterative process involving numerous steps with many decisions being made by the user, beginning with the understanding and definition of problem, ending with the results analysis and knowledge application to make predictions and decisions（David and Heikki,2003). As data mining technology is good at solving complex system with a great deal of data while there is something previously unknown. The procedures of application of data mining technology are as flows: (1) problem definition: describing the problem and ascertain the goal of data mining. (2) Data preparation --- distilling useful data group; pretreatment of data---check the integrity and coherence of data, ruling out the noise and useless data. (3) Data mining: choose appropriate method of data mining technology to analyze the data group according

the characteristic of problem. (4) Result analysis: explaining and evaluating the data mining results which enable people to understand it. (5) Knowledge application: transforming the reasonable results into knowledge to make prediction of object.

To predict the real estate market, we should combine the characteristic of real estate industry with the knowledge founding function of data mining technology although the basic procedures are as same as what are mentioned above. We should collect original data about investment, construction and sale information as much as possible since data mining technology depends on a great deal of data. A database warehouse also should be established to support the mining process. When the result is worked out, it should be tested and explained with professional knowledge since it possibly reflects merely the relationship between the original data as well as the real running mechanism of real estate. The application flow of data mining technology in real estate market prediction is showed in following picture 1:

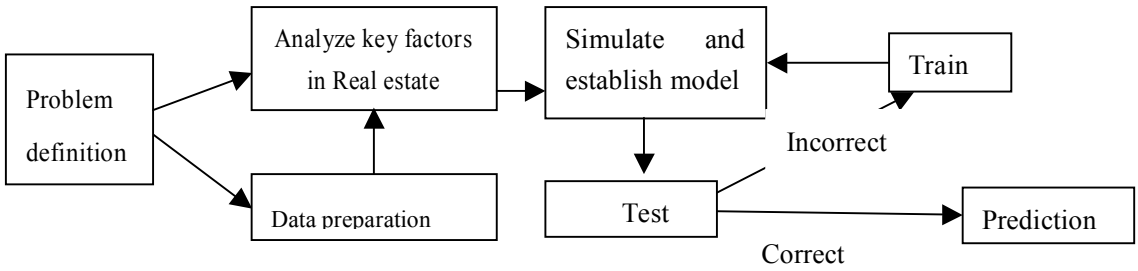


Figure.1 The application flow of data mining technology in real estate industry

4 The application of neural network method based data mining technology in prediction the real estate market in Nanjing city

4.1. Ascertain the influence factors and its data collection

The artificial network normally is adopted in research on the real estate market period and prediction of the market. The first procedure is to analyze the influential factors and collect the data. Although real estate market is a complex nonlinear system; many researches showed that there were a few key factors which can reflect the most change in the market（LI and LI,2006） According to the former research and the writer's analyses, eight factors were chosen to study the real estate market change in Nanjing: the change of national economy; the national income; individual housing area; credit scale; investment in real estate development; area under construction annually; population and resident savings by the end of the year. The change of national economy was shown by the gross domestic products and the national income by the annual individual income. The data of real estate market in Nanjing city from 1999—2005 was adopted in this study which were shown in table 2, All the data were from Jiangsu province statistic information website (http://www.jssb.gov.cn/).

4.2 Neural network structure design and data pretreatment

To design the neural network structure, firstly the input indexes as we mentioned above were analyzed, which were used to train the neural network. According to our practical need to predict the real estate market, the following three indexes were chosen to reflect the running status of different times: Completed area, sold area and house price (average house price). As there are eight input indexes and three output indexes which were required for prediction, two layers network with three nerve cells in every layer was adopted after many testing simulations.

As neural network model of MATLAB was adopted as a tool in this study, normally the original data couldn't meet the requirement of data mining process and they should be standardized. Considering the housing policy reform in 1998, the data of real estate market in 1999 was taken as the basic data and all the

other data in different years were divided by relative data in 1999. The pretreated data used in this study were shown in table 3:

Table 2 The original data for real estate market prediction based on artificial neural network

Year	Input index								Output index		
	GDP (Billions)	Individual income (Yuan)	Population (Millions)	Investment (Billions)	Credit savings (Billions)	Individual housing area (m^2)	Resident savings (Billions)	Construction area ($10^4 m^2$)	Completed area ($10^4 m^2$)	Sold area ($10^4 m^2$)	House price (Y/m^2)
1999	89.942	7694	5.3744	9.791	148.203	9.68	56.805	9.9238	385.10	174	2780
2000	102.13	8233	5.4489	9.934	196.344	10.1	59.67	9.7067	383.11	222.23	2910
2001	115.03	8848	5.5304	11.1	229.302	20	71.613	10.5288	402.76	281.6	3093
2002	129.757	9157	5.6328	13.763	279.475	21.4	103.101	11.844	434.59	382.05	3405
2003	157.62	10195	5.7223	18.380	362.255	21.1	125.385	15.569	392.72	444.42	4255
2004	191	11601	5.84	29.288	401.223	24.5	123.484	23.1052	644.45	607.96	4284
2005	241.3	14997	5.96	29.61	466.221	25	138.068	27.0196	627.05	795.5	4432

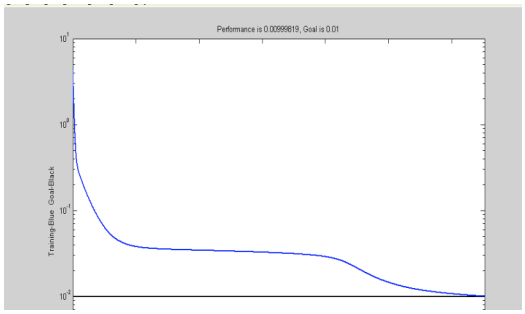
Table 3 Artificial neural network based standardization data for real estate market prediction

Year	Input index								Output index		
	GDP	Individual income	Population	Investment	Credit savings	Individual housing area	Resident savings	Construction area	Completed area	Sold area	House price
1999	1	1	1	1	1	1	1	1	1	1	1
2000	1.136	1.07	1.014	1.015	1.325	1.043	1.050	0.978	0.995	1.277	1.047
2001	1.279	1.15	1.029	1.134	1.547	2.066	1.261	1.061	1.046	1.618	1.063
2002	1.443	1.19	1.048	1.406	1.886	2.211	1.815	1.193	1.129	2.196	1.225
2003	1.752	1.325	1.065	1.877	2.444	2.180	2.207	1.569	1.020	2.554	1.531
2004	2.124	1.508	1.087	2.991	2.707	2.531	2.174	2.328	1.673	3.494	1.541
2005	2.683	1.949	1.109	3.024	3.146	2.551	2.431	2.723	1.628	4.572	1.594

4.3 Establishment, training and testing of net work

As all the data and neural network structure were prepared, the next important procedure was to choose a proper tool to form the real network which can simulate the real estate market. In this study MATLAB was chosen as a tool and input index was defined as p as well as output index as t. The following program was input as MATLAB command to establish and train the neural network:

```
net=newff(minmax(p),[3,3],{'tansig','purelin'},'traingdm');
net.trainparam.epochs=5000;
lp.lr=0.08;lp.mc=0.8;
```



net.trainparam.goal=1e-2;
net=train(net,p,t);

The training error was shown in picture 2，the result indicate that the simulation error reached 0.0099,which is less than the setting error 0.01 and met the requirement.

The trained neural network structure was shown in picture 3: variable p represented the inputs during network training, which were the eight indexes in real estate in this study. Weightiness matrixes of different layers were expressed by w1 and w2, Variable a1 was the output of latent layer and a2 was the final output. Layer deflections were shown by b1 and b2, TANSIG function was adopted in the first layer and PURELIN function for the second one.

The neural network has been tested before prediction to improve the precision of data mining. The MATLAB command for test was: Y=sim(net,p), Variable net was the trained network , p represented the data of random year and Y the output.

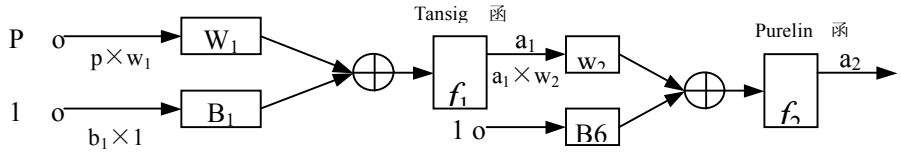


Figure 3 Artificial neural network structure

4.4 Prediction of real estate market in Nanjing based on neural network

As precision of the neural network has been proved, the last but most important procedure is to use it to make prediction. The MATLAB command was: p*= sim(net,p), p* represented the prediction result of the predicting year and p was the input indexes of predicting year, which were the completed and planning input data in real estate industry in Nanjing,2006. According to the communiqué of Nanjing statistic bureau, the investment in real estate industry from January to May was 12.57 billions which increased 10.4% over the same period in last year and total investment in 2006 was estimated about 30.50 billions. Other input indexes also were calculated in this way based on the completed and planning data and the final input standardization indexes in 2006 were as following: Input2006=[3.059 2.186 1.135 3.338 3.052 2.857 2.959 3.246]Input the above data and neural network made prediction after its data mining: Output2006=[1.7094 4.7367 1.5939]This was transformed into the normal data we understood as following:Output2006=[completed area：658.29×10^4㎡ sold area：759.86×10^4㎡ house price：4431Y/㎡].

The result showed that real estate market would bloom both in demand and supply since the GDP of Nanjing city, annual individual income and residential savings kept rising in 2006. Completed area would reach 658.29×10^4m2 and sold area would reach 759.86×10^4m2, house price kept 4431yuan/㎡, which meant the demand slightly exceeded the supply while house price would not rise sharply. The phenomenon should attribute to the national hosing policy change and the vacancy house which haven't been sold during 1999-2002. Policies to advance the real estate development threshold and individual initial payment for buying house instituted by Chinese government in May 2006 leaded the whole market to prosper and stable one without sharp change in house price. The vacancy houses in the market which haven't been sold during former years also contributed to leveling off the house price.

5 Problems and discussions

Real estate industry is a complex nonlinear system with many affective factors and different data. There are many running mechanisms and relationships which people haven't understood up to the present (Sung, 1998). It's not enough and to analyze real estate market with traditional econometric methods and system theory. It's very necessary to adopt some methods and tools like data mining technology with "knowledge discovery" functions. The application of data mining technology in real estate industry is a front field full of challenge although many new data mining methods and models were worked out. There are still some problems puzzled

us, such as the establishment of data warehouse in real estate industry, data mining efficiency especially with large-scale data, the data mining methods appropriate for real estate industry and so on. With more in-depth study on the application of data mining in real estate industry, more useful information and knowledge will be mined out, which is very helpful for government to institute industry policies as well as developers to make right decisions.

References:

David Hand, and Heikki Mannila. (2003). *Principle of data mining*. New York: Prentice hall, 5-17.

Jonathan R.M. Hoskin.(1997). *A statistical perspective on data mining*. Future generation computer systems,13(5),117-134.

LI Xijuan, and LI Bin. (2006).The empirical analysis on real estate and national economic growth, Commercial research, 336(4), 201-206.

Sung Ho Ha, and Sang Chan Park. (1998).Application of data mining tools to hotel data mart on the intranet for database marketing. Expert system with application, 31(1), 3-31

Zhang Yan.(2006). *Macro-demand prediction for real estate market* Journal of Chongqing Three Gorge Institute,22(3), 61-63.