Adaptive Weighted Finite Mixture Model: Identifying the Feature-Influence of Real Estate

XIN XU, Northeast Normal University

YANJIE FU, University of Central Florida

JINGYI WU, YUQI WANG, ZEYU HUANG, ZHIGUO FU, and MINGHAO YIN,
Northeast Normal University

It is significant for real estate investors to understand how the construction environments and building characteristics impact the housing unit price. However, it is challenging for identifying the complex feature-influence from construction environments and building characteristics. It is also hard to alleviate the heterogeneity of real estate. In this article, we propose a framework named Adaptive Weighted Finite Mixture Model to identify the feature-influence and simultaneously alleviate the ill effect of heterogeneity. Applying this framework, we can predict the housing unit price based on the corresponding features. Besides, we discover that the feature-influence exists in the dissimilarity among similar cities. Specifically, we adaptively learn the weights of features to identify the feature-influence, and we model the estimation of the housing unit price with the feature-influence into a finite mixture model. We utilize the Principle Component Analysis algorithm to obtain a low-dimensional representation of housing features. The low-dimensional representation reduces the computational cost of model learning, and it avoids a potential catastrophe of the singular matrix inversion during the process of learning model parameters, which are estimated by the Expectation Maximization algorithm. To avoid the blind search for the latent group number used in the proposed framework, we employ the pre-clustering result as a guide of the searching range of the latent group numbers. We conduct numerous experiments on three real datasets from Shenyang, Changchun, and Harbin, which are the three provincial capital cities that have similar geography, economics, and cultures. The experimental results demonstrate the effectiveness of the proposed framework.

CCS Concepts: • **Information systems** → **Data mining**;

Additional Key Words and Phrases: Real estate, finite mixture model, expectation maximization algorithm, identify feature influence

ACM Reference format:

Xin Xu, Yanjie Fu, Jingyi Wu, Yuqi Wang, Zeyu Huang, Zhiguo Fu, and Minghao Yin. 2020. Adaptive Weighted Finite Mixture Model: Identifying the Feature-Influence of Real Estate. *ACM/IMS Trans. Data Sci.* 1, 3, Article 20 (September 2020), 16 pages.

<https://doi.org/10.1145/3379560>

Authors' addresses: X. Xu, J. Wu, Y. Wang, Z. Huang, Z. Fu, and M. Yin (corresponding author), Northeast Normal University, 2555 Jingyue Street, Changchun city, Jilin Province, PRC, 130117; emails: {xux894, wujy050, wangyq600, huangzy149, fuzg432, ymh}@nenu.edu.cn; Y. Fu (corresponding author), University of Central Florida, 4000 Central Florida Blvd. Orlando, Florida, 32816; email: yanjiefoo@gmail.com.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2020 Association for Computing Machinery.

2577-3224/2020/09-ART20 \$15.00

<https://doi.org/10.1145/3379560>

1 INTRODUCTION

It is important for investors to understand how construction environments and building characteristics impact the housing unit price because of the rapid development of society and the acceleration of urbanization where the real estate economy plays a crucial role [7]. As a long-term illiquid finance [32], investors are willing to pay a high price for a house with excellent community benefits like a first-class location, various property services (like malls, schools, hospitals, bus stations), and a green rate. Besides, the housing characteristics (like the number of rooms, the number of halls, the number of bathrooms, the orientation, the number of floors, the floor type) bring investors a preliminary estimation of the housing unit price. If real estate investors hope to invest in a high potential increasing-value house, they need to understand the feature-influence from construction environments and housing characteristics.

However, the problem of predicting the housing unit price is still far from being perfectly solved due to the following challenges:

- **Identifying Feature-Influence.** The contribution of each feature is considered as equal in most existing works [16, 17, 19, 22]. In practice, various features should have different contributions to the housing unit price. For example, in China, the investors in the southern cities evaluate the green rate more than those in the northern cities. The reason is that the northern cities have fewer green plants in winter, over half the year. Although quite a few works have been proposed, those works consider the features' ranking, which is calculated by their information gain ratio [26] as the feature-influence, which is fixed and not adaptive [33, 35]. Therefore, we need a scheme to adaptively learn the feature-influence.
- **Heterogeneity.** Most of the existing works generally simplify the predicting problem using a hypothesis that all buildings are considered as a group with similar properties. However, the heterogeneity of real estate leads models to incorrectly predict the housing unit price. Buildings, in various sub-markets of real estate, indeed have different properties. Hence, we necessitate alleviating the ill effect of heterogeneity.

To address the above challenges, we intend an Adaptive Weighted Finite Mixture Model (AWFMM) framework, which can more accurately estimate the housing unit price through adaptively learning the feature-influence and addressing the ill effect of heterogeneity. Specifically, we first transfer the original representation of housing features into a low-dimensional representation by the Principal Component Analysis algorithm [1]. The low-dimensional representation reduces the computational cost of model learning, avoids the potential catastrophe of singular matrix inversion in the process of the estimation of model parameters, and alleviates the coupling among features as well. Second, we utilize two clustering algorithms (KMeans [2] and MiniBatchKMeans [28]) to determine the clustering results on the two widely used metrics, which are the Silhouette Coefficient (SC) [27] and the Calinski-Harabaz (CH) [5]. The clustering results can be viewed as a guide to obtain the range of latent group numbers. Using this range, we can avoid a blind search in the process of model learning. Finally, we propose the AWFMM, which can identify the feature-influence to adaptively discover each feature's importance to the housing unit price. Applying this model, we can more accurately predict the housing unit prices based on the features of buildings.

We conduct numerous experiments on the real estate datasets from three provincial capital cities, and the experimental results demonstrate that the proposed framework outperforms several baselines. Besides, we study three cases: (1) We discover the dissimilarity of feature-influence even in the cities with similar geography, economies, and cultures. (2) The adaptive learning feature-influence method extends the capability of learning weight compared with the feature weights

Fig. 1. The distribution of house pricing in different districts.

calculated by the Information Gain Ratio. (3) The low-dimension representation increases the accuracy of the housing unit price prediction.

2 PRELIMINARY

In this section, we introduce some concepts and display the mathematical notation to describe the proposed model.

2.1 Feature-influence

The feature-influence means the importance of construction environments and building characteristics to the housing unit prices. If people prioritize some features, they will accept the higher price when the house includes the features. Each city has unique characters due to its geography, economy, and culture. Grasping the feature-influence can help investors dissect the statue of a local housing market, and it can also improve the predicting accuracy of the housing unit price.

2.2 The Heterogeneity of Real Estate

The heterogeneity of real estate means that the buildings have similar natures in the same latent groups, while different natures in various latent groups. Taking buildings in Changchun as an example, the housing information contains five districts. The buildings are divided into five groups based on their own district. As shown in Figure 1(a)–(e), we observe that the distribution of the housing unit price of buildings in the same group is close to a Gaussian distribution, and the distributions in the different groups hold obvious distinctions. As Figure 1(f) shows, the distribution of the housing unit price of all buildings in Changchun is composed of the above five distributions. The heterogeneity of real estate is ubiquitous, and we should consider alleviating its ill effect.

2.3 Expectation-Maximization Algorithm

The Expectation-Maximization (EM) algorithm is used in statistics to find the maximum likelihood estimate of parameters in probability models that rely on unobservable hidden variables [9]. The

Table 1. Mathematical Notations

Symbol	Size	Description
N	—	The number of houses
C	—	The number of latent groups
d	—	The dimension of a feature
ψ	—	The Lagrange multiplier
$Error$	$1 \times N$	The value of the real house price minus the house price estimated by the model
Y	$1 \times N$	The real housing unit price
X	$d \times N$	The feature vectors
H	$d \times 1$	A learned weight vector
W	$d \times C$	The learning matrix of sub-cluster
σ	$1 \times C$	The common noise variance of this Finite Mixture Model
$\aleph$	$1 \times C$	The mixing coefficient of sub-cluster

EM algorithm estimates model parameters through two alternate steps. The first step (E-step) is to calculate the expectation of hidden variables. The second step (M-step) is to maximize the maximum likelihood value to update the estimated parameters. The updated parameters are used in the next E-step calculation, and this process continues alternately. In this work, we can not observe how many latent groups of real estate are in a city, and we have no ability to obtain to which group each building belongs. For estimating the parameters of models with latent variables, we utilize the EM algorithm to estimate the parameters.

2.4 Mathematical Notations

In this work, we predict the housing unit price of buildings based on their corresponding features. The goal is minimizing the error between the real price and the related predicting price. Formally, the features of buildings are $X = (x_1, x_2, \dots, x_n)$, and the real responding housing unit price is $Y = (y_1, y_2, \dots, y_n)$. More notations are shown in Table 1.

3 THE FRAMEWORK OF ADAPTIVE WEIGHTED FINITE MIXTURE MODEL

In this section, we detail the proposed framework named AWFMM. As Figure 2 shows, there are three critical steps in the framework. First, we use the Principle Component Analysis (PCA) algorithm to get the low-dimensional representation of the features of buildings. Then, we utilize two clustering algorithms to get the range of the number of latent groups for avoiding the blind search during the process of model learning. Finally, we apply the EM algorithm to estimate the parameters of the proposed model.

Principal Component Analysis. The PCA [31] is a statistical procedure, which applies an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables. We utilize the PCA algorithm to get the low-dimensional representation of the features of buildings. In this work, the low-dimensional representation has three advantages. Specifically, the low-dimensional representation first can reduce the computational cost of learning the proposed model. Second, it can avoid the potential catastrophe of singular matrix inversion in the process of the estimation of model parameters. Third, it can alleviate the coupling among features as well.

Fig. 2. An overview of framework.

Clustering analysis finds the range of the number of latent groups. To avoid the blind search for the number of latent groups, we propose two clustering algorithms, which are the K-means algorithm [2] and the MiniBatchKMeans algorithm [28], to initially estimate the range of the number of latent groups. We utilize the SC [27] and the CH [5] as measurements to obtain the best result of clusterings.

Parameter estimates for Adaptive Weighted Finite Mixture Model. According to the statistical results in Figure 1, the housing unit prices obey different Gaussian distributions. Therefore, the maximum likelihood function is defined as:

$$\begin{aligned}
 l(\theta)_{\theta=W, \sigma, \mathbf{N}, H} &= \arg \max_{W, \mathbf{N}, H} \sum_{i=1}^N \ln \sum_{z=1}^C \alpha_z P(\text{Error}_{iz} | \sigma_z^2) \\
 &= \arg \max_{W, \mathbf{N}, H} \sum_{i=1}^N \ln \sum_{z=1}^C \alpha_z \cdot \frac{\exp \left(\frac{-(y_i - w_z^T \cdot \text{multiply}(x_i, H))^2}{2\sigma_z^2} \right)}{\sqrt{2\pi}\sigma_z},
 \end{aligned} \tag{1}$$

where N is the number of houses; C is the number of latent groups; Error_{iz} represents the value of the real unit price of the i^{th} house minus the predicting unit price of the related house when the i^{th} house belongs to the z^{th} latent group, $z \in [1, C]$, $i \in [1, N]$; $y_i \in \mathbb{R}^+$ is the real house price of the i^{th} house, $Y = (y_1, y_2, \dots, y_n)$; $x_i \in \mathbb{R}^{d*1}$ is the feature vector of the i^{th} house, $X = (x_1, x_2, \dots, x_n)$; H is the weight vector, $\text{multiply}(x_i, H)$ means the multiplication of the corresponding elements of the feature vector x_i and the weight vector H ; $W = (w_1, w_2, \dots, w_C)$, w_z is the learning matrix of z^{th} sub-cluster, $w_z \in \mathbb{R}^{d*1}$; $\sigma = (\sigma_1, \sigma_2, \dots, \sigma_C)$, σ_z is a common noise variance of z^{th} component of this Finite Mixture Model; $\mathbf{N} = (\alpha_1, \alpha_2, \dots, \alpha_C)$, α_z is the mixing coefficient of sub-cluster and is satisfied with the following constraint:

$$\sum_z \alpha_z = 1 \quad \alpha_z \in [0, 1]. \tag{2}$$

As Equation (1) shows, the total number of errors gets minimum when the likelihood function gets maximum.

Because of the latent variables $\alpha_z \in \mathbb{N}$, we utilize the EM algorithm to estimate the parameters. For the convenient representation, we deliver

$$s_i = \text{multiply}(x_i, H).$$

In E step, we calculate the probability density function $Q(\alpha_z)$ about the latent variables $\mathbb{N}$. Based on Equation (1) and the Jensen Inequality [34], the maximization likelihood function is defined as:

$$\begin{aligned} l(\theta)_{\theta=W, \sigma, \mathbb{N}, H} &= \arg \max_{W, \mathbb{N}, H} \sum_{i=1}^N \ln \sum_{z=1}^C Q(\alpha_z) \frac{\alpha_z P(\text{Error}_{iz} | \sigma_z^2)}{Q(\alpha_z)} \\ &\geq \arg \max_{W, \mathbb{N}, H} \sum_{i=1}^N \sum_{z=1}^C Q(\alpha_z) \ln \frac{\alpha_z P(\text{Error}_{iz} | \sigma_z^2)}{Q(\alpha_z)} \end{aligned} \quad (3)$$

Because function $Q(\alpha)$ is probability density function, the constraint is the following:

$$\sum_z Q(\alpha_z) = 1 \quad Q(\alpha_z) \in [0, 1] \quad (4)$$

Let the following formula satisfy

$$l(\theta)_{\theta=W, \sigma, \mathbb{N}, H} = \arg \max_{W, \mathbb{N}, H} \sum_{i=1}^N \sum_{z=1}^C Q(\alpha_z) \ln \frac{\alpha_z P(\text{Error}_{iz} | \sigma_z^2)}{Q(\alpha_z)} \quad (5)$$

if and only if

$$\frac{\alpha_z P(\text{Error}_{iz} | \sigma_z^2)}{Q(\alpha_z)} = c, \quad (6)$$

where c is a constant. Due to Equation (4), we get the following equation:

$$c = \sum_z \alpha_z P(\text{Error}_{iz} | \sigma_z^2) \quad (7)$$

Based on Equations (6) and (7), we obtain:

$$r_{iz} = Q(\alpha_z) = \frac{D}{S}, \quad D = \alpha_z \frac{\exp \frac{-(y_i - w_z^T \cdot s_i)^2}{2\sigma_z^2}}{\sqrt{2\pi}\sigma_z}, \quad S = \sum_z \alpha_z \frac{\exp \frac{-(y_i - w_z^T \cdot s_i)^2}{2\sigma_z^2}}{\sqrt{2\pi}\sigma_z} \quad (8)$$

After getting Equation (8), we can maximize Equation (5) to estimate parameters $W, \sigma, \mathbb{N}, H$.

In M step, we first calculate the parameter W . Setting the derivative of Equation (5) with respect to w_k equal to zero gives

$$\begin{aligned} \frac{\partial l(\theta)}{\partial w_k} &= \frac{\partial \left(\sum_{i=1}^N \sum_{z=1}^C Q(\alpha_z) \ln \frac{\alpha_z P(y_i - w_z^T \cdot s_i | \sigma_z^2)}{Q(\alpha_z)} \right)}{\partial w_k} \\ &= \sum_{i=1}^N r_{ik} (y_i - w_k^T \cdot s_i) \cdot s_i^T = 0 \end{aligned} \quad (9)$$

We reorganize Equation (9), then get

$$\sum_{i=1}^N r_{ik} w_k^T \cdot s_i \cdot s_i^T = \sum_{i=1}^N r_{ik} y_i s_i^T \quad (10)$$

Let $A_k = \sum_{i=1}^N r_{ik} s_i \cdot s_i^T$, $A_k \in R^{d \times d}$. We obtain

$$w_k^T \cdot A_k = \sum_{i=1}^N r_{ik} y_i x_i^T \quad (11)$$

As $N \gg d$, it is a zero measure that matrix A is a singular matrix. The matrix A is reversible. We obtain

$$w_k = \left[\left(\sum_{i=1}^N r_{ik} y_i s_i^T \right) \cdot A_k^{-1} \right]^T \quad (12)$$

Then, we set $\beta_z = \frac{1}{\sigma_z}$ for convenient calculation. Setting the derivative of Equation (5) with respect to β_k equal to zero gives

$$\begin{aligned} \frac{\partial l(W)}{\partial \beta_k} &= \frac{\partial \left(\sum_{i=1}^N \sum_{z=1}^C Q(\alpha_z) \ln \frac{\alpha_z P(y_i - w_z^T \cdot s_i | \sigma_z^2)}{Q(\alpha_z)} \right)}{\partial \beta_k} \\ &= \sum_{i=1}^N r_{ik} \left(\frac{1}{\beta_k} - \beta_k (y_i - w_k^T \cdot s_i)^2 \right) = 0 \end{aligned} \quad (13)$$

We obtain the equation for β in the form

$$\sigma_k^2 = \frac{1}{\beta_k^2} = \frac{\sum_{i=1}^N r_{ik} (y_i - w_k^T \cdot s_i)^2}{\sum_{i=1}^N r_{ik}} \quad (14)$$

Next, we maximize Equation (5) with respect to the mixing coefficients α_z while keeping fixed other parameters. There is a constraint about α exhibited in Equation (4). We can utilize the Lagrange multiplier to address this constrained optimization problem. To maximize the following quantity

$$l_{Lagrange}(\mathbf{N}, \psi) = \sum_{i=1}^N \sum_{z=1}^C r_{iz} \ln \frac{\alpha_z P(Error_{iz} | \sigma_z^2)}{r_{iz}} + \psi \left(\sum_{z=1}^C \alpha_z - 1 \right), \quad (15)$$

where ψ is the Lagrange multiplier, setting the derivative of Equation (15) with respect to α_k equal to zero gives

$$\frac{\partial l_{Lagrange}(\mathbf{N}, \psi)}{\partial \alpha_k} = \frac{\sum_{i=1}^N r_{ik}}{\alpha_k} + \psi = 0. \quad (16)$$

We rewrite Equation (16) and put it back to the Constraint (2). We can obtain the new α_k

$$\alpha_k = \frac{\sum_{i=1}^N r_{ik}}{N}. \quad (17)$$

Finally, we estimate the parameter H. For the convenient calculation, let

$$u_z = \text{multiply}(w_z, H).$$

The results of $w_z^T \cdot \text{multiply}(x_i, H)$ and $u_z \cdot x_i$ are equivalent. Setting the derivative of Equation (5) with respect to u_k equal to zero gives

$$\begin{aligned} \frac{\partial l(\theta)}{\partial u_k} &= \frac{\partial \left(\sum_{i=1}^N \sum_{z=1}^C Q(\alpha_z) \ln \frac{\alpha_z P(y_i - u_z^T \cdot x_i | \sigma_z^2)}{Q(\alpha_z)} \right)}{\partial u_k} \\ &= \frac{\sum_{i=1}^N r_{ik} (y_i - u_k^T \cdot x_i) \cdot x_i^T}{\sigma_k^2} = 0. \end{aligned} \quad (18)$$

Table 2. The Statistics of Construction Environments and Building Characteristics

Data Type	Attributes	CC				SY				HRB				Description
		min	max	mean	median	min	max	mean	median	min	max	mean	median	
HouseInfo	Room Number	1	6	2.24	2	1	6	2.27	2	1	5	2.06	2	Number of rooms
	Hall Number	0	4	1.70	2	0	4	1.70	2	0	3	1.18	1	Number of halls
	Bathroom Number	0	4	1.18	1	0	5	1.25	1	0	4	1.12	1	Number of bathrooms
	Area(m^2)	28	390	93.83	89.20	20.10	482	94.40	88.60	21.70	400	91.06	88	Area of house
	Floor Number	1	35	14.02	11	1	65	18.97	18	1	49	14.94	8	Number of floors
	Unit Price(RMB)	3176	27778	9968.86	9524	2620	34256	10228.04	9167	2444	32558	9918.03	9065	Price of the house per square meter
CommunityInfo	Completion Year	1985	2020	2010	2012	1986	2021	2011	2012	1978	2021	2008	2010	Completion year of the community
	Property Costs	0	5.30	1.02	0.95	0	5.50	1.47	1.50	0	5	0.88	1	Property costs of the community

We obtain

$$u_k = \left[\left(\sum_{i=1}^N r_{ik} y_i x_i^T \right) \cdot A_k^{-1} \right]^T, \quad (19)$$

where $A_k = \sum_{i=1}^N r_{ik} x_i \cdot x_i^T$, $A_k \in R^{d \times d}$. Finally, the H is defined as

$$H = \frac{\sum_{z=1}^C \text{divide}(u_z, w_z)}{C}, \quad (20)$$

where $\text{divide}(u_z, w_z)$ means each element of the vector u_z is divided by the corresponding component of the vector w_z .

After getting the estimated learning matrix W , the mixing coefficient $\mathbf{S}$, and the weight vector of features H , the unit price can be predicted. The unit price of a new house can be predicted by Equation (21).

$$Y = \text{Sigmoid} \left(\sum_{z=1}^C \lambda_z \left(w_z^T \cdot \text{multiply}(x_{new}, H) \right) \right) \quad \lambda_z = \begin{cases} 1, & z = \text{Gambling}(\mathbf{S}) \\ 0, & \text{others}, \end{cases} \quad (21)$$

where x_{new} is a feature vector of an expected house. $\text{Gambling}(\mathbf{S})$ is the method to determine that a new house belongs to a certain hidden group. The possibility of a house belonging to a category is proportional to the value of elements in $\mathbf{S}$. The function $\text{sigmoid}(x) = \frac{1}{1+e^{-x}}$ is an activation function.

4 EXPERIMENTS

In this section, we design a series of experimental comparisons to demonstrate the effectiveness of the proposed framework on three real-world datasets.

4.1 Data Description

Table 2 shows three real-world datasets used in the experiments. The datasets contain more than 48 thousand real estate transaction records, including the construction environments, building characteristics, and housing unit prices in Changchun, Shenyang, and Harbin, which are the provincial capital cities all in the northeastern part of China. In addition to the partial information in the table, our house feature information also includes districts, locations, the orientation, the type of floor, the elevator, the decoration, the years of the property, the type of buildings, and the type of property. Specifically, the type of floor represents a description of whether the house is in the high-rise, middle-rise, or low-rise. The elevator means if the house has an elevator. The decoration describes that the house is rough, simple, fine, or luxurious. The years of property

equals to the difference between the present time and the time of obtaining the house property card, which contains the following three categories; there are less than two years, a full two years, and a full five years. The types of houses are divided into three categories that contain the ordinary house, the villa, and the apartment. The kinds of property are the commercial type, the commercial and residential house type, and the economic house type.

To encode the above information into a vector, we directly utilize the value of numerical attributes (e.g., the number of rooms), and apply the one-hot encoding method [35] for vectorizing the nominal attributes (e.g., house location).

4.2 Baselines

The baselines include classical, widely used regression methods and the variant methods based on the Adaptive Weighted Finite Mixture model. As follows:

Classical Widely Used Methods:

- (1) *Linear Regression (LR)*: A classical Linear Regression models the relationship between features and the unit price.
- (2) *Lasso Regression (LA)*: A Lasso Regression is a Linear Regression with L1 regularization [13].
- (3) *Support Vector Regression (SVR)*: A Support Vector Regression is similar to the Support Vector Machine, but the output is a real number [29].
- (4) *KNeighbors Regressor (KNN)*: It is a regression model based on k-nearest neighbors [4].
- (5) *AdaBoost Regressor (ABR)*: An AdaBoost Regression model [15], which is based on a boosting method.

Variant Methods

- (1) *Finite Mixture Model with Information Gain Ratio (FMMIGO)*: The method is our previous work [33], which is a finite mixture model with the fixed feature-influence calculated by IGO.
- (2) *AWFMM50*: This method is the AWFMM, where we transfer the data to 50 dimensions of data by PCA.

4.3 Evaluation Metrics

We utilize four metrics to evaluate the performance of each algorithm from two aspects, which are accuracy and stability. We display the definition of evaluation metrics as follows:

- (1) Mean Square Error (MSE):

$$MSE = \frac{\sum_i^N |Y_i - \hat{Y}_i|^2}{N} \quad (22)$$

- (2) Mean Absolute Error (MAE):

$$MAE = \frac{\sum_i^N |Y_i - \hat{Y}_i|}{N} \quad (23)$$

- (3) Root Mean Squared Error (RMSE):

$$RMSE = \sqrt{\frac{\sum_i^N (Y_i - \hat{Y}_i)^2}{N}} \quad (24)$$

Fig. 3. Cluster results in different cities.

Table 3. The Performance Results with the Different Number of Latent Groups in Different Cities

Method	Latent groups	CC		SY		HRB	
		RMSE	SDE	RMSE	SDE	RMSE	SDE
AWFMM	k=2	0.1346	0.0931	0.2562	0.1581	0.2664	0.1840
	k=3	0.1202	0.0834	0.2562	0.1563	0.2729	0.1920
	k=4	0.1327	0.0920	0.2566	0.1568	0.2782	0.1914
	k=5	0.1227	0.0843	0.2560	0.1570	0.2734	0.1886
	k=6	0.1366	0.0967	0.2717	0.1668	0.2817	0.1954
	k=7	0.1273	0.0883	0.2589	0.1564	0.2862	0.1966
	k=8	0.1224	0.0840	0.2591	0.1564	0.2779	0.1939

(4) Standard Deviation of Error (SDE):

$$SDE = \sqrt{\frac{\sum_i^N (Error^2 - \bar{Error})^2}{N}} \quad (25)$$

The first three metrics, which are MSE, MAE, and RMSE, are used to evaluate the predicting accuracy. The performance of prediction accuracy is better when the metrics are smaller. The Standard Deviation of Error is an index of how closely the individual error points cluster around the mean of errors [30]. The lower SDE metric represents that the proposed method has a stable performance of prediction of the house unit prices on the whole perspective and vice versa.

4.4 Performance of Adaptive Weighted Finite Mixture Model

To avoid the blind search of the number of latent groups, we first set variables from 2 to 8 for two clustering methods, which are K-means [2] and minibatchKmeans [28]. Based on the metrics of the SC [27] and the CH [5], Figure 3 shows that the best number of clusters is 2. With the number of latent groups increasing, the value of metrics decreases. Inspired by the aforementioned process, we perform experiments to find the best number of latent groups among the latent group's number from 2 to 8, making the best performance of our proposed method (AWFMM) in different cities.

Next, we utilize the RMSE and the SDE to evaluate the best latent groups number. As is shown in Table 3, we find that the best latent groups numbers are 3 in Changchun (CC) and 2 in Harbin

Fig. 4. The comparison of performance between baselines.

(HRB). However, in Shenyang (SY), the RMSE and the SDE show inconsistency in evaluating the best latent groups. In this case, we choose the smallest RMSE as the best result because the gap of SDE is slight. Therefore, the best latent group number is 5 in SY.

Figure 4 shows the results of the comparison between our model, the AWFMM, and five baselines. As can be seen in this figure, the AWFMM consistently outperforms other baselines in the accuracy metrics. While for the value of SDE, the AWFMM doesn't have overwhelming superiority, which means that the stability of our framework needs to be improved in rare cases.

4.5 Case Studies

In this section, we first compare two kinds of weighting mechanisms. Then, we discuss the different influence patterns in similar cities. At last, we deliver the reason for utilizing the PCA.

Two kinds of methods of learning weights: We compare the proposed method (AWFMM) with our previous work (FMMIGO) [33] where we utilize the Information Gain Ratio to calculate the weights of features. The results of the two weighting methods are present in Figure 5. The AWFMM dominates FMMIGO on almost all of the datasets on all metrics. The only exception is that the AWFMM performs a little worse on the SDE metric in Harbin. We think the FMMIGO method makes predicting the housing unit price in Harbin city more stable.

Fig. 5. The comparison of performance between weighting methods.

Fig. 6. An overview of IGO ranking in different cities.

Different Patterns of Weight of Features in Different Cities: Figure 6 shows the ranking of all 16 features through its IGO value. The feature is more critical when the IGO value is higher. For example, in Changchun, as can be seen in Figure 6(a), the essential influence on the unit price are house type, area, and location, respectively. As Figure 7 shows, we identify the patterns among three cities. Interestingly, we find that the floor type is the least noticed feature in all the cities.

The Reasons of Utilizing the Principle Component Analysis: We apply the PCA to pre-process our data not only for reducing the dimension but for avoiding facing a singular matrix inversion, which

Fig. 7. An comparison of different patterns of feature-influence based on IGO in different cities.

is caused by the sparsity of the feature matrix. The sparsity is caused by the one-hot decoding method for the nominal attribute.

To find the best transferred dimension of the representation of data, we transfer the original data (107 dimensions), respectively, into 50 dimensions and 16 dimensions. We name the AWFMM with 50-dimensional representation as “AWFMM50,” and the proposed model with 16-dimensional representation as “AWFMM16.” As Figure 8 shows, AWFMM16 obtains better results in comparable experiments, except for some cases on the SDE metric, but the gap is much smaller. Therefore, in our research, we transfer the original data into the 16 low-dimensional representation.

Dissimilarity of feature-influence among similar cities: We discover the existing dissimilarity of feature-influence in similar cities. Figure 9 shows the obvious differentiation among the three cities, although they may have similar geography, economics, and cultures.

5 RELATED WORK

In our study, we propose a model based on a finite mixture model to identify the feature-influence and predict housing unit price. Following this study, we summarize the most relevant research as follows.

5.1 House Price Prediction

A recommendation system has been applied in many fields [23]. House price prediction is an essential role in the real estate recommendation system. Researchers have proposed many related works. Limsombunchai utilized the hedonic model with an artificial neural network model to compare the predictive power on housing price predictions [22]. Mcmanus and Mumey proposed the method to estimate the price index using property value data from refinancing transactions [25]. Clapp and Giaccotto calculated a city-wide house price index using an autoregressive process [6]. Drake and Leigh used a parsimonious dynamic model to estimate UK housing prices [11].

5.2 Finite Mixture Model

The proposed method has a connection with the Finite Mixture Model. Figueiredo and Jain learned a finite mixture model from multivariate data with an unsupervised algorithm [14]. McLachlan and McGiffin found a way to simulate failure times in competitive risk situations in the field of survival analysis [24]. Zivkovic and Heijden proposed a recursive algorithm to estimate both the parameters of the mixture model and the number of components [36]. Bailey et al. discovered how to use the technique of expectation maximization to train a finite mixture model to find successive motifs [3].

5.3 Information Gain Ratio

In the previous work [33], the authors utilized the Information Gain Ratio to calculate the weights of features. Karegowda et al. demonstrated the significance of feature subset selection for classifying Pima Indian diabetic database (PIDD) using Gain ratio and correlational feature selection [21]. Dai et al. proposed an attribute selection method using a fuzzy gain ratio with a fuzzy rough set

Fig. 8. The comparison of performance between different dimensions of data.

Fig. 9. An comparison of patterns of adaptive learning feature-influence.

[8]. Jia et al. developed an attributes reduction algorithm based on mutual information gain ratio, and it obtained sound relative attributes reduction in decision systems [20].

5.4 The Low-Dimensional Representation

The low-dimensional representation reduces the computational cost of model learning, avoids the potential catastrophe of singular matrix inversion in the process of model parameter estimation, and alleviates the coupling among features as well. In numerous research of the PCA, Chris Ding and Xiaofeng found that the unsupervised dimensionality reduction is closely related to unsupervised learning by combining the k-means and the PCA [10]. Jackson and Mudholker learned the treatment of residuals related to principal component analysis and proposed a procedure for testing residuals [18]. Mingjing et al. introduced the PCA based on the Density Peaks Clustering (DPC) algorithm to deal with the high-dimensional dataset, and the results showed their method to be effective [12].

6 CONCLUSION

In this article, we investigated how to identify the feature-influence of real estate. Since the heterogeneity is ubiquitous, we developed an adaptive weighted finite mixture model to adaptively learn weights of features and estimate a housing unit price based on the construction environments and building characteristics. We utilized the PCA algorithm to reduce the computational cost of the proposed model learning, avoid the potential catastrophe of singular matrix inversion in the process of the estimation of model parameters, and alleviate the coupling among features as well. For addressing the blind search, we proposed a pre-clustering method for obtaining a guide of the range of the number of latent groups. Besides, we constructed many experiments for demonstrating that our proposed method dominates other widely used baselines on three real-world datasets. Finally, based on the feature-influence, we discovered the dissimilarity of feature-influence, even among the cities with similar geography, economies, and cultures. In the future, we plan to re-research on the influence of feature-weight continually. We hope the result can be more explainable and more credible.

REFERENCES

- [1] Hervé Abdi and Lynne J. Williams. 2010. Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics* 2, 4 (2010), 433–459.
- [2] David Arthur and Sergei Vassilvitskii. 2007. k-means++: The advantages of careful seeding. In *Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms*. Society for Industrial and Applied Mathematics, 1027–1035.
- [3] Timothy L. Bailey and Charles Elkan. 1994. Fitting a mixture model by expectation maximization to discover motifs in bipolymers. In *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*. AAAI Press, Menlo Park, California, 28–36.
- [4] Jon Louis Bentley. 1975. Multidimensional binary search trees used for associative searching. *Commun. ACM* 18, 9 (1975), 509–517.
- [5] Tadeusz Caliński and Jerzy Harabasz. 1974. A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods* 3, 1 (1974), 1–27.
- [6] John Clapp and Carmelo Giaccotto. 2002. Evaluating house price forecasts. *Journal of Real Estate Research* 24, 1 (2002), 1–26.
- [7] John M. Clapp and Carmelo Giaccotto. 1994. The influence of economic variables on local house price dynamics. *Journal of Urban Economics* 36, 2 (1994), 161–183.
- [8] Jianhua Dai and Qing Xu. 2013. Attribute selection based on information gain ratio in fuzzy rough set theory with application to tumor classification. *Applied Soft Computing* 13, 1 (2013), 211–221.
- [9] Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)* 39, 1 (1977), 1–22.
- [10] Chris Ding and Xiaofeng He. 2004. K-means clustering via principal component analysis. In *Proceedings of the 21st International Conference on Machine Learning*. ACM, 29.

- [11] Leigh Drake. 1993. Modelling UK house prices using cointegration: An application of the Johansen technique. *Applied Economics* 25, 9 (1993), 1225–1228.
- [12] Mingjing Du, Shifei Ding, and Hongjie Jia. 2016. Study on density peaks clustering based on k-nearest neighbors and principal component analysis. *Knowledge-Based Systems* 99 (2016), 135–145.
- [13] Bradley Efron, Trevor Hastie, Iain Johnstone, Robert Tibshirani, et al. 2004. Least angle regression. *The Annals of Statistics* 32, 2 (2004), 407–499.
- [14] Mario A. T. Figueiredo and Anil K. Jain. 2002. Unsupervised learning of finite mixture models. *IEEE Transactions on Pattern Analysis & Machine Intelligence* 3 (2002), 381–396.
- [15] Yoav Freund and Robert E. Schapire. 1997. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences* 55, 1 (1997), 119–139.
- [16] Y. Fu, Y. Ge, Y. Zheng, Z. Yao, Y. Liu, H. Xiong, and J. Yuan. 2014. Sparse real estate ranking with online user reviews and offline moving behaviors. In *Proceedings of the IEEE International Conference on Data Mining*.
- [17] Yanjie Fu, Hui Xiong, Yong Ge, Zijun Yao, Yu Zheng, and Zhi-Hua Zhou. 2014. Exploiting geographic dependencies for real estate appraisal: A mutual perspective of ranking and clustering. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1047–1056.
- [18] J. Edward Jackson and Govind S. Mudholkar. 1979. Control procedures for residuals associated with principal component analysis. *Technometrics* 21, 3 (1979), 341–349.
- [19] Hanwool Jang, Kwangwon Ahn, Dongshin Kim, and Yena Song. 2018. Detection and prediction of house price bubbles: Evidence from a new city. In *Proceedings of the International Conference on Computational Science*. Springer, 782–795.
- [20] Ping Jia, Jian-hua Dai, Yun-he Pan, and Miao-liang Zhu. 2006. Novel algorithm for attribute reduction based on mutual-information gain ratio. *Journal-Zhejiang University Engineering Science* 40, 6 (2006), 1041.
- [21] Asha Gowda Karegowda, A. S. Manjunath, and M. A. Jayaram. 2010. Comparative study of attribute selection using gain ratio and correlation based feature selection. *International Journal of Information Technology and Knowledge Management* 2, 2 (2010), 271–277.
- [22] Visit Limsombunchai. 2004. House price prediction: Hedonic price model vs. artificial neural network. In *Proceedings of the New Zealand Agricultural and Resource Economics Society Conference*. 25–26.
- [23] Qi Liu, Yong Ge, Zhongmou Li, Enhong Chen, and Hui Xiong. 2011. Personalized travel package recommendation. In *Proceedings of the 2011 IEEE 11th International Conference on Data Mining*. IEEE, 407–416.
- [24] G. J. McLachlan and D. C. McGiffin. 1994. On the role of finite mixture models in survival analysis. *Statistical Methods in Medical Research* 3, 3 (1994), 211–226.
- [25] Douglas A. McManus and Sol T. Mumey. 2002. System and method for providing house price forecasts based on repeat sales model. US Patent 6,401,070.
- [26] J. Ross Quinlan. 1986. Induction of decision trees. *Machine Learning* 1, 1 (1986), 81–106.
- [27] Peter J. Rousseeuw. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* 20 (1987), 53–65.
- [28] David Sculley. 2010. Web-scale k-means clustering. In *Proceedings of the 19th International Conference on World Wide Web*. ACM, 1177–1178.
- [29] Alex J. Smola and Bernhard Schölkopf. 2004. A tutorial on support vector regression. *Statistics and Computing* 14, 3 (2004), 199–222.
- [30] David L. Streiner. 1996. Maintaining standards: Differences between the standard deviation and standard error, and when to use each. *The Canadian Journal of Psychiatry* 41, 8 (1996), 498–502.
- [31] Michael E. Tipping and Christopher M. Bishop. 1999. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 61, 3 (1999), 611–622.
- [32] Catherine Tucker, Juanjuan Zhang, and Ting Zhu. 2013. Days on market and home sales. *The RAND Journal of Economics* 44, 2 (2013), 337–360.
- [33] Xin Xu, Zeyu Huang, Jingyi Wu, Yanjie Fu, Na Luo, Weitong Chen, Jianan Wang, and Minghao Yin. 2019. Finding the key influences on the house price by finite mixture model based on the real estate data in Changchun. In *Proceedings of the International Conference on Database Systems for Advanced Applications*. Springer, 378–382.
- [34] Zhi-Hua Zhou and Xu-Ying Liu. 2006. Training cost-sensitive neural networks with methods addressing the class imbalance problem. *IEEE Transactions on Knowledge & Data Engineering* 1 (2006), 63–77.
- [35] Hengshu Zhu, Hui Xiong, Fangshuang Tang, Qi Liu, Yong Ge, Enhong Chen, and Yanjie Fu. 2016. Days on market: Measuring liquidity in real estate markets. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 393–402.
- [36] Zoran Zivkovic and Ferdinand van der Heijden. 2004. Recursive unsupervised learning of finite mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 26, 5 (2004), 651–656.

Received June 2019; revised October 2019; accepted December 2019