Precision Marketing Scheme based on Integrating Spatio-temporal Data Clustering and Neural Network

Su Ying LIU

School of Mechanical Engineering, Shandong University, Jinan Shandong, 250002 China

liusy579@163.com

Abstract. Precision marketing can effectively predict the user's willingness to consume, with nearly one trillion of market space. The current precision marketing strategy still has remarkable problems such as the low correlation between data model and space; the marketing initial data redundancy and so on. Aiming at the low correlation between the model and space and the initial data redundancy in the marketing, this paper proposes a precise marketing scheme combining the spatiotemporal data clustering and the neural network. Through the establishment of multi - dimensional space - time large data model, effectively enhance the geographic data model, spatio - temporal data model and multi - dimensional data superposition. Design the large-scale data normalization clustering scheme, construct the normalized classification Euclidean distance and the adaptive precision marketing objective function, and realize the K-Means data clustering of the initial data redundancy. Through the neural network based spatial and temporal data precision marketing decision, design data input layer; deal with hidden layer and intelligent output layer; provide an effective program for accurate marketing.

1. Introduction

Precise marketing is a strategy mining the most fit of human will from the massive, complex, subtle characteristics, to make the prediction reasoning the closest to the actual purchase [1]. Based on data collection, data processing and large data analysis, precision marketing can perceive the shopping intention of consumers with time and space changes, and also promote the sensor technology, communication network technology, computer processing technology to develop and integrate deeply.

Domestic and foreign enterprises and research institutions are to increase investment in research of more intelligent and accurate precision marketing program [2]. 1999 Dresden University of Technology in Germany has achieved a personalized e-commerce prototype system TELLIM. NEC Research Institute, New York University, IBM and other companies have launched search engine CiteSeer, personalized e-commerce user modeling system, Websphere and so on. In 2007, Yahoo launched the SmartAds advertising program, in 2009, Overstock launched a personalized banner advertising program, can effectively achieve the search engine keyword search, the database regularly issued EDM and demand side platform. With the "Internet plus" economic development and the gradual application of Internet of things, precision marketing will usher in nearly one trillion of market space.

The current precision marketing strategy mainly includes the integration of mobile Internet, public communication platform and large data support and other means. Integrate mobile Internet with 3G, 4G mobile network for fast data transmission, and to achieve the optimal control of advertising content,

presentation and delivery mode. But because of the lack of the ability of mobile Internet to perceive users desire in shopping, cannot be associated with space, so accurate marketing accuracy is not high [3]. With the public communication platform can increase the data collection channels of shopping intention of the users, such as WeChat, microblogging, space, etc. to a certain extent, reflect the user's behavioral characteristics, but a lot of redundant data for the platform processing capacity is a serious challenge. Through a large data processing approach, to achieve a good user behavior data feature acquisition, processing, mining awareness system [4]. However, it is difficult to construct behavioral feature librarya directly, to form a certain size of the potential consumer database in initial state. Followed by the database construction cycle is long, and in the library data's timing requirements are sensitive.

To sum up, the current precision marketing strategy mainly has the following problems:

(1) Correlation between data model and space is low. The existing data model use generally the commodity circulation process as the main model, and spatial correlation is low. And offline shopping behavior and space are closely related, need to establish multi-dimensional space-time large data model.

(2) Marketing initial data redundancy. As the marketing behavior involves many objects, each object has multiple dimensions and multiple eigenvalues, and each eigenvalue has a big difference in the role of marketing reasoning, which makes it difficult to deal with data. So need to simplify process data.

Aiming at the low correlation between the model and space and the initial data redundancy in the marketing, this paper proposes a precise marketing scheme combining the spatiotemporal data clustering and the neural network. Firstly, the multi-dimensional space-time large data model is established to effectively enhance the association between geographic data model, spatio-temporal data model and multidimensional data superposition method. Then, we design a large-scale data normalization clustering scheme, construct the normalized Euclidean distance and the adaptive precision marketing objective function, and realize the K-Means data clustering of the initial data redundancy. Finally, through the neural network based spatio-temporal data precision marketing decision-making, design data input layer, dealing with hidden layer and intelligent output layer, providing an effective solution for accurate marketing.

2.Multidimensional space-time large data model

2.1. Geographic data model

Du Daosheng from Wuhan University of Surveying and Mapping Science and Technology, presents a temporal geographic data model based on synchronous data item group and fragmented topological arc time mark. Based on the spatial unit subdivision, this paper effectively solve the problems of traditional geographic data model in the complex geospatial information expression when the existence of redundant data storage, local changes in information extraction difficulties and geographic entities of the time characteristics of the performance is not obvious and so on. Nanjing University of Technology Zhu Xiaolin proposed a geo-analysis task-driven multi-source geographic data extraction and push method research. Based on MongoDB, the geography data of the existing multi - source geographic data are unified and organized. Through the construction of model requirements template and task model driven, this paper realize the dimension extraction and push of geospatial data, effectively solve the difference between data transformation and geographic model transformation, and reduce the complexity of model coupling. The Geographic data model based on unified semantic expressions is proposed by Qiu Jianni, Guangzhou Geography Institute, and the integrated architecture of traffic geographic data diversity characteristics and unified semantic representation are designed to support the effective representation of different features under different topological conditions, effectively solve the problems of poor coordination and low interaction in the traditional geography.

2.2. Spatio-temporal data model

China University of Geosciences, Liu Fei, an event-based dual-sequence spatiotemporal data model is

proposed. The changes of state and spatial object are expressed by double sequence. Using the change of sequence storage object to solve the problem that the existing spatio-temporal data model is based on the spatial object individual with weak expression of time and space, achieve the effective management of space-time data. Nanjing Geographic Information Application Engineering Technology Research Center Wang Lvhua proposes a recursive algorithm to improve the application of the recursive algorithm in improving the historical backtracking of the space-time data model. The main life cycle management of a single GIS space object is the main line to solve its efficiency and show problems in multi-level historical backtracking. Improving the representation of the spatial and temporal data model of the ground state and the design of the bi-directional recursive algorithm to realize the efficient retrieval of the historical relationship of the spatial data under the frequent change. The GIS tense player is used to describe the change process in the form of animation; to display the historical evolution of spatial objects. Wuhan University Lin Yinchao proposes a hybrid space-based spatial and temporal data storage model based on object and snapshot. The space-time process of plaque objects covered by object-oriented description is used to organize the space-time events and spaces, attribute information, at the same time, it covers the whole spatio-temporal distribution with snapshots, and organizes the snapshots of the grid. The two forms a mixed model based on the logical association relation between time and space position. On the basis of the change of the expression object, increase the expression of spatiotemporal events between the different categorized objects.

2.3. Multidimensional data superposition method

Xi'an University of Electronic Science and Technology of China, Hui Zhuanni, proposes a kind of multidimensional data clustering algorithm based on Global K-means and its GPU acceleration. Based on the attribute weight and entropy weight, the clustering results of the global K-means algorithm are stable and accurate, which can effectively solve the clustering problem caused by data sparsity. Based on the GPU-based parallel Global K-means algorithm-PGKM_Mix algorithm, the most time-consuming clustering center in large-scale data mining is selected to achieve a high speedup ratio, which solves the problem that the Global K-means algorithm is computationally complex. Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences Wang Shaohua has proposed a topological consistency processing for vector data superposition analysis, including the core process of inter-arc topology processing, topological processing between nodes and arcs, and proximity search between nodes. The processing has a higher guarantee of topology consistency processing effect, is a kind of practical strong topology consistency processing algorithm. East China Jiaotong University Hu Hui proposed a fast capture algorithm based on BIT bit correction and data superposition, can effectively suppress the influence of navigation data transitions under long integration time condition and reduce the amount of capture computation, which is much lower than that of delay multiplication algorithm, signal compression algorithm, non-coherent acquisition algorithm, FFT parallel code search algorithm, has a great advantage.

3. Massive data normalized clustering

Modern large-scale commercial area is large, the number of shops and commercial marketing data are massive and complicated, such as Jinan Heng Long Plaza, the construction area of 280,000 square meters, nearly 359 businesses, more than 100,000 different types of goods, commodity features more than 10 million species, the relationship between space, customers and the goods reached one billion order of magnitude [5]. Resulting in data processing difficulties, the need for large amounts of large data normalized clustering is necessary. Through the establishment of multi - dimensional space - time large data model, effectively enhance the relevance of the geographic data model, spatio - temporal data model and multi - dimensional data superposition method.

The K-means clustering algorithm takes the mean of all commercial marketing data samples in each subset of commercial marketing data as the representative point of the cluster [6]. The main idea of the algorithm is to divide the commercial marketing data set into different Category through the iterative process, so that the criteria function of evaluating the clustering performance is optimal, so that each

cluster is compact and independent. The specific processes include the normalized Euclidean distance, the adaptive precision marketing objective function, and the K-Means data clustering process.

3.1. Normalized Euclidean distance
When calculating the distance between commercial marketing data samples, the Euclidean distance [7] can be selected as a measure of similarity between data samples.

Set the given business marketing data set to

$$X = \{x_m | m = 1, 2, \dots, \text{total}\} \tag{1}$$

The samples in X are represented by d descriptive attributes A1、A2···Ad, and the d descriptive attributes are contiguous. Data samples xi=(xi1,xi2,···xid), xj=(xj1,xj2,···xjd). Among them, xi1，xi2，···xid and xj1，xj2，···xjd are the specific value of samples xi and xj corresponding to the d descriptive attributes A1、A2···Ad. The similarity between the samples xi and xj is usually expressed by the distance d(xi,xj) between them. The smaller the distance, the closer the samples xi and xj are, the smaller the difference; the larger the distance, the more the samples xi and xj not similar, the greater the degree of difference.

The Euclidean distance between the commercial marketing data samples is,

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{d}(x_{ik} - x_{jk})^2} \tag{2}$$

3.2. Adaptive precision marketing objective function
As the commercial marketing data is expected to tap the data, and customer shopping behavior characteristics can be effectively matched, therefore, select the user's shopping frequency and the number of goods purchased as the main indicators of data clustering [8].

Given the commercial marketing data set X, where the user's shopping frequency and the number of times the item was purchased contains only the description attribute, the category attribute is not included. Suppose X contains k clustering subsets $X_1, X_2, \dots, X_K$; the number of samples in each clustering subset is $n_1, n_2, \dots, n_k$, respectively. By using the square sum criterion function of error, we can adjust the subspace boundary according to the principle of squared sum of the fitting error declines from the initial division of precision marketing, obtain the new spatial partition and the corresponding regression model, and iterate until the criterion function converges. Here, the user's shopping frequency and the average number of times the purchased number of goods represent points m1 and m2, the error squared and criterion functions are as follows:

$$E = \sum_{i=1}^{k}\sum_{p \in X_i} ||p - m_i||^2 \tag{3}$$

3.3. K-Means data clustering process
The calculation of the similarity is made according to the average of the objects in the m_1 and m_2 clusters. The specific algorithm process is as follows:

Step 1: All objects are randomly assigned to k non-empty clusters using the Euclidean distance in Eq. (1).

Step 2: Calculate the average of m_1 clusters and m_2 clusters and use the average to represent the corresponding clusters.

Step 3: Assign to the nearest m_1 cluster and m_2 cluster according to the distance between each object and the m_1 cluster and m_2 cluster center.

Step 4: Then turn (2) to recalculate the average of each cluster. This process is repeated until the square of the error and the criterion function satisfies the formula (3).

The data clustering process input commercial marketing data set X to the K-Means; set the Euclidean distance between the data samples as the similarity measure [9]. The user's shopping frequency m_1 and the average number of the purchases of goods m_2 are the control relationship of the error square and the criterion function and you can get the commercial marketing data set clustering simplified results shown in Figure 1.

According to the description of Figure 1, the blue dot is the original commercial marketing data set,

4

the green point is the original data center set, the red dot is the clustering result of the shopping frequency m_1 of the user and the average number of purchased times of the goods m_2. It is seen that the number of clustered data points drops significantly compared to the original blue dots.
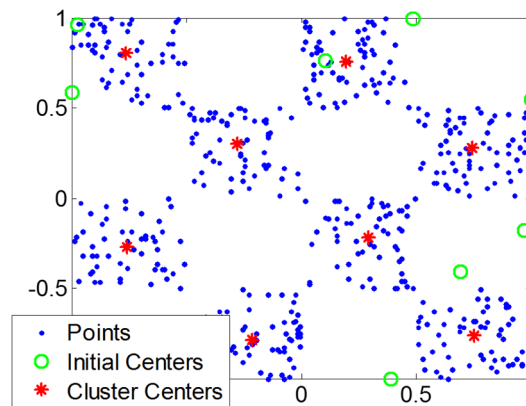

Figure 1 K-Means data clustering for business marketing datasets

4. Precise marketing decision of spatio-temporal data based on neural network

The K-Means data clustering of the initial data redundancy is realized by constructing the normalized Euclidean distance and the adaptive precision marketing objective function through the mass data normalization clustering scheme. After clustering simplifying the commercial marketing data set, the user's shopping frequency m_1 and the average number of purchased items m_2 are obtained. Ordered data can be processed through the neural network-based association processing, you can get accurate marketing strategy of time and space data.

Artificial neural network can simulate the working mode of biological neurons, and carry out mathematical model of distributed parallel information processing. It mainly includes data input layer, processing hidden layer and intelligent output layer. The concrete structure is shown in Fig.2.
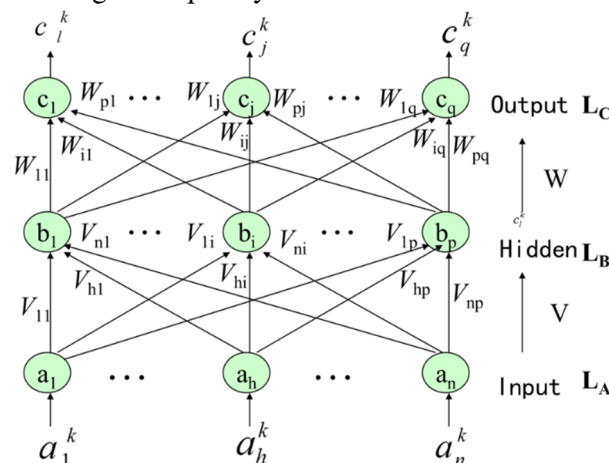

Figure 2 Multi-level structure of neural network

As shown in Figure 2, a_n^k is a subclass of m_1 and m_2, a_n is the control state of the input layer, b_p is the control state of the hidden layer, and c_q is the control state of the output layer. V_h^p is the change from the input state to the implied state, and W_h^p is the change from the implied state to the output state.

4.1. Data input layer design

In order to improve the accuracy of accurate marketing decision-making, it is necessary to extract enough samples from m1 and m2 to support, randomly divided into training samples, examine samples (10% or more) and test samples (10% or more), and give full consideration to the balance between

5

samples patterns.

Since the user's shopping frequency m1 and the average number of purchased times m2 are used as the endogenous variables (influencing factors or independent variables) of the precise marketing decision in the normalized clustering process, it isprocessed [10]. The formation of the value between 0.2 to 0.8, to meet the hidden layer requirements of using Sigmoid function of the unsaturated zone , the normal formula is as follows,

$$\overline{a_n^k} = \frac{1}{n+k}\sum_i^n \sum_j^k a_i^j \tag{4}$$

4.2. Processing hidden layer design

In the neural network design, the hidden layer is the layers other than the input layer and the output layer. Implicit layer design points are the number of hidden layers and the number of hidden nodes. It is necessary to satisfy the following two theorems,

■   Linear Sigmoid approximation

If the input layer and the output layer adopt the linear transformation function, the hidden layer adopts the Sigmoid transformation function, then the MLP network with a hidden layer can approximate any rational function with arbitrary precision.

■   Limited precision of the most compact

In order to avoid the "over-fitting" phenomenon and to ensure a high enough network performance and generalization ability as far as possible, the most basic principle of determining the number of hidden nodes is to make the structure as compact as possible under the premise of satisfying the precision requirement, That is, take as few as possible hidden layer nodes.

Thus, the necessary conditions for determining the number of hidden nodes are obtained:

(1) The number of hidden nodes in precise marketing decision must be less than N-1 (where N is the number of training samples) and the number of nodes of input layer (the number of variables) must be less than N-1.

(2) precision marketing decision-making training samples must be more than the connection of the network model, usually 2 to 10 times.

Sigmoid is set to

$$\text{sgn}(f(x)) = \begin{cases} 1, 0.5 < f(x) < 1 \\ 0, 0 < f(x) < 0.5 \end{cases} \tag{5}$$

4.3. Intelligent output layer design

In order to improve the accuracy of the output, the neural network will be trained , using output data as the next level of neural network input data re-reasoning. By adjusting the network weights by applying the error backtracking principle, the sum of the square errors between the output value of the network model and the known training sample output value is minimized or less than a certain expected value.

The sub-sets of m_1 and m_2 are processed into the neural network of Fig. 2 to obtain the fitting results as shown in Fig. 3.

Can be obtained from the figure, after many fitting times by the user's shopping frequency and the average number of purchases of goods, the user's intention derived can be gradually close to the user's true will.
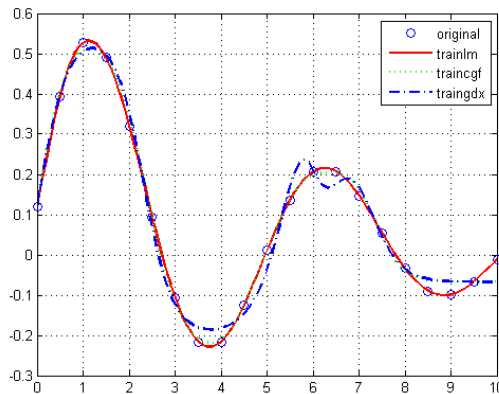
Figure 3 Precision marketing data fitting results

5. Conclusion

In this paper, through the integration of spatio-temporal data clustering and neural network of precision marketing program, designed a large mass data normalized clustering program; built a classification of Euclidean distance and adaptive precision marketing objective function to achieve the K-Means data clustering of the initial marketing data redundancy. Through the neural network based precise marketing decision-making, design data input layer, processing hidden layer and intelligent output layer, which effectively solves the problem of low spatial correlation of model and the initial data redundancy in precise marketing.

References

[1]   Liu, Z. Y. (2007). Study on precision marketing method. Journal of Shanghai Jiaotong University.
[2]   Lin, G. Z., & Fan, P. F. (2009). Precision marketing of telecommunication corporation in 3g times. Journal of Chongqing University of Posts & Telecommunications.
[3]   You, Z., Si, Y. W., Zhang, D., Zeng, X. X., Leung, S. C. H., & Li, T. (2015). A decision-making framework for precision marketing. Expert Systems with Applications An International Journal, 42(7), 3357-3367.
[4]   Zhang, J., & Zhu, J. (2014). Research intelligent precision marketing of e-commerce based on the big data. Journal of Management & Strategy, 5(1).
[5]   Esmin, A. A. A., Coelho, R. A., & Matwin, S. (2015). A review on particle swarm optimization algorithm and its variants to clustering high-dimensional data. Artificial Intelligence Review, 44(1), 23-45.
[6]   Aparna, K., & Nair, M. K. (2015). Comprehensive study and analysis of partitional data clustering techniques. Journal of Clinical Nursing, 16(8), 1405–1416.
[7]   Seera, M., Lim, C. P., Chu, K. L., & Singh, H. (2015). A modified fuzzy min–max neural network for data clustering and its application to power quality monitoring. Applied Soft Computing, 28(C), 19-29.
[8]   Sjöström, P. J., Frydel, B. R., & Wahlberg, L. U. (2015). Artificial neural network-aided image analysis system for cell counting. Cytometry Part A, 36(1), 18-26.
[9]   Graves, A., Wayne, G., Reynolds, M., Harley, T., Danihelka, I., & Grabska-Barwińska, A., et al. (2016). Hybrid computing using a neural network with dynamic external memory. Nature, 538(7626), 471.
[10]   Lin W, Deng Z, Fang Q, et al. A new satellite communication bandwidth allocation combined services model and network performance optimization. International Journal of Satellite Communications & Networking, 2016,35(3): 263-277

www.manaraa.com