Clustering Analysis on E-commerce Transaction Based on K-means Clustering

Xuan HUANG
Cognitive Science Department, Xiamen University, Xiamen, Fujian, China, 361005
Fujian Key Laboratory of the Brain-like Intelligent Systems (Xiamen University), Xiamen, Fujian, China, 361005
Economic Management Department, Zhangzhou Institute of technology, Zhangzhou, Fujian, China, 363000

Zhijun Song
The 28th Research Institute of China Electronics Technology Group Corporation, Nanjing 210007, China

***Abstract*—Based on the density, increment and grid etc, shortcomings like the bad elasticity, weak handling ability of high-dimensional data, sensitive to time sequence of data, bad independence of parameters and weak handling ability of noise are usually existed in clustering algorithm when facing a large number of high-dimensional transaction data. Making experiments by sampling data samples of the 300 mobile phones of Taobao, the following conclusions can be obtained: compared with Single-pass clustering algorithm, the K-means clustering algorithm has a high intra-class dissimilarity and inter-class similarity when analyzing e-commerce transaction. In addition, the K-means clustering algorithm has very high efficiency and strong elasticity when dealing with a large number of data items. However, clustering effects of this algorithm are affected by clustering number and initial positions of clustering center. Therefore, it is easy to show the local optimization for clustering results. Therefore, how to determine clustering number and initial positions of the clustering center of this algorithm is still the important job to be researched in the future.**

***Index Terms*—K-Means Clustering Algorithm; Electronic Commerce Transaction; Clustering Analysis**

I. INTRODUCTION

The improvement of computer network and web technology prompts the rapid development of the e-commerce. In essence, e-commerce is a commercial operation model, which makes consumers realize online shopping and commercial tenants realize the on-line trade and electronic payment between themselves. Thus in everyday transactions of e-commerce, there are a large number of transaction data, which can be automatically obtained by the Web server, and then develop into transaction database through a series of transforms to save everyday transaction data. And in transaction database, the consumers' every purchased goods, data item, can be transformed into one transaction record [1]. Moreover, a large number of purchasing information can be recorded in transaction database, if we analyze the transaction database with different data mining methods, then the potential huge commercial value can be dug out. For example, we can analyze the consumers' buying habits through discovering relationships of different goods put into the shopping basket by consumers when shopping, which can help shopkeepers to know what goods are frequently purchased by consumers at the same time, so that they can develop better marketing strategy. If we can get the often arisen goods together, which is helpful to dig out relationships between goods, and then sellers can make a selective sales [2]. Meanwhile, if consumers with similar purchasing modes can be got together, then the specific sales for specific consumers can be achieved through analyzing characteristics of each consumer, in order to improve sales profit of goods [3].

At present, the clustering analysis of e-commerce transaction is generally finished with clustering method basing on level, increment, density and grid. Meanwhile, theory of clustering method based on level is relatively simple, but it tends to cause problems when selecting or agglomerating breakpoints and can seriously affect the next work, subsequently affecting the clustering effects. It compares one renewed record in table with other records in the database in proper order for clustering method basing on increment, in order to renew the results of clustering. But the clustering efficiency is low, for relationships between the renewed objects are not fully considered when using this method. However, for density based clustering method, sampling technology and partition can be used to reply the massive data, although this method can well deal with high-dimensional data, it can inevitably produce sampling error, the reason of which is that the sampling technology is used [4-6].

It produces a large number of transaction information every day for a transaction database, which may include hundreds of data items. What's more, there are a lot of problems when transforming a transaction database into a high-dimensional database, so that many clustering algorithms cannot be well used [7-8].

The origin of the wavelet is about in the 1930s. Some substantially similar theory in different fields has been proposed. Until 1984 Morlet and Grossmann officially named it wavelet. In 1985, Mallat found the resemblance of theory, and it also developed the application scope of wavelet theory [9]. The most basic element in Wavelet theory is mother wavelet. To be the template wavelet, it needs to meet certain conditions. The physics meaning of

© 2014 ACADEMY PUBLISHER
doi:10.4304/jnw.9.2.443-450

the above two equations represents the component whose frequency of the mother wavelet is zero. That is to say the mother wavelet must have shock wave and the average on time of the mother wavelet is zero [10-11].

Wavelet function is extended from the mother wavelet. Its form is $\Psi_{a,b}(x) = 1/\sqrt{\alpha}\Psi((x-b)/a)$. For different parameters in the equation it conversely produces a new function. The kind of function $\Psi_{a,b}$ is called the wavelet, and the original function Ψ is called the mother wavelet function as before defined. It includes two kinds of operation. When the absolute value is larger than 1, it represents the mother wavelet is compressed.With the tension and compression of mother wavelet, it can conversely represent the low frequency and high frequency signal.So according to the different scale factor, and we can get the different frequency wavelet. And in the equation, b is called translation factors. In order to making the wavelet transform have the time domain analysis ability, we should make the time axis full of wavelet. Translation factor b is to make all the center of the wavelet functions shift from time axis zero arbitrarily and do the shifts of the function. Wavelet transform basically is to use the concept of wavelet function. It decomposed the original function or signal into many related small sections. These decomposed small sections have scale changes and the wavelet can weaken with time. This decomposition was called wavelet decomposition or Wavelet transformation [12]. And thewavelet reversal or wavelet reconstruction is to combine these decomposed small sections together again.

Based on the above clustering algorithm, shortcomings just like the bad elasticity. Weak handling ability of high-dimensional data, sensitive to time sequence of data, bad independence of parameters and weak handling ability of noise usually exist when facing a large number of high-dimensional transaction data. The K-means clustering algorithm based on dividing is presented in this paper, and compared with the above methods this algorithm has the strong ability of dealing with high-dimensional data. In addition, K-means clustering algorithm has very high efficiency and strong elasticity when there are many data items. In the process of clustering, this algorithm will try to find k dividing, which makes the value of square error function minimum. So it will show excellent clustering effects when the final group is concentrated and differences of groups are obvious [13-16].

First, divide the input data points into K initialized groups, then calculate the centers of each group, and next put the objects to centers in order to determine the groups again. Finally, repeat the above operation until the convergence occurs [17].Make experiments by sampling data samples of the 300 mobile phones of Taobao, the following conclusions can be obtained: compared with Single-pass clustering algorithm, K-means clustering algorithm has a high intra-class dissimilarity and inter-class similarity when analyzing e-commerce transaction; meanwhile, K-means clustering algorithm shows the obvious superiority when dealing with high-dimensional data; moreover, K-means clustering algorithm expresses a very high efficiency and the strong elasticity, it can also show the excellent clustering effects when there are obvious differences between groups [18].

II. The Proposed Approach for E-commere Transaction Analysis

In this section, we will propose an approach for E-commerce transaction analysis, based on the clustering analysis and the wavelet neural networks. The proposed approach is composed of 6 procedures: data collection, data preprocessing, network structure determination, feature exaction, clustering analysis and evaluation. The overall framework is presented in Figure 1.

A. Data Collection and Preprocessiong

Taobao has been one of the worldwide electronic trading platforms, and has a rapid development since its establishment. It is the most popular on-line shopping platform in China, having a registered user of nearly 500 million, with the number of everyday on-line products more than 800 million. Thus, it is representative to collect data from Taobao for experiments. We consider the e-commerce transaction data of mobile phones. It mainly contains the following three attributes: (1) the product attribute of mobile phones, such as the product's price, ring, time to market, smart mobile phone or not, color of mobile screen, mobile brand and so on; (2) the attribute of seller, such as the positive feedback rate for sells, registration time of the shop, the seller's credit and after-sale service; (3) the sales status, such as the sale volume of product, sale status of the current stage, the cumulative sale status etc. This paper collects a total of 300 samples of mobile phones, and each phone product item can be used as a sample [19].

For data with numeric values, K-means algorithm has an excellent clustering ability. In this paper, the collected data is preprocessed before the experiment to achieve the transformation from non-numeric data to numeric data. For example, for the registration time of sellers, the form of the collected original data is yyyy-mm-dd; then the registration days can be obtained through using time collecting data to subtract the seller's original registration time. For the phone type, 1 denotes smart phone, and 0 denotes the non-smart phone. For after-sale service of phones, 0 denotes nationwide warranty; 1 denotes three guarantees and 2 denotesthe other form of after-sale service. For the column of phone brand, brand is non-numeric data, so they can be displayed as tags, and no attribute name. The processed data is summarized in Table 1 [11].

B. Feature Via Wavelet Neural Network

The concept of wavelet neural network is that it put forward new network architecture, and the functional of network neurons is achieved by the wavelet transform, which makes the wavelet decomposition and neural network fuse together.

From the wavelet theory, it is known that the function which is made o L2(R) can use projection quantity to express. And the projection quantity was gotten by the functions and the wavelets. This is also the concept of discrete wavelet decomposition. At the same time it

© 2014 ACADEMY PUBLISHER

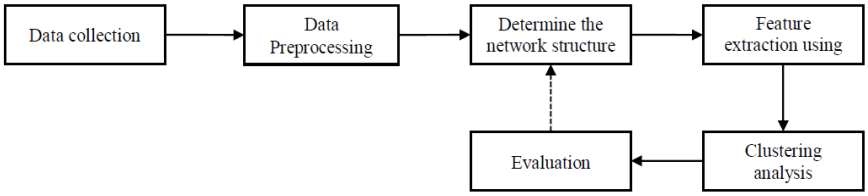

Figure 1. The procedure of the proposed approach for e-commerce transaction

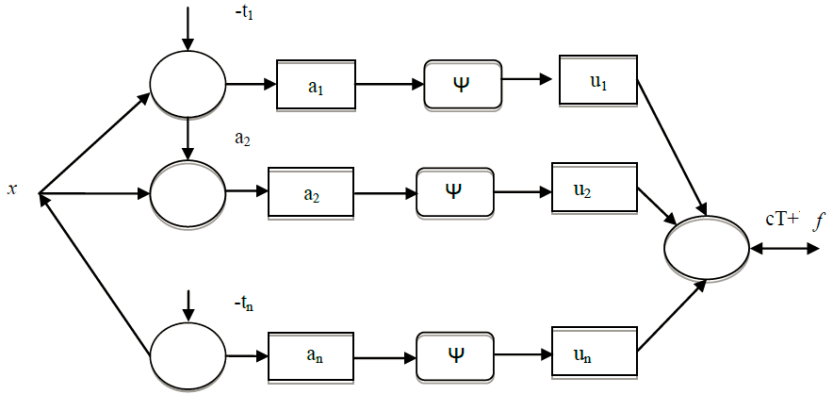

Figure 2. The frame of wavelet neural network

usually chooses the wavelet function which is supported by the time domain and frequency domain. Thus they can make the wavelet function to have time-frequency localization properties. As wavelet has the localization time-frequency character, so arbitrary function can use the intercepted discrete wavelet decompose to approach. For this kind of approximation function, it generally will choose the wavelets which have good distributions and rules in the time-frequency domain. However, for wavelet neural networks, the type of wavelet function depends on the characteristics of network training data. It is not limited by the limited or countable set. So there are many decision ways of wavelet. Wavelet neural network is built by the limited wavelet.

Wavelet can be used for the nonlinear regression and the parametric nonlinear functionfamily. And these parameters can be gotten by the network training process. Usually the decision of the parameters in wavelet neural network uses the back-propagation learning algorithm, gradient descent method and evolutionary procedure of learning algorithm is also often widely used in. When the input or output data find a best fit, which can make the study error minimum, so the parameters that reflect the weights in wavelet neural network were decided. Wavelet neural network combined wavelet transform with RBF network structure, and its structure is shown in Figure 2. In Figure 2, the same straight line that is composed by translation ($-t_i$), dilation (D_i) (dilation, can also be called scale) and wavelet(ψ, i) operator, is called as wavelon, which is similar with the neuron in the neural network.In some cases, the approach function may have part linear characteristics, so in the wavelet neural network we often use linear combination (cT+b) which parallel with the network to improve the accuracy of approach. So the structure of the wavelet neural network can be expressed as followings:

$$f(x) = \sum_i w_i \alpha_i^{-\frac{d}{2}} \Psi\left(\frac{x-t_i}{a_i}\right) + c^T x + b \qquad (1)$$

where Ψ is the mother wavelet. It is usually chosen in accordance with the best space base. We often select the wavelet which has radial, smoothness and considering the time domain and frequency domain.The computational procedure is summarized in Section 3.3.

C. Clustering Analysis Using K-means

Based on the analysis of the shortage about the traditional methods such as wavelet neural network in the analysis of e-commerce transactions, it provided a K-means clustering method.

Clustering is the process of dividing data objects into classes or groups, in order that objects in the same groups have very high similarity and objects in different groups have very high dissimilarity. Clustering stems from many research fields, such as biology, engineering, statistics and so on. Clustering analysis is one of the important contents of data mining, and has been widely used in many research fields, mainly including image processing, pattern identification, market research, data analysis and so on. In commercial affairs, clustering analysis can help market analysts to discover different consumer groups from consumer base according to models, in order to depict characteristics of consumer groups. In biology, clustering can be used to infer classifications of plants and animals, that is to say, classify genes according to the similar function to get cognizance of inherent structure in groups. Meanwhile, clustering can be also used to classify texts in Web.

Characteristics of clustering can be denoted by a triple $CF = (N, LS, SS)$, where N is the quantity of text vector in clustering category. $\mathrm{LS} = \sum_{i=1}^{N} x_i$ is the algebraic sum of N text vectors, and is a vector. $\mathrm{SS} = \sum_{i=1}^{N} x_i' \; x_i$ is the quadratic sum of N text vectors, and SS is a value. One

© 2014 ACADEMY PUBLISHER

TABLE I. DATA FORMATS OF MOBILE PHONE E-COMMERCE TRANSACTION

	Price	Current sales	Totalsales	Total product	Producttime	Screen color	After sale	Smart	Seller credit	Seller Evaluation
Samsung	2500	7	10	200	2012	45	0	1	1345	97
Nokia	1510	3	5	150	2009	1600	0	0	357	87.5
Motorola	1970	9	12	120	2011	673	0	1	3223	88
Sony	1960	3	5	40	2011	120	2	0	342	78.9
Changhong	980	1	2	38	2012	34	2	0	1368	77.4
Coolpad	590	0	2	20	2010	56	1	0	3572	73.0
HTC	2100	7	10	160	2013	24	0	1	456	95.4
ZET	1200	23	25	930	2012	120	1	1	247	80.7
Blackberry	3900	14	15	745	2013	80	0	1	1790	90.3

center of clustering category can be obtained according to the following formula, as shown below:

$$\bar{x} = LS/N \quad (2)$$

The characteristics of clustering has additivity, so we can assume that A and B are respectively two lines, they also are sub-clustering to be merged. Assume that clustering characteristics of the two sub-clustering are respectively shown as $CF_A = (N_A, LS_A, SS_A)$ and $CF_B = (N_B, LS_B, SS_B)$.

Then the merged clustering characteristics can be expressed as follows:

$$CF_C = CF_A + CF_B = (N_A + N_B + LS_A + LS_B + SS_A + SSB \quad (3)$$

Characteristics of a clustering and its sub-clustering have additivity [10].

K-means clustering algorithm stems from the middle of 1950s, the iterative improved heuristics is one of the most common forms in the algorithm. For the iterative improved heuristics, first, divide the input data points into k initialized groups, and these data points can be random or heuristic data. Then calculate the center of each group. Next, put the objects to centers having shortest distance according to the centers' positions, so that the groups can be determined again. Finally, repeat the above operation until the convergence occurs.

K-means algorithm, which is one technology basing on center of mass, takes k as the input parameter, than divide n data point object sets into k groups. The purpose of clustering is to make the inter-group similarity highest, but the intra-group similarity lowest. Similarity of groups can be measured by mean values of objects in groups, which can be deemed as the center of group. The processing procedure of K-means clustering algorithm is described as below. First, we randomly select K objects, which can be deemed as the initial mean value of each group. Then compare the rest objects with the distances of mean values of original groups, each object will be assigned into group having the highest similarity. Next, calculate the new mean value of the new group. Repeat the above process until the criterion function converges. And the function can be described as:

$$J_c(m) = \sum_{j=1}^{k} \sum_{X \in C} |X_i - Z_j|^2 \quad (4)$$

where $J_c(m)$ is the total sum of square error of all objects in the database, $X_j$is data point in the space, expressing the given data object. Z_j is the mean value of group C_j. Meanwhile, X_j and Z_j are both multi-dimensional data.

The overall procedure of K-means algorithm is shown as Figure 3. It is worth noting that: (1)K-means value. K-means value algorithm divided by users, centers of each group can be denoted by mean values of objects of groups. (2) The input is that K is the numbers of group; D is the number set including n objects. (3) The output is the set of K groups.

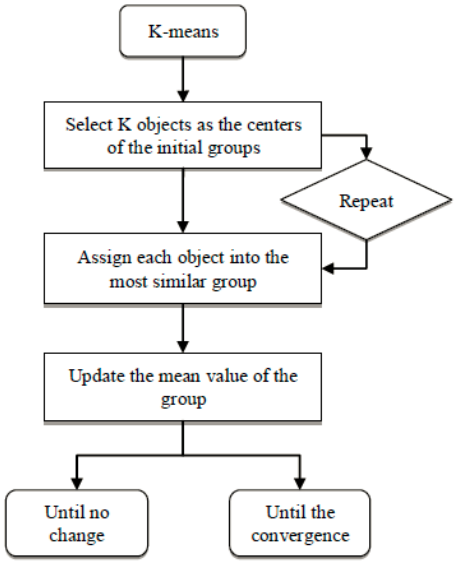

Figure 3. The flowchart of K-means algorithm

III. EXPERIMENTAL ANALYSIS

To evaluate the proposed approach for e-commerce transaction analysis, we design a group of numerical experiments. The over experiment procedure is illustrated in Figure 1. It is composed of 6 procedure, data collection, data preprocessing, network structure determination, feature extraction, clustering analysis and evaluation.

A. Obtain the Data Source

Taobao is the most popular on-line shopping platform in China, having a registered user of nearly 500 million.The number of everyday on-line products has been more than 800 million. In this experiment, it has a strong representativeness to select data from Taobao. We used the e-commerce transaction data of mobile phones for experiment, which mainly includes the following three attributes. (1) The first attribute is product attribute of mobile phones, such as the product's price, ring, time

© 2014 ACADEMY PUBLISHER

to market, smart mobile phone or not, color of mobile screen, mobile brand and so on. (2) The second attribute is the attribute of seller, such as the seller's positive feedback rate, registration time of the shop, the seller's credit and after-sale service. (3) The third attribute is sales status, such as the product's sale volume, sale status of the current stage, the cumulative sale status and so on. This paper takes mobile phones on Taobao as an example, collecting a total of 300 samples; each phone product item can be used as a sample [11].

B. Pretreatment of Data

For numeric values, K-means algorithm has excellent clustering ability. The collected data is pretreated before experiment to achieve the transformation from non-numeric data to numeric data. For example, for the seller's registration time, the collected original data form is yyyy-mm-dd.Then the registration day can be obtained through using time collecting data to subtract the seller's original registration time. For that the phone is smart phone or not, 1 expresses smart phone, and 0, the non-smart phone. For after-sale service of phones, 0 expresses nationwide warranty, 1 expresses three guarantees and 2 expresses the other after-sale service forms. For the column of phone brand, brand is non-numeric data, so they can be displayed as tags, and no attribute name. The treated data form is shown as Table 1 [11].

C. Experimental Process

Based on the thought of K-means algorithm, the object sets of e-commerce transaction data of 300 phones can be deemed as input to be clustered, in order to get K clustering center Z_j and object sets C_j of clustering data. K objects can be randomly selected from data sets as the center of initial group, then assign each object to the most similar group according to the object's mean value of the group, and update the group's mean value, calculating the object's mean value of each group. Repeat the above steps until there is no change for the number of group.

When using wavelet neural network to extract feature, its basic process is shown in Figure 4. First of all, the following parameters are needed to be sure:

(1) Network structure parameters. As forthe number of wavelets, it can use the standard model selection criteria in statistical such as Akaike's Final Prediction Error Criterion (FPEC) or Schwarz Bayesian Information Criterion (BIC) and Akaike Information Criterion (AIC) to decide. In the sequence prediction, in order to avoid overfitting, we often set the number of the wavelet as 1.

(2) Initialization parameters. Such as initializing the scale level of the training sample space and the input sample size of each wavelet. The decisions of the two parameters usually use the thumb rule. The scale level generally was set 4. The input sample size of every little wavelet is usually equal to the number of variables plus 2.

(3) Learning parameters. Such as the largest iterative numberand the end condition, the wavelet in the network will be selected according to the initial parameters and input/output data sample.And then after the regressive selection of the wavelet, the initial wavelet will be built.The flow of the wavelet neural network is as the Figure 4.

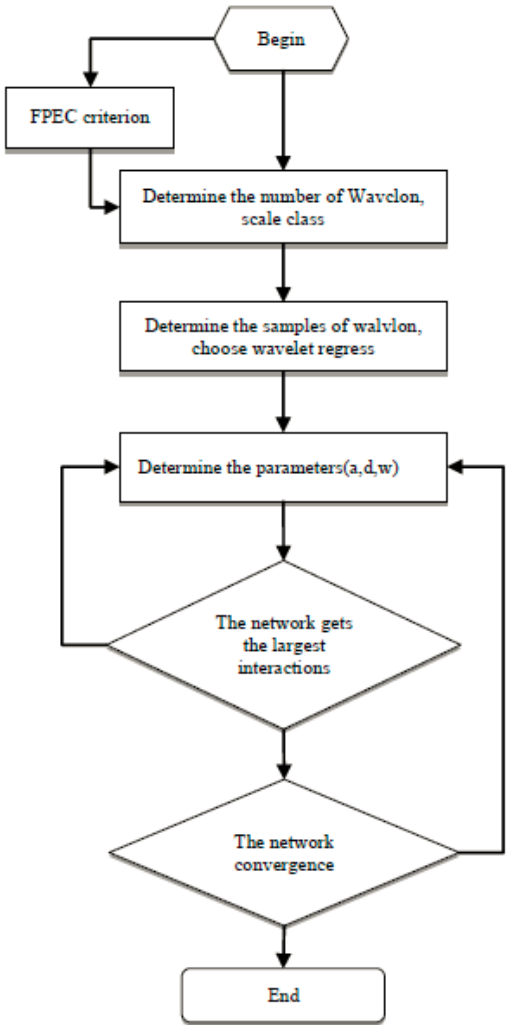

Figure 4. The flow chart of wavelet neural network

We can use Quasi-Newton's algorithmto calculate the weights in the network. After the repeated learning process, the differences between the calculation output value and the actual measurement output value will be decreased. So the construction of wavelet neural network model is finished, which can be used as a simulation or prediction tool. However, in the model construction process, because the input value and output value are the same sequence. They can be expressed as follows:

$$y(t) = f\big(y(t-n_k), K, y(t-n_k-n_\alpha+1)\big) \quad (5)$$

The choice of model order is always very important. For the forecasting of processing time sequence, then $_K$ in the above equation represents the time delay, n_a represents input parameters size of the clusteringneural network. And both the parameters should be positive integer and the minimum value is 1.

The traditional forecasting model can be used to solve the level selection problems. It is distinguished by the graphics of subjective observation autocorrelation and partial autocorrelation function.But for the building of wavelet neural network Prediction model; we take another kind of method to calculate the loss function of

© 2014 ACADEMY PUBLISHER

each level first. Then we decided the ideal level n_a based on the Akaike's Final Prediction Error Criterion (FPEC).

D. Experimental Results

In this paper, the data of 300 phones selected on Taobao is used as data source to make experiment. Taobao, which is one of the most influential e-commerce transaction platforms, is one of the worldwide electronic trading platforms and has a rapid development since its establishment. Therefore, in this experiment, it has a strong representation to select data from this website. Meanwhile, the e-commerce transaction data of phones can be mainly divided into three attributes by analyzing, product attribute, seller attribute and sale status, in order to make experiments. What's more, in this experiment, the two parameters, the average dissimilarity and intra-class similarity, are used to estimate the effects of clustering. The two parameters are shown as follows:

$$\mathrm{H} = -\frac{1}{n\log C}\sum_{i=1}^{n}\sum_{j=1}^{K} u_{ij}\log u_{ij} \qquad (6)$$

where *H* is one evaluating indicator to measure various intra-class dissimilarity. The smaller the *H*, the more indistinct the two categories are, and the worse the classification effects are. *k* is clustering number.

$$\mathrm{F} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{K} {u_{ij}}^2 \qquad (7)$$

where *F* is the similarity of inter-class elements, the larger *F* is, the better classification effects is [12].

With the clustering process introduced in the above sections, it made experiments by the number *K* of input group and data set *D*, then make a contrastive study with Single-pass clustering algorithm. Finally, respectively select 10 clustering results with better clustering effects after respectively repeating experiments using the two clustering algorithms for 20 times. The results are shown in Table 3 and Table 4. It also respectively used the K-means clustering method and single-pass algorithm to do the simulation. In this paper it used dissimilarity degree between the classes and the similarity in a class as the evaluation standards. At last it also make a comparison about the average training time in the 20 rounds, and the results can be seen in Table 5.

In the first experiment, we evaluate the proposed approach for e-commerce transaction analysis, where the wavelet neural network is used to extract features and K-means algorithm is employed to analysis the data. The results are reported in Table 2. We can see the K-means clustering results have higher dissimilarity degree and similarity. We can see the highest similarity in a class can be as high as 98.71%, and the lowest can be 63.48%. The difference between the two classes is concentrate upon 65%. Therefore, we can say that the K-means clustering algorithm can well applied to the e-commerce transactions data analysis.The reasons may come from the three aspects. Firstly, K-means clustering algorithm can effectively overcome the poor scalability and ability, and deal with high-dimensional data in facingof massive and high dimensional, but in facing of vast amounts of e-commerce transactions data it often appears high dimensional data with poor scalability features. Second, for e-commerce transactions data it has very strong time sequence, and K-means clustering algorithm has the characteristic that it is not sensitive to time order, therefore this makes the K-means clustering algorithm having the good clustering effect for the e-commerce transaction data. Finally, because K-means clustering algorithm has strong independent parameters characteristic,this will also reduce the effect brought by the time order of the commerce transaction data [6].

TABLE II. RESULTS OF CLUSTERING OF K-MEANS FOR 10 TIMES

Running times	Dissimilarity	Similarity
1	0.6439	0.8795
2	0.8532	0.7690
3	0.3673	0.6945
4	0.5986	0.6348
5	0.7432	0.6598
6	0.6532	0.9871
7	0.6598	0.8654
8	0.6531	0.7945
9	0.6598	0.6982
10	0.7459	0.9687
Means	0.6589	0.8639

In the second experiment, we use the single-pass clustering method to perform the cluster,where the wavelet neural network is used to extract features and K-means algorithm is employed to analysis the data. The over experimental procedure can be found in Figure 1. The parameters are presented in Section C. Though this experiment we can see the effectiveness of the method. we use the single-pass method to do the cluster. The over experimental procedure can be found in Figure 1. The parameters are presented in Section C. And then we select 10 clustering results with better clustering effects after repeating experiments with the clustering algorithms for 10 times, as shown in Table 3. The experiment parameters are the cluster number K and the critical value of the distance. The average similarity is 56.28%. Besides, the average dissimilarity is only 24.29%, which is much lower than that in Table 3. When dealing with the e-commerce transactions data, the K-means method showed great advantage over the Single-pass method. The reasons counting for these are twofold. Firstly, the Single-pass algorithm is sensitive to the time order of the data, which is the characteristic about the e-commerce transactions data. The second is it is hard to determine the parameters about the clustering algorithm. The method has poor parameter independence.

TABLE III. RESULTS OF CLUSTERING OF SINGLE-PASS ALGORITHM FOR 10 TIMES

Running times	Dissimilarity	Similarity
1	0.4438	0.3792
2	0.2531	0.2691
3	0.2673	0.6944
4	0.1982	0.4342
5	0.2436	0.5596
6	0.4531	0.3877
7	0.3598	0.5651
8	0.4531	0.6942
9	0.1598	0.3989
10	0.6459	0.5681
Means	0.2429	0.5628

© 2014 ACADEMY PUBLISHER

In the third experiment, we evaluated the time cost of the proposed approach for the e-commerce transaction analysis. The over experimental procedure can be found in Figure 1, where the wavelet neural network is used to extract features and K-means algorithm is employed to analysis the data. The parameters are presented in Section C. We make a comparison about the average training time in the 20 rounds, and the results can be found in Table 4.Though this experiment we can identify the effectiveness of the method in this paper. The experiment used the Single-pass and K-means methods to do the comparisons. And the experiment repeated 20 times. The experiment parameters are the cluster number K and the critical value of the distance. The experiment results can be seen in Table 4. In the Table 4, we use T to express the training time. From the result we can see the average training time is 12.38s and 8.54s. The training time in this paper has an acceleration of 4s. The reasons for thesemainly are because the K-means has the parameter dependence and it is not sensitive to the time.Besides, in the process of clustering, this algorithm tries to find the K partitions making the square error function have the minimum value. Therefore, this algorithm can show the excellent clustering effects when the resulted groups are concentrated and there is no obvious difference between groups.

The results of the above three experiments show that: (1) in comparison with single-pass clustering algorithm, K-means clustering algorithm has a very high inter-class dissimilarity and intra-class similarity when performing clustering analysis of e-commerce transaction; (2)the K-means clustering algorithm has an obvious superiority indealing with high-dimensional data;(3) the K-means clustering algorithm puts up a high efficiency, strong elasticity and the excellent clustering effect when there are obvious differences between different crowds [6].

TABLE IV. RESULTS OF THE TRAINING TIME

Running time (s)	Single-pass	K-means
1	13.21	11.02
2	13.69	9.32
3	10.47	7.36
4	14.23	7.69
5	10.47	8.36
6	14.25	9.56
7	13.63	9.32
8	10.23	7.12
9	12.36	6.97
10	11.20	8.69
Means	12.374	8.541

IV. CONCLUSIIONS

Based on density, increment, grid, shortcomings just like the bad elasticity, weak handling ability of high-dimensional data, sensitive to time sequence of data, bad independence of parameters and weak handling ability of noise are usually existed in clustering algorithm when facing a large number of high-dimensional transaction data. K-means clustering algorithm based on dividing is presented in this paper. First, divide the input data points into K initialized groups, then calculate the centers of each group, and next put the objects to centers have the shortest distance to itself, in order to determine the groups again. Finally, repeat the above operation until the convergence occurs.Make experiments by sampling data samples of the 300 mobile phones of Taobao, the following conclusions can be obtained: compared with Single-pass clustering algorithm, K-means clustering algorithm has a high intra-class dissimilarity and inter-class similarity when analyzing e-commerce transaction; meanwhile, K-means clustering algorithm shows the obvious superiority when dealing with high-dimensional data; moreover, K-means clustering algorithm expresses a very high efficiency and compared with the traditional clustering algorithms, and K-means clustering algorithm presented in this paper has the strong ability of dealing with high-dimensional data. In addition, K-means clustering algorithm has very high efficiency and strong elasticity when dealing with a large number of data items.

ACKNOWLEDGMENT

This work was partially supported by National Nature Science Foundation of China (No. 61202143), the Nature Science Foundation of Fujian Province (No. 2011J01367), Xiamen University 985 Project and research funds of Zhangzhou institute of technology (ZZY1307).

REFERENCES

[1] Chen An, Chen Ning. The secondary clustering algorithm of transaction database in electronic business. *Computer science*, 29(8) pp. 126-128, 2001

[2] Zhang Zhushan. Hot topic detection on forum basing on clustering analysis. *Master's thesis of Harbin Institute of Technology*, 2010.

[3] Oussar, Y., I. Rivals, L. Personnaz, G.. "Trainingwavelet networks for nonlinear dynamic input-output modeling", *Neurocomputing*, Vol. 19, No. 20, pp. 173-188, 1998

[4] Pan, Zuohong, A Stochastic Nonlinear Regression Estimator Using Wavelets, *Computational economics*, Vol. 21, No. 11, pp. 90-102, 1998

[5] Strang, Gilbert, Wävelet and filter banks, *Wellesley-Cambridge Press*, JSA, 1996

[6] Kisi, O. ''Wavelet regression model for short-term stream-flow forecasting'', *Journal of Hydrology*, Vol. 46, No. 389, pp. 344-353, 2010

[7] Kisi, O. ''Wavelet regression model as an alternative to neural networksfor monthly streamflow forecasting'', *Hydrological Processes*, Vol. 23, No. 127, pp. 3583-3597, 2009

[8] Kisi, O. and Cimen, M. ''A wavelet-support vector machine conjunctionmodel for monthly streamflow forecasting'', *Journal of Hydrology*, 399, pp. 132–140 2011.

[9] Daubechies, I. ''The wavelet transform, time-frequency localizationand signal analysis'', *IEEE Transactions on Information Theory*, Vol. 36, No. 5, pp. 961-1005, 1990

[10] Huang Yudong, Li Xiang, Lin Xiang. Research and application of active discovery technique of the focus of internet media information. *Computer technique and development*, Vol. 19, No. 5, pp. 1-5, 2009

[11] Li Gang, An Lu. Clustering analysis of transaction of mobile phone electronic business basing on SOM. *Information analysis and research*, Vol. 169, No. 9, pp. 70-77, 2008

© 2014 ACADEMY PUBLISHER

[12] Lu Mingyu, Yan Xiaona, Wei Shanling. Excavation of hot topic on forum basing on fuzzy clustering, Vol. 34, No. 3, pp. 52-55, 2008
[13] Jun Chen, YueshengGu, Yanpei Liu, Grid Service Concurrency Control Protocol, *Journal of Networks*, Vol. 7, No. 4, pp. 707-714, 2012
[14] Min Zhao, Tao Zhang, FangbinGe, Zhijian Yuan, RobotDroid: A Lightweight Malware Detection Framework On Smartphones, *Journal of Networks*, Vol. 7, No. 4, pp. 715-722, 2012
[15] Xiaobo Wang, Xianwei Zhou, Junde Song, Hypergraph based Model and Architecture for Planet Surface Networks and Orbit Access, *Journal of Networks*, Vol. 7, No. 4. pp. 723-729, 2012
[16] Jinliang Wan, Yanhui Liu, Multi-Regions Texture Substitution, *Journal of Multimedia*, Vol. 7, No. 6, pp. 394-400, 2012
[17] Wensi Cao, Jingbo Liu, A License Plate Image Enhancement Method in Low Illumination Using BEMD, *Journal of Multimedia*, Vol. 7, No. 6, pp. 401-407, 2012
[18] Tao Gao, Ping Wang, Chengshan Wang, Zhenjing Yao, Feature Particles Tracking for Moving Objects, *Journal of Multimedia*, Vol. 7, No. 6, pp. 408-414, 2012

© 2014 ACADEMY PUBLISHER