Send Orders for Reprints to reprints@benthamscience.ae

1258 *The Open Cybernetics & Systemics Journal,* **2015,** *9,* **1258-1261**

Classification Optimization Clustering Model Simulation Based on User Interest

Jinhua Zhu*

College of Network Communication, Zhejiang Yuexiu University of Foreign Languages, Shaoxing, China

Abstract: Clustering analysis was carried out on the user's interests is of great significance to the study of consumer psychology. Considering user's interests is a kind of classification optimization clustering model, improve the user's interests using the algorithm of ID3 decision tree classification calculation speed, the attribute of the highest information gain as the test attributes of nodes before, to ensure the result of decomposition users interested in samples required minimum amount of information, building user interest classification optimization of adaptive fuzzy clustering objective function, the update matrix clustering prototype, under adaptive fuzzy clustering model, clustering prototype iterative equation is given directly, guarantee the accuracy of the classification. Experiment result shows that the proposed model is compared with traditional clustering model is not easy to fall into local optimal solution, has higher recall ratio and precision, and has great significance for further user behavior research.

Keywords: Classification, clustering model, optimization, simulation, user interest.

1. INTRODUCTION

User classification [9, 11] is conducted in the mass data. The data in the classes is of similarity operation to get a decision scheme, in which the similarity between discrepant data can be neglected, which is of great significance in reducing the operand of decision scheme and improving the real time of decisions scheme. The application foundation of this process is the clustering analysis problem of users [1-3]. Clustering analysis is an unsupervised learning method and based on the principle of "things of one kind come together" [10]. The individuals of one group are divided into several classes by similarity to minimize the individual difference of same class and maximize the individual difference of different class. The scoring standard for different users is different and the scoring of the same commodity by different users is different under the different standard. The traditional clustering method [4] is based on the user's score of commodity, which cannot effectively describe user's interest in the commodity and easily form inevitable personal error, with such defects as slow convergence speed and low precision.

A classification optimization clustering model considering user interest is put forward in this paper, with high recall ratio and precision ratio.

2. CLASSIFICATION OPTIMIZATION CLUSTERING MODEL BASED ON USER INTEREST

2.1. Rapid Classification Optimization of ID3 Decision Tree Algorithm Integrating User Interest

ID3 decision tree algorithm is used to realize the classification optimization of user interest to provide a reliable basis for building the classification optimization clustering model based on user interest [6]. ID3 algorithm is constructed by Quinlan for summarizing the classification model in the data, which is also called decision tree. The information gain is adopted on each node of tree to measure and choose the test attribute and the attribute with the highest information gain is found as the test attribution of former node. The attribute leads to the minimum information quantity required by sample classification of user interest in the result decomposition [12]. At the same time, it can describe the minimum randomness of decomposition.

Set S represents the set of s user interest samples. If class label attribute has m different values, there are m different classes $G_i(i=1,\cdots,m)$. Attribute A is used to decompose S into v subsets $\{S_1, S_2, \cdots S_V\}$, in which S_j includes the samples in S and there is a_j on A. If A is used as the test attribute, these subsets correspond to the branches grown from nodes including set S. Set s_{ij} is the number of user interest samples of class C_i in subset S_j, the expected information of subset divided from A is shown in equation (1):

$$E(A) = \sum_{j=1}^{v} \frac{s_{1j} + \cdots + s_{mj}}{s} I\left(s_{1j}, \cdots, s_{mj}\right) \quad (1)$$

$s_{1j} + \cdots + s_{mj}$ is the weight of j subset and equal to the number of user interest samples in the subset divides the central samples in S. The smaller the entropy is, the higher the purity of subset decomposition will be. The code message of branches on A is:

$$Cain(A) = I(s_1, s_2, \cdots, s_m) - E(A) \quad (2)$$

*Address correspondence to this author at the College of Network Communication, Zhejiang Yuexiu University of Foreign Languages, Shaoxing, China; Tel: 15967508105; E-mail: uuem@163.com

1874-110X/15

where, $Cain(A)$ is the expected compression of entropy as the value of attribute A is known.

Improvement is made according to the rule generation method, that is, attribute selection standard algorithm. By weighting and increasing the interest parameters, the important attribute annotation is enhanced, the unimportant attribute annotation is reduced and the weight sum is converted into the sum of weight sum and interest parameter to not cover the small number of data upon the generation of decision tree so as to reduce the dependence of decision tree on attribute with more values. Equation (1) is modified as:

$$E_1(A) = -\sum_{j=1}^{v}\sum_{i=1}^{m}(\sum_{i=1}^{m} s_{ij} / s + Q)\frac{s_{ij}}{|s_j|}\log_2(\frac{s_{ij}}{|s_j|}) \tag{3}$$

Correspondingly, equation (2) is converted into:

$$Cain(A) = I(s_1, s_2, \cdots, s_m) - E_1(A) \tag{4}$$

The main idea of the optimization algorithm in this paper is to reduce the operation complexity of equation (5), so

$$E_1(A) = -\frac{1}{S\ln 2}\sum_{j=1}^{v}\sum_{i=1}^{m}(\sum_{i=1}^{m} s_{ij} + SQ)\frac{s_{ij}}{|s_j|}\ln(|S_j| - \sum_{k=1}^{m} s_{kj} / |S_j|) \tag{5}$$

where, s_{ij} is the number of samples of class C_i in subset S_j, $|S_j|$ is the total number of samples of all classes in S_j and $\sum_{k=1}^{m} s_{kj}$ is the total number of samples of all the classes except class C_i in S_j.

Set $a = -\sum_{k=1}^{m} s_{kj} / |S_j|$, so there is a real number β. When $\ln(1+a) < \ln(1+\beta)$ and a and β are very small, $\ln(1+a) \approx a < \ln(1+\beta) \approx \beta$, so it is clear that the entropy or expected information of subset divided from B is:

$$E_1'(B) = -\frac{1}{S\ln 2}\sum_{j=1}^{v}\sum_{i=1}^{m}(\sum_{i=1}^{m} s_{ij} + SQ)\frac{s_{ij}}{|s_j|}(\beta) \tag{6}$$

when $E_1'(A) < E_1'(B)$, $E_1(A) < E_1(B)$ and equation (4) can be converted into:

$$Cain(A) = I'(s_1, s_2, \cdots, s_m) - E_1'(A) \tag{7}$$

It greatly saves the time of rapid classification optimization of ID3 algorithm introducing user interest parameter because after optimization, $\ln(S_{kj} / |S_j|)$ is replaced by $-\sum_{k=1}^{m} s_{kj} / |S_j|$ to greatly simplify the complexity of user interest classification and provide a reliable basis for building the classification optimization clustering model based on user interest [5].

2.2. Classification Optimization Self-adaptive Fuzzy Clustering Objective Function Considering User Interest

Considering the above user interest classification optimization results and integrating reasonable optimization conditions, the classification optimization self-adaptive fuzzy clustering objective function [7, 8] can be set as shown in equation (8) and two optimization constraints are given as shown in equation (9) and (10).

$$J_e(X, U, V, W) = \sum_{j=1}^{N}\omega_j^p\sum_{i=1}^{c}u_{ij}^m d_{ij}^2 \tag{8}$$

$$\text{s.t. } \sum_{j=1}^{n}u_{ij} > 0, \quad \sum_{i=1}^{c}u_{ij} = 1 \tag{9}$$

$$\coprod_{j=1}^{n}\omega_j = 1 \tag{10}$$

where, J_e is the objective function of clustering model; a group of new parameters ω_j are introduced into J_e, called fitness of user interest x_j; ω_j constitutes a vector *W*, called fitness vector; p is an adaptive index and a preset constant to adjust the adaptive value. Equation (10) shows that the adaptive ω_j has the constraint of "plus 1", which describes the intrinsic relationship between user interest classes.

The solving of clustering objective function is a nonlinear optimization problem [7] and it can be solved by Lagrange extremum method. On the premise of considering two constraints, the optimization function can be written as:

$$J_{e,\varphi 1,\varphi 2}(X, U, V, W) = \sum_{j=1}^{n}\omega_j^p\sum_{i=1}^{c}u_{ij}^m d_{ij}^2 + \varphi 1(\sum_{i=1}^{c}u_{ij} - 1) + \varphi 2(\coprod_{j=1}^{n}\omega_j - 1) \tag{11}$$

where, $\varphi 1$ and $\varphi 2$ are two Lagrange multiplication operators, corresponding to the fuzzy membership and adaptive constraints respectively.

The partial derivative of u_{ij} and ω_j is obtained on both sides of equation (11).

$$\begin{cases} \dfrac{\partial J_{AFCM,\varphi 1,\varphi 2}}{\partial u_{ij}} = m \bullet \omega_j^p \bullet u_{ij}^{m-1} d_{ij}^2 + \varphi 1 \\ \dfrac{\partial J_{AFCM,\varphi 1,\varphi 2}}{\partial \omega_j} = p \bullet \omega_j^{p-1} \bullet \sum_{i=1}^{c} u_{ij}^m d_{ij}^2 + \varphi 2(\coprod_{k=1,k\neq j}^{n}\omega_k) \end{cases} \tag{12}$$

In solving the optimal solution of J_e, the clustering prototype matrix $V(c \times q)$ should be updated. Under the adaptive fuzzy clustering model, the clustering prototype iterative equation is given directly as shown in equation (13). $\overline{v}_i$ is the *i* clustering prototype of classification optimization adaptive fuzzy clustering model considering user interest.

$$\overline{v}_i = \frac{\sum_{j=1}^{n} u_{ij}^m x_j}{\sum_{j=1}^{n}\overline{u}_{ij}^m} \tag{13}$$

Table 1. Indexes of objective function of two models in 5 tests.

Model		1	2	3	4	5
Presented model	max	1528	1449	1545	1530	1356
	mean	1260	1140	1221	1036	1002
	Los	No	No	No	No	Yes
Traditional model	max	1203	883	1546	1259	1459
	mean	995	658	1209	1002	1224
	Los	Yes	Yes	No	Yes	No

where, $\overline{u}_{ij}$ is the fuzzy membership of adaptive model and its computational formula is:

$$\overline{u}_{ij}^{m} = \omega_j u_{ij}^{m} \tag{14}$$

Based on equation (13) and (14), the new adaptive clustering prototype iterative equation of user interest classification optimization can be obtained as shown in equation (15):

$$\overline{v}_i = \frac{\sum_{j=1}^{n} \omega_j u_{ij}^{m} x_j}{\sum_{j=1}^{n} \omega_j u_{ij}^{m}} \tag{15}$$

The above analysis adopts ID3 decision tree algorithm, in which the attribute with the highest information gain is taken as the test attribute of current node to ensure minimal information quantity required by user interest sample classification in the result decomposition and greatly simplify the complexity of user interest classification. The adaptive fuzzy clustering objective function of user interest classification optimization is built and the clustering prototype matrix is updated. Under the adaptive fuzzy clustering model, the iterative equation of clustering prototype is given directly and the classification optimization clustering model considering user interest is obtained.

3. EXPERIMENTAL ANALYSIS

To verify the clustering results of presented model and traditional model, the films marked by each user on a film page are divided into three parts by time: the former 70% film data are used as the training set of user's stable interest model; the intermediate 10% data are used to simulate the user's browsing behavior and build user's immediate interest model; and the last 20% data are used as the test set.

Table **1** gives the maximum value of objective function of presented clustering model and traditional clustering model in 5 tests and the mean value of global objective function in the iteration and whether there is local optimal solution (LOS). It is known from Table **1** that the indexes of the presented model are superior to the traditional clustering model and the probability that the presented model falls into local optimal solution is much lower than the traditional model.

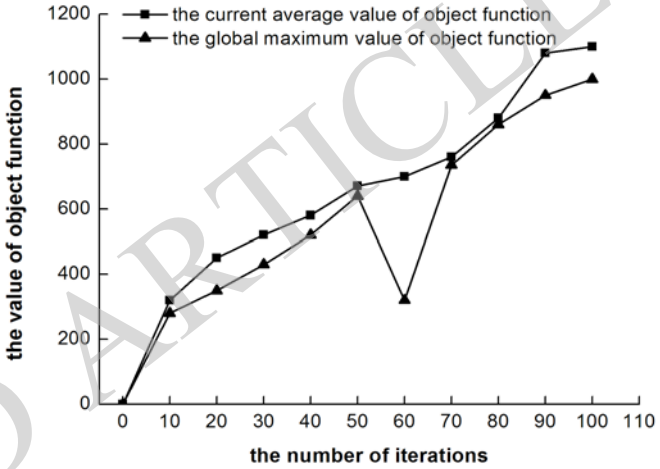

Fig. (1). Changes of maximal fitness of presented model.

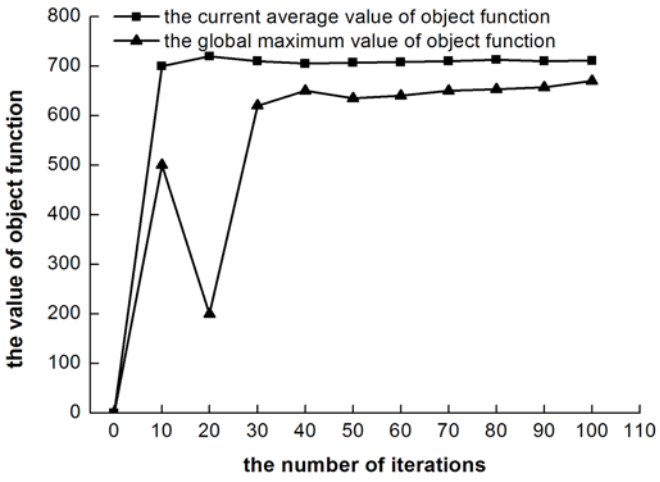

Fig. (2). Changes of maximal fitness of traditional model.

From Figs. (**1** and **2**), the presented clustering model is not easy to fall into local optimal solution compared with the traditional clustering model. When the presented clustering model doesn't fall into the local optimal value and the global optimal value remains unchanged, the presented clustering model can quickly get out of local optimal value. In Fig. (**2**), the traditional clustering model has a slow convergence speed and when the global optimal solution remains unchanged, it cannot be improved and get out of local optimal value.

The experimental results show that the precision ratio and recall ratio of the clustering results of two models decrease with the increase of number of users and the precision ratio and recall ratio of presented model are superior to the traditional model. The presented model has a stable downtrend and the traditional model has obvious fluctuation.

CONCLUSION

A classification optimization clustering model considering user interest is presented in this paper. According to the requirements of system on user clustering, ID3 decision tree algorithm is used to improve the computation speed of user interest classification and the attribute with the highest information gain is taken as the test attribute of former node to minimize the information quantity required by the user interest sample classification in the result decomposition. The adaptive fuzzy clustering objective function of user interest classification optimization is built and the clustering prototype matrix is updated. Under the adaptive fuzzy clustering model, the iterative equation of clustering prototype is given directly and the classification optimization clustering model considering user interest is obtained. The experimental results show that the presented model is not easy to fall into the local optimal solution compared with the traditional clustering model and have higher recall ratio and precision ratio.

CONFLICT OF INTEREST

The author confirms that this article content has no conflict of interest.

ACKNOWLEDGEMENTS

This work is supported by Zhejiang Province Education Department projects (No. Y201330252).

REFERENCES

[1] H. Zhang, X. Zhu, J. Zhao, *et al.* "A user clustering-based optimized filtering strategy," *Computer Systems & Applications*, no. 11, pp. 95-98, 2008.

[2] M. Huang, W. Ni, J. Wang, *et al.* "Clustering-oriented logarithmic spiral data disturbance method," *Journal of Computer*, no. 11, pp. 2275-2282, 2012.

[3] X. Liu, T. Peng, W. Zuo, *et al.* "Clustering-based PU active text classification method," *Journal of Software*, no. 11, pp. 2571-2583, 2013.

[4] Q. Fu, "Modeling and simulation of starting process of a turboshaft engine," *Computer Simulation*, vol. 30, no. 12, pp. 57-60, 2013.

[5] Z. Han, N. Chen, J. Le, *et al.* "Research on hot topic time series-oriented effective clustering algorithm," *Journal of Computer*, no. 11, pp. 2337-2347, 2012.

[6] J. Zhao, R. Zhao, Y. Ding, *et al.* "Nested loop classification-based parallel recognition technology," *Journal of Software*, no. 10, pp. 2695-2704, 2012.

[7] C. Gao, D. Miao, Z. Zhang, *et al.* "Active collaborative semi-supervised rough set classification model," *Pattern Recognition and Artificial Intelligence*, no. 5, pp. 745-754, 2012.

[8] J. Yang, and E. Chen, "Personalized service-oriented user interest shift detection and handling method," *Electronic Technology*, no. 11, pp. 72-76, 2009.

[9] S. Huang, "Analysis on simulation of computational process of complex electromagnetic field model," *Bulletin of Science and Technology*, vol. 29, no. 11, pp. 37-40, 2013.

[10] K. Chen, P. Han, and J. Wu, "User clustering-based heterogeneous social network algorithm," *Journal of Computer*, no. 2, pp. 349-359, 2013.

[11] H. Zhang, and G. Lu, "Evaluation and comparison of machine learning algorithm of classified unbalanced protocol flow," *Journal of Software*, no. 6, pp. 1500-1516, 2012.

[12] Y. Zhang, Q. Zhou, and X. Hu, "Interdisciplinary emotion classification-oriented feature selection method," *Pattern Recognition and Artificial Intelligence*, vol. 11, pp. 1068-1072, 2013.

Received: June 10, 2015 Revised: July 29, 2015 Accepted: August 15, 2015