

International Journal of Internet Protocol Technology

ISSN online: 1743-8217 - ISSN print: 1743-8209

<https://www.inderscience.com/ijipt>

Network big data association recommendation method based on modified entropy and improved FCM algorithm

Xiaomin Liu

DOI: [10.1504/IJIPT.2025.10071756](https://doi.org/10.1504/IJIPT.2025.10071756)

Article History:

Received:	20 December 2024
Last revised:	17 February 2025
Accepted:	20 February 2025
Published online:	10 July 2025

Network big data association recommendation method based on modified entropy and improved FCM algorithm

Xiaomin Liu

School of Mathematics and Physics,
Bengbu University,
Longzihu District, Bengbu, Anhui Province, China
Email: slxylxm520@126.com

Abstract: In order to overcome the problems of low accuracy, low recall and long generation time of recommendation results in traditional recommendation methods, a network big data association recommendation method based on modified entropy and improved FCM algorithm is proposed. Calculate the information entropy of network data and correct it. Use the corrected entropy to perform dimensionality reduction on the data. Calculate the abnormal factor value of the network data after dimensionality reduction to detect and remove abnormal data. Use an improved FCM algorithm to cluster the data after removing anomalies. Build a CLUPCDR model through data augmentation module, user feature extraction module and mapping header module, and use this model to implement network big data association recommendation. The experimental results show that the maximum accuracy of the proposed method is 98.74%, the maximum recall is 99.12%, and the recommended result generation time varies between 0.19 s and 0.57 s.

Keywords: corrected entropy; improved FCM algorithm; network big data; association recommendation; abnormal factor value; CLUPCDR model.

Reference to this paper should be made as follows: Liu, X. (2025) 'Network big data association recommendation method based on modified entropy and improved FCM algorithm', *Int. J. Internet Protocol Technology*, Vol. 18, No. 1, pp.48–57.

Biographical notes: Xiaomin Liu received her Master's degree in Basic Mathematics from Harbin University of Science and Technology in 2006. Currently, she is a Lecturer at the School of Mathematics and Physics, Bengbu University. Her research interests include functional analysis, fuzzy statistical analysis and numerical computational methods. Hosted several provincial and municipal teaching and research projects, and published many academic papers.

1 Introduction

In today's rapidly developing era of informatisation and digitisation, network big data has become the core driving force for promoting innovative development in various industries (Omar et al., 2023). With the widespread application of various digital and information technologies, the speed of data growth has increased rapidly, providing abundant resources and broad application prospects for data association recommendation research (Man et al., 2020). The research on network big data association recommendation not only involves complex data processing techniques, but also integrates advanced algorithms and ideas from multiple fields such as artificial intelligence and machine learning, aiming to provide users with more accurate and personalised recommendation services by deeply mining the correlations between data. Data correlation technology, as a scientific analysis method, provides more comprehensive and profound

basis for decision-making by revealing the mutual influence and correlation between data (Shen et al., 2020). The core principle is to integrate and analyse multidimensional data to identify potential patterns and trends. In the big data environment, the core role of data association technology is reflected in cross system and cross-level data integration, which can effectively combine data from multiple sources and provide a foundation for in-depth data mining. This technology not only improves the efficiency of data analysis, but also significantly optimises the decision-making process, becoming an important driving force for promoting business model innovation. Therefore, researching network big data association recommendation methods is of great significance.

Gao and He (2020) proposed a network big data recommendation method based on bipartite graph multi weight projection. Determine the network data set for analysis, identify the basic information of the data, and use the Lewinstein distance to calculate the correlation and

attribute similarity between network data recommendation items and users. Convert the relationships between multiple network data recommendation items and users into bipartite graphs, and use enhanced bipartite graph projection techniques to predict network links for network big data recommendation. However, after testing, it was found that the accuracy of network big data association recommendation using this method is not high, and there is a significant gap between the expected goals. Li et al. (2020) proposed a network big data recommendation method based on gravity search link prediction and score propagation. Using the generalised Meta Path model to measure the similarity between network users and determine the similarity graph. Using gravity search to optimise subgraphs in similar graphs, in order to achieve link prediction. By combining network transmission strategies based on infectious disease models, hidden patterns can be discovered according to existing patterns, and the association between users and data items can be determined by combining hidden patterns to achieve data recommendation. However, after in-depth analysis, it was found that the recall rate of this method is relatively low, and the actual application effect is not good. Chen and Liu, (2024) proposed a network big data recommendation method based on mining algorithms. Applying collaborative recommendation algorithm to analyse users' historical behaviour, considering the effectiveness of intelligent recommendation through data similarity, optimising the recommendation process using Top-N algorithm and optimising the intelligent recommendation algorithm using neural network-based mining algorithm to achieve data recommendation. However, analysis shows that this method takes a long time to generate recommendation results and overall efficiency is not high.

A new network big data association recommendation method based on modified entropy and improved FCM algorithm is proposed with the expectation of addressing multiple objectives such as low accuracy and recall rate, and long recommendation result generation time of the above methods.

2 Design of network big data association recommendation method

2.1 Abnormal network data removal based on modified entropy

Abnormal data is usually generated due to device failures, data transmission errors, malicious attacks or user misoperation. These data are not only inaccurate, but may also mislead subsequent association recommendations. By removing abnormal data, it can ensure that the data input to the association recommendation method is more authentic and reliable, thereby improving the quality of recommendations.

Information entropy (Wu et al., 2024; Jiao and Duan, 2024) is mainly used to quantify the uncertainty of various possible events occurring in the information source, representing the average uncertainty before the information source sends out the information. The formula for calculating information entropy is as follows:

$$\begin{aligned} H(X) &= p(x_1)I(x_1) + p(x_2)I(x_2) + \dots + p(x_n)I(x_n) \\ &= -p(x_1)\log[p(x_1)] - p(x_2)\log[p(x_2)] \\ &\quad - \dots - p(x_n)\log[p(x_n)] \\ &= -\sum_{i=1}^n p(x_i)\log[p(x_i)] \end{aligned} \quad (1)$$

In the above formula, $p(x_1), p(x_2), \dots, p(x_n)$ represents the probability of network data $x_1, x_2, \dots, x_n$ appearing, and $I(x_1), I(x_2), \dots, I(x_n)$ represents the attribute parameter of $x_1, x_2, \dots, x_n$.

In certain specific network data or application scenarios, traditional entropy measures may not accurately reflect the uncertainty or purity of the data. For example, in recommendation systems, user behaviour data often has high sparsity and dynamism and traditional entropy measures may not fully capture these characteristics. In traditional entropy measurement, although the probability distribution of data is also considered, the impact of the probability of specific data items on the overall entropy value is not fully emphasised. The modified entropy formula may have better interpretability by introducing a calculation term related to the probability of occurrence of the network data. This helps to gain a more intuitive understanding of the sources of data uncertainty or purity, as well as how adjusting data probabilities can affect entropy values. This modification helps to more accurately understand data characteristics, optimise decision-making processes, and improve overall system performance. In order to improve the accuracy of information entropy description, the modified entropy is defined as:

$$H = \begin{cases} -\frac{1}{\lg N} \sum_{i=1}^N p_i \lg p_i, N \geq 2 \\ 1, N = 1 \end{cases} \quad (2)$$

In the formula, p_i represents the probability of the i -th network data appearing (Wang et al., 2024a; Niu et al., 2023), and N represents the total amount of network data.

In order to effectively remove abnormal network data, it is necessary to perform dimensionality reduction on the data. The specific implementation steps are as follows:

- (1) Use formula (2) to calculate the fluctuation degree $H(X_i)$ of each attribute in the network data, and retain the attribute values of $H(X_i) > \varepsilon$. The filtered network data set is represented by $Z = (Z_1, Z_2, \dots, Z_k)$.
- (2) The calculated polynomial coefficient $a_{1j}, a_{2j}, \dots, a_{dj}$ ($j = 1, 2, \dots, d$) is calculated using the following formula:

$$\begin{bmatrix} t_{11} & t_{12} & t_{13} & \dots & t_{1d} \\ t_{21} & t_{22} & t_{23} & \dots & t_{2d} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ t_{m1} & t_{m2} & t_{m3} & \dots & t_{md} \end{bmatrix}_{m \times d} \begin{bmatrix} a_{1j} \\ a_{2j} \\ \vdots \\ a_{dj} \end{bmatrix}_{d \times 1} = \begin{bmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{mj} \end{bmatrix}_{m \times 1} \quad (3)$$

The network data matrix presented in formula (3) is a crucial step in representing raw high-dimensional data in matrix form. This matrix contains information about all network data points, with each row representing a data point and each column corresponding to a feature or dimension. This representation method is convenient for subsequent data processing and analysis. In the process of dimensionality reduction, it is necessary to capture complex relationships in the data through polynomial fitting or other nonlinear methods. The matrix equation in equation (3) provides the basis for this calculation. Through matrix operations, the coefficients of polynomials can be efficiently solved, which reflect the interrelationships of data in different dimensions and are the key basis for subsequent dimensionality reduction processing. By combining the coefficients of polynomials, identify which features are most sensitive to changes in data, and guide feature selection or dimensionality reduction processing. For example, features can be sorted based on the absolute value of coefficients, and features with larger coefficients can be selected for preservation to achieve dimensionality reduction.

- (3) Calculate its eigenvalue λ_i ($i = 1, 2, \dots, m$) and eigenvector $B_{m \times n}^*$ based on Z . If $m = d$, that is, the number of rows in Z is equal to the number of columns, then calculate the eigenvalues λ_i and eigenvectors $B_{m \times m}^* = (b_{ij})$ of matrix M according to formula (4).

$$\begin{cases} |Z - \lambda E| = 0 \\ (Z + \lambda E)B = 0 \end{cases} \quad (4)$$

In the above formula, E represents the error vector.

If the rank of matrix Z is r , according to the theory of singular value decomposition, there exist unitary matrices P and Q . Therefore, the eigenvector $B_{m \times m}^*$ of matrix Z consists of m for $m \times 1$ column vectors to form data set $T_{m \times 1}^*$. At this point, the following relationship holds:

$$Z = P \begin{pmatrix} D & 0 \\ 0 & 0 \end{pmatrix} Q^H \quad (5)$$

- (4) Based on the solved eigenvector matrix, project it at time $T_{m \times 1}^*$ to achieve dimensionality reduction of network data. The specific calculation formula is as follows:

$$T_{m \times 1}^* = B_{m \times m} T_{m \times 1} = \begin{bmatrix} b_{11} & b_{12} & \dots & b_{1m} \\ b_{21} & b_{22} & \dots & b_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ b_{m1} & b_{m2} & \dots & b_{mm} \end{bmatrix} \begin{bmatrix} T_1 \\ T_2 \\ \dots \\ T_m \end{bmatrix} = \begin{bmatrix} W_1 \\ W_2 \\ \dots \\ W_m \end{bmatrix} \quad (6)$$

The specific description of the abnormal network data detection process is as follows:

- *Step 1*: Real time data collection to obtain network data set sample X .
- *Step 2*: Initialise based on sample values and the optimal data set threshold to construct the optimal data set grid;

- *Step 3*: Calculate the information entropy value $H(X_i)$ for each dimension. If $H(X_i) > \varepsilon$, add X_i to the new network data set to form Z ;
- *Step 4*: Calculate the polynomial coefficients a_{ij} ($1 \leq i \leq m, 1 \leq j \leq d$) of the matrix;
- *Step 5*: Calculate the eigenvalues and eigenvectors of the matrix;
- *Step 6*: Use formula (6) to obtain the reduced dimension data set $T_{m \times 1}^*$ and update the data dimension set;
- *Step 7*: Collect the latest data element $X(i)$ at the next moment and establish the nearest data set grid X_p ;
- *Step 8*: Calculate the outlier factor value VOA_p for data point $X(i)$. If $VOA_p > \mu$, the algorithm will identify the point as an outlier. The specific calculation formula is as follows:

$$VOA_p = \text{Var}_{m, n \in S \setminus \{X_p\}} [\theta_{mpn}] \quad (7)$$

In the above formula, θ_{mpn} represents the angular variance between the normal data vector and the abnormal data vector

- *Step 9*: Otherwise, update the best data grid and the nearest data grid according to the first in first out method and the optimal data set threshold and scale.
- *Step 10*: Jump to Step 7 and continue detecting the data points that will arrive at the next moment until all data has been detected. After the detection is complete, remove them to improve the quality of network data.

2.2 Network big data clustering based on improved FCM algorithm

The Fuzzy C-Means (FCM) algorithm was originally extended from the Hard C-Means (HCM) algorithm (Almomany et al., 2023; Wang 2024b). The FCM algorithm mainly determines the degree to which each sample belongs to a certain category based on its membership degree for clustering. Assuming the data to be clustered is $X' = \{x_1, x_2, \dots, x_n\}$, the objective function and its constraints of the FCM clustering algorithm are expressed by the following formula:

$$J = \sum_{j=1}^n \sum_{i=1}^k u_{ij}^m x_j - c_i^2 \quad (8)$$

$$\sum_{i=1}^k u_{ij} = 1, j = 1, 2, \dots, n \quad (9)$$

In the above formula, m represents the membership factor (Hu et al., 2022), u_{ij}^m represents the membership degree of sample j belonging to a certain cluster when the membership factor is m , x_j represents the j -th network data sample, c_i represents the cluster centre and k represents the number of data categories.

If Lagrange multiplication is used to construct the minimisation objective function for the objective function, the following relationship holds:

$$F = \sum_{i=1}^k \sum_{j=1}^n u_{ij} x_j - c_i^2 + \sum_j \lambda_i \left(\sum_{i=1}^n u_{ij} - 1 \right) \quad (10)$$

In the above formula, λ_i represents the i -th Lagrange multiplier.

The value of the objective function is used to measure the non-similarity of samples between clusters. The smaller its value, the less similar the samples between clusters are, and the better the clustering effect (Qian et al., 2022; Chighoub and Saouli, 2022). Based on the constraint conditions, the calculation formula for membership degree u_{ij} and cluster centre c_i can be obtained as follows:

$$u_{ij} = \left(\frac{-\lambda_j}{m} \right)^{\frac{1}{m-1}} \left(\frac{1}{x_j - c_i^{\left(\frac{2}{m-1}\right)}} \right) \quad (11)$$

$$c_i = \frac{\sum_{j=1}^n (x_j u_{ij}^m)}{\sum_{j=1}^n u_{ij}^m} \quad (12)$$

Combining the probability allocation of learning automata and the reward and punishment mechanism in the update process with the FCM algorithm, an improved FCM algorithm Fuzzy c-Means Based on Learning Automata (LAFCM) based on learning automata is proposed to maximise data clustering accuracy.

- 1) *Construction of Q -table*: Set each network data sample to correspond to action 1, allocate the probability of the sample belonging to each class cluster, construct a Q -table consisting of n rows and k columns and

form a membership matrix corresponding Q -table, i.e., $Q = \{q_{ij} | i=1,2,3,...,k; j=1,2,3,...,n\}$. The mapping relationship between the input data and the Q -table is shown in Figure 1.

- 2) *Cluster centre c_i calculation*: The calculation formula for cluster centre c_i is as follows:

$$c_i = \frac{\sum_{j=1}^n (x_j q_{ij}^m)}{\sum_{j=1}^n q_{ij}^m} \quad (13)$$

In the above formula, q_{ij} represents the membership degree of network data sample j to cluster i .

- 3) *Determine the objective function value of J* : The calculation formula for the objective function value J is as follows:

$$J(Q, c_1, c_2, \dots, c_c) = \sum_{i=1}^c \sum_{j=1}^n q_{ij}^m x_j - c_i^2 \quad (14)$$

- 4) *Calculate the average distance within the class $Adis(c_j)$* : The calculation formula for the average distance $Adis(c_j)$ within the class is as follows:

$$Adis(c_i) = \frac{1}{|c_i|} \sum_{j=1}^{|c_i|} \sqrt{\sum_{l=1}^L (x_{jl} - c_{il})^2} \quad x_j \in c_i \quad (15)$$

In the above formula, $|c_i|$ represents the number of samples in the i -th cluster, x_{jl} represents the l -th attribute value of sample j and c_{il} represents the l -th attribute of cluster centre i . The clustering process of network big data based on improved FCM algorithm is shown in Figure 2.

Figure 1 Mapping relationship between input data and Q -table

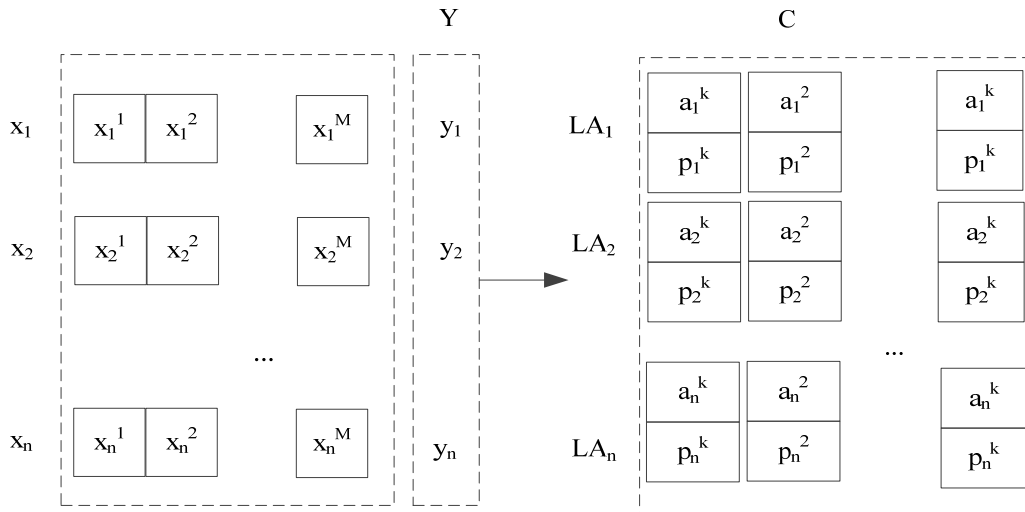
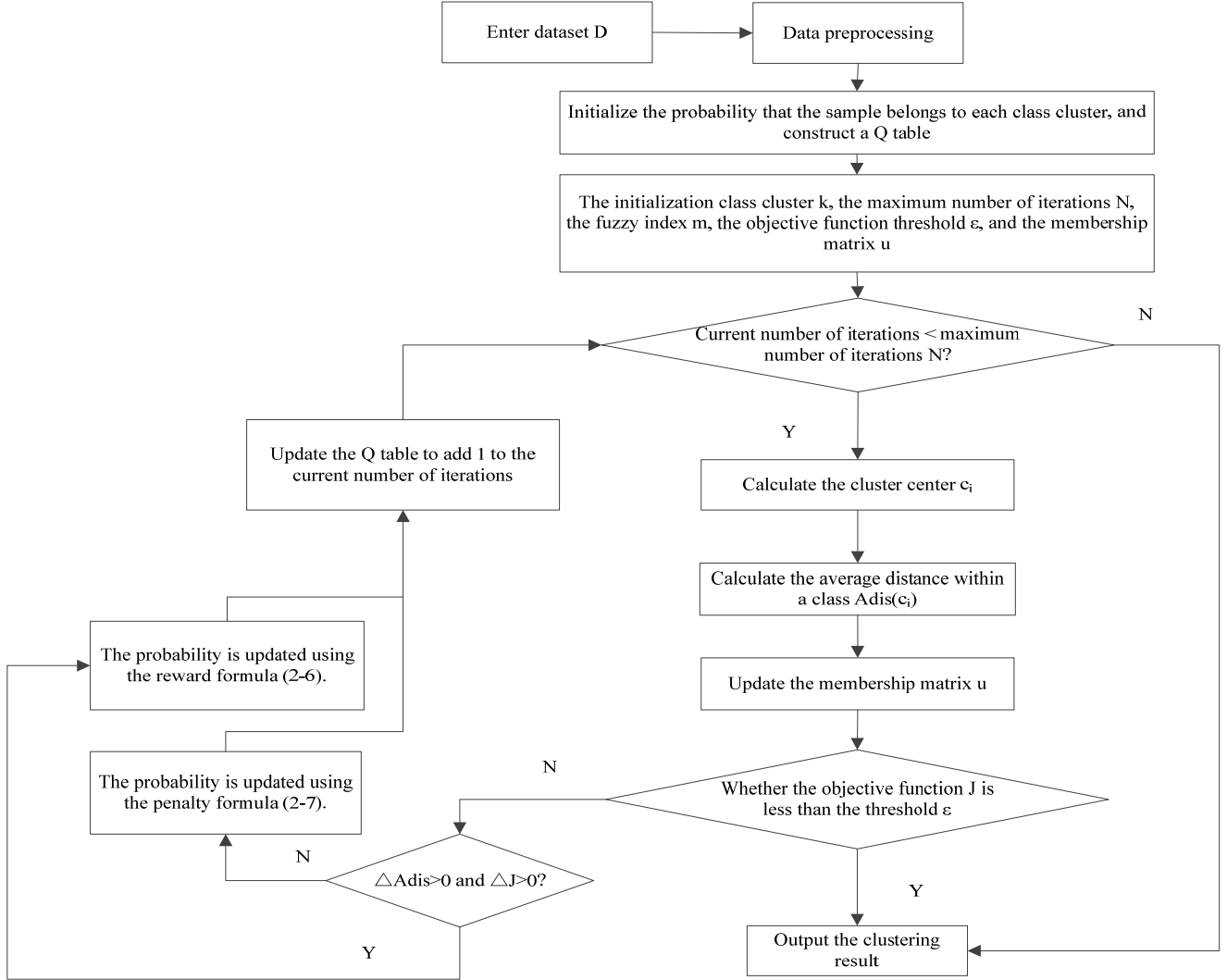
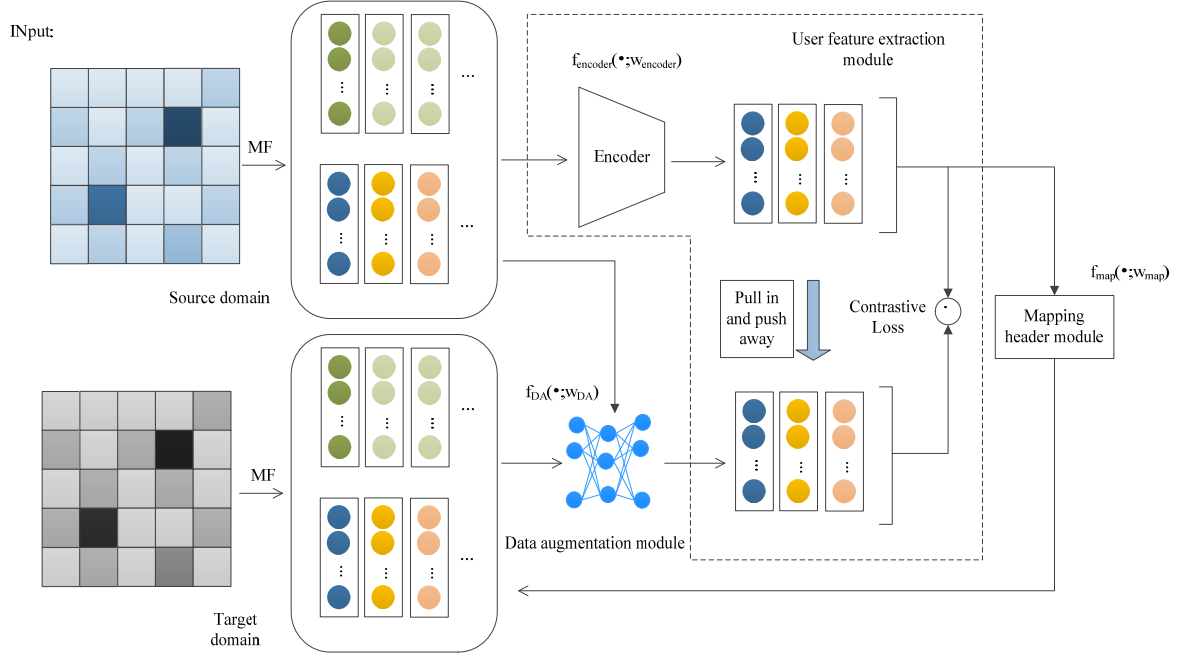


Figure 2 Network big data clustering process based on improved FCM algorithm

2.3 Network big data association recommendation based on CLUPCDR model

CLUPCDR model, namely, contrastive learning for extracting transferable user profiles in cross-domain recommendation system, is an innovative cross-domain recommendation system algorithm. CLUPCDR model mainly aims at the complexity of information transfer between domains in recommendation methods, aiming at maximising the

relevance of overlapping users and minimising the relevance of non-overlapping users in two domains through comparative learning, thus helping the model to capture richer and more differentiated features in cross-domain data and learn the transferable feature representation. This algorithm can improve the performance of cross-domain recommendation methods and effectively alleviate the data sparsity and cold start problems of single domain recommendation methods. The structure of the CLUPCDR model is shown in Figure 3.

Figure 3 CLUPCDR model structure (see online version for colours)

1) *Data augmentation module*: The function of the data augmentation module is to enhance the user feature representation of the target domain to alleviate the problem of imbalance between the source domain user feature representation and the target domain user feature representation caused by insufficient user interaction in the target domain. This module comprehensively considers the user characteristics of the source domain and the information of the target domain to generate an embedded representation that more accurately represents the users in the target domain.

$$u^{t'} = f_{DA}(u^s, v^t, r^t; w_{DA}) \quad (16)$$

In the above formula, $u^{t'}$ represents the user representation in the target domain after data augmentation, $f_{DA}(\cdot)$ represents the data augmentation network, u^s represents the user representation in the source domain and v^t and r^t represent the representation and corresponding rating of the target item, respectively. During the training process of the data augmentation module, the loss function in the EMCDCR model was optimised and the calculation formula for the optimised loss function is as follows:

$$L = \frac{1}{|R_o^t|} \sum_{r_{ij} \in R_o^t} (r_{ij} - f_{u_i}(u_i^s) v_j)^2 \quad (17)$$

In the above formula, $\mathcal{R}_o^t = \{r_{ij} | u_i \in \mathcal{U}^o, v_j \in \mathcal{V}^t\}$ represents the interaction set related to users who coexist in the overlapping user group $\mathcal{U}^o$ and the target domain $\mathcal{V}^t$, r_{ij} represents the correlation coefficient between user i and user j , v_j represents the interest model of user j and u_i represents the attribute parameters of user i .

2) *User feature extraction module*: The function of the user feature extraction module is to mine the essential features of the same users in different fields through contrastive learning. By using contrastive learning to extract the unique features of users, it can pull closer to the target domain space, facilitating accurate cross domain transmission of features. The calculation formula for user feature extraction is as follows:

$$u^{s'} = f_{encoder}(u^s; W_{encoder}) \quad (18)$$

In the above formula, $u^{s'}$ represents the representation of the user after being processed by the encoder, and $f_{encoder}(\cdot)$ represents the encoder network.

By using contrastive learning method, the user representations processed by the encoder in the source domain and the enhanced representations obtained through data augmentation in the target domain are utilised to generate more cross domain transferable user features, in order to achieve effective cross domain knowledge transfer. The specific form of contrastive learning is represented by the following formula:

$$L = \frac{1}{N} \sum_{i=1}^N \left(- \frac{\exp\left(\frac{u_i^{s'} \cdot u_i^{t'}}{\tau}\right)}{\exp\left(\frac{u_i^{s'} \cdot u_i^{t'}}{\tau}\right) + \sum_{k=1}^N \exp\left(\frac{u_i^{s'} \cdot u_k^{t'}}{\tau}\right)} \right) \quad (19)$$

In the above formula, $u_i^{s'}$ represents the representation of user i in the source domain after being processed by the encoder, $u_i^{t'}$ represents the representation of user k in the target domain after data augmentation, τ represents the adjustment factor and N represents the number of sample pairs in the batch.

- 3) *Mapping head module*: The mapping header module can be seen as the simplest cross domain recommendation model, which can be used to detect the quality of cross domain user feature representations generated through contrastive learning. The specific form is represented by the following formula:

$$L_{map} = \frac{1}{|R_o^t|} \sum_{r_{ij} \in R_o^t} \left(r_{ij} - f_{map} \left(f_{encoder} \left(u_i^s; w_{encoder} \right); w_{map} \right) v_j \right)^2 \quad (20)$$

In the above formula, $f_{map}(\cdot)$ represents the mapping head network, and v_j represent the item representation.

The calculation formula for network big data association recommendation is as follows:

$$Q = \min_{u,v} \frac{1}{|\mathcal{R}|} \sum_{r_{ij} \in \mathcal{R}} \frac{u^t u^{s'} (r_{ij} - u_i v_j)^2}{L_{map}} \quad (21)$$

In the above formula, $|\mathcal{R}|$ represents the number of ratings.

3 Experimental design

3.1 Experimental scheme

In order to verify the practical application effect of the network big data association recommendation method based on modified entropy and improved FCM algorithm, experimental tests were conducted.

Experimental standards are the foundation for ensuring the reproducibility and accuracy of experimental results. In the experiment of network big data association recommendation, the following specifications should be followed:

- *Data pre-processing specifications*: including steps such as data cleaning, missing value handling, outlier detection and processing, to ensure data quality and consistency.
- *Model training standards*: Select appropriate algorithms and parameters for model training, and record key parameters and performance indicators during the training process.

The data sources for network big data related recommendation are very diverse, mainly including user behaviour data, product information data, social media data, and other data sources. The publicly available experimental data sets include MovieLens movie rating data set, Amazon product review data set, etc. These data sets contain rich user behaviour information and product information, which are suitable for the research and validation of association recommendation algorithms. On this basis, a custom data set was set up, which collects and constructs its own data set based on specific application scenarios and requirements. Self-built data sets typically contain more specific user behaviour information and product attribute information, which can better reflect the characteristics of practical application scenarios. Randomly allocate data points to the training set, validation set and testing set. This method is simple and easy to implement, but

may not be suitable for data sets with imbalanced class distributions.

- (1) *Experimental data*: In the experimental process of network big data association recommendation, the specific types of experimental data involved mainly include the following:

- 1) *Structured data*: Structured data refers to data that can be stored in relational databases, with clear patterns and structures, typically including numbers, text and dates, and can be queried and manipulated using SQL language. In the association recommendation experiment, structured data includes users' basic information (such as age, gender, occupation, etc.), users' purchase records, product attribute information (such as price, brand, category, etc.), etc.
 - 2) *Semi structured data*: Semi structured data refers to data that has a certain organisational structure but is not as strict as structured data, usually stored in formats such as XML, JSON, etc., containing tags or labels used to define data elements and their hierarchical relationships. In the association recommendation experiment, semi-structured data includes users' browsing history, search history, interactive information on social media, etc.
 - 3) *Unstructured data*: Unstructured data refers to data that does not have a predefined format and typically exists in the form of text, images, audio, video, etc. This type of data is also of great significance in association recommendation experiments. For example, textual data such as user comments, reviews, posts on social media, as well as various types of image and video data.
 - 4) *Time series data*: Temporal data refers to data recorded in chronological order, usually used to reflect the changes of something within a specific time period. In the correlation recommendation experiment, temporal data includes users' purchase time, browsing time, product launch time, etc.
 - 5) *Spatial data*: Spatial data refers to data that contains geographic location information, typically used to describe geographic features and spatial relationships. In the association recommendation experiment, spatial data includes users' geographical location information, product sales areas, etc.
- (2) *Experimental indicators*: We conducted relevant experiments using Gao et al. (2020) method, Li et al. (2020) method and the proposed method, and compared the accuracy, recall and recommendation result generation time of different methods to verify their application effectiveness.

Among them, accuracy is a key evaluation indicator in network big data association recommendation. It mainly measures the proportion of the recommendation results

given by the recommendation system that truly meet the user's interests (that is, the items that the user is actually interested in or will generate positive feedback, such as clicks, purchases, collections, etc.). Simply put, it reflects the accuracy of the recommendation method, that is, how many of the recommended results are truly useful, and its calculation formula is as follows:

$$P_r = \frac{TP}{TP + FP} \quad (22)$$

In the above formula, TP represents the number of items that the recommendation method recommends to the user and that the user is actually interested in. FP represents the number of items recommended by the recommendation method to the user that the user is not interested in.

Recall rate is an important evaluation metric in network big data related recommendation systems. It measures the proportion of items that the recommendation system successfully recommends among all items that users may be interested in (i.e., items that are truly relevant in the test set). Simply put, recall rate reflects the ability of recommendation methods to discover items of interest to users, and its calculation formula is as follows:

$$R_e = \frac{TP}{TP + FN} \quad (23)$$

In the above formula, FN represents the number of items that the user is truly interested in, but the recommendation method does not recommend to the user.

The generation time of network big data related recommendation results refers to the time it takes for the recommendation method to receive user requests (such as queries, browsing, etc., triggering recommendation requirements) and generate the final recommendation results. This indicator reflects the response speed of the recommendation method and is an important aspect of measuring its performance. Its calculation formula is as follows:

$$T = g_i - g_j \quad (24)$$

In the above formula, g_i and g_j , respectively represent the time when the recommendation result is returned and the time when the user makes a recommendation request.

3.2 Experimental result

The comparison results of the accuracy of network big data association recommendation using three methods are shown in Table 1.

According to the analysis of the data in Table 1, the maximum accuracy of network big data association recommendation for Gao et al. (2020) method is 88.59%, and the maximum accuracy of network big data association recommendation for Li et al. (2020) method is 85.33%. The maximum accuracy of network big data association recommendation for the proposed method is 98.74%, which is 10.16% and 13.41% higher than that of Gao et al. (2020)

method and Li et al. (2020) method, respectively. The minimum accuracy of network big data association recommendation for Gao et al. (2020) method is 80.31%, and the minimum accuracy of network big data association recommendation for Li et al. (2020) method is 70.65%. The minimum accuracy of network big data association recommendation for the proposed method is 95.63%, which is 15.32% and 24.98% higher than that of Gao et al. (2020) method and Li et al. (2020) method, respectively. After comparison, it can be seen that the proposed method has a higher accuracy in network big data association recommendation, which can achieve the goal of accurate recommendation for network big data.

Table 1 Comparison results of accuracy of network big data association recommendations

Number of experiments	Accuracy/%		
	Gao et al. (2020) method	Li et al. (2020) method	Proposed method
10	80.31	74.12	98.74
20	84.75	75.63	98.56
30	86.39	74.15	97.49
40	87.41	72.39	98.63
50	87.52	74.63	97.31
60	86.33	85.33	95.87
70	84.14	74.26	95.63
80	88.59	81.47	97.48
90	84.64	70.65	96.35
100	88.17	78.47	95.66

The comparison results of the recall rates of the network big data association recommendation results of the three methods are shown in Table 2.

Table 2 Comparison of recall rates of network big data association recommendation results

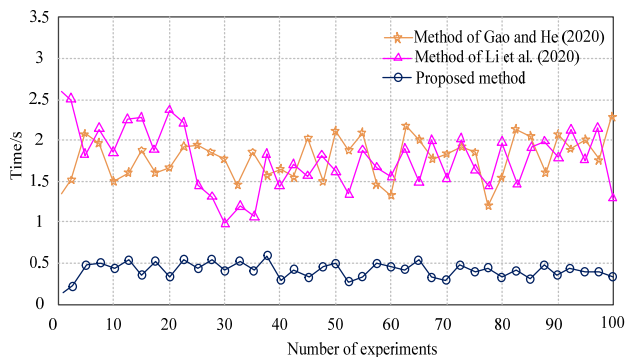
Number of experiments	Recall rate/%		
	Gao et al. (2020) method	Li et al. (2020) method	Proposed method
10	70.63	81.16	97.52
20	74.18	82.33	98.63
30	75.69	87.47	99.12
40	78.41	85.72	94.78
50	76.66	86.34	95.63
60	78.63	84.79	98.41
70	74.55	86.33	98.55
80	75.26	84.72	95.19
90	79.61	82.34	96.37
100	75.28	81.73	98.42

According to the analysis of the data in Table 2, the maximum recall rate of the network big data association recommendation results of Gao et al. (2020) method is

79.61%, and the maximum recall rate of the network big data association recommendation results of Li et al. (2020) method is 87.47%. The maximum recall rate of the network big data association recommendation results of the proposed method is 99.12%, which is 19.51% and 11.65% higher than that of Gao et al. (2020) method and Li et al. (2020) method, respectively. The minimum recall rate of the network big data association recommendation results of Gao et al. (2020) method is 70.63%, and the minimum recall rate of the network big data association recommendation results of Li et al. (2020) method is 81.16%. The minimum recall rate of the network big data association recommendation results of the proposed method is 94.78%, which is 24.51% and 24.15% higher than that of Gao et al. (2020) method and Li et al. (2020) method, respectively. After comparison, it can be seen that the recall rate of the network big data association recommendation results of the proposed method is higher, which means that among all the items that users may be interested in (truly relevant items in the test set), the recommendation method can successfully recommend a larger proportion of items. This indicates that recommendation methods have strong capabilities in mining user interests and can comprehensively capture users' points of interest.

The comparison of the generation time of network big data association recommendation results for the three methods is shown in Figure 4.

Figure 4 Comparison of generation time of network big data association recommendation results (see online version for colours)



Analysing the results in Figure 4, it can be seen that the generation time of network big data association recommendation results for Gao et al. (2020) method varies between 1.2 s and 2.27 s, while the generation time of network big data association recommendation results for Li et al. (2020) method varies between 0.99 s and 2.58 s. The generation time of network big data association recommendation results for the proposed method varies between 0.19 s and 0.57 s. After comparison, it can be seen that the proposed method has a shorter generation time for network big data association recommendation results and higher overall recommendation efficiency.

On this basis, in order to enhance the influence and relevance of network big data association recommendation

methods, it is necessary to adjust and expand them to adapt to a wider range of data types and application scenarios. Specifically, this method can be applied to different fields such as social media analysis, financial data analysis and healthcare, while exploring its applicability to unstructured and time-series data. In addition, it is necessary to add 10 data sets from different fields, scales and structures, and design experimental scenarios based on the experimental data. Ensure its stability and consistency through multiple experiments and cross validation. Through these efforts, we can promote the application and development of this method in a wider range of fields. The size and attributes of the 10 data sets show significant differences, and the test results are shown in Table 3.

Table 3 Adaptability test results on different data sets

Data set name	Accuracy/%		
	Gao et al. (2020) method	Li et al. (2020) method	Proposed method
A	80.36	70.32	98.75
B	85.26	71.45	96.85
C	86.33	78.63	97.41
D	84.17	80.23	97.36
E	85.26	78.26	98.12
F	85.49	71.15	98.63
G	86.32	72.33	97.33
H	81.47	74.89	96.84
I	82.36	71.67	97.19
J	84.78	73.15	96.38
average value	84.18	74.21	97.49

According to the analysis of the data in Table 3, the average accuracy of network big data association recommendation for Gao et al. (2020) method is 84.18%, the average accuracy of network big data association recommendation for Li et al. (2020) method is 74.21% and the average accuracy of network big data association recommendation for the proposed method is 97.49%. After comparison, it can be seen that the accuracy of network big data association recommendation for the proposed method is high and the changes are more stable, indicating that this method has strong adaptability in network big data association recommendation.

In summary, the maximum accuracy of the proposed method for network big data association recommendation is 98.74%, the maximum recall is 99.12% and the generation time of recommendation results varies between 0.19 s and 0.57 s. It has strong adaptability on different data sets and has the dual advantages of high accuracy and high efficiency. This is mainly attributed to the following aspects: firstly, in terms of information entropy correction and dimensionality reduction processing, by correcting the information entropy and effectively performing dimensionality reduction processing, this method not only successfully preserves the key information of the data, but also eliminates noisy data,

thereby significantly improving the accuracy of subsequent clustering analysis and association recommendation. Secondly, the improved FCM algorithm demonstrates stronger global search capabilities and better clustering effects, enabling deeper exploration of potential connections and patterns between data. Furthermore, the CLUPCDR model integrates deep learning techniques and contrastive learning ideas, enabling the model to learn richer and more discriminative transferable feature representations, further improving the accuracy of association recommendations. At the same time, dimensionality reduction effectively reduces data dimensionality and computational complexity, optimising the efficiency of the algorithm. In the process of building the CLUPCDR model, optimised algorithms and modular design were adopted to further improve the operational efficiency of the model. In addition, this method performs well on different data sets and has the ability to flexibly handle various types and scales of network big data. The modular design and strong scalability of the CLUPCDR model make it easy to customise and optimise according to actual needs. In summary, this method has a recommendation accuracy of up to 98.74% and a recall rate of up to 99.12% in network big data association recommendation. The recommendation result generation time is between 0.19 s and 0.57 s, demonstrating significant advantages of high accuracy, high efficiency and strong adaptability.

4 Conclusion

Data association technology is a scientific method that analyses the relationships between data and explores potential information. It reveals the mutual influence and correlation between data, providing more comprehensive and profound basis for decision-making. Therefore, a network big data association recommendation method based on modified entropy and improved FCM algorithm is proposed. The experimental results show that the proposed method has a maximum accuracy of 98.74% for network big data association recommendation, a maximum recall of 99.12% and a recommendation result generation time that varies between 0.19 s and 0.57 s. It has the dual advantages of high accuracy and high efficiency. In the big data environment, this article studies the role of cross system and cross level data integration, effectively combining data from multiple sources to provide a foundation for in-depth data mining.

Acknowledgement

This work was supported by Bengbu University Natural Science Key Project under grant no. 2024ZR05zd.

Declarations

All authors declare that they have no conflicts of interest.

References

- Almomany, A., Jarrah, A. and Al Assaf, A. (2023) 'Accelerating FCM algorithm using high-speed FPGA reconfigurable computing architecture', *Journal of Electrical Engineering and Technology*, Vol. 18, No. 4, pp.3209–3217.
- Chen, Y.Y. and Liu, Y.S. (2024) 'Simulation of user big data periodic intelligent recommendation based on mining algorithm', *Computer Simulation*, Vol. 41, No. 4, pp.456–460,465.
- Chighoub, F. and Saouli, R. (2022) 'Fully integrated spatial information to improve FCM algorithm for brain MRI image segmentation', *Automatic Control and Computer Sciences*, Vol. 56, No. 1, pp.67–82.
- Gao, W. and He, K.Q. (2020) 'Recommendation system of big data based on multi-weight projection of bipartite network', *Application Research of Computers*, Vol. 37, No. 3, pp.712–716.
- Hu, J., Yin, H. and Wei, G. et al. (2022) 'An improved FCM clustering algorithm with adaptive weights based on PSO-TVAC algorithm', *Applied Intelligence: The International Journal of Artificial Intelligence, Neural Networks, and Complex Problem-Solving Technologies*, Vol. 52, No. 8, pp.9521–9536.
- Jiao, Y. and Duan, L. (2024) 'Redundancy analysis in the translation of scientific and technical English texts combined with information entropy', *Applied Mathematics and Nonlinear Sciences*, Vol. 9, No. 1, pp.10–25.
- Li, Z., Wu, Y. and Geng, H.J. (2020) 'Big data recommender system based on gravitational search for link prediction and ratings propagation', *Computer Applications and Software*, Vol. 37, No. 2, pp.39–47.
- Man, N., Wang, K. and Liu, L. (2020) 'Fuzzy sets for data mining and recommendation algorithms', *Journal of Intelligent & Fuzzy Systems: Applications in Engineering and Technology*, Vol. 38, No. 4 Pt.1, pp.3737–3745.
- Niu, G., Hu, D. and Eladdad, Y.Z.E. (2023) 'Evaluation and prediction of the transformer health index based on matter element information entropy and SVM', *Journal of Intelligent & Fuzzy Systems: Applications in Engineering and Technology*, Vol. 45, No. 6, pp.9815–9825.
- Omar, H.K., Frikha, M. and Jumaa, A.K. (2023) 'Big data cloud-based recommendation system using NLP techniques with machine and deep learning', *Telkomnika*, Vol. 21, No. 5, pp.1–10.
- Qian, Z., Cao, Y. and Sun, X. et al. (2022) 'Clustering optimization for triple-frequency combined observations of BDS-3 based on improved PSO-FCM algorithm', *Remote Sensing*, Vol. 14, No. 15, pp.1171–1185.
- Shen, J., Zhou, T. and Chen, L. (2020) 'Collaborative filtering-based recommendation system for big data', *International Journal of Computational Science and Engineering*, Vol. 21, No. 2, pp.219–225.
- Wang, L., He, Y. and He, Y. et al. (2024b) 'Wind turbine blade icing risk assessment considering power output predictions based on SCSO-IFCM clustering algorithm', *Renewable Energy*, Vol. 223, No. 10, pp.1–11.
- Wang, Z., Zhang, J. and Zhao, Q. (2024a) 'Influence of geometric structure on information entropy in spin-1 ferromagnetic dipolar Bose-Einstein condensates', *Modern Physics Letters, B. Condensed Matter Physics, Statistical Physics, Applied Physics*, Vol. 38, No. 15, pp.1–15.
- Wu, H., Chen, Y. and Zhu, W. et al. (2024) 'Feature selection in high-dimensional data: an enhanced RIME optimization with information entropy pruning and DBSCAN clustering', *International Journal of Machine Learning and Cybernetics*, Vol. 15, No. 9, pp.4211–4254.