Fuzzy Information Extraction Based on Improved Chaotic Partitioning Algorithm (Postprint)

Authors: Wan Fucheng

Date: 2018-06-19T00:00:00+00:00

Abstract

Fuzzy information mining and extraction in big data environments is affected by small perturbation inter-class interference between data, leading to poor feature clustering performance. This paper proposes a fuzzy information extraction method based on an improved chaotic partition algorithm. The method performs distributed structural reorganization on high-dimensional data streams, utilizes the Lorenz chaotic attractor as a training and test set for adaptive learning training of big data fuzzy information extraction, employs phase space reconstruction technology to conduct autocorrelation feature matching processing on the chaotic attractor load feature quantities of big data, extracts the average mutual information feature quantity of fuzzy information, and combines association rule fuzzy pairing methods to perform big data chaotic partitioning, thereby achieving optimized clustering of fuzzy information. Based on the data clustering results, accurate fuzzy information extraction is realized. The extracted high-dimensional fuzzy information undergoes feature compression to reduce computational overhead. Simulation results demonstrate that the proposed method exhibits favorable clustering performance for fuzzy information extraction from big data sample sequences, strong resistance to inter-class perturbations, and high accuracy probability for fuzzy information extraction, holding significant application value in data mining and feature extraction.

Full Text

Preamble

Fuzzy Information Extraction Based on Improved Chaotic Partition Algorithm

Wan Fucheng
(Key Laboratory of National Language Intelligent Processing, Lanzhou 730030, China)

Abstract: In big data environments, interference from small disturbances between data affects fuzzy information extraction, leading to poor clustering characteristics of extracted information. This paper proposes a fuzzy information extraction method based on an improved chaotic partition algorithm. The method performs distributed structural reorganization of high-dimensional data information flows and uses the Lorenz chaotic attractor as a training test set for adaptive learning training in big data fuzzy information extraction. Phase space reconstruction technology is employed for autocorrelation feature matching of chaotic attractor load feature quantities in big data, extracting the average mutual information feature quantity of fuzzy information. Association rule fuzzy pairing methods are combined for big data chaotic partitioning to achieve optimal clustering of fuzzy information, enabling accurate fuzzy information extraction based on data clustering results. Feature compression is performed on the extracted high-dimensional fuzzy information to reduce computational overhead. Simulation results demonstrate that this method exhibits good clustering performance for fuzzy information extraction from big data sample sequences, strong resistance to inter-class disturbances, and high accuracy probability for fuzzy information extraction, offering significant application value in data mining and feature extraction.

Keywords: big data; chaos; partition algorithm; clustering; fuzzy information extraction

0 Introduction

Research on related algorithms has received significant attention. The study of fuzzy information extraction methods in big data environments is based on big data clustering and information partitioning. According to the classification attributes of big data, information partitioning is performed to extract clustering information features. Using feature decomposition and association rule mining methods, fuzzy information extraction from big data is achieved. Common big data fuzzy information extraction methods include HPCC (High Performance Computing Cluster) extraction methods, Roxie (HPCC Data Delivery Engine) information clustering feature extraction methods, frequent itemset mining-based feature extraction methods, and fuzzy C-means clustering-based fuzzy data information extraction methods [2,3]. By extracting attribute partition features of fuzzy big data and employing corresponding clustering algorithms, fuzzy information extraction from big data is realized. Combined with correlation-based information fusion methods, fuzzy information detection capability is improved.

Based on the above principles, relevant scholars have conducted research on fuzzy information extraction algorithms and achieved certain results. Literature [4] proposes a hybrid big data similarity information extraction method based on inter-class closed frequent itemset mining, which uses multi-level se-

mantic feature extraction for big data information mining and achieves big data fuzzy partitioning based on user Jaccard similarity information to improve information fusion clustering capability. However, this method suffers from high computational overhead and strong sparsity in similarity feature distribution when performing fuzzy information extraction. Literature [5] proposes a fuzzy big data information extraction method based on discrete Gaussian random testing, which uses piecewise linear fusion for Gaussian random information feature extraction of fuzzy information and employs matched filtering for redundant information removal to enhance the statistical analysis capability of big data fuzzy information extraction. However, this method experiences significant inter-class interference in feature information fusion for fuzzy sets, easily leading to misclassification of fuzzy information.

To address the above problems, this paper proposes a fuzzy information extraction method based on an improved chaotic partition algorithm by leveraging the random clustering and anti-disturbance properties of chaos. First, a fuzzy information flow model for big data is constructed to perform information reorganization and phase space reconstruction. Then, the chaotic partition algorithm is used for feature extraction and data clustering to achieve optimized fuzzy information extraction from big data. Finally, experimental analysis demonstrates the superior performance of the proposed method in improving the accuracy of big data fuzzy information extraction.

1.1 Fuzzy Information Distributed Structure Reorganization

To achieve fuzzy information extraction from big data, chaotic partition methods are employed for association rule mining to extract fuzzy information feature quantities from big data. The first step in chaotic sequence analysis of big data fuzzy information time series is phase space reconstruction [6], which enables distributed structural reconstruction and data structure analysis of fuzzy information. Assuming the observed time series of the big data information flow to be mined, denoted as $X(n)$, is a set of non-stationary broadband time series, structural mapping of fuzzy information is performed in an m-dimensional distributed feature space to obtain the distributed reorganization structure of big data as follows:

$$X(n) = \{x(n), x(n+\tau), \cdots, x(n+(m-1)\tau)\}, \quad n = 1, 2, \cdots, N$$

where τ represents the sampling time delay of big data in high-dimensional space. Using adaptive chaotic training methods for feature fusion and performing phase trajectory evolution analysis in high-dimensional feature space [7], the distribution trajectory of reconstructed big data fuzzy information is obtained as:

$$X = \begin{bmatrix} x(1) & x(1+\tau) & \cdots & x(1+(m-1)\tau) \\ x(2) & x(2+\tau) & \cdots & x(2+(m-1)\tau) \\ \vdots & \vdots & \ddots & \vdots \\ x(N-(m-1)\tau) & x(N-(m-2)\tau) & \cdots & x(N) \end{bmatrix}$$

where all quantities are dimensionless after reduction, t represents the sampling time for fuzzy data; x, y, z denote the partition variables of the Lorenz chaotic attractor; σ, r, b are fuzzy constraint parameters. Using the fourth-order Runge-Kutta method for chaotic partitioning of fuzzy information, for a feature vector X_i of fuzzy information, a neighboring trajectory vector X_j is selected in the finite domain of chaotic partitioning. Let d_{ij} be the distance between X_i and X_j, the Euclidean distance of clustering centers for big data chaotic partitioning is obtained as:

$$d_{ij} = \|X_i - X_j\| = \sqrt{\sum_{k=1}^{m}(x_{ik} - x_{jk})^2}$$

1.2 Big Data Phase Space Reconstruction

Based on distributed structure reorganization of high-dimensional data information flow, big data phase space reconstruction is performed. Using the Lorenz chaotic attractor as the training test set for adaptive learning training of big data fuzzy information extraction [8], the two subspaces in phase space are S and Q respectively. Subspace S is composed of solution vectors s_i from the fuzzy training set, satisfying the conditional probability $P(s_i)$ of clustering attributes of the Lorenz chaotic attractor. Subspace Q is composed of solution vectors q_j from the big data fuzzy test set, with corresponding detection probability $P(q_j)$ for big data fuzzy information extraction. Under chaotic partition training, the phase space reconstruction model for big data fuzzy information is constructed as:

$$H(S) = -\sum_{i=1}^{n_s} P(s_i)\log P(s_i)$$

$$H(Q) = -\sum_{j=1}^{n_q} P(q_j)\log P(q_j)$$

$$I(Q,S) = H(Q) - H(Q|S)$$

where $P(s_i)$ represents the probability of fuzzy information semantic concept set s_i appearing in chaotic partition region S, and similarly, $P(q_j)$ represents the probability of fuzzy information ontology feature concept set q_j appearing in chaotic partition region Q. Under the condition of consistent similarity, the

average mutual information satisfying fuzzy information clustering conditions in phase space is calculated as:

$$H(Q|S) = -\sum_{s_i \in S} \sum_{q_j \in Q} P(s_i, q_j) \log \frac{P(s_i, q_j)}{P(s_i)}$$

representing the classification attribute set satisfying condition $P(s_i|q_j)$. Using the average mutual information fusion method, high-dimensional feature quantities of fuzzy information are extracted based on the above phase space reconstruction structure. Combined with the chaotic partition algorithm for fuzzy clustering, the clustering performance of fuzzy information mining is improved.

2.1 Improved Chaotic Partition Algorithm

Based on the preprocessing of distributed structure reorganization and phase space reconstruction of high-dimensional data information flow, the big data fuzzy information extraction algorithm is optimized. This paper proposes a fuzzy information extraction method based on an improved chaotic partition algorithm. Using the Lorenz chaotic attractor as the training test set, fuzzy attribute partitioning of big data is performed. The Lorenz chaotic attractor expression is given as:

$$\begin{cases} \frac{dx}{dt} = \sigma(y - x) \\ \frac{dy}{dt} = x(r - z) - y \\ \frac{dz}{dt} = xy - bz \end{cases}$$

Adaptive learning is performed in the reconstructed phase space. When the embedding dimension of phase space increases from m to $m + 1$ and time delay increases from τ to $\tau + 1$, the optimized clustering center value of chaotic partition for fuzzy information extraction is expressed as $R_{tol}^{(m+1)}$. After chaotic partitioning, the association rule fuzzy feature vector set R_{tol} satisfies inter-class balance, and the extracted eigenvalue clustering fusion degree is sorted as:

$$R_{tol}^{(m+1)} = \frac{1}{N} \sum_{k=1}^{N} \left[A_k - \frac{1}{N} \sum_{l=1}^{N} x_{nl}^{(m+1)} \right] > 0$$

Analysis shows that the dimension of certain information extracted using the above method is relatively high, requiring feature compression. The steps for feature compression are described as:

a) Calculate the d-dimensional fuzzy information feature vector X_n in the chaotic partition region, and compute the fuzzy information partition dispersion d_{ij} based on partition clustering attribute values. Calculate the l

eigenvalues and corresponding high-dimensional feature quantities of fuzzy information.

b) According to the center vector of the chaotic partition attractor, obtain the dispersion R_{tol} of fuzzy clustering, where *tol* is the partition threshold of the Lorenz chaotic attractor. Based on empirical values, *tol* can be taken as 15, and A represents the decision threshold for fuzzy information chaotic partitioning, typically set to 2.

c) Using decision statistical regression analysis methods, perform feature compression in high-dimensional space and output feature quantities $Y = \{y_1, y_2, \cdots, y_l\}$. Use K-L transform for feature sorting to obtain the feature arrangement.

d) Output the d-dimensional fuzzy information extraction feature quantity, extract the first d feature vectors as the training set, output $W = [y_1, y_2, \cdots, y_d]$, and obtain the fuzzy transformation matrix $W^* = T^T X$. Extract the average mutual information feature quantity of fuzzy information in the closed frequent item region of chaotic partition as $J(X) = \{J_1(X), J_2(X), \cdots, J_l(X)\}$.

e) Output the information extraction result after feature compression as $J_1(X) \geq J_2(X) \geq \cdots \geq J_d(X)$. Through the above processing, the dimension of fuzzy information feature quantities output after chaotic partition is reduced from L to d, thereby decreasing computational overhead.

2.2 Information Extraction and Feature Compression

Assume the distribution time series of fuzzy information $\{X_n\}, n = 1, 2, \cdots, N$ is the original big data feature distribution set to be partitioned. Under chaotic partition processing, the feature distribution after chaotic partitioning is obtained as $X = \bar{X} + \eta$, where η is observation noise. Among d big data distribution sources, phase space reconstruction technology is used for autocorrelation feature matching processing of chaotic attractor load feature quantities in big data [11], obtaining feature matching results.

Using singular value decomposition to decompose fuzzy information eigenvalues, d-dimensional chaotic partitioned big data is obtained. By analogy with the above method, the output feature values of fuzzy information extraction are obtained as:

$$\Sigma = \mathrm{diag}(\sigma_1, \sigma_2, \cdots, \sigma_m) \in \mathbb{R}^{m \times m}$$

---

3 Simulation Experiments and Results Analysis

To verify the application performance of the proposed method for fuzzy information extraction under large datasets, simulation experiments are conducted. The experiments use Visual C++ for algorithm compilation and MATLAB for data processing programming design. Data extraction and information clustering analysis are performed within the MapReduce programming framework. The big dataset originates from the Hadoop cloud platform, with the big data processing tool component being Thor (HPCC Data Refinery Cluster). The big data test sample set is OAEI (Ontology Alignment Evaluation Initiative), which stores multi-version semantic data and serves as a test set with good information coverage capability for this experiment. The data sampling time interval is 2.4s, the test set size is 2400 Gbit, and the sampling length of the big data training set is 1024. The initial parameter values of the Lorenz chaotic attractor are set as $[x, y, z] = [1, 0, -1]$, $[\sigma, r, b] = [16, 45.92, 4.0]$. The training step size for chaotic partition iteration is $h = 0.01$. The information distribution interval for big data clustering is $[0, 1000]$. The interference intensity in information extraction is -10 dB. Based on the above simulation environment and parameter settings, big data fuzzy information extraction simulation experiments are conducted. Three groups of test samples are taken, and the time-domain waveforms are obtained as shown in Figure 1 [Figure 1: see original paper].

Using the data samples from Figure 1 as the research object, the proposed method is applied for chaotic partition clustering and information extraction. The chaotic partition and fuzzy information extraction results for each group of samples under optimal phase space embedding dimension and time delay are obtained as shown in Figures 2-4 [Figure 2: see original paper][Figure 3: see original paper][Figure 4: see original paper].

Figure 2 corresponds to the first group of data samples, whose time-domain waveform fluctuations are relatively stable, indicating strong information coverage capability. As shown in Figure 2, using the proposed method for chaotic partition clustering and information extraction yields clear chaotic partition and fuzzy information extraction results under optimal phase space embedding dimension and time delay, with good attribute clustering performance. This demonstrates that the proposed method can achieve accurate information extraction from big data. This is because the method combines decision criteria and decision statistics for big data fuzzy information extraction and adaptive learning training, along with feature extraction, making it more suitable for the data environment.

Figure 3 corresponds to the second group of data samples, which exhibit fewer overall fluctuations with consistent amplitude, indicating relatively dispersed data distribution and high dimensionality, resulting in poor inherent clustering performance. However, after applying the proposed method for chaotic partition and fuzzy information extraction, the attribute clustering performance of chaotic partitioning is improved, and the inter-class convergence of informa-

tion extraction is enhanced. This is due to the proposed method' s utilization of adaptive chaotic training for feature fusion and phase trajectory evolution analysis in high-dimensional feature space, which reconstructs big data fuzzy information and thereby improves chaotic partitioning and fuzzy information extraction capability.

Figure 4 corresponds to the third group of data samples, whose time-domain waveform fluctuations are irregular with large amplitude, indicating significant inter-class interference. Nevertheless, the proposed method still maintains good chaotic partition and fuzzy information extraction performance, demonstrating strong anti-inter-class interference capability. This is primarily because the proposed method simultaneously performs feature compression when extracting high-dimensional information, and the dimensionality reduction effectively decreases the interference level of processed data samples, thereby improving performance.

Analysis of the above results reveals that the proposed method for big data information extraction exhibits good attribute clustering performance for chaotic partitioning, high inter-class convergence for information extraction, and strong anti-inter-class disturbance capability, providing excellent fuzzy information extraction ability.

For performance comparison, different methods are applied to extract fuzzy information from each group of samples, analyzing extraction accuracy and computational overhead parameters. The performance comparison results are shown in Table 1 and Table 2 . Analysis of the results demonstrates that the proposed method achieves smaller time overhead for fuzzy information extraction, higher extraction accuracy, and superior performance compared to traditional methods.

---

4 Conclusion

In big data environments, accurate feature analysis and extraction of fuzzy information, rapid mining of required data information, data integration, and achievement of information sharing and accurate link transmission are crucial. This paper proposes a fuzzy information extraction method based on an improved chaotic partition algorithm by utilizing the random clustering and anti-disturbance properties of chaos. The Lorenz chaotic attractor is used for big data chaotic partitioning, fuzzy information extraction is implemented in the reconstructed phase space, and dimensionality reduction is performed on the extracted high-dimensional data. The research demonstrates that the proposed method can improve the accuracy of fuzzy information extraction, reduce computational overhead, and provide strong anti-inter-class interference capability with superior performance.

References

[1] 梁聪刚, 王鸿章. 微分进化算法的优化研究及其在聚类分析中的应用 [J]. 现代电子技术, 2016, 39 (13): 103-107. (Ling Conggang, Wang Hongzhang. Optimization research on differential evolution algorithm and its application in clustering analysis [J]. Modern Electronic Technology, 2016, 39 (13): 103-107.)

[2] 米捷, 张鹏, 于海鹏. 粒子群差分扰动优化的聚类算法研究 [J]. 河南工程学院学报, 2016, 28 (1): 63-68. (Mi Jie, Zhang Peng, Yu Haipeng. Large data clustering algorithm based on particle swarm differential perturbation optimization [J]. Journal of Henan University of Engineering (Natural Science Edition), 2016, 28 (1): 63-68.)

[3] 邢淑凝, 刘方爱, 赵晓晖. 基于聚类划分的高效用模式并行挖掘算法 [J]. 计算机应用, 2016, 36 (8): 2202-2206. (Xing Shuning, Liu Fang' ai, Zhao Xiaohui. Parallel high utility pattern mining algorithm based on cluster partition [J]. Journal of Computer Applications, 2016, 36 (8): 2202-2206.)

[4] Palomares I, Martinez L, Herrera F. A consensus model to detect and manage non-cooperative behaviors in large scale group decision making [J]. IEEE Trans on Fuzzy System, 2014, 22 (3): 516-530.

[5] 李梓杨, 于炯, 卞琛, 等. 基于负载感知的数据流动态负载均衡策略 [J]. 计算机应用, 2017, 37 (10): 2760-2766. (Li Ziyang, Yu Jiong, Bian Chen, et al. Dynamic data stream load balancing strategy based on load awareness [J]. Journal of Computer Applications, 2017, 37 (10): 2760-2766.)

[6] 孙力娟, 陈小东, 韩崇, 等. 一种新的数据流模糊聚类方法 [J]. 电子与信息学报, 2015, 37 (7): 1620-1625. (Sun Lijuan, Chen Xiaodong, Han Chong, et al. New Fuzzy-Clustering Algorithm for Data Stream [J]. JEIT, 2015, 37 (7): 1620-1625.)

[7] 邢长征, 刘剑. 基于近邻传播与密度相融合的进化数据流聚类算法 [J]. 计算机应用, 2015, 35 (7): 1927-1932. (XING Changzheng, LIU Jian. Evolutionary data stream clustering algorithm based on integration of affinity propagation and density. Journal of Computer Applications, 2015, 35 (7): 1927-1932.)

[8] 侯森, 罗兴国, 宋克. 基于信息源聚类的最大熵加权信任分析算法 [J]. 电子学报, 2015, 43 (5): 993-999. (Hou Sen, Luo Xingguo, Song Ke. A maximum entropy weighted trust-analysis algorithm based on sources clustering [J]. Chinese Journal of Electronics, 2015, 43 (5): 993-999.)

[9] 毕安琪, 王士同. 基于 Kullback-Leiber 距离的迁移仿射聚类算法 [J]. 电子与信息学报, 2016, 38 (8): 2076-2084. (Bi Anqi, Wang Shitong. Transfer affinity propagation clustering algorithm based on Kullback-Leiber Distance [J]. JEIT, 2016, 38 (8): 2076-2084.)

[10] 刘俊, 刘瑜, 何友, 等. 杂波环境下基于全邻模糊聚类的联合概率数据互联算法 [J]. 电子与信息学报, 2016, 38 (6): 1438-1445. (Liu Jun, Liu Yu, He You, et al. Joint Probabilistic data association algorithm based on all-neighbor fuzzy clustering in clutter [J]. JEIT, 2016, 38 (6): 1438-1445.)

[11] 吴鸿华, 穆勇, 屈忠锋, 等. 基于面板数据的接近性和相似性关联度模型 [J]. 控制与决策,

2016, 31 (3): 555-558. (Wu Honghua, Mu Yong, Qu Zhongfeng, et al. Similarity and nearness relational degree based on panel data [J]. Control and Decision, 2016, 31 (3): 555-558.)

Note: Figure translations are in progress. See original paper for figures.