Anomaly Detection Using Maximum Entropy Fuzzy Clustering Algorithm Enhanced with Soft Computing Techniques

Chunhua Liang[1,2]
[1]Institute for History of Science and Technology, Shanxi University, Taiyuan, 030006, China
[2]Information Institute, Shanxi Finance and Taxation College, Taiyuan, 030024, China
E-mail: lchua2023@126.com

Keywords: anomaly detection, fuzzy clustering, hilbert schmidt independence criterion, unsupervised learning, soft computing

Received: June 2, 2024

With the continuous growth of data volume, anomaly detection has become an important link in the data processing process. In view of the maximum entropy fuzzy clustering algorithm, an anomaly detection method combining soft computing is proposed. During the process, the K-means algorithm was used to construct the algorithm foundation, followed by the establishment of an objective function for maximum entropy calculation and the introduction of the Hilbert Schmidt independence criterion for variable extraction. Then it conducts data migration and calculates the exception score. The experimental results showed that the proposed method could be reduced to 113 in the Iris data set when the convergence curve was tested. When the calculation time was tested, the calculation time of the research method was only 2697ms when the sample size reached 10000. When the accuracy and purity tests were carried out, the accuracy and purity of the research method were 87.7% and 87.6% in the MR Dataset. In the Leaf dataset, the standardized mutual information index reached 0.6837 and the FM index reached 0.3903. The lowest Davies-Bouldin index was 0.71. The area enclosed by the receiver operation characteristic curve and the horizontal coordinate of the research method was the largest. The results indicate that the research method has high accuracy and computational efficiency in data anomaly detection and can provide effective technical references for anomaly detection.

Povzetek: Za odkrivanje anomalij v podatkih avtorji kombinirajo dve metodi: (a) mehko gručenje z maksimalno entropijo ter (b) mehko računanje za doseganje boljših rezultatov.

1 Introduction

As the boost of information technology and the rapid growth of data, anomaly detection (AD) has become an important research topic in the field of data security. The existing AD methods still face difficulties in dealing with complex data environments [1]. Large scale datasets often have high-dimensional features, and traditional feature representation methods often find it difficult to effectively capture useful information in the data, resulting in a decrease in the performance of AD [2, 3]. Abnormal samples are usually rare in real datasets, which makes it difficult for traditional machine learning algorithms to accurately identify abnormal samples and prone to false positives or omissions [4-6]. The K-means algorithm (KMA) can divide data samples into different clusters, providing clustering results as a basis for subsequent AD. The maximum entropy calculation method can effectively estimate the distribution of data and help to accurately detect anomalies. Transfer learning can use the knowledge and experience of the source domain to assist the AD task in the target domain. In view of this, this research proposes an AD method combining soft computing (SC) and maximum entropy fuzzy clustering (FC) algorithm, to provide a feasible reference technology for information technology.

The research mainly focuses on four aspects. The first part discusses the current research results on AD methods and mushroom clustering. The second part is the design of the maximum entropy FC migration AD method in view of SC. The third part is to determine the effectiveness of the research method. The last part is a summary and discussion of the entire text.

2 Related works

As the boost of information technology, the volume of data is becoming larger and larger. Data AD is an important means to protect data security. Some scholars have conducted relevant research on data AD. Ni et al. proposed a detection method in view of convolutional neural network (CNN) to solve the problem of abnormal data in structural monitoring. This method used a neural network to extract features from the signal, followed by automatic encoder structure for data reconstruction. The experiment illustrated that the proposed method had high accuracy [7]. Zhou et al. proposed a detection method in view of neural network to solve the anomaly problem in industrial big data. This method represented the data in reduced dimensions and reconstructed and quantized the

loss function. The experiment demonstrated that the proposed method has high accuracy [8]. Mao et al. put forward a detection method in view of unsupervised learning (UL) method to solve the problem of abnormal monitoring in civil structure detection data. This method introduced adversarial networks and transforms data containing time series into images. The experiment showed that the proposed method had good robustness [9]. Liu et al. proposed a detection method in view of Federated learning to solve the problem of fault data AD in industrial manufacturing process. This method used scattered edge device data for model training, and added attention mechanism and long short-term memory to maintain fine-grained features. The experiment indicated that the proposed method had a high accuracy in AD [10]. Khaledian et al. proposed a detection method in view of UL for data AD in power system. This method marked fault data and introduced the concept of stack integration for algorithm optimization. The experiment demonstrated that the proposed method could perform accurate data classification and AD [11].

Some scholars have conducted research on clustering algorithms. Liu et al. proposed an efficiency improvement method in view of FC for image size adjustment. This method mapped the pixels of the image spatially, then compressed the clustering sample size and introduced the C-Means clustering algorithm for main data extraction. The experiment illustrated that the proposed method could effectively improve the speed of image adjustment [12]. Zhang et al. proposed a processing method in view of FC for image segmentation in medical image analysis. This method initialized the cluster center and searches for approximate data. The experiment indicated that the proposed method had good robustness [13]. Jiang et al. proposed an auxiliary method in view of clustering fuzzy to improve the efficiency of rock image segmentation. This method first preprocessed the original image, and then used CNN to remove irrelevant feature information. The experiment demonstrated that the proposed method has a high computational speed [14]. Hu et al. proposed a calculation method in view of FC for data processing in Complex network. This method introduced the generalized momentum method, which was then trained and tested on multiple datasets. Experimental results showed that the proposed method had high Rate of convergence and accuracy [15]. Feng et al. proposed a partition method in view of FC to solve the accuracy problem of soft partition of complex data. This method established a new feature space to support data reconstruction and introduces neural networks for data recovery encoding. The experiment demonstrated that the proposed method had high partitioning accuracy [16].

In summary, although clustering fuzzy algorithms have been studied and applied in various fields, there is still relatively little research on data AD. In view of this, this research proposes an AD algorithm that integrates the maximum entropy FC algorithm of SC, to provide some reference for data AD. The related works summary Table is shown in Table 1.

Table 1: Related works summary table

Author and year	Method	Data set	Performance index
Ni et al., 2020	CNN-based method	Structural monitoring data	High accuracy
Zhou et al., 2020	Neural network method	Industrial big data	High accuracy, loss function reconstruction
Mao et al., 2021	Unsupervised learning method	Civil structure monitoring data	Good robustness
Liu et al., 2020	Federated learning	Industrial manufacturing process data	High accuracy, incorporation of attention mechanism and LSTM
Khaledian et al., 2020	UL method for power system	Power system data	Accurate data classification and anomaly detection
Liu et al., 2021	Fuzzy clustering efficiency improvement method	Image data	Effective improvement in speed
Zhang et al., 2021	Fuzzy clustering for medical image segmentation	Medical image data	Good robustness
Jiang et al., 2020	Fuzzy clustering auxiliary method	Rock image data	High computational speed
Hu et al., 2021	Fuzzy clustering calculation method	Complex network data	High convergence rate and accuracy
Feng et al., 2020	FC partition method for complex data	Complex data soft partition	High partitioning accuracy

3 Design of maximum entropy FC migration AD method in view of SC

Abnormal detection of data can ensure higher reliability of information systems. This section will focus on the technical means and evaluation system used in the maximum entropy FC migration AD method fused with SC.

3.1 Construction of maximum entropy clustering algorithm introducing hilbert schmidt independent criterion

Cluster analysis is a statistical analysis method, which belongs to UL method. This method can classify unlabeled objects in view of the potential laws and similar features of different entities [17, 18]. The increasing magnitude of data in various systems makes it difficult to extract potential patterns and perform AD in the data. Therefore, using clustering analysis can help with AD [19]. This study uses the KMA as the basis for constructing detection algorithms. The basic process of the KMA is shown in Figure 1.

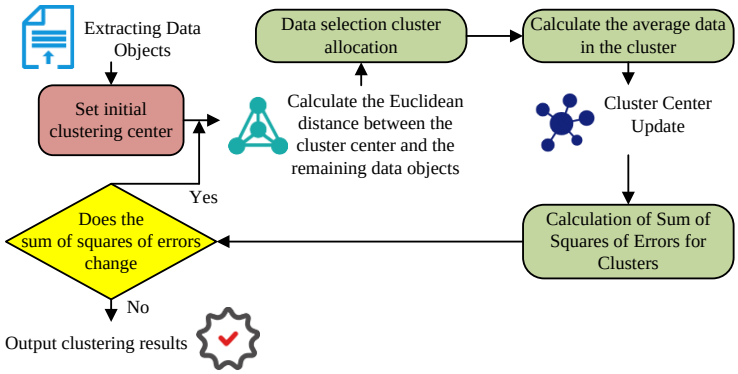

Figure 1: Basic process of KMA

In Figure 1, when running the KMA, it is first necessary to extract data objects from the given dataset for setting the initial clustering center. It calculates the Euclidean distance between the cluster center and the remaining data objects, and allocates the data to the cluster with the closest cluster center. It calculates the average data in each cluster, updates the cluster center, and then calculates the sum of squares of errors for all clusters. If the sum of squared errors changes, calculate and iterate the Euclidean distance again until the clustering is completed and the clustering result is output when the sum of squared errors remains unchanged. The calculation of Euclidean distance is shown in equation (1).

$$d\left(x,C_i\right)=\sqrt{\sum_{j=1}^{m}\left(x_j-C_{ij}\right)^2} \tag{1}$$

In equation (1), d represents the Euclidean distance. x represents the data object. C represents the center of the cluster. m represents the dimension of the data object, i represents the index of the data objects in the dataset. j represents the index that represents the cluster center. The calculation of the sum of squares of errors is shown in equation (2).

$$S_E=\sum_{i=1}^{k}\sum_{x\in C_i}\left|x-M_i\right|^2 \tag{2}$$

In equation (2), S_E represents the sum of squares of errors. M represents the mean of the cluster. k represents the number of cluster clusters. The idea of maximum entropy in machine learning is to maximize the objective function under set conditions. In clustering problems, samples in the same cluster are more similar and have less uncertainty. In the context of clustering, the probability distribution of data points within a given cluster not only adheres to the principle of entropy maximization but also satisfies specific constraints. The objective function for maximum entropy calculation is shown in equation (3).

$$\begin{cases} minimize\ J\left(U,V\right)=\sum_{i=1}^{m}\sum_{j=1}^{m}\mu_{ij}\left\|x_i-v_j\right\|^2+\alpha\sum_{i=1}^{m}\sum_{j=1}^{m}\mu_{ij}\ln\mu_{ij} \\ subject\ to\ \mu_{ij}\in[0,1],\sum_{j=1}^{m}\mu_{ij}=1 \end{cases} \tag{3}$$

In equation (3), μ represents the membership degree of the data object to the cluster center. v represents the clustering center. α represents the coefficient of the regularization term. The larger the coefficient of the regularization term, the greater the impact of entropy on the results. When optimizing and updating the clustering results, the iterative updating formula is designed by using the Lagrange conditional extremum method. The function includes the negative value of the original entropy and the penalty term for violating the constraint. By introducing Lagrange multipliers, constraints are integrated into the optimization process. The objective function belongs to convex function. When designing the updating equation

of the cluster center and membership degree, other variables are set to be in a constant state. The update iteration of the maximum entropy algorithm is shown in equation (4). In equation (4), when calculating v_j, μ_{ij}, fix the other value separately. The process of obtaining maximum entropy algorithm is shown in Figure 2.

$$\begin{cases} v_j = \dfrac{\sum_{i=1}^{m} \mu_{ij} x_i}{\sum_{i=1}^{m} \mu_{ij}} \\ \mu_{ij} = \dfrac{\exp\left(-\dfrac{\left\|x_i - v_j\right\|^2}{\alpha}\right)}{\sum_{k=1}^{m} \exp\left(-\dfrac{\left\|x_i - v_k\right\|^2}{\alpha}\right)} \end{cases} \tag{4}$$

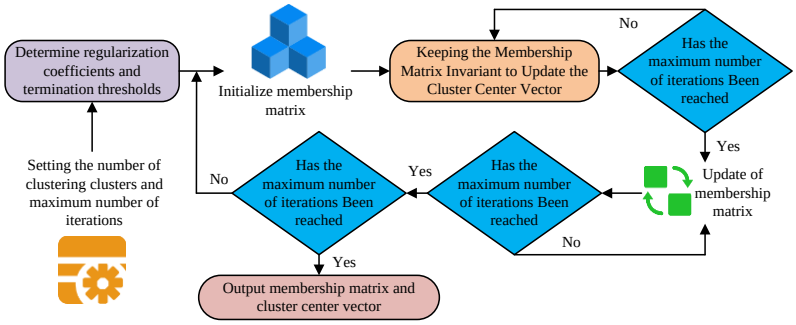

Figure 2: Maximum entropy algorithm process

Figure 2 shows that at runtime, the number of cluster clusters and maximum iteration times are first set, and after determining the regularization coefficient and termination threshold, the input dataset is input for calculation. After initializing the membership matrix, start recording the number of iterations, keep the membership matrix unchanged, and update the cluster center vector until the maximum number of iterations is reached, ending the loop. After entering the update step of the membership matrix, keep the cluster center vector unchanged until the cycle ends when the maximum number of iterations is reached. Finally, output the final membership matrix and cluster center vector. The Hilbert Schmidt independence criterion (HSIC) extracts random variables from the original feature space and maps them into the reconstructed kernel space. It searches for the independent relationships between other variables and random variables, and can estimate the dependency relationships between different clusters [20]. The Cross-covariance operation of the characteristic relationship is shown in Formula (5).

$$R_{y,z} = E_{yz}\left\{\left[\phi(x) - E_y\left(\phi(y)\right)\right] \otimes \left[\phi(z) - E_z\left(\phi(z)\right)\right]\right\} \tag{5}$$

In formula (5), $R_{y,z}$ represents the Cross-covariance matrix. ϕ represents the variable mapping relationship. E_{yz} represents the Joint probability distribution of variables. E_y, E_z represent marginal distribution expectations. $\otimes$ represents tensor product. The HSIC can be expressed by the Hilbert Schmidt norm of the cross-covariance matrix, as shown in equation (6).

$$HSIC\left(P_{yz}, F, G\right) = \left\|R_{y,z}\right\|_{HS}^2 \tag{6}$$

In equation (6), $\|\cdot\|_{HS}$ represents the Hilbert Schmidt norm. P_{yz} is the joint probability between random variables. F, G are the mapping result of the variable. If two variables are independent of each other, the HSIC takes a value of 0. The higher the value of the HSIC, the stronger the correlation between variables.

3.2 Design of migration AD algorithm in view of SC

It integrates the HSIC into maximum entropy clustering to obtain a maximum entropy algorithm in view of minimum dependency. It adds the Hilbert Schmidt independent criterion value of the cluster center to the objective function to ensure the maximum intra cluster similarity of the clustering results. When solving the objective function, the trace operation of the matrix is required, and then the transformed objective function is solved using the Lagrange conditional extreme value method. The calculation of Lagrange equation is shown in equation (7).

$$\begin{cases} L_{lin}(U,V) = \sum_{i=1}^{m}\sum_{j=1}^{m} \mu_{ij} \left\|x_i - v_j\right\|^2 + \alpha \sum_{i=1}^{m}\sum_{j=1}^{m} \mu_{ij} \ln_{ij} + \sum_{i=1}^{m} \beta_i \left(\sum_{j=1}^{m} \mu_{ij} - 1\right) \\ + \lambda \dfrac{1}{(P-1)^2} + \alpha \sum_{i=1}^{m}\sum_{j\neq i}^{m} \left(v_i^T H v_j\right)^2 \end{cases} \tag{7}$$

In formula (7), β_i represents Lagrange multiplier. H represents the concentrated matrix. P represents the feature space dimension. λ represents the equilibrium parameter. The smaller the balance parameter, the less attention is paid to the similarity information between clusters. When it is 0, the original maximum entropy algorithm is equivalent to the algorithm. When

updating the membership matrix, fix the cluster center vector and solve for the partial derivative of the membership matrix to obtain the updated membership matrix as shown in equation (8).

$$\mu_{ij}=\frac{\exp\left(-\frac{\left\|x_i-v_j\right\|^2}{\alpha}\right)}{\sum_{k=1}^{m}\exp\left(-\frac{\left\|x_i-v_k\right\|^2}{\alpha}\right)} \tag{8}$$

In equation (8), exp is obtained from the empirical formula consisting of Lagrange multiplier and the coefficient of the canonical term, and represents the Exponential function with e as the base. When calculating the center of a cluster, the membership degree of the cluster is fixed, and the center vector is solved by partial derivatives to obtain the relationship as shown in equation (9).

$$\begin{cases}-2\sum_{i=1}^{m}\mu_{ij}\left(x_i-v_j\right)+\frac{2\lambda}{\left(P-1\right)^2}\sum_{i=1}^{m}Hv_iv_i^THv_j=0\\ \sum_{i=1}^{m}\mu_{ij}x_i=\left(\sum_{i=1}^{m}\mu_{ij}I+\frac{\lambda}{\left(P-1\right)^2}\sum_{i\neq j}^{m}HK_iH\right)v_j\end{cases} \tag{9}$$

In formula (9), I is the Identity matrix of $P\times P$. K_i surface linear kernel function. When $\sum_{i=1}^{m}\mu_{ij}I+\frac{\lambda}{\left(P-1\right)^2}\sum_{i\neq j}^{m}HK_iH$ is a full matrix, it has a unique Inverse matrix. If it is not a full matrix, the Singular value decomposition method is used to calculate the pseudo-Inverse matrix. The cluster center update is shown in equation (10).

$$v_j=\left(\sum_{i=1}^{m}\mu_{ij}I+\frac{\lambda}{\left(P-1\right)^2}\sum_{i\neq j}^{m}HK_iH\right)^{-1}\sum_{i=1}^{m}\mu_{ij}x_i \tag{10}$$

The improved update formula includes the optimization objective of the original maximum entropy algorithm, as well as the interdependence information of different cluster centers, ensuring sufficient independence of the updated cluster centers. The maximum entropy clustering algorithm in view of the HSIC is shown in Figure 3.

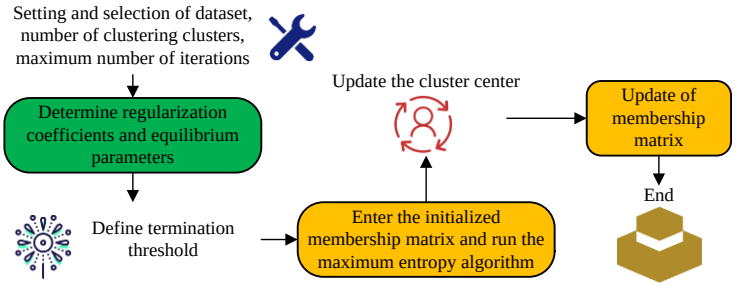

Figure 3: Maximum entropy clustering algorithm process

Figure 3 shows that the maximum entropy clustering algorithm in view of the HSIC needs to first set and select parameters such as the dataset before running, then determine the regularization coefficient and balance parameters, and determine the termination threshold. Then it starts calculating, initializing the membership matrix, zeroing the number of algorithms runs, inputting the initialized membership matrix, and running the maximum entropy algorithm. It takes the output result of the maximum entropy algorithm as the input value for subsequent iterations and sets the number of runs of the new loop to zero. Then it keeps the membership matrix unchanged and updates the cluster center until the maximum number of iterations is reached, ending the cycle. After entering the update step of the membership matrix, keep the cluster center unchanged until the cycle ends when the maximum number of iterations is reached. In each iteration process, the cluster center and membership matrix need to be continuously updated. When updating the cluster center, it is necessary to consider the information of the cluster center and membership matrix in the previous iteration. Moreover, when updating the membership degree, it is necessary to consider the cluster center and sample point information. It outputs the final membership degree and cluster center as the result. When calculating the membership matrix, the time complexity is the same as the Time complexity of the original maximum entropy algorithm. The Time complexity is greatly affected by the sample size. The more samples, the more Time complexity [21]. When performing abnormal detection and analysis of data, the abnormal label of actual data is usually missing or invalid. Similar tasks have commonality, and migrating between different tasks can reduce the burden of data label

collection and reduce time and hardware costs. To address the information omission when discarding the source domain dataset label, the target domain data is combined with the source domain data after adding pseudo labels for instance migration model training. There is a negative transfer problem in UL. This study introduces ensemble learning to train multiple base models, and then combines the output results of all base models to judge the degree of instance migration in the source domain dataset. When training the base classification model of the positive instance, calculate the recall score of the model, as shown in equation (11).

$$recall = \frac{TP}{TP + FN} \tag{11}$$

In equation (11), $recall$ represents the recall score. TP represents the correctly classified instance in the confusion matrix. FN represents the instance in the confusion matrix that is wrongly classified. Update the weights of correctly classified and incorrectly classified instances, as shown in equation (12).

$$\begin{cases} w_{true} = w_i^t * \exp\left(\beta * \left|y_i - \hat{y}_t\right|\right) \\ w_{false} = w_i^t * \exp\left(\beta * \left(1 - \left|y_i - \hat{y}_t\right|\right)\right) \end{cases} \tag{12}$$

In equation (12), w represents the weight. y_i represents the true label. $\hat{y}_t$ represents the prediction label. β is calculated from the recall score. When calculating the base classification model for negative instances, update the weights of correctly classified and incorrectly classified instances after calculating the recall score, as shown in equation (13).

$$\begin{cases} w_i^{t+1}{}_{true} = w_{true} * \exp\left(\beta * \left|y_i - \hat{y}_t\right|\right) \\ w_i^{t+1}{}_{false} = w_{false} * \exp\left(\beta * \left(1 - \left|y_i - \hat{y}_t\right|\right)\right) \end{cases} \tag{13}$$

It comprehensively utilizes the migration data and labels of the source domain, and analyzes unlabeled data in the target domain to generate abnormal scores for the data. If the midpoint of the source domain is in the area near the cluster center, the abnormal score of the data is calculated as shown in equation (14).

$$a_l = \sum \frac{w_j}{d_{ij}^2} \tag{14}$$

In equation (14), a_l represents the abnormal score of the source domain. w_j is the transfer weight. d_{ij} represents the distance between a point in the target domain and its nearest neighbor. The calculation of abnormal scores in the target domain is shown in equation (15).

$$a_u = 1 - \exp\left(-\frac{\sum d_{ij}^2}{\gamma * t}\right) \tag{15}$$

In equation (15), a_u represents the abnormal score of the target domain. γ represents the proportion of abnormal points. The calculation of the comprehensive abnormal score of points is shown in equation (16).

$$a_i = a_l * w_l + a_u * \left(1 - w_l\right) \tag{16}$$

In equation (16), w_l is the contribution value of the object point. The Transfer learning part of the AD algorithm in view of soft instance migration is shown in Figure 4.

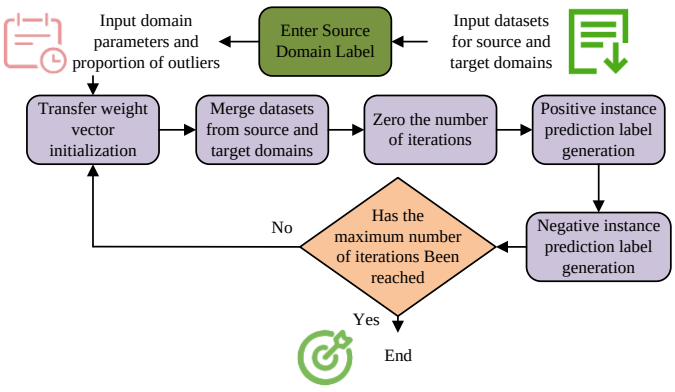

Figure 4: Algorithm Transfer learning part

Figure 4 shows that the Transfer learning part of the AD algorithm in view of soft instance migration first initializes the migration weight vector and merges the data sets of the source domain and the target domain. After resetting the iteration count to zero, it generates prediction labels for both positive and negative instances, and ends the loop when the iteration count reaches the preset upper limit. The AD process is shown in Figure 5.

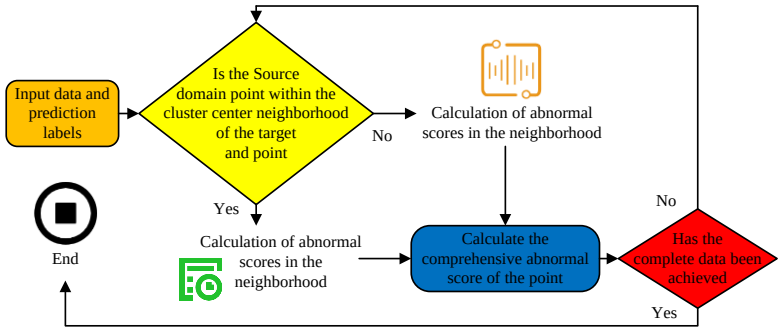

Figure 5: Abnormal detection process

Figure 5 illustrates the process by which, in the context of AD, the initial determination is made as to whether the source domain point falls within the central neighborhood of the target point cluster. Subsequently, the anomaly scores are calculated for both the neighborhood and the domain. Then it calculates the comprehensive anomaly score of the point, continuously loops to obtain complete data, ends the loop, and outputs the final anomaly score as the AD result.

4 Performance test and application analysis of maximum entropy FC migration AD method integrating SC

A good data AD method is the guarantee of accurate information analysis. This section will test the performance of the research method in AD and conduct application analysis to determine the effectiveness of the research method.

4.1 Performance test of maximum entropy FC migration AD method integrating SC

To analyze the effectiveness of the maximum entropy FC migration method fused with SC in data AD, the performance test and application analysis of the research method are carried out. When conducting performance testing, the experiment used the Iris dataset, Leaf dataset, and MR dataset as the experimental datasets. It sets the regularization coefficient on the Iris dataset to 4 and the equilibrium parameter to 40. The regularization coefficient on the Leaf dataset is 1, and the equilibrium parameter is 1. The regularization coefficient on the MR dataset is 1, and the equilibrium parameter is 1. The computing resources used in the experiment includes a personal computer with an Intel Core i7 processor and 16GB of RAM. All experiments are performed on a single computing node without the use of distributed computing resources. The operating system is Ubuntu 20.04 LTS to ensure the consistency of the experimental environment. The main hyperparameters of the LOF algorithm include the calculation radius of the local outlier factor and the significance threshold, which are adjusted according to the local density distribution of the data set. The principal hyperparameters of the DBSCAN algorithm are the number of neighbors and the minimum number of dots. These are selected on the basis of the density distribution of the data set and the detection requirements of outliers. The hyperparameters include regularization coefficient, equilibrium parameter and maximum number of iterations. The selection of hyperparameters is based on the characteristics of the data, the working principle of the algorithm and the objective function. Firstly, the convergence curve of the research method is tested and compared with the local outlier factor (LOF) algorithm and density based spatial clustering of applications with noise (DBSCAN) algorithm, as shown in Figure 6.

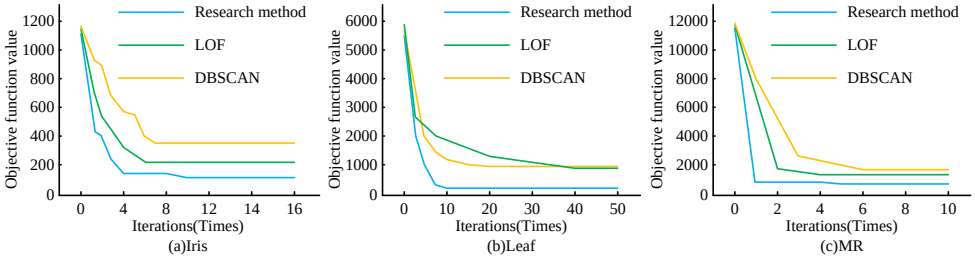

Figure 6: Convergence curve test

Figure 6 (a) shows that in the Iris dataset, the objective function value of the LOF algorithm decreased from 1147 to 214 after 6 iterations, and then remained stable. The objective function value of the DBSCAN algorithm decreased from 1166 to 361 after 7 iterations, and remains stable thereafter. The objective function value of the research method decreases from 1102 to 134 after 4 iterations, maintains until the 8th iteration, and began to decrease in the second stage. After the 9th iteration, it decreases to 113. Figure 6 (b) shows that in the Leaf dataset, the objective function value of the LOF algorithm decreased from 5864 to 892 after 40 iterations, and then remains stable. The objective function value of the research method decreased from 5426 to 264 after 10 iterations, and remains stable thereafter. Figure 6 (c) shows that in the MR dataset, the objective function value of the LOF algorithm decreases from 11876 to 1375 after 4 iterations, and then remained stable. The objective function value of the research method decreased from 11623 to 872 after one iteration, maintained until the fourth iteration, and began to decline in the second stage. After the fifth iteration, it decreases to 793. This indicates that the research method has better convergence performance. Test the calculation time of the research method, as shown in Figure 7.

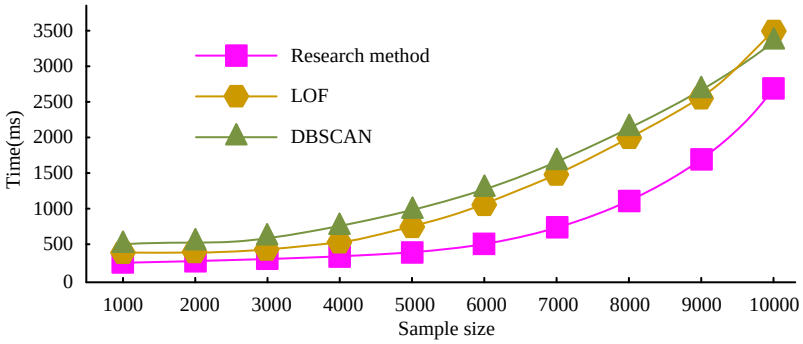

Figure 7: Computing time

Figure 7 shows that the calculation time of all three methods increases with the increase of sample size, and the LOF algorithm reaches 3498ms when the sample size increases to 10000. The DBSCAN algorithm achieves a computation time of 3321ms when the sample size increases to 10000. The research method achieved a calculation time of 2697ms when the sample size increased to 10000. The research method requires less computational time. The accuracy and purity of the research method are tested, as shown in Figure 8.

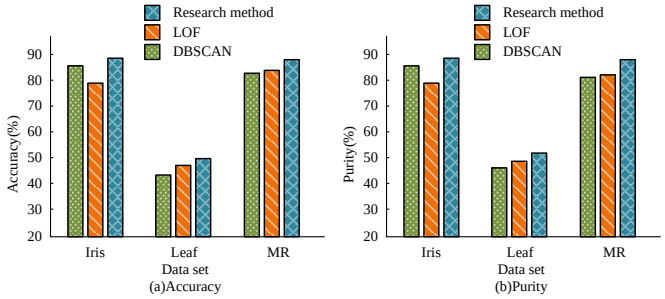

Figure 8: Accuracy and purity

Figure 8 (a) shows that in the Iris dataset, the accuracy of the DBSCAN algorithm is 85.7%, the accuracy of the LOF algorithm is 78.7%, and the accuracy of the research method is 88.4%. In the Leaf dataset, the accuracy of the DBSCAN algorithm is 42.9%, the accuracy of the LOF algorithm is 47.2%, and the accuracy of the research method is 49.5%. The accuracy of the LOF algorithm in the MR dataset is 83.5%, and the accuracy of the research method is 87.7%. Figure 8 (b) shows that in the Iris dataset, the purity and accuracy of the three algorithms are consistent. In the Leaf dataset, the purity of the DBSCAN algorithm is 46.7%, the LOF algorithm is 49.1%, and the purity of the research method is 51.8%. In the MR dataset, the purity of the DBSCAN algorithm is 80.8%, the LOF algorithm is 81.6%, and the purity of the research method is 87.6%. This indicates that the research method has higher accuracy and data purity. It tests the adjusted RAND coefficient, standardized Mutual information index and FM index of the research method, as shown in Table 2.

Table 2: Adjusted RAND coefficient, standardized Mutual information index, FM index

Data set	Method	ARI	NMI	FM
Iris	LOF	0.6639	0.6891	0.7746
	DBSCAN	0.6765	0.7004	0.7834
	Research method	0.7091	0.7308	0.8061
Leaf	LOF	0.3216	0.6599	0.3454
	DBSCAN	0.1220	0.4710	0.2280
	Research method	0.3680	0.6837	0.3903
MR	LOF	0.3896	0.3744	0.7166
	DBSCAN	0.5763	0.5423	0.8001
	Research method	0.5763	0.5423	0.8001

Table 2 shows that the adjusted RAND coefficient of the research method on Iris data set reaches 0.7091, the standardized Mutual information index reaches 0.7308, and the FM index reaches 0.8061, both higher than the LOF algorithm and DBSCAN algorithm. The adjusted RAND coefficient of the research method on the Leaf dataset reaches 0.3680, the standardized Mutual information index reaches 0.6837, and the FM index reaches 0.3903. It is higher than LOF algorithm and DBSCAN algorithm. The adjusted RAND coefficient of the research method on the MR dataset reaches 0.5763, the standardized Mutual information index reaches 0.5423, and the FM index reaches 0.8001. It is higher than the LOF algorithm and consistent with the DBSCAN algorithm. This indicates that the research method can achieve more accurate data partitioning and clustering. The precision, recall rate and F1 score of the research method are shown in Table 3.

Table 3: Precision, recall rate and F1 score test

Data set	Method	Precision	Recall rate	F1 score
Iris	LOF	0.80	0.75	0.77
	DBSCAN	0.82	0.70	0.76
	Research method	0.85	0.83	0.84
Leaf	LOF	0.60	0.55	0.57
	DBSCAN	0.65	0.50	0.57
	Research method	0.70	0.68	0.69
MR	LOF	0.84	0.80	0.82
	DBSCAN	0.78	0.75	0.76
	Research method	0.88	0.87	0.87

In Table 3, the precision, recall rate and F1 score of the research method in the Iris dataset reach 0.85, 0.83, and 0.84 respectively. In MR Data set, the precision of the research method reaches 0.88, the recall rate reaches 0.87, and the F1 score reaches 0.87. The research method demonstrates superior performance compared to other algorithms across all three data sets, including the Iris, Leaf, and MR data sets.

4.2 Application analysis of maximum entropy FC migration AD method in view of SC

When analyzing the application of data AD in research methods, due to the widespread application of data AD in financial anti fraud, the credit card transaction dataset of European cardholders in September 2013 and the credit card transaction dataset of European cardholders in October 2013 are used for application analysis. Both datasets contain sample sizes of over 280k. Firstly, it analyzed the Calinski Harabasz index of the research method, as shown in Figure 9.

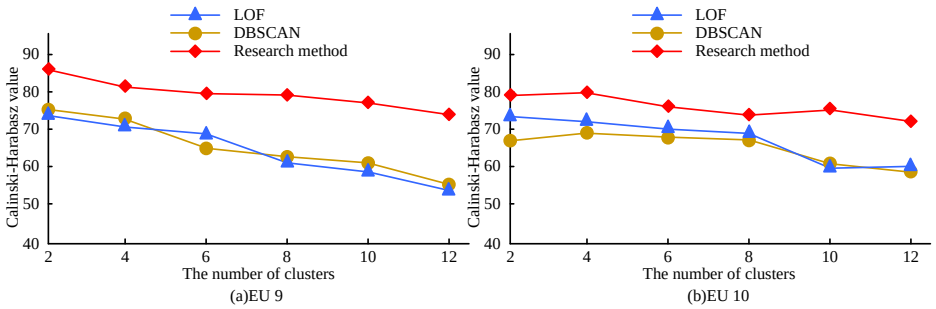

Figure 9: Calinski Harabasz index

Figure 9 shows that the Calinski Harabasz index of the three methods varies with the number of clusters in the September and October datasets. In the September dataset, the highest Calinski Harabasz index of the LOF algorithm is 73.7, while the lowest is 53.4. The highest Calinski Harabasz index of the DBSCAN algorithm is 75.2, and the lowest is 55.3. The highest Calinski Harabasz index for the research method is 86.1, and the lowest is 74.1. In the October dataset, the highest Calinski Harabasz index of the LOF algorithm is 73.6, and the lowest is 59.7. The highest Calinski Harabasz index of the DBSCAN algorithm is 68.8, and the lowest is 58.1. The highest Calinski Harabasz index for the research method is 79.9, and the lowest is 72.0. The Calinski Harabasz index of the research method is higher. It analyzes the Davies Boldin index of the research method, as shown in Figure 10.

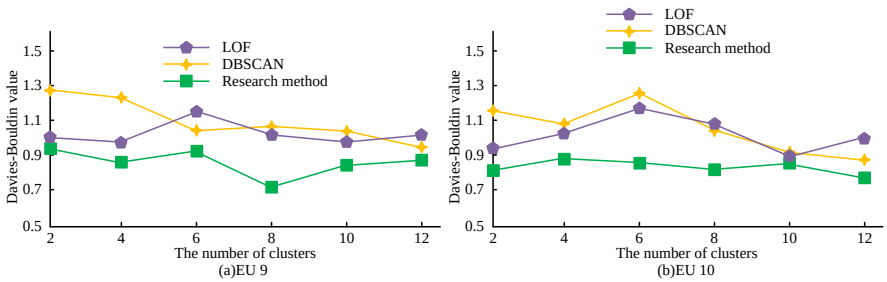

Figure 10: Davies Boldin index

Figure 10 shows that the Davies-Bouldin index of the three methods varies with the number of clusters in the September and October datasets. In the September dataset, the highest Davies Boldin index of the LOF algorithm is 1.13, and the lowest is 0.97. The highest Davies Boldin index of the DBSCAN algorithm is 1.27, and the lowest is 0.94. The highest Davies-Bouldin index of the research method is 0.93, and the lowest is 0.71. In the October dataset, the highest Davies-Bouldin index of the LOF algorithm is 1.16, and the lowest is 0.89. The highest Davies Boldin index of the DBSCAN algorithm is 1.26, and the lowest is 0.87. The highest Davies-Bouldin index of the research method is 0.88, and the lowest is 0.76. The Davies Boldin index of the research method is lower, indicating that the clustering effect of the research method is better. It generates the receiver operating characteristic curve of the research method, as shown in Figure 11.

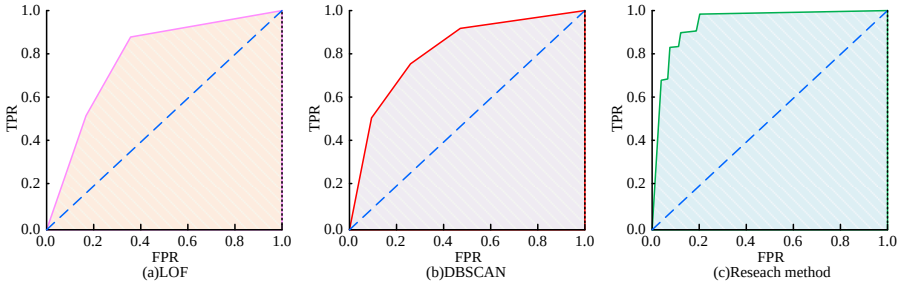

Figure 11: Receiver operating characteristic curve

Figure 11 shows that the receiver feature operation curve of the LOF algorithm increases to 0.88 when the FPR value is 0.4. The receiver characteristic operation curve of the DBSCAN algorithm increases to 0.84 when the FPR value is 0.4. The receiver characteristic operating curve of the research method increases to 0.98 when the FPR value is 0.4. The larger area enclosed by the research method curve and the abscissa indicates that the AD results obtained by the research method are better.

5 Discussion

An AD method based on the maximum entropy FC algorithm in conjunction with SC technology has been devised and its efficacy evaluated using a variety of data sets. On the Iris dataset, the research method showed a faster convergence rate in the convergence curve test, reaching the lowest value of 113 in only 9 iterations. Compared with CNN-based method, UL method, FC auxiliary method and other related technologies, the

research method had a higher accuracy. The study also exhibited an advantage in computation time. When the sample size reached 10,000, the computation time was only 2697 ms, which was lower than the 3498 ms of LOF algorithm and 3321 ms of DBSCAN algorithm. Compared with FC for medical image segmentation, neural network method, federated learning and other related technologies, the research method had more advantages in operation efficiency. The performance advantages of the research method are mainly derived from: (1) By optimizing the entropy of data distribution, the research method can more accurately estimate the potential distribution of data, thereby enhancing the precision of AD. (2) The incorporation of SC methods, particularly fuzzy logic and possibility theory, augments the capacity of algorithms to address uncertainty and fuzziness. The main contributions of the research are as follows: (1) A new AD framework is proposed by combining the maximum entropy principle with SC technology. (2) The flexibility and effectiveness of feature extraction are improved by combining KMA and HSIC.

6 Conclusion

AD can improve the reliability of information system operation. In this study, a maximum entropy FC algorithm combined with SC is proposed to detect data anomalies. Firstly, the algorithm infrastructure was constructed, and then the iteration method of the maximum entropy algorithm was designed. Then it maps the random variables from the original feature space to the kernel space and analyzes the correlation of the variables. Then it reconstructed the updating method of membership degree and center vector, calculated the comprehensive anomaly score, and finally analyzed the effectiveness of the research method. The experiment showed that the research method could decrease to the lowest value within 10 iterations in all three datasets during convergence curve testing. When conducting calculation time testing, the research method only took 2697ms when the sample size reached 10000, which was lower than other methods. When conducting the Rand coefficient test, the adjusted Rand coefficient of the research method in the Iris dataset reached 0.7091. The highest Calinski Harabasz index in the two datasets reached 86.1 for the research method. The receiver operating characteristic curve of the research method increases to 0.98 when the FPR value was 0.4. The results indicate that the research method has better clustering performance in data AD and can perform fast and accurate AD. However, the research only conducts application testing on financial type data, and subsequent application analysis will be conducted on other types of data to enrich experimental results and optimize methods.

Fundings

The research is supported by: Major Projects for National Philosophy and Social Sciences Foundation (China), Research on the Scientific Revolution and Philosophical Revolution of Soft Computing, (No. 17ZDA029).

Reference

[1] Y. Zuo, Y. Wu, G. Min, C. Huang, and K. Pe, “An intelligent anomaly detection scheme for micro-services architectures with temporal and spatial data analysis,” IEEE Transactions on Cognitive Communications and Networking, vol. 6, no. 2, pp. 548-561. https://doi.org/10.1109/TCCN.2020.2966615

[2] A. Blázquez-García, A. Conde, U. Mori, and J. A. Lozano, “A review on outlier/anomaly detection in time series data,” ACM Computing Surveys (CSUR), vol. 54, no. 3, pp. 1-33, 2021. https://doi.org/10.1145/3444690

[3] Y. M. Zhang, H. Wang, H. P. Wan, J. X. Mao, and Y. C. Xu, “Anomaly detection of structural health monitoring data using the maximum likelihood estimation-based Bayesian dynamic linear model,” Structural Health Monitoring, vol. 20, no. 6, pp. 2936-2952, 2021. https://doi.org/10.1177/1475921720977020

[4] E. Šabić, D. Keeley, B. Henderson, and S. Nannemann, “Healthcare and anomaly detection: using machine learning to predict anomalies in heart rate data,” Ai and Society, vol. 36, no. 1, pp. 149-158, 2021. https://doi.org/10.1007/s00146-020-00985-1

[5] P. M. S. Raja, “Brain tumor classification using a hybrid deep autoencoder with Bayesian fuzzy clustering-based segmentation approach,” Biocybernetics and Biomedical Engineering, vol. 40, no. 1, pp. 440-453, 2020. https://doi.org/10.1016/j.bbe.2020.01.006

[6] S. O. Adamsa, E. Azikweb, and M. A. Zubaira, “Artificial neural network analysis of some selected kdd cup 99dataset for intrusion detection,” Acta Informatica Malaysia, vol. 6, no. 2, pp. 55-61. 2022. https://doi.org/10.26480/aim.02.2022.55.61

[7] F. T. Ni, J. Zhang, and M. N. Noori, “Deep learning for data anomaly detection and data compression of a long-span suspension bridge,” Computer-Aided Civil and Infrastructure Engineering, vol. 35, no. 7, pp. 685-700, 2020. https://doi.org/10.1111/mice.12528

[8] X. Zhou, Y. Hu, W. Liang, J. Ma, and Q. Jin, “Variational LSTM enhanced anomaly detection for industrial big data,” IEEE Transactions on Industrial Informatics, vol. 17, no. 5, pp. 3469-3477, 2020. https://doi.org/10.1109/TII.2020.3022432

[9] J. Mao, H. Wang, and B. F. Spencer Jr, “Toward data anomaly detection for automated structural health

monitoring: Exploiting generative adversarial nets and autoencoders," Structural Health Monitoring, vol. 20, no. 4, pp. 1609-1626, 2021. https://doi.org/10.1177/1475921720924601

[10] Y. Liu, S. Garg, J. Nie, Y. Zhang, Z. Xiong, J. Kang, and M. S. Hossain, "Deep anomaly detection for time-series data in industrial IoT: A communication-efficient on-device federated learning approach," IEEE Internet of Things Journal, vol. 8, no. 8, pp. 6348-6358, 2020. https://doi.org/10.1109/JIOT.2020.3011726

[11] E. Khaledian, S. Pandey, P. Kundu, and A. K. Srivastava, "Real-time synchrophasor data anomaly detection and classification using isolation forest, kmeans, and loop," IEEE Transactions on Smart Grid, vol. 12, no. 3, pp. 2378-2388, 2020. https://doi.org/10.1109/TSG.2020.3046602

[12] Z. Y. Liu, F. Ding, and Y. Xu, "Background dominant colors extraction method based on color image quick fuzzy c-means clustering algorithm," Defence Technology, vol. 17, no. 5, pp. 1782-1790, 2021. https://doi.org/10.1007/s41095-021-0239-3

[13] X. Zhang, H. Wang, Y. Zhang, X. Gap, G. Wang, and C. Zhang, "Improved fuzzy clustering for image segmentation based on a low-rank prior," Computational Visual Media, vol. 7, no. 4, pp. 513-528, 2021. https://doi.org/10.1007/s41095-021-0239-3

[14] F. Jiang, N. Li, and L. Zhou, "Grain segmentation of sandstone images based on convolutional neural networks and weighted fuzzy clustering," IET Image Processing, vol. 14, no. 14, pp. 3499-3507, 2020. https://doi.org/10.1109/ICPR.2018.8545649

[15] L. Hu, X. Pan, Z. Tang, and X. Luo, "A fast fuzzy clustering algorithm for complex networks via a generalized momentum method," IEEE Transactions on Fuzzy Systems, vol. 30, no. 9, pp. 3473-3485, 2021. https://doi.org/10.1109/TFUZZ.2021.3117442

[16] Q. Feng, L. Chen, C. L. P. Chen, and L. Guo, "Deep fuzzy clustering-a representation learning approach," IEEE Transactions on Fuzzy Systems, vol. 28 , no. 7, pp. 1420-1433, 2020. https://doi.org/10.1109/TFUZZ.2020.2966173

[17] S. Javadi, S. M. Hashemy Shahdany, A. Neshat, and A. Chambel, "Multi-parameter risk mapping of Qazvin aquifer by classic and fuzzy clustering techniques," Geocarto International, vol. 37, no. 4, pp. 1160-1182, 2022. https://doi.org/10.1080/10106049.2020.1778099

[18] Y. Guo, Z. Mustafaoglu, and D. Koundal, "Spam detection using bidirectional transformers and machine learning classifier algorithms," Journal of Computational and Cognitive Engineering, vol. 2, no. 1, pp. 5-9, 2022. https://doi.org/10.47852/bonviewJCCE2202192

[19] G. Li, G. Kou, and Y. Peng, "Heterogeneous large-scale group decision making using fuzzy cluster analysis and its application to emergency response plan selection," IEEE Transactions on Systems, Man, and Cybernetics: Systems, vol. 52, no. 6, pp. 3391-3403, 2021. https://doi.org/10.1109/TSMC.2021.3068759

[20] G. Muhiuddin, A. Mahboob, and M. E. A. Elnair, "A new study based on fuzzy bi-Γ-ideals in ordered-Γ-semigroups," Journal of Computational and Cognitive Engineering, vol. 1, no. 1, pp. 42-46, 2022. https://doi.org/10.47852/bonviewJCCE19919205514

[21] O. P. Mahela, B. Khan, H. H. Alhelou, and P. Siano, "Power quality assessment and event detection in distribution network with wind energy penetration using stockwell transform and fuzzy clustering," IEEE Transactions on Industrial Informatics, vol. 16, no. 11, pp. 6922-6932, 2020. https://doi.org/10.1109/TII.2020.2971709