Clustering Analysis of Unlabeled Data and Weak-Label Detection Analysis Method Integrating Soft Computing Technology

Chunhua Liang\textsuperscript{1,2}\textsuperscript{*}
\textsuperscript{1}Institute for History of Science and Technology, Shanxi University, Taiyuan, 030006, China
\textsuperscript{2}Information Institute, Shanxi Finance and Taxation College, Taiyuan, 030024, China

Corresponding author: Chunhua Liang (e-mail: lchua2023@126.com).

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)

ABSTRACT With the continuous improvement of digitization, the processing and analysis of massive data has become one of the hot issues. Soft computing technology, as an emerging machine intelligence technology, performs well in handling complex uncertainty problems and is an important component of artificial intelligence. This study takes soft computing technology as the technical core and constructs a fuzzy dynamic clustering model based on improved immune algorithms to process unlabeled data. And an anomaly detection and analysis algorithm is designed based on soft instance transfer learning to handle weakly labeled data. The performance test outcomes denote that the accuracy, recall, and F1 values of the immune optimization fuzzy dynamic clustering algorithm are 91.69\%, 89.27\%, and 92.15\%, respectively, reaching the optimal level of similar intelligent optimization clustering algorithms. The immune optimization fuzzy dynamic clustering algorithm has better computational efficiency, loss function curve performance, and strong global search ability, and avoids the occurrence of local optimal solutions. Compared with other advanced clustering algorithms, the immune optimization fuzzy dynamic clustering algorithm performs well on datasets in various fields, and both external and internal evaluation indicators verify the algorithm's clustering effect. The AUC value of the soft computing instance transfer learning anomaly detection algorithm is 0.913, with a detection accuracy of 91.67\%, which is superior to other anomaly detection algorithms. The unlabeled and weak-label data processing model designed based on soft computing technology can effectively achieve the processing and analysis of real-world data problems.

INDEX TERMS Soft computing, Labeled data; Cluster analysis, Weak-label, Unlabeled

I. INTRODUCTION

Data mining is the discovering of valuable information and patterns in data through the use of techniques such as statistics and machine learning, and involves the extraction of knowledge from large amounts of data for use in prediction and decision support problems in a variety of fields [1-2]. The Internet is filled with various types of data such as numbers, text, images, etc., showing the basic characteristics of diversity, variability, distribution and association, and these characteristics are the key to understanding the characteristics and patterns of data [3]. However, the data processing often faces the quality problem of lack of labels, and the lack of pre-labeled or fuzzy and incomplete labeling information affects the completeness and usability of the data, which is not conducive to the development of stable models [4]. At present, the common method to deal with unlabeled data is to use clustering algorithms to classify and label them, and transform "unlabeled data" into "labeled data"; or use semi-supervised clustering to complete the samples with missing labels. However, the clustering algorithm has lower accuracy in processing unlabeled data and higher sensitivity to unbalanced samples [5-6]. Soft computing (SC) consists of rough sets, fuzzy logic theory, artificial neural networks, evolutionary immune algorithms, genetic algorithms (GAs), simulated annealing (SA), and chaos theory [7]. SC tolerates the uncertainty and partial authenticity of data by simulating human decision-making and reasoning, thus obtaining the performance of close to
the reality, easy to deal with and better robustness, suitable for fuzzy and noisy data. SC can solve uncertainty, imprecision, and approximate solution problems, which are widely used in various fields such as process control, predictive decision-making, fault diagnosis, and data mining, etc. [8].

In response to the shortcomings of existing hard computing processing techniques for unlabeled and weakly labeled data, an improved immune algorithm optimized fuzzy dynamic clustering model (IA-FDC) and an anomaly detection and analysis algorithm based on soft instance transfer learning (SITL) were studied and designed. The study consists of four parts. The first part is a review of the current status of domestic and international research on SC technology and unlabeled data processing. The second part proposes an IA-FDC algorithm and a SITL algorithm. The third part of the algorithm performance is tested and verified. And the fourth part of the study summarizes the results of the research experiments, which promise to describe the uncertainty of the sample more objectively and accurately.

There are two innovative points in the research. Firstly, a new solution for soft clustering models was designed by breaking through the traditional thinking of hard computing techniques in handling unlabeled data. This method utilizes the global search ability of immune algorithms and the fuzziness of fuzzy dynamic clustering to handle the uncertainty of unlabeled data, and has good data classification ability and adaptability to fuzzy scenes; Secondly, the introduction of transfer learning utilizes existing knowledge to improve the ability of soft computing to detect anomalies in weakly labeled data. The main contributions of the research are twofold. On the one hand, it enriches the theory and practice of soft computing and unlabeled data, providing strong guidance for the processing of unlabeled and weakly labeled data; On the other hand, research has promoted the progress and development of soft computing technology, expanding the application fields of soft computing technology.

II. RELATED WORKS
To explore the knowledge hidden in data and develop more effective data support applications, a great deal of research on the processing of unlabeled data samples have been organized by scholars. The application of machine learning and deep learning in mobile sensing largely depends on the availability of the dataset. Based on this, Tang C. I. et al. designed a self-supervised human behavior recognition model trained on unlabeled data. The model effectively learned unlabeled mobile sensing datasets to supplement small labeled datasets. The test results for different human behavior datasets denoted that the model exhibited more advanced performance compared to existing semi-supervised models, with an F1 score improvement of 12% in the presence of a large amount of unlabeled data [9]. At present, the annotation of plant disease images mostly relies on manual labeling by agricultural experts, which consumes a lot of time and labor. Moreover, the self-supervised learning annotation of existing datasets is more suitable for balanced datasets and relies on prior knowledge. Based on this, Uno F. et al. designed a self-supervised clustering framework for plant disease images with the Kernel K-means vulnerability. This cross iterative clustering algorithm has been extensively experimented on different plant disease datasets, including different plants and plant diseases. The research outcomes indicated that the framework had significant advantages on imbalanced datasets [10].

To utilize pairwise similarity and non similarity between data points, Shimada T. et al. designed an empirical risk minimization method that calculated unbiased estimates of classification risk from pairwise similarity and unlabeled data. And the semi-supervised clustering method was integrated with similarity and non similarity points into the clustering framework, and the experimental results verified the practicality of this method [11]. Hierarchical clustering is a commonly used unsupervised method for processing unlabeled data, but a single clustering method is difficult to deal with complex problems. Li T. et al. designed a clustering hierarchical clustering framework based on ensemble methods, which belongs to a meta clustering integration scheme with model selection. The framework created meta clusters by re-clustering the main cluster. It would merge similar clusters and consider thresholds to determine the optimal number of clusters. The research findings showed that this method was superior to existing unsupervised clustering techniques [13]. Cancer variant gene data recorded in public databases such as COSMIC lacks functional and clinical annotations, and unlabeled data is generally redundant with labeled data. In this regard, Z. Ren et al. designed a semi-supervised generative adversarial network by combining the features of the clinical guidelines, and incorporated the features of the labeled and unlabeled data into the construction of the network model. The experimental results demonstrated that the research-designed model had a unique advantages and competitive performance [14].

Existing battery health...
management mostly relies on labeled data, but unlabeled data is still of great importance. C. Lin designed a semi-supervised learning method for processing unlabeled data, using two regression factors to learn the mapping between health indicators and battery status. Based on the semi-supervised co-training, pseudo-labeled prediction was performed on unlabeled data to increase the training samples. Experimental results showed that the method could significantly improve the estimation accuracy of battery health status, broadening the prospect of using unlabeled industrial data in combination with labeled laboratory data to estimate battery health status [15].

Data-driven fault detection and diagnosis often face limited labeled samples, and the distribution of data classes is highly unbalanced. X. Peng et al. integrated active learning and semi-supervised learning methods, and designed a robust and cost-effective fault detection and diagnosis framework based on the adaptive graph construction of graphs, which can be used to predict labels of unbalanced data and detect new categories. Three synthetic data sets and one real data set validated the superior performance of the framework [16].

In recent years, major experts and scholars have conducted extensive research on SC technology, using it to address issues that traditional algorithms are difficult to solve. The prediction of shallow settlement in geotechnical engineering is a complex engineering problem. Ray R. et al. applied three SC techniques: minimax probability machine regression, PSO artificial neural network, and adaptive network fuzzy inference system to study the reliability of shallow foundations. The effectiveness of the model was assessed using multiple indicators, and the experimental findings denoted that minimax probability machine regression was superior to the other two SC techniques, which can be used for nonlinear problem analysis of shallow foundation settlement on soil [17]. Zhang W. G. et al. established a tunnel surface settlement prediction model based on SC. The effectiveness of the model was assessed using multiple indicators, and the results verified the reliability of the model [18]. The early detection of skin cancer can help in its treatment and diagnosis. Xu Z. et al. designed an automatic diagnosis method for skin cancer using SC technology. They optimized the convolutional neural network using a satin bowerbird to construct a skin cancer diagnosis image segmentation method, and verified its effectiveness through a confusion matrix. Finally, simulations were conducted in the American Cancer Society database, and the experimental results expressed the accuracy and sensitivity of the algorithm [19].

Nguyen P. T. et al. designed four integrated SC models with logistic regression combined with bagging, random subspace, and cascade generalization integration techniques for groundwater potential mapping. The model was tested using indicators such as area under the receiver's working characteristic curve, positive and negative predictive values, root mean square error, accuracy, sensitivity, and specificity. The results indicated that all four ensemble learning techniques have successfully improved the performance of the basic logistic regression model, enabling the growth of effective adaptive groundwater management plans [20]. For blasting engineering, Zhang X. et al. designed a prediction model for peak particle velocity and slope stability using PSO and limit gradient enhancement machine based on SC. The effectiveness of the model was assessed using indicators such as average absolute error, coefficient of determination, variance ratio, and root mean square error, and the results verified the reliability of the model [21]. A. D. Skentou et al., based on artificial neural network, predicted the unconfined compressive strength of granite by using three nondestructive testing indexes: pulse velocity, Schmidt hammer rebound number and effective porosity. Experimental results showed that the performance of the Levenberg-Marquardt artificial neural network designed by the study was superior to the existing prediction model [22]. To solve the design problem of multiple heat recovery technology, M. A. Haghghi et al. proposed A new polygeneration model considering parallel and series waste heat recovery, and carried out multi-criteria optimization under four different scenarios by using artificial neural network and multi-objective gray wolf optimization method. Experimental results verified the practicability of this model [23].

In summary, the processing and analysis of unlabeled data, as well as the comprehensive application of SC technology, have been widely used. However, the processing technology of unlabeled data still has the disadvantage of insufficient processing accuracy. A large number of research results have shown that SC is suitable for processing fuzzy uncertainty problems. Therefore, studying the application of SC technology to the processing of unlabeled and weak-label data has certain research value.

**III. ANALYSIS AND DETECTION ALGORITHM DESIGN FOR UNLABELED AND WEAK-LABEL DATA BASED ON SC**

With the continuous improvement of informatization and digitization, big data has become an important electronic resource. Data plays a critical role in the development of artificial intelligence, and the processing of massive data has become a key challenge faced by different fields. Extracting valuable information from various labeled,
unlabeled, and weak-label helps users understand deeper, hidden trends in the data. In real tasks, obtaining labeled data is relatively difficult. However, obtaining a great deal of unlabeled or weak-label data is relatively easy, and giving machines the ability to learn knowledge and patterns from massive unlabeled data is crucial. Two SC models have been established: an IA-FDC algorithm and a SITL algorithm to handle unlabeled and weak-label data.

A. DESIGN OF UNLABELED DATA ANALYSIS MODEL BASED ON FDC

SC is a new type of machine intelligence technology. The emergence of SC has led artificial intelligence towards the development direction of fuzziness, randomness, imprecision, and incompleteness. Artificial intelligence has gradually evolved into computational intelligence. SC is a type of algorithm inspired by human or biological intelligence for problem solving, including neural networks, machine learning, GAs, fuzzy cumulative computing, ant colony algorithms (ACA), immune algorithms, evolutionary algorithms, heuristic algorithms, SA algorithms, hybrid intelligent algorithms, and many other types of algorithms.

The application of SC in big data includes clustering, association rules, rule extraction, and data synthesis. Cluster analysis is suitable for modeling and analyzing data without reliable labels, transforming the problem into a classification or clustering problem with or without labels on the dataset. The clustering algorithm includes different types of clustering methods, including soft and hard clustering. The schematic diagram of the clustering algorithm is denoted in Figure 1. Hard clustering requires that each data point can only belong to a specific cluster, while soft clustering allows data points to belong to multiple clusters simultaneously. Fuzzy clustering belongs to a type of soft clustering, FDC is a method based on fuzzy theory and dynamic clustering for finding fuzzy cluster structures in a data set. Compared to traditional hard clustering algorithms, FDC takes into account the fact that data points may belong to multiple clusters, further increasing the flexibility of clustering. In FDC, each data point can be viewed as a fuzzy set indicating its affiliation to the cluster to which it belongs. The affiliation of each data point to each cluster is calculated based on the distance of the data point from the cluster center. As clustering proceeds, the cluster centers are continuously adjusted until a preset number of iterations is satisfied. The core idea of FDC is to adjust the clustering results by continuously updating the cluster centers and affiliations to make the results more accurate. Compared to the traditional hard clustering analysis, FDC can better adapt to changes in the data and is able to handle the growth, decrease and drift of data samples. The advantage of FDC is that it can handle the ambiguity of the data while taking into account the extent to which the data points contribute to all the clusters. And research is conducted on constructing unlabeled data analysis models based on FDC [24-26].

![FIGURE 1. Schematic diagram of the clustering algorithm process](image)

The fuzzy equivalence relationship $\bar{R}$ means a fuzzy set on the product of non empty set $X$ and Cartesian product $X \times X$, and the fuzzy cut set $\bar{R}_\alpha$ of any dataset $\alpha$ belongs to $X \times X$, which can be considered as the equivalence relationship of $X$. Furthermore, the classification of non empty set $X$ can be derived, and the classification number can be classified to form a dynamic clustering spectrum.

The FDC algorithm should first establish fuzzy similarity relationships, and study uses the angle cosine method to calculate similarity relationships, as shown in equation (1). In equation (1), $r_{ij}$ denotes the similarity relationship between the classification targets $x_i$ and $x_j$; $x_i, x_j \in R^r$, and $R'$ indicates the fuzzy similarity relationship, and $x_k$ expresses the $k$th dimensional feature.

$$R'_k = \left\{ \begin{array}{ll} r_{ij} & 0 \leq r_{ij} \leq 1, i, j = 0, 1, \ldots, n \\ r_{ij}' = \frac{\sum_{k=1}^{n} x_{ik} x_{jk}}{\sqrt{\sum_{k=1}^{n} x_{ik}^2} \sqrt{\sum_{k=1}^{n} x_{jk}^2}} & \end{array} \right.$$  

(1)

It will standardize the calculated similarity, with $N$ classification datasets corresponding to feature parameter $x_k$, and each feature parameter $x_k$ containing $n$ elements $x_{ik}$. The first step of standardization processing is to calculate the mean and variance of $x_k$, as shown in equation (2).

$$\begin{align*}
\bar{x}_k &= \frac{1}{n} \sum_{i=1}^{n} x_{ik} \\
S_k^2 &= \frac{1}{n} \sum_{i=1}^{n} (x_{ik} - \bar{x}_k)^2 
\end{align*}$$

(2)

Finally, the standardized values for the dataset are calculated, as shown in equation (3). In equation (3), $x_{ik}'$ indicates the standardized value.

$$x_{ik}' = \frac{x_{ik} - \bar{x}_k}{S_k}$$

(3)

FDC classifies things without clear boundaries, but different datasets correspond to different dimensions. It is also necessary to standardize the data of different dimensions, transform and compress them within the interval of $[0,1]$, and use the translation range...
transformation method to convert them into a standardized fuzzy matrix. The operation process is shown in equation (4).

\[
x_{ij} = \frac{x_{ij} - \min \{x_{ij} \mid 1 \leq i \leq n \}}{\max \{x_{ij} \mid 1 \leq i \leq n \} - \min \{x_{ij} \mid 1 \leq i \leq n \}}
\]

(4)

After obtaining the fuzzy similarity relationship, it will establish the fuzzy equivalence relationship between the classification dataset. The conversion calculation between the fuzzy similarity relationship and the fuzzy equivalence relationship is shown in equation (5), and \( R_s \) denotes the fuzzy equivalence relationship of the dataset. In summary, different fuzzy cut sets output different thresholds, representing different clustering results, forming a clustering lineage graph.

\[
R_s = R_s \cap R_s = R_s \cap R_s
\]

(5)

**B. DESIGN OF IA-FDC ALGORITHM OPTIMIZATION**

FDC method belongs to unsupervised learning method, which solves the problem of local optimal solution rather than global optimal solution. The threshold value determines the quality of classification results. Research introduces immune algorithms to improve the global search ability of FDC analysis.

Immune algorithm is an information processing technology inspired by the principles of the biological immune system. The immune system has multiple mechanisms to fully utilize limited resources to respond to the invasion of pathogens. The immune mechanism is complex and variable, but mainly composed of innate and adaptive immune systems, completing immune defense and monitoring to achieve immune stability. Immune algorithms combine the mechanisms of the immune system and evolutionary algorithms. The immune system generates a large number of different antibodies through cloning, which continuously recognize and memorize antigens. Immune algorithms exhibit strong adaptability, diversity, dynamism, robustness, and global convergence [27-28].

The immune algorithm mainly includes three main contents: antigen recognition and initial antibody generation, antibody evaluation, and immune operation. The diversity of the immune algorithm and population mainly depends on the introduction of antibody concentration and information entropy. The schematic diagram of allele information entropy is indicated in Figure 2.

\[
E_j(N) = -\sum_{j=1}^{M} p_j \log p_j
\]

(6)

The calculation process of information entropy is shown in equation (6). In equation (6), \( E_j(N) \) means information entropy. \( N \) denotes antibody. \( j \) expresses gene locus. \( p_j \) indicates \( j \) as the frequency of \( k_i \). \( k_i \) represents gene locus. It contains \( S \) gene loci, among which \( x_{i,j} = \begin{cases} 1 & \text{if } \text{antigen } i \text{ binds to } \text{antibody } j \text{ at locus } l \\ 0 & \text{otherwise} \end{cases} \)

\[
E_j(N) = -\sum_{j=1}^{M} p_j \log p_j
\]

(6)

The flowchart of the entire immune algorithm is shown in Figure 3. The implementation of immune algorithms first involves antigen recognition, which involves dealing with the problem to be optimized. And an affinity function is established to complete the setting of constraint conditions. Then, the feasible solutions of the problem to be optimized are binary coded as antibodies, and an initial antibody population is randomly generated. The binary programming method can ensure the stability of the algorithm and the diversity of the population [29].

\[
ax_{vw} = \frac{1}{1 + opt_{vw}}
\]

(7)
The calculation of antibody concentration is shown in equation (8). In equation (8), \( C \) denotes the concentration of antibodies, i.e., the proportion of antibodies in similar antibodies. \( A \) refers to affinity.

\[
C = \frac{1}{N} \sum_{x \in N} A
\]  

(8)

The affinity and concentration values of each antibody are calculated according to equations (7) and (8), and high-quality antibodies are selected based on the affinity between antibodies and individual concentration. Inferior antibodies will be eliminated. It will determine whether the algorithm termination criteria have been met, select high-quality individuals, and perform evolutionary operations to calculate antibody concentration and antibody motivation.

The calculation of antibody concentration is shown in equation (9). The antibody concentration is related to the diversity of the antibody population, and excessive concentration can lead to the algorithm falling into a local optimal solution. In equation (9), \( ab \) denotes the \( i \)th antibody in the population. \( N \) indicates the population size. \( \text{aff}(ab,ab) \) expresses the affinity between antibodies \( i,j \). \( \delta \) means the antibody similarity threshold.

\[
\begin{align*}
\text{den}(ab) & = \frac{1}{N} \sum_{i<j} S(ab,ab) \\
S(ab,ab) & = \begin{cases} 1, & \text{aff}(ab,ab) < \delta, \\ 0, & \text{aff}(ab,ab) \geq \delta, \\ \end{cases}
\end{align*}
\]

(9)

The calculation of antibody stimulation level is shown in equation (10), where \( \text{sim}(ab) \) refers to the stimulation level. \( a,b \) represent the calculation parameter.

\[
\text{sim}(ab) = a \cdot \text{aff}(ab) + b \cdot \text{aff}(ab)
\]

(10)

After completing the calculation of antibody concentration and motivation, immune operations, including immune selection, cloning, mutation, and clone inhibition, are carried out to search for optimization problems. Immune selection is the selection of cloned antibodies based on the calculated value of motivation, and the calculation of cloning is shown in equation (11). In equation (11), \( ab \) refers to the set of clonal antibody structures with the same number of antibody clones as \( ab \). Finally, the population will be refreshed, and the antibodies with lower motivation will be replaced, and the generated new antibodies will repeat the immune algorithm process again.

\[
T(ab) = \text{clone}(ab)
\]

(11)

The immune algorithm optimizes and calculates the FDC algorithm, and the resulting immune optimized FDC algorithm flowchart is shown in Figure 4. Firstly, the division of the initial clustered dataset is obtained according to the initial threshold, fuzzy dynamic clustering assigns the data points to multiple fuzzy clusters and calculates the degree of affiliation of each data point to each cluster. The degree of affiliation is the degree of match between data points and clusters, and FDC divides abnormal samples or samples different from normal patterns into clusters with lower degrees of affiliation. And immunoidentification mainly identifies and discriminates whether the sample to be processed is antigenic or not, i.e., it identifies abnormal samples with lower affiliation degree. Then the immunization algorithm optimal threshold is run. The result of dataset division is determined according to the optimal threshold.

**C. DESIGN OF ANOMALY DETECTION ALGORITHM BASED ON SITL**

Anomaly detection is a technology to identify abnormal conditions and mine non-logical data, which is an important branch of machine learning tasks. In computer vision, data mining and natural language processing tasks, detecting abnormal conditions in data is the key to task processing. Abnormal points in the data set in the clustering process will seriously affect the clustering effect, and there are often abnormal data in common data sets or real-world problems that lack effective anomaly labels. However, traditional anomaly detection algorithms, including Gaussian fitting, semi-supervised learning and deep learning, are hard computing technologies, which are suitable for accurate data, while are not suitable for dealing with common problems in real life. However, SC can flexibly deal with complex, fuzzy and incomplete problems, and can learn and reason under incomplete labeling or feature information, which has certain advantages in anomaly detection. Transfer learning refers to the use of existing knowledge to improve the performance of a target task or domain in different tasks or domains. In anomaly detection, transfer learning can make use of existing anomaly detection models or features to process new data sets. Therefore, the idea of transfer learning and SC technology are introduced in this study to build an anomaly detection algorithm based on SITL for anomaly detection and analysis of weak-label data [30].

Transfer learning is a machine learning method that learns new knowledge by applying existing knowledge. The core task of transfer learning is to find the similarity between
existing and new knowledge, and achieve the purpose of transfer learning through the transfer of similarity. The transfer learning structure is shown in Figure 5. Transfer learning includes source domain $D_s$, target domain $D_t$, source task $T_s$, and target task $T_t$. The design concept of transfer learning is to improve the new learning task $T_t$ through $D_s$ and $T_s$ learning and transferring knowledge. The scale of the source domain is usually larger than the target domain [31-32].

The scale of the source domain is usually larger than the target domain [31-32].

![FIGURE 5. Structure diagram of transfer learning](image)

In anomaly detection problems, there is similarity between different anomaly detection tasks, and the commonality of similar tasks meets the learning concept of transfer learning, reducing the time and cost of collecting data labels for different tasks. The SITL maximizes the utilization of the label information of the source domain $D_s$ in the absence of labels in the target domain $D_t$, resulting in the most usable transfer learning results.

Firstly, it needs to mark pseudo labels on the target domain $D_t$ and train the instance migration model with homologous data simultaneously. During the migration, insufficient utilization of labels in the source domain data can sometimes lead to a decrease in migration performance on the target domain dataset. To avoid this issue, the study introduces the integration method to train multiple base models simultaneously, and judges the degree of migration of source domain instances based on the training results of the base models.

The training base model sets the update weights for positive and negative instances in the source domain, and the calculation of correctly classified instance weights is shown in equation (12).

$$w = w_e \exp \left( \beta \left| y_i - \hat{y}_i \right| \right)$$  \hspace{1cm} (12)

The calculation of instance weights for mis-classification is shown in equation (13). In equation (13), $y_i$ denotes the true label. $\hat{y}_i$ indicates the prediction label. $\exp$ serves as an exponential function with $e$ as the base. Among them,

$$\beta = \log \left( \frac{1 - \text{recall}}{\text{recall}} \right).$$

The calculations of equations (12) and (13) are repeated until a satisfactory transfer weight vector is obtained through updating.

$$w = w_e \exp \left( \beta \left| 1 - \left| y_i - \hat{y}_i \right| \right| \right)$$  \hspace{1cm} (13)

After obtaining the migration weight of the source domain instance, the migrated source domain data is combined with the unlabeled data of the target domain to calculate the anomaly scores of all data. The calculation is denoted in equation (14). In equation (14), $a_i$ and $a_e$ mean the anomaly scores of the source and target domain data. $w_i$ indicates the migration weight of $j$ point. $d_i^j$ expresses the distance between points in the target domain and adjacent points of $k_i$. $\gamma$ represents the proportion of anomaly points in the source domain. $k$ and $k_e$ denote neighborhood parameters.

$$a_i = \sum w_i^j \frac{d_i^j}{\gamma}$$

$$a_e = 1 - \exp \left( - \sum \frac{d_e^j}{\gamma} \right)$$  \hspace{1cm} (14)

Finally, the anomaly score calculation for each object point is shown in equation (15), where $w_i$ expresses the contribution of source domain data to target domain objects.

$$a_l = a_i \cdot w_l + a_e \cdot (1 - w_l)$$  \hspace{1cm} (15)

At this point, the entire SITL has been constructed, and the flowchart of SITL is shown in Figure 6. The ideas of transfer and ensemble learning constantly update transfer weights.

![FIGURE 6. Flow chart of anomaly detection algorithm based on soft instance migration](image)
IV. PERFORMANCE TESTING OF UNLABELED AND WEAK-LABEL DATA PROCESSING MODELS BASED ON SC TECHNOLOGY

To evidence the effectiveness of the SC model designed in the study, different performance testing and clustering effect analysis experiments were designed. The performance and clustering effect of the IA-FDC algorithm were analyzed, as well as the performance of the anomaly detection algorithm for SC instance transfer learning.

A. PERFORMANCE TESTING OF IA-FDC ALGORITHM OPTIMIZATION

To verify the effectiveness of IA-FDC algorithms, several other common intelligent optimization algorithms were utilized to globally optimize the FDC algorithm. The accuracy, recall rate, and F1 value evaluation indicators of different optimization algorithms were compared. The experiment outcomes are indicated in Figure 7. The comparative algorithms included GA, Tabu Search (TS), SA, AC algorithm, and PSO. As shown in Figure 7, the three indicators optimized by immune algorithm were at the highest level, with accuracy, recall, and F1 values reaching 91.69%, 89.27%, respectively 92.15%. The evaluation index values of the other five algorithms were mostly in the range of below 80%, significantly lower than those of immune algorithms. Accuracy represents the true positive probability among all statistically positive samples, recall represents the probability of being predicted as positive in the actual positive samples, and positive samples represent the detection of anomalies. Usually, there is a certain restrictive relationship between accuracy and recall, which is a pair of contradictory indicators. In the classification model, there may be situations where the accuracy is improved while the recall is reduced. However, the IA-FDC algorithm shows a good balance between the two in testing, while other algorithms experience an increase in recall but a decrease in accuracy. The F1 value is the harmonic value of accuracy and recall. The F1 value of the IA-FDC algorithm was 92.15%, while the lowest F1 value of the GA was 59.97%, with a difference of 32.18 percentage points between the two.

Comparing the training time results of different intelligent optimization algorithms for optimizing clustering models, the research findings are expressed in Figure 8. As shown in Figure 8, when processing 6000 data samples, the immune optimization algorithm took the least time, and the final time was only 8.45 seconds. The time consumption statistical curve was significantly lower than other optimization algorithms, and the model's work efficiency was better when searching globally. This ensured the optimal performance indicators while reducing the time cost of algorithm operation, resulting in better overall clustering performance.

Comparing the loss function curves of different intelligent optimization algorithms, the experimental statistical findings are expressed in Figure 9. From Figure 9, the loss function curve of the immune algorithm converged to the minimum value, and the convergence curve steadily decreased with less oscillation. The final convergence value of other optimization algorithms was greater than...
that of immune algorithms, and the number of iterations during convergence was smaller. But the earlier the convergence, the greater the possibility of falling into the local optimal value, indicating that the immune algorithm had a stronger global search ability. The loss function represents the degree of inconsistency between the predicted and true values of the model. The smaller the loss function value, the better the model fitting. Therefore, the IA-FDC algorithm performed better.

![Figure 9](image)

**Figure 9.** Loss function curves of different intelligent optimization algorithms

## B. ANALYSIS OF CLUSTERING EFFECT OF IA-FDC ALGORITHM OPTIMIZATION

To analyze the clustering performance of improved IA-FDC, traditional FDC algorithm, fuzzy c-means algorithm (FCMA), density-based clustering method (DBSCAN), and maximum entropy clustering (MEC) were selected for performance comparison. Considering the dimension, magnitude, complexity and design field of the data set, open low dimensional data sets in many fields such as biology, medicine and agriculture were used, including Iris, Wine, Glass Identification, Breast Cancer Wisconsin, and Mushroom datasets.

The evaluation of fuzzy models is relatively complex, so multiple evaluation indicators were selected for comprehensive evaluation. Firstly, Purity and normalized mutual information (NMI) were selected to analyze the effectiveness of the fuzzy algorithm, with all indicator values normalized within the [0,1] range. The research findings are indicated in Table I. The Purity and NMI indicators of IA-FDC algorithm performed well, with values reaching over 0.8 on different datasets. The values of other clustering algorithms’ index sets ranged from 0.5 to 0.7. The Purity index was similar to the accuracy evaluation index in supervised learning, similar to NMI. The closer the value was to 1, the better the clustering effect.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Iris Purity</th>
<th>Wine Purity</th>
<th>Glass Identification Purity</th>
<th>Breast Cancer Wisconsin Purity</th>
<th>Mushroom Purity</th>
</tr>
</thead>
<tbody>
<tr>
<td>IA-FDC</td>
<td>0.85</td>
<td>0.83</td>
<td>0.890</td>
<td>0.869</td>
<td>0.931</td>
</tr>
<tr>
<td>NMI</td>
<td>0.86</td>
<td>0.91</td>
<td>0.875</td>
<td>0.784</td>
<td>0.806</td>
</tr>
<tr>
<td>FDC</td>
<td>0.74</td>
<td>0.75</td>
<td>0.737</td>
<td>0.756</td>
<td>0.714</td>
</tr>
<tr>
<td>FCMA</td>
<td>0.76</td>
<td>0.76</td>
<td>0.780</td>
<td>0.747</td>
<td>0.632</td>
</tr>
<tr>
<td>DBSCAN</td>
<td>0.80</td>
<td>0.69</td>
<td>0.701</td>
<td>0.614</td>
<td>0.752</td>
</tr>
<tr>
<td>NMI</td>
<td>0.76</td>
<td>0.68</td>
<td>0.623</td>
<td>0.726</td>
<td>0.694</td>
</tr>
</tbody>
</table>

Internal evaluation indicators of the clustering algorithm were added to analyze its performance, including Mean Squared Error (MSE), Xie Beni (XB), Davies-Bouldin Index (DB), and Rand Index (RI). The statistical results of MSE and RI are shown in Table II. The RI value of IA-FDC was relatively large, reaching 0.965 on the Mushroom dataset, an increase of 0.357 compared to the same type dataset. The MSE index value of IA-FDC was relatively small, with a minimum MSE of only 0.136. The MSE indicator represents the square of the average error between the predicted value and the actual value, while the RI reflects the percentage of consistent results in the clustering results, representing the clustering effect and the degree of fitting with the real situation. The comprehensive clustering effect of the IA-FDC model designed in the study was better.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Iris MSE</th>
<th>Wine MSE</th>
<th>Glass Identification MSE</th>
<th>Breast Cancer Wisconsin MSE</th>
<th>Mushroom MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>IA-FDC</td>
<td>0.21</td>
<td>0.13</td>
<td>0.162</td>
<td>0.214</td>
<td>0.136</td>
</tr>
<tr>
<td>RI</td>
<td>0.84</td>
<td>0.91</td>
<td>0.884</td>
<td>0.875</td>
<td>0.965</td>
</tr>
<tr>
<td>FDC</td>
<td>0.30</td>
<td>0.20</td>
<td>0.333</td>
<td>0.274</td>
<td>0.220</td>
</tr>
<tr>
<td>RI</td>
<td>0.74</td>
<td>0.73</td>
<td>0.697</td>
<td>0.741</td>
<td>0.766</td>
</tr>
<tr>
<td>FCMA</td>
<td>0.31</td>
<td>0.30</td>
<td>0.248</td>
<td>0.296</td>
<td>0.278</td>
</tr>
<tr>
<td>RI</td>
<td>0.70</td>
<td>0.71</td>
<td>0.695</td>
<td>0.648</td>
<td>0.675</td>
</tr>
</tbody>
</table>
The statistical results of XB and DB indicators are shown in Figure 10. From Figure 10, the XB and DB indicators of all clustering algorithms were decreasing, but the IA-FDC algorithm had the lowest decrease in indicator values. XB represents a certain balance point between intra class compactness and inter class separation, and the smaller the indicator value, the better the clustering efficiency; The DB index represents the sum of the average distance within a class divided by the distance between two cluster centers. The smaller the value, the smaller the intra-class distance and the larger the inter-class distance. The internal indicators of the IA-FDC algorithm had better evaluation performance and good clustering performance.

<table>
<thead>
<tr>
<th></th>
<th>MSE</th>
<th>RI</th>
<th>MSE</th>
<th>RI</th>
<th>MSE</th>
<th>RI</th>
</tr>
</thead>
<tbody>
<tr>
<td>DBSCAN</td>
<td>0.36</td>
<td>0.62</td>
<td>0.31</td>
<td>0.59</td>
<td>0.35</td>
<td>0.59</td>
</tr>
<tr>
<td>MEC</td>
<td>0.41</td>
<td>0.61</td>
<td>0.32</td>
<td>0.59</td>
<td>0.35</td>
<td>0.59</td>
</tr>
</tbody>
</table>

**FIGURE 10.** Statistical results of internal clustering indicators for different algorithms

C. PERFORMANCE TESTING OF ANOMALY DETECTION ALGORITHM BASED ON SC INSTANCE TRANSFER LEARNING

The Credit Card Fraud Detection was chosen as test dataset, which includes transaction information from European cardholders using credit cards in September 2013. Out of 284807 transactions, there were 492 frauds, and the dataset was highly unbalanced. Positive (fraud) accounted for only 0.172% of all transactions. The dataset has been desensitized and data cleaned. The study used the average value filling method to fill in missing values in the dataset. The features of each dimension of the dataset were standardized using the normal transformation method.

Comparison was done between the SC instance transfer learning anomaly detection algorithm and the common anomaly detection algorithms which includes the semi-supervised learning-based anomaly detection algorithm (SSNO), K-Nearest neighbors-based outliers detection method (KNNO) and semi-supervised K-Nearest neighbors based outliers detection method (SSKNNO). The dataset for anomaly detection learning often exhibited significant high imbalance, and the area under the curve (AUC) under the receiver operating characteristic curve (ROC) was selected as the evaluation indicator. ROC is the ratio of recall rate to false positive rate, while AUC value is the area enclosed by ROC curve and coordinate axis, commonly utilized to evaluate the training effect of binary classification models and evaluate the classification results of imbalanced data. The AUC value results of different anomaly detection algorithms are shown in Figure 11. As shown in Figure 11, the ROC curve of the SC instance transfer learning anomaly detection algorithm was located at the top of the coordinate axis, while the ROC curve of KNNO was located at the bottom of the coordinate axis. The AUC values were 0.913, 0.841, 0.783, and 0.692 in order of magnitude. The larger the AUC value, the better the classification effectiveness of the model, indicating that the performance of the SC instance transfer learning anomaly detection algorithm was superior to other detection algorithms.
Model performance validation experiments were conducted using a large size, high outlier detection dataset (ODDS), which was a multidimensional open dataset. The algorithm iteration number was set to 40, and the detection accuracy curves of different anomaly detection algorithms are shown in Figure 12. In Figure 12, the detection accuracy curve of the SC instance transfer learning anomaly detection algorithm was at the highest level. In two different open datasets, the maximum detection accuracy of the algorithm designed in the study reached 91.67%, which was significantly higher than other anomaly data monitoring models.

Parameter sensitivity analysis on the SC instance transfer learning anomaly detection algorithm was performed. \(k_t\) and \(k_s\) denote the neighborhood parameters of the SITL algorithm. The impact of changes in these two parameters on the model's detection performance was examined. The study set the values of \(k_t\) and \(k_s\) to 5 and 100, respectively, as the initial values for parameter analysis. The variables increased in order of 50. The experiment outcomes are denoted in Figure 13. From Figure 13, the changes in the target domain data neighborhood \(k_t\) had a small impact on the AUC value, and the AUC value curve was relatively stable with almost no significant fluctuations as the neighborhood parameters changed. However, the change in the neighborhood \(k_s\) of the source domain data had a significant impact on the AUC value. As the \(k_s\) value increased, the AUC value curve continuously decreased, with a decrease of more than half. The parameter settings of transfer learning would affect the performance of the model, and it was more appropriate to study the neighborhood parameters set.
V. CONCLUSION

A. Finding

The processing and analysis of unlabeled and weak-label data is the key to mining data information and handling computer information tasks. The research takes SC as the technical core, and constructs an IA-FDC and an anomaly detection and analysis algorithm based on SITL for unlabeled and weak-label data, respectively. The experimental results show that the precision, recall, and F1 value of the IA-FDC algorithm are significantly better than those of other intelligent optimization algorithms, and the optimal balance and reconciliation of precision and recall are achieved. The dynamic clustering algorithm designed in the research has high computational efficiency, the loss function curve converges to the minimum value, and the global search ability is strong. Compared with other advanced clustering algorithms, the dynamic clustering algorithm has better performance on the datasets in various fields, with the NMI, Purity, and RI indexes taking a larger value, and the XB, DB, and MSE indexes taking a smaller value. Internal and external indexes verify the algorithm's excellent clustering effectiveness. The classification effect of the migration learning anomaly detection algorithm of the SC example is the best, with the maximum AUC value of 0.913 and the maximum detection accuracy of 91.67%, but the parameter settings of migration learning have a greater impact on the performance of the model.

B. Future work

Comprehensively, the SC model constructed by the study has better processing ability for unlabeled and weak-label data, but the FDC algorithm designed by the study has not been analyzed for the model parameters. In the parameter sensitivity analysis, some of the parameter settings of migration learning will affect the model performance. Therefore, future research work can continue to explore the joint use of migration learning and anomaly detection analysis, and the performance of the algorithm can be further optimized.

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).

REFERENCES


