Hindawi
Journal of Applied Mathematics
Volume 2022, Article ID 2996725, 10 pages
https://doi.org/10.1155/2022/2996725

Research Article

Classification Algorithm for Heterogeneous Network Data Streams Based on Big Data Active Learning

Lili Zhan

School of Information Engineering, Harbin University, Harbin 150086, China

Correspondence should be addressed to Lili Zhan; zhanlili@hrbu.edu.cn

Received 24 August 2022; Revised 27 September 2022; Accepted 7 October 2022; Published 21 October 2022

Academic Editor: Man Leung Wong

Data classification is one of the main tasks in the current data mining field, and the existing network data triage algorithms have problems such as too small a proportion of labeled samples, a large amount of noise, and redundant data, which lead to low classification accuracy of data stream implementation. Network embedding can effectively improve these problems, but the network embedding itself has problems such as capturing relational honor and ambiguity. This study proposes a SNN-RODE based LapRLS heterogeneous network data classification algorithm to achieve deep embedding of structure and semantics among nodes by constructing a multitask SNN and selecting dead song datasets to perform mining tasks to train the neural network. Then a semisupervised learning classifier based on Laplace regular least squares regression model is designed to use the relative support difference function as the decision method and optimize the function. The simulation experimental results show that the SNN-RODE-LapRLS algorithm improves the performance by 14%-51% over the mainstream classification algorithms, and the consumption time meets the demand of real-time classification.

1. Introduction

In the era of big data, numerous domains of society generate immeasurable continuous data streams on a daily basis [1]. These data streams are often characterized by real-time, continuity, variability, and infinity [2]. In the field of data stream mining, the focus and difficulty of research lies in how to mine effective information from data streams using two basic methods, online learning and active learning, and the sampling method of neural network models is a determinant of classification quality [3]. In order to obtain the corresponding classifier, the standard data stream classification method process is to extract labeled samples, extract a subset of features, input data in the classification where, and train the classifier in the training set [4]. When the labeled samples occupy a low proportion of the dynamic data stream, the training algorithm can identify the labeled samples, and then filter out the other unlabeled samples by similarity measure and extend the dataset to be used as the training set of the classifier to improve the performance of the classifier. In order to solve the problem that labeled samples occupy too low a proportion of the dynamic data stream, this study introduces a heterogeneous network embedding framework based on Siamese Neural Network (SNN) combined with Relation Oriented Deep Embedding (RODE) to solve the problem of capturing relational redundancy and, the active learning mechanism introduces a Laplace regression model to evaluate the least squares error of labeled samples, and then constrains the active learning process with multiple constraint rules, and finally optimizes the decision function of the classifier in order to improve the performance and efficiency of data stream classification.

2. Related Work

As the technology with the largest number of users and the largest scope of influence in today's society, the main research direction of network technology is the analysis and classification of network data streams [5]. In recent years, web embedding has received attention from domestic and foreign scholars, and many scholars have conducted in-

depth research on both web embedding and web data stream classification.

Yi L et al. proposed a scoring prediction algorithm based on nonlinear feature fusion in order to fully exploit and integrate the structural features of heterogeneous information network nodes. The structural features of nodes are first extracted using a metapath-based Heterogeneous Information Network (HIN), then the structural features are transformed using a nonlinear fusion method, and finally the fused features are input into a multilayer perceptron for score prediction. The algorithm was proved to outperform the baseline through simulation experiments [6]. Liu et al. modeled the collected Net Health data as HIN in order to identify individuals vulnerable to depression and anxiety, and then a novel way to redefine the problem of predicting individual mental health status, and finally modeled individual mental health prediction as a node classification in HIN of another problem type by evaluating four node features as proof-of-concept classifiers in the process of logistic regression [7]. Sharma et al. implemented functional encryption of letters in UAV-assisted heterogeneous networks for dense urban areas to protect data from illegal intrusion; the main goal of this technique implementation is to provide proof-of-safe passage against illegal intrusion. The technique was validated using internet security protocols and application automatic verification tools and the results showed that the technique is implementable in UAV-assisted heterogeneous networks [8]. Chang et al. proposed a new metapath extraction heterogeneous graph neural network (Megnn), which can extract meaningful metapaths in heterogeneous graphs and provide data insights and interpretable conclusions for the effectiveness of the model. Megnn combines different bipartite subgraphs corresponding to edge types into a new trainable graph structure by using heterogeneous convolutions. Using the message passing paradigm of GNN through trainable convolutions, Megnn can optimize and extract effective metapaths for heterogeneous graph representation learning. A large number of experimental results on three datasets not only demonstrate the effectiveness of Megnn's method compared with the latest method but also prove that the extracted metapath has good interpretability [9]. Lei et al. used source task models to help reduce the training cost of target task models in order to reduce the neural network training cost. They proposed a new migration learning method which linearly transforms the feature mapping of the target region, increases the weights of feature matching, enables knowledge transfer between heterogeneous networks, and adds discriminators based on adversarial principles to speed up feature mapping and learning [10]. Sharipuddin et al. addressed the intrusion detection system in heterogeneous networks which is easily affected by objective factors such as devices and network protocols, proposed an identification method combining deep learning, and conducted preliminary experiments on denial-of-service attacks, and the experimental results showed that deep learning can improve the detection accuracy in heterogeneous networks [11]. Fu et al. proposed a new model called Metapath Aggregation Graph Neural Network (MAGN) to improve the final performance. MAGN uses three main components, namely, node content transformation to encapsulate input node attributes, aggregation within the metapath to merge intermediate semantic nodes, and aggregation between metapaths to merge messages from multiple metapaths. The results show that, compared with the most advanced baseline, MAGN achieves more accurate prediction results for the three real-world heterogeneous graph datasets of node classification, node clustering, and link prediction [12].

Li et al. address the problem of online learning with delayed feedback in the presence of malicious data generators based on the practical application of network data stream classification. Based on the feedback delay and the transparency or not of the malicious data generator, four algorithms that are sublinear under mild conditions are proposed and simulated for experiments. The experimental results show that the algorithms outperform the offline classification method when tested with a mixture of normal and malicious data [13]. Huang et al. proposed a similarity-based approach to detect unexpected events mentioned in social media, including natural disasters, public health events, and social security events. The method focuses on clustering social media text data based on the attributes of the events, matching time and location using regular expressions, and finally calculating the similarity of the text data. After testing, the method has good accuracy and real-time performance [14]. Liu et al. proposed an intrusion detection system, using CICIDS2017 as a dataset to train it, embedding high-dimensional features and using random forest algorithm to extract redundant features from the original dataset, with an accuracy rate of 99.93% and a false alarm rate of only 0.3%. After converting the textual data into digital features, the training time was reduced with an accuracy of 87.08%, saving 87.97% of the training time [15]. Zdemir et and Turani proposed a machine learning approach applicable to the e-commerce domain, applying Logistic Regression, Parsimonious Bayes, and Support Vector Machines as data classification algorithm, aiming to identify the model with the best accuracy [16]. Qiu et al. proposed a data stream classification model based on distributed processing in order to solve the problem of real-time detection of grid equipment anomalies. The model uses a local node mining method and a global mining model for nonuniform data stream classification, and to ensure the robustness and efficiency of the data stream classification method, a clustering algorithm is used, combined with its expression of local mining and real-time maintenance to improve the speed of information transmission between nodes and nodes [17].

Since most networks in reality are heterogeneous networks, optimization of heterogeneous network embedding framework is essentially an optimization of data preprocessing of network data classifier [18]. This study proposes a SNN-RODE embedding framework to distinguish the semantic relationships of network data based on several types and measure the similarity of similar and dissimilar

nodes to handle the relationships of different types of network structures with a view to improving the accuracy of data classification.

3. Classification Algorithm for LapRLS Network Data Streams Based on SNN-RODE Heterogeneous Network Embedding Framework

3.1. Heterogeneous Network Embedding Framework Combining SNN and RODE. Let the heterogeneous network $G = (V, E, \phi, \varphi)$, V be the set of entity nodes, E be the set of edges connecting two nodes, and each node v and each edge e be associated with their respective mapping functions $\phi(v)$. In the heterogeneous network, let the meta-path be π, which represents the composite relationship between different node and edge types to different node types.

$$\pi = a_1 \xrightarrow{r_1} a_2 \cdots a_{n-1} \xrightarrow{r_{n-1}}. \tag{1}$$

In Equation (1), π is a sequence of abstract types, representing meta-paths. a is a sequence of characteristic nodes, representing meta-path instances, which conforms to the π model. For Heterogeneous Network Embedding (HNE), it is defined as follows: let $R^d (d \ll |V|)$ be a low-latitude latent vector space, let $v(v \in V)$ be a node, and mapping v to R^d be HNE. The basic process is to input a graph, and then generate and output a low-latitude embedding representation corresponding to the graph. There are two main modules in the HNE framework, which are the relation extraction module and network embedding module. Among them, the relationship extraction module models the structural and semantic relationships extracted from the HN using the similarity existing between nodes and nodes. Let the similarity retention matrix be M, if $M_{ij} = 1$, it means that there is similarity between structure and semantics; if $M_{ij} = 0$, it means that there is no similarity between structure and semantics. Besides, when v_i and v_j are connected by the same edge in the single-hop structure, $M_{ij} = 1$, it means there is semantic similarity. Since each special element path in the heterogeneous network contains different types of nodes and edges, after obtaining the structural similarity, in order to approximate the neighborhood structure into a low-dimensional potential space, the loss function needs to be optimized.

$$L_{\text{sim}} = \left\| z_i - z_j \right\|_F^2 M_{ij}. \tag{2}$$

In Equation (2), z_i and z_j are the low-dimensional vectors of the nodes v_i and v_j, respectively, and M_{ij} is the term of the i and j columns in M and the label to determine the similarity of v_i and v_j. For the distance between the two node vectors, the Euclidean distance is measured. When there is similarity in the structure of two nodes, it means that the distance between the vectors corresponding to the nodes is the shortest. At this point, the optimization objective for HNE is to minimize the loss function [19]. After discarding the redundant relations, the similarity between them constitutes a specific semantic relation, regardless of whether the nodes are of the same type or not. When nodes v_i and v_j exist a, it means that the two nodes have semantic similarity. From Equation (2), it can be seen that the semantic similarity is approximated to the nodes that closely express semantic similarity in the embedding space, except for the neighboring nodes of different types, which are also far away from each other in the embedding space. The embedding formula that encodes the differences between different types of nodes is

$$L_{\text{dissim}} = \left(1 - M_{ij}\right)_{\max} \left(0, m - \left\| z_i - z_j \right\|_F\right)^2. \tag{3}$$

In Equation (3), m is a nonnegative constant that represents the marginal value. In theory, the larger the distance between two nodes, the less similar they are. The criterion to determine whether m achieves the desired goal is whether the embedding variance of nodes of different types is greater than the embedding variance of nodes of the same type. Again, because metapath capture to be similar requires determining the length of the metapath, longer metapaths can connect remote nodes that are semantically less relevant, and shorter metapaths can preserve more specific semantics. SNN is a special type of convolutional neural network, compared to standard convolutional neural network, which uses the given data to determine a similarity measure and compare the similarity of new samples without determining to which class each sample label belongs. The structure of a standard SNN is shown in Figure 1.

In Figure 1, since the loss functions of the two subnetworks are the same, the semantic distance action of the original space can be maintained, and the similarity of the two inputs can be determined by calculating the distance of the input vectors with the set distance metric. The formula of the SNN hidden layer is

$$h^{(k)}(\hat{x}) = f^{(k)}\left(W^{(k)}\hat{x} + b^{(k)}\right). \tag{4}$$

In Equation (4), $h^{(k)}$ is the output of the first k hidden layer, f is the nonlinear activation function, and $\hat{x}$ is the representation vector of the input nodes. When $k = 1$, $\hat{x} = x_i (i = 1, 2, \cdots, s, n)$, and when $k > 1$, $\hat{x} = h^{(k-1)}(x_i)$ or $\hat{x} = h^{(k-1)}(x_j)$; the subnetworks share the hidden layer with equal weights W and base b. The equation of SNN output layer is

$$\hat{z} = \text{Relu}\left(W^{(3)}h^{(2)}(\hat{x}) + b^{(3)}\right). \tag{5}$$

In Equation (5), $\hat{z}$ is the low-dimensional node representation in the embedding into space, and Relu is the activation function. SNN randomly selects pairs of nodes and maps them into the low-dimensional embedding space, and in order to embed the similarity and dissimilarity between nodes into the representation, the loss values of the samples are calculated using the objective function,

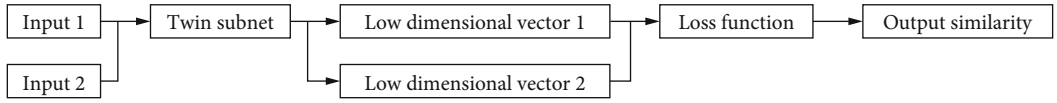


Figure 1: Standard twin neural network structure.

Table 1: Datasets information.

Data set	Basic functions and node types			Total data
Habitat datasets of swan	Swan number	Migration times	Migration distance	1317
The Movies Datasets	Number of users	Film category	Viewing times	48649
Word Net-WN11	Position	Relationship	Semantics	157962

Table 2: Link prediction AUC.

Frame	AUC		
	Habitat datasets of swan	The Movies Datasets	Word Net-WN11
Deep Walk	0.6124	0.7125	0.5721
Node2vec-Megnn	0.6328	0.7841	0.7893
Metapath2vec	0.7782	0.8531	0.8217
LINE-Megnn	0.7894	0.7352	0.6914
SNN-tri	0.9013	0.8427	0.8129
RODE	0.9274	0.9011	0.9173

which contains the unity of similarity and dissimilarity.

$$L = \sum_{k=1}^{K} \left(y\left\|z_i - z_j\right\|\right)_F^2 + (1-y)_{\max}\left(0, m - \left\|z_i - z_j\right\|_F^2\right). \quad (6)$$

In Equation (6), y is the label that distinguishes positive samples from negative samples, $y = M_{ij}$. When $M_{ij} = 1$, the sample is a positive sample; when $M_{ij} = 0$, the sample is a negative sample, where positive samples are the unity of structural similarity and to be similar. In order to optimize the representation of nodes in the RODE framework, the objective function is optimized using Stochastic Gradient Descent (SGD) so that the objective function can more accurately calculate the loss of node similarity and node dissimilarity being mapped to the low-dimensional embedding space, and the optimization formula is

$$\theta \coloneqq \theta - \eta \frac{\partial L}{\partial \theta}. \quad (7)$$

In Equation (7), η is the learning rate. ∂ is the number of iterations, and θ is the randomly selected pairs of nodes from the heterogeneous network.

3.2. Optimized Loss Function for LapRLS Network Data Stream Classification Algorithm. Let the classifier model that determines the classification result by the support function be ψ, and the support formula for each classification is

$$\begin{aligned} F &= \{F_1, F_2, \cdots, F_n\}, \\ \psi(z) &= \max_{k\in\delta}(F_k(z)). \end{aligned} \quad (8)$$

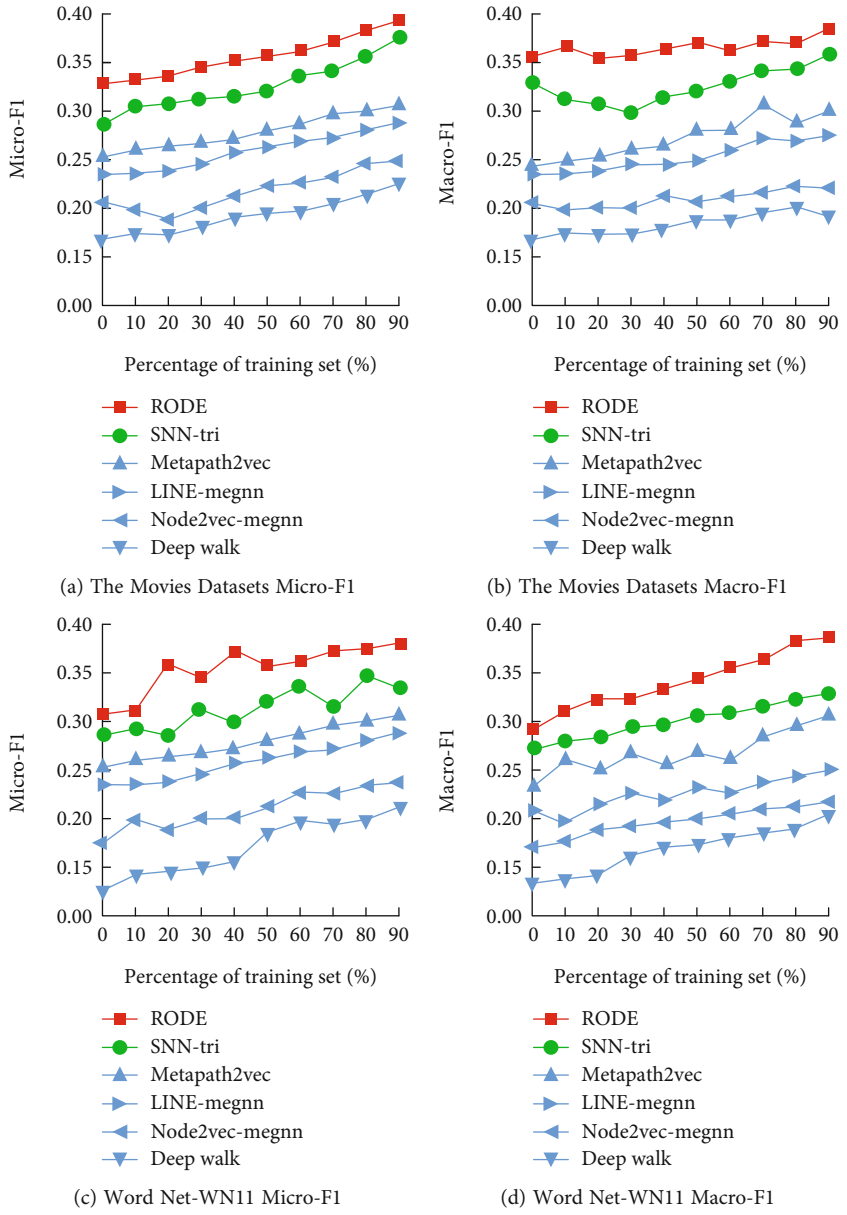
In Equation (8), z is the feature vector, k is the set of class tokens, and $\delta = \{1, 2, \cdots, n\}$ is the classification set. The performance of the classifier model can hardly be improved by increasing the support of the decision, indicating the low importance of the support of the decision. The support is proportional to the correct classification rate, and the difference in support is proportional to the uncertainty. To evaluate the difference between the maximum support of x and other types of support, the Relative Support Difference Function (RSDF) is used.

$$\mathrm{RSDF}(x) = \frac{\sum_{i=1}^{n}\left[\max_{k\in\delta}(F_k(x)) = F_i(x)\right]}{n-1}. \quad (9)$$

Unlike the ordinary support degree method, RSDF divides the boundaries for data points instead of dividing regions for data points. In general, the similarity between neighboring data streams is large, and RSDF generates new decisions when the support difference is large. When the support difference is small, the current classification model is maintained to reduce the number of classifiers trained and to substantially improve the efficiency of the system. Sequence is the usual expression of the data stream, and the order of the samples determines the high performance of the classifier [20]. Usually, a randomized preprocessing of the samples of the data stream is required. This is only a threshold to filter the sample data, and a threshold value greater than the support difference indicates that the classifier performance is insufficient and it needs to be trained again. In order to reduce the cost of collecting labeled samples for the data stream, an active learning mechanism of Laplace least squares regression model is designed, whose linear function is expressed.

$$y = x\beta + \varepsilon. \quad (10)$$

In Equation (10), x and y are the input and output of the variables, respectively, and β and ε are the unknown errors



(a) The Movies Datasets Micro-F1

(b) The Movies Datasets Macro-F1

(c) Word Net-WN11 Micro-F1

(d) Word Net-WN11 Macro-F1

Figure 2: Multilabel node classification.

of the weight vector and the mean of 0, respectively. Let the errors under different observations independently λ^2 have equal variance, x and β are determined, and the output equation is

$$f(x) = x\beta. \tag{11}$$

Let there exist a set of labeled samples $(z_1, k_1), (z_2, k_2), \cdots, (z_n, k_n)$, where k_n is the label of z_n. Calculate β by minimizing the sum of mean squared errors.

$$J_{\mathrm{SSE}}(\beta) = \sum_{i=1}^{m} (z_i\beta - z_i)^2. \tag{12}$$

In Equation (12), $Z = (z_1, z_2, \cdots, z_n)^T$ is the eigen matrix, $K = (k_1, k_2, \cdots, k_n)^T$ is the marker vector, and the covariance matrix formula is

$$\mathrm{cov}\left(\widehat{\beta}\right) = \lambda^2 \left(Z^T Z\right)^{-1}. \tag{13}$$

The objective of OED is to select the most confident sample set $Z = \{z_1, z_2, \cdots, z_n\}$ from the unlabeled dataset $X = \{x_1, x_2, \cdots, x_n\}$, and OED transforms the optimization problem into minimizing the variance of the estimation module. There are two main types of unsupervised learning in OED, one minimizes the trace of cov $(\widehat{\beta})$ and the other minimizes the decision of cov $(\widehat{\beta})$. To solve the problem of insufficient training samples for these two types of unsupervised learning, Laplacian Regularized Squares (LapRLS) are

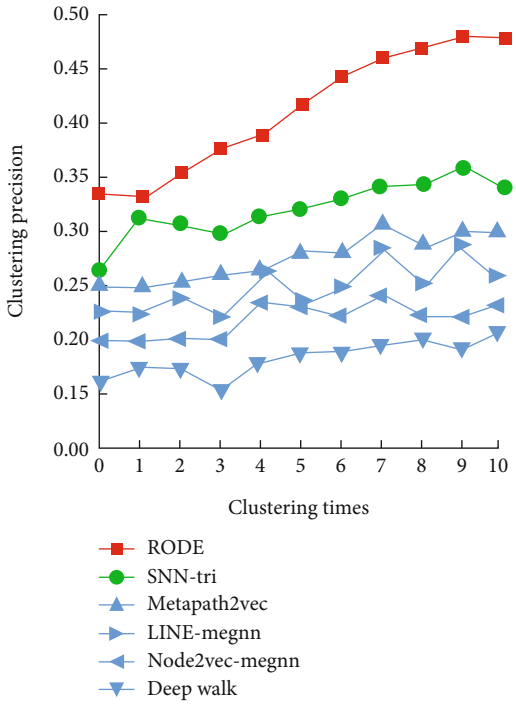


Figure 3: Frame clustering results.

Table 3: Node cluster NMI.

Frame	NMI
Deep Walk	0.178
Node2vec-Megnn	0.298
Metapath2vec	0.225
LINE-Megnn	0.209
SNN-tri	0.317
RODE	0.398

introduced. The data can be utilized simultaneously regardless of whether they are labeled or not. Let the sample $x_i \approx x_j$, then its fitness $f(x_i) \approx f(x_j)$. Let a and b be the total amount of data and the number of labeled data, and W be the similarity matrix, and the loss function of LapRLS is

$$J_{\mathrm{LapRLS}}(\beta) = \sum_{i=1}^{b} (f(z_i) - y_i)^2 + \frac{\kappa_1}{2} \sum_{i,j=1}^{a} \left(f(x_i) - f(x_j)\right)^2 W_{ij} + \kappa_2 \|\beta\|^2. \tag{14}$$

In Equation (14), κ and κ_1 are the regularization parameters and the smoothness penalty terms, which serve to maintain the flow structure of the input space. κ_2 The role of is to control the sparsity of the regression model. If two adjacent points are to be separated, a lower weight value needs to be assigned to the loss function. The equation defining the similarity matrix W_{ij} is

$$W_{ij} = \begin{cases} \exp\left(-\dfrac{x_i - x_j^2}{\lambda^2}\right), x_i \in N(x_j), x_j \in N(x_i), \\ 0, \text{others}. \end{cases} \tag{15}$$

Since the principle of LapRLS is to estimate linear functions to describe the flow structure based on whether the samples are labeled or not, while Laplace regularized OED performs better in the linear model and has average effect in the nonlinear model. To better describe the nonlinear system, let H denote the Reproducing Kernel Hilbert Space (RKHS), and rewrite Equation (14) as the RKHS formula.

$$\min_{f \in H} \sum_{i=1}^{b} (f(z_i) - y_i)^2 + \frac{\kappa_1}{2} \sum_{i,j=1}^{a} \left(f(x_i) - f(x_j)\right)^2 W_{ij} + \kappa_2 \|f\|_H^2. \tag{16}$$

The OED was improved by using the Laplace regular model. When the proportion of labeled samples to all samples is low, unlabeled samples in high-density regions have more influence on the classifier accuracy than unlabeled samples in low-density regions. This indicates that the unlabeled samples in high-density regions are more representative in the feature space of the regression model.

4. Heterogeneous Network Embedding Framework Training and Network Data Stream Classification Algorithm Testing

4.1. SNN-RODE Link Prediction Training and Node Classification Training. In order to evaluate the performance of the SNN-RODE model more comprehensively, three datasets with different node densities and network sizes were selected for training, and the information of the datasets is shown in Table 1. From Table 1, Habitat datasets of swan records the number, migration frequency, and migration distance of a swan colony through GPS, and the semantic information of the swan colony to a specific location can be obtained by analyzing these data; The Movies Datasets records the number of users, movie types, and movie viewing frequency of a movie website, which describes the information of individual The Movies Datasets record the number of users, types of movies, and the number of times a movie is viewed on a movie website, describing individuals' preferences for movies.

The five basic network embedding frameworks compared in this study are Deep Walk, Node2vec-Megnn, Metapath2vec, LINE-Megnn, and SNN-tri. To obtain the best framework performance, trial-and-error experiments were conducted, setting the window size to 10, the node walk to 100, the walk length of each node to 40, and the number of negative samples to 5. SNN was set to one input layer, four hidden layers and one output layer. Among them, the sizes of the four hidden layers are 256, 512, 1024, and

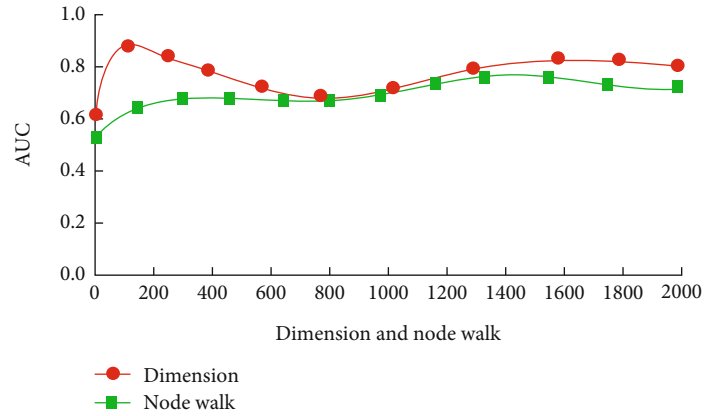


Figure 4: Influence of different embedding dimensions and node routing on AUC.

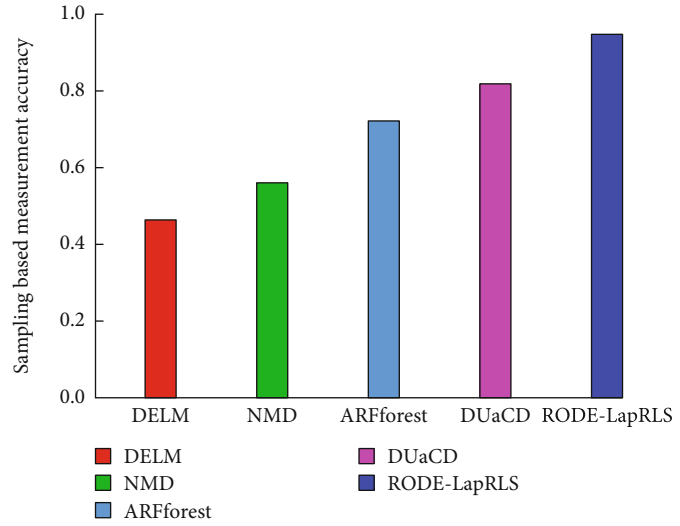


Figure 5: Model performance comparison.

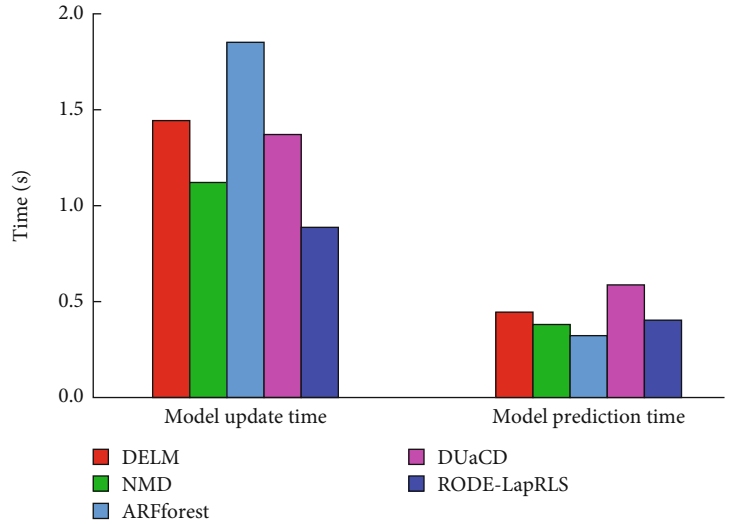


Figure 6: Model efficiency comparison.

2048, the learning efficiency is set to 0.02, and the size of the output layer is set to 64.

Link prediction is a fundamental task in data mining, predicting the edge between two nodes by actively learning the embedding representation [21]. The experimental results of link prediction are shown in Table 2.

As shown in Table 2, LINE-Megnn and Node2vec-Megnn did not consider the type of nodes compared to Deep Walk, but only improved the network structure. As seen from the AUG values, LINE-Megnn and Node2vec-Megnn improve 3.33% and 28.90%, 10.05% and 3.19%, 37.97% and 20.85% on the three datasets, respectively, with limited improvement in the quality of heterogeneous network embedding. While SNN-tri and RODE improve significantly, the best performing RODE improves performance by 51.44%, 26.47%, and 60.30% compared to Deep Walk. While SNN-tri and RODE possess the same parameters of SNN, RODE optimizes the loss function and improves the performance by 2.90%, 6.93%, and 12.84%. The node classification task is usually performed on data with label information and, in this study, the datasets with label information are The Movies Datasets and Word Net-WN11. The data nodes without labels in the datasets are first removed, and then trained using the training set, using Micro-F1 and Macro-F1 as metrics. The experimental results are shown in Figure 2.

As shown in Figure 2, RODE exceeds SNN-tri by 0.01~0.02 in the Micro F1 indicator of The Movies Datasets dataset and exceeds SNN-tri by 0.03~0.04 in the Macro F1 indicator. In the Word Net-WN11 dataset, RODE exceeds SNN-tri by 0.04-0.05 in the Micro-F1 indicator. On the Macro-F1 index, the RODE exceeds SNN-tri by 0.07~0.08. It can be seen that other frameworks will not maintain similarity in the dataset and cannot generate competitive embedded expressions. RODE has better performance. Meanwhile, by comparing SNN-tri with the other four frameworks, it can be seen that Micro-F1 metrics and Macro-F1 metrics lead by 0.08-0.15 and 0.06-0.14, respectively, indicating that SNN can preserve structural and semantic similarity and the improved loss function can improve the embedding quality of heterogeneous networks. And RODE not only captures more comprehensive relationships but also has stability in handling sparsity problems. Besides link prediction, clustering task is also a common task. The training is conducted in Word Net-WN11 dataset, set one node to one label, and the clustering algorithm of synonym nodes is K-means. Normalized Mutual Information (NMI) is selected as the training metric, and the training times are set to 10 times, and the training results are shown in Figure 3.

As shown in Figure 3, the RODE always maintained a high precision in the 10 clustering experiments, and the precision gradually increased with the number of clustering experiments, with the highest precision of 0.48, which was 0.12 higher than the highest precision of SNN-tri. The average NMI scores were recorded as shown in Table 3.

As can be seen from Table 3, the NMI of SNN-tri is improved by 0.027-0.153 compared to the four frameworks without SNN, while the RODE optimized by loss function is improved by 0.081 compared to the NMI of SNN-tri. In practical applications, the size of the label set tends to be large, and the NMI scores of the nodes are generally lower. And the NMI of RODE demonstrates that while in unsupervised clustering training, embedding representations showing structural and semantic similarity can get better structure. Besides exploring the effect of link prediction on AUC, it is also necessary to analyze the effect of different embedding dimensions and node walks on AUC. The Word Net-WN11 with the largest size is selected as the test set, and the obtained experimental results are shown in Figure 4.

As can be seen from Figure 4, the value of AUC increases with the increase of dimensionality when the dimensionality is less than 100; it starts to decrease when the dimensionality is greater than 100, rises again when it is close to 800, and decreases again when the dimensionality is equal to 1800. The trend of node walk is similar to the trend of dimension, but the stability is better, and the change is not as fast as the dimension change, and the magnitude is smaller. The AUC reaches the maximum when the dimension is equal to 100, and the node walk is equal to 200. However, as the number of dimensions and node walks continue to increase, the embedded expression leaves behind some information that does not have value, and the AUC decreases. Therefore, the common expression of the two metrics must find a balance of performance and time cost when performing link prediction.

4.2. RODE-LapRLS Classification Performance and Efficiency Tests. The current approaches for data flow experimentation and evaluation are mainly maintenance evaluation methods and prediction sequence methods [22]. The maintenance evaluation method uses the current classification model to process an independent test set; the prediction sequence method has to predict the marker of each reach sample, and then use this sample to update the learning model. The evaluation metric used is the sampling-based metric accuracy, which needs to consider the time cost in addition to performance. Therefore, the update time and prediction time of the model need to be evaluated, and the classification models are evaluated comprehensively. The models compared are ARForest, NDM, DUaCD, and DELM. The reasons for choosing these models are that ARForest and DELM are both predictive classification models with active learning mechanisms, and both DUaCD and DELM are involved in the concept drift problem, and NDM proposes diversity assessment methods and both can be effectively compared with RODE-LapRLS.

As known from Figure 5, the accuracy of the five models based on the adopted metrics is 0.45 (DELM), 0.58 (NMD), 0.76 (ARForest), 0.82 (DUaCD), and 0.96 (RODE-LapRLS), respectively. The accuracy is improved by 0.51, 0.48, 0.20, and 0.14, respectively, possessing a significant enhancement effect, and the time cost comparison is shown in Figure 6.

As can be seen from Figure 6, the update times of the five models ranged between 0.8 and 1.8 seconds, and the prediction times ranged between 0.3 and 0.7 seconds. In terms of model update time, RODE-LapRLS occupies a large advantage, about 0.4 seconds faster than the second place NMD,

while in terms of model prediction time, the fastest is ARForest taking 0.3 seconds, the slowest DUaCD 0.7 seconds, and RODE-LapRLS with 0.4 seconds, although RODE-LapRLS is not the fastest, the overall difference is not large. Taking into account, RODE-LapRLS is sufficient to meet the demand of real-time and also has better performance.

5. Conclusion

In today's fast-changing Internet technology, we inevitably need to deal with all kinds of network data, such as shopping platforms that generate various transaction data, traffic management systems monitor vehicle movement data, news reports from around the world, etc. How to extract effective information from the complicated and huge amount of data is the focus of data mining work. In this study, we propose a heterogeneous network embedding framework combining twin neural networks and depth-oriented relationships to address the shortcomings of network embedding and data stream classification in data mining work and introduce a Laplace regular least squares regression model in data stream classification to optimize the loss function of the classifier. After training on four datasets, the performance of the heterogeneous network embedding framework improves from 2.7% to 15.3%, and the best performance of the embedding framework is achieved when the dimensionality is around 100 and the node walk is around 200. After the optimization of the data stream classification algorithm, the algorithm classification accuracy is improved by 14%-51%, while in terms of efficiency, the time consumption of the improved algorithm increases by only 0.1 seconds. It can be seen that the web data classification algorithm after embedding the framework is not only substantially improved in performance but also meets the real-time classification requirements. The shortcoming of this study is that the effect of classifier threshold on classification results was not explored in the preprocessing experiments, and future studies can be conducted from this perspective.

Data Availability

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Conflicts of Interest

It is declared by the author that this article is free of conflict of interest.

References

[1] S. Luo, X. Chen, Z. Zhou, and S. Yu, "Fog-enabled joint computation, communication and caching resource sharing for energy-efficient IoT data stream processing," *IEEE Transactions on Vehicular Technology*, vol. 70, no. 4, pp. 3715–3730, 2021.

[2] N. Nagendhiran and L. Kuppusamy, "Adaptive drift detection mechanism for non-stationary data stream," *Journal of Information and Knowledge Management*, vol. 20, no. 1, pp. 2150008–2150925, 2021.

[3] C. Liu, Y. Chen, and L. Zhao, "An adaptive prediction method based on data stream mining for future driving cycle of vehicle," *Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering*, vol. 235, no. 6, pp. 1702–1712, 2021.

[4] R. Xin, Z. Jiang, and Y. Shao, "Complex network classification with convolutional neural network," *Tsinghua Science and Technology*, vol. 25, no. 4, pp. 447–457, 2020.

[5] A. Latif, L. A. Fitriana, and M. R. Firdaus, "Comparative analysis of software effort estimation using data mining technique and feature selection," *JITK (Jurnal Ilmu Pengetahuan dan Teknologi Komputer)*, vol. 6, no. 2, pp. 167–174, 2021.

[6] L. Yi, S. Ji, L. Ren, R. Su, and Y. Liang, "A nonlinear feature fusion-based rating prediction algorithm in heterogeneous network," *IEEE Transactions on Computational Social Systems*, vol. 8, no. 3, pp. 728–736, 2021.

[7] S. Liu, F. Vahedian, D. Hachen et al., "Heterogeneous network approach to predict individuals' mental health," *ACM Transactions on Knowledge Discovery from Data*, vol. 15, no. 2, pp. 1–26, 2021.

[8] D. Sharma, S. K. Gupta, A. Rashid, S. Gupta, M. Rashid, and A. Srivastava, "A novel approach for securing data against intrusion attacks in unmanned aerial vehicles integrated heterogeneous network using functional encryption technique," *Transactions on Emerging Telecommunications Technologies*, vol. 32, no. 2, pp. 1–32, 2020.

[9] Y. Chang, C. Chen, W. Hu, Z. Zheng, X. Zhou, and S. Chen, "Megnn: meta-path extracted graph neural network for heterogeneous graph representation learning," *Knowledge-Based Systems*, vol. 235, article 107611, 2022.

[10] F. Lei, J. Cheng, Y. Yang, X. Tang, V. S. Sheng, and C. Huang, "Improving heterogeneous network knowledge transfer based on the principle of generative adversarial," *Electronics*, vol. 10, no. 13, pp. 1525–1536, 2021.

[11] S. Sharipuddin, B. Purnama, K. Kurniabudi et al., "Intrusion detection with deep learning on internet of things heterogeneous network," *Institute of Advanced Engineering and Science*, vol. 10, no. 3, pp. 735–742, 2021.

[12] X. Fu, J. Zhang, Z. Meng, and I. King, "MAGNN: Metapath Aggregated Graph Neural Network for heterogeneous graph embedding," in *Proceedings of The Web Conference 2020*, pp. 2331–2341, New York, 2020.

[13] Y. Li, B. Liang, and A. Tizghadam, "Robust online learning against malicious manipulation and feedback delay with application to network flow classification," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 8, pp. 2648–2663, 2021.

[14] L. Huang, G. Liu, T. Chen, H. Yuan, P. Shi, and Y. Miao, "Similarity-based emergency event detection in social media," *Journal of Safety Science and Resilience*, vol. 2, no. 1, pp. 11–19, 2021.

[15] I. H. Liu, C. H. Lo, T. C. Liu, J. S. Li, C. G. Liu, and C. F. Li, "IDS malicious flow classification," *Journal of Robotics Networking and Artificial Life*, vol. 7, no. 2, pp. 103–106, 2020.

[16] R. Zdemir and M. Turanl, "Comparison of machine learning classification algorithms for purchasing forecast," *Journal of Life Economics*, vol. 8, no. 1, pp. 59–68, 2021.

[17] Y. Qiu, G. Du, and S. Chai, "A novel algorithm for distributed data stream using big data classification model," *International*

Journal of Information Technology and Web Engineering, vol. 15, no. 4, pp. 1–17, 2020.

[18] R. D. Somashekhara and C. Dr, "Generalized light gradient boost classifier for traffic aware seamless mobility management in heterogeneous network," *Indian Journal of Computer Science and Engineering*, vol. 11, no. 1, pp. 36–47, 2020.

[19] J. Xie, W. Gao, and C. Li, "Heterogeneous network selection optimization algorithm based on a Markov decision model," *China Communications*, vol. 17, no. 2, pp. 40–53, 2020.

[20] Z. Sun, N. Chang, C. F. Chen, C. Mostafiz, and W. Gao, "Ensemble learning via higher order singular value decomposition for integrating data and classifier fusion in water quality monitoring," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, no. 1, pp. 3345–3360, 2021.

[21] M. O. Xin, N. Pang, and N. Liu, "THS-GWNN: a deep learning framework for temporal network link prediction," *Frontiers of Computer Science*, vol. 16, no. 2, pp. 174–176, 2022.

[22] H. Ozkan and E. Kerman, "Comparative evaluation of RNAlater solution and snap frozen methods for gene expression studies in different tissues," *Revista Romana de Medicina de Laborator*, vol. 28, no. 3, pp. 287–297, 2020.