IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received June 23, 2021, accepted July 7, 2021, date of publication July 12, 2021, date of current version July 19, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3096194

Representation Learning With Dual Autoencoder for Multi-Label Classification

YI ZHU[1,2,3], YANG YANG[1], YUN LI[1], JIPENG QIANG[1], YUNHAO YUAN[1], AND RUNMEI ZHANG[4]

[1]School of Information Engineering, Yangzhou University, Jiangsu, Yangzhou 225009, China
[2]Key Laboratory of Knowledge Engineering with Big Data, Ministry of Education, Hefei University of Technology, Hefei 230009, China
[3]School of Computer Science and Information Engineering, Hefei University of Technology, Hefei 230009, China
[4]School of Electronics and Information Engineering, Anhui Jianzhu University, Hefei 230022, China

Corresponding author: Yun Li (liyun@yzu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61906060, in part by the Open Project Program of Key Laboratory of Huizhou Architecture in Anhui Province under Grant HPJZ-2020-02, and in part by the Open Project Program of Joint International Research Laboratory of Agriculture and Agri-Product Safety, the Ministry of Education of China, Yangzhou University, under Grant JILAR-KF202104.

ABSTRACT Multi-label classification aims to deal with the problem that an object may be associated with one or more labels, which is a more difficult task due to the complex nature of multi-label data. The crucial problem of multi-label classification is the more robust and higher-level feature representation learning, which can reduce non-helpful feature attributes from the input space prior to training. In recent years, deep learning methods based on autoencoders have achieved excellent performance in multi-label classification for the advantages of powerful representations learning ability and fast convergence speed. However, most existing autoencoder-based methods only rely on the single autoencoder model, which pose challenges for multi-label feature representations learning and fail to measure similarities between data spaces. To address this problem, in this paper, we propose a novel representation learning method with dual autoencoder for multi-label classification. Compared to the existing autoencoder-based methods, our proposed method can capture different characteristics and more abstract features from data by the serially connection of two different types of autoencoders. More specifically, firstly, the algorithm of Reconstruction Independent Component Analysis (RICA) in sparse autoencoder is trained on patches on all training and test dataset for robust global feature representations learning. Secondly, with the output of RICA, stacked autoencoder with manifold regularization (SAMR) is introduced to ameliorate the quality of multi-label features learning. Comprehensive experiments on several real-world data sets demonstrate the effectiveness of our proposed approach compared with several competing state-of-the-art methods.

INDEX TERMS Multi-label classification, dual autoencoder, RICA, manifold regularization, representation learning.

I. INTRODUCTION

Recent years have witnessed many approaches to solve the problem of that one object may associate with a set of labels, which is also commonly framed as the multi-label classification problem [1]. Different from binary class and multi-class classification in single-label problem, the intrinsic multi-label nature of most real datasets could represent the world more exactly [2]–[4]. In addition, multi-label learning has a wide range of applications in news classification [5], image processing [6] and other fields [7]. For example, a scenery image in MS COCO data set [8] may contain car, person, sky, boat and so on, which can be regarded as a multi-label image. One news article can be classified to several topics like politics and sport due to it reports that one athlete is running for president.

The traditional multi-label classification methods, such as problem transformation and algorithm adaptation methods, either transform multi-label into single-label classification or extend specific algorithm for multi-label data [9]. For example, classifier chains methods [10], [11] built binary classification in the chain based on the previous predictions for the transformation from multi-label classification to a chain of binary classification. Multi-Label k-Nearest Neighbor (ML-kNN) method [12] introduced kNN algorithm for multi-label classification, where maximum a posteriori principle is utilized to predict label information for the instances.

The associate editor coordinating the review of this manuscript and approving it for publication was Chun-Hao Chen.

Considering the shortcomings in the neglect of label correlations, many correlation-enabling methods attempted to improve the generalization ability of system in recent decades [13]. Tahir *et al.* [14] proposed to integrate stacked Spectral Regression based Kernel Discriminant Analysis (SRKDA) with ML-kNN, which can be utilized for dimensionality reduction and multi-label classification simultaneously, both correlation and high dimensionality problems can be tackled in this method. Alali and Kubat [15] proposed to reduce error-propagation and prune unnecessary label dependencies with classifier-stacking method, the stacking structure is used in this method to fulfill controlled label correlation exploitation by pruning the uncorrelated outputs. However, the main problem of these methods is the learning ability of feature representations for multi-label classification, the raw input data that used in the learning system directly may contain non-helpful features and deteriorate the classification performance.

Due to the tremendous success in feature representation learning, there have already been some efforts on devoting deep-based methods to multi-label classification. Wang *et al.* [16] combined recurrent neural networks (RNNs) with deep convolutional neural networks (CNNs) for multi-label image classification, and a joint image-label embedding is learned to model the label co-occurrence dependency in an end-to-end way. Wang *et al.* [17] proposed a label graph superimposing method based on graph convolution network (GCN) for multi-label recognition, the knowledge graph is superimposed into statistical graph for label correlation learning, and lateral connection is conducted for label-feature correlation modeling. However, these methods always suffer from the lack of labeled data, which is often expensive and laborsome in the real world. Recently, the autoencoder based models have achieved sound performance for the superiority of powerful representations learning ability and fast convergence speed [18]. Yeh *et al.* [19] proposed canonical correlated autoencoder based on deep neural networks for more desirable performance on multi-label classification, and a joint feature and label embedding is performed to better relate feature and label domain data. Huang *et al.* [20] proposed a two-encoding layer autoencoder to share knowledge with the second encoding weight matrix, both representation learning and multi-label learning is jointly optimized with the autoencoder model for the improvement of multi-label classification performance. However, these autoencoders-based methods just relied on the single autoencoder model, which pose challenges for multi-label feature representations learning and fail to measure similarities between data spaces.

To address these problems, we propose a novel Representation Learning method with Dual Autoencoder for multi-label classification (RLDA for short), in which we can capture different characteristics and more abstract features from data by the serially connection of two different types of autoencoders. Specifically, firstly, the algorithm of Reconstruction Independent Component Analysis (RICA) in sparse autoencoder is trained on patches on all training and test dataset for robust global features learning. Then, with the output of RICA, a stacked autoencoder with manifold regularization (SAMR for short) is applied to improve the quality of multi-label feature representations. Finally, we can obtain the new feature representations for multi-label classification by serially connecting two different types of autoencoders. Extensive experiments on several real-world data sets demonstrate the effectiveness of our proposed RLDA compared with other state-of-the-art methods. The main contributions of this paper are summarized as follows:

- We propose a novel representation learning method called RLDA, which exacts different characteristics and more abstract features from data by serially connection of two different types of autoencoders for multi-label classification.
- A algorithm of RICA and the method of stacked autoencoder with manifold regularization (SAMR) are introduced to learn more discriminative and abstract features, which can discover latent knowledge of the raw input data for multi-label learning.
- The comprehensive experiments over four real data sets show that our method outperforms state-of-the-art models and evaluate the effectiveness of our method.

The remainder of this paper is organized as follows. Some preliminary knowledge used in our proposed method is reviewed in Section II and details of the proposed RLDA method are provided in Section III. Experimental results and analysis on four real world datasets are presented in Section IV, followed by the related work is introduced in Section V. Finally, our conclusions are summarized in Section VI.

II. PRELIMINARIES

A. AUTOENCODER

The autoencoder model [21] is an unsupervised feature representation learning model, which aims to learn an approximate representation of the input by the encoder and decoder layers. Autoencoder has already been one of the most successful deep neural networks and actively adopted as a multi-label classification model recently. Given the input as $\{x_1, x_2 \ldots, x_i \ldots, x_n\}$, where $x_i \in \Re^m$, the autoencoder model attempts to learn an approximate output $h_{W,b}(x) \approx x$. Specifically, the autoencoder model usually contains one encoder and one decoder layer respectively. In encoder layer, the input is encoded to one or more hidden layers through several encoding processes, then the hidden layers are decoded to the output as $\hat{x}$. The encoder and decode layer in autoencoder model that just includes one hidden layer can be represented as (1) and (2):

$$\xi = f(W_1 x + b_1) \tag{1}$$

$$\hat{x} = g(W_2 \xi + b_2) \tag{2}$$

where $W_1 \in \Re^{k \times m}$ and $W_2 \in \Re^{m \times k}$ are the weight matrixes, $b_1 \in \Re^{k \times 1}$ and $b_2 \in \Re^{m \times 1}$ are the bias vectors, $\xi \in \Re^{k \times 1}$ is

the output of hidden layer, f and g are the nonlinear activation function of encode and decode layers respectively. For succinctness, the original input data are denoted as $\{x_i\}_{i=1}^n$, thus the reconstruction error can be expressed as $\sum_{i=1}^n \left\|\hat{x}_i - x_i\right\|^2$. The crucial problem of the autoencoder model is to minimize the reconstruction error by the parameters learning about W_1, W_2, b_1 and b_2, which can show as (3):

$$\min_{W_1,W_2,b_1,b_2} \sum_{i=1}^{n} \left\|\hat{x}_i - x_i\right\|^2 \tag{3}$$

B. RECONSTRUCTION INDEPENDENT COMPONENT ANALYSIS (RICA)

Reconstruction Independent Component Analysis (RICA) model [22] aims to exact sparse representations of whitened or non-whitened data from unlabeled data, which tries to learn a set of linearly independent basis features to represent input data accurately. Given the input as x, in order to learn the output which is represented in the columns of a weight matrix W as shown in (4):

$$J(W) = \|Wx\|_1 \tag{4}$$

The optimization of the RICA's objective function is represented as (5):

$$\min_W \lambda \|Wx\|_1 + \frac{1}{2}\left\|W^T Wx - x\right\|_2^2 \tag{5}$$

Compared to the above-mentioned objective function of autoencoder, the reconstructive penalty is added for scaling up to over-complete features, which is shown as the second item in (5). Since the objective function of RICA has no analytic solution, the gradient of reconstruction cost is driven with the back-propagation idea for the optimization. The gradient with respect to W^T is transposed to the gradient with respect to W, and the final gradient with respect to W is shown as (6):

$$\begin{aligned}\nabla_W J &= \nabla_W J + (\nabla_{W^T} J)^T \\ &= (W)(2(W^T Wx - x))x^T + 2(Wx)(W^T Wx - x)^T\end{aligned} \tag{6}$$

C. MANIFOLD REGULARIZATION

Manifold regularization aims to construct a graph connecting similar observations for unsupervised or semi-supervised learning, and label information propagates through the graph from labeled nodes to unlabeled ones by finding the minimum energy configuration [23]. In our proposed method, we incorporate manifold learning as a regularization into autoencoder for enforcing neighbors located in the same local structure on the representation space. Given the input as $\{x_1, x_2 \ldots, x_i \ldots, x_j \ldots, x_n\}$, the adjacent matrix K can be shown as (7):

$$K^{ij} = \begin{cases} 1, & x_i \in NN(k, x_j) or x_j \in NN(k, x_i) \\ 0, & Otherwise \end{cases} \tag{7}$$

where $NN(k, x_j)$ is the k nearest neighbors of x_j and k is the hyper-parameter. The distance between x_i and x_j can be calculated by cosine distance. D is denoted as a diagonal matrix where $D = diag(\sum_j K^{ij})$, and Laplacian matrix L is denoted as $L = D - K$. The manifold regularization term can be written as (8):

$$\sum\nolimits_{i,j} K^{ij} \left\|f(x_i) - f(x_j)\right\|^2 = tr(F^T LF) \tag{8}$$

where tr denotes trace operator, $f(x_i)$ is the map function of x_i, F^i denotes the i^{th} row of F and $F^i = f(x_i)$.

III. METHODOLOGY

In this section, firstly, some important notations used in this paper are listed in TABLE 1. Then, the whole framework of our proposed RLDA is presented in detail.

TABLE 1. Important notations used in this paper and descriptions.

NOTATIONS	DESCRIPTIONS
D_r , D_s	The training and test dataset
n_r , n_s	The number of instances in training and test dataset
$x_i^{(r)}$, $x_i^{(s)}$	The i^{th} instance in training and test dataset
$y_i^{(r)}$, $y_i^{(s)}$	The label of i^{th} instance in training and test dataset
$\hat{x}_i^{(r)}$, $\hat{x}_i^{(s)}$	The RICA reconstruction of $x_i^{(r)}$, $x_i^{(s)}$
$\tilde{x}_i^{(r)}$, $\tilde{x}_i^{(s)}$	The SAMR output of $x_i^{(r)}$, $x_i^{(s)}$
m	The features number of input data
c	The label number of nodes
k	The number of nodes in the embedding layer
V , S	The eigenvectors and eigenvalues of the covariance
$\xi_i^{(s)}$, $\xi_i^{(t)}$	The hidden level input of i^{th} instance in training and test dataset
W_{1i}, W_{2i}	The encoder and decode weight matrix of the level i
b_{1i}, b_{2i}	The encoder and decode bias vectors of the level i
M	Weight vectors for manifold regularization
K^{ij}	The adjacent matrix between x_i and x_j
L	Laplacian matrices
tr()	Trace operator
λ,γ	Tuning parameters

A. OVERALL ARCHITECTURE

The proposed representation learning method with dual autoencoder is a deep neural network which is able to learn more robust and higher-level feature representations for multi-label classification. As shown in Fig. 1, the methods contains two different types of models as dual autoencoder, each autoencoder has its own strengths for extracting multiple characteristics of the input data. Specifically, there are two stages in our proposed method: (1) the algorithm of Reconstruction Independent Component Analysis (RICA) in sparse autoencoder is trained on patches for global features learning; (2) based on the results of stage (1), a stacked autoencoder with manifold regularization (SAMR) is applied to improve the quality of multi-label feature representations. After training, the softmax regression is used to predict the label set of each test instance with the learned feature representations. In the following, the details of two stages in our proposed RLDA will be given.

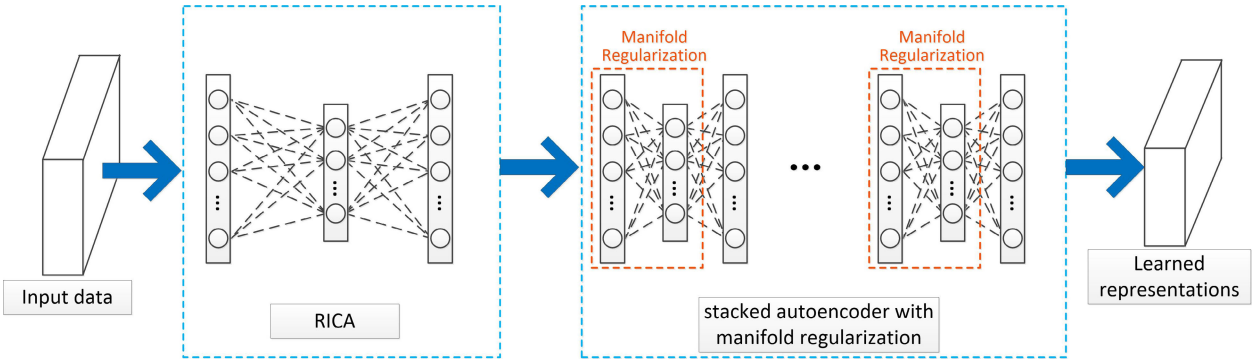

FIGURE 1. The whole framework of our proposed RLDA.

B. REPRESENTATION LEARNING VIA RECONSTRUCTION INDEPENDENT COMPONENT ANALYSIS (RICA)

The first stage of our proposed method is the RICA model, which learns the latent feature representation subspace from the original input data. Given the input data as training dataset $D_r = \{x_i^{(r)}, y_i^{(r)}\}|_{i=1}^{n_r}$ and test dataset $D_s = \{x_i^{(s)}, y_i^{(s)}\}|_{i=1}^{n_s}$, where $y_i^{(r)}, y_i^{(s)} \subseteq \{1, 2 \ldots, c\}$ are sets of relevant labels associated with input data $x_i^{(r)}, x_i^{(s)}$. The objective function of RICA is defined as (9):

$$J = \sum_{t\in\{r,s\}} \lambda \left\| Wx^{(t)} \right\|_1 + \frac{1}{m} \left\| W^T W x^{(t)} - x^{(t)} \right\|_2^2 \tag{9}$$

In our work, $f(x) = \sqrt{(Wx^{(t)})^2 + \varepsilon}$ is used to implement L1 regularization in (9) so that the reconstruction term can be scaled, and $\varepsilon = 0.1$ is a small constant value to avoid the item $(Wx^{(t)})^2 + \varepsilon$ be numerically close to zero. Therefore, the objective function can be defined as (10):

$$J = \sum_{t\in\{r,s\}} \sum_{i=1}^{n} (\lambda \left(\sqrt{(Wx_i^{(t)})^2 + \varepsilon} \right) + \frac{1}{2n} \left\| W^T W x_i^{(t)} - x_i^{(t)} \right\|_2^2) \tag{10}$$

The computational formula of the partial derivatives for L with respect to W and W^T is shown as (11):

$$\nabla_W J = \frac{1}{n} \sum_{t\in\{r,s\}} \begin{array}{l} (W(W^T W x_i^{(t)} - x_i^{(t)}) \left(x_i^{(t)}\right)^T \\ +(WX)(W^T W x_i^{(t)} - x_i^{(t)})^T) \\ +\lambda((Wx_i^{(t)})^2 + \varepsilon)^{-1/2} (Wx_i^{(t)}) \left(x_i^{(t)}\right)^T \end{array} \tag{11}$$

Based on the above partial derivatives, we feed the output feature $\hat{x}_i^{(t)} = W^T W x_i^{(t)}$ as the input into the stacked autoencoder with manifold regularization.

C. REPRESENTATION LEARNING VIA STACKED AUTOENCODER WITH MANIFOLD REGULARIZATION (SAMR)

The stacked autoencoder with manifold regularization (SAMR) has been used to improve the quality of multi-label feature representations, and generalized eigendecomposition is used to optimize the parameters of model and learn higher level feature representations. The main intuition behind the labels is to transform the multi-label task to multi-class task, which converts (instance, labels) into a set of (instance, label) where each (instance, label) contains just one label. The training set $D_r = \{x_i^{(r)}, Y_i^{(r)}\}|_{i=1}^{n_r}$ can be converted as $D_r' = \{x_i^{(r)}, y_j^{(r)} | y_j^{(r)} \in Y_i^{(r)}\}|_{i=1}^{n_r}$. The manifold regularization item can be noted as $f(\hat{x}) = M\hat{x}$, where M is the transformation weight vectors. The loss function for manifold regularization can be shown as (12):

$$Loss = \sum_{t\in\{r,s\}} \sum_{i=1}^{n} (M\hat{x}_i^{(t)} - y_i^{(t)})^2 + \gamma \|M\|^2 \tag{12}$$

where γ is tuning parameter.

In addition, as mentioned in (8), local geometry preserving term can be defined as $\sum_{i,j} K^{ij} \left\| M\hat{x}_i^{(t)} - M\hat{x}_j^{(t)} \right\|^2$, in which K^{ij} records the similarity between $\hat{x}_i^{(t)}$ and $\hat{x}_j^{(t)}$. Integrating this term into (12), the optimization problem can be expressed as (13):

$$\arg\min_M \left[\begin{array}{l} \left(\frac{1}{n_t} \sum_{t\in\{r,s\}} \sum_{i=1}^{n_t} (M\hat{x}_i^{(t)} - y_i^{(t)})^2 + \gamma \|M\|^2 \right) + \\ \frac{1}{2} \sum_{t\in\{r,s\}} \sum_{i,j} K^{ij} \left\| M\hat{x}_i^{(t)} - M\hat{x}_j^{(t)} \right\|^2 \end{array} \right] \tag{13}$$

According to the tricks mentioned in (8), (13) can be rewritten as (14):

$$\arg\min_M \left[\begin{array}{l} \left(\frac{1}{n_t} \sum_{t\in\{r,s\}} \sum_{i=1}^{n_t} (M\hat{x}_i^{(t)} - y_i^{(t)})^2 + \gamma \|M\|^2 \right) + \\ \frac{1}{2} \sum_{t\in\{r,s\}} (M(\hat{x}^{(t)}) L (\hat{x}^{(t)})^T M^T) \end{array} \right] \tag{14}$$

where L is the Laplacian matrix.

D. PREDICTION

After the feature representations are learned with the connection of dual autoencoders, the softmax regression method is

introduced to predict the multi-labels for each test instance. More specifically, following the same strategy adopted by other methods [20], the probability of one instance belongings to every label is estimated firstly, then all the probabilities of label are sorted in descending order, and the difference between two adjacent label probabilities is calculated. Finally, the labels are assigned based on the maximum difference, the labels before the max difference are considered to be the predicted labels for instances. The whole process of our proposed RLDA model is summarized in Algorithm 1.

Algorithm 1 Representation Learning With Dual Autoencoder for Multi-Label Classification (RLDA)

Require: The training dataset $D_r = \{x_i^{(r)}, y_i^{(r)}\}|_{i=1}^{n_r}$ and test dataset $D_s = \{x_i^{(s)}, y_i^{(s)}\}|_{i=1}^{n_s}$, the number of nodes in the embedding layer k, paremeters λ and γ;

Ensure: the predicted label set $Y_i^{(s)}$ for test instance $x_i^{(s)}$.

1: Initialize W randomly respectively.
2: **The stage of RICA:**
3: Compute the partial derivatives of all variables based on (11);
4: Compute $\sum_{t\in\{r,s\}} \hat{x}^{(t)} = \sum_{t\in\{r,s\}} W^T W x^{(t)}$;
5: **The stage of stacked autoencoder with manifold regularization:**
6: Compute M with (13) and (14);
7: Compute $\sum_{t\in\{r,s\}} \tilde{x}^{(t)} = \tanh\left(\sum_{t\in\{r,s\}} M\hat{x}^{(t)}\right)$;
8: Predict the label sets of test instances.

IV. EXPERIMENTS

In this section, we conduct extensive experiments to evaluate the performance of our proposed method. In the following, the details of datasets are presented firstly. Secondly the compared methods and experiment settings are introduced in detail. Then the classification results with observations of our proposed RLDA and other competing methods are given. Finally, the properties and parameter sensitivity of RLDA are analyzed with certain dataset.

A. DATASETS

The datasets including enron, medical, Corel5k and Corel16k001 are selected from Mulan [24], which is an open Java library for multi-label learning.[1] The selected datasets can evaluate the proposed method in different cases including text and image. The statistics of all the datasets are summarized in TABLE 2, where domain denotes the domains of the datasets, instances denote the number of instances, features denote the feature dimension, and labels denote the number of labels. For all the datasets, we use the default division of training and test set as the original dataset.

[1] http://mulan.sourceforge.net/index.html

TABLE 2. Dataset overview.

Data Set	domain	instances	features	labels
enron	text	1702	1001	53
medical	text	978	1449	45
Corel15k	images	5000	499	374
Corel16k001	images	6931	500	153

B. COMPARED METHODS

We compare our proposed RLDA with the following baseline methods:

- Learning multi-label scene classification (Binary relevance, short for BR) [25]. It fits multi-label data into n independent binary classifiers for each label.
- A lazy learning approach to multi-label learning (ML-KNN) [12]. It is based on the traditional k-nearest neighbor method, and it introduces maximum a posteriori principle to determine the label set for the unseen instance.
- Random k-Labelsets (RAKEL) [26]. It is based on random label space projection, and a set of Label Powerset classifiers is trained on an ensemble of k random label subsets for multi-label classification.
- Stacked Denoising Autoencoders (SDA) [27]. In this method, higher level feature extraction is firstly learned by stacked denoising autoencoder for code layer. Then, the features are combined with the labels to construct new feature space. Finally, the method of Bayesian Multinomial Regression (BMR) is adopted for classification on the new feature space.
- Manifold regularized discriminative feature selection for multi-label learning (MDFS) [28]. The manifold regularization is introduced in this method to generate the low-dimensional representations from the original input data for the local and global label correlations exploitation, then the feature selection is conducted for discriminative feature representation learning by involving ℓ_{21}-norm regularization.
- Supervised representation learning for multi-label classification (SERL) [20]. It introduces a two-encoding layer autoencoder with supervised manner to learn global feature representations for multi-label classification, and the softmax regression is utilized to incorporate label knowledge by being jointly optimized with autoencoder for improving the performance of this method.

C. EXPERIMENT SETTINGS

There are three hyper-parameters in our proposed method including tuning parameter λ, γ and the number of nodes k in the embedding layer, and we set $\lambda = 0.05$, $\gamma = 1E-03$, $k = 100$ for all datasets after cross-validations training. LIBSVM with linear kernel is used as the base classifier for all the compared methods except SDA. Moreover, for ML-KNN, the K is set as 10. For RAKEL, we set the value of ensemble and label subset k as $2c$ and 3, where c is the number of labels. For SDA, the number of nodes k in the

embedding layer is set to 100, which is consistent with our proposed method. For MDFS, the influence of parameters β and γ are set to 1 and 100 as conducted in their source code [2] [28]. For SERL, the trading-off parameters α, β and k are set as 15, 0.005, 100 respectively according to [20].

D. EXPERIMENTAL RESULTS

The ranking based evaluation metrics about RankingLoss, Coverage, MacroAUC and the classification metrics about Accuracy, F1-score and MacroF1 are adopted to compare our RLDA with other methods in a more comprehensive way. The ranking and classification results of experiments are reported in TABLE 3-10, and the best results are marked in bold. We have the following observations from experimental results:

- Among all ranking based evaluation metrics, our RLDA performs the best in enron dataset and achieves the best performance on Coverage and MacroAUC in datasets corel5k and corel16k01. Even on the metric of RankingLoss, RLDA obtains an competitive result in all the four results.
- Among all classification evaluation metrics, our proposed RLDA performs better than BR, it indicates the multi-classification methods with autoencoder outperform the standard multi classifiers method. Our proposed RLDA can extract more robust feature representations for multi-label classification. The BR method aims to learn the classifiers for overlap class, which may not be able to learn discriminative features and deteriorate performance of multi-label classification.
- RLDA outperforms ML-KNN and RAKEL, which shows the deep-based methods can learn more abstract feature representations than the shallow architecture for multi-label classification.
- Our proposed RLDA outperforms SDA, MDFS and SERL, it indicates the serially connection of two different autoencoders which captures different features is better than stacking a single autoencoder in multi-label classification.
- Neural network based methods (e.g., SDA, MDFS and SERL) deliver a relatively good result compared to problem transformation and algorithm adaptation methods (e.g., BR, ML-KNN and RAKEL) in most cases, which demonstrates the ability of feature representations learning of neural network in multi-label classification.
- Overall, in all datasets, our proposed RLDA performs best in terms of Accuracy, F1-score and MacroF1 compared to the state-of-the-art methods. The results validate the effectiveness of our proposed method.

E. COMPARISON WITH SINGLE AUTOENCODER

To verify the effectiveness of our proposed method, especially the serially connection of two different types of autoencoders, we compare the RLDA with only RICA and only SAMR.

[2]https://github.com/jiazhang-ml/MDFS

TABLE 3. The ranking results on enron.

ranking metrics	RankingLoss(%)	Coverage(%)	MacroAUC(%)
BR	29.82±0.07	58.01±0.12	57.94±0.07
ML-KNN	9.64±0.03	26.04±0.06	57.03±0.14
RAKEL	20.83±0.06	47.23±0.12	59.63±0.07
SDA	15.72±0.11	31.62±0.10	61.24±0.05
MDFS	10.01±0.14	28.87±0.08	64.18±0.03
SERL	7.46±0.05	23.42±0.10	66.39±0.06
RLDA	**7.18±0.07**	**21.07±0.11**	**70.11±0.05**

TABLE 4. The ranking results on medical.

ranking metrics	RankingLoss(%)	Coverage(%)	MacroAUC(%)
BR	9.17±0.11	22.32±0.16	48.32±0.26
ML-KNN	**4.42±0.09**	**7.14±0.17**	53.49±0.21
RAKEL	6.71±0.15	8.92±0.19	55.27±0.23
SDA	7.35±0.17	17.68±0.17	59.15±0.24
MDFS	7.51±0.08	15.74±0.23	60.43±0.27
SERL	7.12±0.10	11.78±0.21	**64.32±0.26**
RLDA	6.48±0.12	10.83±0.23	63.41±0.25

TABLE 5. The ranking results on Corel15k.

ranking metrics	RankingLoss(%)	Coverage(%)	MacroAUC(%)
BR	12.99±0.10	30.36±0.20	59.47±0.28
ML-KNN	14.31±0.19	30.13±0.28	59.79±0.45
RAKEL	19.12±0.24	43.01±0.45	57.35±0.71
SDA	14.23±0.11	32.30±0.19	62.08±0.29
MDFS	14.35±0.30	29.35±0.16	66.52±0.23
SERL	**10.63±0.14**	25.42±0.32	71.20±0.52
RLDA	11.07±0.31	**25.05±0.29**	**72.08±0.33**

TABLE 6. The ranking results on Corel16k001.

ranking metrics	RankingLoss(%)	Coverage(%)	MacroAUC(%)
BR	16.43±0.14	32.02±0.22	65.23±0.41
ML-KNN	16.65±0.27	31.85±0.19	68.29±0.19
RAKEL	21.28±0.22	40.97±0.37	60.76±0.26
SDA	17.19±0.09	32.80±0.20	60.05±0.22
MDFS	16.83±0.17	31.63±0.24	59.01±0.32
SERL	**12.69±0.18**	24.77±0.27	77.21±0.12
RLDA	13.15±0.21	**23.80±0.23**	**78.37±0.11**

TABLE 7. The classification results on enron.

classification metrics	Accuracy(%)	F1-score(%)	MacroF1(%)
BR	44.91±0.11	47.19±0.27	30.84±0.28
ML-KNN	46.24±0.21	48.63±0.25	44.91±0.11
RAKEL	53.94±0.26	57.91±0.27	49.61±0.16
SDA	60.46±0.31	66.31±0.36	58.55±0.42
MDFS	65.13±0.36	77.43±0.35	64.82±0.28
SERL	67.52±0.35	77.85±0.31	64.91±0.28
RLDA	**68.54±0.29**	**80.32±0.37**	**65.79±0.31**

TABLE 8. The classification results on medical.

classification metrics	Accuracy(%)	F1-score(%)	MacroF1(%)
BR	30.09±0.29	29.69±0.32	28.73±0.32
ML-KNN	40.34±0.49	46.94±0.48	37.41±0.61
RAKEL	40.24±0.47	45.33±0.59	37.21±0.58
SDA	61.53±0.65	65.87±0.63	53.72±0.66
MDFS	70.16±0.57	77.04±0.45	64.16±0.58
SERL	73.83±0.49	80.82±0.51	69.34±0.41
RLDA	**76.27±0.62**	**83.35±0.54**	**70.11±0.67**

The results on medical and Corel15k are listed on TABLE 7. We can see that the performance of proposed RLDA is obviously better than only RICA and only SAMR, which

TABLE 9. The classification results on Corel15k.

classification metrics	Accuracy(%)	F1-score(%)	MacroF1(%)
BR	0.29±0.03	0.54±0.05	6.80±0.53
ML-KNN	1.22±0.04	1.76±0.13	7.64±0.45
RAKEL	1.37±0.05	1.91±0.04	6.86±0.58
SDA	9.79±0.12	14.98±0.19	8.47±0.56
MDFS	11.61±0.18	17.33±0.16	7.94±0.38
SERL	13.99±0.29	20.79±0.36	8.66±0.51
RLDA	**14.11±0.32**	**21.61±0.42**	**8.71±0.53**

TABLE 10. The classification results on Corel16k001.

classification metrics	Accuracy(%)	F1-score(%)	MacroF1(%)
BR	1.36±0.18	1.91±0.25	0.94±0.19
ML-KNN	1.71±0.23	2.14±0.19	1.76±0.11
RAKEL	1.57±0.17	2.22±0.24	0.87±0.13
SDA	12.39±0.08	18.17±0.13	4.43±0.24
MDFS	13.56±0.12	20.49±0.21	5.04±0.19
SERL	16.40±0.23	23.21±0.29	5.73±0.12
RLDA	**16.55±0.25**	**23.34±0.32**	**5.81±0.14**

TABLE 11. The MacroF1 performance of RLDA, only RICA and only stacked autoencoder with manifold regularization on medical and Corel15k dataset (%).

Methods	medical Dataset	Corel15k Dataset
RLDA	70.11±0.67	8.71±0.53
only RICA	56.83±0.59	8.49±0.48
only stacked autoencoder with manifold regularization	65.21±0.52	7.99±0.41

demonstrate that the combing two different types of autoencoders can capture more powerful and abstract feature representations than a single autoencoder in multi-label classification.

F. PARAMETER SENSITIVITY

We investigate the influence of parameters in this section, including λ, γ and k in the objective (11) and (14). When we change one parameter, the rest one is fixed in the experiment. λ is set to {1E-06,5E-06,1E-04,5E-04,0.01,0.05,0.1,0.5,1.5,5}, γ is set to {1E-06,1E-05, 1E-04,1E-03,0.01,0.1,1,10} and k is set to {20,40,60,80,100, 120,140,160,180} respectively. All the results about MacroF1 on enron and medical datasets are reported in Figure 2. From Figure 2, we set $\lambda = 0.05$, $\gamma = 1E-03$ and $k = 100$ to get good and stable results.

V. RELATED WORK

Multi-label classification has been extensively researched and used in many applications such as text categorization [29], music categorization [30] and semantic classification of images [31]. The multi-label classification methods can be mainly divided into two different groups: problem transformation and algorithm adaptation methods [9].

Problem transformation methods solve the multi-label learning problem by transforming it into other well-established scenarios. For example, Binary Relevance [25] and Classifier Chains [32] transformed the multi-label learning tasks into binary classification tasks. Calibrated Label Ranking method [33] aimed to transform the multi-label learning problem into the label ranking problem based on pairwise comparison. Mencía *et al.* [34] proposed Quick Weighted Multi-label Learning (QMWL) method, which transformed a class ranking into a bipartite prediction by introducing an artificial thresholding class with the QWeighted voting for reducing computational costs. Random k-labelsets method [35] learned an ensemble of multi-label classifiers based on the dividing of k random label subsets, which improved computational efficiency and predictive performance compared to the traditional label powerset methods.

Algorithm adaptation methods adapt the existing single label classification algorithms to multi-label data [9]. For example, Rastin *et al.* [36] proposed a prototype weighting method to adapt the distance measure based on the ML-kNN method [12], the prototype weights were adjusted by gradient ascent method in order to maximize the objective function as macro-F1 measure. Kouchaki *et al.* [37] designed the multi-label random forest (MLRF) models for treating tuberculosis resistance classification and mutation ranking in medical. Wu *et al.* [38] jointed Ranking support vector machine and Binary Relevance with robust Low-rank learning (RBRL), which enjoyed the advantages and tackled the disadvantages of Rank-SVM and BR. Xuan *et al.* [39] developed a Bayesian nonparametric model for multi-label learning, which can learn both low-dimensional labels and instances embedding without the fixed of dimensions number. Zhang *et al.* [40] proposed a fully associative ensemble learning method for hierarchical multi-label classification, which built a multi-variable regression model between the global and local predictions of all the nodes.

Recently, the representation learning methods have achieved the encouraging results in multi-label classification. For example, Huang *et al.* [41] proposed to learn label-specific data representation for each class label in a sparse stacking way, which exploited both second-order and high-order label correlations for multi-label classification. Zhang *et al.* [42] proposed a hierarchical and transparent representation learning method to express the semantic information for accurate paper-reviewer recommendation as multi-label classification. Ye *et al.* [43] introduced a dynamic graph convolutional network to project raw input into category-aware representations with semantic attention module, and the final category representations are utilized for multi-label image recognition. Gong *et al.* [44] proposed a hierarchical graph transformer method for multi-label text classification, a multi-layer transformer structure and the hierarchical relationship of the labels are used for feature representations learning in different level. On the other hand, the autoencoder based methods have attracted much attention for the superiority of powerful representations learning ability and fast convergence speed. For example, Huang *et al.* [20] designed the two encoding layers auto-encoder model for multi-label learning, and the knowledge is shared by softmax regression for the performance improvement. Law and Ghosh [45] introduced a stacked autoencoder for a discriminating and reduced input representation learning of the

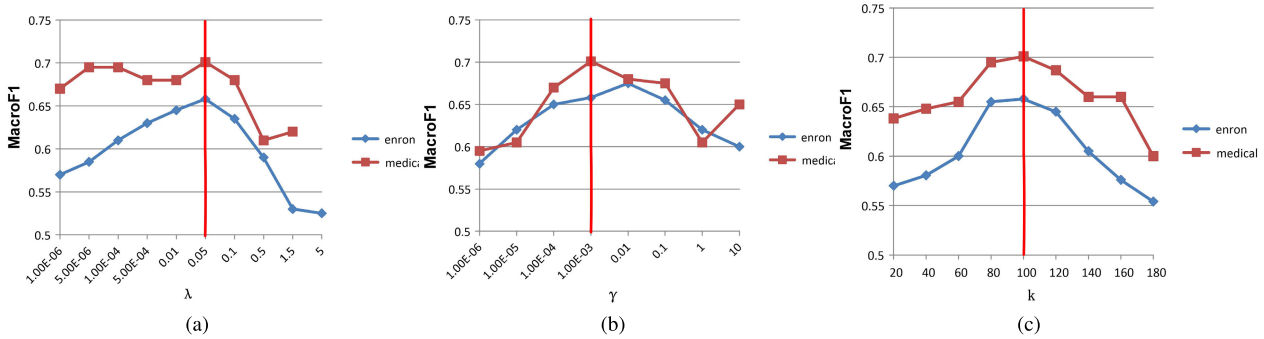

FIGURE 2. Parameter Influence of MacroF1 on enron and medical dataset.

multi-label data. Cheng *et al.* [46] proposed a kernel extreme learning machine autoencoder for the associations learning between the features in the input space. Due to there has no iterative process in the extreme learning machine autoencoder, this method can reduce the computational complexity and improve the classification performance.

VI. CONCLUSION

In this paper, we propose a representation learning method with dual autoencoder (RLDA), which learns richer feature representations by the serially connection of two different types of autoencoders for multi-label classification. In our proposed method, the method of Reconstruction Independent Component Analysis (RICA) is introduced in the first stage for robust global feature representation learning. Furthermore, the stacked autoencoder with manifold regularization (SAMR) is applied in the second stage to extract more powerful feature representations. Extensive experiments conducted on four real-world datasets demonstrate the effectiveness of our proposed method compared with other competing methods.

This study points out the effectiveness of the proposed RLDA on multi-label classification. However, how to determine the types, the numbers and the connection methods of autoencoders is still a challenge. In our future work, we will try to add other different types of autoencodes for discovering more characteristics of data, and multiple types of autoencoders will be connected serially and parallelly to extract more abstract feature representations for multi-label classification.

REFERENCES

[1] J. Huang, L. Xu, K. Qian, J. Wang, and K. Yamanishi, "Multi-label learning with missing and completely unobserved labels," *Data Mining Knowl. Discovery*, vol. 35, no. 3, pp. 1061–1086, May 2021.
[2] G. Tsoumakas and I. Katakis, "Multi-label classification: An overview," *Int. J. Data Warehousing Mining*, vol. 3, no. 3, pp. 1–13, 2007.
[3] P. Prajapati, A. Thakkar, and A. Ganatra, "A survey and current research challenges in multi-label classification methods," *Int. J. Soft Comput. Eng.*, vol. 2, no. 1, pp. 248–252, 2012.
[4] X. Zheng, P. Li, Z. Chu, and X. Hu, "A survey on multi-label data stream classification," *IEEE Access*, vol. 8, pp. 1249–1275, 2020.
[5] P. Yang, X. Sun, W. Li, S. Ma, W. Wu, and H. Wang, "SGM: Sequence generation model for multi-label classification," 2018, *arXiv:1806.04822*. [Online]. Available: http://arxiv.org/abs/1806.04822
[6] T. Durand, N. Mehrasa, and G. Mori, "Learning a deep ConvNet for multi-label classification with partial labels," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 647–657.
[7] T. Baumel, J. Nassour-Kassis, R. Cohen, M. Elhadad, and N. Elhadad, "Multi-label classification of patient notes a case study on icd code assignment," 2017, *arXiv:1709.09587*. [Online]. Available: https://arxiv.org/abs/1709.09587
[8] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.
[9] M.-L. Zhang and Z.-H. Zhou, "A review on multi-label learning algorithms," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 8, pp. 1819–1837, Aug. 2014.
[10] J. Read, B. Pfahringer, G. Holmes, and E. Frank, "Classifier chains for multi-label classification," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases*, 2009, pp. 254–269.
[11] J. Read, B. Pfahringer, G. Holmes, and E. Frank, "Classifier chains for multi-label classification," *Mach. Learn.*, vol. 85, no. 3, p. 333, Dec. 2011.
[12] M.-L. Zhang and Z.-H. Zhou, "ML-KNN: A lazy learning approach to multi-label learning," *Pattern Recognit.*, vol. 40, no. 7, pp. 2038–2048, Jul. 2007.
[13] M.-L. Zhang, Y.-K. Li, X.-Y. Liu, and X. Geng, "Binary relevance for multi-label learning: An overview," *Frontiers Comput. Sci.*, vol. 12, no. 2, pp. 191–202, 2018.
[14] M. A. Tahir, J. Kittler, and A. Bouridane, "Multi-label classification using stacked spectral kernel discriminant analysis," *Neurocomputing*, vol. 171, pp. 127–137, Jan. 2016.
[15] A. Alali and M. Kubat, "Prudent: A pruned and confident stacking approach for multi-label classification," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 9, pp. 2480–2493, Sep. 2015.
[16] J. Wang, Y. Yang, J. Mao, Z. Huang, C. Huang, and W. Xu, "CNN-RNN: A unified framework for multi-label image classification," in *IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 2285–2294.
[17] Y. Wang, D. He, F. Li, X. Long, Z. Zhou, J. Ma, and S. Wen, "Multi-label classification with label graph superimposing," *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 07, pp. 12265–12272, Apr. 2020.
[18] S. Yang, Y. Zhang, Y. Zhu, P. Li, and X. Hu, "Representation learning via serial autoencoders for domain adaptation," *Neurocomputing.*, vol. 351, pp. 1–9, Jul. 2019.
[19] C.-K. Yeh, W.-C. Wu, W.-J. Ko, and Y.-C. F. Wang, "Learning deep latent space for multi-label classification," in *Proc. AAAI Conf. Artif. Intell.*, Feb. 2017, vol. 31, no. 1, pp. 2838–2844.
[20] M. Huang, F. Zhuang, X. Zhang, X. Ao, Z. Niu, M.-L. Zhang, and Q. He, "Supervised representation learning for multi-label classification," *Mach. Learn.*, vol. 108, no. 5, pp. 747–763, May 2019.
[21] Y. Bengio, "Learning deep architectures for AI," *Found. Trends Mach. Learn.*, vol. 2, no. 1, pp. 1–127, 2009.
[22] Y. Zhu, X. G. Hu, Y. Zhang, and P. Li, "Transfer learning with stacked reconstruction independent component analysis," *Knowl.-Based Syst.*, vol. 152, pp. 100–106, Jul. 2018.
[23] N. Pitelis, C. Russell, and L. Agapito, "Learning a manifold as an atlas," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 1642–1649.
[24] G. Tsoumakas, E. Spyromitros-Xioufis, J. Vilcek, and I. Vlahavas, "MULAN: A java library for multi-label learning," *J. Mach. Learn. Res.*, vol. 12, pp. 2411–2414, Jun. 2011.

[25] M. R. Boutell, J. Luo, X. Shen, and C. M. Brown, "Learning multi-label scene classification," *Pattern Recognit.*, vol. 37, no. 9, pp. 1757–1771, Sep. 2004.

[26] G. Tsoumakas and I. Vlahavas, "Random *k*-labelsets: An ensemble method for multilabel classification," in *Proc. Eur. Conf. Mach. Learn.*, 2007, pp. 406–417.

[27] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, P.-A. Manzagol, and L. Bottou, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *J. Mach. Learn. Res.*, vol. 11, no. 12, pp. 3371–3408, 2010.

[28] J. Zhang, Z. Luo, C. Li, C. Zhou, and S. Li, "Manifold regularized discriminative feature selection for multi-label learning," *Pattern Recognit.*, vol. 95, pp. 136–150, Nov. 2019.

[29] G. Chen, D. Ye, Z. Xing, J. Chen, and E. Cambria, "Ensemble application of convolutional and recurrent neural networks for multi-label text categorization," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, 2017, pp. 2377–2383.

[30] S. Oramas, O. Nieto, F. Barbieri, and X. Serra, "Multi-label music genre classification from audio, text, and images using deep features," 2017, *arXiv:1707.04916*. [Online]. Available: http://arxiv.org/abs/1707.04916

[31] W. Ge, S. Yang, and Y. Yu, "Multi-evidence filtering and fusion for multi-label classification, object detection and semantic segmentation based on weakly supervised learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1277–1286.

[32] L. E. Sucar, C. Bielza, E. F. Morales, P. Hernandez-Leal, J. H. Zaragoza, and P. Larrañaga, "Multi-label classification with Bayesian network-based chain classifiers," *Pattern Recognit. Lett.*, vol. 41, pp. 14–22, May 2014.

[33] G. Madjarov, D. Gjorgjevikj, and S. Džeroski, "Two stage architecture for multi-label learning," *Pattern Recognit.*, vol. 45, no. 3, pp. 1019–1034, Mar. 2012.

[34] E. L. Mencía, S.-H. Park, and J. Fürnkranz, "Efficient voting prediction for pairwise multilabel classification," *Neurocomputing*, vol. 73, no. 7, pp. 1164–1176, Mar. 2010.

[35] G. Tsoumakas, I. Katakis, and L. Vlahavas, "Random *k*-labelsets for multilabel classification," *IEEE Trans. Knowl. Data Eng.*, vol. 23, no. 7, pp. 1079–1089, Jul. 2011.

[36] N. Rastin, M. Z. Jahromi, and M. Taheri, "A generalized weighted distance k-Nearest neighbor for multi-label problems," *Pattern Recognit.*, vol. 114, Jun. 2021, Art. no. 107526.

[37] S. Kouchaki, Y. Yang, A. Lachapelle, T. M. Walker, A. S. Walker, T. E. A. Peto, D. W. Crook, D. A. Clifton, and C. Consortium, "Multi-label random forest model for tuberculosis drug resistance classification and mutation ranking," *Frontiers Microbiol.*, vol. 11, p. 667, Apr. 2020.

[38] G. Wu, R. Zheng, Y. Tian, and D. Liu, "Joint ranking SVM and binary relevance with robust low-rank learning for multi-label classification," *Neural Netw.*, vol. 122, pp. 24–39, Feb. 2020.

[39] J. Xuan, J. Lu, G. Zhang, R. Y. D. Xu, and X. Luo, "A Bayesian nonparametric model for multi-label learning," *Mach. Learn.*, vol. 106, no. 11, pp. 1787–1815, Nov. 2017.

[40] L. Zhang, S. K. Shah, and I. A. Kakadiaris, "Hierarchical multi-label classification using fully associative ensemble learning," *Pattern Recognit.*, vol. 70, pp. 89–103, Oct. 2017.

[41] J. Huang, G. Li, Q. Huang, and X. Wu, "Learning label-specific features and class-dependent labels for multi-label classification," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 12, pp. 3309–3323, Dec. 2016.

[42] D. Zhang, S. Zhao, Z. Duan, J. Chen, Y. Zhang, and J. Tang, "A multi-label classification method using a hierarchical and transparent representation for paper-reviewer recommendation," *ACM Trans. Inf. Syst.*, vol. 38, no. 1, pp. 1–20, Feb. 2020.

[43] J. Ye, J. He, X. Peng, W. Wu, and Y. Qiao, "Attention-driven dynamic graph convolutional network for multi-label image recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 649–665.

[44] J. Gong, H. Ma, Z. Teng, Q. Teng, H. Zhang, L. Du, S. Chen, M. Z. A. Bhuiyan, J. Li, and M. Liu, "Hierarchical graph transformer-based deep learning model for large-scale multi-label text classification," *IEEE Access*, vol. 8, pp. 30885–30896, 2020.

[45] A. Law and A. Ghosh, "Multi-label classification using a cascade of stacked autoencoder and extreme learning machines," *Neurocomputing*, vol. 358, pp. 222–234, Sep. 2019.

[46] Y. Cheng, D. Zhao, Y. Wang, and G. Pei, "Multi-label learning with kernel extreme learning machine autoencoder," *Knowl.-Based Syst.*, vol. 178, pp. 1–10, Aug. 2019.

YI ZHU received the B.S. degree from Anhui University, the M.S. degree from the University of Science and Technology of China, and the Ph.D. degree from the Hefei University of Technology. He is currently an Assistant Professor with the School of Information Engineering, Yangzhou University, China. His research interests include data mining and knowledge engineering.

YANG YANG received the B.S. degree from Nanjing Tech University. He is currently pursuing the degree with the School of Information Engineering, Yangzhou University, China. His research interests include knowledge engineering and recommendation systems.

YUN LI received the M.Eng. degree in computer science and technology from the Hefei University of Technology, China, in 1991, and the Ph.D. degree in control theory and control engineering from Shanghai University, China, in 2005. He is currently a Professor with the School of Information Engineering, Yangzhou University, China. He has published more than 70 scientific articles. His research interests include data mining and cloud computing.

JIPENG QIANG received the B.S. degree in computer science from Hefei University, China, in 2010, and the M.S. and Ph.D. degrees in computer science from the Hefei University of Technology, China, in 2013 and 2016, respectively. From October 2014 to June 2016, he was a Visiting Ph.D. Student with the University of Massachusetts Boston. He is currently an Assistant Professor with the School of Information Engineering, Yangzhou University, China. His research interests include data mining and natural language processing.

YUNHAO YUAN received the Ph.D. degree in pattern recognition and intelligence system from the Nanjing University of Science and Technology (NUST), China, in 2013. He is currently an Associate Professor with the School of Information Engineering, Yangzhou University. He is the author or coauthor of more than 60 scientific articles. His research interests include pattern recognition, machine learning, multimedia search, and information fusion. He received two National Scholarships from the Ministry of Education, China, and an Outstanding Ph.D. Thesis Award from NUST.

RUNMEI ZHANG received the B.S. degree from Huaibei Normal University, and the M.S. and Ph.D. degrees from the Hefei University of Technology. She is currently a Professor with the School of Mechanical and Electrical Engineering, Anhui Jianzhu University, China. Her research interests include data mining, complex system modeling, and architecture digitization.