Generative Multi-Task Learning for Text Classification

WEI ZHAO¹², HUI GAO¹, SHUHUI CHEN¹, AND NAN WANG³
¹College of Computer Science, National University of Defense Technology, Changsha 410073, China
²Hunan Police Academy, Changsha 410138, China
³College of Intelligence Science and Technology, National University of Defense Technology, Changsha 410073, China
Corresponding author: Hui Gao (gaohui_baixiang@163.com)

ABSTRACT Multi-task learning leverages potential correlations among related tasks to extract common features and yield performance gains. In this paper, a generative multi-task learning (MTL) approach for text classification and categorization is proposed, which is composed of a shared encoder, a multi-label classification decoder and a hierarchical categorization decoder. In the two decoders, a label-order-independent multi-label classification loss function and a hierarchical structure mask matrix are introduced. Experiments conducted on the real-world public security dataset show that the proposed approach has obvious advantages over the baseline approaches and can enhance the semantic association between the results of the classification and categorization tasks.

INDEX TERMS Generative model, hierarchical classification, multi-label classification, multi-task learning.

I. INTRODUCTION

Text classification or categorization is the process of automatically tagging a textual document with most relevant labels or categories [1]. In real-world applications, many texts to be classified are ambiguous, that is, one instance could belong to multiple labels. The purpose of multi-label text classification is to establish a one-to-many association for certain text to label sets, which is more in line with the laws of the objective world [2]. For example, in the public security scenario, the case acceptance process will generate a large number of police case texts, which generally need to be tagged to facilitate information retrieval, case stringing and connection analysis and the investigation of the subsequent cases. As shown in Figure 1(a), the given police case text is tagged as “fraud” and “foreigner”. Besides, in order to efficiently organize and manage the massive text data, it is common practice to organize the text into hierarchical categories according to a concept or a theme in some scenario (as shown in Figure 1(b)), and each text has a specified category path under the hierarchical categories. For example, the category path corresponding to the police case text in Figure 1(a) is “fraud/contact-fraud/fraud in the name of borrowing”.

The existing research on multi-label text classification and hierarchical classification mainly focus on modeling one-to-many mapping between samples and labels or categories. Each sample will be associated with one or more labels or categories, which can be considered as a sequence of labels or categories. Therefore, the multi-label classification
and hierarchical classification approaches can be abstracted into a Seq2Seq sequence generation model [3], [4]. Since strong semantic relationships and father-child or sibling relationships often exist among labels or categories, making full use of the semantic relationship between them can further improve the classification quality. The Seq2Seq model takes into account the information of the preamble labels and categories for the current labels and categories generation, so it can effectively utilize the semantic relationship between labels and categories.

To perform multi-label classification and hierarchical classification on a text, the conventional way is to train a multi-label classification model and a hierarchical classification model separately and then independent predictions are generated [5], [6]. However, it may result in lack of necessary semantic correlation between the predicted results of the two tasks. For example, the multi-label classification result is “fraud/contact-fraud” for the police case in Figure 1(a). If the hierarchical classification result is mispredicted as “snatch & robbery”, then it will cause great confusion. To solve this problem, this paper integrates the tasks of multi-label classification and hierarchical classification into one multi-task learning framework, which improves the classification performance through the eavesdropping mechanism [7] of the two tasks. In addition, the semantic association between the two tasks is established, which avoids the contradiction between the predicted labels and categories of the same text.

In summary, a generative multi-task text classification approach is proposed in this paper. The approach is based on the Seq2Seq generation model, which is composed of a shared encoder, a multi-label classification decoder and a hierarchical classification decoder. The main contributions of this paper are as follows:

- Multi-label classification and hierarchical classification are jointly modeled into one Seq2Seq model, which makes the description and solution of the problem more concise and consistent. Based on the traditional Seq2Seq model, a sequence-independent loss function and a hierarchical structure mask matrix are proposed for multi-label classification and hierarchical classification respectively. Experiments show that these measures have an obvious positive effect on the classification performance.
- An integrated framework based on multi-task learning (MTL) is applied to train the Seq2Seq model. It can establish semantic associations between tasks by sharing text features and alternative training of related decoders. Hence the performance of both tasks is improved, and the prediction results of the two tasks are more semantically consistent.

II. RELATED RESEARCH

A. MULTI-TASK LEARNING

MTL applies joint learning instead of the traditional single-task independent learning, which extracts the correlation between multiple tasks and common features, such as shared sub-vectors, shared sub-spaces, etc. [8]. Each task can get additional useful information and may achieve better results than single-task learning [9]. In natural language processing applications, MTL can not only use the correlation of related tasks to alleviate the problem of under-fitting due to the small size of the training text, but also improve the generalization ability of the model. Previous work includes proposing an adversarial MTL framework to improve the ability to identify named entities in the case of small samples, applying MTL model to cross-domain sentiment analysis and cross-language text classification [10]–[13]. In this paper, MTL is adopted to enhance the semantic associations between the predicted results of the multi-label classification and hierarchical classification tasks, which is rarely involved in the existing literature.

B. MULTI-LABEL TEXT CLASSIFICATION

Multi-label classification focuses on how to model the one-to-many mapping between samples and labels, and how to make full use of the semantic relationship between labels to improve the classification accuracy [10], [11]. The existing methods can be roughly divided into two categories [12]. One aims to improve the performance by optimizing the classification algorithms, such as the ML-KNN proposed in reference [13] and Rank-SVM proposed in reference [14]. The other aims to solve the problem by converting it into a different but equivalent problem, such as Binary Relevance [15] and Label Powerset [16]. With the development of deep learning techniques, CNN-based multi-label classification models [17], [18] have been proposed to better extract text features, in which a sigmoid activation function is added to the final output layer to output the probability of each label. However, the above methods ignore the association between labels. In recent years, there has been work on making use of the association relationship between labels, which applies the sequence generation method into multi-label classification [2], [19]–[22].

C. HIERARCHICAL TEXT CLASSIFICATION

At present, hierarchical classification methods are divided into three categories [23]. The first one ignores the hierarchical relationship and directly models the problem into flat multi-class classification problem. The second category is the local classification, which is also called the top-down classification [24], [25]. In these approaches, the class of the root node of the hierarchical tree is first predicted, and then the leaf nodes are predicted in sequence based on the tree topology [24]. However, such approaches may cause the error propagation problem and require multiple classifiers. The third category is global classification, which builds a model to learn the hierarchical relationship between categories and to classify hierarchical categories [26]. In these methods, both local and global information are utilized, but the disadvantage is that the models become much more complicated and are prone to overfitting.
corresponding labels and categories are \( L \) and \( M \), respectively. Suppose the text to be classified contains \( n \) Chinese words, the corresponding output sequence \( \ldots, h_1, h_2, \ldots, h_n \) is obtained by the encoder. The feature vector at the \( i \)-th time step of the decoder is \( \hat{y}_i \) and the corresponding numbers of labels and categories are \( L_M \) and \( L_H \) respectively. See Figure 2. The first part is a shared encoder, which encodes the input text for the subsequent decoders. The second part is a decoder for multi-label classification. The third part is a decoder for hierarchical classification. All the encoder and the decoders are implemented by Long Short-Term Memory (LSTM) network, and each decoder has its own attention mechanism.

B. SEQ2SEQ MODEL
1) THE SHARED ENCODER
The shared encoder extracts the semantic representation of the text, which is composed of a bidirectional LSTM (Bi-LSTM). Since Bi-LSTM obtains historical and future information of sequence data, it can better handle the semantic information of long sentences. For a sequence containing \( n \) Chinese words, the corresponding output sequence \( h_1, h_2, \ldots, h_n \) is obtained by the encoder. The feature vector at the \( i \)-th time step can be expressed as follows:

\[
\hat{h}_i = \tilde{h}_i \oplus h_i = Bi-LSTM(x_i, \tilde{h}_{i-1}, h_{i-1})
\]  

(1)

where \( \hat{h}_i \) and \( \tilde{h}_i \) correspond to the hidden state of the LSTM in the two directions before and after the \( i \)-th time, and \( \oplus \) represents the connection of two vectors.

2) ATTENTION MECHANISM
In the Seq2Seq structure without attention mechanism, the encoder encodes all input sequences into a unified semantic feature \( c \) which will be decoded afterward. Therefore, \( c \) must contain all the information in the original sequence, thus the length of \( c \) becomes the bottleneck that limits the performance of the model. The attention mechanism alleviates this problem by using a different \( c_t \) at each decoding time step, which is defined as follows:

\[
c_t = \sum_{i=1}^{n} a_{ti} h_i
\]  

(2)

where \( a_{ti} \) is the learning weight that represents the correlation of \( h_i \) and semantic feature at the \( t \)-th time step.

3) DECODER
The hidden state at the \( t \)-th time step of the decoder is \( S_t \), which is defined as follows:

\[
s_t = LSTM(S_{t-1}, g(y_{t-1}) \oplus c_{t-1})
\]  

(3)

where \( y_t \) represents the output probability distribution over the entire label space \( L \) at time \( t \), the output vector \( g(y_{t-1}) \) represents the embedding vector of label \( l \) that corresponding to the maximum probability in \( y_{t-1} \) at time \( t-1 \). The \( y_t \) is further defined as follows:

\[
o_t = W_{df}(W_ds_t + V_df_t)
\]  

(4)

\[
y_t = \text{soft max}(o_t)
\]  

(5)

where \( W_d, W_s, V_d \) are parameters to be trained.

The loss function in the training process is calculated by the average value of the cross entropy of the output of the decoder and the true value of the corresponding label at each time step, as shown in Equation (6).

\[
\text{loss}(Y, \hat{Y}) = \frac{\sum_{t=1}^{m_M} \text{cross}(y_t, \hat{y}_t)}{m_M}
\]  

(6)

C. MULTI-LABEL CLASSIFICATION LOSS FUNCTION
The traditional Seq2Seq loss function encourages the predicted label order to be consistent with the true label order; otherwise, a large loss value will be generated. However, for multi-label classification problems, it is not necessary to maintain a strict label order in the generated sequence. For example, in the police case of Figure 1(a), the label is considered to be correct when it is predicted to be "fraud, foreigner" or "foreigner, fraud". According to this feature, a multi-label classification loss function MLCLoss is proposed, which is insensitive to the order of the labels:

\[
\text{MLLoss}(Y, \hat{Y}) = \sum_{t=1}^{m_M} \min_{t'=1,2,...,m_M} \left[ \text{cross}(y_{t'}, \hat{y}_{t'}) + \text{cross}(y_{t'}, \hat{y}_{t'}) \right]
\]  

(7)

For the predicted output at time step \( t \), the cross entropy of the predicted output and all labels of the target set is
calculated, and the minimum value of the entropy is selected to calculate the final loss function. Meanwhile, for the labels in the target set at time step $t$, the cross entropy of the labels and all the predicted labels is calculated, and the minimum value of the entropy is involved in the calculation of the final loss function. In order to restrict the length of the predicted label sequence, the loss function value of the end positions in the predicted label sequence and the target label sequence are calculated by Equation (6). Finally, the two part of the loss function values are summed up as the final loss.

**D. HIERARCHICAL CLASSIFICATION DECODER**

The hierarchical classification can utilize the rich semantic information in the hierarchical structure of labels to improve the classification performance. However, the traditional sequence generation model cannot impose sufficient constraints on whether the predicted results conform to the predefined hierarchical categories. To deal with this problem, this paper enhances the hierarchy constraint by adding a hierarchical structure mask matrix $I_{l-1}$ before the decoder, which is defined as follow:

$$y_t = \text{soft} \max (o_t + I_{l-1})$$

where the hierarchical structure mask matrix $I_{l-1} \in R^{H \times H}$ is defined as follows:

$$(I_{l-1})_{lj} = \begin{cases} 0, & \text{When } l_j \text{ is the son node of the generated } l_{l-1} \\
-\infty, & \text{at a previous moment in the classification tree.} \\
\end{cases}$$

Through the introduction of the mask matrix, the search space of the current category is constrained to the sub-layer of the previous generated category. As a result, the output label sequence conforms to the predefined hierarchical categories.

**IV. EXPERIMENT**

**A. DATA SET**

The police case texts generated during the case acceptance process are used as the experimental data. In cooperation with a city’s public security system, 127,142 police case texts from January to August 2018 were collected. Each text was marked by a professional police officer with labels and categories. The size of the label set is 14. The category set is a tree structure with a depth of 7 layers, and the number of labels corresponding to layer 1 to layer 7 is 20, 106, 55, 132, 144, 210 and 62, respectively. Each text may correspond to a full-depth category or a partial depth category for hierarchical classification [27], that is, the classification of the text can be assigned to a non-leaf node, instead of being assigned to a leaf node of the category set. This is accomplished by adding an EOS (End of Sequence) identifier to the category set, as commonly used in Seq2Seq models. When the output $l_t$ of the hierarchical classification decoder at time step $t$ is EOS, the training cycle or prediction ends even if $l_{t-1}$ corresponds to a non-leaf node. Hence the classification approach is qualified for handling the partial depth category.

In this paper, the data set is divided into training set, validation set, and test set according to a ratio of 8:1:1, as shown in Table 1. The Total Samples denotes the total number of samples in data set. Words/Sample is the average number of Chinese words per sample. Label/Sample denotes the average number of labels per sample in multi-label classification and Levels/Sample denotes the average levels per sample in hierarchical classification.

**B. METRICS**

Performance evaluation in multi-label classification and hierarchical classification is much more complicated than traditional single-label setting, as each example can be associated with multiple labels simultaneously. Since hierarchical classification can be regarded as a special case of multi-label classification, the example-based metrics are introduced for performance evaluation, which are defined as the Accuracy, Precision and $F$ [15]. Since the public security application would need higher precision on the predicted results, the metric Precision is modified into $Precision_{Full-Match}(Precision_f$ for short), which reflects the proportion of the completely correct results predicted in the test set to the total samples and is defined as follows:

$$Precision_f = \frac{1}{N} \sum_{k=1}^{N} 1_{L_k = \hat{L}_k}$$

where $L_k$ represents the true label and category of the $k$-th sample, and $\hat{L}_k$ represents the predicted label and category of the $k$-th sample, $N$ is the number of samples in the training set.

Besides, in this paper a common form of $F$, which is also called the Micro-$F1$ is adopted with the parameter $\beta$ of $F$ set to 1.

**C. BASELINE**

For multi-label classification and hierarchical classification problems, the following baselines are selected to compare with GMLC.

CNN [28] multi-label classification: CNN is used to capture text features, its last layer uses sigmoid to output the probability of each label, and marks the samples by labels whose probability is greater than a certain threshold. This approach does not consider the relationship between labels.

SGM [22] multi-label classification: The Seq2Seq model with attention mechanism is used for multi-label text classification. Besides, the concept of “global embedding” is
TABLE 2. Police case text multi-label classification results.

| Model          | Training Set | Validation Set | Test Set |
|----------------|--------------|----------------|---------|
|                | Accuracy     | Precision_3^F  | Micro-F1 | Accuracy | Precision_3^F | Micro-F1 | Accuracy | Precision_3^F | Micro-F1 |
| CNN            | 0.902        | 0.873          | 0.869    | 0.902    | 0.873          | 0.869    | 0.902    | 0.873          | 0.869    |
| SGM            | 0.954        | 0.941          | 0.924    | 0.954    | 0.941          | 0.924    | 0.954    | 0.941          | 0.924    |
| GMLC_M         | 0.954        | 0.933          | 0.917    | 0.954    | 0.933          | 0.917    | 0.954    | 0.933          | 0.917    |
| GMLC(common loss) | 0.952    | 0.956          | 0.916    | 0.952    | 0.956          | 0.916    | 0.952    | 0.956          | 0.916    |
| GMLC           | 0.959        | 0.940          | 0.923    | 0.959    | 0.940          | 0.923    | 0.959    | 0.940          | 0.923    |

introduced, in which labels except the one with the maximum probability at the previous time step are used together for the prediction of the labels at the current time step.

CNN_Flat hierarchical classification: The case categories are tiled into one layer, and the CNN model [28] is used for text classification.

Top-Down (CNN) hierarchical classification: For the tree hierarchy, a CNN text classifier [28] is trained for each non-leaf node to predict the subclass.

Besides, in order to verify the effectiveness of the proposed method, the following models are set up for comparative experiments:

GMLC_M: A sequence generation model used only for multi-label classification. The structure and hyperparameters of the encoder and multi-label classification decoder are consistent with the GMLC model.

GMLC_H: A sequence generation model used only for hierarchical classification. The structure and hyperparameters of the encoder and hierarchical classification decoder are consistent with the GMLC model.

GMLC(common loss): The multi-label classification task in the GMLC model uses the traditional cross-entropy loss function.

GMLC(without mask): The hierarchical classification decoder in the GMLC model does not adopt the hierarchical structure mask matrix.

D. MODEL TRAINING

The hyperparameters used in the experiments are set as follows. The word embedding dimension is 128, the sentence length is 600, the encoder is bidirectional LSTM, and the batch size is 32; the optimizer uses the Adam algorithm, the learning rate is 0.001, the betas is (0.9, 0.999), and the eps is 1e-08, weight_decay is 0, dropout is 0.5, num_epochs is 4, and teacher_forcing_ratio is 0.5. The forecast uses the Beam search strategy with the Beam Width set to 3. In this paper, multi-task alternative training is adopted to perform multi-task learning by alternately calling the optimizer of each task.

E. EXPERIMENTAL RESULTS AND ANALYSIS

Tables II-III show the multi-label classification experimental results of the performance of GMLC and baseline approaches on the police case text dataset. It can be seen that the performance of GMLC is equivalent with SGM (the Precision_3^F and Micro-F1 metrics of GMLC are 0.1 percentages lower than that of SGM, but the Accuracy of GMLC is 0.5 percentages higher than that of SGM on the test set). Moreover, the results show that GMLC has better performance after the introduction of the MLCLoss in Section 3.3.

According to the case text hierarchical classification results (As shown in Table 3), the GMLC outperforms other approaches in all the metrics (its metrics are 2 to 7 percentages higher than Top-Down on the test set). Therefore, the classification performance is proved to be improved by the introduction of the hierarchical structure mask matrix.

1) MULTITASK PREDICTION RESULTS CORRELATION ANALYSIS

Experiments have shown that under the same network structure and hyperparameters settings, the GMLC model outperforms the single multi-label classification (GMLC_M) and hierarchical classification (GMLC_H). Therefore, the proposed multitask learning framework can improve the learning performance of two tasks at the same time. To further verify the effect of the semantic correlation in the multitask learning framework, the prediction results of the single-task models are combined for comparison with GMLC, as shown in Table 4. The classification results are evaluated by the proportion of the predicted results which are complete correct (the two tasks are both correct, marked by Full-Match2) on the test set. It can be seen that although the SGM and GMLC_H achieve relative high values of metrics respectively, the combined results are still worse than GMLC. For
example, for the police case text in Figure 1(a), the police case labels and the categories predicted by SGM+GMLC_H are “Double grab” and “Fraud class/Contact fraud/Fraud in the name of borrowing/Lending mobile phone/Dating type”, while GMLC successfully predicts the correct results “Fraud” and “Fraud class / Contact fraud /Fraud in the name of borrowing / Lending mobile phone / Dating type”.

Besides, the effect of the Attention mechanism on multi-task learning is also analyzed. Table 5 gives a visualization of the Attention weights on the words of GMLC, GMLC_H and GMLC_M for the same piece of police case text. The true values of the labels and categories corresponding to this paragraph are “Involving guns, Involving evil” and “Involving evil/Involving guns”. The GMLC_M assigns higher attention weights to “hit”, while the attention weights assigned to “steel balls” and “air guns” are smaller, which cause the GMLC_M to predict an incorrect label as “harm”. In contrast, the GMLC assigns higher attention weights to the words “steel ball” and “air gun” when predicting the police case labels, which is consistent with the prediction of the categories. It shows that the Attention mechanism in GMLC helps improving the sub-task learning effect and enhancing the correlation of predicted results.

Red and pink highlights indicate assigned attention weights $a > 10^{-1}$ and $10^{-1} > a > 10^{-2}$.

2) MULTI-LABEL CLASSIFICATION LOSS FUNCTION

The MLCLoss function proposed in this paper relaxes the order of the predicted labels in the multi-label classification task, which shows a slightly improvement of the classification performance (Table 2). To further evaluate the effects of MLCLoss function on the prediction performance and the converge speed, the perplexity function [3], [29] which serves as a common measure of language model is introduced. The perplexity is defined as the exponent of the average negative log likelihood per word. It expresses the value “if we randomly picked words from the probability distribution calculated by the language model at each time step, on average how many words would it have to pick to get the correct one?” The lower the perplexity is, the better the language model becomes [29]. Figure 3 shows the evolution of perplexity of GMTC and GMTC (common loss) during training. It can be seen that GMTC tends to converges at about 7000 steps, while the GMTC (common loss) tends to converge at about 10,000 steps. Besides, the perplexity of GMTC is lower than that of GMTC (common loss). The main reason is that the
model of GMTC no longer needs to learn the knowledge related to the label order.

3) **A HIERARCHICAL STRUCTURE MASK MATRIX**

With the introduction of the category structure mask, the current time tag can only be generated from the set of child nodes of the previous time tag in the hierarchy of Ground Truth. Experimental results show that this improvement significantly improves the prediction results of warning situation category in all indicators (table II). In the data set used in this experiment, the number of alert categories at all levels is 729, while the category with the most subclasses is “theft”, and the number of sub-classes is 15. The average number of subclasses included in all categories is 4. It can be seen that the introduction of category structure mask can greatly reduce the search range of labels at each moment, thus improving the classification performance.

However, this strong constraint may lead to the problem of error propagation, that is, if the previous prediction is wrong, the subsequent prediction could not amend. In order to reduce the influence of error propagation, the GMLC has introduced Beam Search strategy in the prediction stage, in which the Beam Width was set to 3. In addition, since the correct category path can be backtracked through the predefined hierarchy of categories as long as the model can predict the bottom category correctly in theory, the accuracy of the bottom category of GMLC(without mask) was recalculated and the result was 0.727. It can be seen that though the introduction of the category structure mask may constrain the prediction results, the accuracy of GMLC is still higher than that of GMLC (without mask). The cause of this phenomenon may be complex and may relate to the distribution of categories in the training samples. However, it can be argued that the negative impact of error propagation is relatively small compared to the performance improvement brought by the category structure mask.

**V. CONCLUSION**

The GMLC based on Seq2Seq model successfully implements multi-label classification and hierarchical classification at the same time. Compared with the traditional Seq2Seq model, a new label-order-independent loss function for multi-label classification and a hierarchical structure mask matrix for hierarchical classification are introduced. Experiments show that the proposed approach outperforms the baseline approaches in handling the classification of police case texts, in which the effectiveness of the multi-label classification and hierarchical classification tasks are both improved. Besides, the results of the two tasks are more semantically consistent. Further study includes evaluation of the performance of the proposed approach in other domain datasets.

**REFERENCES**

[1] A. Elnagar, R. Al-Debsi, and O. Einea, “Arabic text classification using deep learning models,” *Inf. Process. Manage.*, vol. 57, no. 1, Jan. 2020, Art. no. 102121.

[2] G. Tsounakas and I. Katakas, “Multi-label classification: An overview,” *Int. J. Data Warehousing Mining*, vol. 3, no. 3, pp. 1–13, 2007.

[3] I. Sutskever, O. Vinyals, and Q. V. Le, “Sequence to sequence learning with neural networks,” 2014, arXiv:1409.3215. [Online]. Available: http://arxiv.org/abs/1409.3215

[4] J. Cho, M. K. Baskar, R. Li, M. Wiesner, S. H. Mallidi, N. Yalta, M. Karafiat, S. Watanabe, and T. Horii, “Multilingual sequence-to-sequence speech recognition: Architecture, transfer learning, and language modeling,” in *Proc. IEEE Spoken Lang. Technol. Workshop (SLT)*, Athens, Greece, Dec. 2018, pp. 521–527.

[5] X. Zheng, P. Li, Z. Cha, and X. Hu, “A survey on multi-label data stream classification,” *IEEE Access*, vol. 8, pp. 1269–1275, 2020.

[6] J. Gong, H. Ma, Z. Teng, Q. Teng, H. Zhang, L. Du, S. Chen, M. Z. A. Bhaivan, 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.

[7] S. Ruder, “An overview of multi-task learning in deep neural networks,” 2017, arXiv:1706.05098. [Online]. Available: http://arxiv.org/abs/1706.05098

[8] R. Caruana, “Multitask Learning,” *Mach. Learn.*, vol. 28, no. 1, pp. 41–75, 1997.

[9] J. Baxter, “A model of inductive bias learning,” *J. Artif. Intell. Res.*, vol. 12, pp. 149–198, Mar. 2000.

[10] H. Gao, T. Wang, W. Luo, and L. Gui, “Adversarial multilabel learning for technology entity recognition,” in *Proc. IEEE Int. Conf. Prog. Informat. Comput. (PIC)*, Suzhou, China, Dec. 2018, pp. 134–139.

[11] Z.-W. Wang, S.-K. Wang, B.-T. Wan, and W. W. Song, “A novel multi-label classification algorithm based on K-nearest neighbor and random walk,” *Int. J. Distrib. Sensor Netw.*, vol. 16, no. 3, pp. 1–17, 2020.

[12] H. Daumé, III, “Bayesian multitask learning with latent hierarchies,” *J. Mach. Learn. Res.*, vol. 12, pp. 135–142, 2009.

[13] P. Prettenhofer and B. B. Stein, “Cross-language text classification using structural correspondence learning,” in *Proc. 48th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, Uppsala, Sweden 2010, pp. 11–16.

[14] D. Fu, B. Zhou, and J. Hu, “Improving SVM based multi-label classification by using label relationship,” in *Proc. Int. Joint Conf. Neural Netw.*, 2015, pp. 1–6.

[15] 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.

[16] 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.

[17] A. Eliseeef and J. Weston, “A kernel method for multi-labelled classification,” in *Proc. 14th Int. Conf. Neural Inf. Process. Syst.*, Natural Syst., 2001, pp. 681–687.

[18] 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, Apr. 2018.
[19] J. Liu, W.-C. Chang, Y. Wu, and Y. Yang, “Deep learning for extreme multi-label text classification,” in Proc. 40th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr. New York, NY, USA: ACM, Aug. 2017, pp. 115–124.

[20] B. Szalkai and V. Grolmusz, “Near perfect protein multi-label classification with deep neural networks,” Methods, vol. 132, pp. 50–56, Jan. 2018.

[21] Y.-Y. Yang, Y.-A. Lin, H.-M. Chu, and H.-T. Lin, “Deep learning with a rethinking structure for multi-label classification,” 2018, arXiv:1802.01697. [Online]. Available: http://arxiv.org/abs/1802.01697

[22] 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

[23] C. N. Silla and A. A. Freitas, “A survey of hierarchical classification across different application domains,” Data Mining Knowl. Discovery, vol. 22, nos. 1–2, pp. 31–72, Jan. 2011.

[24] D. Koller and M. Sahami, “Hierarchically classifying documents using very few words,” in Proc. ICML, vol. 223, 1997, pp. 170–178.

[25] R. Cerri, R. C. Barros, A. C. P. L. F. de Carvalho, and Y. Jin, “Reduction strategies for hierarchical multi-label classification in protein function prediction,” BMC Bioinf., vol. 17, no. 1, p. 373, Dec. 2016.

[26] J. Wehrmann, R. Cerri, and R. Barros, “Hierarchical multi-label classification networks,” in Proc. Int. Conf. Mach. Learn., 2018, pp. 5225–5234.

[27] H. Li, J. Yan, and H. Weihong, “Research and development of large-scale hierarchical classification problems,” Chin. J. Comput., vol. 35, no. 10, pp. 2101–2115, 2012.

[28] Y. Kim, “Convolutional neural networks for sentence classification,” 2014, arXiv:1408.5882. [Online]. Available: http://arxiv.org/abs/1408.5882

[29] G. Neubig, “Neural machine translation and sequence-to-sequence models: A tutorial,” 2017, arXiv:1703.01619. [Online]. Available: http://arxiv.org/abs/1703.01619

WEI ZHAO received the B.S. degree in computer science and technology from the Cheng Du University of Technology, Chengdu, China, in 2004, and the M.S. degree in control science and engineering from the National University of Defense Technology, Changsha, China, in 2006, where she is currently pursuing the Ph.D. degree in cyberspace security.

She is currently an Associate Professor with the Department of Information and Technology, Hunan Police Academy, Changsha. Her research interests include data mining, machine learning, and cyberspace security.

HUI GAO received the Ph.D. degree in systems engineering from the National University of Defense Technology, Changsha, Hunan, China, in 2011, where he is currently pursuing the Ph.D. degree in computer science.

From 2011 to 2017, he was a Senior Engineer with the China Defense Science and Technology Information Center, Beijing, China. His research interest includes the big data, machine learning, and natural language processing. He has received the Best Paper Reward from the 2018 International Conference on Progress in Informatics and Computing (PIC-2018).

SHUHUI CHEN received the Ph.D. degree in computer science and technology from the National University of Defense Technology (NUDT), Changsha, China, in 2004.

He is currently a Professor with the College of Computer Science and Technology, NUDT.

He has over 120 refereed publications that span the research forums of data mining, machine learning, privacy protection, and cyber forensics.

His research interests include cyberspace security, network architecture, and the high-speed internet measurement technology.

NAN WANG received the B.S. degree in automatic control from the College of Mechatronics and Automation, National University of Defense Technology (NUDT), Changsha, China, in 2003, and the M.S. and Ph.D. degrees in control science and engineering from the College of Mechatronics and Automation, NUDT, in 2005 and 2012, respectively. He is currently an Associate Professor with the College of Intelligence Science and Technology, NUDT. He has published more than ten articles in international journals and conferences and has coauthored one book. His research interests include mission planning, autonomous control, and the collaboration of unmanned systems. He has served as an Academic Editor for the Tactical Missile Technology.

* * *