Imbalance deep multi-instance learning for predicting isoform-isoform interactions

Guoxian Yu\textsuperscript{a,b,c}, Jie Zeng\textsuperscript{b}, Jun Wang\textsuperscript{b,c,e}, Hong Zhang\textsuperscript{b}, Xiangliang Zhang\textsuperscript{d}, Maozu Guo\textsuperscript{e}

\textsuperscript{a}School of Software, Shandong University, Jinan 250101, China
\textsuperscript{b}College of Computer and Information Science, Southwest University, Chongqing 400715, China
\textsuperscript{c}Joint SDU-NTU Centre for Artificial Intelligence Research, Shandong University, Jinan 250101, China
\textsuperscript{d}CEMSE, King Abdullah University of Science and Technology, Thuwal, SA.
\textsuperscript{e}School of Electrical and Information Engineering, Beijing University of Civil Engineering and Architecture, Beijing 100044, China

Abstract

Multi-instance learning can model complex bags (samples) that are further made of diverse instances (sub-samples). In typical multi-instance learning, the labels of bags are known while those of individual instances are unknown and to be specified. In this paper we propose an imbalanced deep multi-instance learning approach (IDMIL-III) and apply it to predict genome-wide isoform-isoform interactions (IIIs). This prediction task is crucial for precisely understanding the interactome between proteoforms and to reveal their functional diversity. The current solutions typically formulate the prediction of IIIs as a multi-instance learning problem by pairing two genes as a ‘bag’ and any two isoforms spliced from these two genes as ‘instances’. The key instances (interacting isoform pairs) trigger the label of the positive (interacting) gene bags, which is important for identifying the IIIs. Furthermore, the prediction task was simplified as a balanced classification problem, which in practice is a rather imbalanced one. To address these issues, IDMIL-III fuses RNA-seq, nucleotide sequence, amino acid sequence and exon array data, and further introduces a novel loss function to separately model the loss of positive pairs and of negative pairs, and thus to avoid the expected loss dominated by majority negative pairs. In addition, it includes an attention strategy to identify positive isoform pairs from a positive gene bag. Extensive experimental results prove the effectiveness of IDMIL-III on predicting IIIs. Particularly, IDMIL-III achieves an F1 value as 95.4\%, at least 3.8\% higher than those of competitive methods at the gene-level; and obtains an F1 as 29.8\%, at least 2.4\% higher than the state-of-the-art methods at the isoform-level. The code of IDMIL-III is available at http://mlda.swu.edu.cn/codes.php?name=IDMIL-III

Keywords: Multi-instance learning, Imbalance data, Data fusion, Deep learning, Isoform-isoform interaction

1. Introduction

Multi-instance learning (MIL) has been extensively studied and applied in diverse domains, such as image annotations\textsuperscript{[1]}\textsuperscript{,} text mining\textsuperscript{[56,57]} and so on. MIL models complex bags (samples) that are made of diverse instances (subsamples), in which the labels of bags are available but those of individual instances are unspecified. The main purpose is to make use of the bag-instance relation and instance/sample features to identify the labels of individual instances. One main key assumption of MIL is that the labels of a bag are responsible by its instances\textsuperscript{[35,58]}\textsuperscript{.} More recently, MIL has been extensively adopted to differentiate the functions of alternatively spliced isoforms\textsuperscript{[53,22,54,31,10,20]} and predict isoform interactions\textsuperscript{[15,14,16]} using gene-level annotations and gene-isoform relation, which is analogous to the bag-instance relation of MIL.

Alternative splicing (AS) is a crucial mechanism to regulate the gene expression and increase the diversity of proteins in a variety of eukaryotes\textsuperscript{[48,49]}\textsuperscript{.} It occurs in pre-mRNA transcripts and enables a gene (bag) to generate at least one alternatively spliced isoform (instance) through selectively including and excluding exons and introns. In human, more than 95\% of multi-exon genes undergo alternative splicing events\textsuperscript{[2]}\textsuperscript{.} Alternative splicing events may result in the insertion or deletion of protein domains and further modulate the protein-protein interactions (PPIs)\textsuperscript{[21]}\textsuperscript{,} which have been widely studied to understand the molecular interactions and explore the functions of proteins in the past decades. However, most studies\textsuperscript{[5,6,47]} are typically performed at the gene level. They assume that the ‘protein’ in PPIs refers to the canonical (or the longest) protein of a gene. Apparently, they neglect the impact of alternative splicing and miss the interaction diversity between isoforms. Given that, developing computational solutions to accurately identify the isoform-level interactions is essential to systematically dissect the molecular interactions and detect novel functions of proteins.

Due to the lack of a systematic and standardized database with experimentally verified IIIs as the ‘gold standard’, and the more complexity of IIIs, the isoform-level interactions are far less studied. Considering the fact that there are a mass of experimentally identified gene-level PPIs, the current studies\textsuperscript{[15,16]} formulate the task of predicting IIIs as a MIL problem, as illustrated in Figure\textsuperscript{1}\textsuperscript{.} A gene pair is regarded as a ‘bag’, isoforms alternatively spliced from the genes are paired as ‘instances’ of that bag. Based on the hypothesis of MIL, for a positive gene pair with interaction, there is at least one positive isoform pair responsible for this interaction; for a negative pair without interaction, none
of its isoform pairs has an interaction. The task of predicting IIIIs can then be modeled to identify the interacting isoform pairs from positive gene pairs with interaction. In general, the positive bags (gene pairs with interactions) can be collected from existing PPI databases (i.e., DIP [9], HPRD [3]), which are validated or curated by wet-lab experiments and experts. As for negative bags (gene pairs without interaction), they can be generated by pairing proteins from different subcellular locations [7,5]. As a result, the number of artificially generated negative samples is much larger than that of positive samples, leading to an imbalanced problem. For a negative bag, all its isoform pairs are negative, so the class imbalance is more serious in predicting IIIIs. The number of samples that belong to majority (negative) class is much larger than that of minority (positive) class, which leads the classifier can not well learn the features of minority groups and performing poorly on minority classes. As a matter of the fact, the minority class should be paid with more attention, incorrectly classifying the minority samples as majoritoy ones leads to a greater cost [11,55,58]. In practice, we focus more on the interacting isoform pairs than the non-interacting ones. However, existing solutions simply under-sample the negative bags to keep data balance and then induce classifier on the manually balanced data to predict IIIIs [15,14,43,16], whose prediction performance drops sharply when dealing with intrinsically imbalanced MIL data. In addition, the key instance (interacting isoform pair) plays an important role in triggering the label of the positive (interacting) gene bags, which is ignored by the current solutions.

To the best of authors knowledge, the study of identifying IIIIs has not been well studied yet. In this paper, IDMIL-III, an imbalance deep multi-instance learning method is proposed to predict IIIIs. To obtain the isoform-level features, IDMIL-III integrates genomic and proteomic data, including RNA-seq, nucleotide sequence, amino acid sequence and exon array data. It organizes all isoform pairs from the same gene pair as a ‘bag’ and employs a convolution neural network to extract high-level abstract for each isoform pair of the gene bag. Meanwhile, an attention model is utilized to calculate the confidence for each isoform pair. Through the element-wise product operation, an attentive feature map is generated to increase the difference of different isoform pairs of the same gene bag. Then, IDMIL-III calculates the interacting score for each isoform pair and utilizes the multi-instance learning pooling layer to bridge the isoform-level predictions and gene-level labels. The key instance (interacting isoform pair) plays an important role in triggering the label of the positive (interacting) gene bags, which is ignored by the current solutions.

Besides the application of IDMIL for predicting IIIIs, our key
technical contributions are summarized as follows:

- We propose a novel loss function to handle the imbalanced data, this loss function considers the expected loss for each category and introduces a modulating factor to more focus on hard samples.

- We incorporate the attention mechanism into IDMIL to differentiate isoform pairs of the same gene bag, and to identify the interacting isoform pairs from a positive gene bag.

This paper is organized as follows. Section 2 introduces the related works on predicting IIIs using MIL. Section 3 elaborates on the methodology, the attention mechanism and the proposed loss function for handling imbalance data. Section 4 presents the experiments and analysis from different aspects. Section 5 concludes our work.

2. Related Work

As the critical role of alternative splicing in cell functioning has been apparent, recent years have witnessed the increase of isoform-level researches. Due to the lack of isoform-level gold standard, the current studies can be categorized as two groups: single-isoform and genome-wide interactions prediction. The prediction of interactions for single-isoform can be seen as a supervised learning task, which is based on the hypothesis that for an interacting gene pair, if both of the genes generate only one isoform, this isoform pair must have an interaction. Tseng et al. [14] utilized a logistic regression model to predict isoform interactions by integrating RNA-seq, domain-domain interaction and protein-protein interaction data. At the same time, an isoform-isoform interactions database (IIIDB) is constructed for further understanding protein functions and biological pathways. By integrating multiple tissue-specific RNA-seq datasets and sequence information, Kandoi et al. [43] employed a Random Forest model to learn tissue-specific mRNA isoform functional networks (TENSION) following a leave-one-tissue-out strategy.

To expand the isoform interactions prediction from single-isoform gene to genome-wide scale, Ghadie et al. [42] proposed a domain-based method (DIIP) to predict the interactions between isoforms from the reference proteins. In their study, an isoform pair is predicted to have an interaction, if it remains the domain-domain interactions (DDI) modulating interactions from the reference proteins. If an isoform pair loses the DDI modulating interactions from the reference proteins, it is predicted to have no interactions. On the other hand, considering the fact that isoform-level interactions are unknown, Li et al. [15] and Zeng et al. [16] employed MIL framework to predict isoform-level interactions. In the MIL framework, gene pairs with multiple isoform pairs are considered as ‘bags’, and isoform pairs of the gene bag are seen as ‘instances’. Generally, for an interacting gene pair, there must be an interacting isoform pair; while for a non-interacting gene pair, none of its isoform pairs has an interaction. Based on this hypothesis, Li et al. [15] developed a single-instance bag MIL (SIB-MIL) algorithm by integrating multiple data, including RNA-seq, exon array, protein docking and pseudo-amino acid composition. Zeng et al. [16] proposed a deep multi-instance learning method (DMIL-III) to predict interactions between isoforms, which fuses RNA-seq, nucleotide sequence, domain-domain interaction and exon array data.

Apparently, studies [14, 43] on single-isoform interactions ignore the genes with multiple isoforms, which occupy a very large proportion of human transcriptome. DIIP [42] relies on domain annotations on known reference PPIs and isoforms to infer the isoform interactions. Thus, it can not work when predicting a novel isoform pair without known reference PPIs. Although the MIL-based solutions [15, 16] expand the isoform interactions prediction from single-isoform to genome-wide, they equally treat all isoform pairs of a gene bag, which is difficult to distinguish the interacting isoform pairs from non-interacting ones of a positive gene bag. In addition, these aforementioned solutions [14, 43, 15, 16] implicitly target for a balanced III dataset and ignore the intrinsic imbalanced problem. To address these issues, we proposed IDMIL-III for predicting isoform interactions from genome-wide. IDMIL-III includes a novel Category Focal Cross Entropy (CFCE) loss to handle imbalanced data by separately modeling the loss of minority and majority classes, and by more focusing on hard-to-classify samples. Meanwhile, it incorporates the attention mechanism to distinguish different isoform pairs of the same gene bag.

3. Methodology

In this section, we describe the framework of IDMIL-III, detail the attention mechanism and the proposed loss function for handling imbalance data.

3.1. Problem Formulation

Due to the lack of a large number of experimentally verified isoform pairs with/without interactions, while the bag-level interactions between proteins/genes are readily available, the prediction of IIIs is typically reformulated as a multi-instance learning task [15, 16]. Particularly, a gene pair $G_1$ and $G_2$ is regarded as a ‘bag’, an isoform pair respectively spliced from the two genes is considered as an ‘instance’. We formulate a gene bag with multiple isoform pairs as:

$$X_i = \{x_{i1}, \ldots, x_{ik}\}, x_{ij} \in \mathbb{R}^d,$$  \hspace{1cm} (1)

where $x_{ij}$ represents the $j$-th isoform pair of the $i$-th gene bag $X_i$. $d$ is the dimension of feature vector for an isoform pair, $k_i$ is the number of isoform pairs of the $i$-th gene bag, which is varying from different bags.

In the typical MIL, only the bag-level labels $Y_i \in \{0, 1\}$ are available, while individual label of isoform pair $y_{ij}$ is unknown. Our aim is to learn an isoform-level classifier based on hypotheses that for an interacting gene pair, there must be an isoform pair interacting; while for a non-interacting isoform pair, none of its isoform pairs is interacting. This assumption can be mathematically expressed as follows:

$$Y_i = \begin{cases} 1, & \text{if } \exists j, \ s.t. \ y_{ij} = 1 \\ 0, & \text{if } \forall j, \ s.t. \ y_{ij} = 0 \end{cases},$$  \hspace{1cm} (2)
where \( y_{ij} \) refers to the \( j \)-th isoform pair (instance) of the \( i \)-th gene bag. \( y_{ij} = 1 \) (\( y_{ij} = 0 \)) means this isoform pair with (without) an interaction.

### 3.2. Isoform-level Feature

To characterize isoform from diverse aspects, we integrated RNA-seq, nucleotide sequence, amino acid sequence and exon array data. Details are described as follows.

**RNA-seq data:** The RNA-seq data was downloaded from ENCODE project [37], which contains 51 tissue-specific data corresponding to 298 RNA-seq datasets. For each RNA-seq dataset, we employed the HISAT2 [38] to align the reads against the human reference genome build GRCh38.90 from Ensembl database [39]. Then, the resulting mapped read files were processed by StringTie [40] to calculate the relative abundance of the transcript in terms of FPKM (Fragments Per Kilobase of exon per Million fragments). To calculate the correlations, we removed tissue-specific data with less than 4 datasets. As a result, we obtained 43 tissue-specific data including 282 RNA-seq datasets. For each tissue-specific data, we calculated the Pearson correlation coefficients for all isoform pairs by FPKM values. Finally, we got a 43-dimensional feature vector, which is utilized as isoform-level feature when building model.

**Nucleotide Sequence:** We downloaded the nucleotide sequences of isoform from NCBI Nucleotide database. To encode the nucleotide sequence as numeric feature, the conjoint triad (CT) method [41] was employed in our work, which considers three continuous bases as a unit and results in \( 4 \times 4 \times 4 \) combinations. By counting the frequencies of each combination and normalizing it, each isoform can be characterized by a 64-dimensional feature. Then, we concatenated sequence feature of an isoform pair as a 128-dimensional feature for experiment.

**Pseudo Amino Acid Sequence**: Protein sequence carries a massive of information such as binding sites, which is essential to analyze the molecule interaction networks [50]. Although the protein sequences largely overlap with the nucleotide sequences, they are still complementary for each other, since the genomic data and peptide data have their own distinctions, the same genomic sequence do not always give the same peptide level sequence. In our work, we downloaded the protein sequence data from NCBI database. Then, the pseudo amino acid composition (PseAAC) descriptor [44] was employed to encode the amino acid sequence as numeric features. PseAAC not only considers the composition of amino acids in a protein sequence, but also takes into account interactions between amino acids at different distances. It firstly calculates the composition of the 20 standard amino acids in a protein sequence, which results in a 20-dimensional feature vector. Then, strengths of interactions between different amino acids in different distances are described by a 20-dimensional vector. Thus, a protein sequence was characterized by a 40-dimensional feature vector. Finally, the Pearson correlation coefficient is calculated as the feature data for an isoform pair.

**Exon Array Data:** We downloaded 61 exon array datasets from the NCBI GEO (Gene Expression Omnibus) database. For each dataset, the R package MEAP [35] was employed to analyze the expression value of isoforms. To measure the correlation of a pair of isoforms, we calculated the Pearson correlation by isoform expression data in each exon array dataset. By doing this, we obtained a 61-dimensional feature for isoform interactions prediction.

To make a comprehensive feature view of isoforms from these data sources, we concatenated them as the feature vector of an isoform pair. In this way, we obtained a 233-dimensional feature vector for an isoform pair \( x_i \); 43 from RNA-seq data, 128 from nucleotide sequence data, 1 form protein amino acid sequence and 61 form exon array data. Due to the lack of identified isoform-level interaction, we organize a gene pair with multiple isoform pair as ‘bag’, denoted as \( X_i = [x_{i1}, \cdots, x_{ik}] \), and \( x_{ik} \) is the \( k \)-th isoform pair of the gene bag \( X_i \). Since the number of isoform pairs \( k_i \) is varying for different bags, in this paper, we unify the number of isoform pairs in each gene bag as \( n \), which is the maximal number of isoform pairs of all gene bags. A gene bag with less than \( n \) isoform pairs is padded with \( n - k_i \) zero vectors. Then, the gene bag with padded isoform pairs is represented as \( X_i = [x_{i1}, \cdots, x_{ik}], x_i \in \mathbb{R}^{n} \), where \( d \) denotes the feature dimension of an isoform pair. Meanwhile, to distinguish the true isoform pairs with the padded ones, we defined a masking matrix as \( M_i = [m_{i1}, \cdots, m_{im}], \) where \( m_{ij} = 1 \) means the \( j \)-th isoform pair of the \( i \)-th gene bag is not padded, while \( m_{ij} = 0 \) means a padded vector.

To predict the interactions between pairwise isoforms spliced from two genes, we propose an imbalance deep multi-instance learning (IDMIL-III) model to predict IIIs, as sketched in Figure 2. IDMIL-III is consisted with three sub-networks for feature extraction, for important isoform pairs detection and for interaction prediction. Next, we describe the three sub-networks in detail.

### 3.3. Feature Extraction

IDMIL-III is fed with the pairs \( (X_i, Y_i) \) as the input. To capture the high-level representation for each isoform pair \( x_i \) of the gene bag \( X_i \), the convolution neural network is employed as the feature extraction model, which is consisted with three identical convolution modules. Each convolution module is composed with a convolution layer, a batch normalization layer, a ReLU (Rectified Linear Unit) layer, a max-pooling layer. The convolution layer is characterized with its local connection and weight sharing, which allows us to extract the high-level features for each isoform pair of the gene bag with the same parameters. In our model, each convolution layer employs a size of \( 1 \times 3 \) kernel to extract the feature for each isoform pair. Through the convolution operation on all isoform pairs of a gene bag, one convolution kernel generates one feature map. To get multiple feature maps, 64 kernels are utilized in the convolution layer of the first two convolution modules. Following the convolution layer, a batch normalization layer transforms the extracted feature maps to obey the normal distribution, and to accelerate convergence. Then, a ReLu layer is leveraged as the activation function. Although a mass of information has been extracted through convolution layer, over-detailed features are not conducive to differentiate isoform pairs of a gene bag. To remain the significant features and reduce the dimension of the input
features, a max-pooling layer with pooling size of $1 \times 3$ and stride of $1 \times 3$ is utilized in each convolution module.

Each isoform pair $x_{ij}$ is represented by a high-level abstract $h_{ij}$ via the feature extraction model as follow:

$$h_{ij} = CM(x_{ij}),$$

where $CM$ denotes the convolution module. The gene bag can be represented as $H_i = [h_{i1}, \ldots, h_{in}], H_i \in \mathbb{R}^{m \times 7}$.

3.4. Attention Model

Considering the importance of the key (interacting) isoform pairs of a positive (interacting) gene bags, we incorporate the attention mechanism to identify the interacting isoform pairs from a positive gene bag. The attention model is a two-layer neural network with the number of neurons in hidden layers as $32$ and $1$, seen in Figure 2. For each isoform pair $x_{ij}$, the attention model is responsible to automatically calculate a weight $a_{ij}$ as follows:

$$a_{ij} = \frac{1}{1 + \exp \left(-\varphi(x_{ij})\right)},$$

where $\varphi$ is the attention operation. The weights of all isoform pairs of a gene bag are denoted as $A_i = [a_{i1}, \ldots, a_{in}], A_i \in \mathbb{R}^{m \times 1}$.

Then, we integrate the attentive weight with the feature of each isoform pair (instance) to generate attentive features by an element-wise product operation as follows:

$$v_{ij} = a_{ij} \odot h_{ij},$$

Then the gene bag is denoted as $\Gamma_i = [v_{i1}, \ldots, v_{in}].$

3.5. Interaction Prediction

To get the isoform-level interaction label, a convolution layer with a size of $1 \times 7$ kernel is employed on the attentive features $\Gamma_i$ to calculate a value for each isoform pair of a gene bag as follows:

$$\mu_{ij} = \text{Conv}(v_{ij}),$$

where $\text{Conv}$ is the convolution operation, $v_{ij}$ is the attentive feature for the $j$-th isoform pair. Then, the Sigmoid function is utilized as the activation function to map the feature value to $(0, 1)$:

$$p_{ij} = \frac{1}{1 + e^{-\mu_{ij}}},$$

where $p_{ij}$ represents the interacting probability for the $j$-th isoform pair of the $i$ gene bag.

In this way, we obtain the isoform-level interaction prediction $P_i = [p_{i1}, \ldots, p_{in}]$. Note, the padded zero vector also obtains a probability. To mask the prediction generated for the padded zero vectors, we perform an element-wise product operation as follows:

$$\hat{P}_i = P_i \odot M_i,$$

Through this operation, we can mask the predicted values for padded zero vectors as $0$.

Since the isoform-level interactions are not available, it is necessary to learn a function to map the isoform-level prediction $\hat{P}_i$ to the corresponding gene-level label $Y_i$. For different purposes, researchers proposed different MIL pooling operations, including instance-based [13], bag-based [17], which respectively correspond to the max and mean operator. Among these operations,
the max-operation is the best choice to train an instance-level classifier \[^{[12]}\]. Therefore, the max-operation is adopted in our model to map the bag-level (gene pair) label to instance-level (isoform pair). Particularly, the gene-level score is defined as the maximum score of all isoform pairs in the gene bag. Isoform pair with the maximum score of a positive gene bag called ‘witness’ is considered as the responsible III for the bag. As for a negative gene bag without interactions, we regard the isoform pair with the maximum score as the ‘hard’ sample, whose label should be consistent with the label of a non-interacting gene bag. In other words, this hard isoform pair should be non-interacting (negative). Both the ‘witnesses’ from positive gene bags and ‘hard’ samples from the negative gene bags are selected as the isoform pairs responsible for the interacting/non-interacting gene bags to train the isoform-level classifier. This operation is made by a global max-pooling layer as the MIL layer as follows:

\[ \hat{Y}_i = \max(p_i). \] (9)

3.6 Imbalanced Loss Function

Through the MIL pooling operation, the prediction of IIIs turns to a binary classification problem. The expected loss of binary classification is typically calculated by the classical Cross Entropy (CE) loss as follow:

\[ CE(\hat{Y}, Y) = \frac{1}{N} \sum_{i=1}^{N} \left( -Y_i \log \hat{Y}_i + (1 - Y_i) \log (1 - \hat{Y}_i) \right). \] (10)

where \( N \) is the total number of gene bags. For simplicity, we define that

\[ \hat{Y}_i = \begin{cases} \hat{Y}_i, & Y_i = 1 \\ \hat{Y}_i - \hat{Y}_i, & Y_i = 0 \end{cases}. \] (11)

Then, CE can be simplified as:

\[ CE(\hat{Y}, Y) = \frac{1}{N} \sum_{i=1}^{N} \left( -\log \hat{Y}_i \right). \] (12)

From Eq. (12), we can see that CE equally calculates the negative log-likelihood loss for each sample and then averages them as the model loss, so it implicitly targets for balanced datasets. However, when dealing with class imbalance data, it faces the issue that the loss is dominated by the majority class and thus performs poorly on minority class. To overcome this problem, we introduce a novel cross entropy loss function, called Category Cross Entropy loss (CCE), which calculates the cross entropy loss for each class and sums up them as the expected loss. CCE is computed as:

\[ CCE(\hat{Y}) = \sum_{c=1}^{C} \frac{1}{N_c} \sum_{i=1}^{N_c} \left( -\log \hat{Y}_{ci} \right), \] (13)

where \( C \) is the number of classes, \( N_c \) is the number of samples of the \( c \)-th class.

While CCE balances the importance of samples of majority/minority classes, it does not differentiate between easy/hard samples. For this purpose, we further propose a Category Focal Cross Entropy (CFCE) loss function by introducing a modulating factor \((1 - \hat{Y}_i)\) to down-weight the easy-to-classify samples and focus more on hard-to-classify samples as follows:

\[ CFCE(\hat{Y}) = \sum_{c=0}^{C} \frac{1}{N_c} \sum_{i=1}^{N_c} \left( -\hat{Y}_i - \hat{Y}_i \log \hat{Y}_i \right). \] (14)

To minimize the CFCE loss function, the Adam algorithm \[^{[19]}\] is utilized to optimize the parameters of the model. As for the binary classification problem, the CFCE is consisted with the two parts: positive focal cross entropy (PFCE) and negative focal cross entropy (NFCE), which are respectively defined as:

\[ PFCE(\hat{Y}_i) = \frac{1}{N_1} \sum_{i=1}^{N_1} \left( -\hat{Y}_i - \hat{Y}_i \log \hat{Y}_i \right), \] (15)

\[ NFCE(\hat{Y}_0) = \frac{1}{N_0} \sum_{i=1}^{N_0} \left( -\hat{Y}_i - \hat{Y}_i \log (1 - \hat{Y}_i) \right), \] (16)

where \( N_1 \) and \( N_0 \) are the number of positive and negative samples, respectively.

Then, the category focal cross entropy loss (CFCE) can be rewritten as:

\[ CFCE(\hat{Y}) = PFCE(\hat{Y}_i) + NFCE(\hat{Y}_0). \] (17)

The derivative of CFCE is defined as:

\[ \frac{\partial CFCE}{\partial X} = \frac{\partial PFCE}{\partial X} + \frac{\partial NFCE}{\partial X}. \] (18)

\[ \frac{\partial PFCE}{\partial X} = \frac{1}{N_1} \sum_{i=1}^{N_1} \left( -\hat{Y}_i \log \hat{Y}_i \frac{\partial \hat{Y}_i}{\partial \hat{Y}_i} \frac{\partial \hat{Y}_i}{\partial X} \right), \] (19)

\[ \frac{\partial NFCE}{\partial X} = \frac{1}{N_0} \sum_{i=1}^{N_0} \left( -\hat{Y}_i \log (1 - \hat{Y}_i) \frac{\partial \hat{Y}_i}{\partial \hat{Y}_i} \frac{\partial \hat{Y}_i}{\partial X} \right), \] (20)

The overall procedure of IDMIL-III is given in Algorithm \[^{[1]}\].

4. Experiments and Analysis

In this section, we first introduce the experimental setup, and then detail the experimental results and analyze IDMIL-III from different evaluation protocols.
In this way, we got a number of 572 interacting isoform pairs and 2,772 non-interacting ones. For each gene pair, isoforms spliced from the genes as ‘instances’ of the gene bag. In the end, we obtained a total of 189,771 single-isoform gene pairs in multi-isoform dataset. Finally, we got a totally of 150,013 isoform pairs in HPRD dataset (seen as Table 2).

### Evaluation Metrics

The evaluation metrics (i.e., accuracy) commonly used for balance data may be dominated by majority classes, when dealing with class imbalance data. In this paper, representative metrics for imbalanced data are employed to comprehensively evaluate the performance of IDMIL-III, including Precision, Recall, Specificity, F-Measure (F1), G-Mean, Balance Accuracy (Balance-Acc). They are defined as:

\[
\begin{align*}
\text{Precision} & = \frac{TP}{TP + FP} \\
\text{Recall} & = \frac{TP}{TP + FN} \\
\text{Specificity} & = \frac{TN}{TN + FP} \\
F1 & = 2 \times \frac{\text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} \\
G - \text{Mean} & = \sqrt{\text{Recall} \times \text{Specificity}} \\
\text{Balance} - \text{Acc} & = \frac{1}{2} \times (\text{Recall} + \text{Specificity}),
\end{align*}
\]

where \(TP\) and \(TN\) refer to the number of interacting isoform pairs predicted as ‘positive’ (interacting) and ‘negative’ (non-interacting), \(FP\) and \(FN\) refer to the non-interacting isoform pairs that predicted as ‘positive’ and ‘negative’. In addition, the area under the receiver operating characteristic curve (AUROC)
and area under the precision-recall curve (AUPRC) are also adopted to measure the performance.

**Parameters of Methods:** IDMIL-III is implemented on the widely-used Pytorch \[20\]. It is impractical to try all parameter configurations and select the optimal ones for a deep learning model. The recommended solution is to optimize the most important parameters. We trained IDMIL-III model by randomly selecting 80% of the data in the training set as the pre-training set, and used the remaining 20% of the training set as the validation set to optimize the parameters of IDMIL-III. Then, models with the best performance were selected for experiments. The hyper-parameters configurations of IDMIL-III are listed as Table 3.

<table>
<thead>
<tr>
<th>Name</th>
<th>Range</th>
<th>Recommend</th>
</tr>
</thead>
<tbody>
<tr>
<td>Learning Rate</td>
<td>0.001, 0.005, 0.01, 0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>Batch Size</td>
<td>64, 128, 256</td>
<td>128</td>
</tr>
<tr>
<td>Dropout</td>
<td>0.1, 0.2, 0.3</td>
<td>0.1</td>
</tr>
<tr>
<td>Epoch</td>
<td>100, 300, 500</td>
<td>500</td>
</tr>
<tr>
<td>Kernel Size</td>
<td>1×3, 1×5, 1×7</td>
<td>1×3</td>
</tr>
</tbody>
</table>

In addition, we compare IDMIL-III against IIIDB \[14\], TENSION \[23\], SIB-MIL \[15\] and DMIL-III \[16\]. All the parameters are kept the same as the authors reported or optimized through grid search. Parameters of compared methods are shown in Table 4. SIB-MIL is not listed in Table 4 since its adopted naïve Bayesian network has no required input parameters.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>IIIDB</td>
<td>C</td>
</tr>
<tr>
<td>TENSION</td>
<td>n-estimators</td>
</tr>
<tr>
<td></td>
<td>Penalty</td>
</tr>
<tr>
<td></td>
<td>L2 criterion</td>
</tr>
<tr>
<td></td>
<td>gmi</td>
</tr>
<tr>
<td>DMIL-III</td>
<td>Learning Rate</td>
</tr>
<tr>
<td></td>
<td>Batch Size</td>
</tr>
<tr>
<td></td>
<td>Dropout</td>
</tr>
<tr>
<td></td>
<td>L2 Regularization</td>
</tr>
</tbody>
</table>

From Table 5, we can see that all the loss functions achieve similar results with a balance positive and negative samples. As the number of non-interacting isoform pairs increase to IR of 4, we can find that model with traditional cross-entropy loss function does well on classifying non-interacting isoform pairs with Specificity of 99.3%, while it performs poorly on positive samples, only 74.3% interacting isoform pairs are correctly recognized as ‘positive’. Compared with the traditional CE loss, CCE loss reduces the bias to majority (negative) samples and achieves a Recall as 85.9%. In addition, it obtains an F1 as 85.7%, G-Mean as 91.0%, Balance-Acc as 91.3% and AUROC as 97.5%, which are 2.0%, 5.3%, 4.6% and 0.3% higher than those of CE. Our proposed CFCE also achieves a superior performance on the dataset with IR=4, reaching an F1 as 87.6%, G-Mean as 92.2%, Balance-Acc as 92.3% and AUPRC as 94.5%. The key difference between CCE and CFCE is that CFCE introduces the modulating factor \(1 − y\) to over-weight the loss of hard-to-classify samples and down-weight the loss of easy-to-classify samples, which improves the performance of classification. From Table 5 we can find that CFCE can accurately predict more interacting and non-interacting pairs with Recall as 87.8% and Specificity as 96.8%, which is 1.9% and 0.4% higher than those of CCE.

**4.2. Ablation Study**

**4.2.1. Contribution of Different Types of Features**

Since isoform pairs spliced from single-isoform genes can be seen as the isoform-level ‘gold standard’, we employed the support vector machine (SVM) to evaluate the performance of different types of feature data on single-isoform gene pairs from HPRD dataset (seen as Table 1). The results in terms of AUROC are shown in Figure 3.

From Figure 3, we can see that the AUROC of all features range from 55.1% to 75.9% with nucleotide sequence data ranking the highest. It is because that nucleotide sequence can provide a mass of information, such as binding sites. In addition, we can find that the integrated data achieves the best performance with AUROC as 77.8%, which meets the expectation that integrating multiple biological data related with alternative splicing and isoforms contributes to a more accurate prediction.

**4.2.2. Analysis of Different Loss Functions**

To verify the effectiveness of our CFCE loss function for handling class imbalance data, experiments are carried out on single-isoform dataset with the *imbalance ratio* (IR) as 1, 4, 9. We set the number of interacting (positive) isoform pairs to 4900 and the number of non-interacting (negative) isoform pairs to a multiple of positive isoform pairs with reference to the IR.

In addition, we carried out experiment on all 4,963 interacting isoform pairs and 150,013 non-interacting isoform pairs with an imbalance ratio of 30. As comparisons, the Cross Entropy (CE) and Category Cross Entropy (CCE) loss were also employed for experiments. According to the principle adopted in IIIDB \[14\], labels of an isoform pair is consistent with its hosting gene pair when this gene pair has only one isoform pair. Therefore, we evaluate the performance of all methods by assigning the label of a gene pair to its isoform pair and report the Results in Table 5.

From Table 5 we can see that all the loss functions achieve a Recall as 85.9%, which is 2.0% higher than those of CE. Our proposed CFCE also achieves a superior performance on the dataset with IR=4, reaching an F1 as 87.6%, G-Mean as 92.2%, Balance-Acc as 92.3% and AUPRC as 94.5%. The key difference between CCE and CFCE is that CFCE introduces the modulating factor \(1 − y\) to over-weight the loss of hard-to-classify samples and down-weight the loss of easy-to-classify samples, which improves the performance of classification. From Table 5 we can find that CFCE can accurately predict more interacting and non-interacting pairs with Recall as 87.8% and Specificity as 96.8%, which is 1.9% and 0.4% higher than those of CCE.

When the IR increases to 9, the above trend is more obvious. We can find that IDMIL-III with CE loss function can accurately predict almost all the non-interacting isoform pairs with a Specificity as 99.8%, while it performs the most worse
on the positive samples. Most interacting pairs are identified as ‘negative’, except some easy-to-classify interacting pairs with a Precision as 97.1% but Recall as 68.3%. In addition, we can see that CCE and CFCE can reduce the impact of imbalanced data to a certain extent. IDMIL-III with CCE loss achieves an F1 as 84.3%, G-Mean as 92.4%, Balance-Acc as 92.5%, AUROC as 97.8% and AUPRC as 92.6%, which are 6.0%, 10.1%, 8.7%, 4.1%, 0.4%, 0.4% and 0.2% and 0.3% higher than those of CCE, and 1.9%, 0.2%, 0.3%, 0.1% and 0.9% higher than those of CE. The CFCE loss reaches an F1 as 86.2%, G-Mean as 92.6%, Balance-Acc as 92.8%, AUROC as 97.9% and AUPRC as 92.4%, which is 4.1%, 9.9%, 8.4%, 0.1% and 0.3% higher than those of CE. The CFCE loss function achieves an F1 as 79.1% and 86.6%. By summarizing the results in Table 5, we can see that compared with CE and CCE, the CFCE not only reduces the impact of imbalanced data on minority samples to a certain extent with a Precision as 66.4% and Recall as 86.3%, but also achieves a better performance with a Precision as 72.8% and Recall as 86.2%, G-Mean as 92.6%, Balance-Acc as 92.8%, AUROC as 97.8% and AUPRC as 92.0%, which is 4.1%, 9.9%, 8.4%, 0.1% and 0.3% higher than those of CE. The CFCE loss function obtains a Precision as 96.0%, Recall as 88.7%, Specificity as 96.2%, F1 as 92.2%, G-Mean as 92.4%, Balance-Acc as 92.4%, AUROC as 97.0% and AUPRC as 97.4%, which are 1.6%, 1.5%, 1.5%, 1.6%, 1.5%, 0.2% and 0.2% higher than those of IDMIL-III without the attention module. This advantage is that the attention mechanism can differentiate isoform pairs of a gene bag with different weights, which makes the difference of isoform pairs from the same gene pair more obvious, and thus to easily identify the important isoform pairs that are responsible for the interactions of gene pairs. This ablation study confirms the necessity and benefit of incorporating attention mechanism into IDMIL-III for predicting IIIs.

When the IR reaches to 30, due to the high imbalance data, models with different loss functions achieve a prominent performance on negative (majority) samples with Specificity as 100%, 98.6% and 98.9%, but they perform poorly on positive (minority) samples. Although CE achieves a Precision as 97.6%, only 60.3% positive samples are correctly recognized. CCE reduces the impact of imbalanced data on minority samples to a certain extent with a Precision as 66.4% and Recall as 86.3%. Compared with CE and CCE, the CFCE not only reduces the impact of imbalanced data to a certain extent, but also achieves a better performance with a Precision as 72.8% and Recall as 86.6%. By summarizing the results in Table 5, we can see that IDMIL-III with CFCE loss function achieves an F1 as 79.1%, G-Mean as 92.6%, Balance-Acc as 92.8%, AUROC as 97.8% and AUPRC as 92.4%, which are 4.6%, 15.0%, 12.7%, 0.7% and 2.2% higher than those of CE, 4.1%, 0.4%, 0.4%, 0.2% and 1.9% higher than those of CCE. In addition, we employed the singed-rank test [59] to further check the statistical difference between CFCE and different loss functions, all the p-value are smaller than 0.0397.

Overall, this ablation study confirms the effectiveness of our CFCE on handling imbalanced data for predicting IIIs.

### 4.2.3. Analysis of Attention Mechanism

To validate the effectiveness of the attention module on differentiating different isoform pairs of a multi-isoform gene bag, we selected the gene pairs with 4 isoform pairs from multi-isoform dataset for experiments, which results in a total of 4,228 interacting gene pairs and 4,149 non-interacting gene pairs. In addition, IDMIL-III without the attention module (base) is employed for comparison. Since the isoform-level interactions are not available, we evaluate the performance at the gene-level by assigning the maximum score of its isoform pairs to the hosting gene bag and plot the results in Figure 4.

From Figure 4, we can see that IDMIL-III with the attention module obtains a Precision as 96.0%, Recall as 88.7%, Specificity as 96.2%, F1 as 92.2%, G-Mean as 92.4%, Balance-Acc as 92.4%, AUROC as 97.0% and AUPRC as 97.4%, which are 1.6%, 1.5%, 1.5%, 1.6%, 1.5%, 0.2% and 0.2% higher than those of IDMIL-III without the attention module. This advantage is that the attention mechanism can differentiate isoform pairs of a gene bag with different weights, which makes the difference of isoform pairs from the same gene pair more obvious, and thus to easily identify the important isoform pairs that are responsible for the interactions of gene pairs. This ablation study confirms the necessity and benefit of incorporating attention mechanism into IDMIL-III for predicting IIIs.

### 4.3. Comparison with State-of-the-Art Methods

#### 4.3.1. Predicting IIIs on Single-Isosof orm Dataset

To access the performance of IDMIL-III on data with different imbalance ratios, we conduct experiments on single-isoform dataset adopted from HPRD. Four state-of-the-art methods for predicting IIIs are employed for comparison, including IIIDB [14], SIB-MIL [15], TENSION [43], DMIL-III [16]. Results are reported in Table 6.

From Table 6 we can see that on the dataset with IR = 1, IDMIL-III achieves an F1 as 88.1%, G-Mean as 88.5%, Balance-Acc as 86.9%, AUROC as 94.7% and AUPRC as 95.3%, which
are 9.7%, 9.4%, 9.3%, 8.4% and 6.9% higher than those of IIIDB; and 13.9%, 12.5%, 12.0%, 10.3% and 8.4% higher than those of SIB-MIL; 4.1%, 5.4%, 5.4%, 3.8% and 4.3% higher than those of TENSION; and 2.1%, 1.9%, 1.7%, 0.3% and 0.7% higher than those of DMIL-III. We observe DMIL-III and IDMIL-III achieve a superior performance than others, that is because they utilize the Convolution Neural Network (CNN) to capture more essential information from different data for identifying IIIs. In addition, we find that IDMIL-III achieves a higher Specificity as 93.1% but a lower Recall as 80.6%. That is because DMIL-III performs poorly on some hard-to-classify positive isoform pairs and tends to predict them as negative, resulting in a higher Specificity and Precision, but a lower Recall.

We further increase the number of non-interacting isoform pairs with IR=4, IR=9 and IR=30, respectively. We can find that all these compared methods are impacted by imbalanced data and have a good performance on the majority negative isoform pairs, while performs poorly on the positive isoform pairs. Unlike these compared methods, IDMIL-III takes the imbalanced data into consideration and reduces the impact of majority negative isoform pairs to the loss. It has an F1 as 87.6%, 86.2% and 79.1% when IR=4, IR=9 and IR=30, respectively, which are 24.6%, 30.5% and 27.2% higher than those of IIIDB; 26.9%, 39.1% and 52.3% higher than those of SIB-MIL; 16.8%, 23.7% and 20.7% higher than those of TENSION; 4.8%, 6.1% and 6.4% higher than those of DMIL-III. These results again confirm that it is necessary to consider the imbalanced problem in predicting IIIs. We can conclude that our proposed CFCE loss is effective in dealing with the imbalanced isoform data.

### 4.3.2. Predicting IIIs on Multi-Isoform Dataset

To further evaluate the performance on multi-isoform gene pairs, IDMIL-III and these compared methods are applied on the whole HPRD dataset. Here, IIIDB [14] and TENSION [43] are excluded for this experiment, since they are only applicable to gene pairs with single isoform pairs. Since the isoform-level interactions are unknown, the predictions of isoform pairs are finally mapped to the gene level by assigning the maximum score of all its isoform pairs to the hosting gene bag to evaluate the performance. Figure[5] reveals the results.

![Figure 5: Results of Predicting IIIs on the Whole HPRD Dataset.](image-url)

From Figure[5] we can see that IDMIL-III outperforms the other two compared methods, with F1 as 95.4%, G-Mean as 95.4%, Balance-Acc as 95.4%, which are 42.2%, 37.3% and 35.3% higher than those of SIB-MIL; and 3.8%, 3.5% and 3.2% higher than those of DMIL-III. The superior performance of IDMIL-III on predicting IIIs can be attributed to two factors: i) IDMIL-III improves the traditional cross-entropy loss function by calculating the mean loss for each category and introducing the modulating factor. The former can reduce the impact of easy-to-classify samples and up-weight hard-to-classify samples, which conduces to a better performance with a Recall as 84.8%, Precision as 91.7%.

![Table 6: Results of predicting IIIs on single-isoform dataset from HPRD with different imbalance ratios (IR).](table-url)
the same gene pair and find the important isoform pairs that are responsible for the interactions between gene pairs.

4.4. Validation of Predicted Isoform Interactions

In the previous subsection, we have demonstrated the effectiveness of IDMIL-III by surrogating the evaluation at the gene level, due to the lack of isoform-level label. Here, we evaluate the isoform-level performance of IDMIL-III by three kinds of experiments.

4.4.1. Predicting on Experimentally Identified IIIIs

To evaluate the performance on the isoform-level predictions, a set of experimentally identified isoform pairs (seen as Table 2) were collected in our experiment. We firstly removed the gene pairs that appeared in Yang’s dataset from HPRD dataset. Then, we trained IDMIL-III and other comparing methods on the updated HPRD dataset to predict IIIIs on Yang’s dataset. Results are shown in Table 7.

From Table 7, we can see that IDMIL-III accurately predicts 386 out of 495 interacting gene pairs with Accuracy as 78.0%, which is 31.1%, 13.4% higher than that of SIB-MIL and DMIL-III. As to the isoform level, IDMIL-III accurately predicts 380 out of 572 interacting isoform pairs, which outperforms SIB-MIL and DMIL-III by a margin of 193, 82, respectively. SIB-MIL only identifies 187 out of 572 interacting isoform pairs and has a Precision as 21.5%, which is 2.9% and 2.1% higher than that of DMIL-III and IDMIL-III. That is because SIB-MIL identifies most interacting isoform pairs as ‘negative’, and only correctly recognizes a few easy-to-classify ones as ‘positive’. As a consequence, SIB-MIL has a lower Recall as 32.7%. In contrast, IDMIL-III achieves a Recall as 66.4%, which is 33.7%, 14.3% higher than that of SIB-MIL, DMIL-III. Precision and Recall measure the performance from different aspects, so their harmonic F1 value is reported also. Particularly, IDMIL-III achieves an F1 as 29.8%, which is 4.0%, 2.4% higher than that of SIB-MIL, DMIL-III.

These experimental results suggest that IDMIL-III can more credibly identify IIIIs than other compared methods, since it incorporates the attention mechanism into the deep multi-instance learning model, which helps to find the important isoform pairs of a gene bag. In addition, IDMIL-III proposes a modulating factor to force the classifier to pay more attention on hard samples, and the category focal cross entropy loss to handle imbalance data, both of which improve the prediction accuracy.

4.4.2. Isoform Function Prediction

Alike traditional PPI network based gene function prediction [22][46][55], identifying the interactions between isoform pairs is vital to predict the functions of isoforms [22][10]. Here, the III network constructed by IDMIL-III from HPRD data is leveraged to predict the functions of isoforms. Firstly, we downloaded the latest version of Gene Ontology (GO) annotation file (archived on 2020-03-23), which stores the known functional annotations of genes. To avoid the change of GO structure, the contemporary GO file was downloaded as well. It records the relationships between GO terms and divides the biological functions of genes into Biological Process Ontology (BPO), Cellular Component Ontology (CCO) and Molecular Function Ontology (MFO). Due to the large number of GO terms, too rare/general GO terms were excluded, which are annotated to fewer than 10 genes or more than 300 genes. As a result, we got a number of 4,214 (BPO), 562 (CCO), 774 (MFO) GO terms for experiments. Then, we employed iMILP (instance-oriented Multiple Instance-based Label Propagation) [22] to predict isoform functions. For reference, the similar IIIIs network individually constructed by SIB-MIL and DMIL-III are also utilized in our experiment. Since the isoform-level functional annotations are not available, we evaluate the performance at the gene level by following the principle in [22][10][50]. Results are reported in Table 8.

To reach a comprehensive comparison, three recommended metrics for protein function prediction [23] are used to quantify the performance, they are Smin, Fmax, AvgAUC. Smin leverages the remaining uncertainty and missing information to calculate the minimum semantic distance between predictions and ground-truth across all possible thresholds by using the ontology hierarchy. Fmax refers to the overall maximum harmonic mean of precision and recall of all possible thresholds of the prediction likelihoods. AvgAUC calculates the area under the receiver operating curve (AUROC) for each GO term, and then takes the average of these areas. In addition, AUPRC is also employed in our experiment, which calculates the area under the precision-recall curve for each term and then takes the average.

In fact, isoform function prediction can be modeled as a multi-label learning problem [33][34]. Thus, three widely used metrics in multi-label learning are also utilized in our experiments, including MicroAvgF1, MacroAvgF1, 1-Rankingloss. MicroAvgF1 calculates the F1 score across all the GO terms. MacroAvgF1 computes the F1 value for each term, and then averages on these terms. Rankingloss calculates the average value of incorrectly predicted annotations ranking ahead of the ground-truth annotations of genes. Note that, except Smin, the higher values of these metrics, the better the performances is.

From Table 8 we can see that IDMIL-III generally manifests a better performance than other compared methods across the BPO, MFO and CCO, respectively, though these performance values are close to those of some compared methods. In addition, we use the signed rank test to further check the statistical difference between IDMIL-III and other compared methods across different GO branches and metrics, all the p-values are smaller than 0.0006. The study again suggests the effectiveness of IDMIL-III, which can construct a better III network that enables more accurate isoform function prediction.

4.4.3. Case Study

We collected four exemplar genes (BCL2L1, CRK, PLP1 and AXIN1), as listed in Table 9, to further investigate the practical usage of IDMIL-III for differentiating the individual interaction partners of isoforms spliced from the same gene. For this purpose, we manually checked the interaction partners of isoforms spliced from these genes. By manually searching the isoform-related literature from PubMed, we found that Bcl-x(L) interacts with BAK1 [24], BAX [25], NLRP1 [28], HRK [27] and BAD [26], plays a role of inhibiting apoptosis, while Bcl-x(S) promotes apoptosis in cells by interacting with BCL2 [29].
In addition, Yang et al. [21] tested a small number of isoform-level interactions through yeast two-hybrid (Y2H) screens. We selected three genes (CRK, PLP1 and AXIN1) for case study. CRK produces two isoforms, namely CRK_1 and CRK_2. CRK_1 interacts with ELK3 and PPP1FB2, while CRK_2 interacts with TP53BP2, DPPA4 and ASAP3. Both the two isoforms of PLP1 interact with BCL2L13. PLP1_1 interacts with PTPRN, while PLP1_2 does not. AXIN1 generates three isoforms (AXIN1_1, AXIN1_2 and AXIN1_3). AXIN1_3 has interactions with ETV6, NBR1 and TSGA10, but AXIN1_2 has interactions with RNFI11 and ZFYVE19. We apply IDMIL-III and other compared methods trained on HPRD dataset to predict individual interaction partners of isoforms hosted by these genes, and report the results in Table 10.

From Table 10, we can see that all the methods correctly identify the interaction between Bcl-x(S) and BCL2. Compared with the predictions of IIIDB, IDMIL-III accurately predicts the interactions between Bcl-x(L) and BAK1, HRK and BAD. IDMIL-III recognizes the interactions between Bcl-x(L) and BAK1, BAX, HRK and BAD, while SIB-MIL does not. Meanwhile, compared with the predictions of DMIL-III, IDMIL-III correctly identifies the interactions between Bcl-x(L) and BAX and HRK. As to the predictions for two isoforms of CRK, the interactions between CRK_2 and TP53BP2, ASAP3 are successfully identified by all methods. IDMIL-III correctly recognizes the interaction between CRK_2 and DPPA4, while others do not, and IDMIL-III fails to identify the interactions between CRK_1 and ELK3, but others do. Both SIB-MIL and IDMIL-III accurately predict the interaction between CRK_1 and PPP1FB2, while others do not. In addition, we can find that both IDMIL-III and DMIL-III accurately predict that PLP1_1 interacts with PTPRN and BCL2L13, and PLP1_2 interacts with BCL2L13. Compared with the predictions of IIIDB, IDMIL-III identifies the interactions between BCL2L13 and PLP1_1. PLP1_2. SIB-MIL does not identify any interacting partner of PLP1_1 and PLP1_2, while IDMIL-III does. Similar to the case of PLP1, SIB-MIL does not predict the interactions between AXIN1 and its interacting partners, while IDMIL-III identifies the interactions between AXIN1_1 and ETV6, TSGA10, and interactions between AXIN1_2 and RNFI11 and ZFYVE19. Compared with IIIDB, IDMIL-III correctly predicts that AXIN1_3 interacts with TSGA10, and AXIN1_2 interacts with RNFI11 and ZFYVE19. Last but not least, IDMIL-III identifies the interaction between AXIN1_3 and TSGA10, while DMIL-III does not.

By taking the isoform pairs without recorded interactions as ‘negative’, we further calculate the Precision, Recall and F1 for all compared methods. We can find that IDMIL-III achieves an F1 as 65.3%, which is 26.4% higher than that of IIIDB, 29.6% higher than that of SIB-MIL, and 8.2% higher than DMIL-III. Meanwhile, IDMIL-III achieves a higher Recall as 84.2%. We note that SIB-MIL achieves a higher Precision, since SIB-MIL tends to identify the interacting isoform pairs as non-interacting ones, which results in a higher Precision. In summary, we can conclude that IDMIL-III gives a more comprehensive prediction of IIIs.

5. Conclusion

In this paper, we proposed an imbalance deep multi-instance learning model (IDMIL-III) to reduce the impact of imbalance data by a novel category focal cross entropy loss function. In addition, IDMIL-III considers the different relevance of instances, and incorporates the attention mechanism to identify the important instances responsible for the label of a bag. Experiments on public datasets have shown that IDMIL-III can not only effectively handle imbalanced data by separately averaging the cross entropy loss for minority and majority classes, and by introducing a modulating factor to more focus on the hard-to-classify samples. In addition, the inclusion of attention mechanism is helpful to differentiate different instances from the same bag, which is beneficial to find the important isoform pairs responsible for the interaction between genes. The experiments on validated IIIs further confirm that IDMIL-III not only outperforms the state-of-the-art methods in recognizing the interactions between isoforms, but also constructs a more competent IIIs network for isoform function prediction. In addition, the case study on exemplar genes with known isoform level interactions again demonstrates its effectiveness in predicting IIIs.

---

**Table 7: Results of prediction on validated IIIs dataset**

<table>
<thead>
<tr>
<th>Methods</th>
<th>Test Pairs</th>
<th>Identified Pairs</th>
<th>Accuracy</th>
<th>Test Pairs</th>
<th>Identified Pairs</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIB-MIL</td>
<td>495</td>
<td>252</td>
<td>0.469</td>
<td>572</td>
<td>187</td>
<td>0.213</td>
<td>0.327</td>
<td>0.285</td>
</tr>
<tr>
<td>DMIL-III</td>
<td>495</td>
<td>320</td>
<td>0.646</td>
<td>572</td>
<td>298</td>
<td>0.186</td>
<td>0.521</td>
<td>0.274</td>
</tr>
<tr>
<td>IDMIL-III</td>
<td>495</td>
<td>346</td>
<td>0.780</td>
<td>572</td>
<td>380</td>
<td>0.192</td>
<td>0.664</td>
<td>0.298</td>
</tr>
</tbody>
</table>

---

**Table 8: Results of Isoform Function Prediction**

<table>
<thead>
<tr>
<th>Methods</th>
<th>Smin</th>
<th>Smac</th>
<th>Fmax</th>
<th>AvgAUC</th>
<th>AUPRC</th>
<th>MicroAvgF1</th>
<th>MacroAvgF1</th>
<th>1-Rankingloss</th>
</tr>
</thead>
<tbody>
<tr>
<td>BPO</td>
<td>3.162</td>
<td>0.126</td>
<td>0.435</td>
<td>0.701</td>
<td>0.689</td>
<td>0.710</td>
<td>0.719</td>
<td>0.725</td>
</tr>
<tr>
<td>SIB-MIL</td>
<td>3.489</td>
<td>0.141</td>
<td>0.328</td>
<td>0.501</td>
<td>0.466</td>
<td>0.492</td>
<td>0.501</td>
<td>0.501</td>
</tr>
<tr>
<td>DMIL-III</td>
<td>3.484</td>
<td>0.141</td>
<td>0.328</td>
<td>0.501</td>
<td>0.466</td>
<td>0.492</td>
<td>0.501</td>
<td>0.501</td>
</tr>
<tr>
<td>IDMIL-III</td>
<td>3.484</td>
<td>0.141</td>
<td>0.328</td>
<td>0.501</td>
<td>0.466</td>
<td>0.492</td>
<td>0.501</td>
<td>0.501</td>
</tr>
</tbody>
</table>

---

In addition, Yang et al. [21] tested a small number of isoform-level interactions through yeast two-hybrid (Y2H) screens. We selected three genes (CRK, PLP1 and AXIN1) for case study. CRK produces two isoforms, namely CRK_1 and CRK_2. CRK_1 interacts with ELK3 and PPP1FB2, while CRK_2 interacts with TP53BP2, DPPA4 and ASAP3. Both the two isoforms of PLP1 interact with BCL2L13. PLP1_1 interacts with PTPRN, while PLP1_2 does not. AXIN1 generates three isoforms (AXIN1_1, AXIN1_2 and AXIN1_3). AXIN1_3 has interactions with ETV6, NBR1 and TSGA10, but AXIN1_2 has interactions with RNFI11 and ZFYVE19. We apply IDMIL-III and other compared methods trained on HPRD dataset to predict individual interaction partners of isoforms hosted by these genes, and report the results in Table 10.

From Table 10, we can see that all the methods correctly identify the interaction between Bcl-x(S) and BCL2. Compared with the predictions of IIIDB, IDMIL-III accurately predicts the interactions between Bcl-x(L) and BAK1, HRK and BAD. IDMIL-III recognizes the interactions between Bcl-x(L) and BAK1, BAX, HRK and BAD, while SIB-MIL does not. Meanwhile, compared with the predictions of DMIL-III, IDMIL-III correctly identifies the interactions between Bcl-x(L) and BAX and HRK. As to the predictions for two isoforms of CRK, the interactions between CRK_2 and TP53BP2, ASAP3 are successfully identified by all methods. IDMIL-III correctly recognizes the interaction between CRK_2 and DPPA4, while others do not, and IDMIL-III fails to identify the interactions between CRK_1 and ELK3, but others do. Both SIB-MIL and IDMIL-III accurately predict the interaction between CRK_1 and PPP1FB2, while others do not. In addition, we can find that both IDMIL-III and DMIL-III accurately predict that PLP1_1 interacts with PTPRN and BCL2L13, and PLP1_2 interacts with BCL2L13. Compared with the predictions of IIIDB, IDMIL-III identifies the interactions between BCL2L13 and PLP1_1. PLP1_2. SIB-MIL does not identify any interacting partner of PLP1_1 and PLP1_2, while IDMIL-III does. Similar to the case of PLP1, SIB-MIL does not predict the interactions between AXIN1 and its interacting partners, while IDMIL-III identifies the interactions between AXIN1_1 and ETV6, TSGA10, and interactions between AXIN1_2 and RNFI11 and ZFYVE19. Compared with IIIDB, IDMIL-III correctly predicts that AXIN1_3 interacts with TSGA10, and AXIN1_2 interacts with RNFI11 and ZFYVE19. Last but not least, IDMIL-III identifies the interaction between AXIN1_3 and TSGA10, while DMIL-III does not.

By taking the isoform pairs without recorded interactions as ‘negative’, we further calculate the Precision, Recall and F1 for all compared methods. We can find that IDMIL-III achieves an F1 as 65.3%, which is 26.4% higher than that of IIIDB, 29.6% higher than that of SIB-MIL, and 8.2% higher than DMIL-III. Meanwhile, IDMIL-III achieves a higher Recall as 84.2%. We note that SIB-MIL achieves a higher Precision, since SIB-MIL tends to identify the interacting isoform pairs as non-interacting ones, which results in a higher Precision. In summary, we can conclude that IDMIL-III gives a more comprehensive prediction of IIIs.

5. Conclusion

In this paper, we proposed an imbalance deep multi-instance learning model (IDMIL-III) to reduce the impact of imbalance data by a novel category focal cross entropy loss function. In addition, IDMIL-III considers the different relevance of instances, and incorporates the attention mechanism to identify the important instances responsible for the label of a bag. Experiments on public datasets have shown that IDMIL-III can not only effectively handle imbalanced data by separately averaging the cross entropy loss for minority and majority classes, and by introducing a modulating factor to more focus on the hard-to-classify samples. In addition, the inclusion of attention mechanism is helpful to differentiate different instances from the same bag, which is beneficial to find the important isoform pairs responsible for the interaction between genes. The experiments on validated IIIs further confirm that IDMIL-III not only outperforms the state-of-the-art methods in recognizing the interactions between isoforms, but also constructs a more competent IIIs network for isoform function prediction. In addition, the case study on exemplar genes with known isoform level interactions again demonstrates its effectiveness in predicting IIIs.
Table 9: Details for Isoforms of Different Genes

<table>
<thead>
<tr>
<th>Gene</th>
<th>Isomorph</th>
<th>RefSeq ID</th>
<th>mRNA-length</th>
<th>Protein-length</th>
<th>Unprot ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>BCL2L1</td>
<td>Bcl-x(L)</td>
<td>NM_13578</td>
<td>2575</td>
<td>233</td>
<td>Q07817-1</td>
</tr>
<tr>
<td>Bcl-s(L)</td>
<td>NM_001191</td>
<td>2386</td>
<td>170</td>
<td>Q07817-2</td>
<td></td>
</tr>
<tr>
<td>CRK</td>
<td>CRK-1</td>
<td>NM_010823</td>
<td>972</td>
<td>304</td>
<td>P40108-1</td>
</tr>
<tr>
<td></td>
<td>CRK-2</td>
<td>NM_005206</td>
<td>742</td>
<td>204</td>
<td>P40108-2</td>
</tr>
<tr>
<td>PLP1</td>
<td>PLP1-1</td>
<td>NM_199427</td>
<td>725</td>
<td>242</td>
<td>P09201-2</td>
</tr>
<tr>
<td></td>
<td>PLP1-2</td>
<td>NM_0011285344</td>
<td>831</td>
<td>277</td>
<td>P09201-1</td>
</tr>
<tr>
<td>AXIN1</td>
<td>AXIN1-2</td>
<td>NM_003502</td>
<td>2586</td>
<td>862</td>
<td>O15169-1</td>
</tr>
<tr>
<td></td>
<td>AXIN1-3</td>
<td>NM_181050</td>
<td>287</td>
<td>826</td>
<td>O15169-2</td>
</tr>
</tbody>
</table>

Table 10: Results of Different Methods for Case Study

<table>
<thead>
<tr>
<th>Gene</th>
<th>Isoform</th>
<th>Interacting Partner</th>
<th>SHH-MI</th>
<th>DMML-III</th>
<th>DML-III</th>
</tr>
</thead>
<tbody>
<tr>
<td>BCL2L1</td>
<td>Bcl-x(L)</td>
<td>BAK1</td>
<td>√</td>
<td>√</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>BAX</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>NLRP1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>HRK</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>BAD</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>CRK-1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>ELK1</td>
<td>√</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>CRK-2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>PPFIBP2</td>
<td></td>
<td>√</td>
<td></td>
</tr>
<tr>
<td>PLP1</td>
<td>PLP1-1</td>
<td>TPRN</td>
<td></td>
<td>√</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>BCL2L1</td>
<td></td>
<td>√</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>PLP1-2</td>
<td></td>
<td>√</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>AXIN1-3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>ETV6</td>
<td></td>
<td>√</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>NBR1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>TSGA10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>AXIN1-2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>RNF111</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>ZFYVE19</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Precision: 0.856
Recall: 0.623
F1: 0.653

Acknowledgements

This work is supported by Natural Science Foundation of China (61872300 and 62031003).
