24th European Conference on Artificial Intelligence - ECAI 2020
Santiago de Compostela, Spain

A Graph-based Measurement for Text Imbalance Classification

Jiachen Tian[1] and **Shizhan Chen**[1] and **Xiaowang Zhang**[1 2] and **Zhiyong Feng**[1 3]

Abstract. Imbalanced text classification, as practical and essential text classification, is the task to learn labels or categories for imbalanced text data. Existing imbalanced text classification approaches are mostly based on the Imbalance Ratio (i.e. ratio of sizes between categories). Recently, some researchers verified that the imbalance ratio severely affects the performance of classifiers when intrinsic characteristics of data such as class overlapping and small disjuncts occur. However, since the distribution of real-world data is unknown, it is difficult to describe above intrinsic characteristics directly. In this paper, we transform the unknown distribution of data into a graph model and present a graph-based imbalance index named $\mathcal{G}_{\mathrm{IR}}$ to predict the impact of imbalanced text data on classification performance. Firstly, we introduce an environmental factor that makes the imbalance index sensitive to the intrinsic characteristics of data. Secondly, we propose a graph-based method to calculate this environmental factor. Finally, we use the imbalance index to analyze the performances of imbalanced learning methods and the impact of imbalanced data on text classifiers. The experimental results evaluated on both synthetic data sets and real-world data sets demonstrate the effectiveness of our approach.

1 Introduction

Imbalanced text classification is the task of classifying the imbalanced text data into one or more defined classes [17]. Imbalanced data refers to data in which the size of one class is significantly larger than the size of other classes. In the binary-classification task, the class with the larger size is called the majority class, while the other one is called the minority class [1]. Imbalanced data often exist in real-world applications, such as sentiment analysis [8], topic labeling [36] and spam detection [26]. The main difficulty in analyzing imbalanced data is that the distribution of data is unknown. This issue severely hinders us from synthesizing ideal data set (balanced data set) for training the machine learning model.

To analyze imbalanced data set, previous studies defined Imbalance Ratio (IR), which is obtained by dividing the size of the majority class by that of minority one [12]. Standard imbalanced learning methods are mainly to improve IR and focus on the preprocessing of data and the improvement of the algorithm. For example, oversampling techniques attempt to rebalance data by synthesizing data for minority class [5, 16], while undersampling techniques reduce data for majority class [20]. Cost-sensitive methods assign different costs for majority class and minority class [18, 25].

[1] College of Intelligence and Computing, Tianjin University, China
[2] Corresponding author: xiaowangzhang@tju.edu.cn.
[3] College of Intelligence and Computing, Shenzhen Research Institute of Tianjin University, Tianjin University, China

However, some data intrinsic complexity (DIC) also affect the performance of the classifier as well as IR. More than that, DIC such as class overlapping (CO) and small disjuncts (SD) deteriorate the imbalance problem [31]. CO in text classification happens when the common area of the input space includes texts from more than one class [24]. SD is the heterogeneous area phenomenon of texts belonging to the same class in the input space [23]. Therefore, it becomes interesting to investigate a proper imbalance index which takes into account IR, CO and SD. Because we cannot intuitively know the overlapping area among different classes and the smallest cluster in the minority class, it is difficult to calculate the degrees of CO and SD. Therefore, how to represent the unknown distribution of imbalanced text data is a challenging study.

In this paper, we propose a graph-based text imbalance index named $\mathcal{G}_{\mathrm{IR}}$ with an environmental factor that fully considers CO and SD of imbalanced texts. To characterize CO and SD, our approach converts text data set into an undirected, weighted and labelled graph. Therefore, we can use the structural information of the graph to approximate represent the unknown distribution of imbalanced texts. The experimental results show our approach is efficient in capturing text imbalance problem. Main contributions as follows:

- We propose an imbalance index $\mathcal{G}_{\mathrm{IR}}$, which is sensitive to DIC via an environmental factor and be highly correlated with the classification error rate (ER).
- We present a graph-based method to calculate the environmental factor by transforming the unknown distribution of input data into a graph model.
- $\mathcal{G}_{\mathrm{IR}}$ can reflect the imbalance of data sets without training and testing models and can be used to evaluate the performances of imbalanced learning methods. In particular, $\mathcal{G}_{\mathrm{IR}}$ is the first imbalance index that can be applied to the NN-based model of text classification.

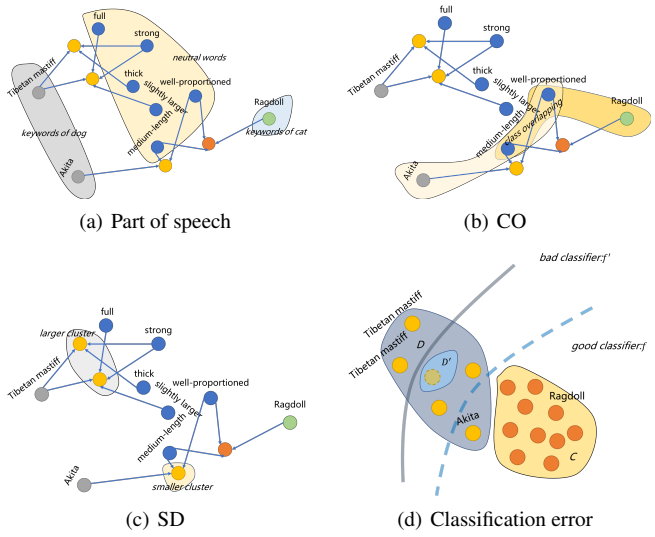
2 Imbalance Index of Texts

In this section, we formalize our imbalance index of the text. First, we give an example in Subsection 2.1 to illustrate the imbalance problem. Then, we formalize the imbalance problem in Subsection 2.2 and propose an imbalance index based on the trend line of ER changes with IR in Subsection 2.3.

2.1 Motivating Example

Suppose there is a task that classifies texts describing cat and dog:

- *The Tibetan mastiff has a strong body and thick limbs. (dog)*
- *The Ragdoll is well-proportioned and has medium-length limbs. (Cat)*



(a) Part of speech (b) CO

(c) SD (d) Classification error

Figure 1. The golden nodes, orange nodes, grey nodes, green nodes and blue nodes represent the embedding vectors of dog and cat's texts, the keywords of dog and cat, and neutral words. The embedding vectors of each texts are obtained by a combination of the embedding vectors of keywords and neutral words.

- *The Akita is well-proportioned and has medium-length limbs. (dog)*
- *The length of Tibetan mastiff is slightly larger than the body height, and the muscles are full and strong. (dog)*

As shown in Figure 1.(a), we divide words into keywords and neutral words. keywords are defined as words that are much more probable for one class than the other classes. Neutral words are defined as words that are equal probable for all classes. Since "*Tibetan mastiff*", "*Akita*" and "*Ragdoll*" can directly determine the categories of texts, and these words never appear in the texts of the opposite category, we define them as keywords. "*strong*", "*thick*", "*well-proportioned*", "*full*", "*slightly larger*" and "*medium length*" can be used not only to describe cat and dog but also to describe other animals, so they are the neutral words.

CO: In Figure 1.(b), the texts of "*Akita*" and "*Ragdoll*" overlap because the neutral words are the same. We difine the co-occurrence neutral words are the neutral word that coexists in two texts. For example, "*well-proportioned*" and "*medium-length*" are co-occurrence neutral words of "*Akita*" and "*Ragdoll*".

SD: In Figure 1.(c), samples of dog are overly concentrated in its sub-cluster. This issue makes the classifier hard to learn the smaller sub-cluster.

DIC: The above CO and SD belong to DIC. In Figure 1.(d), IR is equal to 2 and the classifier learns a poorly hyperplane $f^{'}$ due to the existence of CO and SD problems. However, if the $D^{'}$ region exists, IR at this time is equal to $\frac{5}{3}$. Using the $D^{'}$ region, due to the larger sub-cluster can connect the smaller ones, the CO and SD problems are improved, and the classifier learns a better hyperplane f.

In summary, IR does not consider the impact of DIC. This issue causes IR not to reflect the impact of the data set on the performance of the classifier.

2.2 Formalization of Class Imbalance

Given the training data $\mathrm{X} = [\mathrm{x}_1, ..., \mathrm{x}_n]^\top \in \mathbb{R}^{n\times m\times d}$ and their labels $\mathrm{Y} = [y_1, ..., y_n]^\top \in \{0,1\}^n$, where x_i is a text sample containing m words, n is the size of training data, d is the number of features regarding with one word. Among the binary data sets, the positive data are represented as $\mathrm{X}^+ = [\mathrm{x}_1^+, ..., \mathrm{x}_{n^+}^+]^\top \in \mathbb{R}^{n^+\times m\times d}$ with n^+ being the size of positive class, and the negative data are represented as $\mathrm{X}^- = [\mathrm{x}_1^-, ..., \mathrm{x}_{n^-}^-]^\top \in \mathbb{R}^{n^-\times m\times d}$ with n^- being the size of negative class. The imbalance problem refers to $n^+ \gg n^-$, which means that the size of positive class is much bigger than that of negative class. We also call X^+ as the majority class, and X^- as the minority class. Therefore, IR is defined as $\frac{n^+}{n^-}$.

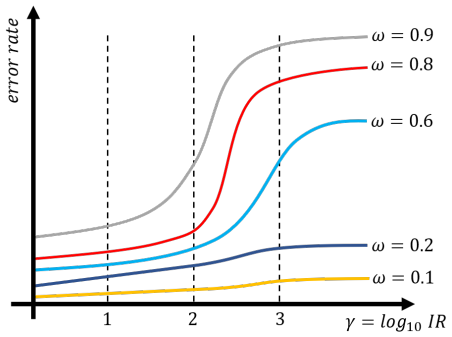


Figure 2. The trend line of ER with IR.

2.3 Proposed Imbalance Index: $\mathcal{G}_{\mathrm{IR}}$

As the degree of DIC increases, the impact of larger IR on the performance of classifier is greater. Therefore, we use DIC as a factor, namely ω, to describe the basic environment of $\mathcal{G}_{\mathrm{IR}}$. The trend of ER with IR is shown in Figure 2.

- With the decrease of ω, DIC decline, the trend line of ER is flat.
- With the increase of ω, DIC rise, and the trend line of ER is steep.
- When the IR is fixed, the higher the ω, the higher the ER.

The objective of $\mathcal{G}_{\mathrm{IR}}$ is to fit the trend line of ER under different ω. In the experiment, the arctangent function is the closest to the changing trend of ER compared with other functions. Therefore, $\mathcal{G}_{\mathrm{IR}}$ is defined as follows:

$$\mathcal{G}_{\mathrm{IR}} = \arctan(\omega\gamma) + \frac{\pi}{2}, \tag{1}$$

where γ is an imbalanced level. $\omega \in (0,1)$ is the score of the environmental factor. When the data is very complicated, ω is close to 1 and vice versa. We define ω as

$$\omega = f\big(\mathrm{D}(\mathrm{X}^-), \mathrm{sim}(\mathrm{X}^+, \mathrm{X}^-)\big), \tag{2}$$

where $\mathrm{sim}(\mathrm{X}^+, \mathrm{X}^-)$ is the degree of CO, and $\mathrm{D}(\mathrm{X}^-)$ is the degree of SD. f is a mean function used to trade-off $\mathrm{D}(\mathrm{X}^-)$ and $\mathrm{sim}(\mathrm{X}^+, \mathrm{X}^-)$. It can be arithmetic, geometric or quadratic means.

3 Graph-based Measurement for Text Imbalance Index

Characterizing the unknown distribution of data set is the necessary condition for calculating the degrees of CO and SD. Based on the inspiration from graph similarity problem [11], we transform the above data distribution into a graph model with structured information and propose a graph-based method to calculate the text imbalance index.

3.1 Formalization of Graph

Given an undirected, weighted and labelled graph $G = (V, E, \lambda, t)$, where V is a set of nodes connected by edges E, λ is a function that mapping a label value $\lambda(v)$ to each node $v \in V$ and t is a weight function of each node $v \in V$. For a binary text classification data set, $\lambda(v) \in \{P, M, N\}$, $\lambda(v) = P$ iff $v \in x_i^+$ and $v \notin x_j^-$; $\lambda(v) = N$ iff $v \in x_j^-$ and $v \notin x_i^+$; $\lambda(v) = M$ iff $v \in x_i^+$ and $v \in x_j^-$, where $i \in \{1, ..., n^+\}$ and $j \in \{1, ..., n^-\}$. A subset $V^+ \subseteq V$ induces a positive subgraph $G(V^+) = (V^+, E, \lambda^+)$ of G, where $\forall\lambda(v \mid v \in V^+) \in \{P, M\}$. A subset $V^- \subseteq V$ induces a negative subgraph $G(V^-) = (V^-, E, \lambda^-)$ of G, where $\forall\lambda(v \mid v \in V^-) \in \{N, M\}$. An example is shown in Figure 3, the text of "*Akita*" can be transformed into a subgraph ⟨"*Akita*" → "*well-proportioned*" → "*medium-length*"⟩, where "*Akita*" is the keyword of dog, the "*well-proportioned*" and "*medium-length*" are the neutral words.

3.2 Graph-based Measurement of CO

In our defined graph, CO problem can be described as the similarity problem of subgraphs from different classes, e.g. the texts of "*Akita*" and "*Ragdoll*", the category of the subgraph is mainly dominated by the keywords (i.e. "*Akita*" and "*Ragdoll*"), which also dominate that of above texts. Therefore, the degree of CO can be written as follows:

$$\begin{aligned} \mathrm{sim}(X^+, X^-) &= \sum_{i,j} \mathrm{sim}(\mathrm{x}_i^+, \mathrm{x}_j^-) \\ &= \sum_{i,j} \mathrm{sim}\big(G(V_i^+), G(V_j^-)\big) \approx \sum_{i,j} \mathrm{sim}(v^+, v^-), \\ \text{s.t. } \lambda(v^+) = P, \lambda(v^-) &= N \quad \forall\, v^+ \in V_i^+, \forall v^- \in V_j^-. \end{aligned} \tag{3}$$

Where $\approx$ means approximate equality. Inspired by information theory, the similarity between v^+ and v^- can be calculated based on their commonality and differences [7]. These intuitions as following: The similarity between v^+ and v^- is related to the amount of information they share. When more information is shared, their commonality is greater, and conversely, their difference is greater. Therefore, we define the similarity $\mathrm{sim}(v^+, v^-)$ as follows:

$$\mathrm{sim}(v^+, v^-) = \frac{\mathrm{common}(v^+, v^-)}{\mathrm{description}(v^+, v^-)}, \tag{4}$$

where $\mathrm{common}(\cdot)$ and $\mathrm{description}(\cdot)$ represent the commonality and difference, respectively.

Based on intuition 1, we define $\mathrm{common}(\cdot)$ as:

$$\begin{aligned} \mathrm{common}(v^+, v^-) = \underbrace{\mathrm{dis}(v^+, v^-)\big(t(v^+) + t(v^-)\big)}_{\star} \\ + \underbrace{\mathrm{ct\_com}(v^+, v^-)}_{\dagger}, \end{aligned} \tag{5}$$

Due to word often have different semantics in different contexts. Therefore, the similarity between v^+ and v^- includes individual similarity and contextual one. The individual similarity is the distance between the words in the data space, e.g. the Euclidean distance of the keywords of "*Akita*" and "*Ragdoll*" in data space. The contextual similarity is determined by the co-occurrence neutral words, which belong to both $G(V_i^+)$ and $G(V_j^-)$. The co-occurrence neutral words of "*Akita*" and "*Ragdoll*" (i.e. "*well-proportioned*" and "*medium-length*") are the same, which means that the semantics of

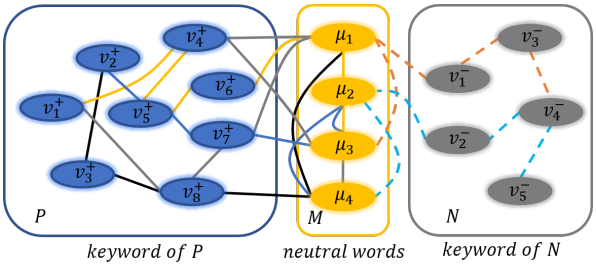


Figure 3. The undirected, weighted and labelled graph. Lines of the same color connect words that appear together in the same text. The solid line represents the text in the majority class and the dashed line represents the text of the minority class.

their texts are highly similar. In Equation (5), $\mathrm{dis}(\cdot)$ represents the degree of individual similarity between v^+ and v^-, e.g. Euclidean distance. The weight value $t(\cdot)$ indicates the amount of information that the node contains, e.g. TF-IDF. $\star$ considers the amount of common information existing in the pair of keywords (i.e. individual similarity). † considers the amount of information contained in the co-occurrence neutral words connected to the pair of keywords (i.e. contextual similarity). † is defined as follows:

$$\begin{aligned} \mathrm{ct\_com}(v^+, v^-) = \sum_{\mu \in (V_i^+ \cap V_j^-)} \big(\mathrm{dis}(\mu, v^+) \\ + \mathrm{dis}(\mu, v^-)\big) t(\mu), \end{aligned} \tag{6}$$

where μ is the co-occurrence neutral node (i.e., the node of the co-occurrence neutral word) in V_i^+ and V_j^-. For "*Akita*" and "*Ragdoll*", † calculates the amount of information that "*well-proportioned*" and "*medium-length*" provide to "*Akita*" and "*Ragdoll*", respectively.

Based on intuition 2, we define $\mathrm{description}(v^+, v^-)$ as:

$$\begin{aligned} \mathrm{description}(v^+, v^-) = \underbrace{t(v^+) + t(v^-)}_{\hat{\star}} \\ + \underbrace{\mathrm{ct\_dif}(v^+, v^-)}_{\hat{\dagger}}, \end{aligned} \tag{7}$$

$\hat{\star}$ considers the amount of information of each keyword, and $\hat{\dagger}$ is the contextual different, which considers the amount of information of neutral words connected to each keyword, e.g. "*Tibetan mastiff*" and "*Ragdoll*" do not have the co-occurrence neutral words, so $\hat{\dagger}$ calculates the amount of information that "*strong*" and "*thick*" provide to "*Tibetan mastiff*", and the amount of information that "*well-proportioned* " and "*medium-length* " provide to "*Ragdoll*". $\hat{\dagger}$ is defined as follows:

$$\begin{aligned} \mathrm{ct\_dif}(v^+, v^-) = \sum_{\hat{\mu} \in V^+} \mathrm{dis}(v^+, \hat{\mu}) t(\hat{\mu}) \\ + \sum_{\check{\mu} \in V^-} \mathrm{dis}(v^-, \check{\mu}) t(\check{\mu}). \end{aligned} \tag{8}$$

When v^+ and v^- are the same node and $\mathrm{dis}(\cdot) \in [0, 1]$ is the normalized space distance. We conclude $\mathrm{dis}(v^+, v^-) = 1$, and $\mu = \hat{\mu} = \check{\mu}$. Then $\mathrm{ct\_com}(v^+, v^-) = \mathrm{ct\_dif}(v^+, v^-)$. Therefore, the maximum of $\mathrm{sim}(\mathrm{X}^+, \mathrm{X}^-) = 1$. Conversely, $\mathrm{sim}(\mathrm{X}^+, \mathrm{X}^-) = 0$.

3.3 Graph-based Measurement of SD

In our graph, SD can be transformed into the problem of over-concentration of keyword weight. When $t(v \mid \lambda(v) \neq M)$ is significantly greater than others, it means that the same polarity subgraphs containing v is over-represented by v, e.g. the dog's samples are excessively skewed to the "*Tibetan mastiff*", so the frequency of the keyword of "*Tibetan mastiff*" appears relatively high. The small disjuncts factor is defined as

$$\begin{aligned} \mathrm{D}(\mathrm{X}^-) \propto \mathrm{D}\big(V^- \mid \lambda(V^-) \neq M\big) \\ = \sum_m |t(\tilde{v}_m^-) - t(\tilde{v}_{m+1}^-)|, \\ \text{s.t. } \tilde{v}_m^- \in \mathrm{sorted}\big(V^- \mid \lambda(V^-) \neq M\big), \end{aligned} \quad (9)$$

where $\tilde{v}_m^-$ is the node in the sorted V^-, which does not contain the neutral node. Due to the normalization of $t\big(V^- \mid \lambda(V^-) \neq M\big)$, the maximum of $\mathrm{D}(\mathrm{X}^-)$ is 1 when $t(\tilde{v}_m^-)$ is equal to 1 and the weight of other nodes is 0. On the contrary, if all $\tilde{v}_m^-$ get the same weight then $\mathrm{D}(\mathrm{X}^-)$ is 0.

3.4 Fine-grained Text Imbalance Level

Unlike other types of data, such as images, etc. Text classification data is a set of texts consisting of multiple words. If the number of keywords in a text is small, its polarity is difficult to determine by the classifier. Therefore, we define a more fine-grained imbalance level γ that considers the number of inter-class keywords and the contribution of each text for training the classifier. In our graph model, We first calculate the contribution of each subgraph.

$$\begin{aligned} r(V_i^+) &= \sum_{p=1}^{n_i^+} \frac{t\big(v_p^+ \mid \lambda(v_p^+) = P, v_p^+ \in V_i^+\big)}{t(v_p^+ \mid v_p^+ \in V_i^+)}, \\ r(V_j^-) &= \sum_{p=1}^{n_j^-} \frac{t\big(v_p^- \mid \lambda(v_p^-) = N, v_p^- \in V_j^-\big)}{t(v_p^- \mid v_p^- \in V_j^-)}, \end{aligned} \quad (10)$$

where n_i^+ and n_j^- are the number of nodes in the positive and negative subgraphs, respectively. Therefore, the imbalance factor can be defined as follows:

$$\gamma = \log_{10}\left(\frac{\sum_i r(V_i^+)}{\sum_j r(V_j^-)}\right). \quad (11)$$

The reason we use $\log_{10}(\cdot)$ is that if the value of the imbalance factor is too large, it will cause the arctangent function to enter a smooth region.

Finally, our method is summarized in Algorithm 1.

4 Experiments and Evaluation

To evaluate the correlation between the imbalance index $\mathcal{G}_{\mathrm{IR}}$ and the classification performance, we introduce two steps: (1) we conduct correlation analysis experiments on synthetic data sets and real-world data sets; and (2) we use $\mathcal{G}_{\mathrm{IR}}$ to evaluate the performance of existing imbalance learning methods.

Algorithm 1 The $\mathcal{G}_{\mathrm{IR}}$ algorithm

Input: $Dataset = \{\mathrm{X}^+, \mathrm{X}^-, \mathrm{Y}\}$
Arg: displacement hyper-parameter β
Output: $\mathcal{G}_{\mathrm{IR}}$ score

1: Convert $Dataset$ to graph $G = (V, E, \lambda, t)$
2: Compute the degree of CO $\mathrm{sim}(\mathrm{X}^+, \mathrm{X}^-)$ between X^+ and X^- with Eq.3
3: Compute the degree of SD $\mathrm{D}(\mathrm{X}^-)$ with Eq.9
4: Compute the imbalance factor γ between X^+ and X^- with Eq.11
5: Bring $\mathrm{sim}(\mathrm{X}^+, \mathrm{X}^-)$ and $\mathrm{D}(\mathrm{X}^-)$ into Eq.2 to compute the environmental factor ω
6: Bring ω and γ into Eq.1 to compute $\mathcal{G}_{\mathrm{IR}}$
7: **return** $\mathcal{G}_{\mathrm{IR}}$

4.1 Correlation Analysis Method

The Spearman's rank correlation coefficient ρ is a non-parametric index to measure the dependence between two variables a and b, where a and b have the same sample size n. ρ is defined as follows:

$$\rho = \frac{\sum_i (a_i - \bar{a})(b_i - \bar{b})}{\sqrt{\sum_i (a_i - \bar{a})^2 \sum_i (b_i - \bar{b})^2}}. \quad (12)$$

With the increase of the monotonic correlation between a and b, $|\rho|$ is close to 1. In our experiments, the two variables are $\mathcal{G}_{\mathrm{IR}}$ and ER of the minority class.

4.2 Metrics of ER

We select $F1$ to measure the ER of the majority class, assuming that the real label is y and the predicted label is $\hat{y}$, therefore, the numbers of true positive (TP), false positive (FP), false negative (FN) and true negative (TN) samples are formally defined as follows:

$$\begin{aligned} TP &= \#\{x_i \mid y_i = +1 \wedge \hat{y}_i = +1\}, \\ FP &= \#\{x_i \mid y_i = -1 \wedge \hat{y}_i = +1\}, \\ FN &= \#\{x_i \mid y_i = +1 \wedge \hat{y}_i = -1\}, \\ TN &= \#\{x_i \mid y_i = -1 \wedge \hat{y}_i = -1\}, \\ &\text{s.t. } i = 1, ..., m. \end{aligned} \quad (13)$$

Here "$\#$" represents the size of the set, m is the size of the minority class. Thus $F1$ is defined as follows:

$$F1 = \frac{2 \cdot TP}{2 \cdot TP + FN + FP}. \quad (14)$$

Therefore, the ER of the minority class is 1-$F1$.

4.3 Compared Imbalance Indexes

We conduct comparative experiments with IR and BI^3, respectively.

- **IR** is traditional imbalanced ratio, and IR can intuitively compare the sizes of majority class and minority class. Thus IR can be further used to measure its correlation with all machine learning methods.
- ***BI*^{3}** uses the KNN algorithm to estimate the distribution of the input data. The inputs of the KNN algorithm are the feature vectors of texts, not that of words. The inputs of the NN-based model are the feature vectors of words. Therefore, BI^3 cannot be used to measure its correlation with the NN-based model.

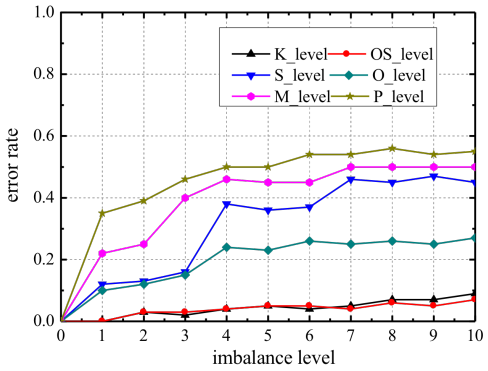


Figure 4. The trend line of ER with imbalance level.

4.4 Text Classification Methods and Settings

Four traditional machine learning algorithms (**SVM**, **KNN**, and **Adaboost**) are selected to perform the comparative experiment of BI^3. All of these algorithms are derived from the *scikit-learn* library[4] and their parameters are all official defaults.

The above four traditional machine learning models and three well-known basic NN-based models are used to compare with IR. The parameters and model structures are as follows. We repeat each of the experiments 10 times, and the mean of 10-folds cross-validation obtains each of its results.

- **CNN**: The number of channel and kernel_size are set by 100 and 1, respectively. To analysis the basic CNN model, we set the number of hidden layer to 1 [33].
- **RNN**: We adopt the binary-directional LSTM model. The number of hidden layers in the model is the same as CNN and is set to 1 [22].
- **Self-attention**: We use the encoder module, which has 6 encoder layers, 8 scaled dot-product attention modules and one layer of feedforward NN [4].

4.5 Experiments on Synthetic Data

To evaluate the sensitivity of the $\mathcal{G}_{\mathrm{IR}}$ to the single DIC that can cause deterioration in classification performance. We construct six levels of 3-dimensional data sets. These data sets contain **P**erfect (P_level), **M**ulti_distributed (M_level), **OC** (O_level), **SD** (S_level) and **OC+SD** (OS_level) and **K**eyword_insufficient (K_level) factors resp.

4.5.1 *Synthetic Method*

Each data set consists of two classes. The sample in each class is a text, which contains keywords and neutral words. In the binary classification problem, keywords can be divided into positive K_p and negative K_n clusters. K_p, K_n and neutral words K_m are sampled from different distributions, each of which contains two different normal ones $\{N_1(\mu_1^\top, \sigma_1^\top) \bigcup N_2(\mu_2^\top, \sigma_2^\top)\}$. The size of K_p and K_m is fixed to 5000 on each level data set, the size of K_n varies in the set $\{50, 100, 1000, 3000, 5000\}$ on S_level data set for analysing the SD problem and be fixed to 5000 on each other. A larger K_n can better express the distribution of its data space. To construct imbalance, the number of majority class is fixed to 5000, and the number of minority class varies in the set $\{\frac{5000}{2^n}, n = 1, ...10\}$. From Figure 5. The details of each data set are described below.

[4] Scikit-learn website: https://scikit-learn.org/stable/.

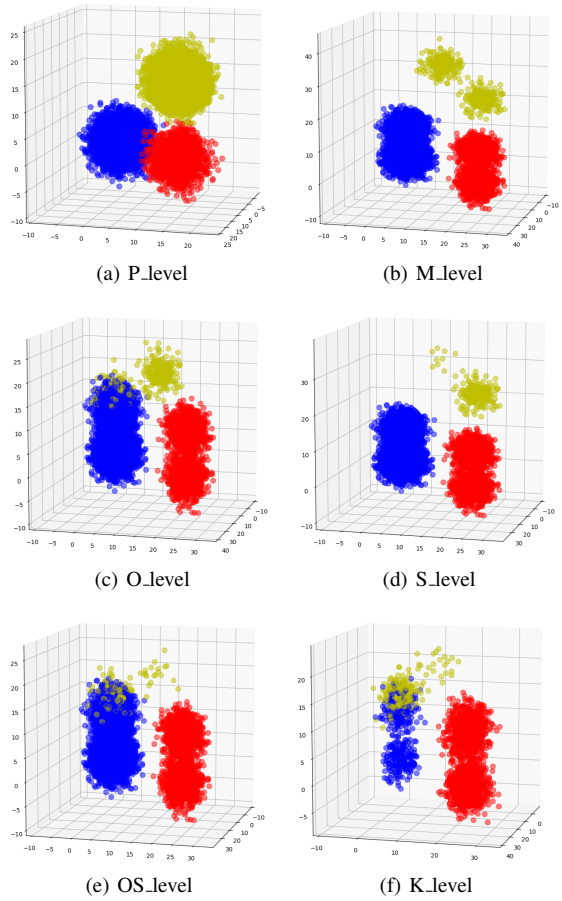


Figure 5. 3D visualization of synthetic data sets. Corresponding to Figure 2, the blue dot indicates the keyword of P, the red dot indicates the keyword of N, and the green dot indicates the neural word.

- **P_level**: No CO and SD in the distributions of K_p, K_n and K_m. i.e. $\mid \mu_1^\top - \mu_2^\top \mid = 1^\top$ and $\sigma_1^\top = \sigma_2^\top = 5^\top$.
- **M_level**: K_p, K_n and K_m have a much wider distribution than they in P_level. Each distribution consists of two clusters. i.e. $\mid \mu_1^\top - \mu_2^\top \mid = 10^\top$.
- **O_level**: The distribution of K_p and K_n overlapped partially. i.e. $\mu_1^\top(K_p^1) = \mu_2^\top(K_n^2)$ where $K_\diamond^i$ refers to the i subcluster of $K_\diamond$.
- **S_level**: K_n are concentrated in one cluster, while the other one is too sparse. As the size of K_n decreases, the SD problem has a high probability of occurring. i.e. $\frac{\#K_p^1}{\#K_n^2} = 0.001$ where $\#$ means the size of $K_\diamond^i$.
- **OS_level**: OS_level is the combination of O_level and S_level. i.e. $\mu_1^\top(K_p^1) = \mu_2^\top(K_n^2)$ and $\frac{\#K_n^1}{\#K_n^2} = 0.001$.
- **K_level**: On the basis of OS_level, the number of keywords in all texts is obviously less than that of neutral words, i.e., $\frac{\#keyword}{\#neutral\ word} = 0.02$ in all texts.

4.5.2 *Result and Analysis*

The results of correlation analysis on synthetic data sets are presented in Table 1.

- In the case of P_level and M_level in Figure 4, ERs of all methods are less than 5% and without obvious rise. Therefore, when inter-class keywords have a higher separability, even if the data set is

imbalanced, it would not have a worse impact on the classifier. This situation can be described by BI^3 and $\mathcal{G}_{\mathrm{IR}}$, except for IR.

- In the case of O_level in Figure 4, ERs of all methods increase gradually. It shows that the classification performance is affected by IR. Because BI^3 is based on the KNN algorithm, which can capture the area of CO, BI^3 performs very well.
- In the case of S_level, unlike the KNN algorithm, $\mathcal{G}_{\mathrm{IR}}$'s environmental factors ω can capture SD problems. Therefore, $\mathcal{G}_{\mathrm{IR}}$ achieves the best score.
- OS_level is a combination of O_level and S_level, so it is more complex than them. Because it contains overlapping problems, BI^3 has better scores than which on S_level. However, due to the small disjuncts problem, $\mathcal{G}_{\mathrm{IR}}$ still gets the best score.
- K_level takes into account the proportion of neutral words and keywords in texts. We find that this is a crucial reason that affects the performance of classifiers. Especially for the NN-based model, when there are enough keywords in texts, even if there is a serious overlapping problem, the classification performance is still very high.

Table 1. Correlation analysis of synthetic data sets.

	classifier	IR	BI^3	$\mathcal{G}_{\mathrm{IR}}$
P_level	SVM	0.1032	0.4163	**0.5219**
	KNN	0.0911	**0.6127**	0.5901
	Adaboost	0.1321	0.4199	**0.4231**
	CNN	0.1073	–	**0.3261**
	RNN	0.1002	–	**0.3210**
	Self-attention	0.0894	–	**0.3288**
M_level	SVM	0.1453	0.4201	**0.5313**
	KNN	0.1322	**0.5299**	0.5127
	Adaboost	0.1532	0.4147	**0.5275**
	CNN	0.1196	–	**0.8222**
	RNN	0.1175	–	**0.8318**
	Self-attention	0.1004	–	**0.8240**
O_level	SVM	0.6800	0.8005	**0.8014**
	KNN	0.6275	**0.8145**	0.8132
	Adaboost	0.7079	0.8111	**0.8195**
	CNN	0.3011	–	**0.8066**
	RNN	0.2906	–	**0.8137**
	Self-attention	0.2727	–	**0.7901**
S_level	SVM	0.2111	0.6557	**0.7281**
	KNN	0.2050	0.6722	**0.7304**
	Adaboost	0.2397	0.6316	**0.7026**
	CNN	0.1983	–	**0.7309**
	RNN	0.1744	–	**0.7238**
	Self-attention	0.1639	–	**0.7066**
OS_level	SVM	0.5437	0.6988	**0.7131**
	KNN	0.5201	0.7094	**0.7190**
	Adaboost	0.5513	0.7064	**0.7122**
	CNN	0.5101	–	**0.7047**
	RNN	0.5081	–	**0.7203**
	Self-attention	0.5033	–	**0.7158**
K_level	SVM	0.5190	0.7210	**0.7305**
	KNN	0.5055	**0.7282**	0.7234
	Adaboost	0.5222	0.7192	**0.7267**
	CNN	0.5852	–	**0.7351**
	RNN	0.5840	–	**0.7116**
	Self-attention	0.5639	–	**0.7032**

4.6 Experiments on Benchmark Data

We adopt four data sets to cover as many scenarios as possible, such as overlapping, small disjuncts and different degrees of imbalance level [2][5]. The details of the data sets are shown in Table 2.

Table 2. Details of benchmark data sets.

	topic	*#train*	*#test*	IR
20 News	rec.sport.hockey	600	399	–
	rec.sport.baseball	597	397	1.005
	misc.forsale	585	390	1.026
	comp.os.ms-windows.misc	572	394	1.049
	talk.politics.misc	465	310	1.299
	talk.religion.misc	377	251	1.592
R8	earn	2840	1083	–
	acq	1596	696	1.780
	crude	253	121	11.225
	trade	251	75	11.315
	money-fx	206	87	13.786
	interest	190	81	14.947
	ship	108	36	26.296
	grain	41	10	69.268
Cade12	servicos	5627	2846	–
	sociedade	4935	2428	1.140
	lazer	3698	1892	1.522
	internet	1585	796	3.550
	noticias	701	381	8.027
	compras-online	423	202	13.303
WebKB	student	1097	544	–
	faculty	750	374	1.463
	course	620	310	1.769
	project	336	168	3.265

- **20 News**: 20_news is a relatively balanced data set, but topics are similar, such as *comp.os.ms-windows.misc*, *talk.politics.misc* and *talk.religion.misc*. It makes classification more difficult because they are likely to have an overlapping problem.
- **R8**: Reuters-21578 is an imbalance data set about news, and the *grain* topic only has 41 samples, so it is prone to have a small disjuncts problem.
- **Cade12**: Compared with R8, Cade12 has much larger topic sizes, and the IR changes more smoothly.
- **WebKB**: Compared with other data sets, WebKB has the fewest number of topics, and the topics are not very large.

4.6.1 Result and Analysis

The results of correlation analysis on benchmark data sets are presented in Table 3.

- $\mathcal{G}_{\mathrm{IR}}$ achieves optimal or very close to optimal results on each data set. Especially on the 20 news data set, $\mathcal{G}_{\mathrm{IR}}$, containing CO factor, obtains the best correlation on different classifiers. It illustrates that $\mathcal{G}_{\mathrm{IR}}$ can still capture the fatal factor affecting the performance degradation of the classifier when the IR does not change much.
- Due to BI^3 uses the k-nearest neighbour algorithm to estimate the distribution of input data, it is suitable for the KNN classifier and achieves optimal or sub-optimal results.
- The NN-based model has a lower ER than the traditional machine learning algorithm, and $\mathcal{G}_{\mathrm{IR}}$ still has a high correlation. Among them, the score for the self-attention model is the smallest, which also indicates that the classification performance of it is the state-of-the-art level.

4.7 Evaluation of Imbalance Learning Methods

Random **O**versampling (OS), **R**andom **U**ndersampling (US), **S**ynthetic **M**inority **O**versampling **T**echnique (SMOTE) and **S**ampling **W**eighting (SW) are performed to improve the benchmark

[5] Data sets website: http://ana.cachopo.org.

Table 3. Correlation analysis of benchmark data sets.

	classifier	IR	BI^3	$\mathcal{G}_{\mathrm{IR}}$
20 News	SVM	0.0751	0.1421	**0.2456**
	KNN	0.3136	0.3259	**0.3348**
	Adaboost	0.0299	0.1641	**0.2789**
	CNN	0.0736	–	**0.2593**
	RNN	0.1096	–	**0.2125**
	Self-attention	-0.1862	–	**0.1230**
R8	SVM	0.5714	0.8928	**0.9285**
	KNN	0.4286	**0.8214**	0.8176
	Adaboost	0.4286	0.8571	**0.8626**
	CNN	0.3806	–	**0.4355**
	RNN	0.3913	–	**0.5230**
	Self-attention	0.3214	–	**0.3701**
Cade12	SVM	0.6187	0.6746	**0.6875**
	KNN	0.9000	0.9182	**0.9182**
	Adaboost	0.6636	0.6818	**0.6912**
	CNN	0.5593	–	**0.5761**
	RNN	0.5313	–	**0.5532**
	Self-attention	0.4909	–	**0.5333**
WebKB	SVM	0.4178	**0.8969**	0.8813
	KNN	0.4454	**0.8988**	0.8795
	Adaboost	0.4187	0.8870	**0.8902**
	CNN	0.2936	–	**0.4249**
	RNN	0.3127	–	**0.4507**
	Self-attention	0.2500	–	**0.3258**

data sets and we conduct correlation experiments between the classification performance and the imbalance index. These imbalance learning methods are implemented by $imbalanced\text{-}learn$ toolbox[6].

4.7.1 Result and Analysis

The evaluation results of imbalance learning methods are presented in Table 4. The experimental results show that $\mathcal{G}_{\mathrm{IR}}$ achieves a high correlation score.

- OS and US have equal effects on the improvement of the data set since they are the basic addition and deletion operations of the data set. However, they are unable to alleviate the DIC problem, e.g. although US may reduce the amount of data in OC areas, US deteriorates the SD problem and OS has the opposite effect.
- SMOTE and OS are also oversampling methods. However, it does not resample data repeatedly, which makes the data set to get better training data possibly. Therefore, it has a useful data set recovery effect.
- The score of SW is not high or low, because SW is an algorithm-level method. Only when the distribution between the training set and testing one is inconsistent, SW is significantly affected. Therefore, SW has no high correlation with IR.

5 Related Works

Existing researches on imbalance problem are mainly around IR [3, 15, 19]. IR is an extrinsic characteristic of data that does not consider the distribution of data. Most standard imbalanced learning methods assumed that the deterioration of classification performance is caused by IR, and can be roughly divided into data-level resampling methods [9, 20, 34, 28] and algorithm-level cost-sensitive learning methods [18, 35].

[6] Imbalanced-learn website: http://imbalanced-learn.org.

Table 4. Evaluation of imbalance learning methods.

	data set	IR	BI^3	$\mathcal{G}_{\mathrm{IR}}$
OS	20 News	0.0765	0.1913	**0.2344**
	R8	0.1924	0.7259	**0.7335**
	Cade12	0.2719	0.7283	**0.7529**
	WebKB	0.2203	0.6021	**0.6188**
US	20 News	0.0264	0.1482	**0.2923**
	R8	0.1376	0.7066	**0.7192**
	Cade12	0.2457	0.7200	**0.7221**
	WebKB	0.2342	**0.5862**	0.5700
SMOTE	20 News	0.1221	0.3109	**0.3318**
	R8	0.2948	0.7218	**0.7626**
	Cade12	0.2161	0.7350	**0.7411**
	WebKB	0.3535	0.6324	**0.6515**
SW	20 News	0.0588	0.1517	**0.2856**
	R8	0.1963	0.6066	**0.6104**
	Cade12	0.2313	0.6375	**0.6983**
	WebKB	0.2561	0.5921	**0.6470**

However, some theoretical studies have recently pointed out that IR is more sensitive to DIC [14, 29, 30, 21, 6]. In more detail, if the DIC of the data is low, higher IR does not necessarily affect the performance of the classifier, and vice versa [13]. Some of these studies focused on the assessment of DC areas of input space [37], and the R-value-based metric that estimates the degree of DC with the k neighbors of each instance in a given class [27], the others focused on intra-class imbalances, where the training data of sub-cluster is severely missing due to IR [23]. Smaller cluster aggregates a large number of classification errors [10]. This is because class imbalance results in smaller disjunct fails to represent sub-concepts [31]. Research on DIC requires exploring the distribution of data that cannot be visually described. Some studies cleverly avoid direct exploration of data distribution, using Bayesian optimal classifiers to study the effects of imbalance data from a theoretical perspective [32].

In this paper, we use a graph-based imbalance index to investigate the imbalance of text. This index uses a novel approach to fuse DIC and IR to more fully reflect the characteristics of imbalanced data.

6 Conclusion

In this paper, we present a novel text imbalance index $\mathcal{G}_{\mathrm{IR}}$ using an environmental factor to characterize CO and SD as two important DIC to predict the impact of data on classification performance. Moreover, our proposed graph-based imbalance measurement built on $\mathcal{G}_{\mathrm{IR}}$ can figure out the unknown distribution of real-world imbalanced text well. In this sense, our approach provides an alternative way to analyze DIC of the imbalanced text, and we believe that our results would be helpful to optimize text imbalanced classifiers even NN-based models.

7 Acknowledgments

This work is supported by the National Key Research and Development Program of China (2017YFC0908401), the National Natural Science Foundation of China (61972455,61672377), and Shenzhen Science and Technology Foundation (JCYJ20170816093943197). Xiaowang Zhang is supported by the Peiyang Young Scholars in Tianjin University (2019XRX-0032).

REFERENCES

[1] A. Fernández, S. García, M. Galar, R. C. Prati, B. Krawczyk, and F. Herrera, *Learning from imbalanced data sets*, Reading, Mass.: Springer, (2018).
[2] A. M. G. C. Cachopo, Improving methods for single-label text categorization, Ph.D.thesis, Universidade Técnica de Lisboa, (2007).
[3] A. Sen, M. M. Islam, K. Murase, and X. Yao, Binarization with boosting and oversampling for multiclass classification, *IEEE Trans. Cybernetics*, 46(5):1078–1091, (2016).
[4] A. Vaswani, N. Bengio, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and J. Uszkoreit, Attention is all you need, In *Proc. of NIPS*, 6000–6010, (2017).
[5] C. Bellinger, C. Drummond, and N. Japkowicz, Manifold-based synthetic oversampling with manifold conformance estimation, *Mach. Learn.*, 107(3):605–637, (2018).
[6] C. H. Lim, and S. J. Wright, K-support and ordered weighted sparsity for overlapping groups: hardness and algorithms, In *Proc. of NIPS*, 284–292, (2017).
[7] D. Lin, An information-theoretic definition of similarity, In *Proc. of ICML*, 296–304, (1998).
[8] F. Wu, C. Wu, and J. Liu, Imbalanced sentiment classification with multi-task learning, In *Proc. of CIKM*, 1631–1634, (2018).
[9] G. Lemaître, F. Nogueira, and C. K. Aridas, Imbalanced-learn: a python toolbox to tackle the curse of imbalanced datasets in machine learning, *J. Mach. Learn. Res.*, 18(1):559–563, (2017).
[10] G. M. Weiss, Learning with rare cases and small disjuncts, In *Proc. of ICML*, 558–565, (1995).
[11] G. Nikolentzos, P. Meladianos, and M. Vazirgiannis, Matching Node Embeddings for Graph Similarity, In *Proc. of AAAI*, 2429–2435, (2017).
[12] H. He, and E. A. Garcia, Learning from imbalanced data, *IEEE Trans. Knowl. Data Eng.*, 21(9):1263–1284, (2009).
[13] J. Luengo, A. Fernández, S. García, and F. Herrera, Addressing data complexity for imbalanced data sets: analysis of smote-based oversampling and evolutionary undersampling, *Soft Comput.*, 15(10):1909–1936, (2011).
[14] K. Napierala, and J. Stefanowski, Types of minority class examples and their influence on learning classifiers from imbalanced data, *J. Intell. Inf. Syst.*, 46(3):563–597, (2016).
[15] L. Abdi, and S. Hashemi, To combat multi-class imbalanced problems by means of over-sampling techniques, *IEEE Trans. Knowl. Data Eng.*, 28(1):238–251, (2016).
[16] L. Yang, Y. -N. Guo, and J. Cheng, Manifold distance-based over-sampling technique for class imbalance learning, In *Proc. of AAAI*, 10071–10072, (2019).
[17] M. Alejandro, E. Andrea, and S. Fabrizio, Distributional random oversampling for imbalanced text classification, In *Proc. of SIGIR*, 805–808, (2016).
[18] M. Liu, C. Xu, Y. Luo, C. Xu, Y. Wen, and D. Tao, Cost-sensitive feature selection via F-Measure optimization reduction, In *Proc. of AAAI*, 2252–2258, (2017).
[19] M. Pérez-Ortiz, P. A. Gutiérrez, P. Tino, and C. Hervás-Martínez, Oversampling the minority class in the feature space, *IEEE Trans. Neural Netw. Learning Syst.*, 27(9):1947–1961, (2016).
[20] M. Peng, Q. Zhang, X. Xing, T. Gui, X. Huang, Y. -G. Jiang, K. Ding, and Z. Chen, Trainable undersampling for class-imbalance learning, In *Proc. of AAAI*, 4707–4714, (2019).
[21] M. R. Smith, T. R. Martinez, and C. G. Giraud-Carrier, An instance level analysis of data complexity, *Mach. Learn.*, 95(2):225–256, (2014).
[22] M. Sundermeyer, R. Schlüter, and H. Ney, LSTM neural networks for language modeling, In *Proc. of ISCA*, 194–197, (2012).
[23] R. C. Holte, L. Acker, and B. W. Porter, Concept learning and the problem of small disjuncts, In *Proc. of IJCAI*, 813–818, (1989).
[24] R. C. Prati, G. E. A. P. A. Batista, and M. C. Monard, Class imbalances versus class overlapping: an analysis of a learning system behavior, In *Proc. of MICAI*, 312–321, (2004).
[25] R. McBride, K. Wang, Z. Ren, and W. Li, Cost-sensitive learning to rank, In *Proc. of AAAI*, 4570–4577, (2019).
[26] S. Liu, Y. Wang, J. Zhang, C. Chen, and Y. Xiang, Addressing the class imbalance problem in twitter spam detection using ensemble learning, *Computers & Security*, 69:35–49, (2017).
[27] S. Oh, A new dataset evaluation method based on category overlap, *Comput. Biol. Med.*, 41(2):115–122, (2011).
[28] T. Guo, X. Zhu, Y. Wang, and F. Chen, Discriminative sample generation for deep imbalanced learning, In *Proc. of IJCAI*, 2406–2412, (2019).
[29] T. K. Ho, and M. Basu, Complexity measures of supervised classification problems, *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(3):289–300, (2002).
[30] V. García, R. A. Mollineda, and J. S. Sánchez, On the k-NN performance in a challenging scenario of imbalance and overlapping, *Pattern Anal. Appl.*, 11(3–4):26–280, (2008).
[31] V. López, A. Fernández, S. García, V. Palade, and F. Herrera, An insight into classification with imbalanced data: empirical results and current trends on using data intrinsic characteristics, *Inf. Sci.*, 250:113–141, (2015).
[32] Y. Lu, Y. -M. Cheung, and Y. -Y. Tang, Bayes imbalance impact index: a measure of class imbalanced dataset for classification problem, *arXiv*, CoRR abs/1901.10173, (2019).
[33] Y. Kim, Convolutional neural networks for sentence classification, In *Proc. of EMNLP*, 1746–1751, (2014).
[34] Y. Yan, M. Tan, Y. Xu, J. Cao, M. Ng, H. Min, and Q. Wu, Oversampling for imbalanced data via optimal transport, In *Proc. of AAAI*, 5605–5612, (2019).
[35] Y. Zhang, P. Zhao, J. Cao, W. Ma, J. Huang, Q. Wu, and M. Tan, Online adaptive asymmetric active learning for budgeted imbalanced data, In *Proc. of SIGKDD*, 2768–2777, (2018).
[36] Y. Zuo, J. Zhao, and K. Xu, Word network topic model: a simple but general solution for short and imbalanced texts, *arXiv*, CoRR abs/1412.5404, (2014).
[37] Z. Borsos, C. Lemnaru, and R. Potolea, Dealing with overlap and imbalance: a new metric and approach, *Pattern Anal. Appl.*, 21(2):381–395, (2018).