An Ensemble Classification Algorithm for Short Text Data Stream with Concept Drifts

Gang Sun, Zhongxin Wang*, Zhengqi Ding, and Jia Zhao

Abstract—With the rapid development of information technology and instant messaging, many short text data streams are continuously generated. Due to short texts’ inherent features such as sparse features, weak concept description, and concept drifts, most of the existing data stream classification methods are difficult to adapt and decrease performance. In this paper, we propose a new short text data stream classification algorithm to handle the problem. Specifically, we first select the distinguishing features from short texts to form the feature space and use the similar features in the feature space to expand. Then, a concept drift detection method based on topic distribution is used to detect and adapt to the concept drifts in this short text data stream. Finally, the ensemble classification model is updated in accordance with whether concept drift happening or not. The experimental results show that the classification performance of the short text data stream algorithm proposed in the paper is better than other algorithms.

Index Terms—Short text data stream, ensemble classification algorithm, feature extension, concept drift

W

I. INTRODUCE

ITH the rapid development of information technology and instant messaging, network users and servers generate a large amount of short text data at all times, such as Weibo, online comments, instant messages. Short text data streams play a vital role in user intent understanding, question-answering systems, and intelligent information retrieval. Short text data streams have three characteristics: sparsity, weak feature expression, and concept drifts. Due to the three characteristics mentioned above, it is not easy to apply traditional data stream classification methods to short text data streams. Traditional algorithms pay rare attention to the characteristics of the short text data stream and the concept drift hidden in the data stream at the same time [1-4]. Therefore, in accordance with the features of short text data streams, this paper proposes a new short text data stream classification algorithm to handle the corresponding problems. First, the algorithm selects the distinguishing features from short texts to form the feature space and use the similar features in the feature space to expand. Then, in order to effectively detect and adapt to concept drifts in the short text data stream, this paper uses a concept drift detection method based on topic distribution. It can determine whether the concept drift occurs in the new data block, through accumulating the semantic distance between each short text in the new and present data block. Finally, according to whether concept drift happening or not, update the ensemble classification model will be updated.

The main contributions of this paper are as follows:

1) First of all, this paper proposes a short text feature extension method based on the features of short texts. This method selects the distinguishing features from the short texts themselves to form the feature space and uses the similar features to expand the space.

2) Secondly, this paper proposes a concept drift detection method based on topic distribution, so as to detect the concept drifts hidden in the short text data stream. By accumulating the semantic distance between each short text in the new and present data block, it is determined whether the concept drift occurs in the new data block.

3) Finally, compared with most of existing data stream classification methods and short text classification methods, the ensemble classification model proposed in this paper can effectively detect the concept drifts hidden in the short text data stream and has better classification performance than other classification models.

II. RELATED WORKS

Short text feature extension methods are mainly divided into two categories: the first category is an extension method based on its resources, and the second one is an extension method based on external resources. Based on its resources, the feature extension method mainly uses the text’s context or semantics to perform feature extension on short texts. Yuan et al. [5] improved the naive Bayes algorithm to adapt and deal with short text’s sparse feature. Wang et al. [6] proposed a new short text method based on semantic clustering and convolutional neural networks. Gao et al. [7] used a structured sparse representation classifier to classify short text data effectively. Haddoud et al. [8] proposed more than 80 metrics for the weighting problem of vocabulary labeling. Bicalho et al. [9] presented a topic classification model, which uses the words “co-occurrence” and “vectors” to create an important document in the original document. Doulamis et al. [10] used a fuzzy time feature sequence to detect similar and related words in Weibo events. Fislar et al. [11] showed an improving short text classification using information from DBpedia ontology.
At present, there are few researches on short text data stream classification. Bouaziz et al. [12] introduced the IGLM model, which improves data stream classification by continuously updating the classifier. Ren et al. [13] proposed hierarchical multi-label short text data stream classification, which uses a block-based structure optimization strategy to classify short texts. Li et al. [14] proposed an incremental ensemble model to adapt to the short text data stream, which compensates for data sparsity by introducing more semantic contextual information hidden in a short text.

There have been also some researches on concept drift detection methods. Gmama et al. [15] proposed a method to detect the change of sample probability distribution. Kurlej et al. [16] proposed an active learning method, which detects concept drift by calculating the closest distance. Sun et al. [17] came up with the financial distress concept drift method, which creates a dynamic FDP model to select and predict instances. Loeffel et al. [20] proposed an online algorithm that deals with various types of concept drifts and did not use a fixed threshold. Chen et al. [21] proposed to use a genetic algorithm to obtain fuzzy concept drift mode. Mirza et al. [22] proposed the ESOS-ELM method, which can handle unbalanced data streams. Sethi and Kantardzic et al. [23] proposed the MD3 method, a drift detection algorithm for unmarked stream data. Tennant et al. [24] used the minimum cluster neighbor method to calculate statistical data. Silva et al. [25] proposed the FEAC-Stream algorithm, which uses k-means clustering and automatically estimates the value of k from the stream. Deng et al. [26] proposed a concept drift detection method based on rough set and granular computing.

The algorithms mentioned above pay rare attention to the characteristics of the short text data stream, and the concept drifts hidden in the short text data stream at the same time. Therefore, it is not easy to meet the short text data stream’s classification requirements with concept drifts.

III. SHORT TEXT DATA STREAM ENSEMBLE CLASSIFICATION ALGORITHM

A. Problem Definition

The classification of short text data streams is defined as follows. Given a short text data stream D, the data stream D is divided into N data blocks of the same size, so D can be expressed as D=[D1,D2, D3, ...,DN] (N→∞): Each short text data block Di can be defined as Di=[d1,d2, d3, ...,dM] (|Di|=M); each d can be defined as di=(vi, y) ϵ E, y= Y, where E represents the feature space of the data block, and Y represents the class label set. The goal of short text data stream classification is to train a dynamic classifier: E→ Y, which maps the feature vector to a set of labels. This classifier automatically adapts to changes in short text data streams and detects concept drifts in the short text data stream. Figure 1 shows the basic framework of the proposed algorithm in this paper.

B. Feature Space

Distinguishing features are chosen to form the feature space F, and are tried to be contained in each short text appear in the feature space, so as to ensure that each short text is directly related to the feature space. In addition, the selected features should be shared by many texts in a category, which can ensure the balance of feature distribution in the text and avoid sparseness. For this reason, this paper focuses on the difference in the number of texts in each category and the relevance between the feature and the category when making feature selection. The features that have a large contribution to the category are selected and used to represent the corresponding category. The selected features are combined to form a feature space except the repetitive features. The feature space will cover the text of each category to the greatest extent, and the selected features are also better to the classification of short texts.

This paper selects the same ratio of feature words for each category. The advantage of the method is that when the number of categories A and B is extremely unbalanced, the small category is not easily covered by the large category. Besides, this paper uses the intra-category dispersion CD, to represent the distribution of features in a category, and the calculation is shown in formula (1).

![Fig. 1. The framework of data streams classification.](image-url)
To calculate the semantic distance \( dist(d_j',D_i) \) in formula (3), according to the class distribution, the current data block is first divided into class clusters \( \{L_c\} \), where \( L_c \) represents the short text set of class label \( c \); then the distance between the short text and all class clusters \( \{L_c\} \) is calculated. The semantic distance \( dist(d_j',D_i) \) between the short text and the current data block is the value of the smallest semantic distance to a certain class cluster. The calculation is shown in formula (4):

\[
dist(d_j',D_i) = \arg \min_{c \in \{L_c\}} \text{dist}(d_j',L_c)
\]

(4)

Given a threshold \( \mu \), it can be judged whether a concept drift occurs. If \( dist(D_{j,i},D_i) \) is greater than \( \mu \), we consider that the concept drift occurs in the new data block.

### E. Construction and Update of Ensemble Classification Model

K base classifiers are constructed on K expanded data blocks, respectively. SVM is used as the base classifier because it is widely used for text classification. When the new data block \( D_{new} \) arrives, we first perform feature extension on the short text to obtain the expanded short text data block \( D'_{new} \). The semantic distance is calculated between the new data block \( D'_{new} \) and the data block \( D_i' \) that constructs the base classifier \( C_i \). When the data block \( D'_{new} \) has a concept drift relative to each data block in the ensemble model, if the number of base classifiers in the ensemble model \( EC \) is less than \( K \), the classifier \( C_{new} \) is added to the ensemble model \( EC \); if it is equal, the classifier \( C_{new} \) is replaced as the oldest base classifier in the ensemble model \( EC \); otherwise, the classifier \( C_{new} \) is used to replace the base classifier constructed from the data block with the smallest semantic distance from the data block \( D'_{new} \).

#### Our algorithm

Input: short text data stream \( D \), the number of base classifiers \( K \), the threshold of concept drift detection \( \mu \);
Output: an ensemble classification model \( EC \);
For \( D_i \) in \( D \) do:

1. Performing feature extension on the short text data block \( D_i \) to obtain the expanded short text data block \( D_i' \);
2. Constructing a base classifier \( C_i \) in the short text data block \( D_i' \);
3. The weight of the base classifier \( C_i \) is the classification accuracy of the base classifier \( C_i \) on the data block \( D_i' \);
4. If \( (i<K) \):
   a. Calculating the semantic distance between the short text data block \( D'_{new} \) and \( D_i' \);
   b. Determining whether concept drift occurs according to the threshold of concept drift detection \( \mu \);
   c. If (concept drift occurs):
      i. Constructing a base classifier \( C_{new} \) in the data block \( D'_{new} \) and updating the ensemble classification model \( EC \);
      d. End for.

The paper proposes a concept drift detection method based on topic distribution. We accumulate the semantic distance between each short text in the new and current data block to determine whether the concept drift occurs in the new data block. The calculation is shown in formula (3):

\[
dist(D_{j,i+1},D_i) = \frac{1}{|D_{j,i+1}|} \sum_{j=1}^{|D_{j,i+1}|} \text{dist}(d_j',D_i)
\]

(3)
IV. EXPERIMENTAL RESULTS AND ANALYSIS

This section mainly compares the experimental results of the algorithm proposed in this paper with the benchmark algorithm. First, the short text data sets and the benchmark algorithms used in the experiment are given [27-31], and then the evaluation indexes and parameter settings of the experiment are explained. Finally, the effectiveness of the algorithm proposed in this paper in concept drift detection and classification performance is verified.

A. Experimental Data Sets and Evaluation Indexes

Table 1 shows the detailed information of the experimental data sets:

Snippets data set: Snippets is the search result of putting a preset expression on the Internet search engine, which has 8 categories and 12340 short texts.

News data set: News comes from TagMyNews data set, which has 7 categories. Each news contains a short headline, description, and a date, and other information. In the experiment, we only select descriptions as a data set, and there are 32604 short texts.

Tweets data set: With the help of the Twitter keyword tracking API, the data is obtained from Twitter from November to December 2012, which has 4 categories. In the experiment, we only used the data from December 2012, and there are 803613 short texts.

This paper presents 4 classical data stream classification algorithms, 4 short text data classification algorithms, and 4 concept drift detection algorithms to verify the effectiveness of the algorithm proposed in this paper in classification and concept drift detection. Table 2 gives detailed information on the benchmark algorithms.

B. Parameter Setting

On the Snippets and News data sets, the data block size is set to 100. The Tweets data set is too large, so the data block size is set to 1000. In concept drift detection, we set a threshold, $\mu = 1 - 0.5/C$, where $C$ represents the number of categories. When using an ensemble framework, too many or too few buffer data blocks are not appropriate. This paper selects 6 buffer data blocks to construct an ensemble classifier. Finally, we choose the SVM from libsvm as the base classifier, and the corresponding parameters are $C$-SVC and linear kernel function.

C. Parameter $R$

The feature selection ratio $R$ is an essential parameter in the algorithm, which directly affects the feature space’s dimension. If the selection ratio is small, the dimension of the feature space is small. Otherwise, the dimension of the feature space is enormous. Figure 2 shows this method’s variation curve with the feature selection ratio on the data sets Snippets, News, and Tweets. It can be seen from Figure 2 that as the ratio of feature selection increases, the classification accuracy increases first and then stabilizes. This is because the selected features are highly representative. After reaching a certain accuracy, as the selection ratio decreases, the feature dimension becomes smaller. The sparsity of short texts causes a downward trend in classification accuracy. In the experiment, the feature selection ratio $R$ is set to 1/60.

D. Concept Drift Detection

To verify the effectiveness of the concept drift detection method in this paper, three statistical evaluation indexes are used. False alarms: the error probability of concept drift detection; Missing: the probability that there is concept drift but not detected; Delay: the average number of short texts that delay detection of concept drift when concept drift occurs. Table 3 shows the evaluation results of the concept drift detection method proposed in this paper and the benchmark concept detection method. The experimental results show that this paper’s concept drift detection method can effectively detect concept drift.

<table>
<thead>
<tr>
<th>TABLE 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>DATA SETS</td>
</tr>
<tr>
<td>Data set</td>
</tr>
<tr>
<td>Snippets</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>News</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Tweets</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

Fig. 2. The curve of classification accuracy with parameter $R$.  

Volume 48, Issue 4: December 2021
TABLE 2
BENCHMARK ALGORITHMS.

<table>
<thead>
<tr>
<th>Category</th>
<th>Method</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Stream Classification</td>
<td>Naive Bayes</td>
<td>Data stream classification method based on single-sample Bayesian model</td>
</tr>
<tr>
<td></td>
<td>Spegasos</td>
<td>Spegasos method based on random variables</td>
</tr>
<tr>
<td></td>
<td>KNN+PAW+ADWIN</td>
<td>K-nearest neighbor method based on probability approximation window and adaptive sliding window</td>
</tr>
<tr>
<td>Short Text Data Classification</td>
<td>S-SVM</td>
<td>Support vector machine method based on self-information</td>
</tr>
<tr>
<td></td>
<td>MaxEnt</td>
<td>Classification method based on maximum entropy</td>
</tr>
<tr>
<td></td>
<td>INPath</td>
<td>Classification method based on semi-supervised learning</td>
</tr>
<tr>
<td>Concept Drift Detection</td>
<td>EDDM</td>
<td>Early concept drift detection method</td>
</tr>
<tr>
<td></td>
<td>ADWINChangeDetector</td>
<td>Concept drift detection method based on adaptive sliding window</td>
</tr>
<tr>
<td></td>
<td>HDDM-A-Test</td>
<td>On-line concept drift detection method based on Hoeffding boundary</td>
</tr>
<tr>
<td></td>
<td>HDDM-W-Test</td>
<td>On-line concept drift detection method based on McDiarmid boundary</td>
</tr>
</tbody>
</table>

TABLE 3
STATISTICS OF CONCEPT DRIFT DETECTION IN DATA SETS.

<table>
<thead>
<tr>
<th>Measure</th>
<th>EDDM</th>
<th>ADWIN</th>
<th>HDDM-A-Test</th>
<th>HDDM-W-Test</th>
<th>Ours</th>
</tr>
</thead>
<tbody>
<tr>
<td>Snippets False Alarms</td>
<td>4.16</td>
<td>13.64</td>
<td>8.53</td>
<td>5.42</td>
<td>1.78</td>
</tr>
<tr>
<td>Delay</td>
<td>28.34</td>
<td>23.42</td>
<td>9.84</td>
<td>11.73</td>
<td>5.92</td>
</tr>
<tr>
<td>News False Alarms</td>
<td>6.98</td>
<td>12.35</td>
<td>7.59</td>
<td>7.64</td>
<td>6.75</td>
</tr>
<tr>
<td>Missing</td>
<td>36.42</td>
<td>28.54</td>
<td>16.36</td>
<td>16.23</td>
<td>6.42</td>
</tr>
<tr>
<td>Delay</td>
<td>28.56</td>
<td>20.36</td>
<td>9.87</td>
<td>10.21</td>
<td>3.14</td>
</tr>
<tr>
<td>Tweets False Alarms</td>
<td>5.38</td>
<td>9.82</td>
<td>7.85</td>
<td>8.14</td>
<td>4.92</td>
</tr>
<tr>
<td>Missing</td>
<td>32.41</td>
<td>18.93</td>
<td>34.36</td>
<td>32.52</td>
<td>9.64</td>
</tr>
<tr>
<td>Delay</td>
<td>67.52</td>
<td>62.46</td>
<td>82.24</td>
<td>78.43</td>
<td>1.86</td>
</tr>
</tbody>
</table>

E. Classification Performance

Figure 3 shows the classification accuracy of the algorithm proposed in this paper and the benchmark algorithm on the three data sets. The algorithm proposed in this paper has achieved a certain degree of improvement on the data sets Snippets, News and Tweets. It shows that the algorithm proposed in this paper is effective for short text data stream classification.

When concept drift occurs in the short text data stream, the algorithm uses multiple different data blocks to construct multiple base classifiers, containing more concept drifts that have occurred, which can more effectively adapt to the concept drifts. The algorithm proposed in this paper alleviates the high-dimensional and sparse problems of short text data streams with feature extension, adopts the semantic distance between data blocks to judge whether concept drift occurs, and uses the latest data blocks to update the
ensemble model. Therefore, the method proposed in this paper is more suitable for the classification of short text data streams and has strong robustness and better classification performance.

V. CONCLUSIONS

This paper presents a new short text data stream classification algorithm. For the high-dimension and sparseness of short text, the feature space selected in this paper takes into account the difference in data distribution on the data set and reduces the dimension of the feature space, thereby alleviating the classification impact caused by sparse features and weak concept description. To effectively detect and adapt to the concept drifts in short text data streams, a concept drift detection method is also used based on topic distribution. By accumulating the semantic distance between each short text in the new and current data block, it is judged whether concept drift occurs in the new data block. The ensemble classification model is updated according to whether the concept drift has occurred. Experimental results show that the classification performance of the short text data stream classification algorithm proposed in this paper is more effective than other algorithms on the same data sets.

REFERENCES