Design Engineering

ISSN: 0011-9342
Issue: 1 | Pages: 147 - 154

An Ensemble Classification Algorithm based on Semantics for Text Data Streams with Concept Drifts

Gang Sun[1], Zhongxin Wang[1,*], Jia Zhao[1], Hao Wang[1], Xiaowen Guan[2]

[1]School of Computer and Information Engineering, Fuyang Normal University, Fuyang, China

[2]Fuyang Cigarette Factory, China Tobacco Anhui Industry Co., Ltd., Fuyang, China

*Corresponding author: Zhongxin Wang

Abstract:

How to mine user-interested information from text data streams with concept drifts is one of the hot topics in natural language processing research, therefore, a new ensemble text data streams classification algorithm based on semantics is proposed. The algorithm first uses the minimum redundancy and maximum relevant feature selection method to remove irrelevant features and redundant features in the text data stream; then, uses the topic model calculates the semantic similarity in the text data stream and detects the concept drifts; finally, the ensemble classification model is used to classify the text data stream. Experimental results show that the ensemble classification algorithm proposed in this paper can effectively detect the concept drifts and has good classification performance for text data streams.

Keywords: *Text data stream, Ensemble classification model, Concept drift, Feature selection, Topic model.*

I. INTRODUCTION

At present, it is the information age, people work and live through the Internet, and the online social media and e-commerce in real life are widely used, such as online chat, product evaluation, online news, etc., which generates massive text data streams in the information society. In practical applications, the topics of these text data streams will change with time. For example, the content of sports news will change with different sports, such as from ball sports to track and field sports, from diving sports to swimming sports, from gymnastics sports to shooting sports, and so on. The comments of products will change with the interests of users, and the content of online chat will also change with social hot issues. The change is called the concept drift of text data stream [1-2]. The general dimensions of text data streams are relatively high, and not all features are helpful for text data stream classification. There are some irrelevant and redundant features, which will not only help classification, but will

ISSN: 0011-9342
Issue: 1 | Pages: 147 - 154

seriously affect the performance of text data stream classification. The concept drift detection of text data streams is an important problem in the classification of text data streams. In the past, concept drift detection was based on measuring the degree of change in the error rate of some kind of information in the process of data stream classification, rarely considering the implied semantic information in text data stream. The semantic information of the text is very important for text classification. In different environments, the same vocabulary may have different meanings.

To solve the above problems, a new ensemble classification algorithm based on semantics is proposed for the classification of text data stream with concept drifts. This algorithm first uses the minimum redundancy and maximum relevant feature selection method (mRMR) to remove irrelevant features and redundant features in the text data stream; then, uses the topic model calculates the semantic similarity in the text data stream and detects the concept drifts; finally, the ensemble classification model is used to classify the text data stream.

II. RELATED RESEARCH

Reference [3] proposed an effective classification method to classify high-speed text streams, so that only one scan of the text stream can effectively classify the text data stream. Reference [4] aims to solve semi-supervised text stream classification problems, and uses support vector machines as sub-classifiers to classify text data streams. Reference [5] proposed an extended model to dynamically identify useful patterns on the text stream. Reference [6] proposes to assign weight to each representative sample, indicating the subjection degree between it and the related class. Reference [7] proposes that when the sample is judged as a non-conceptual drift sample, the original sub-classifier is continuously updated incrementally according to the new sample; when the sample is judged as a conceptual drift sample, new sub-classifier is established based on the sample. Reference [8] proposed an ensemble data stream classification algorithm, which uses random decision trees to establish an ensemble classifier and uses a double-layer threshold to distinguish noise and concept drift. Reference [9] proposed a concept drift detection method, which compares the changes in the original data distribution and the current data distribution, and uses a two-layer window mechanism to detect concept drift to improve adaptability. Reference [10] proposed an ensemble classification algorithm based on decision trees and Bayes, using Hoeffding boundary and μ test to realize the concept drift detection. Reference [11] proposed a new method to realize concept drift detection in the data stream, which is combined with other detection methods to work together to detect concept drift in the data stream. In practical applications, most algorithms use different error rate evaluation methods to reduce the noise interference, and the performance is mainly limited by the label information and the classifier. The above studies rarely consider the impact of useless features on the classification performance in the text data stream, nor do they consider the semantic information implied in the text data stream.

III. AN ENSEMBLE CLASSIFICATION ALGORITHM BASED ON SEMANTICS FOR

ISSN: 0011-9342
Issue: 1 | Pages: 147 - 154

TEXT DATA STREAMS WITH CONCEPT DRIFTS

For the problems of irrelevant features and redundant features, and concept drifts, a new ensemble classification algorithm based on semantics for text data streams with concept drifts ECASTC is proposed. The algorithm first uses the mRMR method to delete irrelevant features and redundant features in the text data stream, then uses semantic similarity to detect the concept drifts in the text data stream, and finally uses the ensemble classification model classifies the text data stream.

3.1 Minimum Redundancy and Maximum Correlation Feature Selection Method

Mutual information (MI) as the evaluation function of feature selection. The greater the mutual information for a certain feature and category, the greater the contribution of the feature to the classification. It assumes that the features are relatively independent, without considering the redundancy between features. In real data, there is a very high dependency between features. The mRMR feature selection method was first proposed by Peng et al, which uses mutual information as a feature selection method for redundancy features and correlation features. For features, the new feature is selected by considering the maximum dependency between the new feature and category and the minimum dependency between the new feature and the selected feature.

The formula of mutual information for processing discrete data sets is shown in formula (1):

$$I(X,Y)=\sum_{x_i\in X,y_j\in Y}P(x_i,y_j)\ln\frac{P(x_i,y_j)}{P(x_i)P(y_j)} \tag{1}$$

The principle of minimum redundancy reflects the correlation between different features. For minimizing the redundancy between features, the minimum similarity between features needs to be considered, specifically, it means researching a set S containing | S | features so that each feature in set S is the most dissimilar among each other, that is minimum similarity. The minimum redundancy condition is shown in formula (2):

$$\min R(S), R=\frac{1}{|S|^2}\sum_{w_i,w_j\in S}I(w_i,w_j) \tag{2}$$

The principle of maximum relevance reflects the ability of features to distinguish categories. Specifically, it means researching a set S containing | S | features, which maximizes the correlation between all features in S and categories. The maximum correlation condition is shown in formula (3):

$$\max D(S,C), D=\frac{1}{|S|}\sum_{w_i\in S}I(w_i,C) \tag{3}$$

Using the mRMR feature selection method to select features of text data stream [12], the useless features in the text data stream can be removed to reduce the impact of the useless features on the performance of the text data stream classification.

3.2 Concept Drift Detection Method

The concept drift detection principle of the ECASTC algorithm is based on the semantic similarity implied in the text data stream, and the semantic similarity is calculated by the LDA

ISSN: 0011-9342
Issue: 1 | Pages: 147 - 154

model to judge whether the text data stream has concept drift. If the topics of two adjacent data blocks are relatively similar, it is considered that the concept drift has not occurred, otherwise, it has occurred [13]. The following describes the implementation process.

The LDA model is used for calculation on the data block Di to obtain the topic-word probability distribution of each topic $\varphi_{t_i} = \{P(w_1|t_i), P(w_2|t_i), \cdots, P(w_m|t_i)\}$, $t_i \in T_i$. The concept drift is detected by determining the proportion of similar topics in T_i and T_{i+1}. Given $t_i \in T_i$ and $t'_j \in T_{i+1}$, the first *g* words with larger semantic weight values in probability distributions φ_{t_i} and $\varphi_{t'_j}$ form the keyword sets W_{t_i} and $W_{t'_j}$ of the corresponding topic, and the similarity of the topic is determined by the weight and proportion of the same words in the two keyword sets. Since the probability of the word w_i in φ_{t_i} and φ_{t_j} in the intersection W is different, that is, the semantic weight of the word w_i for the topics t_i and t'_j is different, they need to be summed separately. The specific calculation formula is shown in formula (4).

$$sim(t_i, t'_j) = \frac{\sum_{w_i \in W} P(w_i|t_i) + P(w_i|t'_j)}{\sum_{w_i \in W_{t_i}} P(w_i|t_i) + \sum_{w_i \in W_{t'_j}} P(w_i|t'_j)} \tag{4}$$

When $sim(t_i, t'_j)$ is greater than the threshold β, it can be determined that the topics t_i and t'_j are similar.

Considering that the essence is that the semantics of the next text data block has changed with respect to the previous text data block, so for any $t'_j \in T_{i+1}$, if there is no topic with similar semantics in T_i, then t'_j is considered a new topic, and the semantics have changed; if there is topic with similar semantics in T_i, the semantics have not changed.

Use $l_{t'_j}$ to mark whether t'_j has changed semantically. A value of 1 means no change, and a value of 0 means change, as shown in formula (5).

$$l_{t'_j} = \begin{cases} 1, & sim(t_i, t'_j) > \beta \\ 0, & other \end{cases}, \quad t_i \in T_i, t'_j \in T_{i+1} \tag{5}$$

Considering the proportion of similar topics as the semantic similarity of adjacent data blocks, the following calculation can be used to detect concept drift. Specifically, it is shown in formula (6), where *k* is the number of topics.

$$r_t(D_i, D_{i+1}) = \frac{\sum_{t'_j \in Z_{i+1}} l_{t'_j}}{k} \tag{6}$$

When r_t is greater than the threshold α, no concept drift occurs because the semantic similarity of the two data blocks is high, otherwise, the concept drift occurs.

3.3 Algorithm Description

The symbols used by the ECASTC algorithm proposed in this paper are explained, D

ISSN: 0011-9342
Issue: 1 | Pages: 147 - 154

represents the current data block, EC represents the ensemble classifier, Ei represents the i-th data instance, Ci represents the i-th base classifier.

The framework of the ECASTC algorithm is shown in Fig 1:

```
ECASTC algorithm
Input: Text data stream TDS
Output: The ensemble classifier EC
Begin
    While (read new data instance) {
        Read d data instances to form a data block;
        Form K data blocks;
        Create K base classifiers Ci on K data blocks;
        The weight of the base classifiers is the classification accuracy of the base classifier Ci on the data block D;
        if (concept drift occurs )
            Create a new base classifier Cnew on data block D;
            Calculate the classification accuracy of the new base classifier Cnew on data block D;
            Calculate the classification accuracy of all base classifiers Ci in the ensemble classifier EC on the data
block D;
            Find the base classifier Cm with the smallest classification accuracy;
            if(Cnew's classification accuracy> Cm's classification accuracy)
                Replace the base classifier Cm with a new base classifier Cnew;
            Update the weight of each base classifier Ci in the ensemble classifier EC;
        }
    }
End
```

Fig 1: ECASTC algorithm

IV. EXPERIMENTAL RESULT AND ANALYSIS

4.1 Data Sets

In this paper, three commonly used benchmark text data sets are selected in the experiment, and the text data stream is simulated by text data through different combinations. The three data sets are Amazon website shopping data (Amazon) and Reuters news data (Reuters) and 20 news groups data (20-Newsgroups), where Amazon website shopping data is composed of shopping reviews of many products, while Reuters news data and 20 news groups data are composed of many news reports in multiple fields.

Because the benchmark data contains many different fields, the data from different fields have different concepts. By selecting data from different fields to form data blocks and simulate text data streams with concept drifts. The specific construction method is as follows:

(1) Amazon data set: There are 4 concepts in the data set. Each concept is repeatedly sampled to obtain 20 data blocks, and each data block is composed of 500 data. The experimental data is composed of 80 data blocks, and is set to 15 concept drifts.

(2) Reuters data set: There are 5 concepts in the data set. Each concept is repeatedly sampled to obtain 20 data blocks, and each data block is composed of 500 data. The experimental data is composed of 100 data blocks, and is set to 16 concept drifts.

(3) 20-Newsgroups data set: There are 4 fields in the data set, and each field contains 3 to 5 concepts. A concept is randomly selected from each field, and each concept is repeatedly sampled to obtain 20 data blocks, and each data block is composed of 500 data. The experimental data is composed of 80 data blocks, and is set to 15 concept drifts.

ISSN: 0011-9342
Issue: 1 | Pages: 147 - 154

4.2 Parameters Analysis

If the feature dimension of the text data is too high, many irrelevant features and redundant features will exist; if the feature dimension of the text data is too low, the text information can not be completely represented by the features. Through some experiments, the feature dimension set by this algorithm is 200. Taking into account both classification performance and time efficiency, this algorithm sets the numbers of topics and keywords to 20 and 50 respectively. When performing concept drift detection, the greater the threshold of similarity, the greater the number of mis-detections of concept drift; the smaller the threshold of similarity, the greater the number of misses of concept drift. Through some experiments, the algorithm sets the threshold α and the threshold β to 0.2 and 0.5 respectively. The ensemble classifier contains multiple base classifiers, and it is not suitable to have more or less base classifiers. A larger number will increase the space-time overhead of the algorithm, while a smaller number will not easily reduce the interference of noisy data. Through some experiments, this algorithm sets the number of base classifiers to 6.

4.3 Concept Drift Detection Analysis

There are usually two evaluation indicators for the concept drift detection method, one is the probability of not being correctly detected, and the other is the number of not being detected. TABLE I is the statistical value of the experimental results on the data sets Amazon, Reuters and 20-Newsgroups, where: Misreported probability represents the probability of not being correctly detected, and Missed number represents the number of not being detected. From the statistical results in TABLE I, it can be seen that the method can effectively detect the change of concept, and most of the concept drift can be detected, and the probability of not being correctly detected is relatively low. Analyse the reason: misreported concepts generally occur at the beginning, and insufficient training data can easily cause concept drift to be incorrectly reported. Because some same words may have different semantics, the algorithm in this paper uses the text semantic information, which enhances the concept discrimination of adjacent data blocks, and can accurately detect the change of concept, and the concept drift has not rarely been detected.

TABLE I. Statistics of the experimental results

Data set	Misreported probability	Missed number
Amazon	6.28%	2
Reuters	5.72%	3
20-Newsgroups	5.36%	3

4.4 Classification Performance Analysis

The experiments were carried out by the ECASTC algorithm and other benchmark algorithms DDM, HT-DDM, CDRDT, DWCDS on the data sets Amazon, Reuters and 20-Newsgroups In the experiments, the experimental data is the average of the results of multiple experiments. The experimental results are shown in Fig 2, which show that the ECASTC

ISSN: 0011-9342
Issue: 1 | Pages: 147 - 154

algorithm not only has better accuracy on the Amazon data set, but also has better classification accuracy on the Reuters and 20-Newsgroups data sets. Analyze it: the mRMA method can remove useless features, reducing the impact of noise on classification performance; the concept drift detection method based on semantic similarity makes full use of the semantic information hidden in the text data, and it can effectively detect the change of concept. If the concept has changed in the text data stream, the ensemble classification model is updated according to the current data block. The updated classification model can quickly adapt to changes of concept, improving classification performance. Therefore, the ECASTC algorithm has better classification performance than other algorithms.

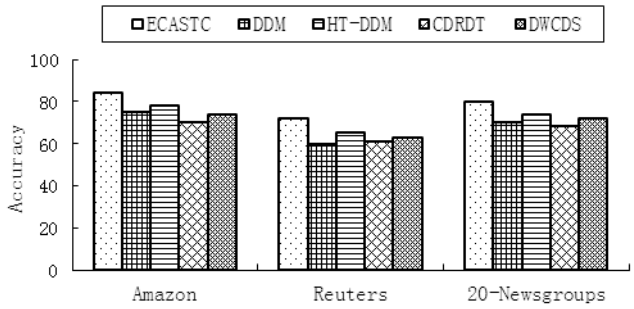

Fig 2: Classification accuracy on data sets

V. CONCLUSION

Text data streams generally have higher dimensions, and there exist concept drifts, therefore, a new ensemble classification algorithm based on semantics was proposed for text data streams with concept drifts. The algorithm first uses the mRMR method to remove useless features in the text data stream; then, uses the topic model calculates the semantic similarity in the text data stream and detects the concept drifts; finally, the ensemble classification model is used to classify the text data stream. The experimental results show that the ECASTC algorithm can effectively detect the concept drifts and has a better classification performance for text data streams with concept drifts.

ACKNOWLEDGEMENTS

The work was supported by the fund project of NSFC (61906044), and the cooperation research project of Fuyang Government and Fuyang Normal University (XDHX2016024), and the fund project of Anhui Education Department Fund Project (KJ2018A0328, KJ2019A0532 and KJ2019A0542).

REFERENCES

[1] Vishwakarma A, Shibu S (2012). Text stream classification techniques and research issues: a review. International Journal of Advanced Research in Computer Science 3(3): 266-278

[2] Lee J H, Lee Y J (2011) Concept drift detection for evolving stream data. ICE Transactions on

ISSN: 0011-9342
Issue: 1 | Pages: 147 - 154

Information and Systems 94(11): 2288-2292
[3] Fung G P C, Yu J X, Lu H (2003) Classifying high-speed text streams. The 4th International Conference on Web-Age Information Management Chengdu China 17-19
[4] Li X, Yu P (2009). Positive unlabeled learning for data stream classification. Proceedings of the SIAM International Conference on Data Mining Nevada USA 259-270
[5] Burdisso S G, Errecalde M, Montes-y-Gómez, Manuel (2019) T-SS3: a text classifier with dynamic n-grams for early risk detection over text streams. Expert Systems with Application 22(11): 1-8
[6] Liu B, Xiao Y, Cao L (2010) Vote-Based LELC for positive and unlabeled textual data streams. Proceedings of the 10th IEEE International Conference on Data Mining Workshops Sydney Australia 951-958
[7] Katakis I, Tsoumakas G (2010) Tracking recurring contexts using ensemble classifiers: an application to email filtering. Knowledge and Information Systems 22(3): 371-391
[8] Li Peipei, Wu Xindong, Hu Xuegang (2010). A random decision tree ensemble for mining concept drifts from noisy data streams. Applied Artificial Intelligence 24(7): 680-710
[9] Zhu Qun, Zhang Yuhong, Hu Xuegang (2011) A conceptual drift data stream classification algorithm based on double-layer windows. Journal of Automation 37(9): 1077-1084
[10] Gui Lin, Zhang Yuhong, Hu Xuegang (2012) A method for detecting drift of data stream concept based on hybrid integration method. Computer Science 39(1): 152-155
[11] Dongre S S, Malik L G, Thomas A (2019) Detecting concept drift using HEDDM in data stream. International journal of intelligent engineering informatics 7(2): 164-179
[12] Shi Qingwei, Cong Shiyuan (2016) Research on text classification based on mRMR and LDA topic model. Computer Engineering and Applications 52(5): 127-133
[13] Chu Guang, Hu Xuegang, Zhang Yuhong (2018) Semantic-based concept drift detection algorithm for text data stream. Computer Engineering 44(2): 24-30