Open Access

Coal Mine Safety Evaluation Method Based on Incomplete Labeled Data Stream Classification

Sun Gang[1,2,*], Zhou Huaping[3] and Sun Kelei[3]

[1]*Data Mining and Intelligent Computing Laboratory, Hefei University of Technology, Hefei, China;*

[2]*School of Computer and Information, Fuyang Teachers College, Fuyang, China;*

[3]*School of Computer Science and Engineering, Anhui University of Science and Technology, Huainan, China;*

Abstract: Monitoring data in coal mine is essentially data stream, and missing coal mine monitoring data is caused by harsh coal mine environment, therefore coal mine safety evaluation can be seen as incomplete labeled data stream classification. The method is proposed for unlabeled data and concept drift in incomplete labeled data stream in this paper that uses semi-supervised learning method based on k-Modes algorithm and incremental decision tree model and concept drift detection mechanism based on clustering concept-cluster. Experimental results show the method can better label unlabeled data and detect concept drift in incomplete labeled data stream, and it has better classification accuracy for incomplete labeled data stream, and it provides a new practical approach for coal mine safety evaluation.

Keywords: Coal mine, safety evaluation, incomplete label, concept drift, data stream classification.

1. INTRODUCTION

Coal mine safety problem is an important problem to be resolved. In order to prevent the occurrence of coal mine accidents, coal mine invested a lot of manpower, material and financial resources to construct a variety of monitoring systems to monitor various data about coal mine safety. Through these monitoring data, coal mine safety can be judged. Coal mine monitoring data have characteristics of data stream as being real-time, continuous, orderly, time-varying and infinite [1], therefore, coal mine safety judged through monitoring data can be seen as data stream classification, and classification label are safety and unsafety.

Data stream Classification is widely applied in network monitoring, sensor networks, e-commerce, but missing class label is existed in actual data stream, such as unlabeled pages in Web, network packets in intrusion detection and online product reviews. Due to performance defects in monitoring equipments and harsh coal mine environment resulting in the failure of monitoring equipment and the fault of transmission line, mine monitoring data can not been collected, therefore, it is difficult to judge coal mine safety state by missing monitoring data. Coal mine safety state judged by missing monitoring data can be seen as incomplete labelled data stream classification.

For unlabeled data and concept drift in incomplete labeled data stream, the method is proposed in this paper that uses semi-supervised learning method based on k-Modes algorithm and incremental decision tree model and concept drift detection mechanism based on clustering concept-cluster. The incomplete labelled data stream classification algorithm designed in this paper can label unlabeled data effectively, and can adapt to the concept drift in incomplete labelled data stream. Coal mine monitoring data stream is classified by this method, and coal mine safety can be judged by the results of data stream classification. Experimental results show the method can better label unlabeled data and detect concept drift in incomplete labeled data stream, and it has better classification accuracy for incomplete labeled data stream, and it can be applied to coal mine safety evaluation.

2. INCOMPLETE LABELED DATA STREAM CLASSIFICATION ALGORITHM

The incomplete labeled data stream classification algorithm proposed in this paper is that it uses semi-supervised learning method based on k-Modes algorithm and incremental decision tree model, and it uses concept drift detection mechanism based on clustering concept-cluster.

2.1. Semi-Supervised Learning Method Based on K-Modes Algorithm and Increamental Decision Tree Model

In order to label unlabeled data in incomplete labeled data stream, it is proposed in this paper that it uses semi-supervised learning method based on k-Modes algorithm and incremental decision tree model. The algorithm proposed in this paper is different from k-Median algorithm in reference [2], this paper uses k-Modes clustering algorithm [3] to label unlabeled data in the process of incrementally creating decision tree. Because k-Modes clustering algorithm is suitable for dealing with discrete attributes, we use discretization method [4] for dealing with continuous attributes.

Semi-supervised learning method [5] for incomplete labeled data stream includes two processes: 1) Incrementally create decision tree; 2) Use k-Modes clustering algorithm to deal with data in process of creating decision tree, and then use maximum class method to label unlabeled data. The brief description of the process is as follows: The root node of the decision tree is initialized firstly. With the incoming data stream, data is traversed to the leaf node of decision tree. If the number of instances reaches the threshold, k-Modes clustering algorithm is used in leaf node, and then the unlabeled instance is labeled by using the maximum class method. If the number of instances reaches the minimum division threshold, the split point is determined by using Hoeffding Bounds inequality and entropy, and then the branch of the decision tree is grown.

2.2. Concept Drift Detection Mechanism Based on Clustering Concept-Cluster

Concept drift detection mechanism based clustering concept-cluster is that it calls one time clustering algorithm at intervals of one data block. Mnew is a clustering cluster formed by the data block in the current cycle, called new concept-cluster. Mhis is a clustering cluster formed by the data block in the previous cycle, called historyical concept-cluster. According to statistical theory, if the distribution of the instances does not change, the Bayes classification error rate will be reduced, otherwise, the Bayes classification error rate will be increased [6]. The change of data distribution means the change of attribution dimension, and the change of the Bayes classification error rate means the change of classification distribution. The algorithm detects different types of concept drift by using the difference between new clustering concept-cluster and historical clustering concept-cluster and using the change of classification distribution.

In order to measure the difference between the new clustering concept-cluster and the historical clustering concept-cluster, the algorithm defines three variables: rnew represents the radius of clustering concept-cluster Mnew, rhis represents the radius of clustering concept-cluster Mhis, dis represents the average distance between the two clustering concept-clusters. 1) If the value of dis is less than the minimum of the radiuses of rnew and rhis, it is considered that a potential concept drift is existed, and the new clustering concept-cluster Mnew complicates with the historical clustering concept-cluster Mhis. If the value of dis is between rnew and rhis, it is considered that the noise is

```
Incomplete labeled data stream classification algorithm:
Input: Data stream; Initial height of tree; The minimum number of split instances; Detection period;
Output: Classification error;
Generate a single leaf for the current decision tree;
For each training instance do
    Sort the instance into a leaf
    Increase the total number of instances arrived at tree;
End for
If it is a labeled instance then
    Update the statistics relevant;
End if
If the count of arrived instances reaches threshold then
    Create concept clusters in k-Modes;
    Label unlabeled data at leaf in majority-class;
    Grow child leaves;
    Record the information of new leaves;
End if
If the number of arrived instances satisfies detection period
then
    For each leaf from bottom to top do
        If there are new instances at the current leaf
        then
        Create concept clusters in k-Modes;
        Label unlabeled data at leaf in majority-class;
        End if
    Detect concept drifts using concept clusters;
    End for
End if
For each testing instance do
    Classify with the current tree in majority-class;
End for
Return the error rate of classification;
```

Fig. (1). The description of incomplete labeled data stream classification algorithm.

existed, and the historical clustering concept-clusters Mhis is maintained, and the new clustering concept-cluster Mnew is discarded. In addition to the above two cases, it is considered that concept drift is existed, and the new clustering concept-cluster Mnew replaces historical clustering concept-cluster Mhis.

2.3. Incomplete Labeled Data Stream Classification Algorithm

Based on the above description of the theory and technology, the general framework of incomplete labeled data stream classification algorithm is shown in (Fig. **1**). With the coming data stream, decision tree is incrementally created. When the number of instances satisfies the division threshold, unlabeled data is labeled in leaf node by using k-Modes clustering Algorithm. In order to track concept drift in data stream, the algorithm will call the clustering algorithm again at intervals of one data block. Potential concept drift is detected by using the difference between new clustering concept-cluster and historical clustering concept-cluster. After training is completed, the algorithm will use the majority-class method to classify test data.

3. COAL MINE SAFETY EVALUATION METHOD BASED ON INCOMPLETE LABELED DATA STREAM CLASSIFICATION

In order to prevent the occurrence of coal mine accidents, coal mine monitoring systems monitor various data about coal mine safety, and using these data judge coal mine safety.

3.1. Coal Mine Monitoring Data

Five natural disasters of coal mine are gas, coal dust, water, fire and roof. Gas refers to various toxic and flammable gas in mine; coal dust refers to coal dust causing explosion; water refers to gushing water leading to flood and casualties; fire refers to various fire underground mine; roof refers to various rock collapse in roadway or workplace. In order to prevent the occurrence of five natural disasters, monitoring systems real-timely monitor various data about coal mine safety, which are gas, carbon monoxide, oxygen, smoke, fan, air door, wind speed, wind pressure, flow, temperature, water level, negative pressure, feed power, voltage, current, power, of which 13 attributes are continuous and 3 attributes are discrete.

3.2. Coal Mine Safety Evaluation Method Based On Incomplete Labeled Data Stream Classification

Description of coal mine safety evaluation method based on incomplete labeled data stream classification is as follows:

(1) Normalize coal mine monitoring data stream;

(2) Create decision tree according to semi-supervised learning method based on k-Modes algorithm and incremental decision tree model in chapter 2.1;

(3) Label unlabeled data according to semi-supervised learning method based on k-Modes algorithm and incremental decision tree model in chapter 2.1;

(4) Detect concept drift in coal mine monitoring data stream according to concept drift detection mechanism based on clustering concept-cluster in chapter 2.2;

(5) If concept drift is occurred, then return (2), and re-create decision tree; if concept drift is not occurred, data stream is classified;

(6) Evaluate coal mine safety according to the results of coal mine monitoring data stream classification.

Coal mine safety evaluation method based on incomplete labeled data stream is shown as (Fig. **2**):

4. EXPERIMENTAL RESULTS AND ANALYSIS

In order to verify the validity of coal mine safety evaluation method based on incomplete labeled data stream classification, experiments are made in simulated data sets and real data sets.

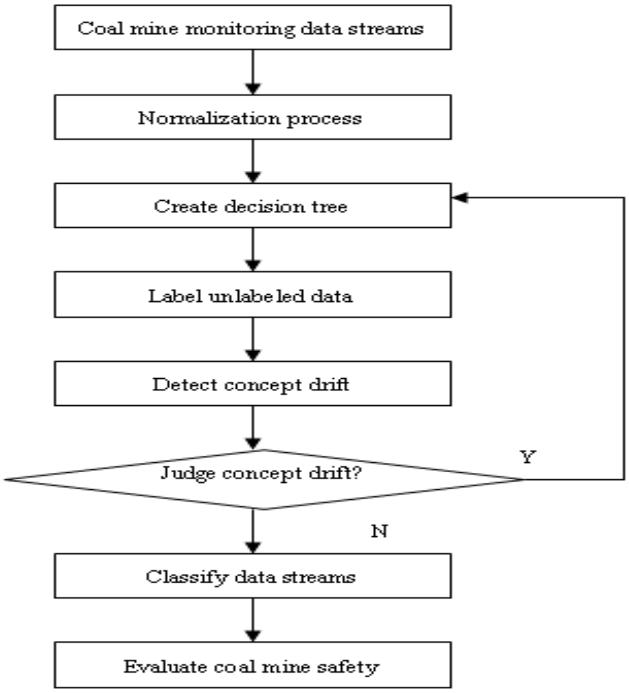

Fig. (2). Coal mine safety evaluation flow diagram.

4.1. Experimental Data Sets

SEA concept drift data set [7]: The data set is a classic abrupt concept drift data set, including two class label and three attributes. In the data set, the first two attributes are related, the third attribute is irrelevant, and all attributes range in [0, 10]. The database contains four data blocks, each data block satisfies function f1 + f2≤θ, f1 and f2 represent the first two attributes, θ is a threshold. Depending on the value of θ four concepts can be formed.

STAGGER data set: The standard database STAGGER [8] contains three Boolean features, each feature contains three values. In STAGGER database, each sample has three attribute values: colour {green, blue, red}, shape {triangle, circle, rectangle} and size {small, medium, large}. Three Boolean features are colour = red size = small, colour = green shape = large and size = medium size = large.

LED data set: MOA open-source platform encapsulates LED data generator with concept drift [9]. For 24-dimensional LED data set, you can set up seven-dimensional drift attributes, but also you can set up different proportions of category noise.

MINE data set: MINE data set is coal mine monitoring data that come from 12 months monitoring data of workface in Huainan coal mine. Each data includes 17 properties, as being gas, carbon monoxide, oxygen, smoke, fan, air door, wind speed, wind pressure, flow, temperature, water level, negative pressure, feed power, voltage, current, power, and safety state. Among them there are 13 continuous attributes and 4 discrete attributes. Safety state is as categories attribute, and it is divided into 2 categories: safe and unsafety.

4.2. Labeled Correct Rate

In order to verify the labeled correct rate of the algorithm, the labeled correct rate (lc) and the classification error rate (ce) is experimented under the difference value (rlp) that the unlabeled data proportion of the total labeled data in this paper. Experimental results show that with the increase of the value of rlp, the labeled correct rate slowly decreased, while the classification error rate slowly increase. When the value of rlp is not higher than 50%, the labeled correct rate remains at 90%, and the error rate rise no more than 2%. Table **1** shows statistics of labeled correct rate in simulated data sets SEA, STAGGER and LED. These experimental results show that the algorithm still has a higher labeled correct rate and classification accuracy in the case of a large number of the missing class label data.

4.3. Concept Drift Detection

In order to verify the effectiveness of the algorithm for concept drift detection, experiments were carried out for concept drift when ulp is 40%. The change of concept drift detection under other cases is same to the change of concept drift detection when ulp is 40%. Usually evaluating concept drift detection algorithm use 4 parameters as being the average number of concept drift detection, false probability of concept drift detection, the number of concept drift not detected, and the sample number of concept drift detected.

Table **2** shows statistics of concept drift detection in simulated data sets SEA, STAGGER and LED. Detection is represented as the average number of concept drift detection; false alarms is represented as the false probability of concept drift detection; Missing is represented as the number of concept drift not detected; Delay is represented as the number of concept drift detected. As can be seen from the statistics, the concept drift detection algorithm can detect most of concept drift in the case of consuming fewer samples.

4.4. Classification Accuracy

Coal mine monitoring data is classified by the coal mine safety evaluation method proposed in this paper, and the method is denoted by CC-SSDSC. Experimental result is the average value of 20 times experiments, and the error rate of classification is 4.38%. Coal mine monitoring data is classified by other classical data stream classification algorithm CVFDT [10], Bag10-ASHT-W + R [11], Clustering-training [12] and Self-training [13]. All experimental results are the average value of 20 times experiments, and the error rates of classification are 10.85%,

Table 1. Statistics of labeled correct rate in simulated data sets

Database	ulp=10%		ulp=20%		ulp=30%		ulp=40%		ulp=50%	
	1c(%)	ce(%)	1c(%)	ce(%)	1c(%)	ce(%)	1c(%)	ce(%)	1c(%)	ce(%)
Sea	95.3	4.2	94.8	4.5	94.2	4.9	93.7	5.1	90.5	5.8
Stagger	94.9	5.3	94.4	5.6	93.8	6.0	93.6	6.2	91.3	6.5
Led	95.1	5.1	94.2	5.5	94.0	5.9	92.6	6.3	90.4	6.7

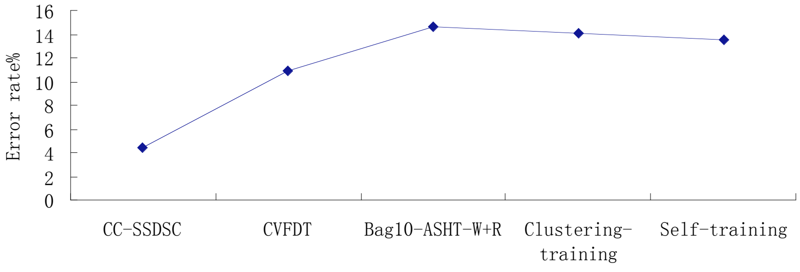

Fig. (3). Error rate of classification in MINE data set.

Table 2. Statistics of concept drift detection in simulated data sets.

Database	ulp=10%		ulp=20%		ulp=30%		ulp=40%		ulp=50%	
	1c(%)	ce(%)	1c(%)	ce(%)	1c(%)	ce(%)	1c(%)	ce(%)	1c(%)	ce(%)
Sea	95.3	4.2	94.8	4.5	94.2	4.9	93.7	5.1	90.5	5.8
Stagger	94.9	5.3	94.4	5.6	93.8	6.0	93.6	6.2	91.3	6.5
Led	95.1	5.1	94.2	5.5	94.0	5.9	92.6	6.3	90.4	6.7

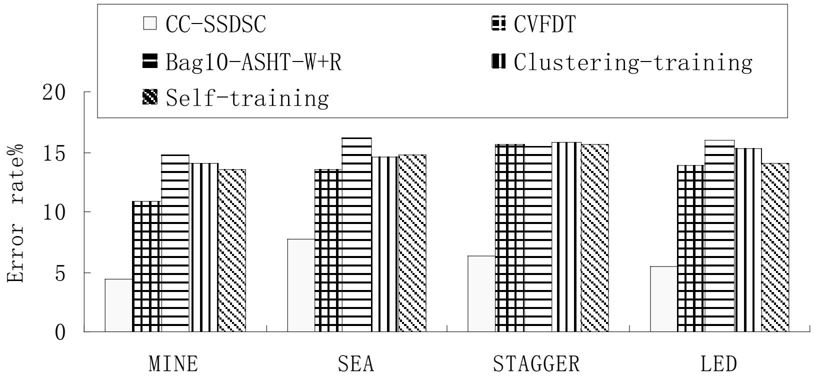

Fig. (4). Error rate of classification in simulated and real data sets.

14.68%, 14.12% and 13.46%, as shown in (Fig. **3**). As can be seen from Fig. **3**, the method is superior to the other algorithms on classification accuracy for coal mine monitoring data set.

In order to verify the effectiveness of the method proposed in this paper, experiments are taken in simulated data sets SEA, STAGGER and LED. Experimental results are shown in Fig. **4**. As can be seen from Figure 4, the method has good classification accuracy in other data sets. It shows the method is effective not only for real data set, but also for other simulated data sets.

As can be seen from the above experiment results, incomplete labelled data steam is better classified by CC-SSDSC method proposed in this paper, and classification accuracy is higher. Coal mine monitoring data is classified by this method for judging coal mine safety. The higher classification accuracy of the method is, the more accurate safety evaluation of coal mine is. By comparison with other algorithms in simulated data sets and real data set, the method proposed in this paper has higher better adaptability. Therefore, the method can satisfied the needs of coal mine safety evaluation, and it can be used for coal mine safety evaluation.

5. CONCLUSIONS

Monitoring data in coal mine is essentially data stream. Due to performance defects in monitoring equipments and harsh coal mine environment resulting in the failure of monitoring equipment and the fault of transmission line, mine monitoring data can not been collected, and missing monitoring data is produced. Coal mine safety evaluation can be seen as incomplete labeled data stream classification, and classification label are safety and unsafety. The method is proposed for unlabeled data and concept drift in incomplete labeled data stream in this paper that uses semi-supervised learning method based on k-Modes algorithm and incremental decision tree model and concept drift detection mechanism based on clustering concept-cluster. Experimental results show the method can better label unlabeled data and detect concept drift in incomplete labeled data stream, and it has better classification accuracy for incomplete labeled data stream. Coal mine monitoring data stream is classified by this method for coal mine safety evaluation and it provides a new practical approach for coal mine safety evaluation.

CONFLICT OF INTEREST

The authors confirm that this article content has no conflict of interest.

ACKNOWLEDGEMENTS

The work was jointly supported by the National Natural Science Foundation of China and Shenhua Group Co., Ltd. (51174257/E0422), and the work was supported by the Fundamental Research Funds for the Central Universities (2013BHZX0040).

REFERENCES

[1] Golab L, Ozsu M T. Issues in data stream managerment. ACM SIGMOD Record, 2003, 32(2): 5-14.

[2] M.-M. Masud, J. Gao, L. Khan, J. Han, and B. Thuraisingham. A Practical Approach to Classify Evolving Data Streams: Training

with Limited Amount of Labeled Data. In: Proceedings of ICDM'08, pp. 929-934, 2008.

[3] M.-K. Ng, M.-J. Li, Z.-X. Huang, and Z.-Y. He. On the Impact of Dissimilarity Measure in k-Modes Clustering Algorithm. IEEE Transactions on Pattern Analysis and Machine Intelligence, 29: 503-507, 2007.

[4] X.-G. Hu, P.-P. Li, X.-D. Wu, and G.-Q. Wu. A Semi-Random Multiple Decision-Tree Algorithm for Mining Data Streams. Journal of Computer Science & Technology, 22(5): 711-724, 2007.

[5] Li-Peipei. Concept Drifting Detection and Classification on Data Streams. Hefei: HeHefei University of Technology, 2012.

[6] O. Chapelle, B. Schölkopf, and A. Zien, eds. Semi-Supervised Learning. Cambridge, MA: MIT Press, 2006.

[7] W.-N. Street and Y. Kim. A Streaming Ensemble Algorithm (SEA) for Large-scale Classification. In: Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD’01), pp. 377-382, 2001.

[8] J.-C. Schlimmer and R.-H. Granger Jr. Incremental Learning from Noisy Data. Machine Learning, 1(3): 317-354, 2004.

[9] W. Fan., H.-X. Wang, P.-S. Yu, and S. Ma. Is Random Model Better? On Its Accuracy and Efficiency. In: Proceedings of the 3rd IEEE International Conference on Data Mining (ICDM’03), pp. 51, 2003.

[10] G. Hulten, L. Spencer, and P. Domingos. Mining Time-changing Data Streams. In: Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining(KDD’01), pp. 97-106, 2001.

[11] A. Bifet, G. Holmes and B. Pfahringer. Leveraging Bagging for Evolving Data Streams. In: Proceedings of ECML PKDD’10, pp. 135-150, 2010.

[12] M. Harries. Splice-2 Comparative Evaluation: Electricity Pricing. Technical report, University of South Wales, 1999.

[13] S. Wu, C. Yang, and J. Zhou. Clustering-training for Data Stream Mining. In: Proceedings of ICDMW'06, pp. 653-656, 2006.

Received: September 16, 2014 Revised: December 23, 2014 Accepted: December 31, 2014