Hindawi
Journal of Electrical and Computer Engineering
Volume 2020, Article ID 4027423, 8 pages
https://doi.org/10.1155/2020/4027423

Research Article

A Classification and Novel Class Detection Algorithm for Concept Drift Data Stream Based on the Cohesiveness and Separation Index of Mahalanobis Distance

Xiangjun Li,[1,2] **Yong Zhou**,[2] **Ziyan Jin**,[3] **Peng Yu**,[2] **and Shun Zhou**[2]

[1]*School of Software, Nanchang University, Nanchang 330047, China*
[2]*Department of Computer Science and Technology, Nanchang University, Nanchang 330031, China*
[3]*Information and Communication Branch, State Grid Jiangxi Electric Power Co. Ltd., Nanchang 330096, China*

Correspondence should be addressed to Xiangjun Li; lxjun_alex@163.com and Ziyan Jin; 578184763@qq.com

Received 15 June 2019; Revised 11 January 2020; Accepted 15 February 2020; Published 19 March 2020

Academic Editor: Hector E. Nistazakis

Data stream mining has become a research hotspot in data mining and has attracted the attention of many scholars. However, the traditional data stream mining technology still has some problems to be solved in dealing with concept drift and concept evolution. In order to alleviate the influence of concept drift and concept evolution on novel class detection and classification, this paper proposes a classification and novel class detection algorithm based on the cohesiveness and separation index of Mahalanobis distance. Experimental results show that the algorithm can effectively mitigate the impact of concept drift on classification and novel class detection.

1. Introduction

In recent years, with the continuous popularization of the Internet and the continuous development of the Internet of Things and data acquisition technology, data has exploded. A constantly changing time-stamped data model, the data stream, has emerged in the Internet, finance, medicine, and ecological monitoring. After the advent of the Internet and wireless communication networks, data flow as a new type of data model has attracted more and more attention from the society [1, 2]. The data stream has characteristics different from traditional datasets. It has chronological, rapid changes and massive, potential infinite, etc. characteristics. It is precisely because of the unique characteristics of the data stream that the data processing model of the data stream is very different from the traditional data mining technology. The data processed by the traditional data mining technology are static datasets, which can be permanently stored in the medium and can be scanned and used multiple times during the process of data analysis. Unlike traditional static databases, the data processing model of the data stream is updated at a faster rate and continuously flows into and out of the computer system. Accordingly, the two biggest challenges in processing data from a data stream are its inherently infinite length and the concept drift that occurs in real-time data changes. Concept drift means that the statistical properties of the target variables that the model attempts to predict change over time in an unpredictable manner. Therefore, using traditional data mining techniques, it is impractical to store and use all historical data for training, which makes it necessary to change existing data mining techniques and design new mining algorithms for this new data model.

Data flow novel class detection is a technique for detecting new categories in a data stream. Many traditional data stream classification algorithms use fixed class numbers to train data stream classifiers. However, in reality, outliers and novel class will appear in the data stream over time, which will lead to a gradual decline in the accuracy of the traditional data stream classification algorithm. Therefore, it

is urgent to design a novel class detection algorithm for the characteristics of data flow.

The rest of this paper is organized as follows: Section 2 introduces the relevant research on data stream classification and novel class detection. Section 3 details the C&NCBM algorithm. Section 4 describes the experimental results and detailed analysis in different datasets. The conclusion of the research as well as challenges and directions for future research is presented in Section 5.

2. Related Work

2.1. Data Stream Classification in the Presence of Concept Drift. In the literature [3], various learning algorithms in the context of concept drift in recent years are reviewed. In 1986, Schlimmer and Granger [4] first proposed the "concept drift," which was followed by the increasing attention of the academic community. From 1986 to 2000, research focused on the use of a single classifier to implement concept drift data stream classification. Widmer and Kubat proposed CBBIT [5], and Hulten et al. proposed methods such as FLORA [6]. At the same time, researchers began to pay attention to the theoretical problem of concept drift data stream classification.

Due to the need to continuously update the classification model when using the single classifier to process the concept drift data stream and the fact that the generalization ability of the classifier is not high [7], Black and Hickey [8] proposed the introduction of integrated learning into the concept drift data stream classification for the first time and proposed the AES algorithm. Therefore, after about 2000, people began to turn to the integrated classifier for the study of concept drift data streams. At this time, the concept drift data stream classification research entered a period of rapid development and began to study the concept drift data stream closer to the reality. Klinkenberg and Lanquillon earlier studied the concept drift in some cases with user feedback or with no feedback [8–11]. In 2004, the Intelligent Data Analysis Journal published the concept drift data stream special issue [12] that mainly discussed how to use the incremental learning method to make the existing classifier use concept drift at a small cost. Subsequently, more attention has been paid to issues such as class imbalance learning [13, 14], concept repetitive learning [15, 16], semisupervised learning [17, 18], and active learning [19, 20] in the classification of concept drift data streams. Table 1 summarizes the main three types of concept drift data stream classification techniques from 2000 to 2016.

2.2. Novel Class Detection in the Presence of Concept Drift. In the literature [33], Masud et al. proposed a novel class detection method in the data stream with concept drift and infinite length. However, this method does not address the problem of feature evolution. In the literature [34], the problem of evolution of the concept is solved while solving the problem of conceptual evolution, but the literature [33, 34] still has too high false alarm rate for some datasets and cannot distinguish different novel class problems. Masud et al. [35] proposed a method to solve the concept evolution caused by the emergence of novel classes. This method adds an auxiliary classifier set to the main classifier set. When each arriving instance in the data stream is determined to be a secondary outlier by the primary classifier set and the associated classifier set, it is temporarily stored in a buffer. When there are enough instances in the buffer, the novel class detection module is called for detection. If a novel class is found, the novel class instance is marked accordingly. In the literature [36], the feature space transformation technique is proposed to deal with the evolution of data stream feature. The traditional data stream integration classifier is combined with the novel class detection technology to solve the feature evolution problem in the data stream.

Chandak [37] proposed a string-based data stream processing method, which mainly solves the problem of data stream concept evolution through the CON_EVOLUTION algorithm. Miao et al. [38] solved the problem that only the numerical data can be solved in the framework of MineClass algorithm. A novel class detection algorithm that can process mixed-attribute data is proposed, and the processing time and model size of the algorithm framework are optimized by using VFDTc classifier. ZareMoodi et al. [39] used local patterns and neighbor graphs to solve the concept evolution problem in data streams. Local patterns are Boolean feature groups that affect sequential features and classification features, which are used to improve classification accuracy. At the same time, in candidate novel class classes, neighbor graphs are used to analyze interrelated objects to improve the accuracy of novel class detection.

After many researchers have continuously explored it, novel class detection has achieved many results. However, most of the novel class algorithms cannot solve the problem of multiple novel class problems at the same time and also do not consider the interaction of different attributes in the instance to determine the novel class. Therefore, based on the previous studies and considering the role of attributes, this paper proposes a novel class detection algorithm that can distinguish different categories of novel class.

3. Classification and Novel Class Detection Algorithm Based on Mahalanobis Distance (C&NCBM)

3.1. Cohesion and Separation Index Based on Mahalanobis Distance. Based on the Mahalanobis distance [40] and the cohesive separation index N-NSC proposed by Masud et al. [33], a novel class detection index is proposed. The relevant definitions are as follows.

Definition 1 (R-outlier) (see [33]). Let x be the test point and $C_{\min}$ be the clustering result point closest to x. If x is outside the range determined by the feature space contained in $C_{\min}$, then x is an R-outlier.

Definition 2 (F-outlier) (see [33]). If x is an R-outlier for all classifiers E_i in the classification set E, then x is an F-outlier.

Table 1: Representative research achievements of concept drift data stream in 2000–2016.

Type	Algorithm	Year	Characteristics	Reference
Incremental learning	VFDT	2000	The leaf node is replaced with a split node, and the algorithm uses less memory and time.	[21]
	HAT	2009	Hoeffding trees are combined with a sliding time window based techniques; there is no need to predict when concept drift occurs in the data stream.	[22]
	OHT	2014	The misclassification rate is used to control node splitting, and the concept drift is solved based on misclassification classes and false alarm rates.	[23]
	Hoeffding-ID	2016	Bayes' theorem is combined with traditional Hoeffding trees. The new spanning tree is continuously used in the classification process to replace the old spanning tree so that the classifier maintains high accuracy and adapts to the data flow concept drift.	[24]
Cluster-based	CluStream	2003	Extending the traditional clustering algorithm BIRCH to the data flow scenario has strong flexibility and scalability, but it is sensitive to outliers.	[25]
	DenStream	2006	Microclusters are used to capture summary information about a data stream, which can find clusters of arbitrary shapes in the data and have the ability to process noise objects.	[26]
	IEBC	2014	The clustering framework is integrated with the classified data stream using sliding window technology and data marking technology, which is excellent in clustering results and detection concept drift but can only process classified data.	[27]
	MuDi-Stream	2016	The multidensity classification problem in the concept drift data stream is solved by a hybrid method based on network and microclusters, but it is not suitable for high-dimensional data streams.	[28]
Integrated learning	AWE	2003	K classifiers are fixedly constructed, and a new classifier is trained in batch mode using the new arrival data object. Subsequently, the k most accurate classifiers are selected to form a classifier set, and each classifier is weighted according to the accuracy.	[29]
	AE	2011	It mainly solves the problem of data stream mining noise and is a collection of horizontal and vertical integration framework methods. The time complexity is high.	[30]
	EM	2013	Concept drift and novel class in the data stream can be automatically detected, but only concept drift under dynamic feature sets can be handled.	[31]
	CLAM	2016	It uses a class-based integrated classifier to efficiently classify data flow loop classes and novel classes, but it cannot classify multiclass data.	[32]

Definition 3 (λ_c-neighbor) (see [33]). The λ_c-neighbor of the F-outlier x is the set of n neighbors closest to x in class c, denoted by the symbol $\lambda_c(x)$, where n is a user-set parameter.

According to the above definition, we give the definition of the cohesiveness and separation index MN-NSC based on Mahalanobis distance.

Definition 4 (MN-NSC). Let $\mathrm{ma}(x)$ be the average Mahalanobis distance of F-outlier x to $\lambda_o(x)$, $\mathrm{mb}_e(x)$ be the average Mahalanobis distance of F-outlier x to $\lambda_e(x)$, and $\mathrm{mb}_{\min}(x)$ be the minimum of $\mathrm{mb}_e(x)$; then MN-NSC is defined as follows:

$$\mathrm{MN-NSC} = \frac{\mathrm{mb}_{\min}(x) - \mathrm{ma}(x)}{\max\left(\mathrm{mb}_{\min}(x), \mathrm{ma}(x)\right)}, \quad (1)$$

where $\lambda_o(x)$ represents the λ_c-neighbor of x to other F-outliers and $\lambda_e(x)$ represents the λ_c-neighbor of x to its existing class.

By definition, the value of MN-NSC is within the interval $[-1, 1]$. When MN-NSC is negative, it means that x is closer to the existing class and it is far away from the F-outlier; when MN-NSC is positive, it means that x is farther from the existing class and close to the F-outlier. When at least N ($>n$) F-outliers have an MN-NSC value greater than 0, this indicates that a new heterogeneity is generated in the data stream.

3.2. Algorithm. This section will elaborate the algorithmic process of classification and novel class detection algorithms based on the Mahalanobis distance cohesive separation index, and it will analyze the concept drift processing in the data stream.

First, the data stream is divided into data blocks of the same size, and the last arriving data block D_i, the currently optimal m classifier sets M, the nearest neighbor n, and the novel class threshold β are taken as input of the algorithm. Then, the instances in the data block are classified to determine whether the instance is R-outlier. If the instance is R-outlier, it will be added to the exception set F. k-means is used to cluster the instances in the set F and create a cluster point Fp_k for each cluster. The Fp_k saves the cluster center and clustering radius of each cluster and calculates MN-NSC value for each cluster point Fp_k. If the number of cluster points with MN-NSC value greater than zero is greater than the set threshold, the algorithm determines that novel class is generated and classifies it. When all data in D_i is marked, D_i is used to train a new model M_{m+1}. M_i, the model with the lowest classification accuracy, is selected from the set M and replaced with M_{m+1}. Through the above method, the classification model of the current latest concepts can be maintained at any time, so as to solve the concept drift problem in the data flow (Algorithm1). The pseudocode for the algorithm is shown below.

Input: *Data block* D_i , *Classifier set* $M = \{M_1, M_2, \ldots, M_m\}$, *Nearest neighbor* n , *Threshold* β
Output: *Updated classifier set* M'
(1) **for** Each instance x in block D_i **do**
(2) **Classify** (M, x)
(3) **if** x is an R-outlier for all classifiers M_i in the classification set M **then**
(4) Add x to the set F
(5) **end if**
(6) **end for**
(7) Clustering F by k-means $(k = n * |F|/|D_i|)$ and creating a cluster point Fp_k for each cluster
(8) **for** Each cluster in F **do**
(9) Compute MN-NSC (Fp_i)
(10) **if** MN-NSC (Fp_i) is greater than 0 **then**
(11) count = count + 1
(12) **end if**
(13) **end for**
(14) **if** count greater than β **then**
(15) Put all instances x belonging to novel class in block D_i into class **C**
(16) **end if**
(17) **if** All instances x in D_i is classified **then**
(18) M_{m+1} = **Train** (D_i)
(19) M' = **Replacement** (M, M_{m+1})
(20) **end if**

Algorithm 1: Classification and novel class detection algorithm based on Mahalanobis distance.

4. Experiment and Analysis

In order to verify the classification and novel class detection algorithm based on the Mahalanobis distance cohesive separation index proposed in this paper, three sets of experiments were performed on two real datasets and one synthetic dataset. KNN (*K*-Nearest Neighbor) [41] was selected as the total data stream classifier of C&NCBM algorithm to confirm the final prediction category of the instance. The essence of the algorithm proposed in this paper is based on KNN. In order to verify the effectiveness of the algorithm, the algorithm that uses KNN to classify the data flow alone and MineClass [33] algorithm proposed by Masud et al. are selected for comparative experiments.

4.1. Experimental Datasets. The KDD Cup 1999, Covertype, and ArtificialCDS datasets were selected as experimental datasets. The number of classes, number of dimensions, and total number of dataset samples for each dataset are shown in Table 2.

4.1.1. KDD Cup 1999 Dataset. (http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html). The KDD Cup 1999 dataset is the dataset used by ACM's annual competition in 1999. The dataset consists of 3 categories of a total of 494,021 data samples, each of which contains 42 attributes. This article uses the 10% version of the KDD Cup dataset.

4.1.2. Covertype Dataset. (http://archive.ics.uci.edu/ml/datasets/Covertype). The Covertype dataset is Resource Information System (RIS) data for the US Forest Service (USFS) Region 2. The dataset contains 7 types of a total of 581,012 instances, each with an attribute dimension of 54.

4.1.3. ArtificialCDS Dataset. (https://moa.cms.waikato.ac.nz/). The ArtificialCDS dataset is a random concept drift data stream that is automatically generated by MOA. The data stream contains 5 classes with a total of 100,000 instances, and the attribute dimension of each sample is 27.

4.2. Performance Index

4.2.1. Classification Accuracy. This experiment uses the accuracy [42] and evaluation time [33] of the classification algorithm to evaluate the quality of different algorithms, which is a widely used evaluation standard in the field of classification algorithms. We expect a good classification algorithm to satisfy the short evaluation time while ensuring high classification accuracy.

4.2.2. Kappa Statistic. Kappa Statistic [43] is an indicator for assessing classification accuracy.

$$\text{Kappa statistic} = \frac{p_o - p_e}{1 - p_e}, \tag{2}$$

where p_o is the proportion of the classifier's agreement, that is, the total number of samples of each correct classification divided by the total number of samples, and p_e is the proportion of the random classification agreement.

4.3. Experimental Results and Analysis. This section separately compares and verifies the proposed algorithm classification performance and the algorithm's effect on the concept drift, giving the result analysis.

Table 2: Parameter of different datasets.

Datasets	Number of classes	Number of dimensions	Number of samples
KDD Cup 1999	3	23	494021
Covertype	7	54	581012
ArtificialCDS	5	27	100000

Table 3: Parameter settings of the three compared algorithms.

Parameter	Covertype	KDD Cup	ArtificialCDS
n	10	10	10
β	45	40	20
Chunk	58102	49402	10000

Table 4: Experimental result data in KDD Cup dataset.

C	C&NCBM accuracy (%)	MineClass accuracy (%)	KNN accuracy (%)	C&NCBM evaluation time (s)	MineClass evaluation time (s)	KNN evaluation time (s)
1	**99.8227**	99.6988	99.6174	**26.8109**	25.95	22.4859
2	**99.8468**	99.7375	99.5071	**54.1313**	51.12	45.0750
3	**99.8078**	99.6569	99.4002	**83.9797**	77.80	69.3438
4	**99.8541**	99.7427	99.5081	**117.6719**	109.33	97.3688
5	**99.8819**	99.7942	99.6065	**152.3984**	138.35	123.5469
6	**99.8890**	99.8130	99.6721	**190.0578**	170.77	154.1313
7	**99.8950**	99.8156	99.6966	**213.7469**	203.87	185.8859
8	**99.9028**	99.8290	99.7249	**246.0797**	228.09	207.2969
9	**99.9020**	99.8245	99.7101	**259.0313**	258.43	234.4469

Table 5: Experimental result data in Covertype dataset.

C	C&NCBM accuracy (%)	MineClass accuracy (%)	KNN accuracy (%)	C&NCBM evaluation time (s)	MineClass evaluation time (s)	KNN evaluation time (s)
1	**88.9832**	87.8438	87.0177	**21.5172**	18.2156	15.7438
2	**91.1561**	89.8646	89.4737	**39.0094**	32.5750	30.7797
3	**91.5212**	89.8546	89.5844	**57.2563**	49.2234	46.3016
4	**92.0270**	90.2581	89.9857	**75.3641**	65.4875	64.8359
5	**91.4612**	89.4413	88.9274	**96.3563**	85.3375	84.2594
6	**91.4837**	89.2808	88.8317	**118.4125**	104.6641	102.9203
7	**91.4707**	89.1186	88.5925	**143.8547**	128.4422	121.6672
8	**91.7367**	89.4059	88.8483	**165.1609**	148.0813	139.3656
9	**91.8901**	89.3561	88.9612	**185.7063**	166.9156	156.5406
10	**92.2249**	89.8698	89.4625	**204.0609**	183.3109	173.3313

4.3.1. Experiment 1. According to the experimental objectives described above, we selected the Covertype, KDD Cup 1999, and ArtificialCDS datasets as experimental datasets and compared the classification accuracy and evaluation time of C&NCBM, MineClass, and KNN alone in the above three datasets. In this experiment, the specific values of the algorithm parameters of different datasets are shown in Table 3. The experimental results on the three datasets are shown in Tables 4–6.

It can be seen from the experimental results in Tables 4–6 that, in the whole data stream classification process, compared with the other two algorithms, the classification accuracy of C&NCBM is very stable throughout the experiment and is significantly higher than that of the other two. The algorithm MineClass also has a better classification effect than that of using KNN alone. The evaluation time of C&NCBM is significantly longer than that of the other two algorithms, and the difference between the evaluation time of MineClass and the time of using KNN alone is small. C&NCBM has higher accuracy than MineClass, but it also requires more evaluation time.

The results of three sets of experiments on two real datasets and one artificial dataset show that the algorithm proposed in this paper is used to deal with the classification of data streams with concept drift and novel class, which has the following characteristics. (1) It is able to make timely

Table 6: Experimental result data in ArtificialCDS dataset.

C	C&NCBM accuracy (%)	MineClass accuracy (%)	KNN accuracy (%)	C&NCBM evaluation time (s)	MineClass evaluation time (s)	KNN evaluation time (s)
1	**76.0900**	74.8900	74.1500	**7.7969**	7.3750	7.3088
2	**76.2050**	75.2600	74.4200	**16.7031**	15.0781	14.6625
3	**76.3500**	75.7800	74.6300	**25.2500**	22.8438	21.9838
4	**76.3775**	75.8050	74.5575	**34.2031**	30.4844	29.3313
5	**76.4500**	75.8640	74.4540	**44.4844**	38.2344	36.6325
6	**76.5467**	75.9917	74.4083	**53.9219**	45.9063	44.2265
7	**76.5686**	75.9557	74.4586	**63.6094**	53.6250	51.7875
8	**76.6013**	75.9388	74.4088	**73.2969**	61.2969	59.4813
9	**76.6444**	76.0044	74.3711	**82.1719**	69.0625	67.0588
10	**76.6870**	75.9800	74.3850	**90.7500**	76.7188	74.5263

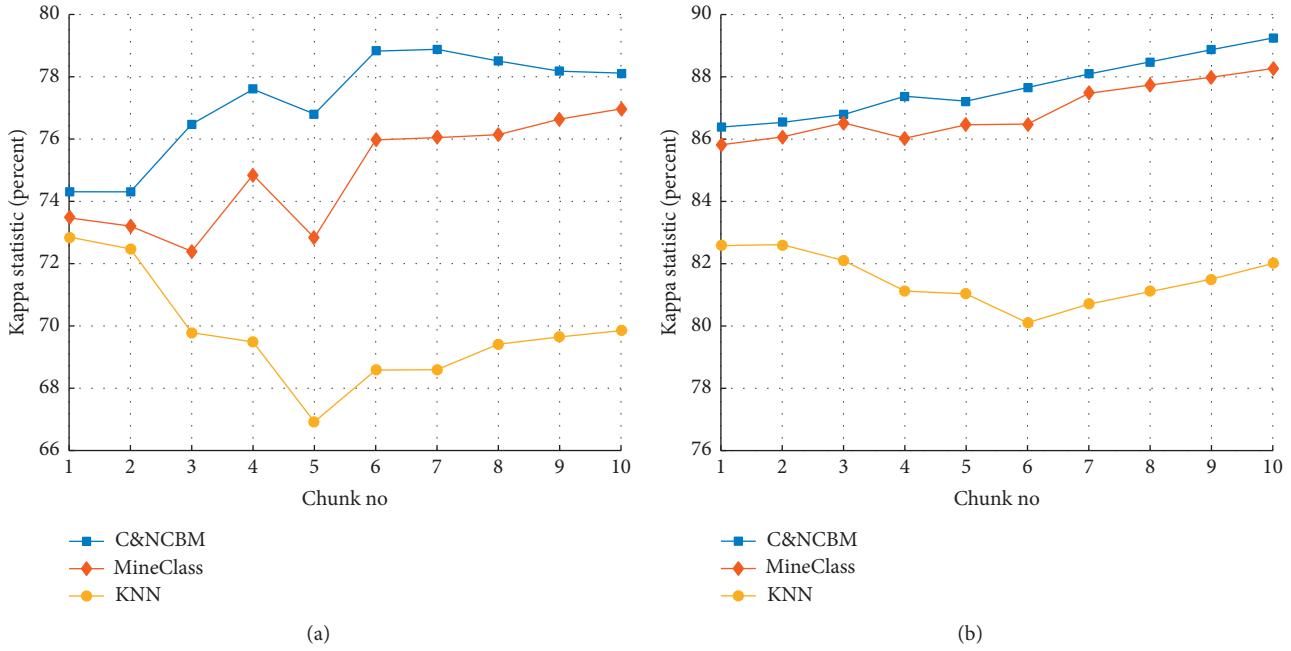

Figure 1: Comparison of Kappa Statistic on datasets. (a) The chunk size in the Covertype dataset is set to 58102, and (b) the chunk size in the ArtificialCDS dataset is set to 10000.

judgments when a novel class appears in the concept drift data stream, and adaptively update the original model after making it, which has stronger classification robustness to novel class occurrences in the concept drift data stream. (2) Compared with the use of ordinary classifiers, there is a significant improvement in classification accuracy, and the classification accuracy is improved to a certain extent compared with the classification and novel class detection algorithms MineClass [33] based on Euclidean distance. (3) The evaluation time is slightly longer than that of the other algorithms.

4.3.2. Experiment 2. The appearance of concept drift in the data stream indicates that the mapping relationship between attributes and categories has changed, and the classifiers on the data stream are based on this mapping relationship. When the attribute-to-category mapping relationship changes, the classification accuracy index Kappa Statistic of the classifier will inevitably change significantly. Therefore, in this section, we will use the difference of classification accuracy of the classifier to determine the sensitivity of different algorithms to the concept drift.

We selected Covertype and ArtificialCDS datasets as experimental datasets and compared C&NCBM, MineClass, and KNN classification accuracy index Kappa Statistic in these two datasets, respectively. The comparison results on the datasets are shown in Figure 1.

In order to introduce the concept drift, we rearranged the Covertype dataset so that at most 3 and at least 2 categories appear in any block at the same time, and new categories appear randomly. The concept drift of the arranged Covertype dataset is mainly in blocks 3 and 5. The ArtificialCDS dataset automatically generated by MOA is incremental drift, which mainly appears in blocks 4 and 6. The results of Figure 1 show that KNN has the fastest decline

in classification accuracy index Kappa Statistic because of the lack of concept drift processing mechanism. MineClass is partially affected, but the decrease is smaller than KNN. C&NCBM is the least affected by concept drift, and the classification accuracy curve is the most gradual. When the concept drift occurs in the data stream, all the three algorithms will be affected to a certain extent. The C&NCBM algorithm proposed in this paper has better concept drift adaptability and can reduce the influence of concept drift on classification to some extent.

5. Conclusion

In this paper, an MN-NSC based on the Mahalanobis distance cohesive separation index is proposed. On this index, a classification and novel class detection algorithm, C&NCBM, based on Mahalanobis distance is proposed. Different from the traditional distance measurement between the examples using Euclidean distance, this method pays more attention to the similarity between instances and can sensitively test small changes between outliers. In the comparative experiment using KNN algorithm and MineClass algorithm, the effectiveness of the classification algorithm is verified. The C&NCBM algorithm, KNN algorithm, and MineClass algorithm classification accuracy Kappa Statistic are also compared. The results show that the proposed C&NCBM algorithm is the best. The concept of drift adaptability can deal with the influence of concept drift on classification in the data stream to some extent. However, due to the problem of adding Mahalanobis distance, the algorithm proposed in this paper requires slightly longer time compared to the other algorithms. How to improve the computational time while ensuring the validity of algorithm classification is the future research direction of this paper.

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was funded by the National Natural Science Foundation of China (nos. 61862042 and 61762062, 61601215); Science and Technology Innovation Platform Project of Jiangxi Province (no. 20181BCD40005); Major Discipline Academic and Technical Leader Training Plan Project of Jiangxi Province (no. 20172BCB22030); Primary Research & Development Plan of Jiangxi Province (no. 20192BBE50075, 20181ACE50033, 20171BBE50064, 2013ZBBE50018); Jiangxi Province Natural Science Foundation of China (nos. 20192BAB207019 and 20192BAB207020); and Graduate Innovation Fund Project of Jiangxi Province (nos. YC2019-S100 and YC2019-S048).

References

[1] U. R. Salunkhe and S. N. Mali, "Security enrichment in intrusion detection system using classifier ensemble," *Journal of Electrical and Computer Engineering*, vol. 2017, Article ID 1794849, 6 pages, 2017.

[2] W. Li, P. Yi, Y. Wu, L. Pan, and J. Li, "A new intrusion detection system based on KNN classification algorithm in wireless sensor network," *Journal of Electrical and Computer Engineering*, vol. 2014, Article ID 240217, 8 pages, 2014.

[3] J. Lu, A. Liu, F. Dong, F. Gu, J. Gama, and G. Zhang, "Learning under concept drift: a review," *IEEE Transactions on Knowledge and Data Engineering*, vol. 31, no. 12, pp. 2346–2363, 2018.

[4] J. C. Schlimmer and R. H. Granger, "Incremental learning from noisy data," *Machine Learning*, vol. 1, no. 3, pp. 317–354, 1986.

[5] G. Widmer and M. Kubat, "Effective learning in dynamic environments by explicit context tracking," in *Proceedings of the Sixth European Conference on Machine Learning*, pp. 69–101, Vienna, Austria, 1993.

[6] G. Hulten, L. Spencer, and P. Domingos, "Mining time-changing data streams," in *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining-KDD '01*, pp. 97–106, San Francisco, CA, USA, 2001.

[7] Y. Wen, B. Qiang, and Z. Fan, "A survey of the classification of data streams with concept drift," *CAAI Transactions on Intelligent Systems*, vol. 46, no. 11, pp. 2656–2665, 2013.

[8] M. Black and R. J. Hickey, "Maintaining the performance of a learned classifier under concept drift," *Intelligent Data Analysis*, vol. 3, no. 6, pp. 453–474, 1999.

[9] W. Nick Street and Y. Kim, "A streaming ensemble algorithm (SEA) for large-scale classification," in *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining-KDD '01*, pp. 377–382, San Francisco, CA, USA, 2001.

[10] R. Klinkenberg, "Using labeled and unlabeled data to learning drifting concepts," in *Proceedings of the Work-shop Notes of the IJCAI-1 Workshop on Learning from Temporal and Spatial Data*, pp. 16–24, Menlo Park, CA, USA, 2001.

[11] R. Klinkenberg and T. Joachims, "Detection concept drift with support vector machines," in *Proceedings of the Seventeenth International Conference on Machine Learning*, pp. 487–494, Stanford, CA, USA, 2000.

[12] C. Lanquillon, "Information filtering in changing domains," in *Proceeding of the 16th International Joint Conference on Artificial Intelligence*, pp. 41–48, Stockholm, Sweden, 1999.

[13] M. Kubat, J. Gama, and P. Utgoff, "Special issue on incremental learning systems capable of dealing with concept drift," *Intelligent Data Analysis*, vol. 8, no. 3, 2004.

[14] H. He and S. Chen, "Towards incremental learning of non-stationary imbalanced data stream: a multiple selectively recursive approach," *Evolving Systems*, vol. 2, no. 1, pp. 35–50, 2011.

[15] S. Ramamurthy and R. Bhatnagar, "Tracking recurrent Concept drift in streaming data using ensemble classifiers," in *Proceedings of the of the 6th International Conference on Machine Learning and Applications*, pp. 404–409, Cincinnati, OH, USA, 2007.

[16] P. Li, X. Wu, and X. Hu, "Mining recurring concept drifts with limited labeled streaming data," in *Proceedings of the 2th Asian Conference on Machine Learning*, pp. 241–252, Tokyo, Japan, 2010.

[17] J. C. Xue and G. M. Weiss, "Quantification and semi-supervised classification methods for handling changes in class distribution," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 897–905, Paris, France, 2009.

[18] P. Li, X. Wu, and X. Hu, "Learning from concept drift data streams with unlabeled data," in *Proceedings of the 24th AAAI Conference on Artificial Intelligence*, pp. 1945-1946, Atlanta, GA, USA, 2010.

[19] I. Ž.t Avail, A. Bifet, B. Pfahringer, and G. Holmes, "Active learning with evolving streaming data," in *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, pp. 597–612, Athens, Greece, 2011.

[20] L. Nan, "Classification algorithm for data streams with concept drift and its applications," Master's thesis, Fujian Normal University, Fuzhou, Fujian, 2013.

[21] P. Domingos and G. Hulten, "Mining high-speed data streams," in *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining-KDD '00*, pp. 71–80, ACM, Boston, MA, USA, 2000.

[22] A. Bifet and R. Gavaldà, "Adaptive learning from evolving data streams," in *Advances in Intelligent Data Analysis VIII*, N. M. Adams, C. Robardet, A. Siebes, and J.-F. Boulicaut, Eds., pp. 249–260, Springer Berlin Heidelberg, Berlin, Germany, 2009.

[23] S. R. Kumari and P. Kumari, "Adaptive anomaly intrusion detection system using optimized Hoeffding tree," *Journal of Engineering and Applied Sciences*, vol. 95, no. 17, pp. 22–26, 2014.

[24] C. Yin, L. Feng, and L. Ma, "An improved Hoeffding-ID data-stream classification algorithm," *The Journal of Supercomputing*, vol. 72, no. 7, pp. 2670–2681, 2016.

[25] C. C. Aggarwal, P. S. Yu, J. Han, and J. Wang, "A framework for clustering evolving data streams," in *Proceedings of the 29th international conference on Very large data bases-Volume 29*, pp. 81–92, VLDB Endowment, Berlin, Germany, September 2003.

[26] F. Cao, M. Ester, W. Qian, and A. Zhou, "Density-based clustering over an evolving data stream with noise," in *Proceedings of the of the Sixth SIAM International Conference on Data Mining*, pp. 328–339, Bethesda, MD, USA, 2006.

[27] A. Amini, H. Saboohi, T. Herawan, and T. Y. Wah, "MuDi-stream: a multi density clustering algorithm for evolving data stream," *Journal of Network and Computer Applications*, vol. 59, no. 1, pp. 370–385, 2016.

[28] Y. Li, D. Li, S. Wang, and Y. Zhai, "Incremental entropy-based clustering on categorical data streams with concept drift," *Knowledge-Based Systems*, vol. 59, no. 2, pp. 33–47, 2014.

[29] Q. Wei, Z. Yang, Z. Junping, and W. Yong, "Mining multi-label concept-drifting data streams using ensemble classifiers," in *Proceedings of the 2009 Sixth International Conference on Fuzzy Systems and Knowledge Discovery*, vol. 5, pp. 275–279, Tianjin, China, 2009.

[30] P. Zhang, X. Zhu, Y. Shi, L. Guo, and X. Wu, "Robust ensemble learning for mining noisy data streams," *Decision Support Systems*, vol. 50, no. 2, pp. 469–479, 2011.

[31] D. M. Farid, L. Zhang, A. Hossain et al., "An adaptive ensemble classifier for mining concept drifting data streams," *Expert Systems with Applications*, vol. 40, no. 15, pp. 5895–5906, 2013.

[32] T. Al-Khateeb, M. M. Masud, K. M. Al-Naami et al., "Recurring and novel class detection using class-based ensemble for evolving data stream," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 10, pp. 2752–2764, 2016.

[33] M. M. Masud, J. Gao, L. Khan, J. Han, and B. Thuraisingham, "Integrating novel class detection with classification for concept-drifting data streams," in *Machine Learning and Knowledge Discovery in Databases*, pp. 79–94, Springer Berlin Heidelberg, Berlin, Heidelberg, 2009.

[34] M. M. Masud, Q. Chen, J. Gao, L. Khan, J. Han, and B. Thuraisingham, "Classification and novel class detection of data streams in a dynamic feature space," in *Machine Learning and Knowledge Discovery in Databases*, pp. 337–352, Springer Berlin Heidelberg, Berlin, Heidelberg, 2010.

[35] M. M. Masud, T. M. Al-Khateeb, L. Khan et al., "Detecting recurring and novel classes in concept-drifting data streams," in *Proceedings of the 2011 IEEE 11th International Conference on Data Mining*, pp. 1176–1181, Vancouver, BC, Canada, 2011.

[36] M. M. Masud, Q. Chen, L. Khan et al., "Classification and adaptive novel class detection of feature-evolving data streams," *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 7, pp. 1484–1497, 2013.

[37] M. Chandak, "Role of big-data in classification and novel class detection in data streams," *Journal of Big Data*, vol. 3, no. 1, pp. 1–9, 2016.

[38] Y. Miao, L. Qiu, H. Chen, J. Zhang, and Y. Wen, "Novel class detection within classification for data streams," in *Proceedings of the International Symposium on Neural Networks*, pp. 413–420, Springer, Berlin, Heidelberg, Germany, 2013.

[39] P. ZareMoodi, H. Beigy, and S. Kamali Siahroudi, "Novel class detection in data streams using local patterns and neighborhood graph," *Neurocomputing*, vol. 158, pp. 234–245, 2015.

[40] P. C. Mahalanobis, "On the generalised distance in statistics," in *Proceedings of the National Institute of Sciences*, Calcutta, India, 1936.

[41] N. S. Altman, "An introduction to kernel and nearest-neighbor nonparametric regression," *The American Statistician*, vol. 46, no. 3, pp. 175–185, 1992.

[42] D. J. Hand and R. J. Till, "A simple generalization of the area under the ROC curve to multiple class classification problems," *Machine Learning*, vol. 45, no. 2, pp. 171–186, 2001.

[43] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, 1960.