International Journal of Multimedia and Ubiquitous Engineering
Vol.9, No.5 (2014), pp.363-372
http://dx.doi.org/10.14257/ijmue.2014.9.5.37

Study on a Classification Model of Data Stream based on Concept Drift

[1]Li Xiaofeng and [2]Gao Weiwei

[1,2] Department of Informatic Science,
Heilongjiang International University,
Harbin 150025, China
[1]mberse@126.com, [2]gvv0451@163.com

Abstract

In the data stream classification process, in addition to the solution of massive and real-time data stream, the dynamic changes of the need to focus and study. From the angle of detecting concept drift, according to the dynamic characteristics of the data stream. This paper proposes a new classification method for data stream based on the combined use of concept drift detection and classification model. The data stream classification model can't adapt to concept drift problem to solve. Before the model classification, the use of information entropy to judge the data block concept drift, the concept of history to have appeared, the use of a classifier pool mechanism to save it, to makes the classification model has stronger resistance to concept drift.

Keywords: *Data Stream Classification, Concept Drift, Incremental Learning, KL-distance*

1. Introduction

In recent years, with further investigation about the concept drift, a few of classification methods for data flow have been proposed against the concept drift. Ross G. J. *et al.* [1], to address the problem of concept drift with data stream, raised the method based on exponentially weighted moving average (EWMA) charts [2], to monitor the accuracy rate of classification model against data flow. When the rate reduces, the concept drift exists and it's necessary to adjust the model. The uniqueness of that model rests with the utilization of concurrent computation, increasing monitoring levels of the model in each computing node, which enables the monitoring of several classifiers simultaneously for the purpose of decreasing monitor load. In the paper, it uses information theory, *i.e.* KL-distance [3], as the method for detecting concept drift to judge whether concept drift happens according to the distribution of data. Moreover, the proposed model designs a pooling mechanism for classifiers, which saves concepts that were ever detected to control the update frequency of the model and thus speed up the classification. That is the most distinctive point of the proposed model from others in previous works. In the end, after visualization of the concept drift with the application of real data set, it conducts in-depth analysis from the conceptual level to provide foundations for data mining conceptually in the future [4].

Concept drift refers to the moving of data in the stream along with time. In many real application of data stream discovery, concept drift is one of the reasons for the failure of classification models. The classification models for data stream have been so far proposed based on models, reducing the impact of concept drift on the performance of such models through the adjustment of model structure (*e.g.* integration learning), or update process of the

ISSN: 1975-0080 IJMUE

model (like incremental learning). On the basis of concept drift detection method, the paper puts forward classification model for data stream and visualized method for concept drift. By introducing concept drift detection to the classification model, the update performance of classification model can be effectively controlled, lowering the update frequency, guaranteeing the appropriateness of the update and enhancing the classifying efficiency with the advantage of pooling mechanism to provide specific classifier for different concept data. Besides, based on statistics of concept drifts, it analyzes and comprehends data stream from conceptual respect, which is helpful to data mining in conceptual level in the future [5-6].

2. Detection Method for Concept Drift

Due to complexity and randomness of concept drift itself, the research on it is still on the exploration stage, which accounts for the fact that it has become one of the top concerns in the field of data mining [7]. After consultation of all related literatures, the proposed methods for concept drift can be divided into:

2.1. Classification model based on integration learning

Integration learning model solves the problem of concept drift by integrating many classification models into a framework and aggregating the classification result by every single classifier to eventually obtain the outcome [8]. For individual classifier in the integration learning model, it is significant that they must be different from one another.

2.2. Classification model based on incremental learning

To detect concept drift, the model takes the idea of updating and adjusting constantly the model to become adaptive to the environment of data stream [9].

2.3. Method for detecting concept drift

The method for detecting concept drift consists of explicit detection method and the implicit. The former designs specifically an algorithm to detect the drift. When concept drift is detected, it will send a message to the model to perform related operations. It works just like a guard monitoring the variation of data flow. The implicit method does not adopt a definite detection mechanism against the concept drift, but a heuristic strategy, to determine whether there is concept drift by monitoring related parameters about the performance, *e.g.* accuracy rate [10]. In a word, this type of method involves four parameters:

- Probability distribution of data
- Correlation of characteristic attributes
- Internal features of classification model
- Accuracy rate

3. Classification Model and Visualization Approach based on Concept Drift Detection Method

According to the description of different methods for detecting concept drift in the above section, we can know the method is still on the theoretical research stage, which has positive meanings to the settlement of concept drift and classification of data stream. With reference to methods mentioned above, we propose a new detection method based on KL-distance algorithm, combining it with classification models to create the model for data stream, so as to weaken the influence of concept drift on such a model for the steady accuracy rate of the

classification. In addition, a visualization method is invented for the concept drift in order to make visualized analytics of the relationship among different concepts.

3.1. Classification Model based on KL-distance

The main idea of KL-distance is to employ KL-distance method to detect concept drift of different data blocks before classification. If the drift is found, the classification model will be updated; otherwise, the model won't be changed but still executes classification of data flow.

3.2. Visualization Method for Concept Drift

At present, concept drift concerns mainly concept drift detection of data stream, classification of concept drift and construction of classification model appropriate for concept drift. Since concept can highly abstract the basic form of data, it's significant for high-level understanding of data and abstraction of knowledge contained in data to investigate data from a conceptual aspect. With regards to the visualization of concept drift, the paper seeks to reconsider concept drift notionally and reveals the feasibility in the learning process. The visualized strategy for concept drift is demonstrated in Figure 1.

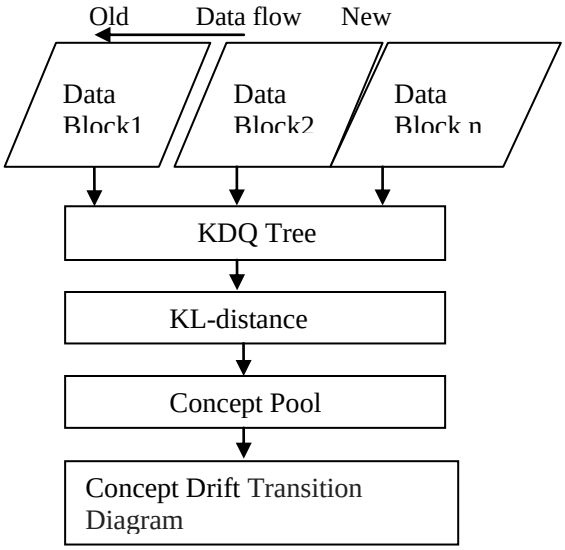

Figure 1. Process of Concept Drift Visualization Method

3.2.1. Firstly, the dynamic data flow is static, here is identical with the above method, still using the sliding window technique, the dynamic data flow is cutting into static data blocks, and each block contains the number of samples are set in advance.

3.2.2. For detecting concept drift of data block. This includes the data block by KDQ tree structure transformation, the formation of virtual data corresponding to the set. In addition, adopting the KL-distance method to detect the occurrence of concept drift, using the Bootstrap method to find the significant parameters of concept drift

3.2.3. Inspired by the classifier pool, this paper designs a concept pool mechanism. Unlike the classifier pool mechanism, the classifier model does not contain the concept of pool, it only can represent a concept. The core idea is to pool, the new data block using the KL-distance method, similarity comparison and concept in the pool the concept of saving, if similar

concepts exist, then the concept of weight updating the concept in the pool (transition probabilities), otherwise the data block to insert a new concept to concept pool. Here in order to better describe the concept of the transfer process, without forgetting mechanism, hoping to more comprehensive understanding of the concept drift of data. Pool mechanism such as shown in Figure 2.

The weight of each concept, namely the concept of transition probability is given by the traditional Bias method, the formula is as follows:

$$P(C_j \mid C_i) = \frac{P(C_i, C_j)}{P(C_i)} = \frac{P(C_j)P(C_i \mid C_j)}{\sum_j P(C_i \mid C_j)} \quad (1)$$

Where, $P(C_i)$ and $P(C_j)$ Represent the probability i and j concepts, $P(C_j \mid C_i)$ said the i concept, a transfer to the j concept of conditional probability.

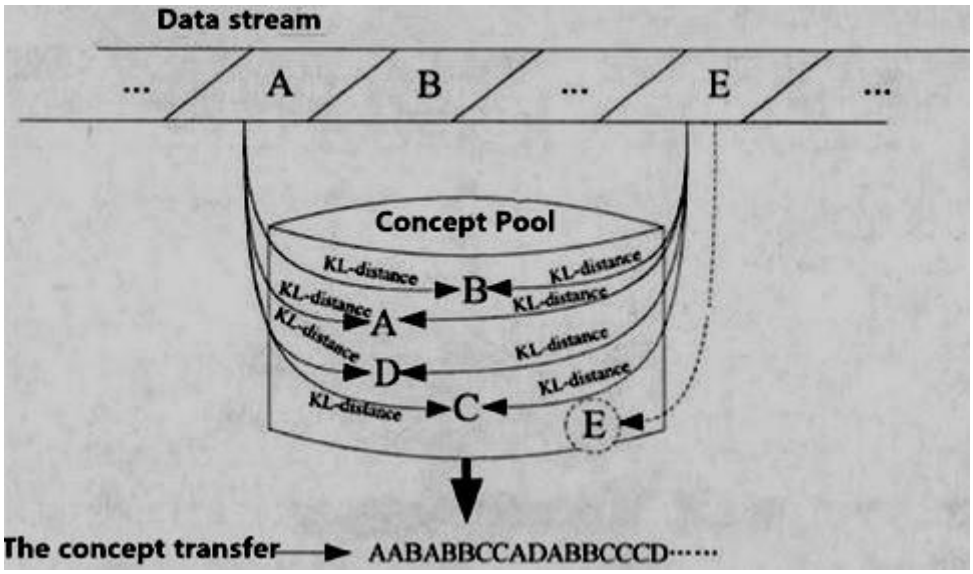

Figure 2. Operating Mechanism of Concept Pool

In summary, to detect the concept drift of data blocks by KL-distance method and the concept of pool mechanism, the frequency of collection and statistics for each concept appeared, finally draw the concept drift transfer diagram to provide help for the next step in the concept of data mining.

4. Experimentation and Result Analysis

The experiment is divided into two parts: one for validation of the classification model for data stream based on concept drift detection and the verification of visualization method for concept drift. The hardware environment is Core Duo 2GHz CPU, 4G DDR, Windows operating system. What's more, in order to conduct comparative study among different classification models of data flow, which is proposed based on the detection of concept drift, we use four classification models to compare them with the proposed model as to validate its effectiveness. Those models are shown as follows:

- Information retrieval support vector machine, IR-SVM
- Incremental support vector machine, I-SVM
- K-Nearest neighbor (K-NN) algorithm
- Decision tree, DT

The reason why we utilize those classification models is that they are fundamental methods which cover almost all classifications of data stream and they are representative, especially for concept drift, validity and correctness can be discovered through comparison.

4.1. Data for experiment

For the test, we use two types of data set, which is respectively artificial and real data set, making artificial data possible to construct concept drift before hands and verifying the validity of the proposed method for detecting concept drift. Real data is applied to determine the performance of such models. For visualized method, we only use real data to do the test in the aim of presenting concept drift of the data through visualization to support the following analysis. In general terms, the data includes 11 data sets, of which there are five artificial data sets and six real data sets. Hereunder is the separate introduction of them.

4.1.1. Artificial data set: We used artificial data sets based on the following reasons:

The experiment data won't be well known if only real data set is used, for example, when will the drift happen to the used data set? Where? And what type? And even whether the drift occurs in the data? They are all totally unknown in advance. Thus, it's impossible to make effective analysis of experimental results and not helpful. But by using artificial data set, the problem can be well solved through presetting the location and time of concept drift to prove the availability of the method proposed here; it's likely to simulate different classes of concept drift and detect well the universality of the new method. For the consideration above, we used an open data generator of concept drift to produce five data sets about the concept drift. The generator compiles in MATLAB environment, easy to adjust all parameters during the generation of data and to control the type of data. Parameters used by the generator are graphically shown in Table 1.

Table1. Synthetic Data Set Generator Parameters

Data name	Generate formulas	Coefficient initial value	Drift process	Category proportion
Circle	$(x-a)^2+(y-b)_2<=r^2$	a=0.5 i=0.5	r = 0.3->0.4->0.5->0.4->0.2	41%
SineV	$y<=a*\sin(b*x+c)+d$	a=1,b=1 c=0	d=-2->-2->-1->0->1->2->2->1 ->-0>1->2->-2	44%
SineH	$y<=a*\sin(b*x+c)+d$	a=1,b=1 d=0	c = 0->π/2->π/2->π->3π / 2->3π/2 ->π->2π->2π->3π/2->π->3π/2->0	50%
Line	$y<=-a_0+a_1*x_1$	$a_1=1$	a_0 = 0->0->1->2->1->1-0->-1 ->-2>-1->-2->-2	50.2%
Plane	$y<=-a_0+a_1*x_1+a_2*x_2$	$a_1=1$ $a_2=1$	a_0 = 1->1->1->2->3->4->5->3->3-> 4->1->3->3	46.6%

With the help of those parameters, five types of concept drift data established by the generator are Circle, SineV, SineH, Line and Plane. Each type of artificial data set consists of

160000 data and every 4000 data moves one time. The data obtained are used to test the classification model for data stream based on concept drift.

4.1.2. Real data set: Despite artificial data sets can control well concept drift and validate materially the effectiveness of the detection mechanism of concept drift, problems with the real data can't be overcome. So for the experimentation, we use not only artificial data and also six real data. They are all collected from open UCI Machine Learning Repository. The following will to these six kinds of data are introduced:

(a) MAGIC gamma telescope dataset
The motion process of MAGIC data set is produced using Y simulation of high energy particle image technology in Cerenkov gamma telescope in. This data set contains 19020 sample data, each data sample contains 11 attributes and two class

(b) Shuttle dataset
The Shuttle data set contains 58000 sample data, 9 properties and 7 categories. The data of all the attribute values are numeric.

(c) Sensor dataset
The Sensor data set is 54 sensor nodes INTEL of the Berkeley laboratory in laboratory the location of each continue to gather and get
(d)Page block dataset
Page block data sets to 54 basic document as the basis, through the analysis of constructing the data set data set consists of 5473 samples, 10 as and 5 categories of this piece of information for each document Chinese.

(e) Power supply dataset
The Power data set describes the short supply situation of one power supply company backbone network and branch network. Global data containing 29928 samples, 2 properties and 24 different categories

(f) KDD99 dataset
In this experiment, adopting the classic KDD99 data is set as the experimental data sets. The content of this data set has been introduced in the previous chapter, so here is not to praise the. That is, the KDD99 data are used in this experiment are processed, 1 normal and 22 categories of original data set of network attack classification, compressed into 3 categories, only the Normal, Neptune, Smurf class, the amount of data from the 494021 original data is reduced to 485269

4.2. Experiment results of artificial data

We used five kinds of artificial data to determine the validity of the proposed model of data stream and since the size of different data block had impacts on the detection of concept drift, blocks in five different volumes were used respectively for the test, like 100, 200, 500, 1000 and 2000. Results are seen in Figure 3.

It can be noted from the picture2 that the accuracy rate of the proposed model changes along with the type of data and volume of block. For the same artificial data, while the block becomes bigger and bigger, the accuracy rate of classification by the model is improved. The rate is the best when the size of a block is 2000, which, however, is not the same for all data. We take SineH data set for instance. On some individual stages during the classification,

when the volume of a block is 100, the accuracy rate is higher than when it's 2000. By contrast, the rate tends to be more stable for blocks in smaller size. After analysis, we found when a block is in small size, concept drift is sensitive. The model adjusts timely to find a suitable classifier as to ensure the stability of the classification accuracy rate. But because the block in small size contains less information, the rate can't be enhanced greatly. When the block is big, although the rate can be increased, due to too many samples and concept drift' insensitivity to concept drift, the classifier won't be adjusted promptly when the drift exists, as a result, the accuracy rate slumps in one moment and there are big changes.

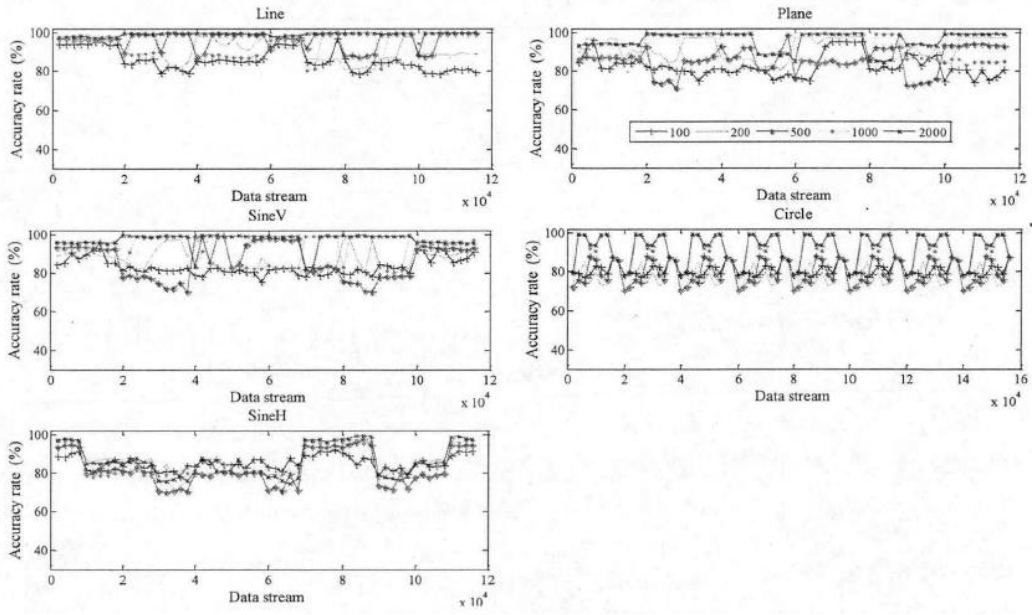

Figure 3. Results of the synthetic data sets

In addition, the artificial data is combined with the four other classification model for experiment, the experimental results is as shown in Table 2, the proposed model is represented by Omodel.

Table 2. Synthetic data set comparing experiments

Size	Omodel	IR-SVM	I-SVM	KNN	DT
100	85%	84%	84%	83%	81%
200	91%	84%	84%	83%	86%
500	96%	84%	84%	84%	86%
1000	92%	84%	84%	84%	84%
2000	95%	85%	84%	85%	85%

(a) Line

Size	Omodel	IR-SVM	I-SVM	KNN	DT
100	81%	71%	75%	80%	82%
200	91%	74%	76%	81%	82%
500	86%	76%	76%	82%	81%
1000	86%	82%	69%	82%	81%
2000	92%	79%	77%	82%	82%

(b) Plane

Size	Omodel	IR-SVM	I-SVM	KNN	DT
100	84%	70%	68%	80%	79%
200	85%	71%	68%	81%	82%
500	82%	74%	71%	81%	82%
1000	85%	74%	70%	82%	83%
2000	83%	74%	71%	82%	83%

(b) Sineh

Size	Omodel	IR-SVM	I-SVM	KNN	DT
100	82%	69%	69%	77%	78%
200	87%	68%	69%	78%	78%
500	86%	68%	69%	80%	80%
1000	90%	68%	68%	79%	79%
2000	93%	67%	68%	79%	80%

(d) Sinev

Size	Omodel	IR-SVM	I-SVM	KNN	DT

100	79%	56%	56%	77%	78%
200	81%	56%	56%	78%	78%
500	80%	56%	60%	78%	78%
1000	89%	57%	60%	78%	78%
2000	90%	80%	77%	81%	84%

(e) Circle

In Table 2, it puts forward the concept drift of data stream classification model based on the classification, classification effect is better than the other classifiers.

Through the analysis can get three conclusions: firstly, IR-SVM and I-SVM on artificial data classification effect is poor, the reason is that although the incremental learning can satisfy the data environment concept problems. But the concept drift suddenly cannot completely meet; secondly, K-NN and DT model, under different data block classification accuracy is relatively stable. Finally, in general, model classification accuracy rate increased along with the data block size.

4.3. Experiment results of real data

After validation with the use of real data, we made comparison between the proposed model and the other four ones to demonstrate the performance and validity. Results obtained by real data are portrayed in Figure 4.

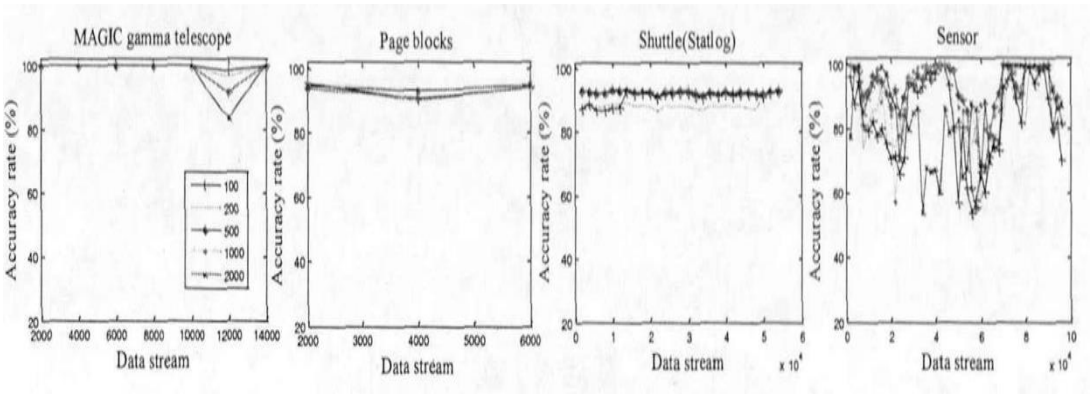

Figure 4. Results of the real-world data sets

The accuracy rate of the first three data, MAGIC, Page and Shuttle reaches to approximately 95%. But in MAGGIC, it's obvious that a sudden drift happened to the data in the later part and the rate is slightly reduced. With data block getting bigger, the rate becomes lower, suggesting that blocks in small size are highly sensitive to concept drift and thus the model adjusts the classifier more instantly than those in big size. In Sensor data set, since the data includes noise, the accuracy rate fluctuates a lot, but in general, when the volume is 1000, the rate has satisfactory compatibility and stability.

4.4. Experiment results of forgetting mechanism

In order to verify the effectiveness of the forgetting mechanism classifier pool, using Sensor data as the background, using five different data block experiment, and forget the formula 1 w parameter is equal to 2, ε is equal to 0.85. The experimental results are shown in Figure 4.

From Figure 5, the number of individual classifiers in the classifier pool decreases with the increase of the data block size. Specifically, when the block size is 100, the number of individual classifier pool fluctuates acutely, up to 17 classifier. However, when the block size

is 2000, the number of individual classifier is relatively stable, up to 10. Through the analysis, it can get two conclusions: firstly, forgetting mechanism can effectively control the number of individual classifier in the pool, it can reduce the occupied storage space: secondly, small block size is more likely to produce individual classifiers, so there are many classifiers, which also proved again that the small data block to concept drift is sensitive, Therefore, it frequently needs to create and forget the individual classifier.

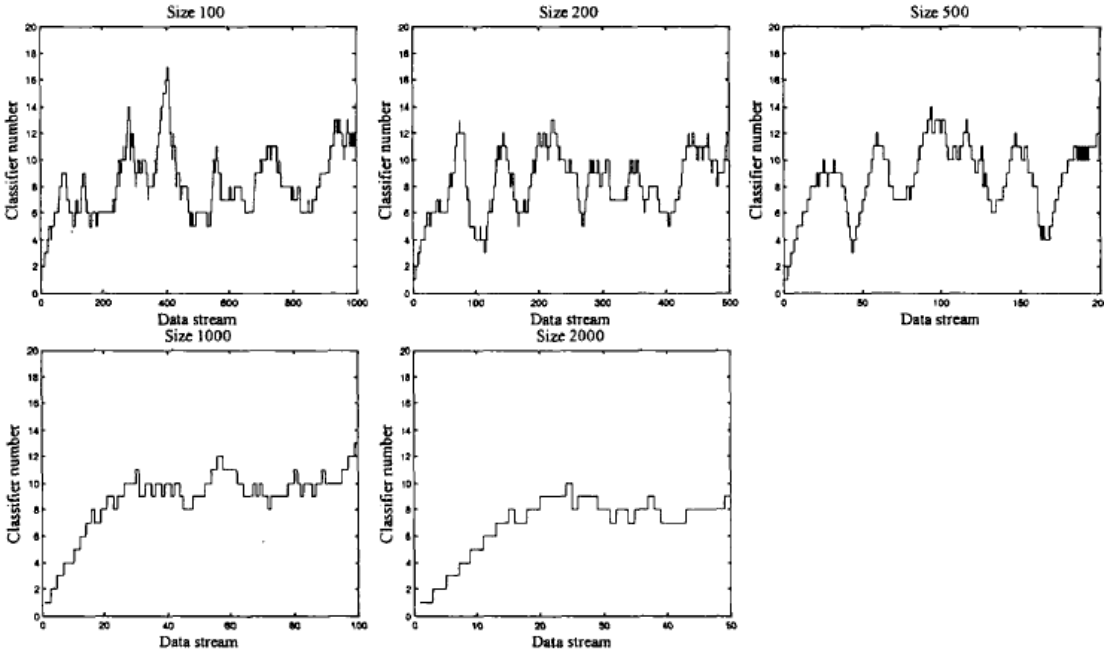

Figure 5. Results of the classifier pool forgetting strategy

5. Conclusion

On the basis of concept drift detection. It discussed the feasibility of the detecting technique for concept drift for different types of concept drift in data stream. Concept drift is detection methods by KL-distance algorithm.it proposed the classification model of concept drift for data flow. The model used KL-distance strategy to discover whether concept drift appeared in two blocks and invented a classifier pooling mechanism to preserve concepts. The proposed model reduced the updating times of classification model through the detection of concept drift, accelerated the speed of classification and avoided the influence of repeated concept drift on the model through the application of historical data as much as possible.

In addition, this paper also study on the method on concept drift visualization, and how to understand the concept from the data, through analysis of the conceptual level, it can be more in-depth understanding of the data stream itself contains information. It is proved by experiments that the above content, and prove the effectiveness of concept drift of data stream classification model and concept drift of visualization based on the proposed. An effective solution was proposed to construct and improve the classification model for data stream and a new idea was created on the data mining.

References

[1] I. Zliobaite, "Combining similarity in time and space for training set formation under concept drift", Intelligent Data Analysis, vol. 15, no. 4, (**2011**), pp. 589-611.

[2] M. M. Masud, C. Qing and L. Khan, "Addressing Concept-Evolution in Concept-Drifting Data Stream", 2010 IEEE 10th International Conference on Data Mining (ICDM), Sydney, Australia, (**2010**) December 13-17, pp. 929-934.
[3] G. J. Ross, N. M. Adams and D. K. Tasoulis, "Exponentially weighted moving average charts for detecting concept drift", Pattern Recognition Letters, vol. 33, no. 2, (**2012**), pp. 191-198.
[4] J. B. Gomes, P. A. C. Sousa and E. Menasalvas, "Tracking recurrent concepts using context", Intelligent Data Analysis, vol. 16, no. 5, (**2012**), pp. 803-825.
[5] S. H. Wang, S. Schlobach and M. Klein, "Concept drift and how to identify it", Journal of Web Semantics, vol. 9, no. 3, (**2011**), pp. 247-265.
[6] Z. Peng, Z. Xingquan and T. Jianlong, "Classifier and Cluster Ensembles for Mining Concept Drifting Data Streams", 2010 IEEE 10th International Conference on Data Mining (ICDM), Sydney, Australia, (**2010**) December 13-17, pp. 1175-1180.
[7] M. M. Masud, T. M. Al-Khateeb and L. Khan, "Detecting Recurring and Novel Classes in Concept-Drifting Data Streams", 2011 IEEE 11th International Conference on Data Mining (ICDM), Vancouver, Canada, (**2011**) December 11-14, pp. 1176-1181.
[8] M. Looks, A. Levine and G. A. Covington, "Streaming Hierarchical Clustering for Concept Mining", Aerospace Conference, 2007 IEEE, (**2007**) March 3-10, pp. 1-12.
[9] M. Masud, J. Gao and L. Khan, "Classification and novel class detection in concept-drifting data streams under time constraints", IEEE Transactions on Knowledge and Data Engineering, vol. 23, no. 6, (**2011**), pp. 859-874.
[10] T. Dasu, S. Krishnan, and S. Venkatasubramanian, "An information-theoretic approach to detecting changes in multi-dimensional data streams", In Proc. Symp. on the Interface of Statistics, Computing Science, and Applications, vol. 12, no. 2, (**2006**), pp. 1-22.

Author

Li Xiaofeng

He is an advanced member of China computer federation and he is the associate professor at Heilongjiang International University. His research interest includes Data mining, Text mining, intelligent algorithm so on.

