A Dynamic Drilling Sampling Method and Evaluation Model for Large-Scale Streaming Data

Peng Zhang, Zhaohui Zhang*, Chaochao Hu, Pengwei Wang
School of Computer Science and Technology
Donghua University
Shanghai 201600, China
zhzhang@dhu.edu.cn

Abstract—**The sampling method for real-time and high-speed changing streaming data is prone to lose the value and information of a large amount of discrete data, and it is not easy to make an efficient and accurate streaming data valuation. The SDSLA (Streaming Data Drilling Sampling Method Under Limited Access) sampling method based on mineral drilling exploration can streaming data valuation containing many discrete data in real-time, but when the range of discrete data in streaming data is irregular, it has low sampling accuracy for discrete data. Based on the SDSLA algorithm, we propose a dynamic drilling sampling method SDDS (Streaming Data Dynamic Drilling Sampling). This method takes well as the analysis unit dynamically changes the size and position of the well, and accurately predicts the position and range of discrete data. A new model SDVEM (Streaming Data Value Evaluation Model), is further proposed for data valuation, which evaluates the sample set from discrete, centralized, and overall dimensions. Experiments show that the method proposed in the paper uses neural network training and testing with a small sampling rate to obtain accuracy, recall, and F1 scores above 90%, which is higher than that of the SDSLA algorithm. In summary, the SDDS sampling method is beneficial to the training neural network models and evaluating the value characteristics of streaming data, which has essential research significance in big data valuation.**

Keywords-streaming data; dynamic drilling; data valuation; neural network

I. INTRODUCTION

In the era of big data, the data valuation is one of the most essential requirements of big data. C.R. Yin et al. [1] proposed data value can be considered a new form of value, and how to evaluate its value has become a new problem. For big data valuation, A.F. Haryadi et al. [2] proposed more than half of financial services organizations report that big data is not delivering the expected value. Big data usually exists in the form of streaming data, and it has the characteristics of high-speed and real-time, which brings severe challenges to sampling high-value data. Therefore, methods for sampling and streaming data valuation are needed.

Sampling is a required data analysis method in big data, and it plays an irreplaceable role in big data valuation. Currently, sampling methods for streaming data are mainly divided into two categories. The first category is unbiased sampling methods: stratified sampling [3], random sampling [4], reservoir sampling [5]. Unbiased sampling is random, and the streaming data obtained by sampling will lose some key information. The second category is biased sampling methods: probability density sampling[6]. Biased sampling can preserve many discrete data in streaming data but amplifies the impact of discrete data in the sample set.

To sum up, there are still some difficulties: 1. How to accurately predict the position and range of discrete data in streaming data? 2. Too much discrete data in the sample is not easy to use to evaluate the overall characteristics of streaming data. Given the above problems, the contributions of the paper are summarized as follows:

- We propose a dynamic sampling method SDDS, which takes well as the analysis unit dynamically changes the size and position of the well, and accurately predicts the position and range of discrete data.
- We propose a new streaming data valuation model SDVEM which evaluates the sample set from discrete, centralized, and overall dimensions.

Section II introduces related work. Section III depicts the SDDS method. Section IV depicts the streaming data valuation model. Section V describes the dataset and experimental analysis. Section VI summarizes the research results of the paper and expectations.

II. RELATED WORK

At this stage, some sampling methods can collect discrete data so that the sample set can be used as the training set of the neural network. T. Li et al. [7] proposed a new data transformation method, the KSB algorithm, that improved machine learning models' performance.

Assessing the value of data, P. J [8] pointed out how we can objectively, systematically, and quantitatively assess the value of data. F.J. Xu et al. [9] proposed a streaming data drilling sampling method (SDLSA) and an overall feature evaluation model of streaming data sets. The main idea of the SDLSA method is to use the skewness coefficient to locate the following well-drilling position, to perform in-well sampling and inter-well sampling respectively. The disadvantage is that when the range of discrete data in streaming data is irregular, the sampling accuracy of discrete data is low. Based on the SDLSA method, we propose a complete SDDS method which dynamically changes the size of the wells for predicting the range of discrete data and changes the size of the well-interval

Zhaohui Zhang is the correspongding author.

DOI reference number: 10.18293/SEKE2023-184

adaptively. Experiments show that the sampling accuracy of the SDDS method in discrete data is higher than that of SDLSA, the effect of evaluating the SDDS sample set from three dimensions of discrete, centralized, and overall is better than that of SDLSA, and the training effect of the sample set of SDDS for neural network models is better.

III. DYNAMIC DRILLING SAMPLING METHOD

Definition 1 (Streaming Data): The streaming data S is represented as:

$$S = \{(id_i,\ time_i,\ value_i)|1 \le i \le N \ and \ i \ \epsilon \ N^+\} \qquad (1)$$

Where id_i is the order of the i^{th} data, $time_i$ is the arrival time of the i^{th} data, $value_i$ is the value of the i^{th} data, N is the size of streaming data. An example of S is shown in Figure 1.

Definition 2 (Well): We use well as analysis units.The i^{th} well W_i is expressed as:

$$W_i = \{(id_j, time_j, value_j)|1 \le j \ \le WS_i \ and \ j \ \epsilon \ N^+\} \qquad (2)$$

Where WS_i is the size of the i^{th} well.

Definition 3 (Well-Interval): The i^{th} well interval WI_i is expressed as:

$$WI_i = \{(id_j, time_j, value_j)|id_{wi\_max} + 1 \le id_j \le id_{wi+1\_min} - 1\} \qquad (3)$$

Where id_{wi_max} is the largest id in the i^{th} well, id_{wi+1_min} is the smallest id in the $i + 1^{th}$ well.

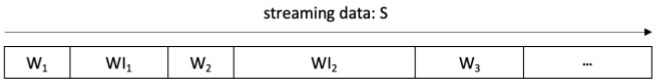


Figure 1. Streaming data schematic diagram

We propose the SDDS sampling method to obtain valuable discrete data in the streaming data. Firstly, setting the initial well and initial well-interval, k-means clustering [10] in the well. Secondly, calculating each class's sampling rate and well-interval by deviation coefficient. Thirdly, using intra-class unbiased sampling and inter-class biased sampling in the well, sampling equidistant in the well-interval. Finally, predicting the size of the following well through the correlation coefficient and the coefficient of variation. The specific sampling method is shown in Figure 2.

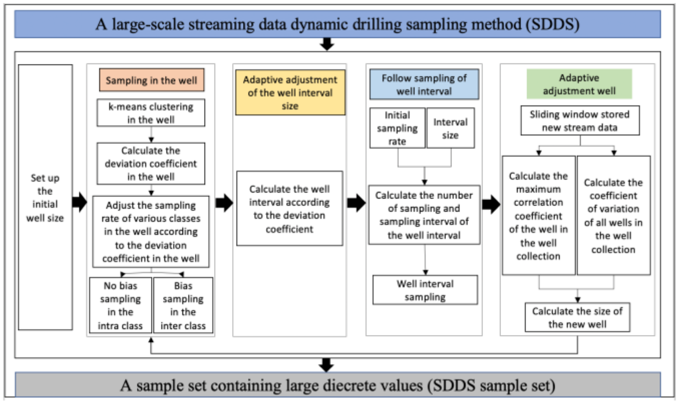

A large-scale streaming data dynamic drilling sampling method (SDDS)
Set up the initial well size
Sampling in the well
k-means clustering in the well
Calculate the deviation coefficient in the well
Adjust the sampling rate of various classes in the well according to the deviation coefficient in the well
No bias sampling in the intra class
Bias sampling in the inter class
Adaptive adjustment of the well interval size
Calculate the well interval according to the deviation coefficient
Follow sampling of well interval
Initial sampling rate
Interval size
Calculate the number of sampling and sampling interval of the well interval
Well interval sampling
Adaptive adjustment well
Sliding window stored new stream data
Calculate the maximum correlation coefficient of the well in the well collection
Calculate the coefficient of variation of all wells in the well collection
Calculate the size of the new well
A sample set containing large diecrete values (SDDS sample set)

Figure 2. The large-scale streaming data sampling

A. Dynamic Adjust Sampling Rate And Well-Position

To accurately predict the position of discrete data in the streaming data. Firstly, we use the k-means clustering to divide well-data into three categories, then carry out intra-class unbiased and inter-class biased sampling, increasing the sampling rate of the minority class. Assuming that the current is the i^{th} well, the initial sampling rate is p_{init}, and the adjustment formula of the sampling rate is as follows:

$$p = \begin{cases} 2 \times p_{init} \times |SK_i| , SK_i \ \epsilon \ (\text{-}1,\text{-}0.5) \text{ or } (0.5,1) \\ 2 \times p_{init} \times |SK_i| , SK_i \ \epsilon \ (\text{-}\infty,\text{-}1) \text{ or } (1,+\infty) \\ p_{init} , SK_i \ \epsilon \ [-0.5,0.5] \end{cases} \qquad (4)$$

If $SK_i \ \epsilon \ [-0.5,0.5]$, the sampling rate for all classes $p = p_{init}$; if $SK_i \ \epsilon$ (-∞,-1) or (1,+∞), increase the sampling rate of the two classes with fewer numbers p=$2 \times p_{init} \times |SK_i|$; if $SK_i \ \epsilon$ (-1,-0.5) or (0.5,1), increase the sampling rate for the least number of classes p=$2 \times p_{init} \times |SK_i|$.

Secondly, we use the deviation coefficient [11] to adjust the well-interval size dynamically, and the size of i^{th} well-interval formula WIS_i is as follows:

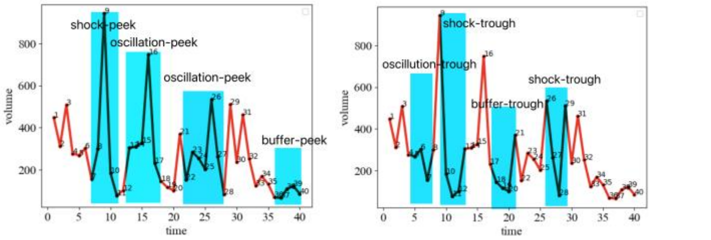
$$WIS_i = \begin{cases} \left\lceil \frac{WIS_{init}}{|SK_i|} \right\rceil , SK_i \ \epsilon \ (\text{-}\infty,\text{-}1) \text{ or } (1,+\infty) \text{ or } \ (\text{-}1,\text{-}0.5) \text{ or } (0.5,1) \\ 2 \times WIS_{init} , SK_i \ \epsilon \ [-0.5,0.5] \end{cases} \qquad (5)$$

Where WIS_i stands for the i^{th} well-interval size, SK_i stands for the deviation coefficient of the i^{th} well.

Specifically divided into two situations, if $SK_i \ \epsilon \ [-0.5,0.5]$, increase well-interval $WI_i = 2 \times WI_{init}$; if $SK_i \ \epsilon$ (-∞,-1) or (1,+∞) or (-1,-0.5) or (0.5,1) , reduce well-interval $WI_i = \lceil WI_{init}/|SK_i| \rceil$.

B. Dynamic Adjust The Well Size

To accurately predict the range of discrete data in the streaming data. Firstly, we define the peaks and troughs of streaming data through three features: 1.The slope changes very large; 2.Periodic changes; 3.The degree of dispersion is low. As shown in Figure 3(a), the peaks are divided into SP(Shock-Peek), OP(Oscillation-Peek), and BP(Buffer-Peek). As shown in Figure 3(b), the troughs are divided into ST(Shock-Trough), OT(Oscillation-Trough), and BT(Buffer-Trough). Where the SP and ST have a huge slope and the degree of dispersion will be huge; the OP and OT have periodic changes and the degree of dispersion will be relatively large; BP and BT have the lowest degree of dispersion of the well. When two different wells contain the same kind of peak or trough, the two wells have a certain self-similarity, and the degree of dispersion of the two wells will be very close.



(a) (b)

Figure 3. Streaming data peak and trough classification diagram

Secondly, we propose an AWS algorithm(Adaptive well sizing) to predict the range of discrete data in streaming data accurately, combining pearson correlation coefficient [12] and variation coefficient [13]. The AWS records the representative wells in the well-set, uses the sliding window to accept the newly arrived data, traverses the well-set, sets the sliding window size to the size of the different wells in the well-set, and set the size of the next well to the one with the highest correlation coefficient with the sliding window in the well-set. The specific algorithm is as follows:

Algorithm AWS: Adaptive Well-Size

```
Input: $W_{init}$ - Init Well; $\delta$ - Threshold Value of Correlation Coefficient; $WC$ - Well Collection; $SW$ - Sliding Window; $PCC$ - Pearson Coefficient Collection.
Output: $WS$
$WC$.add ($W_{init}$)
for data in $WC$:
    $SW$.clear ()
    $SW$.size = data.size
    $SW$.add ($SW$.size data after well interval)
    PC = Pearson correlation coefficient of data and $SW$
    $PCC$.add (PC)
PC_MAX = max(PCC)
if PC_MAX ≥ $\delta$:
  index = PCC.index(PC_MAX)
  $WS = WC[index]$.length
  $WC$.add(Latest well-data)
else:
  $SW$.clear()
  $SW$.size = data.size
  $SW$.add($SW$.size data after well interval)
  $SW\_COV$ = Variation coefficient of $SW$
  $AWC$ = Covariance of all wells
  if $SW_{COV} \geq 75\%\ AWC$:
    $WS = 2 \times W_{init}$.size
  elif $SW_{COV} \geq 50\%\ AWC$:
    $WS = 1.5 \times W_{init}$.size
  elif $SW_{COV} \geq 25\%\ AWC$:
    $WS = W_{init}$.size
  else:
    $WS = W_{init}$.size / 2
  $WC$.add(Latest well-data)
return $WS$
```

C. Dynamic Sampling Algorithm

We propose a SDDS algorithm to accurately predicts the position and range of discrete data. Firstly, setting the initial well and initial well-interval, k-means clustering in the well. Secondly, calculating each class's sampling rate and well-interval by deviation coefficient. Thirdly, using intra-class unbiased sampling and inter-class biased sampling in the well, sampling equidistant in the well-interval. Finally, predicting the size of the following well through the AWS algorithm.

Algorithm SDDS: Streaming Data Dynamic Sampling

```
Input: $WS_{init}$ - Init Well-Size; $S$ - Streaming Data; $WIS_{init}$ – Init Well Interval; $p_{init}$ – Init Sampling Rate
Output: $SS$ – Sample Set
1.  Set $W_{init}$.length = $WS_{init}$
2.  Set $WI_{init}$.length = $WIS_{init}$
3.  K-means clustering for $W_{init}$, get three classes of data
4.  Calculate the p for each class in $W_{init}$ according to formula (4)
5.  Sample by steps 13-23 of this algorithm and add to SS
6.  Equidistant sampling of well interval data
7.  while S is generating:
8.    Get the size of the next well by AWS algorithm: $WS_{next}$
9.    Set $W_i$.length = $WS_{next}$
10.   K-means clustering for $W_i$, get three classes of data
11.   Calculate the $SK_i$ of $W_i$
12.   Calculate the p of three classes in $W_i$ by formula (4)
13.   if $SK_i \epsilon [-0.5,0.5]$:
14.     Sampling rate for all classes is $p_{init}$
15.     $SS$.add(Data obtained from $W_i$ by reservoir sampling)
16.   elif $SK_i \epsilon$ (-∞, -1) or (1, +∞):
17.     The two smaller classes' sampling rate is 2 × $p_{init}$ × $|SK_i|$
18.     The largest class's sampling rate is $p_{init}$
19.     $SS$.add(Data obtained from $W_i$ by reservoir sampling)
20.   elif $SK_i \epsilon$ (-1, -0.5) or (0.5,1):
21.     The smallest class's sampling rate is 2 × $p_{init}$ × $|SK_i|$
22.     The other two classes' sampling rate is $p_{init}$
23.     $SS$.add(Data obtained from $W_i$ by reservoir sampling)
24.   Calculate the $WIS_i$ of $WI_i$ by formula (5)
25.   Set $WI_i$.length = $WIS_i$
26.   $SS$.add(Data obtained from $WI_i$ by equidistant sampling)
27.   $SS$.sort()
28. return SS
```

IV. SDVEM Evaluation Model

We propose SDVEM evaluation model to streaming data valuation. The specific model is shown in Figure 4.

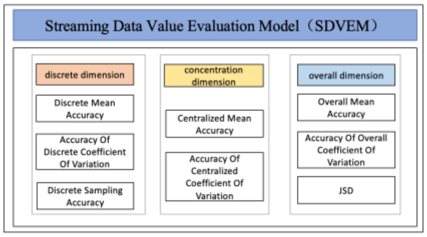

Streaming Data Value Evaluation Model（SDVEM）
discrete dimension
Discrete Mean Accuracy
Accuracy Of Discrete Coefficient Of Variation
Discrete Sampling Accuracy
concentration dimension
Centralized Mean Accuracy
Accuracy Of Centralized Coefficient Of Variation
overall dimension
Overall Mean Accuracy
Accuracy Of Overall Coefficient Of Variation
JSD

Figure 4. SDVEM Evaluation Model

A. Discrete Dimension Evaluation Sample Set

The discrete data of the raw streaming dataset $DDRD$ and the discrete data of the sample set $DDSS$ are as follows:

$$DDRD = \left\{RD_i \middle| RD_i \geq \overline{RD} \times \delta_{upper}\ or\ RD_i \leq \overline{RD} \times \delta_{down}\right\} (6)$$

$$DDSS = \left\{SS_i \middle| SS_i \geq \overline{SS} \times \delta_{upper}\ or\ SS_i \leq \overline{SS} \times \delta_{down}\right\} \quad (7)$$

Where RD_i is the i^{th} data of raw streaming data, δ_{upper} and δ_{down} are the threshold to decide whether it is discrete data, $\overline{RD}$ is the mean of raw streaming data, SS_i is the i^{th} of the sample set, and $\overline{SS}$ is the mean of the sample set.

Definition 4 DMA(Discrete Mean Accuracy): refers to the accuracy rate of estimating the mean value of the $DDRD$ attribute value with the mean value of the $DDSS$ attribute value. The formula of DMA is as follows:

$$DMA = 1 - \left(\frac{|\overline{DDRD} - \overline{DDSS}|}{\overline{DDRD}}\right) \times 100\% \tag{8}$$

Definition 5 ADCV(Accuracy of Discrete Coefficient of Variation): refers to the accuracy rate of estimating the coefficient of variation of the $DDRD$ attribute value by using the coefficient of variation of the $DDSS$ attribute value. The formula of $ADCV$ is as follows:

$$ADCV = 1 - \left(\frac{|CV_{DDRD} - CV_{DDSS}|}{CV_{DDRD}}\right) \times 100\% \tag{9}$$

Definition 6 DSA(Discrete Sampling Accuracy): refers to the ratio of the intersection number of the $DDSS$ attribute value and the $DDRD$ attribute value to the $DDSS$ length. The formula of DSA is as follows:

$$DSA = \frac{len(DDSS \cap DDRD)}{len(DDSS)} \times 100\% \tag{10}$$

B. Centralized Dimension Evaluation Sample Set

The centralized data of the raw streaming dataset $CDRD$ and the centralized data of sample set $CDSS$ are as follows:

$$CDRD = \left\{RD_i \middle| \overline{RD} \times \delta_{down} \leq RD_i \leq \overline{RD} \times \delta_{upper}\right\} \tag{11}$$

$$CDSS = \left\{SS_i \middle| \overline{SS} \times \delta_{down} \leq SS_i \leq \overline{SS} \times \delta_{upper}\right\} \tag{12}$$

Definition 7 CMA(Centralized Mean Accuracy): refers to the accuracy of estimating the mean value of the $CDRD$ with the mean value of the $CDSS$. The calculation formula of CMA is as follows:

$$CMA = 1 - \left(\frac{|\overline{CDRD} - \overline{CDSS}|}{\overline{CDRD}}\right) \times 100\% \tag{13}$$

Definition 8 ACCV(Accuracy of the Centralized Coefficient of Variation): refers to the accuracy of estimating the coefficient of variation of the $CDRD$ with the coefficient of variation of the $CDSS$. The calculation formula of $ACCV$ is as follows:

$$ACCV = 1 - \left(\frac{|CV_{CDRD} - CV_{CDSS}|}{CV_{CDRD}}\right) \times 100\% \tag{14}$$

C. Overall Dimension Evaluation Sample Set

Definition 9 OMA(Overall Mean Accuracy): refers to the accuracy of estimating the original streaming data mean by using the sample set mean. The calculation formula of OMA is as follows:

$$OMA = 1 - \left(\frac{|\overline{RD} - \overline{SS}|}{\overline{RD}}\right) \times 100\% \tag{15}$$

Where $\overline{RD}$ is the mean of the raw streaming data, $\overline{SS}$ is the mean of the sample set.

Definition 10 AOCV(Accuracy of the Overall Coefficient of Variation): refers to the accuracy of estimating the coefficient of variation of raw streaming data with the coefficient of variation of sample set. The calculation formula of $AOCV$ is as follows:

$$AOCV = 1 - \left(\frac{|CV_{RD} - CV_{SS}|}{CV_{DDRD}}\right) \times 100\% \tag{16}$$

V. Experiments and Analyses

In this section, we introduced the dataset, analyzed the impact of the SDVEM model on different parameters, and compared the SDDS algorithm with the SDSLA algorithm.

A. Experimental Dataset

TABLE I. EXPERIMENTAL DATASET

DataSet	Data Volume	Class
HSI	87645	0
NEWS	276336	0
HSI1	55442	3
HSI2	55736	3

As shown in TABLE I, we selected four datasets HSI, NEWS, HSI1 and HSI2 to verify the effectiveness of the SDDS algorithm, among them HSI and NEWS are real datasets, HSI1 and HSI2 are synthetic datasets. In addition, we according to the degree of dispersion of experimental dataset, δ_{upper} is set to 1.5, and δ_{down} is set to 0.5.

B. Influence Analysis of Different Parameters

To prove that the AWS algorithm can predict the position and range of discrete data, the sample set obtained by the SDDS algorithm can preserve the discrete data and reflect the centralized and overall characteristics of the original streaming data. We use the SDVEM model to conduct experimental evaluations on HSI and NEWS datasets under different parameters, a detailed analysis of parameter δ in the AWS algorithm and parameters WS_{init} , WIS_{init} and p_{init} in the SDDS algorithm. The larger the parameter δ, the higher the similarity standard between wells. Parameters WS_{init} and WIS_{init} should be set based on the distribution characteristics of discrete data in the streaming data. The larger the parameter p_{init}, the smaller the proportion of discrete data in the sample set.

Firstly, WS_{init}, WIS_{init}, and p_{init} are set as 20, 20, and 0.1, respectively, the parameter δ in the AWS algorithm was adjusted differently. The experimental results are shown in TABLE II.

TABLE II. EVALUATION RESULTS UNDER DIFFERENT PARAMETERS δ OF THE HSI AND NEWS (%)

	HSI					NEWS				
δ	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**
DMA	94	93	90	91	87	96	93	96	97	98
ADCV	92	94	96	94	95	99	99	98	98	97
DSA	96	96	96	97	97	99	99	99	99	99
CMA	94	94	92	92	91	95	94	94	93	93
ACCV	1	1	1	1	1	97	97	1	1	1
OMA	79	69	71	71	72	75	64	65	65	65
AOCV	90	89	87	89	88	90	90	89	89	89

Secondly, δ, WIS_{init}, and p_{init} are set as 0.2, 20, and 20, respectively, different adjustments are made to the parameter WS_{init} in the SDDS algorithm, and the experimental results are presented in TABLE III.

TABLE III. EVALUATION RESULTS UNDER DIFFERENT PARAMETERS WS_{init} OF THE HSI AND NEWS (%)

	HSI					NEWS				
WS_{init}	**10**	**20**	**30**	**40**	**50**	**10**	**20**	**30**	**40**	**50**
DMA	79	94	96	98	99	98	94	89	86	84
ADCV	1	94	90	85	89	99	95	95	96	93
DSA	96	96	95	96	97	99	99	99	98	97
CMA	91	94	93	92	93	93	94	95	96	97
ACCV	1	1	1	1	94	1	97	97	97	93
OMA	84	69	65	60	60	66	60	60	59	55
AOCV	86	89	91	92	92	89	90	92	94	94

Thirdly, δ, WI_{init}, and p_{init} are set as 0.2, 10, and 0.1, respectively, different adjustments are made to the parameter WIS_{init} in the SDDS algorithm, and the experimental results are presented in TABLE IV.

TABLE IV. EVALUATION RESULTS UNDER DIFFERENT PARAMETERS WIS_{init} OF THE HSI AND NEWS (%)

	HSI					NEWS				
WIS_{init}	**10**	**20**	**30**	**40**	**50**	**10**	**20**	**30**	**40**	**50**
DMA	89	79	80	81	78	98	98	98	97	92
ADCV	98	99	1	98	99	1	96	96	96	95
DSA	97	96	98	97	97	99	99	99	96	99
CMA	95	91	91	92	92	95	93	93	92	92
ACCV	1	1	1	1	1	1	1	1	1	1
OMA	89	87	89	86	89	65	71	72	73	79
AOCV	85	82	83	82	84	89	88	88	88	89

Fourthly, $\delta = 0.2$, $WS_{init} = 20$, and $WIS_{init} = 20$ are set as 0.2, 20, and 20, respectively, and different adjustments are made to the parameter p_{init} in the SDDS algorithm, and the experimental results are exhibited in TABLE V.

TABLE V. EVALUATION RESULTS UNDER DIFFERENT PARAMETERS p_{init} OF THE HSI AND NEWS (%)

	HSI					NEWS				
p_{init}	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**
DMA	94	84	78	75	72	94	97	91	87	85
ADCV	94	99	99	1	98	99	99	96	95	94
DSA	95	96	96	98	99	99	99	99	99	99
CMA	94	92	91	90	90	95	93	93	92	91
ACCV	1	1	1	1	1	97	1	1	1	1
OMA	69	79	85	89	92	62	75	82	85	87
AOCV	89	89	89	90	92	89	90	91	91	91

It can be seen from TABLE I, TABLE II, TABLE III, TABLE IV and TABLE V that the evaluation accuracy of the SDDS algorithm in the discrete and centralized dimension is almost all above 90%, and the evaluation accuracy of AOCV is almost above 85%. It proves that the sample set obtained by the SDDS algorithm can reflect the discrete, centralized and overall characteristics of the raw streaming data.

C. Comparison of Experimental Results of Streaming Data Value Evaluation Model

To prove that the SDDS algorithm can evaluate the value of streaming data from three dimensions: discrete, centralized, and overall. We use real data sets HSI and NEWS to conduct comparative experiments on SDDS and SDSLA sampling. The experimental results are shown in Figure 5 and Figure 6.

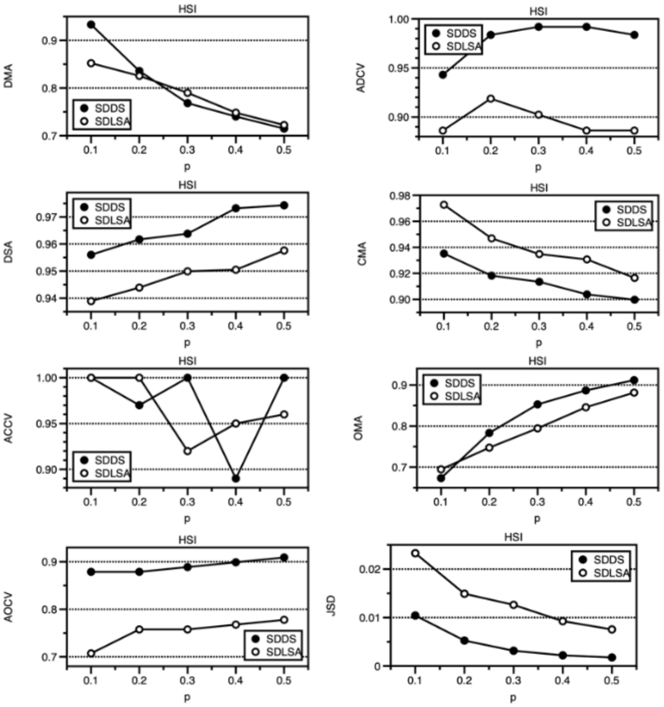


Figure 5. Evaluation results under different parameters p_{init} of the HSI

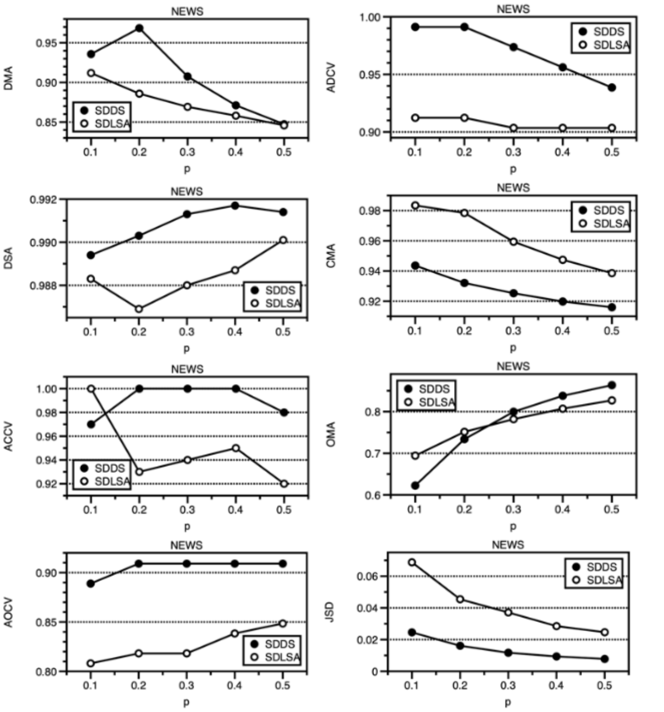


Figure 6. Evaluation results under different parameters p_{init} of the NEWS

It can be seen from Figure 5 and Figure 6 that the evaluation accuracy of the sample set obtained by the SDDS algorithm is very high in the discrete, centralized, and overall dimensions, and the accuracy of the five evaluation indicators ADCV, DSA, CMA, ACCV, and AOCV almost both are above 90%, the evaluation accuracy of DMA is almost above 85%, and the JSD is also very low, indicating that the probability distribution of the sample set and the original streaming dataset is very close. Compared with the SDLSA algorithm, the evaluation accuracy of the SDDS algorithm is almost higher in the three dimensions discrete, centralized, and overall.

D. Comparison of Neural Network Training Effect

To prove that the sample set obtained by the SDDS algorithm is beneficial to reduce the amount of data required for neural network model training. We use HSI1 and HSI2 to divide the dataset into 80% training set and 20% test set, then samples the training set, and conducts comparative experiments on the datasets before and after sampling.

Based on the above experiments, we select parameters with the best evaluation result, setting δ, WS_{init}, WIS_{init} as 0.2, 20, and 20, respectively, different adjustments were made to the parameter p_{init} in the SDDS algorithm, the experimental results are shown in Figure 7.

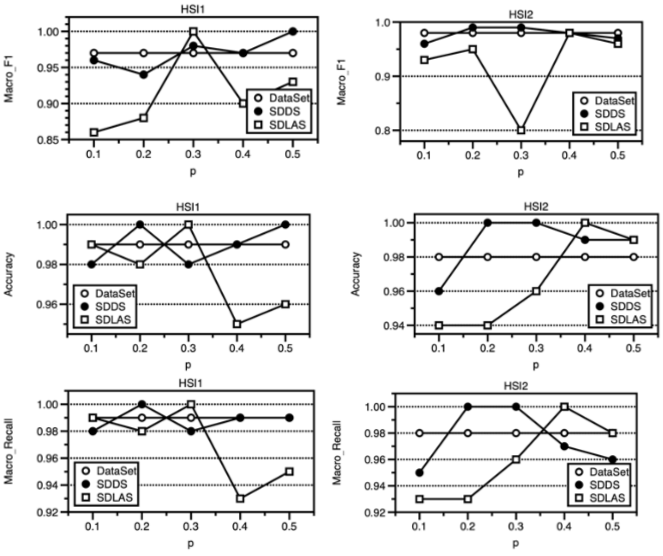


Figure 7. Training F1, Recall and Accuracy values of raw streaming data, the SDSLA and SDDS sample sets on the HSI1 and HSI2

It can be seen from Figure 7 that the F1 value and recall rate of the neural network model trained with the sample set obtained by the SDDS algorithm are almost higher than 90%, and the accuracy is almost higher than 95%, which is almost the same as using the original streaming dataset training. Moreover, the sample set obtained by the SDDS algorithm is better than the SDLSA algorithm in the effect of training the neural network model.

VI. CONCLUSION AND FUTURE WORK

In this paper, we propose the SDDS algorithm to predict the position and range of discrete data and obtain a sample set containing more discrete data. It can describe the original stream data set's value characteristics and be well used in training neural network models. The streaming data value evaluation model SDVEM fully and detailedly analyzes the sample set from three dimensions: discrete, centralized, and holistic, which has essential research significance for the value evaluation of streaming data in the field of big data. In future work, we will further increase the dimension of value evaluation of streaming data and more comprehensively evaluate the value of streaming data.

ACKNOWLEDGMENT

This work was supported in part by the Shanghai Science and Technology Innovation Action Plan Project (22511100700).

REFERENCES

[1] Chuanru YIN, Tao JIN, Peng ZHANG, Jianmin WANG, Jiayi CHEN. Assessment and pricing of data assets:research review and prospect[J]. Big Data Research, 2021, 7(4): 14-27.

[2] A. F. Haryadi, J. Hulstijn, A. Wahyudi, H. van der Voort and M. Janssen, "Antecedents of big data quality: An empirical examination in financial service organizations," 2016 IEEE International Conference on Big Data (Big Data), Washington, DC, USA, 2016, pp. 116-121, doi: 10.1109/BigData.2016.7840595.

[3] C. Cervellera and D. Macciò, "Distribution-Preserving Stratified Sampling for Learning Problems," in IEEE Transactions on Neural Networks and Learning Systems, vol. 29, no. 7, pp. 2886-2895, July 2018.

[4] Jeffrey Scott Vitter, "Random Sampling With Reservior", ACM Transactions On Mathematical Software, vol. 11, no. 1, pp. 37-57, March 1985.

[5] Z.D. Hu, Y. G. Ren and X. Yang, “Bias sampling data stream based on sliding window density clustering algorithm research,” Computer Science, vol. 40, no. 9, pp. 254-256, 269, 2013.

[6] Tang B, He H. KernelADASYN: Kernel based adaptive synthetic data generation for imbalanced learning[C]//2015 IEEE congress on evolutionary computation (CEC). IEEE, 2015: 664-671.

[7] T. Li, S. Fong, Y. Wu and A. J. Tallón-Ballesteros, "Kennard-Stone Balance Algorithm for Time-series Big Data Stream Mining," 2020 International Conference on Data Mining Workshops (ICDMW), 2020, pp. 851-858.

[8] Pei J. A survey on data pricing: from economics to data science[J]. IEEE Transactions on knowledge and Data Engineering, 2020, 34(10): 4586-4608.

[9] Fujuan Xu. A large-scale mathematics-based method for giving a large-scale mathematical model to Koi's formula[D]. Donghua University, 2022.DOI:10.27012/d.cnki.gdhuu.2022.000338.

[10] Y. Zhang, K. Tangwongsan and S. Tirthapura, "Fast Streaming kk-Means Clustering With Coreset Caching," in IEEE Transactions on Knowledge and Data Engineering, vol. 34, no. 6, pp. 2740-2754, 1 June 2022, doi: 10.1109/TKDE.2020.3018744.

[11] Bowley A L. The standard deviation of the correlation coefficient[J]. Journal of the American Statistical Association, 1928, 23(161): 31-34.

[12] Cohen I, Huang Y, Chen J, et al. Pearson correlation coefficient[J]. Noise reduction in speech processing, 2009: 1-4.

[13] C. -M. Hsu and M. -S. Chen, "On the Design and Applicability of Distance Functions in High-Dimensional Data Space," in IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 4, pp. 523-536, April 2009.

[14] Barz B, Rodner E, Garcia Y G, et al. Detecting regions of maximal divergence for spatio-temporal anomaly detection[J]. IEEE transactions on pattern analysis and machine intelligence, 2018, 41(5): 1088-1101.