# A Data Flow Model to Solve the Data Distribution Changing Problem in Machine Learning

Bo-Wen SHANG[1,a], Ke WANG[1]

[1]*College of Computer Science, National University of Defense Technology Changsha, 410072, China*

**Abstract.** Continuous prediction is widely used in broad communities spreading from social to business and the machine learning method is an important method in this problem.When we use the machine learning method to predict a problem.We use the data in the training set to fit the model and estimate the distribution of data in the test set.But when we use machine learning to do the continuous prediction we get new data as time goes by and use the data to predict the future data,there may be a problem. As the size of the data set increasing over time, the distribution changes and there will be many garbage data in the training set.We should remove the garbage data as it reduces the accuracy of the prediction.The main contribution of this article is using the new data to detect the timeliness of historical data and remove the garbage data.We build a data flow model to describe how the data flow among the test set,training set ,validation set and the garbage set and improve the accuracy of prediction .As the change of the data set,the best machine learning model will change.We design a hybrid voting algorithm to fit the data set better that uses seven machine learning models predicting the same problem and uses the validation set putting different weights on the learning models to give better model more weights.Experimental results show that,when the distribution of the data set changes over time,our time flow model can remove most of the garbage data and get a better result than the traditional method that adds all the data to the data set;our hybrid voting algorithm has a better prediction result than the average accuracy of other predict models.

## 1 Introduction

Machine learning and deep learning are the most commonly used methods through training by large data set with high dimensions to predict the unknown target. As some data is continuous, in many cases, we use the training data set to predict the unknown test data set. Over a period of time, the test data set will be known and added to the training data set,Then the new training data set is used to predict the new unknown test data, as in [1]. So we need a predictive model that can evolve with the time.However, with the change of the data set, the best machine learning model and the best hyper parameters will change. The common method is to split the data set into three parts: training set(used for model fitting),validation set(used for model test and get the estimated score),and test set(used for final model assessment).This method needs to refit all the data set and try all the machine learning methods to find the best model,so it costs too much if the data set updates frequently. In [2],they propose a new methods bases on a new data splitting strategies, but this method needs a lot of data.

In some prediction problems, such as the trend prediction,the distribution of the data set changes as the time goes by, so we should remove the garbage data. In this paper, we propose a data flow predictive model based on the voting algorithm. We use different machine learning models to fit the data sets, and use a new validation set to give every model a score, then we use the history data to give each methods a weight for each method to vote the result. We use our new data to test the history data and let the data flow among different data sets to use data efficiently(e.g. [3]–[6]).

## 2 The Data Flow Model

The main research of the data flow model is to find the best training set through the way that let the garbage data flow out of the training set and let the useful data flow into the training set.

The process is like getting new knowledge every day that we use our experiments to checkout if the message is true or not.We will forget the error message that we think it's wrong and remember the message that we think it's true.The message that we think it's true will become our

[a] Corresponding author: 819089115@qq.com

knowledge. Then we use our knowledge to predict new message and to judge whether it is true or not.

When we use a machine learning model to continuous predict the unknown data set that updates with the time, we can use the data flow model. The process is using the training set to predict the unknown data in the test set ,over a period of time, the unknown data will be known,then we use the data to update our training set and adjustment with our machine learning model.

As time goes by,the unknown data set will be known and the distribution of the data set may change. When we use all the known data to train our machine learning model and predict the new unknown data, there may be the a problem: whether all the known data is useful to our prediction? Research shows that for some data with strong timeliness, the distribution of the data will change with the time, and the distribution of the new data is different from the old data, so some old data will not be useful to predict the new data. Then the training set should be updated following the coming of the new data.

Machine learning model fits to the training data set to get the distribution of the data, but some data in the training data set can't reflect the distribution of the data in the test set. For example, some error data in the data set and some data out of date. Here we call them garbage data and use db to express the garbage data set.

Let some of the data in the test set at time T flow into the training set at time T+1 as the target column in the test set is known at time T+1.In order to distinguish the data join the training set at different times, we give them a time label. For example, the test set at time T will have a time label T.To promptly clean up the garbage data, we use the data with the new time label to test the other data in the data set and find the garbage data.Then we remove the garbage data and use the training set to predict the test set.

In data flow model, there are two problems to solve. The first one is that finding the data that flow from the training set to the garbage data set as we want to drop the garbage data that may be harmful to the predict. The second one is that finding the data that flow from the test set to the training set as we can use the useful data to predict the future data.

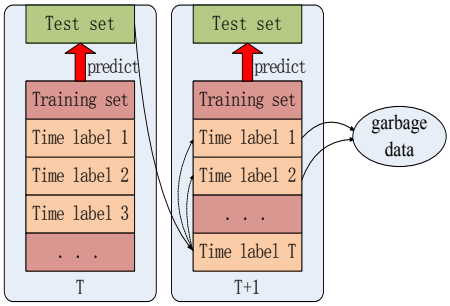


**Figure 1.** The date flow model

As time T+1,the target of the test set with time label T will be known.Then we can add some of the data to the training set. As the distribution of the data maybe change, the old data may have different distribution from the new data.At the same time,the distribution of the data will change.But only the test data with the same distribution with the training data that can be predict.It will cause a problem that the data in the training set has different distribution from the test data will reduce the accuracy of prediction.

In order to indicate the time that the data joins the training set,we give the data different time labels.For example,the data that joins the data set first will have the time label as t1,and the data set join the data set last have the time label tn.This time label is different from the time label T.As the distribution of the data changes over time,the oldest data may have the most different distribution from the new data.Here we define a difference function D to measure the difference between two data set's distribution:

$$D(t_1, t_n) \tag{1}$$

Firstly,we calculate the difference function between the data that have the same target with different time labels and use d(ti,tn) to express the result.Then we calculate the difference function between the data have the different target and use r(dp,dq,ti,tj) to express the value.If the value of D(ti,tn) is bigger than the biggest r(dp,dq,ti,tj), the data with the time label t flow is out.

But in practice,it's very hard to calculate the difference function as the distribution function of the data set is unknown.However,we can validates the history data with different targets and labels reversely. If there is enough new data,we use the new data as the training data and use all the history data as the test date.Then we divide the test error by the time label.If the test error of a data set is bigger than a threshold, the data with the label flow is out.

In the data flow model,we use the new data set as the training set and all the history data set with different time labels as the test set and use the predict result to decide the data flow.Before that we define a simple two classification machine learning model.The data set of the model have two kinds of data that with the target A and B,the distribution of class A is a and the distribution of class B is b.We use the confusion matrix of the predict result to present the test error.In the confusion matrix,each column of the matrix represents the instances in a predicted class while each row represents the instances in an actual class as in [5]. We build a confusion matrix for each data set with different time label.In the confusion matrix:

**Table 1.** The confusion matrix table

| **Predictive _class_A** | **Predictive _class_B** | |
|---|---|---|
| True_class _A | False_class _A | Recall of class_A(RA) |

| False_class _B | True_class _B | Recall of class_B(RB) |
|---|---|---|
| Precision of class_A(PA) | Precision of class_B(PB) | Likelihood of class_A(LA) Likelihood of class_B(LB) |

Recall of class_A:

$$\frac{\sum True\_class\_A}{\sum Class\_A} \tag{2}$$

Recall of class_B:

$$\frac{\sum True\_class\_B}{\sum Class\_B} \tag{3}$$

Accuracy:

$$\frac{\sum True\_class\_A+\sum True\_class\_B}{\sum Total\_population} \tag{4}$$

Precision of class_A:

$$\frac{\sum True\_class\_A}{\sum predictive\_class\_A} \tag{5}$$

Precision of class_B:

$$\frac{\sum True\_class\_B}{\sum predictive\_class\_B} \tag{6}$$

Likelihood of class_A:

$$\frac{RA}{RB} \tag{7}$$

Likelihood of class_B:

$$\frac{RB}{RA} \tag{8}$$

Through RA we can know the different between the new data set and the data set with the special time label. If RA is lower than the half of the maximum value, we can think that the data set with the time label is out of date.Some times,the LA or LB is very high,and one of the PA or PB is very low, then we should retain one kind of data and make the other kind flow out.

In the data flow model, we use the unknown data set as the test set and the history data set as the train when we predict the unknown target. After a period of time, the unknown target will be known and then the data set will be used as a part of the training set and we call it the new data set. Before using the new data set as the training set ,we use it to find which data set is out of date.After that ,the new data set and the history data set will be used to train the machine learning model.

## 3 Hybrid Voting Algorithm Based on History Data

In practice, the data set that we use to predict increases every day. Such as in some shopping website, users do different things every day. If we want to use the behavior data to predict their future behaviors, we should add them to our data set. However, in many cases ,the best machine learning method we use is based on experiment ,it will change follow the data set's change. To solve this problem, a method is to add the daily data to the data set and fit all the machine learning methods to find the best one. But this method is not only complex, but also lose a very important score for every method's accuracy rate in history. To solve this problem, we use a hybrid voting algorithm. In this method, we use different processes to run different machine learning models and use the new data to improve them in parallel. All the models will take part in the vote to decide the predict result, but the weights of them is different. Weight of models is decided by the score in the history. In the supervised machine models, the data set is divided into three parts: the training set d, the validation set v and the test set t. As we get the new data set n, the data in all data set will flow. The training data set d consists of n instances.

$$d = \{(x_i, y_i)\}_{i=1}^{n} \tag{9}$$

In the training data set,xi is a d-dimensional feature vector and yi is the corresponding known label.Our task is to find a function f:

$$f : X \to Y, x_i \in X, y_i \in Y \tag{10}$$

Then we want to get the f can perform better on an independent test data set.we define a loss function to measure the test errors(errors between target response y and the prediction f(x)):

$$L(y, \hat{f}(x)) \tag{11}$$

Because the training data set d is fixed,so the test error is refer to the special training data set d.So we define the test error when we use the training data set d :

$$E_d = T_{(x,y)}[L(y, \hat{f}(x)) \mid d] \tag{12}$$

In the formula above , T(x,y) is the test data set we randomly select.Assessment the Ed is a major method to measure the performance of the machine learning model and tune the parameters of the model.Our target is to chose the best machine model and tune the parameters to reduce

the test error. In the same time,the training error often be ignored.

$$E_t = \frac{1}{n}\sum_{i=1}^{n} L(y_i, \hat{f}_\theta(x_i)) \qquad (13)$$

The training error Et is the average loss of the training data. We can drop the training error to zero if we improve the model. But the training error isn't a good measure of the test error and a zero training error is over-fit to the training data set.

We want to choose a machine learning model that can have less test error, but we should measure the model with the training error. The training error is decided by the training set.With the change of the training set, we must change our machine learning model to get a better result. If we change our machine learning model, we can't compare the results on different data sets and we don't know the prediction result when we use the other machine learning model.

So we use a hybrid voting algorithm to reduce the impact of the using of different models on the results of the experiment .In this method ,we split the behavior data into three parts: training set(used for model fit),validation set(used for model test and get the estimated score),and test set(for final model assessment) as in [4].We use different predictive models to fit the behavior features and use the validation set to get the estimated score of each models, at last, predict the test set through voting by all the models. The weight of each model in voting is based on the score that they get when they predict the validation set.

| Training set | Validation set | Test set |
|---|---|---|

**Figure 2.** The structure of data sets

For example,method A get score 24,35,13,24,32 in the last five days,so today method A's score is :

(1×24+2×35+3×13+4×24+5×32)/15

In the same time ,the score is method's weight.Then we use all the method's results to predict a binary problem,all the methods' result and weight is:

**Table 2.** The methods' result and weight

| method | A | B | C | D | E |
|---|---|---|---|---|---|
| weight | 20 | 30 | 25 | 35 | 15 |
| result | 0 | 1 | 0 | 1 | 0 |

The result is that three methods votes 0,two methods votes 1.Add the weight of different methods ,the result 0's total weight is 60 and the total weight of 1 is 65,so we think 1 is the better result.

## 4 Case Study

We use a time varying data sets and contract two models. The data set contains two kinds of data with the target A and B. Each kind of data has 5 different time labels. Same kind of data with different time labels have the distribution changes continuously over time. Each item of the data has two features, a target and a time label.

We use the scikit-learn kit,the make_blobs function to generate data set.The centers of the class A with different time label is[[0,3],[1,2.8],[2,2.6],[3,2.4],[4,2.2],[5,2]]. The centers of the class B with different time label is[[5,2],[4,1.8],[3,1.6],[2,1.4],[1,1.2],[0,1]].The cluster is 0.5 and every time label will use 1000 samples.

In the traditional machine learning model,we use the history data as the training set and use the new data as the test set.As the time goes by,we have more and more new data,if we add the data to our data set without choose and not reduce the history data,our predict accuracy will drop.

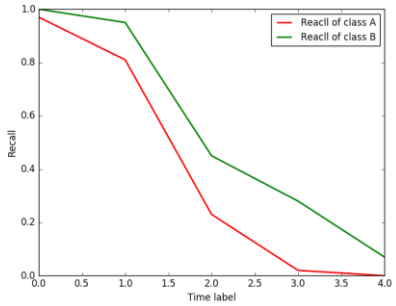


**Figure 3.** The recall of the traction methods

In this chart,we add the new data to the history data set,as the distribution of the data changes continuously over time,the recall of the data is dropping.

Why we have more data but the predict accuracy is down?That's because the distribution of the data set is changing with the time goes by,the new distribution of one kind of data may be more like that of a data set in the history,and some of the history data won't be useful or even harmful to our prediction result.Then we should find the useless data and remove them.

Here we design a method that use the new data set to test the history data set.We use the new data set as the training set and the history data set as the test set.

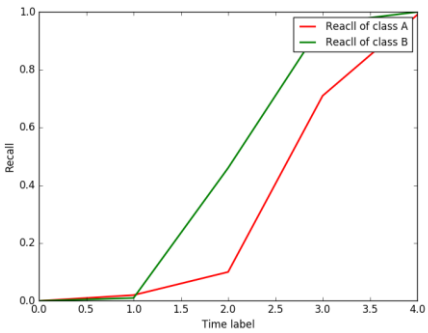


**Figure 4.** The recall of the data with different time label

We can see that the new data has higher test recall.The old data has low recall.We can use the threshold T=0.5,if the data set's average recall less than T,the data set with the time label can be remove.

When we use this methods,we can get a new result.

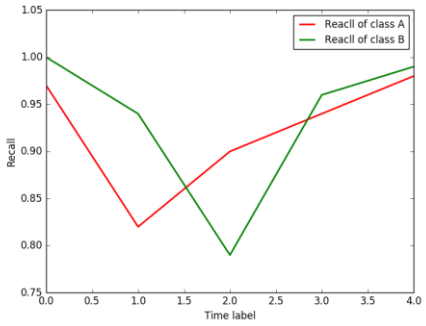


**Figure 5.**The recall of the predict result based on the data flow model

The recall of the prediction result is higher than the traditional method that we add all the data to the data set.

## 5 Conclusions

In this paper,we propose a data flow model that use the new data set to update the old data as the time goes by.In this way,we can add the new data to the history data set and remove the garbage data.At the same time ,the best machine learning model is changing as the update of the data set,we use the hybrid vote algorithm to get a better machine learning model. According to the experiment result,in the case of the distribution of the data set changes with time,we use the data flow model can get a better result than the traditional method that add all the data to the data set.

## Acknowledgment

In this paper, the research was sponsored by the Nature Science Foundation of china (project number 61402511, project number 61572514)

## References

1. L. Cao,T. Joachims, C. Wang, IEEE Intelligent Systems, **29**, 62-80 (2014)
2. V. Craykar,A. Saha, European Conference on Principles of Data Mining and Knowledge Discovery, (2015)
3. S. Valsan ,Backward Sequential Feature Elimination And Joining Algorithms In Machine Learning,(2014)
4. S.Chen,J. Lmoore,D. Turnbull,Knowledge Discovery and Data Mining,(2012)
5. M. David,Journal of Machine Learning Technologies,**2**, 37-63(2011).
6. Z.Abawab, G.Hmills,J.Crespo,,Knowledge Discovery and Data Mining, (2012)
7. B. Springer,S. Martin,Principal Component Analysis,**87,**41-64(2010)