Update of Kernel Density Estimation Model Based on Human Forgetting Mechanism

Yiwen Xu, Jing Chen, Zhongwen Zhuang, Liqun Lin

Fuzhou University, College of Physics and Information Engineering, China, Fuzhou

t07063@fzu.edu.cn

Keywords: Kernel Density Estimation (KDE); Probability Density Function (PDF); human forgetting mechanism; Internet of Vehicles (IoV)

Abstract: Kernel Density Estimation (KDE) is an effective method to estimate the probability density of random data, and is often used in the data processing in applications. In the long-term scenario, it is necessary to update KDE model to ensure its accuracy. However, the most traditional update method in KDE, random resampling, has the weakness of ignoring the contributions of the KDE models used before. Thus, it has negative influence on its accuracy and computation complexity. To solve these problems, we present a new update method based on human forgetting mechanism in this paper. Furthermore, we conduct comparative experiments based on the scenario of Internet of Vehicles (IoV) to verify the performance of the proposed method. The experimental results show that our proposed method is superior to the random resampling.

1. Introduction

Kernel Density Estimation (KDE) is a traditional method on estimating the Probability Density Function (PDF) of random data. It is a non-parametric method to estimate directly from the data without prior knowledge. As a result, it is widely used in various occasions.

Liao *et al.* [1] used non-parametric KDE to calculate the probability density error of wind power prediction, which made up for the shortcomings of the traditional prediction methods based on normal distribution. Cao *et al.* [2] proposed and tested the fast background subtraction method based on KDE, which was the basic step of identifying objects from video sequences in vision system. Yılan *et al.* [3] used KDE to estimate traffic density by assuming that velocity information for a given region is known. Elgammal *et al.* [4] adopted non-parametric KDE techniques to model the statistical representation of the background and foreground of the videos in surveillance system.

KDE is applied in many different situations. In the long-term scenario, it needs to update KDE for the reason that the PDF of the data may change with the continually update of the data. It is well known that random resampling is the most traditional update method in KDE [5]. In this method, a new round of random sampling was performed on the latest data set, and the sampled data was solved to obtain the latest KDE model. Although this method was simple to be implemented, it has an apparent weakness of increasing computation complexity with the increase number of random data. Moreover, it ignored the contribution of historical data sets during the update process, which left the gap of improving the performance of the update method of KDE.

To solve the above problems, we present a new update method on KDE model using the human forgetting mechanism to make full use of the historical KDE model in the update process. We also verify the performance of our proposed method via some experiments in the application of crowdsensing-based Internet of Vehicles (IoV).

The rest of the paper is described as follows. In section II, we introduce the traditional update method of KDE and its weakness. In section III, we describe in detail the specific steps of the proposed method of the KDE model. In section IV, we examine the performance of the update method proposed in this paper. Section V concludes the paper.

  DOI: 10.25236/icemit.2019.006

2. Traditional Update Method of KDE and Its Weakness

In practice, the data are continuously updated as the scene changes. If the KDE model remains unchanged, the accuracy of the KDE model may decline. Therefore, it is necessary to update KDE model to ensure its accuracy.

Random resampling, as the most conventional approach of KDE, is simple and easy to implement. However, it also has some disadvantages as follows:

a) In the long-term application process, as the amount of data continues to increase, we need to collect a large number of data samples to ensure the accuracy of the estimated PDF. Therefore, it leads to an increase in the computational complexity of KDE, and is difficult to achieve the real-time requirements.

b) The main purpose of the update of KDE model was to adjust the model based on the data collected during different time periods. Since data samples from different time periods had different contributions to solving the latest model. By considering the temporality of the data samples, the latest model could be estimated more accurately. However, the traditional random resampling method treats all samples as non-discriminatory data without considering their temporality, which negatively effects the accuracy of the update result.

In order to overcome the above disadvantages of the traditional method, we propose a new update method of KDE model based on human forgetting mechanism.

3. KDE Update Method Based on Human Forgetting Mechanism

The KDE update method is mainly composed of two steps. One is to judge whether the KDE model needs to be updated (that is, update condition of KDE), and the other is how to update the KDE model (that is update procedure of KDE). The implementation of the above two steps is explained separately below.

3.1 Update Condition of KDE

In statistic theory, the KDE model of dataset $\boldsymbol{X}$ is shown as follows:

$$\widehat{f}_{\mathrm{n}}(x) = \frac{1}{\mathrm{nh}} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \tag{1}$$

where n is the number of sample data from $\boldsymbol{X}$, and h is the window width. K (•) is the kernel function, which is described as:

$$K(u) = \frac{3(1 - u^2)}{4}, \quad |u| \leq 1 \tag{2}$$

As the detail calculation process of the KDE model had been demonstrated in our previous work [6], we would not repeat it here for space limitation.

The main purpose of updating the KDE model is to adjust the model in time to adapt to the new dataset. Hence, it is possible to judge whether the model needs to be updated by the difference between the current KDE model and the previous model. Divergence is a probable way to measure the difference between two KDE models [7]. Jansen-Shannon divergence (JS-D) has the characteristics of symmetry and non-negative constant with the range of [0,1]. It is convenient to set a threshold to determine the difference scale between the two probability distributions. Therefore, we use JS-D to measure the difference between two KDE models. The general expression of JS-D can be described as:

$$D_{\mathrm{JS}}(p_i, q_i) = \sum_{x \in X}\left[p_i \ln\left(\frac{2p_i}{p_i + q_i}\right) + q_i \ln\left(\frac{2q_i}{p_i + q_i}\right)\right] \tag{3}$$

where p and q are the two KDE models. The closer the value is to 0, the smaller the difference between the two models.

3.2 Update Procedure of KDE

Currently, random resampling is the most common update method of KDE [5]. It performed a new round of random sampling on the latest dataset to obtain new data samples which were used to update the KDE model. Although this method was simple and easy to be implemented, it had some disadvantages such as ignoring the contributions of the KDE models used before and high computation cost. To solve these problems, this paper uses the human forgetting mechanism to implement a new KDE model update method.

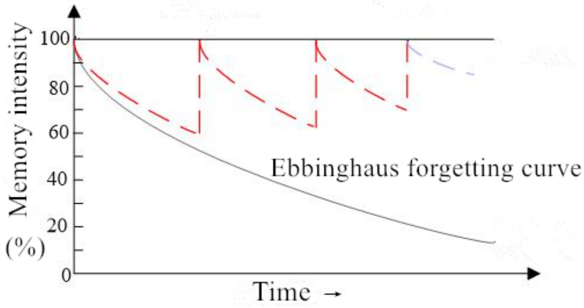

Fig. 1. Ebbinghaus forgetting curve.

1) Ebbinghaus forgetting curve: Ebbinghaus is the first scientist to study human memory, and his research can be summarized as the Ebbinghaus forgetting curve [8] as shown in Fig. 1.

From Fig. 1, we can draw the following conclusions:

a) Memory intensity decays with time. As time goes by, if the acquired knowledge is not recalled, its memory intensity will gradually decay until it disappears.

b) Recalling has a recovery effect on memory intensity. After acquiring knowledge, recalling knowledge in a short time can restore the memory intensity to 100%. Furthermore, after many recalling, the memory intensity tends to be stable.

c) There is a correlation between the recalling know-ledge and the dealing events. When dealing with events by recalling knowledge, only the knowledge related to the event will be recalled.

In practical applications, the calculation and update of the KDE model are always done in the central server. In the central server, the acquired knowledge is equivalent to the KDE model. The central server achieves the purpose of KDE update by recalling the gained knowledge. When the KDE model needs to be updated, the KDE model associated with the current update is recalled and consequently the latest kernel density model is calculated.

2) Update of KDE: Based on the analyses in the previous section, the KDE model update method can be divided into three stages: knowledge learning, recalling and forgetting.

a) Knowledge learning: We set up a knowledge base in the central server. When the latest KDE model is uploaded to the central server, each model of the central server performs parameter initialization and stores the model in the knowledge base. In the parameter initialization step, there are two parameters in each KDE model of the knowledge base need to be initialed. One is the memory strength w whose initial value is 1, and the other is the forgetting factor v whose initial value is the difference between the current KDE model and its previous model.

b) Knowledge recalling: When the value of JS-D is greater than its threshold, which means the KDE model changes significantly, we start updating the KDE model. At first, we calculate the differences between the previous KDE model and all the models in the knowledge base, which are denoted by JS-D as shown in (3). Then the top 50% models with the smallest difference are selected as the useful knowledge. Finally, their average result is taken as the latest KDE model.

c) Knowledge forgetting: After selecting and using the model in the knowledge base, the parameters of the selected model are updated. The update operation involves updating each of the selected KDE models in the knowledge base with the following three parameters.

The first parameter is the number of times selected by the model, and it is directly added 1 when updating.

The second parameter is the memory strength w, which is updated as

$$w = e^{-v(t'-\tau)} \quad (4)$$

where τ represents the time when the model is last selected, and t' denotes the current time.

The third parameter is the forgetting factor v which is updated as

$$v = \frac{\beta}{\lambda + 1} \quad (5)$$

where β represents the initial value of the forgetting factor which can be described by the JS-D value of the KDE model and its previous model. The forgetting factor v indicates the memory stability of the model and is a non-negative number. The smaller the value, the more stable the memory of the model is.

When the number of models exceeds the capacity of the knowledge base in central server, the model with the lowest memory intensity is removed from the knowledge base, which is the forgetting process of knowledge.

In the update method based on human forgetting mechanism, the knowledge most relevant to the current KDE model will participate in the update procedure. The memory intensity w is used as an evaluation criterion for the usefulness of the KDE model, which fully takes into the historical contribution of the KDE model account. Thereby, the most valuable model in history is stably stored in the base, improving the stability of the update.

4. Experiments and Result Analysis

In order to verify the performance of our proposed method, we compare it with conventional method, random resampling [5], via experiments by taking the IoV application as an example.

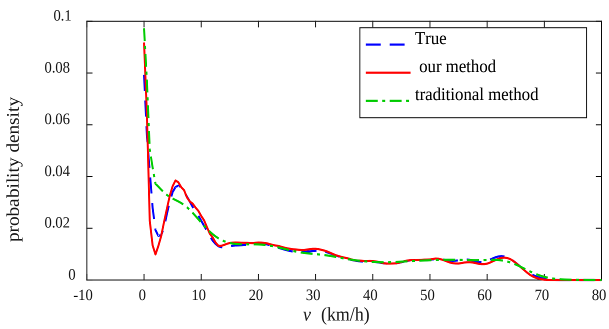

Fig. 2. Results of KDE obtained by our proposal method and the traditional method.

In our previous research, as shown in [8], we designed a traffic data collection software to simulate the data acquisition process in crowdsensing-based IoV system. In this paper, we also utilize this software to obtain enough data to be the data source of the following experiments.

The main parameters in the experiments are set as follows:

a) General parameters: The data volume in central server is initiated as 300000, the detection duration is 30min, and the JS-D threshold is 0.3.

b) Forgetting mechanism parameter: The knowledge base capacity is set as 1200.

c) Resampling parameter: The ratio of data in each re-sampling procedure is set as 10%.

Fig. 2 shows the KDE results based on the above settings. The red solid line in the figure represents the latest kernel density model obtained by our proposed method. The green dotted line represents the model obtained by the traditional random resampling method, and the blue dotted line is the true probability density curve calculated from the data source. It can be seen from the figure that the probability density curve obtained by our proposed method is apparently closer to the true probability density curve. Therefore, we can conclude that the KDE model update algorithm based on human forgetting mechanism proposed can achieve better performance than the traditional random resampling method.

5. Conclusions

In this paper, we propose a KDE model updating method based on human forgetting mechanism, which makes up for the limitation of the traditional random resampling method. It not only reduces the computational complexity of the update procedure, but also fully considers the correlation and the contribution of the data collected at different acquisition time. Comparison experimental results show that the proposed method achieves better performance than the traditional method. It is worth mentioning that, although our experiments take IoV system as an example, the proposed update model can be used in other applications, due to the fact that the derivation process of our method does not refer to a special application.

Acknowledgement

This research is supported by the National Natural Science Foundation of China (Grant 61671152).

References

[1] G. Liao, J. Ming, B. Wei, H. Xiang, *et al.*, “Wind power prediction errors model and algorithm based on non-parametric kernel density estimation,” 2015 5th International Conference on Electric Utility Deregulation and Restructuring and Power Technologies (DRPT), 2015, pp. 1864-1868.

[2] J. Cao, O.C. Victor, O.M. Gilbert, and C. Wang, “A fast background subtraction method using kernel density estimation for people counting,” 2017 9th International Conference on Modelling, Identification and Control (ICMIC), 2017, pp. 133-138.

[3] M. Yılan, and M.K. Özdemir, “A simple approach to traffic density estimation by using Kernel Density Estimation,” 2015 23nd Signal Processing and Communications Applications Conference (SIU), 2015, pp. 1865-1868.

[4] A. Elgammal, R. Duraiswami, D. Harwood, and L.S. Davis, “Background and foreground modeling using nonparametric kernel density estimation for visual surveillance,” Proceedings of the IEEE, vol. 90, no. 7, pp. 1151-1163, 2002.

[5] Y. Yang, Z. Sun, and J. Zhang, “Finding Outliers in distributed data streams based on kernel density estimation,” Journal of Computer Research and Development, vol. 42, no. 9, pp. 1498-1504, 2005.

[6] Y. Xu, N. Xu, Z. Zhuang, *et al.*, “An Abnormal Data Detection Algorithm for Internet of Vehicles Based on Crowdsensing,” Journal of Hunan University: Natural Sciences, vol. 44, no. 8, pp. 145-151, 2017.

[7] A. Giantomassi, F. Ferracuti, S. Iarlori, G. Ippoliti, and S. Longhi, “Electric Motor Fault Detection and Diagnosis by Kernel Density Estimation and Kullback–Leibler Divergence Based on Stator Current Measurements,” IEEE Transactions on Industrial Electronics, 2015, pp. 1770-1780.

[8] Q. Zhao, Y. Jiang, and Y. Lu, “Ensemble Model and Algorithm with Recalling and Forgetting Mechanisms for Data Stream Mining,” Journal of Software, vol. 26, no. 10, pp. 2567-2580, 2015.