THINK INDIA JOURNAL
ISSN: 0971-1260
Vol-22-NO-30 Special Issue- December-2019

Digital Data Forgetting: A Machine Learning Approach

Mir Mohammad Yousuf[1], Shravan Kumar[2],Mamoon Rashid[3],Bisma Majid[4]
[1,3,]Assistant Professor, Department of Computer Science and EngineeringLovely Professional University
[2]Research Scholar, School of Computer Science and Engineering
Lovely Professional University
[4]Research Scholar, Department of electronics and communication
N.I.T Srinagar

Abstract

From the last 20 years digital transformation throughout the world goes on increasing very fast. Nowadays very important and powerful is data. In earlier days, to collect the data magnetic tapes were used after that digital data storage was used. But now the extremely popular method is Big Data with its tool ML in each industry and literature. In order to obtain a powerful/meaningful/valuable data, people use machine learning. The results which come are very valuable for planning. However, these days it's quite difficultto collect the digital data and utilize it in computers. With this expensive process, in the future,there would be enough storage space to store the data. Therefore, we want and propose the phrase "Digital knowledge forgetting" with the ML approach. Because of this software or the digital solution, we can erase the data which is not required and get some more valuable data. We referred to this process as "Big Cleaning". In this article, the set of datathat we will use is to get and extract the additional valuable and meaningful information with Deep Autoencoder, k-nearest neighbor (k-NN), and Principal Component Analysis (PCA) machine learning algorithms for the experimental and analysis purpose.

***Keywords - D*igital Data Forgetting, ML, Big Cleaning, Deep Autoencoder, k-nearest neighbor (k-NN), and PCA**

I. INTRODUCTION

In today's digital world, the development of knowledge is always represented by data. The essential steps for processing thedigital world are Data storing, Data sharing, Data producing, and Data processing[1][2]. Currently, the researchers primarily focus, particularly on processing. In earlier days, hard copies were enough to store the limited data, but, indeed, it was difficult to share that data. However, from the last century, the increasing level of the digital transformation of the data producing has been accelerated[4][8]. Today's nearly everybody is manufacturing the info in some ways via e-commerce, a social platform like Facebook, Instagram, and YouTube or calculating the values like the weather forecast, the pressure level of gas and therefore theothers[5]. These days, the data which is produced we always try to store that information. Though we've got tremendous storage capability, it has a boundary and in the future, we will reach that boundary [9]. In that situation, the data which

ISSN: 0971-1260
Vol-22-NO-30 Special Issue- December-2019

is not required we need to forget / delete / erase that data. Therefore, in digital life, this will become a new area for research.

Nowadays, the volume of information is so high, as a result, it is referred to as large statistics. Big records contain the records on absolutely exceptional significance stages. Nowadays, the strategies which can be especially targeted are device learning, deep mastering, records mining, and artificial intelligence for finding the treasured records (functions) in it[2][4]. However, there are not many papers that don't specify the situation inside literature about digital records forgetting [1].

II. BACKGROUND WITH RESEARCH QUESTIONS

Despite the actual fact that the usage of records with minimum cost and the increasing of prices for storing the consequences of the fact in a lack of it slow inside the system studying models, there's no literature using algorithms for records cleansing. However, like this difficulty, the community has created solutions for statistics troubles which are just like this type [7][8].

For the selection of important real-life data, issues are typically done by the specialists / experts. Let take an example of a library, which of the books should be kept on the shelves and which one should be archived [6][10]. For the given problem, by making the use of the librarian past experience who have expert knowledge in this field makes some criteria for determining it. Once it is decided, the shelf arrangement is done.

Since this type of study wasn't enclosed within the literature, we tend to propose our methodology supported real-life issues like within the library example. So as to search out a replacement approach to search out solutions for the issues that have an effect on the success of the ML model which arises from the gathered dataset, with this we tend to present our methodology by producing some answers to some set of queries[3][10]. The related queries for the research are:

i. Is there any datathat is inconsistent with the dataset even after classifying the data set? Is there any over fitting in the structure of the given dataset?
ii. Which part of datasets is useful for gathering the most valuable attributes?

We tend to finish with a way that solves the above-related queries for the research, whichcan increase the success of ML models & additionally keep the researchers or the corporations from expensive resources for the data storage.

III. DIGITAL FORGETTING AND MACHINE LEARNING

In this paper, we tend to use 3 techniques that are extremely popular in the machine learning field. k-Nearest Neighbors (k-NN) [2][9], Principal Component Analysis (PCA) [8], and Deep Autoencoders (future generation of Neural Networks)[1][3][4]. Before explaining why we decided these 3 techniques, we should understand some basics of these techniques which is important to know the logic concerning however we are able to forget the digital data by using these 3 techniques.

ISSN: 0971-1260
Vol-22-NO-30 Special Issue- December-2019

In Machine Learning, one of the most popular methods is the k-Nearest Neighbors (k-NN) algorithm. This method is often used for the analysis of classification and regression. In the k-NN model, the class is already known [6].This model can be constructed by using the data's location from the given testing data set. In order to measure the distance_between the data and the center of the classified data, we make use of the euclidean distance formula. According to the nearest data class, we identify the class for new data.[1][7][10]

Another method that is commonly used in Machine Learning is Principal Component Analysis (PCA)[5]. PCA is a way of identifying styles in facts & expressing the statistics in such a manner to focus on their similarities and variations. Especially, if we have a tendency to work on the big data, probably all needed is to reduce the existing dimensions which are set for the data [9]. Nowadays, the availability of enough memory for processing and storingbig data is more complex if we don't have powerful processors with very high capacity memory. Using the features of the PCA algorithm, lets us identify the important patterns which represent the data. PCA algorithm is based on the method of the Eigenvector. Eigenvectors are the vectors that are found in the datamatrix [7]. The most important eigenvector would have the direction in which the variables are strongly co-related i.e., its direction doesn't modification even when the transformation of the data takes place. The data of the PCA projects on the Eigenvectors. Because even after the transformation of the data there is a minimum loss of data, and the direction of the Eigenvectors will never change. Moreover, whenever the original data is needed, we can bring it back by using the Eigenvectors. Let us take an example by assuming the dataset with N data and M features (NxM). With this, the dimension of the dataset can be reduced by selecting the primary P components which represent the best data with the condition: P <M[8].

Deep Autoencoder another method that is one of thetypes of neural networks. It is a method that is usually used for reducing the dimension of the data set. With this algorithm, we can easily detect the data which is irrelevant in the data set. This technique is a part of Unsupervised Learning which has 2 components viz., encoder&decoder. The data is used as an input in the encoder to produce a code / model for encoding or transforming. The data which is transformed is also known as more qualified dataset than the input data

Therefore, for reducing the dimension we commonly used the algorithm like PCA and Deep Autoencoders and whereas KNN is used for the classification. In this study, our main purpose is to find the non-relevant data for forgetting or deleting them. Therefore, our main purpose is to delete irrelevant using PCA and Deep Auto Encoders[5][7]. Moreover, using the k-NN algorithm we can also delete the data itself from the data set which is difficult for classification by comparing the Euclidean distances with data or analyzing with the thresholds.

IV. PROPOSED APPROACH

In this study, as mentioned by the author they use the Iris data set for testing the (PCA), k-Nearest Neighbor (k-NN), and Deep Autoencoder algorithms. Three different types of iris have been used in the experiment (Setosa, Versicolour, and Virginia) where the iris dataset

ISSN: 0971-1260
Vol-22-NO-30 Special Issue- December-2019

which is used contains 50 samples from each type and every sample has further classified into four more features [8].For erasing or deleting the data from the data storage, we will try to extract the features which are at least. Because in future storage it'll be not possible to save all the data into digital storage [10]. Therefore, the irrelevant data need to be deleted and the data with only the most important features need to be saved. The three different techniques of experimental analysis are discussed below.

A. Application of PCA on Digital Data Forgetting:

To find the features which are most important from the given dataset, we generally use the PCA algorithm. Here, we have a tendency to manipulate the PCA technique to seek out the least necessaryfeatures/records. We can erase n variety of the data which are least necessary [7]. The main goal during this approach is to seek out the least important data & erase them [5].

The steps for this algorithm is given below in the summarized form:

i. Finding the mean of the given data.
ii. Scaling the data with the help of the calculated mean.
iii. Calculate the covariance matrix of the data.
iv. Find the diagonals of the covariance matrix & also find the variances of the data.
v. Sort the variances of the data in descending order
vi. Finally, erase the data which comes first with minimum variance.

B. Use of KNN on Digital Data Forgetting:

In this section, we're the use k-NN method to hunt down the class of every single statistic. Therefore, if the statistics lie in the middle ± threshold of the cluster, then we can erase it [5]. During this approach, every class will try to constitute their personal elegance once we erase some of the information of that magnificence[9]. The Steps for this algorithm and defined variables are given below:

D_f: *Sample Vector' distance to own cluster's centroid*
D_m: *Mean of* D_f
ξ: *Threshold Value*
D_i: *Sample's distance to own class centroid*
Pseudo code for this approach is:

```
for i=1 to n:
        if | Di-Dm | < ε
                erase the record
        end
end
```

C. Use of Deep Autoencoder on Digital Data Forgetting::

ISSN: 0971-1260
Vol-22-NO-30 Special Issue- December-2019

In this approach, we have a tendency to use Deep auto encoders for thecompression of the data and that we need to seek out the farthest records / features for the compression of the data [5][6][7]. While identifying the farthest records / features we have a tendency to erase them.

The Steps for this algorithm is given below:

i. Normalizing the data.
ii. Design a Deep Autoencoder, there should be only 1 node in the last layer of the encoder.
iii. Train Deep Autoencoder and predict with the layer of an encoder. When this step is completed, then we get the compressed data.
iv. Identify the distance between the normalized data and the compressed data.
v. Sort the distances which are found.
vi. The feature / record with maximum distance hasthe least important feature / record.
vii. Finally, erase the feature / record which is least important.

Therefore, with the use of machine learning techniques, we can manipulate the data. After identifying the most valuable features, we can save these data and erase the remaining data.

We have a propensity to apply iris statistics set and okay-NN technique for the type after the method of our digital information forgetting. Firstly, we must now not observe the method of digital information forgetting. Initially, we should train our version with a k-NN set of rules with the use of the iris dataset and observe the accuracy of ok-NN prediction. Then after this step, we follow the technique of virtual statistics forgetting technique with the Deep Auto Encoder approach [9][10]. In this method, the Deep Auto Encoder approach has a hundred and fifty nodes in the enter layer, one node in the center layer (Encoder's ultimate layer) and one hundred and fifty nodes inside the outputlayer. When 10% of the information has forgotten we have a propensity to identify the accuracy of the growing prediction.

Table 1. Accuracy after applying digital data forgetting

Accuracy on k-NN Algorithm				
Without Digital Forgetting	***5% Forgetting***	***10% Forgetting***	***15% Forgetting***	***20% Forgetting***
95.33%	95.07%	94.81%	95.27%	95.83%

V. CONCLUSION

In recent years, the quantity of data being created and saving globally is been increasing rapidly. This brings several issues like the price and the space for the storage of data. During

ISSN: 0971-1260
Vol-22-NO-30 Special Issue- December-2019

this study, we have a tendency to propose a machine learning technique with valuable information with the method of digital data forgetting.

With the help of 3 types of every iris data set: iris setosa, iris virginica and iris versicolor. Firstly, we should not apply the process of digital data forgetting. Initially we should train our model with the k-NN algorithm. Then, we apply the process of the Deep Auto Encoder technique. In this process, the Deep Auto Encoder technique has 150 nodes within the input layer, 1 node within the middle layer (i.e. Encoder's last layer) and 150 nodes within the output layer. We erasing the data with 5%, 10%, 15%, 20% severally. When doing with the forgetting processes with each data, we need to classify the other data with k-NN. At last, we compare the results of data before the digital data forgetting with 5%, 10%, 15%, and 20% after the digital data forgetting process. We observed that the accuracy has increased after performing the 10% of digital data forgetting.

During this study, we presented a method of digital data forgetting for the big data which is referred to as 'big cleaning'. After this process, we can store more data in our digital data storage or in any other software / hardware device. Moreover, we ignore the operations which prices high for digital storage operations. Therefore, this can be an important method with many advantages by using the process of digital data forgetting.

REFERENCES

[1] M. Günay, "DIGITAL DATA FORGETTING : A Machine Learning Approach," pp. 0–3, 2018.

[2] L. M. Zouhal and T. Denoeux, "An evidence-theoretic k-NN rule with parameter optimization," *IEEE Trans. Syst. Man Cybern. Part C Appl. Rev.*, 1998.

[3] Y. Lecun, Y. Bengio, G. Hinton, M. I. Jordan, U. C. Berkeley, and G. Hinton, "Deep Learning Authors ' Relationships," vol. 444, no. May, pp. 436–444, 2015.

[4] 2016. Goodfellow I, Bengio Y, Courville A. Deep learning[M]. MIT press, "Deep Learning - MIT," *Nature*. 2016.

[5] A. J. Holden *et al.*, "Reducing the Dimensionality of," vol. 313, no. July, pp. 504–507, 2006.

[6] C. Hong, J. Yu, J. Wan, D. Tao, and M. Wang, "Multimodal Deep Autoencoder for Human Pose Recovery," vol. 24, no. 12, pp. 5659–5670, 2015.

[7] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2013.

[8] I. Jolliffe, "Principal component analysis. In: International encyclopedia of statistical science", Springer, Berlin, Heidelberg, 2011. p. 1094-1096.

[9] E. Alpaydın, "Introduction to machine learning (adaptive computation and machine learning)", Mass: MIT Press, Cambridge, 2004.

ISSN: 0971-1260
Vol-22-NO-30 Special Issue- December-2019

[10] M. Lantz, "Why the Future of Data Storage Is (Still) Magnetic Tape," IEEE Spectrum: Technology