International Journal of Research in Engineering, Science and Management
Volume-3, Issue-6, June-2020
www.ijresm.com | ISSN (Online): 2581-5792

Digital Data Forgetting using a Machine Learning Approach

Chaithra[1], S. Pavithra[2*], Nitin Mundhara[3]
[1]*Associate Professor, Department of Computer Science and Engineering, Sapthagiri College of Engineering, Bangalore, India*
[2,3]*Student, Dept. of Computer Science and Engineering, Sapthagiri College of Engineering, Bangalore, India*
*Corresponding author: pavithrapavi2497@gmail.com

***Abstract*: Regularly evolving technology has become an important part of our lives. In the last 2 decades, the Digital transformation of the world has increased rapidly. Our life is dependent on Data. Previously, people used magnetic tapes to store data. With evolving technology, people started using storage devices such as hard disk, to store data. In the recent years, big data and its machine learning tools have become prominent in both literature and industry. Machine Learning is used in order to obtain meaningful information from big data. However, it is impossible to collect and store all the data that is produced. This process has its own limitations. It is expensive and time consuming and moreover, we do not have unlimited storage space. In order to tackle this problem, we propose "Digital Data Forgetting" phrase. Using this solution, we can extract more valuable data and hence we will be able to erase the rest of them. We call this operation as "Big Cleaning". We use a knowledge set and employ principal component analysis (PCA), deep auto encoder and k-nearest neighbor machine learning methods in order to extract valuable data.**

***Keywords*: Machine Learning, Deep Auto encoder, Big cleaning, Big Data, K-Nearest Neighbor, PCA.**

1. Introduction

Data can be defined as the quantitative or qualitative values of a variable. Data is thought to be the lowest unit of information from which other measurements and analysis can be done. Data can be numbers, images, words, figures, facts or ideas.

Data has always represented the development of knowledge in today's rapidly growing digital world. In order to pass to next level of digital life, Data storing, sharing, producing, processing are the essential steps. At one time, data was limited. During this period papers were the best means to store data. Data production has extensively accelerated with the increasing level of digital transformation especially in the last century. Today's almost everyone is producing the data in many ways via electronic trade, web user including browsing, devices like weather forecast, gas pressure and others. Every second, more and more data is being created and needs to be stored and analyzed in order to extract value. Furthermore, data can be stored easily, and the process has become cheap.

As a result, organizations need to extract as much value as possible from this huge amount of stored data. There has been a tremendous surge in the use of digital storage such as Hard disks, DVDs, flash memory etc., as well as a drop in its price within the last 20 years. This has eliminated the requirement of clearing out previous data. Moreover, it has increased the storage of metadata, i.e. data about the data, as well as made backup storage a very common practice to avoid data loss. Additionally, in this growing world of technologies, companies and individuals possess more and more new technologies and devices which create and capture more data in different categories.

A single user today can own a desktop, laptop, smartphone, tablet, and more, where each device carries very large amounts of data related to that individual. Nowadays, people store all data that they produce every day in order to use it in a later point of time. Although we have tremendous storage capacity, it has bound, and we will reach that bound very soon. In that case, we will have to forget / delete / erase data that we do not need to reduce the size of data. Therefore, this will soon be a new research area in digital life.

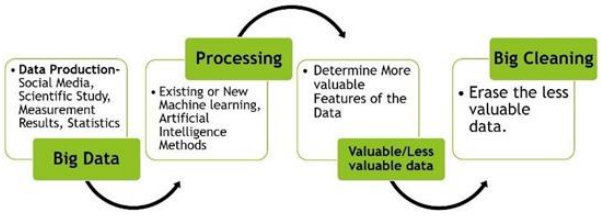

Fig. 1. Data Forgetting Process

Today, data generated is very high, and hence termed as big data. It contains data on different importance levels. Data in itself cannot be understood and to get information from the data one must interpret. There is a need to analyze such sheer amounts of data properly. It is necessary that the pertaining useful information is extracted from it. Nowadays, different techniques such as artificial intelligence, machine learning, data mining, deep learning methods focus on finding most important information from big data. In this study, we intend to open a new research area and we call it 'Digital Data Forgetting'.

Table 1
Literature Survey

S. no.	Author name	Research Paper Title	Abstract or Conclusion	Advantages	Disadvantages
1.	Y. Freeund, R. Sneddon 2015	A Decision-Theoretic Generalization of on-Line Learning and an Application to Boosting	It shows how the resulting learning algorithm can be applied to a variety of problems, including gambling, multiple- outcome prediction, repeated games, and prediction of points in Rn	It will be used to predicting the misbehaving in online games.	Decision Tree complexity is huge, and accuracy is less.
2.	I. Goodfellow, Y. Bengio, A. Courville 2016	Deep learning	The book is aimed at an academic research audience with prior knowledge of calculus, linear algebra, probability, and some programming capabilities. A non-mathematical reader will find this book difficult.	It will useful for to learn about deep learning	There is only theoretical analysis on various methods
3.	G. Hinton, R. R. Salakhutdinov	Reducing the dimensionality of data with neural networks	High-dimensional data can be converted to low-dimensional codes by training a multilayer neural network with a small central layer to reconstruct high- dimensional input vectors. Gradient descent can be used for fine-tuning the weights in such ‘‘autoencoder’’ networks.	very effective for nonlinear dimensionality reduction, provided that computers were fast enough, data sets were big enough to a good solution	works well only if the initial weights
4	C. Hong, J. Yu, J. Wan, D. Tao, M. Wang	Multimodal deep autoencoder for human pose recovery	we propose a novel pose recovery method using non-linear mapping with multi-layered deep neural network. It is based on feature extraction with multimodal fusion and back-propagation deep learning.	It avoids the over- simplified assumption of linear mapping	It is only used for image data
5	X. Lu, Y. Tsao, S. Matsuda, C. Hori	Speech enhancement based on deep denoising autoencoder	In this study, we further introduce an explicit denoising process in learning the DAE.	It is used to noise reduction and speech enhancement	Complex implementations
6	M. Lantz	Why the Future of Data Storage Is (Still) Magnetic Tape	It’s true that tape doesn’t offer the fast access speeds of hard disks or semiconductor memories. Still, the medium’s advantages are many. To begin with, tape storage is more energy efficient: Once all the data has been recorded, a tape cartridge simply sits quietly in a slot in a robotic library and doesn’t consume any power at all.	It is extendable, Reliable	Vulnerable to attacks.
7	Chen Junli; Jiao Licheng	Classification mechanism of support vector machines	This paper is to provide an introductory tutorial on the basic ideas behind support vector machines (SVM). The paper starts with an overview of structural risk minimization (SRM) principle and describes the mechanism of how to construct SVM.	SVMs are attractive approach to data modeling	It supports only for binary classification. Accuracy is less
8	T. Denoeux	A k-nearest neighbor classification rule based on Dempster-Shafer theory	In this paper, the problem of classifying an unseen pattern on the basis of its nearest neighbors in a recorded data set is addressed from the point of view of Dempster Shafer theory.	It provides a natural way of modulating Easy to implement	Accuracy is less.

2. Related Works and Research Questions

The Data is increasing rapidly today. Though, there has been a decrease in the cost of storing the data, this has its own limits of maximum data capacity. There is a need to reduce the storage space consumed by the data. Till now, there is not any data cleaning approach in the literature using machine learning. However, similar to this issue, the community has produced solutions for similar to these kind of data problems.

It is very difficult to choose valuable data in real-life and is often done analogously by experts. For example, consider the example of Library. It is necessary to make decisions regarding which books needs to be archived, which ones should be on the shelves. In this problem, there is a need of a librarian who is an expert in this field. The librarian must define some criteria by making use of his/her past knowledge to sort out the book. Moreover, he/she must also be able to choose between these criteria that best fits the situation. After these choices, the shelf arrangement is determined.

Since this kind of study is necessary in today’s growing digital world, we propose our method based on real-life problems such as in the library example. Moreover, such studies have not been included in the literature. In order to find a new approach to find solutions for the problems that affect the success of the machine learning model arising from the knowledge set, we present our method by producing answers to some questions. These research questions may include:

- Is there any inconsistent data in the dataset, even when the class of the record is specified? (If the machine learning problem is a classification problem).
- Does the data set have a structure that causes over fitting?
- Which parts of the data set give us more valuable attributes? (Extraction of Valuable Data)

For more details about previous works, refer to table 2.1.

We can conclude that a method that is capable of solving the above research questions and giving appropriate results will increase the success of machine learning models. This would save the researchers or companies from high storage costs.

3. Algorithm Used

To Extract the features PCA is used and to perform the classification we are using KNN classification algorithm.

Study done on the algorithms:

Principal Component Analysis (PCA) algorithm is generally used to find most important features of the data set. PCA is an unsupervised, non-parametric statistical technique that is primarily used for dimensionality reduction in the field of machine learning. Here, we manipulate PCA algorithm in order to find the least important features/records. We delete n number least important records. Our goal in this approach is to find the least important records and delete them. We can summarize the steps of PCA with these items:

- Find the mean of the data.
- Scale the data using mean of data
- Calculate the Covariance matrix of Data
- Get diagonals of covariance matrix and find variances of data
- Sort variances of data in descending order
- Finally delete records which is first n minimum variance

We use k-NN algorithm to find each record's class. Therefore, if the record lies within middle ± threshold of cluster, then we can delete that record. In this approach, each class still will represent their own classes when we delete some of them. The definitions and steps of this algorithm can be seen from here:

Rf: Vector of samples' distance to its own cluster's centroid

Rm: Mean of Rf ε: Threshold value

Ri : Sample's distance to its own class' centroid Pseudocode of this approach is:

```
for i = 1 to n :
    if | Ri-Rm | < ε
        delete record
    end
end
```

A Deep autoencoder is a type of artificial neural network. It is used to learn efficient data coding in an unsupervised manner. Deep autoencoder aims to learn a representation (encoding) for a data set. It is typically used for dimensionality reduction which is done by training the network to ignore certain features. In this approach, we use deep autoencoders for compression the data and we want to find farthest records/features to compressed data. After finding farthest records/features we delete them.

Steps of this algorithm can be seen from here:

- Normalize data
- Design deep autoencoder, last layer of encoding part must have 1 node
- Train deep autoencoder and predict with encoding layer. After this step we will have compressed data
- Find distance between compressed data and normalized data
- Sort found distances
- Then which record/feature has maximum distance, that is least important record/feature
- Finally delete least important records/features

Purpose of using this algorithm:

It is used increase the accuracy of identifying data that is not important from the dataset.

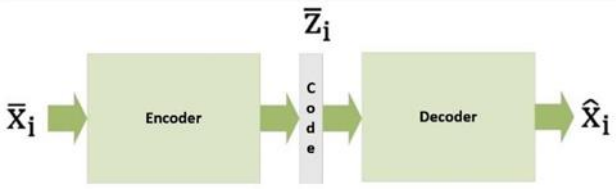

Fig. 2. Deep Autoencoder Architecture

4. Implementation

The Actor/User browses the CSV dataset or performs the search operation for the file using SearchCSV(). User then selects the dataset and the path of the CSV file is stored.

The user can then perform the Data Forgetting Process by using the 3 algorithms namely, KNN, PCA, and Deep autoencoder. So as to perform this, the user is given the option to run any of the 3 algorithms, KNN by calling KNNProcess() method, PCA by calling PCAProcess() method and Deep autoencoder by calling the AEProcess() methods. These algorithms produce the results in the form of graphs and the accuracy can be displayed. The same results are also stored on to the data store or Dard disk. Moreover, the Not selected data and the selected data can be stored separately in two different CSV files. Thus, giving the user an option to delete the data once he/she verifies it. Fig 3 shows the Sequence Diagram for the proposed application.

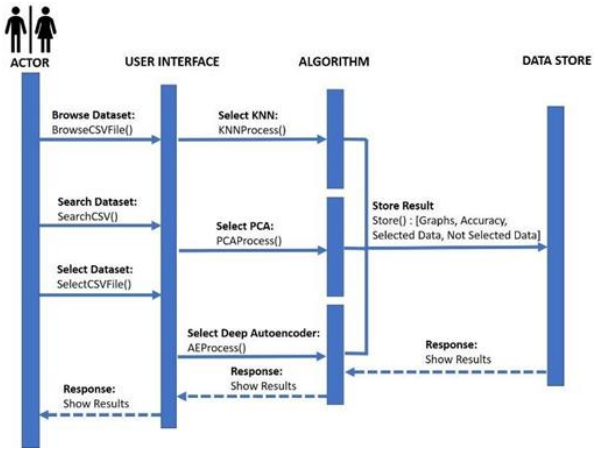

Fig. 3. Sequence Diagram

Similarly, fig. 4 shows an overview flowchart of the

proposed method. Firstly, the dataset is uploaded to the system. This can be done by either browsing for the file or entering the path of the file. Once the file has been uploaded, the various algorithms can be employed over the selected dataset in order to perform the "Big Cleaning" process. The user is given an option to run the algorithm based upon his choice. This would do a data forgetting operation on the dataset by extracting the essential data from the selected dataset.

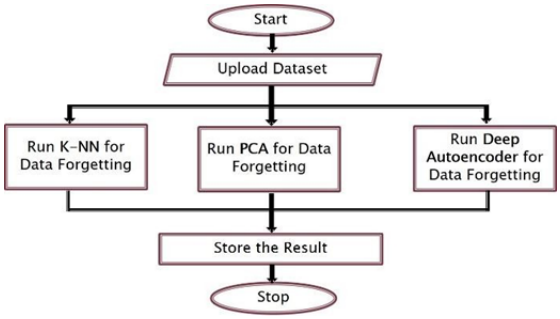

Fig. 4. Flow chart of digital data forgetting

Numerical analysis and their results can be seen from table 2.

Table 2
Accuracy after applying Digital Data Forgetting

Accuracy on KNN Algorithm				
Without Digital Forgetting	5% Forgetting	10% Forgetting	15% Forgetting	20% Forgetting
94.33%	94.07%	93.81%	94.27%	94.83%

5. Conclusion

Over the recent years, the amount of data being produced and stored at global level is tremendously increasing. This brings many problems such as storage space and storage cost since these are limited in their own sense. In this study, we propose a machine learning approach to extract more valuable data from a data set with digital data forgetting process. Thus, eliminating the less required/essential/valuable data.

In this study, we present digital data forgetting approach using machine learning for big data and term the process as 'Big Cleaning'. By employing this process, digital storage devices and other hardware material will be able to store more data occupying less space. Furthermore, we avoid high costs of the digital storage operations. Therefore, we will have many advantages with proposed digital forgetting approach.

References

[1] Y. Lecun, Y. Bengio, G. Hinton, "Deep learning", Nature, 2015, 521.7553: 436.
[2] I. Goodfellow, Y. Bengio, A. Courville, "Deep learning", Cambridge: MIT press, 2016.
[3] J. M. Zurada, "Introduction to artificial neural systems", St. Paul: West publishing company, 1992.
[4] G. Hinton, R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," Science, 2006, 313.5786: 504-507.
[5] C. Hong, J. Yu, J. Wan, D. Tao, M. Wang, "Multimodal deep autoencoder for human pose recovery," IEEE Transactions on Image Processing, 2015, 24.12: 5659-5670.
[6] X. Lu, Y. Tsao, S. Matsuda, C. Hori, "Speech enhancement based on deep denoising autoencoder," In: Interspeech. 2013. p. 436-440.
[7] T. Denoeux, "A k-nearest neighbor classification rule based on Dempster-Shafer theory," IEEE transactions on systems, man, and cybernetics, 1995, 25.5: 804-813.
[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. [10] M. Lantz, "Why the Future of Data Storage Is (Still) Magnetic Tape," IEEE Spectrum: Technology, Engineering, and Science News, IEEE Spectrum, Aug. 2018.
[10] Yoav Freund, Robert E. Schapire, A Decision- Theoretic Generalization of on-Line Learning and an Application to Boosting.
[11] Chen Junli, Jiao Licheng, "Classification Mechanism of Support Vector Machines", WCC 2000 - ICSP 2000. 2000 5th International Conference on Signal Processing Proceedings. 16th World Computer Congress, 2000.