IETC – 670 Database and Big Data

Final paper

Topic:

Big Data and Machine Learning

Student: Sheng Kai Liao

Fall 2021

Introduction

With the development of internet, people are able to exchange information, interact with each other and experience new thing online easily, and tons data is created regarding such behaviors. For example, according to research, today a person can generate at least 2.5 quintillion bytes of data online per day (Price, n.d.). To conclude such big, mess and complex data, Big Data has been defined and play a big role in several domains in industries and academic fields, especially for data analysis. Big Data also has highly impacted the way we live, we work, and we think. Also, many companies are trying to analysis Big Data in order to benefit themselves, such as catch the tendency of the market for making sale strategies or catch the trend of economy in order to evaluate their investment.

In addition, as the most effective data analysis method, with the help of Big Data, Machine Learning is able to accomplish some achievements that were unimaginable in the past. By combining the Big Data with Machine Learning, many innovational applications have been carried out in real life. For example, Tesla has been developing automated driving technology for many years and applying the technology to its own products, as well as traditional car manufacturing brands such as Ford, Toyota, and Nissan. On the other hand, as another popular application of Machine Learning and Big Data, facial recognition is also popular in these ten years which not only been applying in authentication system but also been used in surveillance system. Those innovation technologies are not only improving the convenience of human life but also make the world safer.

In short, in this report, I deliver a brief introduction about Big Data and Machine Learning, the applications of combining both and point out the advantages and challenges when Machine Learning collaborates with Big Data with some real-world examples.

What is Big Data

In short, Big Data means a great amount of data which a usual computer cannot handle for analysis. However, Big Data is very useful resource for companies, industries and researchers to gain a better understanding of market tendency or achieve particular purpose by analyzing the insight of the data. In order to process such big amount of data, the most popular way in industry is to breakdown the Big Data into several part and hand them to different computers to achieve parallel processing, such as Hadoop. The major difference between Traditional Data and Big Data is not only that Big Data contains large amount of data but also it is able to store more diverse data type from text, article and numeric to image, audio and video (Jena, 2020). In general, Big Data can be defined by three words: Volume, Velocity and Variety ("What Is Big Data?," n.d.).

Volume means the quantity of the data, Big Data can be high volumes dataset which has high or low dimensionality, such as sensor corresponding data, clicking streams from social media and network transform information. For Traditional Data, the volume of data can be Gigabytes to Terabytes. Yet, in general, the volume of Big Data can be Petabytes to Zettabytes and Exabytes.

Velocity means that how fast the data been collected. Thanks for the efficiency of internet service, companies could easily collect their customer's data online with a very short time and able to analysis it right away in order to evaluate their sale strategy or observing the

trend of industry immediately. However, in traditional way, computer stores the data into its memory in order to deliver the data into system rapidly, yet, since Big Data have high velocity of data collection, the memory could not hold such big amounts of data and leads to system failures while the data has been collected too fast.

Variety means the types of the data could be various, such as images, videos, audio, articles, and metadata. Excepting regular storing purpose, those kinds of data have been proofed as very useful samples for Machine Learning in many applications, such as automated driving, facial recognition, person's sound identification, NLP, and several AI applications. More detail will be represented in this paper.

What is Machine Learning

Machine Learning is a professional domain in computer science which including which involving algorithms design, computational statistics implementation and mathematical optimization. Briefly speaking, Machine Learning is a way to analysis the data with statistical methodologies in order to obtain the insight of the data. With different objective function, the training data could be used as structured or unstructured data with different dimensionality in order to achieve the purpose. As well as the models, depends on the task, different machine learning models have been designed by different methodologies, the most popular models included: Decision Tree, Multi-Layer Perceptron (MLP), Support Vector Machine (SVM) and Neuro Network (NN). The best example about the machine learning application is facial recognition, by processing a person's image, the machine learning model would be able to determine who is the person in real world. In this case, the training data type would be unstructured image pixel matrix data with multi-dimension attributes. By analyzing those

attributes, the machine learning model will be able to identify what attributes are essential for recognizing the target person face.

Furthermore, in order to obtain high accuracy prediction or decision making from the machine learning model, the amount of the data is important. In most of the cases, large data quantities help the machine learning model clearly understanding the attributes benefitted for identifying the data more. As the reason, Big Data has already a part of factor and in some cases playing essential role in Machine Learning.

The application of Big Data and Machine Learning

Since Big Data provide larger and much diverse of data than Traditional Data for companies and researchers for analyzing, Big Data has been used in many real-world applications across financial services, retail, health care, automotive and policy design. In order to analyze such mass and various types of data, Machine Learning become a most popular and fundamental approach due to the statistics attributes in algorithms and the similarity of statistical analysis. By analyzing the attributes of data, Machine Learning is able to find the most reasonable hypothesis for the data in order to fulfill your objective function.

For example, again, facial recognition has been popular implemented in authentication system or surveillance system, such as face key on your computer or custom facial surveillance system in airport. By analyzing the human face's features, the machine learning model capture the most significant points in the face image data which provide affective value for identifying the human face (Singhal et al., 2022). In order to achieve this task, machine learning developers usually require tons of human facial images with or without labels for training. The reason why you need such high quantities of data is because in this case, the machine not only needs to fully

understand the persons' facial attributes will be appeared in different shooting angle, shading and the degree of focus but also needs to point out the human facial attributes between different persons. Other similar task could be done by this patten such as animal, vehicle and art recognition, or breed, age and gender classification, and so on.

In addition, relying on the development of the Internet of Things (IoT), the demand for machine learning applications in this field is also rapidly increasing. The best example would be smart speaker. When you talk to the smart speaker, let take an example, Amazon Echo, it will understand the words you said and give you the feedback or ask you questions for more details. Here we have two machine learning applications in this scenario which included Automatic Speech Recognition, NLP. For Automatic Speech Recognition, the machine learning model learns from tons of audio data from different persons, accents, gender and even languages in order to obtain the precise transformation from speeches to texts (Deng & Li, 2013). Afterward, NLP is involved. The whole name of NLP is Natural Language Processing, which focuses on the analysis the meaning of words, the applications including spam and ham email filter, language translation, text analysis, smart assistant and etc (Brownlee, 2017). In this situation, NLP model is trained by words, sentences or articles in order to understand the meaning of it. By implementing NLP, Amazon Echo is able to understand the words form the speaker in order to do further action.

The other domain about the IoT is cybersecurity. With Big Data, Machine Learning contributes significantly to prevent the potential malicious behavior on internet. Through training the machine learning model to detect online malicious behavior pattern, the model would be able to real time checking the data steam online in order to achieve the first phase of detection and prevention (Brownlee, 2019).

In other fields, such as healthcare, the accuracy of prediction or classification needs to be reliable and precise because the risk of misbehavior is high and can sometimes be fatal (Beam & Kohane, 2018). To instance, by analyzing the symptoms of flu, the machine learning model is able to determine the patients have flu or not based on the symptom they got. The risk of misbehavior sounds really low because it is only about flu. Yet, what if we are going to diagnose cancer, covid or other disease that has high fatal rate. Therefore, by using Big Data, the machine learning model is able trained with more samples of cases in order to improve the accuracy. In addition, some diseases need to be diagnosed by other means which may need to represent by unstructured data, such as lung cancer or broken bones. Regarding the two examples, doctors usually need to perform an X-ray of your body to check your physical condition. In this case, X-ray images would be a perfect training data for Machine Learning to achieve the diagnosis. Other healthcare applications, such as person medical condition detection and medical devices, also benefited from Machine Learning combined with Big Data.

The advantages and challenges of Big Data and Machine learning

Indeed, Big Data helps Machine Learning to have a chance to learn the insight of the data more comprehensively and explore more opportunities in different applications. Generally, Big Data has promoted the development of Machine Learning. However, on the other hand, even though Machine Learning is quite powerful to analysis large data, some limitations and challenges still exist due to the limitations of hardware and the algorithms. Thus, in this section, we will introduce several advantages and challenges based on Big Data's three V: Volume, Velocity and Variety.

Advantages:

Rely on Big Data raising, the data collection is getting easier and more flexible. Also, since Big Data is able to store various types of data, it allows companies and researchers to discover potential application of Machine Learning. In this section, I will introduce you several advantages that Big Data bring to Machine Learning domain in three V's attributes.

Regarding the volume of Big Data, large quantity and more complex dataset help the machine learning model to gain a better understanding of data insight. More data means more samples from real world, and by training more data, it usually helps the machine learning model to come with more realistic hypothesis for solving the tasks.

In addition, as I mentioned before, data collection always hard for certain realms. Due to the convenience of internet, the velocity of Big Data has huge improvement from traditional data, and it helps machine learning developers efficiently collecting the data as their needed. Thanks for the internet, Data steaming is also a very new data collection technics from traditional data which not only allow developers to obtain to newest data in real-time but also provides a wider range opportunities of machine learning application, such as real-time physical condition detection, cybersecurity malicious behavior detection and stock prediction.

With the development of computer vision, audio applications, network security, and the Internet of Things (IoT), Machine Learning requires more and more data types in order to learn from a more realistic perspective. Instead of text, article or simple numeric data, Big Data is not only able to store those kinds of structured data but also able to store the unstructured data such as image, audio, video, log file, sensor respond data and social media post. By implement

different types of data with Machine Learning, many amazing machine learning applications have been realized, and more are worth exploring.

Challenges:

About the volume of Big Data, the first thing we have to consider implementing such large quantity of data into the machine learning model is the processing performance. Typically, the machine learning model requires more time cost for training large dataset. Especially some machine learning models has high time complexity (L'Heureux et al., 2017). For example, SVM (Support Vector Machine) algorithm's time complexity is O(n^3). The n represents the number of input data. Therefore, for SVM to deal with 10 data, it will take 1000 time-units. In this case, training large dataset such as Big Data is very unrealistic and time consuming. In addition, theoretically, when the machine learning model trained with more data, it usually gains better performance, yet it is not always a case since sometimes too much data may cause the model too focus on the pattern of the training data and fail to make the right decision. It is what we called "Over fitting". On the other hand, high dimension data isn't always helping the machine learning model to understand well about the insight of the data. If a machine learning model is trained with a high dimension dataset, we have to make sure that the dataset contains reasonable quantities of data in order to help the model fully understand the pattern of the data, nor it may lead model can't capture the insight of the data, which we called "under fitting", and the phenomenon about training high dimension data with not enough quantities is called " course of dimensionality" (L'Heureux et al., 2017). Moreover, class imbalance is also an issue for Big Data. While the data getting larger, it is hard to make sure the classes are uniformly distributed. Class imbalance may lead to negative affect for machine learning algorithms since the model can

have more understanding about the classes with more data points but less understanding about the classes with less datapoint [6].

About the velocity of Big Data, Machine Learning commonly benefited from it due to the real time analysis demand. However, several issues may occur when the velocity of the data is too high. For example, the machine learning algorithm usually store the data into memory, such as RAM, in order to import the data and export the training results efficiently (L'Heureux et al., 2017). However, if the velocity of data collection is too high, the computer may not be able to store and analysis the data in time and leads to failure. Yet, some solutions have been delivered, such as cloud computing and parallel processing. Also, since traditional machine learning algorithm is not designed to real time real-time analysis. The machine learning model for real time analysis could be limited and required to modify before implementation. Furthermore, while the data steaming, some noise or anomaly data may be collected in real time and those kinds of data could be harmful for Machine Learning. Therefore, how to verify the data while data steaming is also a big challenge.

About the variety of Big Data, it provides the chance to allow Machine Learning to train with structured or unstructured data such as image and video. Yet, in machine learning domain, some limitations and challenges still curse by this attribute of Big Data. As mentioned, traditional machine learning algorithm usually expect the whole dataset is stored into the local memory or single disk file (L'Heureux et al., 2017). However, in Big Data case, local computer cannot hold such large file due to the memory shortage. Even if we distribute such large dataset into several local disks, transferring data from multi-computing locations would cause processing delays and may cause a lot of network traffic (L'Heureux et al., 2017). Moreover, data heterogeneity is also considerable as potential issue in Big Data that may negatively affect the

machine learning model. Due to the variety of Big Data, the diversity of data, such as data types, dataset formatting, data encoding and data model, may require for pre-processing in order to reconcile the variations before feeding the data into the machine learning model (L'Heureux et al., 2017). Also, with different dataset, the interpretations of data maybe very different and such semantic heterogeneity is required for integration (L'Heureux et al., 2017). Finally, in traditional data, dirty and noise data frequently trouble the machine learning algorithm. While the quantity of data increased substantially, dirty and noise data also increases accordingly. Those noise and uncorrelated data be treated as very bias sample in statistic-wise consideration and result in misleading the machine learning model to understand the fact of the data and find the true hypothesis for the objective.

Conclusion

Overall, many machine learning applications in real-world are relying on Big Data. By combining both, companies and researchers accomplished marvelous jobs that highly impact human's life. In addition, by training with Big Data, Machine Learning not only can be benefited from performance enhancement but also have wide-range application opportunities in real world. However, even though Big Data benefit Machine Learning in many aspects, some challenges and disadvantages still exist due to the limitation of hardware and statistic methodologies behind the machine learning algorithm. Thus, in this paper, several real-world applications, advantages and challenges of implementing Machine Learning with Big Data have been discussed.

In my opinion, Big Data offer a lot of potential for machine learning developers to discover the possibilities of data world in order to create something innovational. Therefore, I hope my paper deliver a clear vision about implementing Big Data into Machine Learning in order to help people to understand the importance, the possibilities and the challenges of it.

Reference List

Beam, A. L., & Kohane, I. S. (2018). Big Data and Machine Learning in Health Care. *JAMA, 319*(13), 1317. https://doi.org/10.1001/jama.2017.18391

Brownlee, J. (2017, September 22). *What Is Natural Language Processing?* https://machinelearningmastery.com/natural-language-processing/

Brownlee, J. (2019, December 23). A Gentle Introduction to Imbalanced Classification. *Machine Learning Mastery*. https://machinelearningmastery.com/what-is-imbalanced-classification/

Deng, L., & Li, X. (2013). Machine Learning Paradigms for Speech Recognition: An Overview. *IEEE Transactions on Audio, Speech, and Language Processing*, *21*(5), 1060–1089. https://doi.org/10.1109/TASL.2013.2244083

Jena, S. (2020, September 7). Difference between Traditional data and Big data. *GeeksforGeeks*. https://www.geeksforgeeks.org/difference-between-traditional-data-and-big-data/

L'Heureux, A., Grolinger, K., Elyamany, H. F., & Capretz, M. A. M. (2017). Machine Learning With Big Data: Challenges and Approaches. *IEEE Access*, *5*, 7776–7797. https://doi.org/10.1109/ACCESS.2017.2696365

Polyakov, A. (2018, October 4). Machine Learning for Cybersecurity 101. *Toward Data Science*. https://towardsdatascience.com/machine-learning-for-cybersecurity-101-7822b802790b

Price, D. (n.d.). INFOGRAPHIC: HOW MUCH DATA IS PRODUCED EVERY DAY? *CloudTweaks*. https://cloudtweaks.com/2015/03/how-much-data-is-produced-every-day/

Qiu, J., Wu, Q., Ding, G., Xu, Y., & Feng, S. (2016). A survey of machine learning for big data processing. *EURASIP Journal on Advances in Signal Processing*, *2016*(1), 67. https://doi.org/10.1186/s13634-016-0355-x

Singhal, P., Srivastava, P. K., Tiwari, A. K., & Shukla, R. K. (2022). A Survey: Approaches to Facial Detection and Recognition with Machine Learning Techniques. In D. Gupta, A. Khanna, V. Kansal, G. Fortino, & A. E. Hassanien (Eds.), *Proceedings of Second Doctoral Symposium on Computational Intelligence* (Vol. 1374, pp. 103–125). Springer Singapore. https://doi.org/10.1007/978-981-16-3346-1_9

What is Big Data? (n.d.). *Oracle*. https://www.oracle.com/big-data/what-is-big-data/