MACHINE LEARNING IN BIG LIDAR DATA: A REVIEW

S. S. Teri *, I. A. Musliman

Dept. of Geoinformation, Faculty of Built Environment and Surveying,
Universiti Teknologi Malaysia, 81310 Skudai, Johor,
Malaysia – sazlieya.teri@gmail.com, ivinamri@utm.my

KEY WORDS: Machine Learning, Big LIDAR Data, Deep Learning, Supervised Learning, Unsupervised Learning

ABSTRACT:

Machine Learning used to refer as one Artificial Intelligence subsection that perform self-learning computational algorithms either supervised learning or unsupervised learning tasks. Machine Learning can compute a prediction onto hidden data patterns that hardly for human to detect. This valuable information and predictions able to help companies or researchers make a crucial decision making especially in natural disasters. In Geographic Information Systems (GIS), the advancement of Machine Learning used literally on satellite imagery analyses and fewer on LIDAR point clouds. In this paper presents an overview of Machine Learning definitions, big geospatial data, Machine Learning types and models, and advancement researches using Machine Learning in big LIDAR data.
.

1. INTRODUCTION

1.1 Motivation

Machine learning (ML) is one of the extension of Artificial Intelligence (AI) that perform computational algorithms and have ability of self-learning and modifying when ML is feed with high scale of data (Bini, 2018; Dhande, 2017). ML can make prediction and estimation by using real world datasets (Moorthy & Gandhi, 2017). These datasets at first needed to be used as training datasets so that the machine capable to learn and create pattern recognition and then making their own decisions (Bini, 2018). The algorithms accuracy can be quantifying when the genuine outcomes is used from the real world datasets. ML algorithms will become more precise and predictive when machine is supplied with more and more training datasets and increasing of testing repetitions numbers (Assefi et al., 2018).

Advances in ML have emerged computation processing power to work with real big data (Zhou et al., 2017). Machine learns to predict an inference on data inputs scopes and have been evolved into another learning known as Deep Learning (DL) thus existing models increase its complexity and convolution of the networks dimensions. DL algorithms are more likely appointed to handle huge volumes of datasets and have undoubtedly now penetrated all aspects in our daily lives (Park et al., 2016)

ML have been developed in various of fields (Tehrany, Jones, & Shabani, 2019; Topol, 2019) and its importance to geospatial field has grown but now yet mature. The previous researchers have developed algorithms in such as Flood Susceptibility Mapping (Sachdeva, Bhatia, & Verma, 2017), Flood Risk Assessment (Opella & Hernandez, 2019) and even in development of personalised services in Smart Cities (Chin, Callaghan, & Lam, 2017). Besides, software companies have move forward towards to big data processing and visualization for establish data-driven analysis. The dawn of data increment nowadays its becoming difficult for people to analyse without help, dependency on machines has become essential to execute the analyses (Topol, 2019). Specifically, LiDAR point clouds has become concerns in geospatial field as these LiDAR data are enriched as unstructured data, increasing numbers of points cloud and the complexity for the processing (García-Gutiérrez et al., 2015).

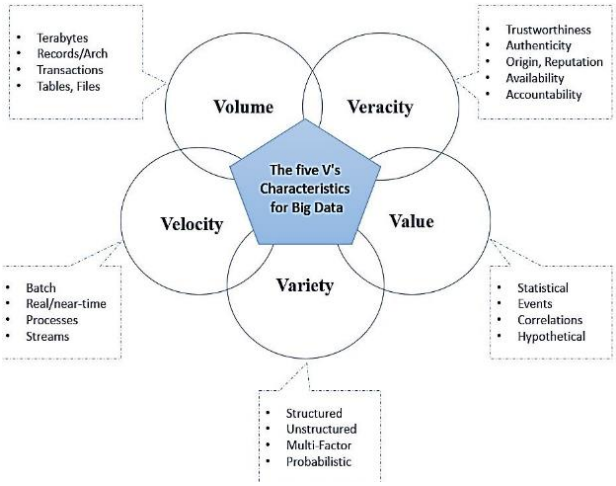
Light Detection and Ranging (LiDAR) has become essential in today's land survey measurement as LiDAR data is able to visualize the two dimensional form of shape into a three dimensional shape that rich in height, shape, type and dimension topology information. LiDAR scanner that was previously built to copy and visualize real earth terrain conditions using UAV tools or aviation (light aircraft) has been evolved to laser scanning either mobile scanning (MLS) or terrestrial laser (TLS) that used to scan indoor buildings. Results from scanning are able to generate a large scale of point clouds for one per square meter. If using a precise scanner, point clouds generation can be generated up to millions of points per square metre and reaches up to Gigabytes up to Terabytes storage.

1.2 Big Geospatial Data

Big Data are often portrayed by its magnitudes, which are indicated to it. Previous meanings of Big Data concentrated on three Vs; volume, velocity, and variety (Narasimhan & Bhuvaneshwari, 2014). However, Big Data definitions have been evolved and accepted into five Vs; volume, velocity, variety, value and veracity (Jasim Hadi et al., 2015). The Digram 3 shows the characteristics of 5Vs.

The definition for the 5Vs can be concluded as:

1. Volume: refers to the amount of data collected
2. Velocity: refers to the time in which Big Data can be processed.
3. Variety: refers to the type of data that contain in Big Data. This data maybe structured or unstructured
4. Value: refers to the important feature of the data which is defined by the added-value of the collected data
5. Veracity: refers to the degree in which a leader trusts Big Data information in order to make a decision.



Digram 3. Characteristics of 5Vs (Jasim Hadi et al., 2015)

1.3 Machine Learning Types and Models

There are three types of machine learning tasks; Supervised and Unsupervised Learning and

1.3.1 Supervised Learning

Supervised Learning is where machine learning is taught to classify data based on desired sample output, where this output data sample is used as data training and machine learning model trained to predict and classify the types of information and variables desired. Next, machine learning will classify new data based on previously trained information and variables. Examples of Supervised Learning flow charts in diagram 1.

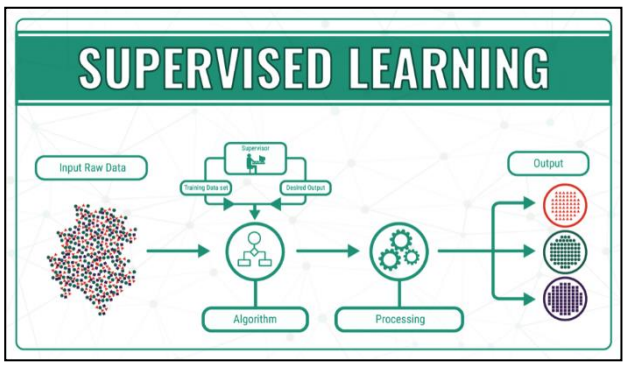


Diagram 1. Supervised Learning Flowchart (Google, 2019)

Supervised Learning is always used for classification and regression context. Algorithms found in Supervised Learning include:

1. Artificial Neural-Network
2. Naïve Bayer Classifiers
3. Support Vector Machine
4. Decision Trees
5. Nearest-Neighbor Classifiers
6. Fuzzy K-Nearest Neighbor

1.3.2 Unsupervised Learning

Unsupervised Learning is where the machine learning model learns about data without guidance such as Supervised Learning. Unsupervised Learning recognizes data with a method of pattern less dataset labels or data references and often

Unsupervised Learning is used to examine and investigate the basic structure of data where we do not have or do not know the desired results. Examples of Supervised Learning flow charts in diagram 2.

Unsupervised Learning is always used for clustering datasets. Algorithms found in Unsupervised Learning are:

1. K-means
2. Hierarchical Clustering
3. Fuzzy C-Means

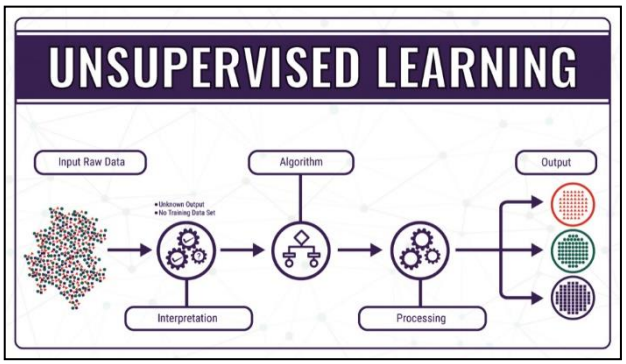


Diagram 2. Supervised Learning Flowchart (Google, 2019)

2. MACHINE LEARNING IN BIG LIDAR DATA

There are several applications in Geographic Information Systems (GIS) used machine learning as a technique to gain automation output.

2.1 Used of Decision Trees (DT) and Support Vector Machine (SVM) for flood conditions factors using LiDAR Dataset.

Researchers (Tehrany et al., 2019) using machine learning supervised algorithms; DT and SVM to study the impact of individual conditioning factors on flood susceptibility mapping and their importance in the construction of precise mapping of potential flood regions. DT and SVM were tested to see spatial correlations between flood conditioning factors and rate their level of importance for mapping the flood prone areas.

Two datasets were used; dataset 1 (DS1): Light Detection and Ranging (LiDAR) derived factors of altitude, slope, aspect, curvature, Stream Power Index (SPI), Topographic Wetness Index (TWI), Topographic Roughness Index (TRI), and Sediment Transport Index (STI) and dataset 2 (DS2): a combination of LiDAR derived factors supplemented by geology, soil, land use/cover (LULC), distance from roads and distance from rivers parameters. DT and SVM were used in this study as these two algorithms techniques are the most popular used by other researchers mainly in natural disaster with minimal accuracy differences and show comparable strengths. The study has concluded that, DT and SVM algorithms techniques offer similar performance as the testing points assessment (prediction) demonstrate DT testing point is 89% while SVM testing points is 87%.

2.2 Comparison of Supervised Learning regression techniques – Neural Network (kNN), Support Vector Machine (SVM), Nearest Neighbor, Gaussian Processes (GP) and Random Forests (RF) with Multiple Linear Regression (MLR) using LiDAR-derived dataset for forest variable estimation.

(García-Gutiérrez et al., 2015) has perform regression techniques comparison; kNN, SVM, Nearest Neighbor, GP, RF and MLR on two study site; Guitiriz and Trabada in province of Lugo, Spain. They have acquired 60 datasets where 20 datasets are extracted from LiDAR data and fieldwork-derived forest variable and another 40 datasets accumulated from power and exponential transformation. Respectively regression techniques were performed on individually 60 datasets for 100 times using altered configuration setups. Additionally, each tuple; configuration setup; algorithm; dataset was performed twice: one with feature selection (CFS) and one without feature selection (CFS). The results from the iteration of these regression techniques studied have shown that, SVM and GP regression technique beat other technique by giving Mean R and Mean RMSE accuracy a superior result.

2.3 Use of Machine Learning to classify tree species from co-registered LiDAR and hyperspectral data.

(Marrs & Ni-Meister, 2019) applied six machine learning techniques which were Decision Trees (DT), Random Forest (RF), k-Nearest Neigbor (kNN), CN2 Rules, Neural Network and Support Vector Machine (SVM) to classify trees species by using G-LiHT thermal imager, hyperspectral and LiDAR datasets from NASA. Two experimental forest data were used to perform this study, which were Howland Experimental Forest with 225-ha forest and Penobscot Experimental Forest is an around 1578-ha forest. Tree species data, tree height, and diameter at breast height (DBH) are obtained from Howland Experimental Forest and Penobscot Experimental Forest. While LiDAR data is obtained as raster format with 13 metre resolution and reflectance data spectrum coverage 418 and 918 nm. The results show that, Neural Network, kNN and RF are the best algorithms for reduce dimensionality and multisource of datasets.

3. CONCLUSIONS

While the revolution ML is exponentially thrusting, it has some kind of weaknesses as a primary concern. The issuing with others as they said developing bias, information misuse, cybercrime and job loss. Additionally, in the field of geospatial, with the vast developments and data increasing have make difficult to analyses, it is important that ML be carefully studied, managed, and validated.

Regardless of its consequences and potential, ML gives a special capacity to make important decision making and predictions in minimal time. As the ML develops to another new level and the technologies become more matured, geospatial field will experience advancements in high performance of geospatial with multiple applications especially for natural disasters decision making.

REFERENCES

Assefi, M., Behravesh, E., Liu, G., & Tafti, A. P. (2018). Big data machine learning using apache spark MLlib. *Proceedings - 2017 IEEE International Conference on Big Data, Big Data 2017*, *2018-Janua*, 3492–3498. https://doi.org/10.1109/BigData.2017.8258338

Bini, S. A. (2018). Artificial Intelligence, Machine Learning, Deep Learning, and Cognitive Computing: What Do These Terms Mean and How Will They Impact Health Care? *Journal of Arthroplasty*. https://doi.org/10.1016/j.arth.2018.02.067

Chin, J., Callaghan, V., & Lam, I. (2017). Understanding and personalising smart city services using machine learning, the Internet-of-Things and Big Data. *IEEE International Symposium on Industrial Electronics*, 2050–2055. https://doi.org/10.1109/ISIE.2017.8001570

Dhande, M. (2017). What is Artificial Intelligence Machine Learning and Deep Learning. *Geospatial World*. Retrieved from https://www.geospatialworld.net/artificial-intelligence-machine-learning-and-deep-learning/

García-Gutiérrez, J., Martínez-Álvarez, F., Troncoso, A., & Riquelme, J. C. (2015). A comparison of machine learning regression techniques for LiDAR-derived estimation of forest variables. *Neurocomputing*, *167*, 24–31. https://doi.org/10.1016/j.neucom.2014.09.091

Jasim Hadi, H., Hameed Shnain, A., Hadishaheed, S., & Haji Ahmad, A. (2015). Big Data and Five V'S Characteristics. *International Journal of Advances in Electronics and Computer Science*, (2), 2393–2835. Retrieved from http://www.iraj.in/journal/journal_file/journal_pdf/12-105-142063747116-23.pdf

Marrs, J., & Ni-Meister, W. (2019). Machine learning techniques for tree species classification using co-registered LiDAR and hyperspectral data. *Remote Sensing*, *11*(7), 1–18. https://doi.org/10.3390/rs11070819

Moorthy, U., & Gandhi, U. D. (2017). *A Survey of Big Data Analytics Using Machine Learning Algorithms*. 95–123. https://doi.org/10.4018/978-1-5225-2863-0.ch005

Narasimhan, R., & Bhuvaneshwari, T. (2014). Big Data – A Brief Study. *International Journal of Scientific & Engineering Research*, *5*(9), 350–353. https://doi.org/10.1111/j.1468-1331.2005.01068.x

Opella, J. M. A., & Hernandez, A. A. (2019). Developing a Flood Risk Assessment Using Support Vector Machine and Convolutional Neural Network: A Conceptual Framework. *Proceedings - 2019 IEEE 15th International Colloquium on Signal Processing and Its Applications, CSPA 2019*, (March), 260–265. https://doi.org/10.1109/CSPA.2019.8695980

Park, S. J., Choi, K. H., Park, J., & Kim, J. B. (2016). A study on spatial analysis using R-based deep learning. *International Journal of Software Engineering and Its Applications*, *10*(5), 87–94. https://doi.org/10.14257/ijseia.2016.10.5.09

Sachdeva, S., Bhatia, T., & Verma, A. K. (2017). Flood susceptibility mapping using GIS-based support vector machine and particle swarm optimization: A case study in Uttarakhand (India). *8th International Conference on Computing, Communications and Networking Technologies, ICCCNT 2017*. https://doi.org/10.1109/ICCCNT.2017.8204182

Tehrany, M. S., Jones, S., & Shabani, F. (2019). Identifying the

essential flood conditioning factors for flood prone area mapping using machine learning techniques. *Catena*, *175*(April 2018), 174–192. https://doi.org/10.1016/j.catena.2018.12.011

Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, *25*(1), 44–56. https://doi.org/10.1038/s41591-018-0300-7

Zhou, L., Pan, S., Wang, J., & Vasilakos, A. V. (2017). Machine learning on big data: Opportunities and challenges. *Neurocomputing*, *237*(December 2016), 350–361. https://doi.org/10.1016/j.neucom.2017.01.026

Revised August 2019