Towards a Taxonomy of Data Mining Algorithms for Internet of Things

Kathleen E. Lange

Regis University

Towards a Taxonomy of Data Mining Algorithms for Internet of Things

Internet of Things (IoT) is at the beginning stages of revolutionizing the current world into a “smart world” where sensors collect data to facilitate wide ranging machine automation of previously human decision oriented tasks (Stankovic, 2014). IoT technology consists of internet enabled devices which can take data input, transmit it to other devices, and provide feedback for decision making (Tsai, Lai, Chiang, & Yang, 2014). Both researchers and the general public anticipate that this budding technology will improve daily activities ranging from transportation, manufacturing, healthcare, and entertainment. Currently, the IoT ecosystem still relies heavily on humans to review acquired IoT data and make determinations. One way to improve daily lives with IoT is to remove the human factor so that decisions can be made quicker and more efficiently. In the Machine-to-Machine (M2M) model of IoT, devices are connected to each other, transmit data between networked objects and automatically make decisions based on the processed data (Chen, 2012).

Due to the great potential of IoT to create a smart world, new devices are coming online at an increasing rate. Evans (2011) estimates that there will be on the order of 50 billion devices by the year 2020. With more devices, data volume is increasing exponentially and could be streamed at terabytes (TB) per hour (Chen, Deng, Wan, Zhang, Vasilakos, & Rong, 2015). In addition to the amount of data, M2M devices have the ability to stream data at rapid speed. Also, due to a many types of devices leads to diverse data. Therefore, Big Data (defined as data that is great in volume, velocity and/or variety) techniques need to be utilized in order to make any sense of the information M2M devices are collecting.

Appropriate machine learning approaches need to be selected for M2M devices in order to process Big Data and make decisions. To choose an appropriate method, device characteristics, such as available memory, processing power and energy usage, along with algorithm characteristics, such as type, accuracy and execution time, need to be considered.

**Research Problem**

For humans to put their trust in M2M outputs, the machine decisions made must be trustworthy. It is critical to choose the best analysis technique which will provide the most accurate results because the decisions made could lead to monetary loss, manufacturing line down time, or possibly even if a person lives or dies. One example from the transportation sector is a recent death due to a driverless car not correctly analyzing the object in its path. The preliminary report from the National Transportation Safety Board (NTSB) stated “as the vehicle and pedestrian paths converged, the self-driving system software classified the pedestrian as an unknown object, as a vehicle, and then as a bicycle with varying expectations of future travel path” (National Transportation Safety Board, 2018, p. 2). This recent failure of a driverless car to correctly classify objects indicates there is great room for improvement in the reliability of M2M decision making.

Generating timely, accurate results is a current challenge when integrating IoT hardware and data mining algorithms (Chen, 2012). Many research questions arise when beginning to analyze the best algorithm. Does there exist a taxonomy for classifying IoT application, physical device characteristics and algorithm features (Mahdavinejad, et al., 2017)? Are current machine learning methodologies insufficient? How does a developer understand the requirements for choosing the best data mining strategy for their M2M device?

**Purpose Statement**

The aim of this study is to develop a guideline, if possible, for identifying the appropriate data mining method based on M2M device type and machine learning algorithm. There are five main analysis approaches of machine learning algorithms: classification, clustering, association, time series and outliers (Chen et al., 2015). The specific IoT application and the device hardware characteristics will both strongly influence the data mining approach taken.

**Significance of Research**

There exist numerous established data mining techniques, however it is untested if these algorithms will be sufficient for IoT or big data streams. A sufficient method would require that the algorithm produces quick and accurate enough results for the application at hand. For devices that could impact a person’s survivability, greater than 99% accuracy is critical. On the other hand, if the device supports a consumer recommendation engine, then appropriate accuracy levels could be sufficient at 70%. In the current state of IoT, there is no single source for developers to reference when determining what the best data mining method or what key characteristics need consideration. By creating a classification system that includes physical device characteristics, IoT application and algorithm features, IoT researchers can start to standardize or understand where gaps exist and develop new mining techniques if necessary.

**Literature Review**

One of the main aims of the IoT is to extract raw data from real world scenarios and then analyze that data to mine information that will be beneficial to enhancing human experiences. Before any meaningful analysis of the raw data can be made, a specific data mining strategy needs to be decided upon. Data mining features for a future IoT world need to include scalability to handle large amounts of data, energy-optimized solutions, and proper data management (Miorandi, Sicari, Pellegrini, & Chlamtac, 2012). Throughout the IoT literature, there are five key interrelated attributes to consider when determining the best data mining method. These attributes are raw data type, location of data analysis, device application, device hardware characteristics, and machine learning algorithm.

**Raw Data Type**

Mahdavinejad et al. (2017) propose three characteristics to initially look for when figuring out what data the incoming data is: streaming, massive or historical. Depending on the data type, it may be more efficient to process the data directly on the edge (i.e. on an individual IoT device or group of closely connected devices) or on a cloud computing service. Furthermore, to choose the most effective type of algorithm, the developer must understand if the data is numerical (i.e. voltages, times), categorical (i.e. dog, fork, house), positional (i.e. GPS coordinates), etc.

Balazinska et al. (2007) describe how current data processing software, like MATLAB and R, and current database structures do not sufficiently meet the need for big data being generated from web connected sensors. Sensors can easily collect data on the millisecond timescale which directly relates to the massive data volume and rapid streaming velocity. Big data also encompasses data variety. Data to be analyzed and compared could be coming from many different times and places or in non-traditional formats such as video images or audio recording (Balazinska et al., 2007).

The data streams from IoT sensors can have many quality issues due to missing data, packet losses, noise, and calibration status (Balazinska et al., 2007). Therefore, there may be a need for pre-processing of raw data before data mining techniques can be effectively implemented. Pre-processing may include eliminating duplicate information, checking for missing data, and transforming data (Tsai et al., 2014). Also, since many IoT devices and sensors are resource constrained then distributed data processing techniques need to be considered (Balazinska et al., 2007; Li, Oikonomou, Tryfonas, Chen, & Xu, 2014).

**Location of Analysis**

When considering where the analysis is going to occur, developers must address processing power of each IoT device, the extent of the data needing mined, and any privacy or security concerns related to the transmission of the data between device nodes. Li et al. (2014) present a robust method for allowing IoT devices to make more efficient and reliable decisions. They suggest using a network of devices that can all transmit data collected to each other. The data is first preprocessed in a local group and then it is transmitted to further away devices until the final decision is available globally across all devices. By constructing the IoT devices in this manner, it relieves some of the resource constraints on individual IoT nodes while still delivering highly accurate decisions. The study also proposes how nodes should be connected by type of application that they will serve.

**Device Application**

One way to determine the data mining needs is to review the IoT application where the device will operate. Different forms of applications include e-commerce, industry, health care and city governance (Chen et al., 2015) among others. No mining algorithm can be 100% accurate which is important for the developer to explore any consequences of a “wrong” decision. Also, as the complexity of both the data and the algorithm increase, so does the time it takes to achieve a final inference (Alam, Mehmood, Katib, & Albeshri, 2016). The developer also needs to consider how fast the decision needs to be relayed back to the system or the user. These are two critical benchmarks to consider when choosing the correct algorithm.

Bi, Xu, and Wang (2014) explore how data mining methods are useful for the manufacturing industry. To improve manufacturing times and quality, process decisions and operator tasks can be automated through the use of M2M technology. In order to do so, the data must be either analyzed on the monitoring/sensor device itself or through network or cloud computing. Similar to all applications, Bi et al. (2014) point out that the critical components of IoT data mining within manufacturing are accuracy, timeliness and any signal/data loss. These data quality metrics should be reported along with any results.

Sun, Song, Jara, and Bie (2016) expound on the city governance application with their description of smart connected cities (SCCs). They propose that data mining methods should not only be used to achieve immediate answers, but also need to promote remembering the past and planning for the future. Therefore, more memory is required in order to save the additional information so that machine learning methods can pull from past and present data to make future predictions. Other SCC challenges include real-time decision making and combining and analyzing different data formats.

One use of machine learning in e-commerce is a recommendation engine. For this case, although accuracy is important, it is not critical to the safety of lives and the timeliness of the decision can be within a couple of seconds. However, in the health care sector accuracy could be extremely important and directly related to patient survivability. Mining IoT device data could assist doctors’ diagnoses and health care plan development. Another thing to consider within the health care sector is privacy and security concerns with transmitting data onto a network or to the cloud. If it is not feasible for the data to be transmitted through encrypted means, then the analysis must occur directly on the device and hardware resource constraints must be considered.

**Device Hardware Characteristics**

For some purposes it may be beneficial to collect the data and perform the analysis to make inferences on the device itself. A developer may choose this model if there are significant privacy concerns related to transmitting the data over a network. Also, network lag times and reliability or packet loss during transfer for very large data streams can play a role in choosing to analyze the data on the device itself. Therefore, the M2M device resource constraints related to processing power, speed, memory and battery life all need to be considered.

Lane, Bhattacharya, Georgiev, Forlivesi, and Kawasar (2015) offer a preliminary study of how data mining algorithms can be tested on M2M-like devices such as smartphones. The authors argue that deep learning methods are the best for making accurate decisions or inferences. However, these types of algorithms also require the longest execution times (Alam, Mehmood Katib, & Albeshri, 2016) and therefore processing power.

In the study, Lane et al. (2015) analyzed four deep learning algorithms on three different M2M representative hardware platforms. The metrics tested were execution time and energy consumption during mining. They also measured total battery life of the device, specifically while the sensors are collecting data, to ensure that the device can operate for at least 16 to 18 hours (a standard human’s awake time) before needing recharged. The results of the testing indicated that for certain types of deep learning algorithms, current IoT device hardware was sufficient to run them effectively. As the algorithm became more complex, both execution time and energy consumption increased, possibly making the machine-made decisions unusable (Lane et al., 2015).

Another key part of the study was to measure the power consumed during different processing stages within the deep learning algorithm. Manipulating energy usage during different parts of the algorithm may help to ease the resource constraints (Lane et al., 2015). This can be another path for future research to reduce execution times and power consumption while increasing accuracies.

Lane et al. (2015) provide a good quantitative study that can be performed in order to start choosing mining algorithms based on the physical limitations of the actual IoT hardware. These types of studies can be useful to the IoT developer to answer questions about how much memory the device has and compare that to how much is needed to execute an algorithm. What sort of processing power does the device have and how much of that is dedicated to data collection, decision making, and other competing factors? Does the device have a sufficient battery life to collect and analyze data throughout a typical use day?

**Machine Learning Algorithm Type**

As discussed in a previous section of this literature review, the first step in choosing the correct algorithm is to fully understand the data type being used. Next the developer must understand if the learning method they need to use is supervised, unsupervised, or reinforcement. For supervised learning the input training set contains a list of data along with the correct output data. In unsupervised learning no output data labels are given with the training set. Finally, reinforcement learning is a method where the best actions are formulated to maximize output results or reward. Most machine learning methods used so far on IoT datasets have been of the supervised and unsupervised types (Mahdavinejad et al., 2017).

There are five main methods or types of algorithms depending on the desired output: classification, clustering, regression, feature detection, and outlier detection. For supervised learning, if the desired output variable is a discrete number or category then a classification machine learning method is used. On the other hand, if the desired output variable is continuous then a regression method is used (Goodfellow, Bengio, & Courville, 2017). As opposed to supervised learning which provides an exact output label from the data, unsupervised learning’s main goal is to cluster the data into appropriate groups using input parameters. If the data is preprocessed and transformed in some manner with a subsequent clustering analysis performed, then this is known as feature extraction or outlier detection (Mahdavinejad et al., 2017). There have been numerous algorithms developed for each of these types of machine learning methods.

Table 1 gives an example of commonly used algorithms alongside what type of method(s) each of those algorithms addresses.

Table 1

*Overview of frequently used machine learning algorithms for smart data analysis – Data adapted from original source Mahdavinejad et al.* *(2017, p. 18).*

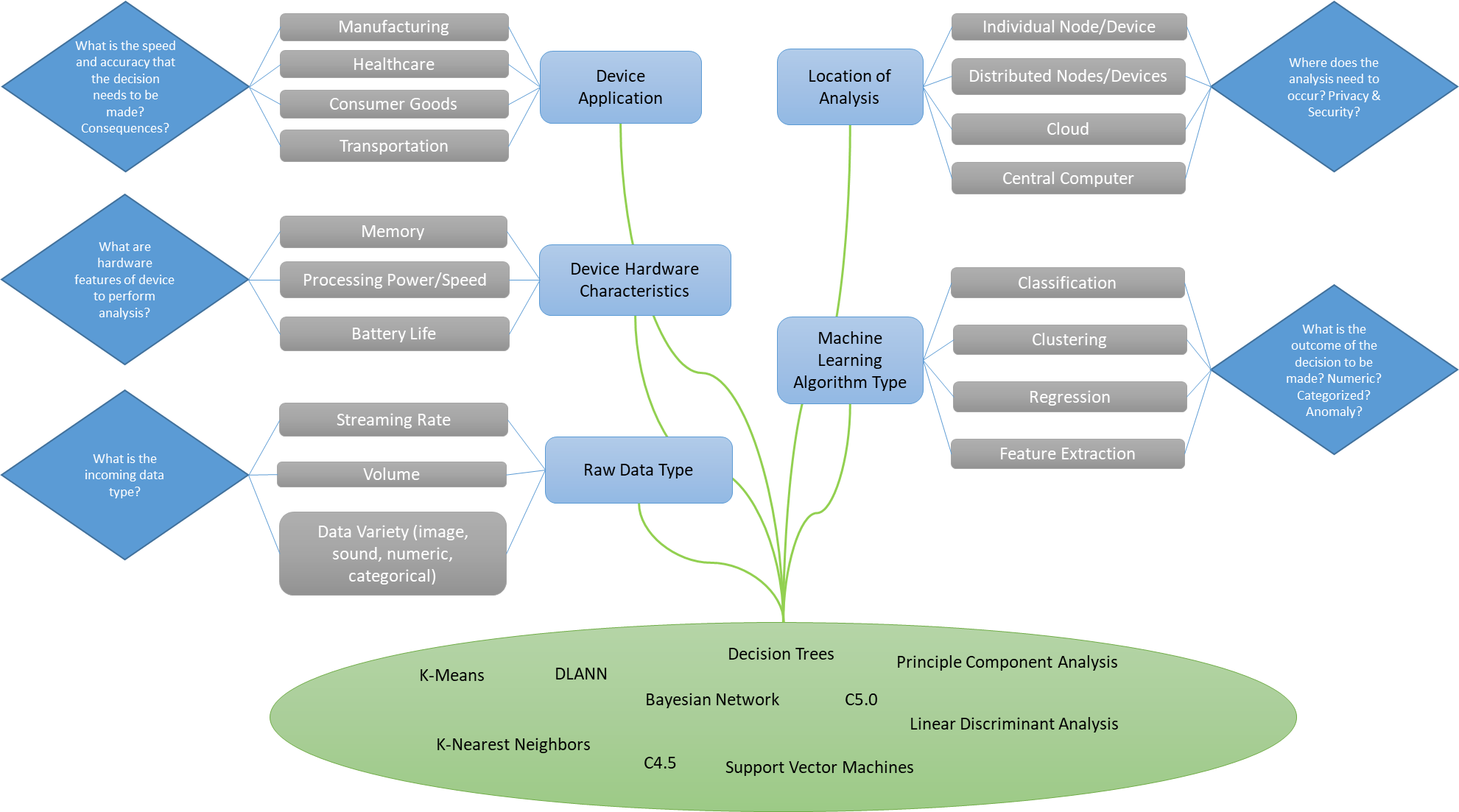
|  |  |  |  |
| --- | --- | --- | --- |
| **Machine learning algorithm** | **Data processing tasks** | **Section** | **Representative references** |
| K-Nearest Neighbors | Classification | 5.1.1 | [58] [59] |
| Naïve Bayes | Classification | 5.1.2 | [60] [61] |
| Support Vector Machine | Classification | 5.1.3 | [62] [63] [64] [65] |
| Linear Regression | Regression | 5.2.1 | [66] [67] [68] |
| Support Vector Regression | Regression | 5.2.2 | [69] [70] |
| Classification and Regression Trees | Classification/Regression | 5.3.1 | [71] [72] [73] |
| Random Forests | Classification/Regression | 5.3.2 | [74] |
| Bagging | Classification/Regression | 5.3.3 | [75] |
| K-Means | Clustering | 5.4.1 | [76] [77] [78] |
| Density-Based Spatial Clustering of Applications with Noise | Clustering | 5.4.2 | [79] [80] [81] |
| Principle Component Analysis | Feature Extraction | 5.5.1 | [82] [83] [84] [85] [86] |
| Canonical Correlation Analysis | Feature Extraction | 5.5.2 | [87] [88] |
| Feed Forward Neural Network | Regression/Classification/Clustering/Feature Extraction | 5.6.1 | [89] [90] [91] [92] [93] [57] |
| One-class Support Vector Machines | Anomaly Detection | 5.8.1 | [94] [95] |

The table also includes a column referring to the section in which the algorithm is defined in detail and references to case studies of where each method has been used in an IoT application. Some of the algorithms deal with large data sets better, while others are more adept at handling different types of data or faster data velocity.

One case study for outlier detection provided by Mahdavinejad et al. (2017) was performed by Souza and Amazonas (2015). Souza and Amazonas (2015) use a Mahout machine learning clustering algorithm to detect outlier events. Mahout is a distributed linear algebra framework that was created to specifically deal with big data such as IoT datasets. The authors only performed clustering from a single sensor’s data for this study, but since they are using the Mahout architecture there are also built in algorithms for classification, regression and other types of data mining methods. This study indicates that there already exist current methods to analyze and mine big data. These methods now need to be tested for accuracy, timeliness, and whether they must be distributed or can occur directly on an edge node device.

To test the different algorithms on an IoT dataset one can initially look at the device hardware characteristics (as discussed in a previous section of this literature review) and investigate the accuracy and execution time of each algorithm on representative datasets. Alam et al. (2016) performed a preliminary research study focused on a quantitative analysis of current standard data mining algorithms using IoT datasets. The benchmarks studied were the classification accuracy and the execution time of eight of the most commonly used algorithms. The results showed the C4.5, C5.0, and neural network algorithms were the most accurate. The authors believe that by optimizing input parameters the neural network mining methods could be the most accurate, but these also take extensive amounts of time. Another current issue facing neural networks is that data scientists struggle with interpretability (Mahdavinejad et al., 2017). That is, because there are many layers to each decision and the calculations to get from one layer to the next are typically non-linear, it can be hard to decipher why an neural network algorithm made the decision it did. This is important information to know when attempting to implement a new algorithm into a safety critical IoT application such as healthcare or self-driving cars.

**Conclusion**

In this short literature review, a full taxonomy for how to choose the correct data mining method for a specific application could not be achieved. However, there were five key interrelated attributes that will be necessary in defining any sort of classification system for the use of data mining methods for IoT datasets. These key attributes are the type of raw or preprocessed data being analyzed, if the analysis will occur on a single device or be distributed over a larger set of connected nodes, the application in which the device will be used, the hardware characteristics of the devices doing the analysis, and finally the type of data mining and machine learning method. Figure 1 shows a preliminary structure of how each of the five key components feeds into choosing the most appropriate algorithm.

*Figure 1.* Questions and key characteristics that go into choosing an effective machine learning algorithm for an IoT device.

Presented in this literature review were specific case studies pointing towards future research that can be performed in order to create a detailed classification system. These case studies included how to measure the accuracy and execution time of the algorithm, how to measure device hardware characteristics, and how to create a distributed network of nodes to perform analyses. Through these studies, current data mining methods have been shown sufficient. Data scientists and developers should focus on optimizing these methods for use on IoT datasets prior to creating new algorithms.

References

Alam, F., Mehmood, R., Katib, I., & Albeshri, A. (2016). Analysis of Eight Data Mining Algorithms for Smarter Internet of Things (IoT). *Procedia Computer Science, 98*, 437-442. doi:10.1016/j.procs.2016.09.068

Balazinska, M., Deshpande, A., Franklin, M. J., Gibbons, P. B., Gray, J., Hansen, M., . . . Tao, V. (2007). Data management in the worldwide sensor web. *IEEE Pervasive Computing, 6*(2), 30-40. doi:10.1109/mprv.2007.27

Bi, Z., Xu, L. D., & Wang, C. (2014). Internet of things for enterprise systems of modern manufacturing. *IEEE Transactions on Industrial Informatics, 10*(2), 1537-1546. doi:10.1109/tii.2014.2300338

Chen, F., Deng, P., Wan, J., Zhang, D., Vasilakos, A. V., & Rong, X. (2015). Data Mining for the Internet of Things: Literature Review and Challenges. *International Journal of Distributed Sensor Networks,11*(8), 431047. doi:10.1155/2015/431047

Chen, Y.K. (2012, January). Challenges and opportunities of internet of things. Proceedings of the Asia and South Pacific Design Automation Conference (pp. 383-388).

Evans, D. (2011, April). *The internet of things: How the next evolution of the internet is changing everything.* Retrieved from Cisco: https://www.cisco.com/c/dam/en\_us/about/ac79/docs/innov/IoT\_IBSG\_0411FINAL.pdf

Goodfellow, I., Bengio, Y., & Courville, A. (2017). *Deep learning*. Cambridge, MA: MIT Press.

Lane, N. D., Bhattacharya, S., Georgiev, P., Forlivesi, C., & Kawsar, F. (2015). An early resource characterization of deep learning on wearables, smartphones and internet-of-things devices. *Proceedings of the 2015 International Workshop on Internet of Things towards Applications - IoT-App 15*. doi:10.1145/2820975.2820980

Li, S., Oikonomou, G., Tryfonas, T., Chen, T. M., & Xu, L. D. (2014). A distributed consensus algorithm for decision making in service-oriented internet of things. *IEEE Transactions on Industrial Informatics, 10*(2), 1461-1468. doi:10.1109/tii.2014.2306331

Mahdavinejad, M. S., Rezvan, M., Barekatain, M., Adibi, P., Barnaghi, P., & Sheth, A. P. (2017). Machine learning for Internet of Things data analysis: A survey. *Digital Communications and Networks*. doi:10.1016/j.dcan.2017.10.002

Miorandi, D., Sicari, S., Pellegrini, F. D., & Chlamtac, I. (2012). Internet of things: Vision, applications and research challenges. *Ad Hoc Networks, 10*(7), 1497-1516. doi:10.1016/j.adhoc.2012.02.016

National Transportation Safety Board (NTSB). (2018, May 24). *Preliminary report highway HWY18MH010*. Retrieved from https://www.ntsb.gov/investigations/AccidentReports/Reports/HWY18MH010-prelim.pdf

Souza, A. M., & Amazonas, J. R. (2015). An outlier detect algorithm using big data processing and internet of things architecture. *Procedia Computer Science, 52*, 1010-1015. doi:10.1016/j.procs.2015.05.095

Stankovic, J. A. (2014). Research Directions for the Internet of Things. *IEEE Internet of Things Journal,1*(1), 3-9. doi:10.1109/jiot.2014.2312291

Sun, Y., Song, H., Jara, A. J., & Bie, R. (2016). Internet of things and big data analytics for smart and connected communities. *IEEE Access, 4*, 766-773. doi:10.1109/access.2016.2529723

Tsai, C., Lai, C., Chiang, M., & Yang, L. T. (2014). Data Mining for Internet of Things: A Survey. *IEEE Communications Surveys & Tutorials,16*(1), 77-97. doi:10.1109/surv.2013.103013.00206