Attribute-Based K-Means Algorithm

Anand Prakash
Institute for Systems Studies and Analyses
DRDO
Delhi, India
adprakash2006@gmail.com

Y. S. Chungkham
Institute for Systems Studies and Analyses
DRDO
Delhi, India
ysangeeta@issa.drdo.in

Mohd. Yousuf Ansari
Defence Scientific Information and Documentation Centre
DRDO
Delhi, India
md_ya@yahoo.com

***Abstract*— Clustering is a method to discover hidden natural structure in a dataset of a phenomenon. In this study, we have extended K-Means algorithm for spatiotemporal dataset by introducing attribute-based mass function to calculate center of mass of cluster instead of calculating geometry-based centroid in the dataset. The proposed modification in traditional K-Means algorithm produces more meaningful clusters and converges faster than traditional K-Means. In our study, we have used a real 'fire dataset' to conduct experiments on the proposed approach for clustering.**

Keywords*— *Clustering, K-Means, Data analytics, Mass function, Center of mass of a cluster.

I. INTRODUCTION

The aim of this study is to apply mathematical methods like clustering on forest fires dataset with the use of R to support data-driven decision making. This would result in allowing to make predictions based on past trends in data, about which area is more likely to face forest fires in which season so that necessary precautions can be taken to reduce the likelihood if possible. It would help with better maintenance and planning as the government would be aware, more safety measures can be taken in advance. Not only does this have environmental benefits, but also economic benefits as it would help the local fire authority in firefighting resource management in terms of prioritizing targets for air tankers and ground crews. Looking at the social factor, the habitants of the specific region will feel more secure knowing that safety measures have been taken already.

One of the major environmental issue are forest fires. In the United States, each year 100,000 forest fires have been reported, due to which over 9 million acres of land have already been destroyed. Forest fires severely affect forest preservation, cause human suffering and creates economic and ecological damage. The severity of the forest fires can be visualized by looking at the graph plotted in figure 1. Human lives and wildlife are endangered due to this phenomenon. There are multiple causes, including both human negligence and nature. Whilst nature, i.e. lightnings, volcanic eruption, account for 10% of forest fires in the U.S, man-made causes consisting of burning debris, cigarettes, fireworks, account for 90% of forest fires. Each year, the occurrence of forest fires has been increasing as well as the state expenses to control this disaster. It is difficult to predict and detect forest fire, especially in sparsely populated forest area however it is even more difficult to make predictions. Fast detection by predicting forest fires occurrences in a specific area is the key for controlling such phenomenon [1].

There are many phenomena in nature which can be modeled in space and time and thus producing huge amount of spatiotemporal data. The spatiotemporal data can be categories into five different types: events, geo-referenced data items, geo-referenced time series, moving objects, and trajectories [2]. Spatiotemporal data can also be expressed as changing geometries in space with respect to time. Mathematically, spatiotemporal event can be expressed by triplet <spatial coordinates, time, attribute >.

Clustering can be used to transform a spatiotemporal data into useful information that, in turn, can be used to get some meaningful conclusion about the phenomenon under study. K-means clustering is a method of unsupervised learning in Machine Learning, aims at finding groups in data and clustering them based on feature similarity. It can be used to make predictions on likelihood of forest fires in a specific area in relationship to the season of the year. This would allow to effectively understand how season and the occurrences of forest fires are related, through which decisions can be made by the local fire authorities to reduce state expenses by proper planning.

The performance of traditional K-Means algorithm is highly depended on initial values of cluster's centroids [3]. The traditional K-Means algorithm produces good results only when the initial chosen centroids are close to final centroids of the clusters [4]. The proposed Attribute-Based K-Means algorithm, however, doesn't solve the initialization problem but it finds out weighted means of the cluster by utilizing 'mass function' in finding center of masses of clusters which is described in definition (2) and (3). The idea of calculating center of masses instead of geometry-based centroids gives a sensible guidance to reach final centers of clusters. Critical attribute in a data object acts as mass of the object and helps in forming meaningful clusters.

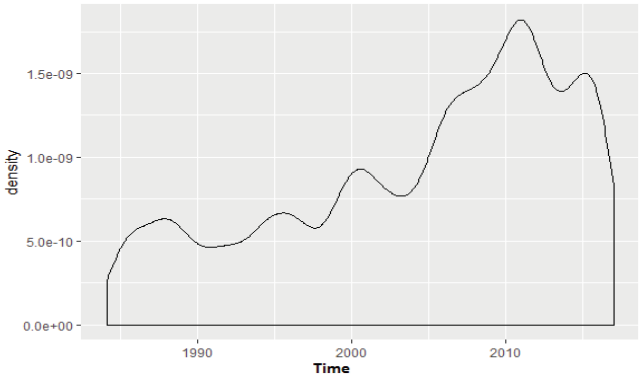

Figure 1. Forest fire density of Alaska region in U. S.

ISBN: 978-1-7281-4826-7/19/$31.00 ©2019 IEEE

II. Related Work

Traditional data analytics focused on the use of simple techniques including data warehouse which is a system used in the collection of data from various sources to provide; data classification of structured/semi-structured/unstructured data, and the use of statistical analysis such as linear regression and clustering. However, with the size and complexity of data increasing, which is now referred to as 'Big Data', new methods were needed to be introduced to handle such data. The concept of data mining applies data analysis algorithms to create predictive models. Nowadays various techniques have been devised from the concepts of data mining, predictive analysis, artificial intelligence, machine learning, and deep learning. In this section, conventional, present and future methods of data analysis are discussed in the context of literature reviews written by many researchers.

Tsai and Lai, et al [5] discuss the conventional methods used in data analysis and provide a review of various representative algorithms such as clustering, classification; used to classify inputs, sequential patterns and association rules; used to determine relationships. These methods were sufficient to provide efficient results.

A study by Ahuja and Jani [6], focused on the ways in which traditional methods are different from big data analytics, as sensor data can be classified as big data due to its volume generated. The disadvantages of traditional methods were as follows; there is unscalability and centralization as these methods are not suitable for large-scale and complex datasets. It is assumed that the methods are performed in a single machine however with increased volume of big data, this is a limitation to big data. There is an issue with non-dynamism as the classifiers are treated as being fixed, and cannot be adjusted with different situations. This factor is not concerned with the velocity problem of big data. Lastly, methods assume that there is uniform data structure and that the format of the input data is the same, with big data there is a variety in the input data and hence traditional methods cannot be applied directly.

Baum, Laroque, et al [7] have conducted a study in which they clustered scientific papers based on the implementation approaches that have been used in production and maintenance. With the help of digitalization, the use of sensors and manufacturing execution systems (MES) can collect data, enabling the use of statistical and machine-learning methods to improve productivity in the maintenance process. Baum and Laroque concluded that based on the number of publications published, there has been an increase in predictive analytics followed by descriptive analytics in production systems.

A study conducted on Social media analytics by Stieglitz, Stefan et al [8], analyzed the use of social media data to gain insights regarding trends, issues, influential actors and other vital information. Through social network analysis, influencers or opinion leaders can be identified which helps reveal the reach of such individual by examining their follower network on social media sites such as Twitter, Facebook. Furthermore, by examining the behavior of the roles, the causes and effects of the key role in network can be understood.

The field of supply chain management (SCM) uses Twitter Analytics to analyze the supply chain tweets to develop insights about supply chain practice and research. Chae and Bongsug [9] used different methodologies to extract useful information from 22,399 tweets. The findings from these findings provided useful insights about the role of Twitter in SCM and analysis the supply chain risk management, product and service development, professional networking and stakeholder engagement.

Another research that used Twitter as the basis for analysis is a research by Singh and Shukla [10] wherein the supply chain of beef was analyzed using three weeks of data from Twitter. In this research, text analysis using support vector machine (SVM) and hierarchical clustering with multiscale bootstrap resampling was done to extract useful information. The findings helped a lot to inform about the customer feedback to top level supply chain decision makers to make the flow and quality of food products better. It has been applied well in supply chain management system.

In the field of healthcare and medical science, big data analytics has revolutionized the industry by deriving useful insights to increase the effectiveness of health information systems. Due to recent advancements in the field, record keeping has been made easier by automated means such as electronic medical report and personal healthcare record systems. Taking motivation from many studies, Adenuga, Kayode & Muniru et al. [11] devised their own model from a dataset consisting of three different parts; bio-data, personality measure and drug use history.

Velinov and Zdravev [12] study analyze Big Data from Moodle Database and Logs, Moodle is an e-learning system used at the University Goce Delcev in Stip. Through the use of analysis of Big Data and clustering tools, knowledge was gained about the behavior of the teachers and hence improve teaching standards, learning process by changing educational process and improve the way in which the which are organized. The teaching methods and the effect of the behavior of the students was examined by big data analysis by looking at the number of downloaded resources, student activity, providing a vital feedback for teachers.

Another study on big data analytics services and healthcare services was examined by Sakr and Elgammal [13]. Big data analytics can be useful to monitor and detect vital signs that can support physicians and healthcare providers in the diagnosis process. In the recent time, diagnosis can be automated to minimize or eventually eliminate the requirement of doctor's visit time to time for simple illnesses such as flu.

Zafeiropoulos and Fotopoulou, et al. [14] proposed the need of analysis design and implementation that supports the overall lifecycle of the analysis process rather than being dependent on a tool or technique that is currently being used. Behavioral analytics is the future trend in data analytics as it is based upon a dynamic concept which is a matter of change and can be applied to various domains. The study focused on the energy and behavioral analytics to help increase efficiency of energy in smart buildings by looking at the change in behavior of the population.

Furthermore, Kalamaras et al [15] aimed their study on visual analytics and how interactive ITS visual analytics tools can help reduce traffic by focusing on historical data and making prediction of future traffic with the use of unified interactive interface. The approach suggested the use of various advanced data analysis techniques such as anomaly detection, road behavioral visualization and

clustering and traffic prediction to help examine the behavioral similarities between roads, the testing of hypotheses and predicting the flow after hypothetical incidents and provide a visual of unusual events detection.

Kibria, Nguyen et al. [16] discussed the role of Big Data analytics, artificial intelligence and machine learning in next-generation wireless networks. They examined how wireless networks has evolved as a complex system, and how predictive networking is needed for cost effective operation and optimization. They examined the opportunities that wireless networks system brings with it, in the field of Big Data analytics. Data-driven wireless networks are envisioned in which the network operators employ and adopt advanced data analytics method, artificial intelligence and machine learning in order to make the system intelligent in terms of being self-adaptive, self-aware, prescriptive and proactive. Additionally, different network designs and optimization schemes were presented concerning data analytics.

Sumathi, Santharam and Selvalakshmi [17] examined the future role of data analytics in the field of intelligent agriculture, to help construct a platform that helps farmers in decision making information about how to improve the yield, details about pesticide for their farm, irrigation control, and agriculture market of their goods. With the use of machine learning techniques and algorithms, information could be provided regarding weather forecasts, crop yield productivity, avoiding unnecessary cost in harvesting. Spatial data mining techniques are proposed for extracting information from spatial data sets for decision making. Along with this, predictive analytics were proposed to allow them to make smart decisions from the information collected and examined by intelligent, and real time data from field sensors.

Another future application of data analytics is in renewable energy as explored by a study conducted by Jha, Bilalovic et al. [18]. An important research and development field has been the exploitation of sunlight and air as a substantial renewable energy. The future comprises of developing technology for optimum production of renewable energy from the natural resources that are available. This may be achieved with the use of artificial intelligence such as statistical and biological method. They proposed how the idea of hybrid artificial intelligence approaches can help achieve the common and future aims of renewable energy.

There has been a lot of discussion of the recent trends of Internet of Things (IoT) and Big Data analytics. This trend was discussed in detail by Mohammadi and Al-Fuqaha et al in their study [19]. With the increase in IoT, tremendous amount of data has been collected and generated from sensing devices which has many applications, one of which is IoT streaming data analytics. The data streams provide valuable information about the future and in making control decisions. Furthermore, the use of advanced machine learning techniques such as deep learning was explored in the domain of IoT and why it helps achieve the desired analytics. For example, how smart IoT devices have incorporated deep learning, likewise the implementation of deep learning in fog and cloud centers has brought potential in the field of IoT and Big Data analytics.

III. ATTRIBUTE-BASED K-MEANS ALGORITHM

The traditional K-Means clustering algorithm is centroid-based but we propose to use center of mass of a cluster to measure the dissimilarity. It has been observed that in some phenomena in nature like forest fire, heavy fire area tends to produce more fire events around it. Hence, it is sensible to use burn area attribute of a fire event as mass to form clusters. We can assume that a critical attribute like burn area acts as a mass for the data object. By adopting this idea, we propose to extend the traditional K-Means algorithm as follows:

Definition 1. Clustering is a process of partitioning a given dataset $X = \{x_1, x_2, \ldots, x_n\}$ that contains n number of data objects into a set of K number of partitions called clusters $C = \{c_1, c_2, \ldots, c_K\}$ such that $c_i \neq \phi$, $(i = 1, 2, \ldots, K)$ and $c_i \cap c_j = \phi$, $(i \neq j)$ and $c_i \cup c_j = X$, $(i \neq j)$.

Definition 2. A mass function is a real valued function that quantifies j^{th} attribute of the i^{th} object $x_i \in X$ i.e. $M_i: A_{ji} \rightarrow R$, where $A_{ji} \in x_i$

Definition 3. Center of mass μ_i of a cluster c_i is calculated as per equation (1)

$$\mu_i = \sum (M_i \times \text{Spatiotemporal Coordinates}) / \sum M_i \quad (1)$$

Input: Dataset $X = \{x_1, x_2, \ldots, x_n\}$ containing n objects and number of clusters i.e. value of K

Output: A set of clusters $C = \{c_1, c_2, \ldots, c_K\}$

Algorithm

1. Initialize: Randomly select K number of attributes to generate initial K center of masses.
2. Iterate:
   a. Assign data objects to the closest cluster
   b. Update center of masses of each cluster according to equation (1) in definition 3.
3. Halt: Stop when center of masses remain unchanged

Figure 2. Attribute-Based K-Means

Our proposed algorithm (shown in Figure 2), takes two parameters a dataset (X) and the number of clusters (K) as input and produces K number of clusters as output. The algorithm has three phases namely initialization, iteration and halt. In initialization phase we generate K random center of masses by choosing critical attributes of K number of data objects in the dataset. In iteration phase we assign data objects to the closest cluster by calculating the Euclidean distance of each data object from the available center of masses in the dataset to decide which the closest cluster is. We update center of masses of the clusters in subsequent iterations. In halt phase, we use some stopping criteria to converge the algorithm finally; one of the criteria can be when center of masses in i^{th} and $(i+1)^{th}$ do not change; or we can also use the objective function of the traditional K-Means algorithm to converge the algorithm.

IV. EXPERIMENT

The proposed algorithm is applied to fire data set [20], which was prepared by Department of Agriculture, USA.

The data pertains to the frequency, extent and magnitude of all wildland fires of different parts of USA from 1984 to 2016. Each fire event is having 18 attributes. We select the relevant attributes such as latitude, longitude, date, fire type and burnt area. The dataset contains 21,673 events of fire, on which our proposed algorithm is applied (shown in figure 3).

We execute the proposed algorithm on the dataset by taking K = 2, 3 and 4 initially because we don't have precise knowledge about number of clusters to be formed in the dataset. Thereafter, we calculate ratio of two statistical validity indices namely SSB (Sum of Squared Between) and SST (Sum of Squared Total) to assess the quality of clusters. The experiment starts by taking K = 2 and after executing the algorithm it is found that two clusters are formed having 1915 and 19758 fire events respectively with SSW/SST = 33.5%. Taking K = 3, it is found that three clusters are formed with 5300, 7098 and 9275 fire events with SSW/SST = 91.9%. Taking K = 4, it is found that four clusters are formed having 5300, 7360, 1915 and 7098 fire events with SSW/SST = 100.0%. Thus, we conclude that the value of K should be 4 for this experiment.

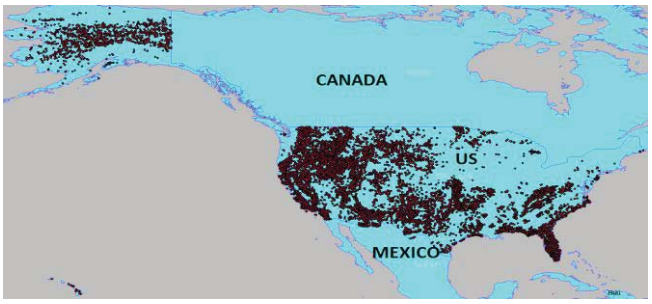

Figure 3. Fire dataset of Alaska region in United States

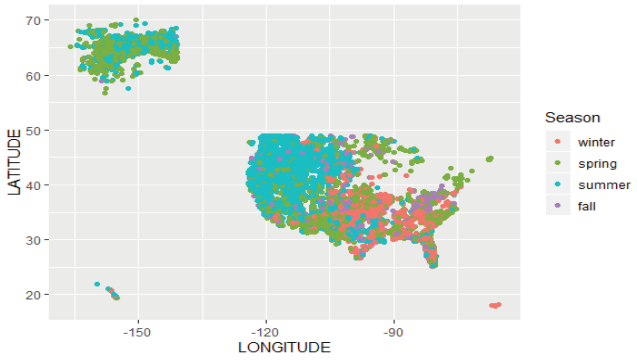

Figure 4. Map showing the seasons of the forest fires

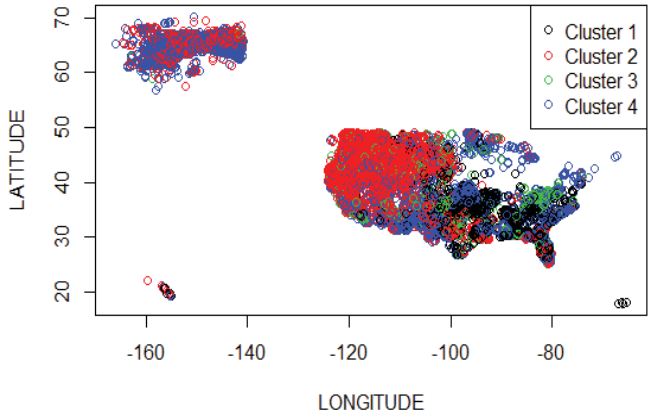

Figure 5. Spatiotemporal clusters in 2D

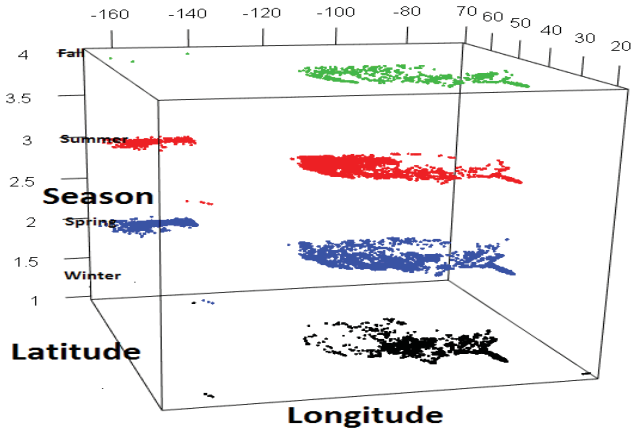

Figure 6. Spatiotemporal clusters in 3D

From the experiment, it is clear that forest fires occur during a specific season in a year, in a specific region allowing to make predictions about the future in the sense which areas will be susceptible to fires in which seasons (shown in figure 4). Taking Alaska as the monitoring area, obtained clusters are shown in 2D and 3D in figure 5 and 6 respectively; each cluster represents the similarity in season of occurrence. The cluster (2) is shown in green color and cluster (3) is shown in black color. It has been observed in the monitoring area that lightning fires are common after the dry seasons when the vegetation is dry and due to this reason 400 fires per year are reported which is evident by cluster (1) which is shown in blue and cluster (4) which is shown in red colors. As the clusters are formed according to the seasons forest fire occurrence happens maximum during the season of spring and summer in the region.

During the experiment it has been found that proposed algorithm has better performance than traditional K-Means algorithm in terms of number of passes required to converge the algorithm. The performance graph is shown in figure 7.

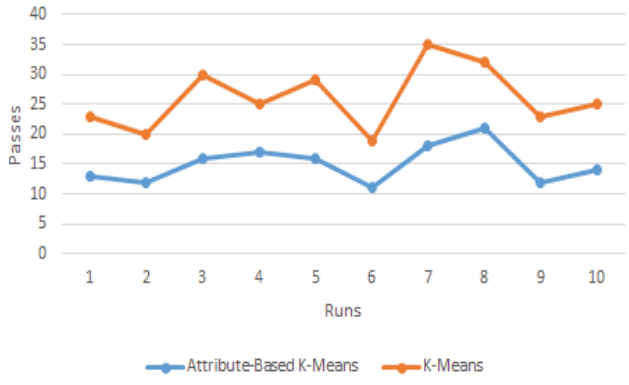

Figure 7. Performance graph

V. Conclusion And Future Work

As it could be seen, K-means clustering is a powerful algorithm to predict patterns based on creating clustering on the similarity criteria, in this case it is the occurrence during the four seasons in aforementioned monitoring area. Using a critical attribute as mass of the data object and adopting the idea of center of mass in the proposed algorithm, rather than using geometrical centroids used in traditional K-Means algorithm, can produce more accurate clusters semantically. It is fast, robust and easy to understand the patterns present through the analysis. However, the biggest disadvantage is

that the number of clusters (K) needs to be pre-specified; but this problem can be simplified when additional information like temperature profile is available to predict number of clusters for the algorithm. For the forest fire dataset and analyzing the data based on the seasons, it is clear that there were four seasons hence four clusters must be formed by looking at the temperature profile of the reason. This claim is also substantiated experimentally to reach an accurate value of 100% for the ratio between sum of squares and total sum of squares in each cluster in case when K = 4.

By extending this study of predicting the occurrence of forest fires and how to monitor them, future work can be done at looking at the causes of man-made forest fires and using this information in a dataset to analyse how man-made activities needed to be monitored carefully. Another more interesting possibility is comparing Big Data collected on forest fires from different approaches of recording data and looking at the accuracy of the predictions concluded from them.

ACKNOWLEDGMENT

We would like to show our gratitude to Ms Niharika Pannu, Department of Computer Science & Engineering, Dr. B. R. Ambedkar National Institute of Technology Jalandhar, Punjab, India for sharing her pearls of wisdom with us during the course of this research.

REFERENCES

[1] Jerry Gao, Kshama Shalini, Navit Gaur and Xuan Guan, "Data-driven Forest Fire analysis," *2017 IEEE International Conference on Smart City and Innovation*, East Bay of Silicon Valley, Fremont, California, USA, Volume: 1, 2017.

[2] M.Y. Ansari, A. Ahmad, S.S. Khan, G. Bhushan, and Mainuddin, "Spatiotemporal clustering: a review," *Artificial Intelligence Review*, 2019. (In Press). https://doi.org/10.1007/s10462-019-09736-1

[3] J. M. Pena, J. A. Lozano, and P. Larranaga, "An empirical comparison of four initialization methods for the K-means algorithm," *Pattern Recognition Letters*, vol. 20, pp.1027– 1040, 1999.

[4] A. K. Jain and R. C. Dubes, "Algorithms for Clustering Data," Prentice Hall, Englewood Cliffs, 1988.

[5] C. W. Tsai, C. F. Lai, H. C. Chao, and A. V. Vasilakos, "Big data analytics: a survey," *J. Big Data*, vol. 2, no. 1, pp. 1–32, 2015.

[6] K. Ahuja and J. N.N, "A Study of Traditional Data Analysis and Sensor Data Analytics," *Int. J. Inf. Sci. Tech.*, vol. 6, no. 1/2, pp. 185–190, 2016.

[7] J. Baum, C. Laroque, B. Oeser, A. Skoogh, and M. Subramaniyan, "Applications of big data analytics and related technologies in maintenance-literature-based research," *Machines*, vol. 6, no. 4, 2018.

[8] S. Stieglitz, M. Mirbabaie, B. Ross, and C. Neuberger, "Social media analytics – Challenges in topic discovery, data collection, and data preparation," *Int. J. Inf. Manage.*, vol. 39, no. December 2017, pp. 156–168, 2018.

[9] Chae and Bongsug, "Insights from hashtag #supplychain and Twitter Analytics: Considering Twitter and Twitter data for supply chain practice and research," *International Journal of Production Economics*, Volume 165, July 2015, Pages 247-259, 2015.

[10] Akshit Singh, Nagesh Shukla and Nishikant Mishra, " Social media data analytics to improve supply chain management in food industries," *Transportation Research Part E: Logistics and Transportation Review*, Volume 114, June 2018, Pages 398-415, 2018.

[11] Kayode I. Adenuga, Idris Oladele Muniru, Fatai Idowu Sadiq, Rahmat O. Adenuga and Muhammad J. Solihudeen, "Big Data in Healthcare: Are we getting useful insights from this avalanche of data?," in *8th International Conference on Software and Information Engineering, ICSIE 2019,* Pages 196-199, ACM New York, NY, USA, 2019.

[12] Velinov, Aleksandar, Ljupce Janevski, & Zoran Zdravev, " Analyzing Teachers Behavior Using Moodle Data And Big Data Tools," *Balkan Journal of Applied Mathematics and Informatics*, vol. II, August, 2018, 2018.

[13] Sakr, Sherif and Amal Elgammal, "Towards a Comprehensive Data Analytics Framework for Smart Healthcare Services," *Big Data Research*, Volume 4 Issue C, June 2016, Pages 44-58, 2016.

[14] A. Zafeiropoulos, E. Fotopoulou, A. González-Vidal and A. Skarmeta, "Detaching the design, development and execution of big data analysis processes: A case study based on energy and behavioral analytics," *2018 Global Internet of Things Summit (GIoTS)*, Bilbao, June 2018, pp. 1-6, 2018.

[15] I. Kalamaras *et al*., "An Interactive Visual Analytics Platform for Smart Intelligent Transportation Systems Management," in *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 2, pp. 487-496, Feb. 2018.

[16] M. G. Kibria, K. Nguyen, G. P. Villardi, O. Zhao, K. Ishizu and F. Kojima, "Big Data Analytics, Machine Learning, and Artificial Intelligence in Next-Generation Wireless Networks," in *IEEE Access*, vol. 6, pp. 32328-32338, 2018.

[17] K. Sumathi, Kundhavai Santharam and N. Selvalakshmi, "Data Analytics platform for intelligent agriculture," *2018 2nd International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*, 2018.

[18] Jha, Sunil Kr. & Bilalovic, Jasmin & Jha, Anju & Patel, Nilesh & Zhang, Han, "Renewable energy: Present research and future scope of Artificial Intelligence," *Renewable and Sustainable Energy Reviews*, Elsevier, vol. 77(C), pages 297-317, 2017.

[19] M. Mohammadi, A. Al-Fuqaha, S. Sorour and M. Guizani, "Deep Learning for IoT Big Data and Streaming Analytics: A Survey," in IEEE Communications Surveys & Tutorials, vol. 20, no. 4, pp. 2923-2960, Fourthquarter 2018.

[20] "Monitoring Trends in Burn Severity Fire Occurrence Locations (Feature Layer)", https://catalog.data.gov/dataset/monitoring-trends-in-burn-severity-fire-occurrence-locations-feature-layer/resource/31610169-50d8-46fa-8f32-41feaa1c7fc3.