Labeling Big Spatial Data: A Case Study of New York Taxi Limousine Dataset

Fawaz AlBatati **Louai Alarabi**

Umm Al-Qura University, College of Computer and Information Systems, Department of Computer Science,
Makkah, Kingdom of Saudi Arabia

Summary
Clustering Unlabeled Spatial-datasets to convert them to Labeled Spatial-datasets is a challenging task specially for geographical information systems. In this research study we investigated the NYC Taxi Limousine Commission dataset and discover that all of the spatial-temporal trajectory are unlabeled Spatial-datasets, which is in this case it is not suitable for any data mining tasks, such as classification and regression. Therefore, it is necessary to convert unlabeled Spatial-datasets into labeled Spatial-datasets. In this research study we are going to use the Clustering Technique to do this task for all the Trajectory datasets. A key difficulty for applying machine learning classification algorithms for many applications is that they require a lot of labeled datasets. Labeling a Big-data in many cases is a costly process. In this paper, we show the effectiveness of utilizing a Clustering Technique for labeling spatial data that leads to a high-accuracy classifier.

Key words:
Unsupervised Learning, K-means Clustering Algorithm, Unlabeled data, Spatial-data, Trajectory.

1. Introduction

Supervised learning classification consists of learning a predictive model using a set of labeled Spatial-data. It is accepted that predictive model accuracy usually increases as more labeled Spatial-data are available. Labeled Spatial-data are generally difficult to obtain as the labeling step is often performed manually. On the contrary, unlabeled Spatial-datasets easily available. As the labeling task is tedious and time-consuming, users generally provide a very limited number of labeled Spatial-datasets. However, designing approaches able to work efficiently with a very large number of unlabeled samples is highly challenging.

In this paper, we present a novel method for converting unlabeled spatial-datasets into labeled spatial-datasets using unsupervised learning, the K-means clustering algorithm. This technique (unsupervised learning using the clustering algorithm) has not been used before to my knowledge. But there are two problems: Firstly, finding the best clusters number (K-value). Secondly, defining and describing the clustering task, which means that the K-means algorithm will divide the spatial-datasets into (for example 3 or 2 Clusters) and thus there will be 3 or 2 class labels, so the challenge how we accurate description for (Class 0, Class 1 and Class 2). To overcome the first problem finding the best clusters number (K-value) we will use the elbow method to identify the elbow point by SSE measurement (Some of Square Error). To overcome the second problem how we defining and describe the clustering task (accurate description for Class labels) we will calculate the center of each cluster (class label) by the centroids algorithm, then draw each cluster with the center point of this cluster, then we carefully study the output of the centroids algorithm, so that the description of each cluster (class label) is assigned according to the highest value of the centroids in each Attribute.

In the experiments section, we will implement the clustering algorithm by building a clustering model using the K-Means algorithm, through which labels are obtained indicating the number of clusters we obtained. Then we will apply the elbow algorithm to make sure that the number of clusters we got is the best and optimal, and this algorithm is summarized in calculating SSE measurement inside a loop of the number of clusters starting with 1 and ending with an optional number that is changed with each experiment, then a graph is made to represent the number of clusters intersecting with His SSE, and looking at the graph it becomes clear that the graph is broken at a specific point called the elbow point at which the number of the clusters are the best and optimal number. Finally, we will implement the centroids algorithm and summarize it calculate the center of each cluster, then graph each cluster with the center point of this cluster, then we carefully study the output of the centroids algorithm, so that the description of each cluster is assigned according to the highest value of the centroids in each Attribute.

In the methodology section, we will discuss the following subsections: Section 3.1 Main Idea Algorithm and its Issues: we will provide a full explanation of Clustering, its benefits, and its uses. Then we will provide a full description of the K-means algorithm, its uses, advantages, some disadvantages, and how to implement and activate it in our experience. Then we provide issues of using K-means algorithm. Section 3.2 Describe the using Techniques: We will provide a full and detailed explanation of the technique we used and how it was modified at each

Manuscript received June 5, 2021
Manuscript revised June 20, 2021
https://doi.org/10.22937/IJCSNS.2021.21.6.27

stage of the experiment, with explaining the two stages. Section 3.3 Dataset and Attribute Information: we will provide a brief description of the dataset used in our experience and how it was obtained, as well as a detailed description of all Attributes, a description, and a datatype of its. section 3.4 Data Quality: we will present a method for verifying whether the data set used in our experiment has any data quality issues such as Missing values, Outliers, Duplicate data, or Wrong data, and what (if any) appropriate strategies are used to deal with any problem. section 3.5 Data Preprocessing: we will explain the importance of applying preprocessing techniques in improving data extraction analysis in terms of time, cost, and quality, such as dimensional reduction, Feature Subset Selection, or Discretization (converting datatype of attribute). Section 4.1 Stage1 (Experiment1): in this stage, we will provide a full and detailed explanation of implementing the clustering model by using the K-Means algorithm, then we apply the elbow algorithm to make sure that the number of clusters we got is the best and optimal, then we apply the centroids algorithm and we are carefully studying the output of this algorithm to describing the clustering task. Section 4.2 Stage2 (Experiment2): depending on the stage1, we reimplementing the clustering model by using the K-Means algorithm but after adjusting of (K-value) to the new value, then we reapply the centroids algorithm and we are carefully studying the output of this algorithm to describing the clustering task by new results. Section 4.3 Results Discussion: Finally, we will present the final outcome and objective of our experiment by identifying describe the clustering task by accurate description for every cluster (Class labels) we obtained.

2. RELATED WORK

Spatial data has always been receiving attention from both academia and industry, Mahmood, A. R., et al. [1], Liu, Y., et al. [2]. Dritsas, E., et al. [3], Wang, M., et al. [4], Alarabi, Louai., et al. [5]. Several research studies investigated the data engineering part of dealing with spatial data, including data management and processing. Yet, most of these systems and frameworks are processing data as they are collected in their raw format form. The main challenge data scientist face is the fact of lacking labeling spatial data. In this study we investigated a real dataset that is collected from NYC Taxi & Limousine Commission trajectories [6]. The dataset contains over a billion of a daily commute in single city of New.

Vittaut JN, et al. [7] In the presence of labeled and unlabeled input, they have implemented a new discriminant algorithm for training classifiers. This algorithm was developed as part of the CEM algorithm architecture and is fairly general in that it can be used for any discriminant classifier. They conducted an experimental evaluation of the proposed approach for text classification and text summarization in terms of the ratio of labeled to unlabeled data in the training collection, and they found that using unlabeled data for supervised learning can also improve classifier accuracy. They also contrasted discriminant and generative semi-supervised learning methods, finding that the former is obviously superior to the latter, especially for small sets. We notice that they used labeled data with Semi-supervised learning, but in our paper, we using unlabeled data with unsupervised learning.

Blum, A., et al. [8] When only a small collection of labeled examples is available, they consider the problem of using a huge unlabeled sample to improve the efficiency of a learning algorithm. They presume that if we had enough labeled data, either view of the example would be sufficient for learning, but our aim is to combine the two views to allow affordable unlabeled data to supplement a much smaller collection of labeled examples. The existence of two distinct views of each case, in particular, implies techniques in which two learning algorithms are trained separately on each view, and then the predictions of each algorithm on new unlabeled examples are combined to expand the training set of the other. The purpose of their paper is to include a PAC-style overview for this situation, as well as a PAC-style system for learning from both labeled and unlabeled data in general. They also present analytical findings based on actual webpage evidence, demonstrating that using unlabeled examples will boost hypotheses significantly in reality. We notice that they used a small collection of labeled data with supervised learning, but in our paper, we using only unlabeled data with unsupervised learning.

De Sa, et al. [9] The final error parameter is usable during training, which is one of the benefits of supervised learning. The algorithm will explicitly reduce the number of misclassifications on the training set for classifiers. Supervisory labels are frequently unavailable or prohibitively expensive when modeling human learning or developing classifiers for autonomous robots. They demonstrate in their paper that they can use structure between pattern distributions of various sensory modalities to substitute for labels. They demonstrate that minimizing the disagreement between the outputs of networks processing patterns from these various modalities is a reasonable approximation to minimizing the number of misclassifications in each modality, and that the findings are comparable. They show that the algorithm performs well in finding suitable placement for the codebook vectors using the Peterson-Barney vowel dataset, particularly when the confusable classes are different for the two modalities. In their paper, they didn’t use the Clustering Algorithm but they used supervised learning. In contrast to it in our paper, we used unsupervised learning by applying the Clustering Algorithm.

Dara, R. et al. [10] Unlabeled data was clustered using a self-organizing map, and potential labeling were inferred from the clusters. When inferred labels are combined with labeled data in a multilayer perceptron, output is better than when only labeled data is used. The results of a variety of common real-world benchmark problems from domains other than text are discussed. Unlabeled data can be used to improve supervised learning in a general-purpose neural network in this way. We notice that they used labeled data with supervised learning, but in our paper, we using unlabeled data with unsupervised learning.

Forestier, G. et al. [11] It is well agreed that the accuracy of predictive models improves as more labelled samples become usable. Labeled samples are difficult to come by since the marking process is always done by hand. Unlabeled samples, on the other hand, are readily available. Since marking is a time-consuming and repetitive process, users usually only have a small number of labeled objects. Designing methods that can perform reliably with a small number of labeled samples, on the other hand, is extremely difficult. Semi-supervised methods have been suggested in this sense to benefit from both labeled and unlabeled results. The emphasis of their paper is on situations where the number of labelled samples is extremely small. They examine and formalize eight semi-supervised learning algorithms, as well as a new approach that combines supervised and unsupervised learning to use both labeled and unlabeled results. Their method produces good results in the experiments, particularly when the number of labeled samples is small. It also proves that mixing classified and unlabeled data for pattern recognition is extremely beneficial. In their paper, they used labeled data with Semi-supervised learning, while being in our paper, we are using unlabeled data with unsupervised learning.

In the previous research referred to in the above LITERATURE REVIEW, the table below (Table 1) shows the difference in the last column of the table between all the previous research study mentioned Including our investigation presented as a summary in this scientific paper.

Table 1: Difference between previous researches and our approach

Research	Method	ML Algorithm	Result	Difference
Vittaut JN, et al. [**7**]	use small number of labeled data with a large number of unlabeled data	A New Discriminant Semi-supervise: Classification Maximum Likelihood (CML)	can create a high-accuracy classifiers	Use labeled data with Semi-supervised learning
Blum, A., et al. [**8**]	use large unlabeled data when only small set of labeled data is available	Co-Training: Hyperlink-Based & Page-Based	High performance of learning algorithm withe decrease Error	Use labeled data with supervised learning
De Sa, et al. [**9**]	substitute labels by making use of structure between the pattern distributions to different sensory modalities	MULTI-MODALITY NN	Minimizing the number of misclassifications in each modality	Didn't use Clustering Algorithm and use supervised learning
Dara, R. et al. [**10**]	Clustering unlabeled data and infer possible labelings from the clusters	A multilayer perceptron (NN)	Performance is improved over using only the labeled data by using supervised learning	Use labeled data with supervised learning
Forestier, G. et al. [**11**]	semi-supervised with both labeled and unlabeled data when the number of labeled data is very limited	Static labeling (SL) Dynamic labeling (DL) Cluster labeling by majority (CLM) (SRIDHCR) (SCEC) Rened clustering (RC) Seeded-Kmeans (SK) Constrained-Kmeans (CK)	Combining labeled and unlabeled data is very useful in pattern recognition.	Use labeled data with Semi-supervised learning

3. METHODOLOGY

3.1 Main Idea and Algorithm and its Issues

Given the repetitive and time-consuming process of labelling spatial-data, there are usually very small amounts of marked spatial-data. However, it is extremely difficult to devise methods that can work effectively with a vast amount of unlabeled spatial details. The K-means clustering algorithm is presented in this paper as a novel approach for transforming unlabeled spatial-datasets into classified spatial-datasets using unsupervised learning.

Cluster Analysis
The Clustering algorithm is a kind of unsupervised learning. Cluster analysis seeks to partition the input data into groups of closely related instances so that instances that belong to the same cluster are more similar to each other than to instances that belong to other clusters.

K-means Clustering algorithm
The k-means clustering algorithm represents each cluster by its corresponding cluster centroid. The algorithm would partition the input data into k disjoint clusters by iteratively applying the following two steps (see Figure 1):

1- Form k clusters by assigning each instance to its nearest centroid.

2- Recompute the centroid of each cluster.

```
Select K points as the initial centroids.
Repeat
    Form K clusters by assigning all points to the closest centroid.
    Recompute the centroid of each cluster.
Until The centroids don't change.
```

Figure 1: The two steps of K-means Clustering algorithm

Issues of using K-means algorithm
Two issues exist: First of all, the best number of clusters (K-value). Secondly, definition and explanation of the cluster task, i.e. that a K-means algorithm divides the spatial-datasets into (3 clusters as cleared in the Experiment section) and hence 3 class labels exist (Class 0, Class 1 and Class 2).

3.2 Describe the using Techniques

To face the first issue: (Eq. 1) shown how we can use the measurement of SSE (Some of Square Error) in the elbow method to identify the elbow point, and then find the best clusters number (K-value).

$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} dist^2(m_i, x)$$

x is a data point in cluster C_i and m_i is the representative point for cluster C_i

Equation 1: The Equation to compute SSE

To face the second issue "describing the clustering task (accurate description for class labels)": we use the centroids algorithm to calculated the center of each cluster, where center of each cluster is calculated using by this algorithm, then a cluster with its central point of that cluster is represented by a suitable graph representing each cluster, as shown in the experiment section below.

3.3 Dataset and Attribute Information

In this section, we perform k-means clustering on NYC Taxi and Limousine Commission dataset. TLC Trip Record Data The yellow and green taxi trip records include fields capturing pick-up and drop-off dates/times, pick-up and drop-off locations, trip distances, itemized fares, rate types, payment types, and driver-reported passenger counts. The data used in the attached datasets were collected and provided to the NYC Taxi and Limousine Commission (TLC) by technology providers authorized under the Taxicab & Livery Passenger Enhancement Programs (TPEP/LPEP). The trip data was not created by the TLC, and TLC makes no representations as to the accuracy of these data. The For-Hire Vehicle ("FHV") trip records include fields capturing the dispatching base license number and the pick-up date, time, and taxi zone location ID (shape file below). These records are generated from the FHV Trip Record submissions made by bases. Note: The TLC publishes base trip record data as submitted by the bases, and we cannot guarantee or confirm their accuracy or completeness. Therefore, this may not represent the total amount of trips dispatched by all TLC-licensed bases. The TLC performs routine reviews of the records and takes enforcement actions when necessary to ensure, to the extent possible, complete and accurate information. We can download dataset from NYC Taxi & Limousine Commission web site [6].

Attribute Information
Attribute 1: VendorID
Attribute 2: pick-up dates/times
Attribute 3: drop-off dates/times
Attribute 4: store_and_fwd_flag
Attribute 5: RatecodeID
Attribute 6: pick-up locations
Attribute 7: drop-off locations
Attribute 8: passenger counts
Attribute 9: trip distances
Attribute 10: itemized fares (fare_amount)
Attribute 11: extra
Attribute 12: mta_tax
Attribute 13: tip_amount
Attribute 14: tolls_amount
Attribute 15: ehail_fee
Attribute 16: improvement_surcharge
Attribute 17: total_amount
Attribute 18: payment types
Attribute 19: trip_type
Attribute 20: congestion_surcharge
There are more details and definitions about Attribute which are not recommended to be presented here [12].

3.4 Data Quality

We must check whether the selected dataset has any data quality issues, and must choose suitable strategies to deal with any issue (if exists).

- Missing Values.
- Outliers.
- Duplicate Data.
- Wrong data.

3.5 Data Preprocessing

The goal for applying Preprocessing techniques is to improve the data mining analysis with respect to time, cost, and quality.

- Dimensionality Reduction.
- Feature Subset Selection.
- Discretization (converting datatype of attribute).

4. EXPERIMENT and RESULTS ANALYSIS

4.1 Stage1 (Experiment1)

First Issue: finding the best clusters number (K-value), we used the elbow method to identify the elbow point by calculate SSE measurement (Some of Square Error).

After implementing the clustering model by using the K-Means algorithm, we got three clusters (K = 3), i.e. three class labels. Now we applied the elbow algorithm to make sure that the number of clusters we got is the best and optimal, and this algorithm is summarized in calculating SSE measurement inside a loop of the number of clusters starting with 1 and ending with 9 (an optional number), then we drew a graph to represent the number of clusters intersecting with his SSE, and looking at the graph it becomes clear that the graph is broken at 2 then at 3 (Figure 2), these points called the elbow point at which the number of the clusters is the best and optimal number.

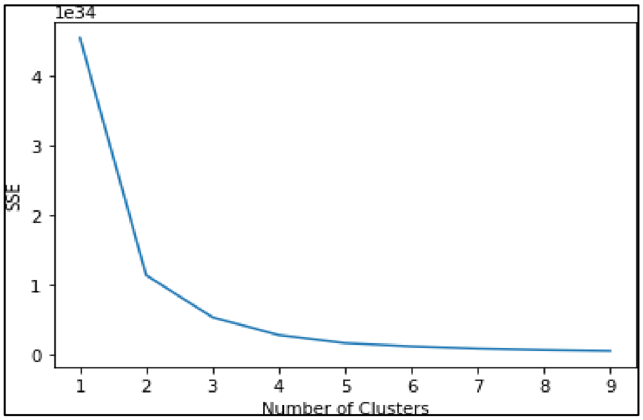


Figure 2: The elbow points

Second Issue: defining and describing the clustering task (accurate description for Class labels), we calculated the center of each cluster (class label) by the centroids algorithm, and this algorithm is summarized in calculating the center of each cluster, then we drew a graph for each cluster with the center point of this cluster (Figure 3).

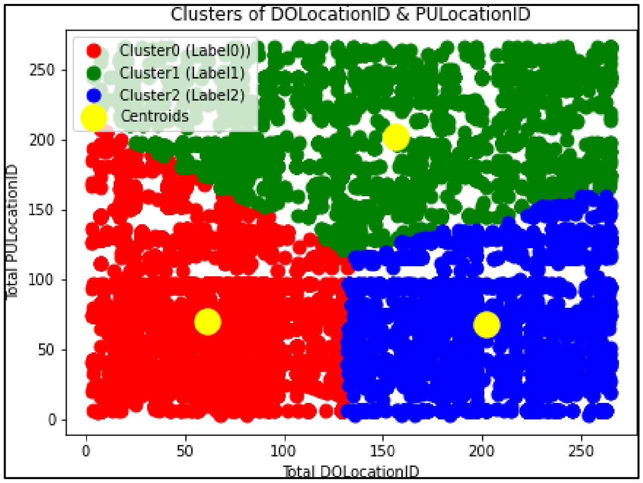
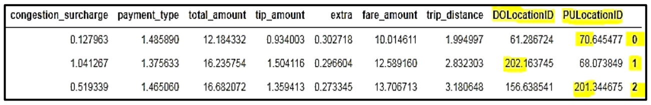


Figure 3: The three clusters with center point of each cluster

By carefully studying the output of the centroids algorithm (Figure 4), so that the description of each cluster is assigned according to the highest value of the centroids in each attribute, we found that the highest centroids value in cluster0 is 70.645477 in the PULocationID attribute, for cluster1 it is 202.163745 in the DOLocationID attribute, and for cluster2 it is 201.344675 in the PULocationID attribute. This means that cluster0, and cluster1 the highest centroids in the attribute PULocationID, meaning that they should have the same cluster, and this will harm us to reapply the algorithm of K-Means clusters again, but at the value of K = 2, i.e. two clusters.



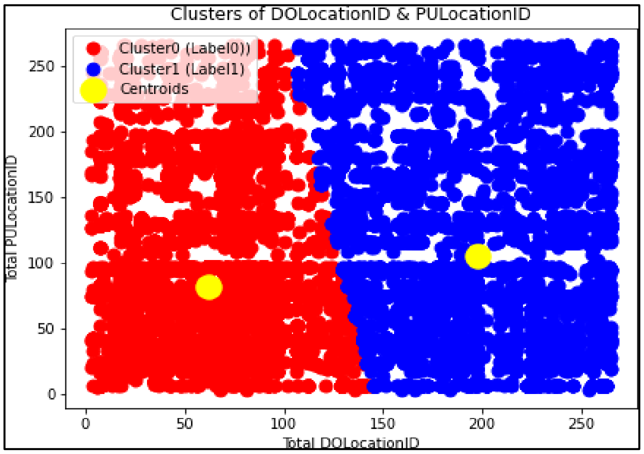
congestion_surcharge	payment_type	total_amount	tip_amount	extra	fare_amount	trip_distance	DOLocationID	PULocationID	
0.127963	1.485890	12.184332	0.934003	0.302718	10.014611	1.994997	61.286724	70.645477	0
1.041267	1.375633	16.235754	1.504116	0.296604	12.589160	2.832303	202.163745	68.073849	1
0.519339	1.465060	16.682072	1.359413	0.273345	13.706713	3.180648	156.638541	201.344675	2

Figure 4: The output of the centroids algorithm: 3 clusters

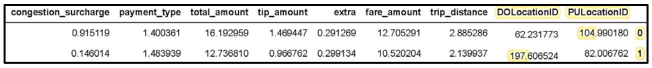
4.2 Stage2 (Experiment2)

Of course, we can choose K = 2 because with reference to (Figure 2) we notice that the elbow point broken is also at point 2, meaning that we can take it as the best value for the number of clusters possible. After implementing the clustering model by using the K-Means algorithm with (K = 2), and calculating the center of each cluster we drew a graph for each cluster with the center point of this cluster (Figure 5).

Figure 5: The two clusters with center point of each cluster



By carefully studying the output of the centroids algorithm (Figure 6), so that the description of each cluster is assigned according to the highest value of the centroids in each attribute, we found that the highest centroids value in cluster0 is 104.990180 in the PULocationID attribute, and for cluster1 it is 197.606524 in the DOLocationID attribute. This means that we have finally reached the desired goal of dividing the dataset, and the best division is on two clusters, as cluster0 (Label 0) indicates the PULocationID (Pick-Up Locations), and cluster1 (Label 1) indicates the DOLocationID (Drop-Ooff Locations). This means that the dataset was done divided into two (2) Clusters based on the TLC-Trip path (Trajectory Spatial-datasets).



congestion_surcharge	payment_type	total_amount	tip_amount	extra	fare_amount	trip_distance	DOLocationID	PULocationID	
0.915119	1.400361	16.192959	1.469447	0.291269	12.705291	2.885286	62.231773	104.990180	0
0.146014	1.483939	12.736810	0.966762	0.299134	10.520204	2.139937	197.606524	82.006762	1

Figure 6: The output of the centroids algorithm: 2 clusters

4.3 Results Discussion

Dividing the dataset into Clusters based on the TLC-Trip path (Trajectory Spatial-datasets). The Experiment showed that all clusters we founded (0 and 1) have a higher centroid points on Attribute (6): pick-up locations, and Attribute (7): drop-off locations. This means that the dataset was done divided into two (2) Clusters based on the TLC-Trip path (Trajectory Spatial-datasets).

5. CONCLUSION and Futurework

Unlabeled Spatial-datasets are easily available, because the labeling task is tedious and time-consuming; users generally provide a very limited number of labeled Spatial-datasets. This is in addition to that labeling a Big-data in many cases is a very costly process. In this paper, we proved by experiment, how using unsupervised learning of K-means clustering algorithm can converting unlabeled spatial-datasets into labeled spatial-datasets. In conclusion, it can be said that our approach appears to have significant merit for converting unlabeled data by unsupervised learning.

In the future work, we will try to use alternative algorithms for K-means such as DBSCAN (Density-based clustering), in order to overcome the problem of determining the number of clusters (K-value), because in the DBSCAN algorithm the determining of (K-value) is not required.

References

[1] Mahmood, A. R., Punni, S., & Aref, W. G. (2019). Spatio-temporal access methods: a survey (2010-2017). GeoInformatica, 23(1), 1-36.

[2] Liu, Y., Singleton, A., Arribas-Bel, D., & Chen, M. (2021). Identifying and understanding road-constrained areas of interest (AOIs) through spatiotemporal taxi GPS data: A case study in New York City. Computers, Environment and Urban Systems, 86, 101592.

[3] Dritsas, E., Kanavos, A., Trigka, M., Vonitsanos, G., Sioutas, S., & Tsakalidis, A. (2020). Trajectory Clustering and k-NN for Robust Privacy Preserving k-NN Query Processing in GeoSpark. Algorithms, 13(8), 182.

[4] Wang, M., Ji, G., Zhao, B., & Tang, M. (2015, October). A parallel clustering algorithm based on grid index for spatio-temporal trajectories. In 2015 Third International Conference on Advanced Cloud and Big Data (pp. 319-326). IEEE.

[5] Alarabi, Louai., Mokbel, M. F., & Musleh, M. (2018). St-hadoop: A Mapreduce Framework for Spatio-Temporal Data. GeoInformatica journal, 22(4), 785-813.

[6] NYC Taxi & Limousine Commission web site. [CrossRef]

[7] Vittaut JN., Amini MR., Gallinari P. (2002) Learning Classification with Both Labeled and Unlabeled Data. In: Elomaa T., Mannila H., Toivonen H. (eds) Machine Learning: ECML 2002. ECML 2002. Lecture Notes in Computer Science, vol 2430. Springer, Berlin, Heidelberg.

[8] Blum, A., & Mitchell, T. (1998, July). Combining labeled and unlabeled data with cotraining. In Proceedings of the eleventh annual conference on Computational learning theory (pp. 92-100).

[9] De Sa, V. R. (1994). Learning classification with unlabeled data. In Advances in neural information processing systems (pp. 112-119).

[10] Dara, R., Kremer, S. C. and Stacey, D. A. (2002) 'Clustering unlabeled data with SOMs improves classification of labeled real-world data', Proceedings of the 2002 International Joint Conference on Neural Networks. IJCNN'02 (Cat. No.02CH37290), Neural Networks, 2002. IJCNN '02. Proceedings of the 2002 International Joint Conference on, Neural networks, IJCNN'02, 3, p. 2237. doi: 10.1109/IJCNN.2002.1007489.

[11] Forestier, G. and Wemmert, C. (2016) 'Semi-supervised learning using multiple clusterings with limited labeled data', Information Sciences, 361–362, pp. 48–65. doi: 10.1016/j.ins.2016.04.040.

[12] More information about the Attribute. [CrossRef]

Fawaz AlBatati Bachelor's degree in Computer Science/Statistics from King Abdulaziz University with a grade of distinction with first class honors, he holds a Nano-degree certificate in data analysis from the Misk Foundation, currently a master's student at Umm Al-Qura University in Computer Science (Artificial Intelligence). His research interests include data science, cybersecurity, and computer network. He is a member of the AI Society, and Saudi Council of Engineers (membership of an engineering specialist). He devises a new indexing system other than the current Dewey Decimal System (a patent for the new indexing system is in progress).

Louai Alarabi received the Ph.D. degrees in computer science and engineering from the University of Minnesota–Twin Cities, MN, USA, in 2018. He is currently an Assistant Professor with the Department of Computer Science, Umm Al-Qura University, Saudi Arabia. His research is published in prestigious research venues. His research interests include database systems, spatial data management, big data management, large-scale data analytics, indexing, main-memory management, and distributed systems. His research recognized by the first place and a gold medal award in student research competition at ACM SIGSPATIAL/GIS 2018, among the Best Paper Award at SSTD 2017, the Finalist of Student Research Competition at ACM SIGMOD 2017, and the Best Demonstration Award at the U-Spatial Symposium in 2014. He has served as a Program Committee Member for ACM SIGSPATIAL/GIS from 2019 to 2021 and the IEEE Big Spatial Data Workshop (BSD), from 2019 to 2021.