A Novel Approach for Knowledge Discovery from AIS Data: An Application for Transit Marine Traffic in the Sea of Marmara

Yunus DOĞAN[1], Özge KART[1], Burak KUNDAKÇI[2] , Selçuk NAS[2]
[1]*Department of Computer Engineering, Dokuz Eylül University, Izmir, Turkey*
[2]*Department of Marine Transportation Education, Dokuz Eylül University, Izmir, Turkey*
yunus@cs.deu.edu.tr

***Abstract*—This paper addresses the discovery of hidden patterns in the data of Automatic Identification Systems by a novel clustering model using data processing and data mining methods. It reveals the transit tracks and the transit vessels on these tracks in the Sea of Marmara which has a dense marine traffic. In this study, improved Density Based Spatial Clustering of Applications with Noise and KMeans++ clustering algorithms have been used together with complex database queries. This proposed approach has been compared to other clustering algorithms such as Self-Organizing Map, Hierarchical Clustering with Single-Link and Genetic Clustering. It has been observed that these alternative algorithms could not reach high accuracy values and they could not give the expected tracks. The proposed approach has five steps and experimental results demonstrate that when this novel approach has been applied step by step, the results can match the observed data by The Republic of Turkey, Ministry of Transport, Maritime and Communications by 95%. Finally, this novel approach is suggested to maritime authorities for all the seas in the world to manage the vessel traffic which has big and complex data.**

***Index Terms*—Clustering algorithms, genetic algorithms, knowledge discovery, machine learning, radar signal processing.**

I. INTRODUCTION

Monitoring of the vessel movements has become very important for maritime situational awareness with the continuous growth in the world maritime transportation [1]. Nowadays, the Automatic Identification System (AIS) has enabled the monitoring of vessel movements with live data. Moreover, according to the International Convention for the Safety of Life at Sea (SOLAS) Chapter V Regulation 19, AIS has been made compulsory gradually for ships [2]. There are two types of AIS on ships as Class A and Class B. While the merchant vessels over 300 GRT and any sized passenger vessels carry Class A, the smaller commercial vessels, fishing vessels, etc. carry Class B [1]. AIS sends the identity, type, position, course, speed, navigational status, and other safety-related information of a ship to the other ships. The ship stations also receive similar information from other ships and shore stations [2]. Current studies show that AIS data and data mining algorithms are used successfully together. Implementation of data mining algorithms such as Density-Based Spatial Clustering of Applications with Noise (DBSCAN) [3], Artificial Neural Network [4], Random Forest [5], Support Vector Machines [6], Window-based encounter clustering [3] and KMeans clustering [7] on AIS data, can be shown as main examples. There are also examples in the literature suggesting techniques for processing AIS data as big data [8-10]. Apart from that, visualization and simulation studies have been also carried out [11, 12]. Finally, examples of decision support systems are also seen in the literature [13, 14]. They handle data mining algorithms or database queries for maritime traffic operations separately. Additionally, these studies do not aim at exploring transit vessel tracks routes. The Sea of Marmara is one of the regions with the most intensive transit ship traffic in the world. This study deals with AIS data on the Marmara Sea. The transit vessel tracks routes are determined by the International Maritime Organization (IMO); whereas, captains can change the routes instantaneously due to the air, the sea, and the other conditions. These movements of ships are detected as outliers. In this study, a novel approach considering these movements has been implemented by means of data mining algorithms.

Some studies used AIS data for mapping the global shipping density [15] and for mapping fishing efforts around European Union [16]. This mapping operation is not an easy process. For example, Wu et al. studied with around 20 billion distinct records, and deriving the density map including data retrieval, computation, and updating of the database process lasts 56 hours [15]. AIS data is not only used for mapping, they are also used for traffic analysis in some areas [17]. For example, Xiao et al. investigate the traffic distribution in narrow and wide waterways [18]; Breithaupt et al. also investigate transverse vessel distribution across the selected routes [19]. Zhang et al. investigate the possible near-miss collisions from AIS data [20] and Li et al. mention that AIS data can be used for avoiding collisions [21]. Especially, the monitoring of big vessels like tankers is very important [22]. Pan et al. propose a novel visualization model for maritime traffic decision making [23] and the management of marine authorities. They applied the proposed model on Xiamen Bay and Meizhou Wan. The AIS data may also use for berthing speed control for large vessels [24]. Arguedas et al. proposed some spatiotemporal data mining methods to improve the Maritime Situational Awareness, handling existing difficulties for example automatic maritime route extraction and synthetic representation, mapping vessel

This work was supported by "Turkey's Directorate General of Coastal Safety" in 2014.

Digital Object Identifier 10.4316/AECE.2021.03009

activities, outlier detection, or position and track prediction. They aimed to give more abilities to operational experts and decision makers to support the decision-making process. The proposed methods were evaluated on various areas from the Dover Strait to the Icelandic coast [25]. In addition, the storage of large volume AIS data is also important. There is another study about this subject, which proposes an infrastructure that handles input rates of about two billion vessel reports per month [26]. In this study, large volume AIS data is filtered and data mining methods are applied to discover motion patterns of transit vessels. The maritime authorities can utilize the results for vessel traffic control of the Marmara Sea.

The contributions of this study are threefold: first, a new filtering approach for AIS data has been proposed using the composition of hybrid clustering algorithms and complex database queries to solve the big data problem in the AIS data warehouse. While KMeans++ is preferred for data cleaning and reduction, the improved DBSCAN algorithm is used to extract the track pattern, including the route information. Complex database queries have been used to manage large data sets. Second, illegal transit vessel tracks have been discovered using AIS data. Maritime authorities can use the discovered tracks for any sea on Earth to detect outliers. Thus, the collusions, pirate, and illegal ships can be detected in real-time. Third, the traditional DBSCAN algorithm has been extended as improved DBSCAN (iDBSCAN) to discover and connect the disconnected tracks which are caused by interruptions of AIS signals.

This paper details our study in four sections. In Section 2, the methodology of the proposed approach is detailed; in Section 3, the experimental studies are explained, and their accuracy results are given; and finally, Section 4 presents conclusions about the proposed methods.

II. METHODOLOGIES

The study has focused on two main problems. The first one is the big data problem of AIS data. Every type of ship sends AIS signals in different time periods. The AIS radar scans its coverage area. It collects each movement as 19 features; Maritime Mobile Service Identities (*MMSI*), Longitude (*Lng*), Latitude (*Lat*), Report Date (*RD*), Speed Over Ground (*SOG*), Course over ground (*COG*), Heading (*HDG*), Rate of Turn (*ROT*), International Maritime Organization (*IMO*), Name (*NM*), Call Sign (*CS*), Ship and Cargo (*SAC*), Draft (*DT*), Type (*T*), Navigational Status (*NS*), Dimension of A part of a ship (*DimA*), Dimension of B part of a ship (*DimB*), Dimension of C part of a ship (*DimC*) and Dimension of D part of a ship (*DimD*). Each signal is a new instance added to the data warehouse. In a little while, lots of instances are collected in the data warehouse. The system records several millions of instances (depending on the intensity of the concerned sea) during a few weeks of observation. This situation causes a big data problem to analyze. For discovering transit tracks, a novel approach containing vertical and horizontal dimension reduction techniques combined with database query operations on AIS data have been implemented in this study.

In an example scenario, an AIS radar is working for a sea where the average of 1000 ships travels. Each ship is sending an instance with 19 features (these features are detailed in Section 3) every 10 seconds on average. It means this AIS data warehouse is collecting 100 instances and 19 x 100 = 1900 feature values. When the observation takes two weeks, it means a big data with 14 x 24 x 6 = 2016 periods, 2016 x 1000 = 2,016,000 instances and 2,016,000 x 19 = 38,304,000 feature values. The second problem is to be able to distinguish transit vessel traffic from other local marine traffic and to discover transit vessel tracks hidden in big AIS data. The AIS data representation in Figure 1 shows all trajectories of all vessels in the Marmara Sea. In this form of the pattern, it can be noticed that the transit vessel tracks are lost in the other traffic.

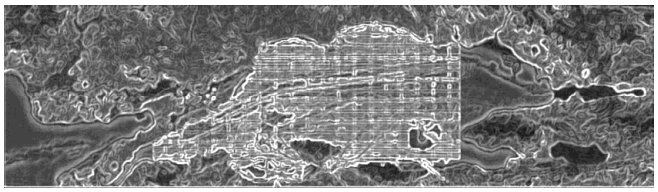

Figure 1. The AIS data pattern in the raw format

AIS uses "Time Division Multiple Access" for sending the data to the respectively [27]. The time slots of the AIS systems are shown in Figure 2. The AIS sends the message in one slot and reserves the next slots. There are approximately 2,250 slots in one minute and for one hour, 135,000 AIS information could be sent by AIS systems in sequence.

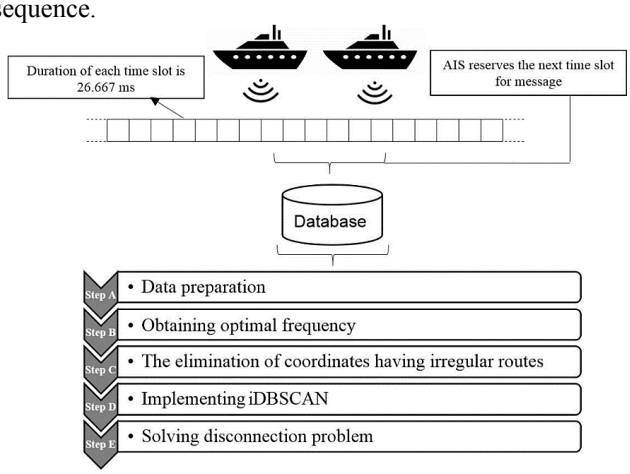

Figure 2. AIS time division multiple access working principle [23]

Shore stations record the AIS information sent from the ships. Historical AIS data becomes useful for analyzing maritime traffic in an area. However, even one-day historical AIS data may be too big to handle for analysis. Vessels send too much data to the other stations every time. For example, if we assume the vessel speed 10 knots and navigating on the same route (no course alteration), this ship sends AIS signal 360 times (with 10 seconds interval) to other stations in one hour. In a heavy traffic area, too many vessels send too much data to other stations [27]. To analyze big data, data mining methods can be useful. For example, clustering algorithms have been used to filter AIS data. Thus, the most preferred routes have been filtered from the whole dataset. KMeans++ algorithm in Table I has been used [28] for implementing Natural Breaks Classification [29], and iDBSCAN algorithm in Table II [30] has been used for clustering the coordinates. COG data have been used in order to specify the current direction of the progress of a ship, between two coordinates, with respect to the

surface. The AIS data filtering consists of two main modules. The first module is data preparation. It includes database management system operations for the knowledge discovery from AIS data. This module has explained in Step A. The second one is knowledge discovery module which includes data mining operations. This module has 4 steps and these steps have been explained below as Step B, Step C, Step D, and Step E.

TABLE I. KMEANS++ ALGORITHM PSEUDO CODE

```
K-MEANS++ ( { x_1, ... , x_N }, K)
//x is the set of coordinates and K is the
number of clusters
Select one coordinate as the candidate centre
uniformly at random among the coordinates.
Repeat
   Foreach (coordinate in x),
      Find the distance D(x) between the
current coordinate and the
      nearest centre that has already been
chosen as the candidate centre.
    Select one new coordinate at random as a
new centre, using a
    weighted probability distribution where a
point x is selected with
    probability proportional to D(x)^2.
Until K centres have been chosen.
(s_1, s_2, ..., s_k) as the centroid set discovered.
For k ← 1 to K
            μ_k ← s_k
End For
For k ← 1 to K
              ∝_k ← {}
             For n ← 1 to N // N is the number
of coordinates
                        j ← argmin_j |μ_j − x_n|
                      ∝_j ← ∝_j ∪ {x_n}
                        For k ← 1  to K
                                μ_k ← (1/|∝_k|) Σ_{x∈∝_k} x
                            End For
                End For
End For
return { μ_1, ..., μ_k }
```

In Step A, the database management system (DBMS) operations have been implemented. Three tables have been created in a database as seen in Figure 3.

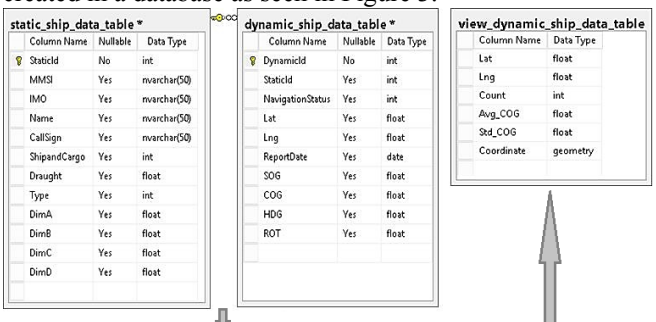

Figure 3. The creation flows of the tables in the database to record and filter the AIS data

The *static_ship_data_table* was created for the static tuples like the features of the ships, and the *dynamic_ship_data_table* was created for the dynamic tuples like the features of each movement. They are two relational tables which store all AIS data. The reason for this separation is to minimize the capacity of AIS data due to the third normal form (3NF) in DBMSs. Thus, AIS data, which is big data, can be managed by reducing to almost half of its size. The other one is a view called *view_dynamic_ship_data_table* which contains some statistical summarized values and the truncated *Lat* and *Lng* values. The data has been transferred from the other two tables by using the queries in Figure 3. At the end of Step A, the coordinates in *view_dynamic_ship_data_table* are printed on the map with various colors according to their frequencies.

In Step B, the points with the high frequencies have been extracted. Sub-steps of Step B can be listed as follows. The frequency values of the coordinates reduced in this step are sorted in an array data structure in descending order; The breakpoint value in the array is calculated by the Jenks Natural Breaks (JNB) method; The coordinates having the frequency under the breakpoint value are eliminated; The remaining coordinates are printed on the map with various colors, which have different meanings, according to their frequencies. A vessel having an unsteady *COG* value is assumed as she has a large standard deviation value among her *COG* values. In other words, this vessel changes her direction irregularly.

TABLE II. IDBSCAN ALGORITHM PSEUDO CODE

```
iDBSCAN(N, minpts, ε) //N is the number of
coordinates
For i←1 to N //The first part
      E = ε_neighborhood set of i^th coordinate
      If  (size(E) < minpts) noise += i
      Else
         new_transit_way += i
         Foreach (coordinate in E)
        E′ = ε_neighborhood set of the current
coordinate
       If  (size(E′) >= minpts) E += the
coordinates in E′
              End If
         End Foreach
     End If
End For
Foreach (transit_way in the pattern) //The
second part
    Foreach (transit_way in the pattern)
        Declare a list named lstCoordinates1 in
the first transit_way;
        Declare a list named lstCoordinates2 in
the second transit_way;
        Declare a counter named lst_index;
        Connect the database; Select the
coordinates from the first
         transit_way order by Lat and Lng;
        Repeat
             If (the Lat of the current
coordinate is equal to the Lat
             of lstCoordinates1[lst_index-1]
and, the longitude of the
             current coordinate is equal to the
Lng of
             lstCoordinates1[lst_index-1]+0.01)
                  Insert the coordinate with
the latitude of
                  lstCoordinates1[lst_index-1]
and
                  the longitude of
```

```
lstCoordinates1[lst_index-1]+0.01 into
                    lstCoordinates1;
               Else
                    Insert the current coordinate
with own Lat and Lng values
                    into lstCoordinates;
               End If
               lst_index=lst_index+1;
          Until all coordinates are evaluated;
     End Foreach
End Foreach
```

In Step C, the view of the coordinates having steady *COG* values have been obtained. The sub-steps are detailed as follows. The standard deviation values of *COG* of the remaining coordinates are sorted in descending order in an array; The breakpoint value in the array is calculated by the JNB method; The coordinates having the standard deviation values of *COG* over the breakpoint value are eliminated; The remaining coordinates are printed on the map.

In Step D, the coordinates revealing the tracks have been detected. The sub-steps are as follows: iDBSCAN clustering has been implemented on the dataset obtained in Step C, with different values of epsilon and *minPoint* parameters. The coordinates in clusters discovered as the tracks are differentiated from outliers using the optimized epsilon ε and *minPoints* parameters by the first part of iDBSCAN algorithm given in Table II; The obtained tracks are printed on the map in various colors.

In Step E, the transit tracks have been formed with coordinates in clusters obtained by Step D. The sub-steps are as follows: A local completion algorithm is applied to complete the disconnections in the pattern after step D. The algorithm is given in the second part of Table II. In the following algorithm, coordinates are selected as ordered by *Lat* and *Lng* for each cluster separately. If there are two coordinates having the differences of latitude and longitude values as higher than 0.01, new interval values are derived and inserted into the table to complete the disconnections in the pattern; the completed version of the tracks is printed on the map with various colors.

III. Experimental Studies and Results

The AIS data used in the study is provided to the trainees for their academic studies by Turkey's Directorate General of Coastal Safety in the IALA Risk Management Toolbox Seminar held by IALA World-Wide Academy on September 8-12, 2014 in Istanbul. Since the data are prepared for academic purposes, the instances in the dataset are a sample in the whole data, and it contains 15-hour parts for each day. For 15 days, 9,509,110 movements for a total of 1,990 ships have been observed and recorded by the AIS.

A. Data Preparation (Step A)

Firstly, AIS data have been recorded in the MSSQL database to manage the big data clearly and for quick filtering operations by the means of the database queries. The features of *Lng*, *Lat*, *RD*, *SOG*, *COG*, *HDG*, and *ROT*, which do not change throughout travel for each ship, were kept in a database table (*dynamic_ship_data_table* in Figure 3). Each row in this table has represented a unique ship. The other features; *MMSI*, *IMO*, *NM*, *CS*, *SAC*, *DT*, *T*, *NS*, *DimA*, *DimB*, *DimC*, and *DimD* have a dynamic structure and the values can change in each movement. Therefore, they have been stored in another table (*static_ship_data_table* in Figure 3). Each row in this table has represented an instantaneous movement for any ship. In this study, the results of the proposed method have been visualized to show the influences of each step by one. Google Maps API has been used for the visualization coding in Microsoft Visual Studio .NET platform.

To capture the density according to the frequency of 9,509,110 movement instances in the table of *dynamic_ship_data_table*, data compression which is one of the data reduction methods has been applied [31]. Thus, the decimal part of the *Lat* and *Lng* values have been reduced from 13 digits to 2 digits. New values have been recorded in a separate table (*view_dynamic_ship_data_table* in Figure 3) in the database. For example, *Lat* and *Lng* values of 40.5733604431152 and 28.077220916748 have been reduced respectively as 40.57 and 28.07. In addition to *Lat* and *Lng* that have been stored as "*geometry :: Point*" data type; it has been aimed to detect the regular (transit) tracks by keeping the standard deviation (*Std_COG*) values and the average (*Avg_COG*) values of the *COG* data in the table of *view_dynamic_ship_data_table* as a comprehensive summary of the movement instances. Thus, a new table named *view_dynamic_ship_data_table*, containing *Lat*, *Lng*, Count, Coordinate, *Std_COG* and *Avg_COG* features, has been obtained. After this dimensionality reduction operation, the number of movement instances decreased from 9,509,110 to 10,020. The data points have been visualized as shown in Figure 4. 10,020 coordinates have been split into 10 slices according to the frequency values to color each slice with a different color. Each slice has contained 1,000 coordinates (1,020 remaining in the last slice). After this colored visualization, two tracks in the darkest colored and high density have appeared in the middle of Marmara Sea.

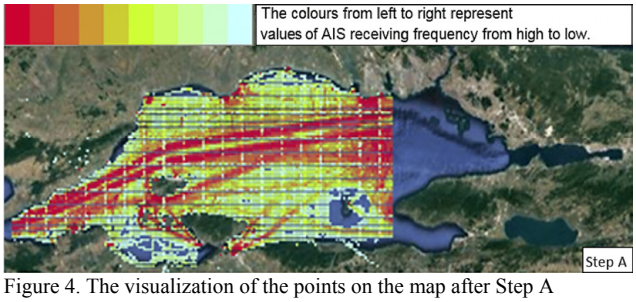

Figure 4. The visualization of the points on the map after Step A

B. Obtaining Optimal Frequency (Step B)

The frequency distributions of 10,020 coordinates are shown in Figure 5. In this figure, a significant breakpoint has been noticed in the frequency distribution. The JNB method has been used to group the 10,020 coordinates into the clusters, the number of which varies between 1 and 7 respectively. Thus, this point has been able to be discovered.

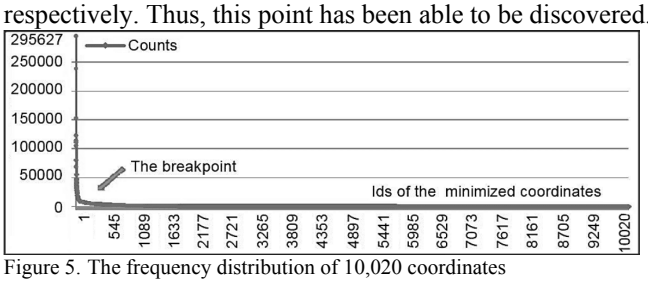

Figure 5. The frequency distribution of 10,020 coordinates

The sum of square error (*SSE*) formulated in (1) has been calculated for each cluster pattern separately as shown in Figure 6.

$$SSE = \sum_{i=1}^{k} \sum_{x \in Si} (x - \mu i)^2 \quad (1)$$

where k is the number of clusters, $k \leq n$, n is the number of coordinates, and $S = \{S1, S2, ..., Sk\}$, μi is the average of the frequencies in Si.

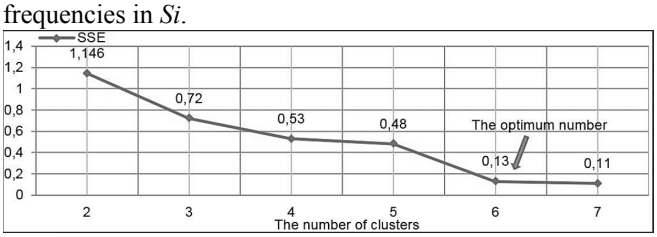

Figure 6. The SSE distribution for the cluster patterns obtained by JNB

The optimum number of clusters has been determined by Elbow method [32,33]. In Figure 6, a significant decrease is observed between the SSE values of 5 clusters and 6 clusters. Since the optimum number of k is obtained as 6, the number of coordinates and the centroids of these 6 clusters given by JNB has been examined in detail. It has been observed that the highest number of coordinates and the lowest average frequency value are in the first cluster. Table III shows that the average frequency value of the 8,826 coordinates in the first cluster is 271.79. It is lower than the values of the other clusters.

TABLE III. THE CENTRE POINTS AND DATA NUMBERS GIVEN BY JNB FOR 6 CLUSTERS

Clusters	The Number of Coordinates	Average frequency of the Clusters
1	8,826	271.79
2	963	3,329.36
3	196	8,803.39
4	27	35,436.92
5	6	114,503
6	2	267,407

Since the average frequency value of this cluster is 271.79; the coordinates with a lower number of the frequencies than 2 x 271.79 ≈ 544 have been eliminated from the dataset. In Figure 7, a colored view of 2,588 coordinates, which have higher frequency values than 544, is given.

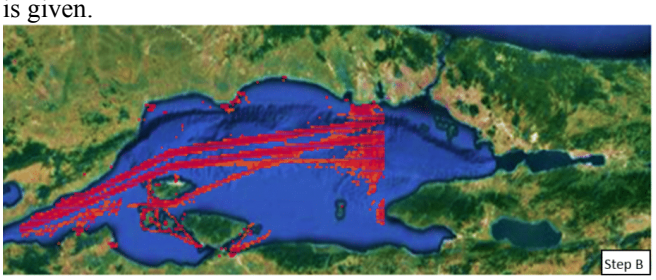

Figure 7. The visualization of the points on the map after Step B

C. The Elimination of Coordinates Having Irregular Routes (Step C)

The next operation is the elimination of coordinates which do not have regular routes, in other words, the elimination of coordinates having high *Std_COG* values. Since the *COG* in the AIS data denotes the angle of the direction of a vessel, the vessels having irregular routes can be discovered by obtaining their *Std_COG* value. The JNB method has been used again for this purpose. Firstly, the *Std_COG* distribution of 2,588 selected coordinates has been obtained as shown in Figure 8.

It is clear that there are two important breakpoints in the *Std_COG* distribution in Figure 8. The second breakpoint , which can be estimated in the range of 50 to 70, has been preferred since expected coordinates have the lowest *Std_COG* values. To be able to discover that point, JNB has been used to group 2,588 values of *Std_COG* into clusters, the number of which varied from 1 to 11.

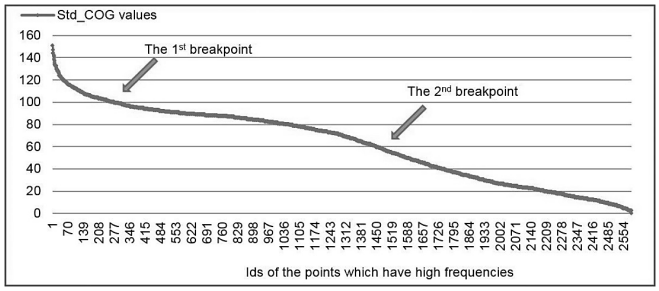

Figure 8. The distribution of Std_COG values

The plot in Figure 9 shows that the optimum number of k is 6.

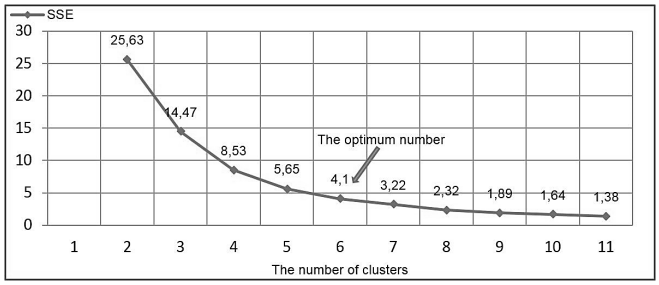

Figure 9. The SSE distribution for the cluster patterns obtained by JNBs

The *Std_COG* value of the second cluster, given in Table IV, has the optimum *Std_COG* value which is expected between 50 and 70. Therefore, since the average *Std_COG* value of this cluster is 29.76, the coordinates, which have higher values of *Std_COG* than 2x29.76 ≈ 60, have been eliminated from the dataset.

TABLE IV. THE CENTRE POINTS AND DATA NUMBERS GIVEN BY JNB FOR 6 CLUSTERS

Clusters	The Number of Coordinates	Average frequency of the Clusters
1	421	12.89
2	440	29.76
3	332	51.72
4	436	74.52
5	720	90.57
6	239	112.77

A colored view of 1,145 coordinates having higher frequency values than 544 and lower *Std_COG* values than 60, is given in Figure 10.

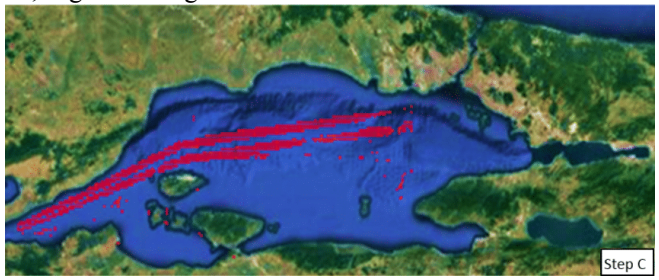

Figure 10. The visualization of the points on the map after Step C

D. Implementing iDBSCAN (Step D)

In this phase, 4 clustering algorithms have been compared to obtain the optimal density for the dataset having 3 features as *Lat*, *Lng* and *Avg_COG*. These algorithms are iDBSCAN proposed in this study, Self-Organizing Map (SOM) used for mapping clustering problems, Hierarchical Clustering with Single-Link (SL-HC) having an available structure to discover the densities and Genetic Clustering (GC) as an optimization clustering algorithm. SOM has been tested with different dimensions of neurons from 2x2 to 10x10. SOM algorithm is a kind of neural network algorithm where each neuron corresponds a cluster [34].

The maps obtained from the implementation of SOM algorithm contain the coordinates of 9, 16 and 100 transit tracks (clusters) respectively. The optimal *SSE* value has been obtained with the map having 3x3 neurons where the most homogeneous distribution of number of coordinates has been observed. The distribution of the coordinates has been shown in Figure 12 (d).

$$f(X,Y)=\sum_{i=1}^{n} X_i Y_i \tag{2}$$

$$f(X,Y)=10n+\sum_{i=1}^{n} X_i Y_i - 10\cos(2\pi X_i) \tag{3}$$

$$path1=-a\times exp\left[-b\left(\sum_{i=1}^{n}\frac{X_i Y_i}{n}\right)^{-\frac{1}{2}}\right] \tag{4}$$

$$path2=-\exp\left(\sum_{i=1}^{n}\frac{cX_i}{n}\right)+a+\exp(1) \tag{5}$$

$$f(X,Y)=path1+path2 \tag{6}$$

Another clustering approach, GC has been tested with 3 different fitness functions as the De Jong's function (2), the Rastrigin's function (3), and the Ackley's path function (6) [35].

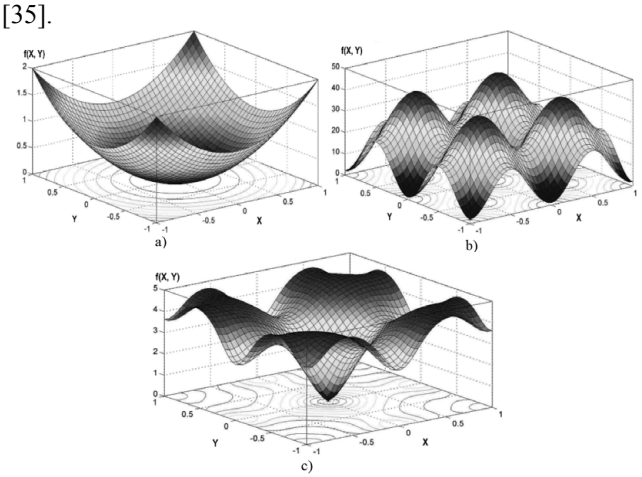

Figure 11. a) De Jong's function, b) Rastrigin's function, c) Ackley's path function

These functions have been represented in Figure 11. De Jong's function obtained 1 cluster with 27.09 *SSE* value (Figure 11(a));Rastrigin's function obtained , 4 clusters with 12.82 *SSE* value (Figure 11(b)); Ackley's path function obtained 7 clusters with 3.07 *SSE* value which is the lowest value (Figure 11(c)). The optimal clustering pattern has been obtained by the Ackley's path function since it supports extreme behaviors and the vessel movements with many variations could be detected. With this approach, 7 transit tracks have been obtained and the distribution of the coordinates has been shown in Figure 12 (b).

Hierarchical clustering with Single Link approach has been preferred for a density-based clustering analysis in the literature. In [36], density detection has been implemented successfully on big data by the means of using SL-HC. In our study, SL-HC has been implemented to obtain the hidden transit tracks, and 8 transit tracks have been detected. The distribution of the coordinates has been shown in Figure 12 (c). In Figure 12, the distributions of coordinates are represented according to the clustering algorithms. In "b" part of the figure, the result of GC algorithm with Ackley's path function has been given. There are 7 transit tracks detected and the densest cluster has 38% of the whole coordinates while the second densest cluster has 17%. It has been expected that the other clusters had to join these 2 big clusters because the expected number of the transit tracks is 2 as the first from the Canakkale Strait to the Istanbul Strait, and the second from the Istanbul Strait to the Canakkale Strait. The same distribution problem has been observed at the usages of SL-HC and SOM. There are 8 transit tracks detected by SL-HC and the densest cluster has 36% of the whole coordinates while the second densest cluster has 21% as shown in the "c" part. Also, there are 9 transit tracks detected by SOM and the densest cluster has 23% of the whole coordinates while the second densest cluster has 15% as shown in the "d" part.

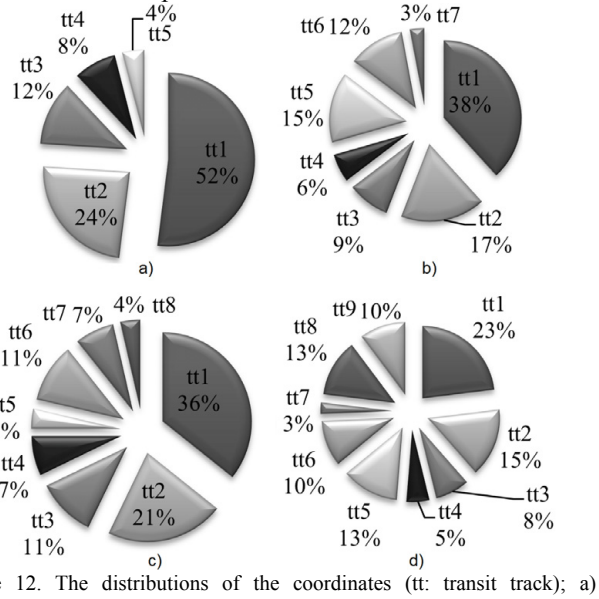

Figure 12. The distributions of the coordinates (tt: transit track); a) iDBSCAN, b) GC, c) SL-HC, d) SOM

In this study, iDBSCAN algorithm given in Table II has been preferred. Because it could eliminate the noise points. The most accurate distribution has been obtained with 5 clusters as shown in Figure 12(a). In the next step, the 4 clusters have been joined by using the second part of iDBSCAN. Thus, 2 transit tracks have been obtained having 52% and 48% distributions of coordinates.

TABLE V. DATA ANALYSIS FOR TWO SETS OF *EPSILONS* AND *MINPOINTS* VALUES WITH IDBSCAN ALGORITHM

	Analysis 1	**Analysis2**	**Analysis3**
E	0.07	0.075	0.08
minPoints	30	30	30
# coordinates in Cluster 1	523	539	539
# coordinates in Cluster 2	543	545	553
# noise coordinates	79	61	53

iDBSCAN analysis has been performed with the different values of ε and *minPoints*, which are the parameters of the DBSCAN algorithm. The *minPoints* value has been fixed to 30. Because, if it becomes lower than 30, the number of clusters increases to 3, and if it becomes higher than 30, the number of clusters decreases to 1. Furthermore, when ε values have been taken as 0.07, 0.075, and 0.08 respectively, it has been observed that the expected 2 clusters and the 60 noise points could be obtained with the ε as 0.075. The *minPoints* are given in Table V. The visualization of the points on the map after the operations of Step D is given in Figure 13.

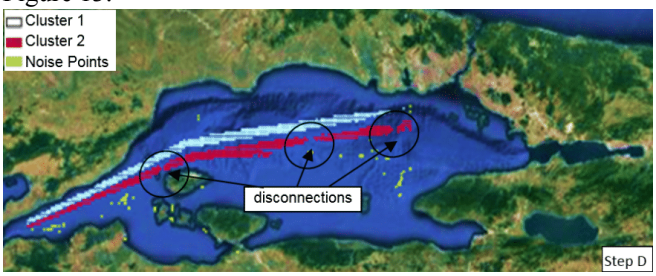

Figure 13. The visualization of the points on the map after Step D

E. Solving Disconnection Problem (Step E)

Some disconnection problems occurred after Step D. Because of the unsteady *COG* values, the points, which must be in the pattern, have been eliminated by the operation in Step C according to the *Std_COG*. Therefore, a local completion operation taking part in the second step of iDBSCAN has been applied to complete connections between disconnected points. After this operation, the count of the points in the first cluster has stabled to 586 (52%) and the count of the points in the second cluster has increased to 540 (48%). In Figure 14, a colored view of 1,126 coordinates is given. These coordinates have corresponded to 3,455,284 movement instances in the pure phase.

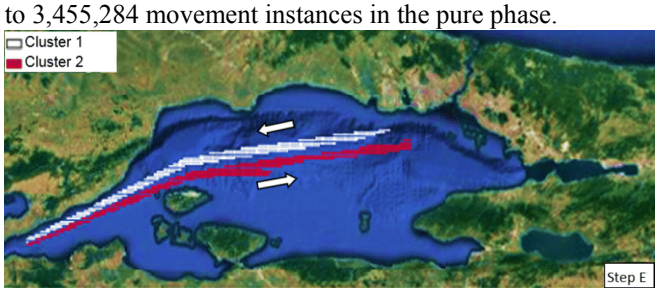

Figure 14. The visualization of the points on the map after Step E

F. Analysis Results

After the operations of five steps, the AIS data has been filtered and the transit tracks have been discovered by the hybrid usage of KMeans++ and iDBSCAN.

TABLE VI. THE COUNTS OF THE COORDINATES IN THE PURE PHASE AND THE REDUCED COORDINATES STEP BY STEP

	The counts of the coordinates in the pure phase	The counts of the reduced coordinates
Beginning	9,509,110	-
After Step A	9,509,110	10,020
After Step B	8,486,359	2,588
After Step C	3,455,284	1,145
After Step D	3,386,433	1,084
After Step E	3,432,334	1,126

Table VI shows that after all operations, 9,509,110 movement instances have decreased to 3,432,334 and 10,020 reduced coordinates have decreased to 1,126. While the decrease for the coordinates is approximately 89%, the decrease for movement instances is approximately 64%. The decreases of these two features have not become proportionally, and it shows that obtained 1,126 coordinates have kept the densest movement points. In the analysis, the efficiency of the obtained two tracks in the marine traffic of the Marmara Sea has been evaluated. These two tracks have had 3,455,284 movement instances and it has been observed that these movement instances have belonged to 1,935 ships. Moreover, it has been observed that 97% of all ships have passed through from at least one of these tracks. Finally, in this study, it has been seen that the number of the ships is approximately equal to the count of transit ships, which had reported by the Republic of Turkey Ministry of Transport, Maritime and Communications. The rate for the official numbers and obtained numbers given in Table VII is 0,588 / 0,617 ≈ 95%. It means that these results can match by 95%.

TABLE VII. OFFICIAL NUMBERS AND OBTAINED NUMBERS

	All ships	Transit ships
Official numbers	3,742	1,990
Discovered numbers	2,312	1,172
Rounded Rates	0,617	0,588

Finally, the types of 1,172 transit ships discovered are 766 cargo ships, 187 tankers, 45 passenger ships and 174 others.

IV. CONCLUSIONS

The main goal of the proposed model using hybrid data mining algorithms is to automatically detect the hidden transit vessel tracks. In this model, two consecutive modules have been proposed. The first module is about the database storage of AIS data and the second module is about the knowledge discovery step using clustering algorithms. iDBSCAN, which is in the clustering part of the proposed approach has been compared to other clustering algorithms such as Self-Organizing Map, Hierarchical Clustering with Single-Link and Genetic Clustering. Experimental results demonstrate that the results obtained by the proposed approach match the observed data by The Republic of Turkey, Ministry of Transport, Maritime and Communications by 95%. The most important advantage of this model is to analyze the AIS data as big data efficiently and accurately. In this study, the model has been implemented and tested on the AIS data with *Lat*, *Lng*, and *COG* parameters for the Sea of Marmara. This model allows the transit vessel traffic data to be filtered out according to the determined parameters within the Marmara Sea. As a result, two central tracks have been obtained as the southwest-northeast line, the route of which is from the Canakkale Strait to the Istanbul Strait and the northeast-southwest line, the route of which is from the Istanbul Strait to the Canakkale Strait. The maritime authorities can use this result for vessel traffic control and for supporting decision- making process. This method can be applied to any area where the AIS data is recorded, and the specific tracks can be discovered practically. Also, the other features of AIS data can be addressed in another study to discover hidden information other than transit ships and their routes, by using other data mining algorithms such as Association

Rule Mining. Furthermore, if a transit ship deviates from the determined tracks and causes suspicion for an accident, that ship can be detected by applying an anomaly detection technique over the obtained patterns, thus can be warned.

ACKNOWLEDGMENT

The AIS data used in the study is provided to the trainees by Turkey's Directorate General of Coastal Safety in the IALA Risk Management Toolbox Seminar held by IALA World-Wide Academy on September 8-12, 2014 in Istanbul. The authors acknowledge the contributions of IALA World-Wide Academy and Turkey's Directorate General of Coastal Safety's to the study and thoroughly thanks the distinguished institutions.

REFERENCES

[1] J. Hoffmann, "Review of maritime transport 2016," in Proc. United Nations Conf. on Trade and Development, Geneva, Switzerland, pp. 1-118, 2016.

[2] I.,I. SOLAS. "International Convention for the Safety of Life at Sea," Consolidated Edition, London, GBP, 2014, pp. 1-10.

[3] P. R. Lei, "Mining maritime traffic conflict trajectories from a massive AIS data," Knowledge and Information Systems, vol. 62, no. 1, pp. 259-285, 2020. doi:10.1007/s10115-019-01355-0

[4] M. Liang, R. W. Liu, Q. Zhong, J. Liu, J. Zhang, "Neural network-based automatic reconstruction of missing vessel trajectory data," in Proc. IEEE 4th International Conf. on Big Data Analytics (ICBDA); Suzhou, China, pp. 426-430, 2019. doi:10.1109/ICBDA.2019.8713215

[5] L. Westerdijk, "Classifying vessel types based on AIS data," MSc, Vrije University, Amsterdam, Holland, pp. 49-61, 2019.

[6] Y. Zhou, W. Daamen, T. Vellinga, S. P. Hoogendoorn, "Ship classification based on ship behavior clustering from AIS data," Ocean Engineering, vol. 175, pp. 176-187, 2019. doi:10.1016/j.oceaneng.2019.02.005

[7] Z. Hanyang, S. Xin, Y. Zhenguo, "Vessel sailing patterns analysis from S-AIS data dased on K-means clustering algorithm," in Proc. IEEE 4th International Conference on Big Data Analytics (ICBDA), Suzhou, China, pp. 10-13, 2019. doi:10.1109/ICBDA.2019.8713231

[8] M. Mustaffa, S. Ahmad, A. M. Ali, N. Ahmad, M. H. Mohd Jais, "Data mining analysis on Ships collision risk and marine traffic characteristic of Port Klang Malaysia waterways from automatic identification system (AIS)," in Proc. International MultiConference of Engineers and Computer Scientists; Hong Kong, pp. 242-246, 2019

[9] D. Yang, L. Wu, S. Wang, H. Jia, K. X. Li, "How big data enriches maritime research–a critical review of automatic identification system (AIS) data applications," Transport Reviews, vol. 39, no. 6, pp. 755-773, 2019. doi:10.1080/01441647.2019.1649315

[10] R. J. Bye, P. G. Almklov, "Normalization of maritime accident data using AIS," Marine Policy, vol. 109, 103675, 2019. doi:10.1016/j.marpol.2019.103675

[11] K. Wang, M. Liang, Y. Li, J. Liu, R. W. Liu, "Maritime traffic data visualization: A brief review," in Proc. IEEE 4th International Conference on Big Data Analytics (ICBDA); Suzhou, China, pp. 67-72, 2019. doi:10.1109/ICBDA.2019.8713227

[12] M. Fujii, H. Hashimoto, Y. Taniguchi, E. Kobayashi, "Statistical validation of a voyage simulation model for ocean-going ships using satellite AIS data," Journal of Marine Science and Technology, vol. 24, no. 4, pp. 1297-1307, 2019. doi:10.1007/s00773-019-00626-3

[13] Y. Liu, R. Song, R. Bucknall, "Intelligent tracking of moving ships in constrained maritime environments using AIS," Cybernetics and Systems, vol. 50, no. 6, pp. 539-555, 2019. doi:10.1080/01969722.2019.1630566

[14] Z. Liu, Z. Wu, Z. Zheng, "A novel framework for regional collision risk identification based on AIS data," Applied Ocean Research, vol. 89, pp. 261-272, 2019. doi:10.1016/j.apor.2019.05.020

[15] L. Wu, Y. Xu, Q. Wang, F. Wang, Z. Xu, "Mapping global shipping density from AIS data," The Journal of Navigation, vol. 70, no. 1, pp. 67-81, 2017. doi:10.1017/S0373463316000345

[16] F. Natale, M. Gibin, A. Alessandrini, M. Vespe, A. Paulrud, "Mapping fishing effort through AIS data," PloS one, vol. 10, no. 6, e0130746, 2015. doi:10.1371/journal.pone.0130746

[17] M. Mustaffa, M. Abas, S. Ahmad, N. Ahmad Aini, W. F. Abbas, S. A. Che Abdullah, M. Y. Darus, "Marine traffic density over Port Klang, Malaysia using statistical analysis Of AIS data: A preliminary study," Journal of ETA Maritime Science, vol. 4, no. 4, pp. 333-341, 2016. doi:10.5505/jems.2016.60352

[18] F. Xiao, H. Ligteringen, C. Van Gulijk, B. Ale, "Comparison study on AIS data of ship traffic behavior," Ocean Engineering; vol. 95, pp. 84-93, 2015. doi:10.1016/j.oceaneng.2014.11.020

[19] S. A. Breithaupt, A. Copping, J. Tagestad, J. Whiting, "Maritime route delineation using AIS data from the Atlantic Coast of the US," The Journal of Navigation, vol. 70, pp. 379-394, 2017. doi:10.1017/S0373463316000606

[20] W. Zhang, F. Goerlandt, J. Montewka, P. Kujala, "A method for detecting possible near miss ship collisions from AIS data," Ocean Engineering, vol. 107, pp. 60-69, 2015. doi:10.1016/j.oceaneng.2015.07.046

[21] J. Li, H. Wang, W. Zhao, Y. Xue, "Ship's trajectory planning based on improved multiobjective algorithm for collision avoidance," Journal of Advanced Transportation, 4068783, 2019. doi:10.1155/2019/4068783

[22] Z. Jiang, D. Chen, Z. Yang, "A synchronous optimization model for multiship shuttle tanker fleet design and scheduling considering hard time window constraint," Journal of Advanced Transportation, 1904340, 2018. doi:10.1155/2018/1904340

[23] P. Jiacai, J. Qingshan, H. Jinxing, S. Zheping, "An AIS data visualization model for assessing maritime traffic situation and its applications," Procedia Engineering, vol. 29, pp. 365-369, 2012. doi:10.1016/j.proeng.2011.12.724

[24] Z. Huang, Z. Shao, J. Pan, X. Ji, Q. Zhao, "Berthing speed control law for large vessels based on AIS data," in Proc. International Conference on Transportation Engineering; Dalian, China, pp. 1322-1330, 2015. doi:10.1061/9780784479384.166

[25] V. F. Arguedas, F. Mazzarella, M. Vespe, "Spatio-temporal data mining for maritime situational awareness," in Proc. OCEANS, Genova, Italy, pp. 1-8, 2015. doi:10.1109/OCEANS-Genova.2015.7271544

[26] A. W. Isenor, M. O. St-Hilaire, S. Webb, M. Mayrand, "MSARI: A database for large volume storage and utilisation of maritime data," The Journal of Navigation, vol. 70, no. 2, pp. 276-290, 2017. doi:10.1017/S0373463316000540

[27] ITU-R. "Technical characteristics for an automatic identification system using time division multiple access in the VHF maritime mobile frequency band," Recommendation ITU-R M.1371-5. Geneva, Switzerland: Electronic Publication, 2014

[28] D. Arthur, S. Vassilvitskii, "K-means++: The advantages of careful seeding," in Proc. SODA'07: 18th Annual ACM-Society for Industrial and Applied Mathematics Symposium on Discrete Algorithms; Philadelphia, USA, pp. 1027–1035, 2007. doi:10.5555/1283383.1283494

[29] G. F. Jenks, F. C. Caspall, "Error on choroplethic maps: definition, measurement, reduction," Annals of the Association of American Geographers, vol. 61, no. 2, pp. 217-244, 1971. doi:10.1111/j.1467-8306.1971.tb00779.x

[30] M. Kantardzic, "Data mining: concepts, models, methods, and algorithms," USA: John Wiley & Sons, pp. 289-296, 2011.

[31] H. Zou, Y. Yu, W. Tang, H. M. Chen, "Improving I/O performance with adaptive data compression for big data applications," In: IEEE International Parallel & Distributed Processing Symposium Workshops; Phoenix, Arizona, USA, pp. 1228-1237, 2014. doi:10.1109/IPDPSW.2014.138

[32] S. H. Jung, K. J. Kim, E. C. Lim, C. B. Sim, "A novel on automatic K value for efficiency improvement of KMeans clustering," in Proc. Advanced Multimedia and Ubiquitous Engineering; Singapore, pp. 181-186, 2017. doi:10.1007/978-981-10-5041-1_31

[33] D. H. Stolfi, E. Alba, X. Yao, "Predicting car park occupancy rates in smart cities," in Proc. International Conference on Smart Cities, Cham, Germany, pp. 107-117, 2017. doi:10.1007/978-3-319-59513-9_11

[34] X. Wang, C. Wang, Z. Chaobiao, "Early warning of debris flow using optimized self-organizing feature mapping network," Water Supply, ws2020142, 2020. doi:10.2166/ws.2020.142

[35] P. Vasant, "Handbook of research on modern optimization algorithms and applications in engineering and economics," USA: IGI Global, pp. 682-690, 2016.

[36] Z. Halim, J. H. Khattak, "Density-based clustering of big probabilistic graphs," Evolving systems, vol. 10, no. 3, pp. 333-350, 2019. doi:10.1007/s12530-018-9223-2