A Distance-Based Trajectory Outlier Detection Method on Maritime Traffic Data

Bao Lei[1]
1 Computer Science Department
Wuhan East Lake College
Wuhan,China
email:blnj2000@nuaa.edu.cn

Du Mingchao[2]
2 College of Electronics Engineering
Navy University of Engineering
Wuhan, China
email:4630598@qq.com

***Abstract*—As a result of establishment of Automatic Identification System(AIS) networks, maritime vessel trajectories are becoming increasingly available.Finding outliers in a collection of patterns among AIS trajectories is critical for real time applications ranging from military surveillance to transportation management. In this paper we present a distance-based approach for trajectory outlier detection on trajectory data.First, we apply Density Based Spatial Clustering of Applications with Noise(DBSCAN) to generate patterns from original trajectories.Second, we extract gravity vectors and sample stop points from clusters to retain stop and move information and to reduce the further computation cost.To measure the similarity between trajectory points and clusters,cluster relative distance and cluster angular distance are proposed. Finally, we present experiments on real AIS data at Chinese QiongZhou strait.The experiment results show that our methods can detect distance anomaly and speed/heading anomaly effectively and greatly reduces computation cost.**

Keywords-AIS; outlier detection; trajectory data mining; cluster

I. INTRODUCTION

The Automatic Identification System(AIS) is a tracking and self-reporting system used by maritime vessels to exchange information with other ships, AIS base stations, and satellites. The International Maritime Organization (IMO) adopted performance standards for AIS and made AIS installation compulsory on all large maritime platforms around the world at 2000, which enable AIS to provide a wealth of valuable surveillance data for vast decision support applications. Consequently, mining on AIS data has become an increasingly important research theme, attracting the attention from numerous areas, including computer science and geography.

Trajectory outlier detection is to detect trajectory outliers (a.k.a. anomalies) that is significantly different from other items in terms of some similarity metric,such as isolated items, devious items, exception items and novelty items. It can also find out events or observations that do not conform to an expected pattern, so as to provide effective support to applications such as urban planning, traffic management, and security controlling.Due to the characteristics of AIS data such as vast,uncertain, incomplete, skew distribution , the traditional methods are facing many challenges and thus somewhat unsuitable. Aiming at solving the problems mentioned above, this paper proposed a trajectory outlier detection method based on clusters distance measure, which can detect distance anomalies,heading anomalies and speed anomalies among AIS trajectories.

II. RELATED WORKS

There are four major categories of trajectory outlier detection methods: prediction-based approach,statistical model approach,grid model and distance-based approach.

The prediction-based approach predict the vessels future status including location, speed and heading, then detect anomalies from the comparison between the real data and predictions.Nevel[1] proposed a Bayesian network approach to detect anomalies within the network context such as changes of destination, inconsistent or unexpected routing.Other approaches generate statistical models of normal trajectories from given data sets, then use these models to measure anomalies.Such as the classifier model ROAM[2], the history trajectory pattern mining model MT-MAD[3],which explore the movement behavior from historical trajectories and build a maritime trajectory model for anomaly detection.There are also some grid models split the map to certain number of buckets, then use some features on each bucket to detect anomalies.Wei[4] partitioned a city into uniform grids and count the number of vehicles arriving in a grid over a time period, then identify contiguous set of cells and time intervals which have the largest statistically significant departure from expected behavior.Other researches present in iBOAT[5] and TOPEYE[6] are also belong to this category.The distance-based methods define anomalies as the trajectories which have relatively long distance with majority trajectories, including Knorr's distance-based trajectory outlier detection[7], Lee's split-detection framework TRAOD[8] and distance calculation approach using R_tree in [9].

Among these approaches, statistical model approaches' performance rely on the pattern mining or statistical model training and usually draw vast calculation on the update of patterns when new data comes.The grid models are not efficient on single trajectory anomaly because they are designed for the events anomalies such as disasters,accidents etc rather than the trajectory itself.The distance-based methods are simple and effective but usually require huge expense on the distance measurement between all the trajectories.

In this paper, we present a distance-based trajectory outlier detection method on AIS data.The method use

978-1-5386-6338-7/18/$31.00 ©2018 IEEE

DBSCAN[10] clustering on historical AIS trajectory point data, produce the normal clusters of moving points and stay points,then extract Gravity Vector(GV) and Sample Stop Point(SSP) from these clusters.The size of GV and SSP is far smaller than the original trajectory data sets, and the expense of distance calculation reduced greatly.

III. CLUSTERING AND GV/SSP EXTRACTION

The purpose of clustering on trajectory data is to reduce the size of original data set and generate the normal pattern of the majority trajectories.Firstly, we present the distance measurement between trajectory points, which is needed for clustering.

The trajectory point is an AIS data item, in this paper we choose 6 attributes of it for further calculation: Maritime Mobile Service Identify(MMSI),Time Stamp(TS),Longitude(LON),Latitude(LAT),Speed Over Ground(SOG) and Course Over Ground(COG).

The geodetic distance between points is the euclidean distance between them.The angular distance between moving points' heading or speed is defined as below.

Definition 1 The Cluster Angular Distance(CAD) is the distance on speed and heading between trajectory points.CAD not only measure the speed difference ,also take the heading difference between moving points or gravity vectors into account.

$$CAD(p,q) = \cos(p.COG - q.COG) \times \frac{p.SOG}{q.SOG} \tag{1}$$

A. Clustering

The DBSCAN algorithm cluster points that are close together(an ε-neighborhood), and surrounded by sufficiently many points. It requires two parameters: a real ***ε,*** and the minimum number of points required to form a cluster MinPts . The ε-neighborhood of a point ***p*** consists of all the points ***q*** that makes distance(p,q)<ε.If the ε-neighborhood of a point ***p*** contains more than MinPts, a new cluster is started,with ***p*** as a core object.

In this paper, we split trajectory points into two kinds according to their speed:moving points and stop points.The moving points clusters represent the sea routes or the often used passages, and the stop points clusters describe the stay area or anchoring area.When determine the ε-neighborhood of stop points, only geodetic distance is taken into account, while the ε-neighborhood of moving points need to use CAD to measure the heading and speed of trajectory points as well.The DBSCAN algorithm is shown in algorithm 1.

B. GV/SSP Extraction

The clusters from DBSCAN algorithm result can represent the normal trajectory pattern, but the amount of trajectory points in the clusters is still huge. To solve this problem,we use certain number of gravity vectors to represent moving points cluster and extract sample points from stop points clusters.

The moving points cluster can be partitioned into some grids, and on each grid, a gravity vector can be extracted to represent the cluster in the distance measurements with other items.

ALGORITHM 1 THE DBSCAN ALGORITHM ON TRAJECTORY POINTS

Algorithm1： DBSCAN(T,**ε**,**Minpts**)

Input： **T**:Trajectory points, **ε**:neighborhood size,**Minpts**:number

Output ： stop points cluster cores **S** and moving points cluster **C**

```
1：For each unvisited point p in T
2：  mark p as visited;
3：  D(p)={};
4：  If p is a stop point
5：   For all stop point q that distance between p,q <ε
6：    generate density reachable set D(p)
7：  else
8：   For all moving point q that angular distance between p,q<ε
9：    generate density reachable set D(p)
10：  If sizeOf(D(p)) < Minpts then
11：   mark p as Noise;
12：  else
13：   create a new cluster from (p,D(p));
14：   Add the core point p of new cluster to S or C;
15：End
```

Definition 2 The gravity vector ***GV*** of a cluster C is a vector with 5 components:average heading,average speed, location of gravity point and inward-distance median.

$$GV_c = \{COG_{avg}, SOG_{avg}, LAT_{avg}, LON_{avg}, D_{median}\} \tag{2}$$

The extraction of gravity vectors from moving points cluster is simple: calculate the average heading of cluster, then partition the whole cluster evenly alongside the average heading, then generate gv on each partition.

The stop points cluster can represent the area where vessels usually stay,like the stay area or anchoring area.To reduce the calculation expense in anomaly detection, we use random sampling on stop points cluster to extract SSP to do further distance calculation instead of the whole cluster.

Definition 3 The Sample Stop Points SSP is a random sampling of a cluster, its number is determined by sampling radius *r* and the area of cluster.

$$SSP_c = random(C,n), n = \overline{area(C)/\pi r^2} \tag{3}$$

To make the sampling well-distributed,only if the new sample point is far enough from all the sample points in SSP, the new point is added to SSP.

IV. ANOMALY DETECTION

Before we can do anomaly detection, one more distance measurement is needed.

Definition 4 The Cluster Relative Distance(CRD) is the relative distance between a moving point and a cluster's GV.Since the moving points cluster usually occupy a certain

area on the map, the CRD is used to measure the distance between a point ***p*** and the cluster ***C*** relative to the size of ***C*** itself.

$$CRD(p,C)=\frac{D(p,(C.LAT_{avg},C.LON_{avg}))}{C.D_{median}} \quad (4)$$

The anomaly detection model is shown in fig.1.The clustering algorithm generate clusters of moving points and stop points from history trajectories.Then we extract the GV and SSP from clusters.We use geodetic distance to measure stop points with SSP, while use CRD and CAD to measure moving points with GV.The anomaly detection rule is straightforward: if the distance between point and cluster is further than threshold, it is an anomaly.The distance threshold is determined by the training data sets, where the threshold can make the majority of trajectory points be marked as normal.Finally,trajectories are evaluated with the percentage of anomalies among them, if the percentage of a certain trajectory is big enough, it will be marked as an abnormal trajectory.

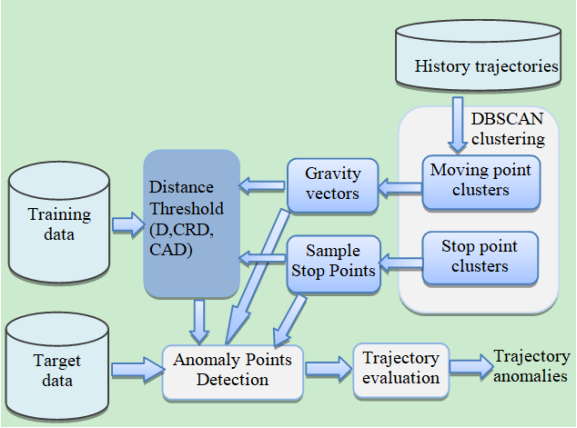

Figure 1. Anomaly detection model based on cluster distance.

The algorithm of anomaly detection is shown in algorithm 2.

V. EXPERIMENTS

A. Data Preparation

The experiment data comes from real AIS data over the Chinese Qiong Zhou strait area on 2012 January , each item in data set is a trajectory point of a certain vessel.The spatial area is latitude 20.0°N to 20.28°N,longitude 109.7°E to 110.4°E, time range from 2012-1-01 to 2012-1-10.

B. Clustering and GV/SSP Extraction

There are more than 50000 data points in Qiong Zhou strait area during 2012-1-01 to 2012-1-05. After data cleaning, we use 30000 points during 2012-1-01 to 2012-1-03 for clustering and get 6 moving points clusters and 1 stop points cluster, which have 13312 moving points and 8021 stop points respectively,shown in fig.2.

After the extraction of gravity vectors and sample stop points, we got 701 GV and 5 SSP,shown in fig.3.The GV and SSP can represent the traffic pattern clearly and its size in compare with original trajectory points data reduced greatly, shown in Table 1.

ALGORITHM 2 THE ANOMALY DETECTION ALGORITHM

Algorithm 2: Anomaly Detection(**D**,**GV**,**SSP**,**Thresholds**)

Input: **D**:target trajectory,**Threshold**: distance thresholds

Output: Trajectory's abnormality

```
1: Split D into moving dataset D_m and stop dataset D_s;
2: Initialize all labels of points in D_m and D_s as Normal;
3: for each data point S in D_s do
4:   geo_D = minimum(D(S, SSP ));
5:   if geo_D > Thresholds.geo then
6:     S.label = Abnormal;
7: for each data point M in D_m do
8:   CRD_D =minimum(CRD(M, GV ))
9:   if CRD_D >Thresholds.crd then
10:    M.label = Abnormal;
11:  else
12:    CAD_D = maximum(CAD(M, GV ))
13:    if CAD_D <Thresholds.cad then
14:      M.label = Abnormal;
15: count_ab =the number of the abnormal points in D;
16: count_all = the total number of all points in D;
17: abnormality = count_ab/count_all;
18: End
```

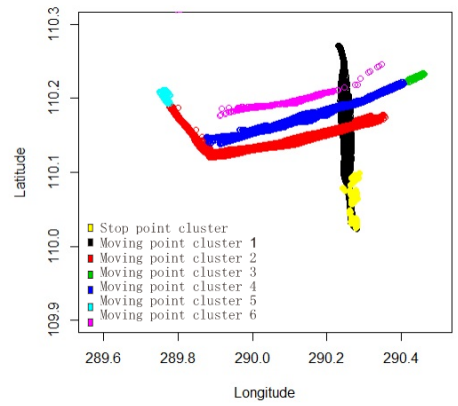

Figure 2. Clustering result.

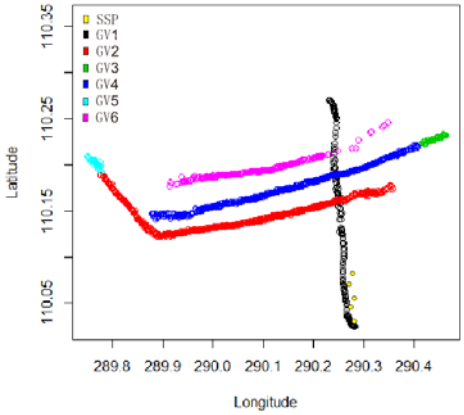

Figure 3. GV/SSP extraction result.

TABLE I. THE SIZE OF DATA SETS

	Original data	After clustering	After extraction
Moving points	20000	13312	701
Stop points	10000	8021	5

C. *Distance Threshold Calculation on Training Set*

The training data set is the AIS data on QiongZhou strait during 2012-1-04 to 2012-1-05.By computing the geodetic distance, CRD and CAD between each trajectory points and GV/SSP, we got the distance distribution.From the distribution, we choose the thresholds to make 95% of the training set have distance below it.

Threshold=(geo,crd, cad)=(169.73,1.243,0.5118).

D. *Anomaly Detection on Target Data*

The target data is the AIS data of Qiong Zhou strait on 2012-1-10. The detection result is shown in figure 4, in 15261 trajectory points, 1482 anomalies are found. Among the anomalies there are 932 distance anomalies(shown in blue points) and 550 speed/heading anomalies(shown in red points).

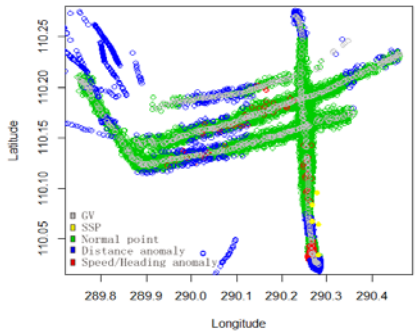

Figure 4. Anomaly detection result.

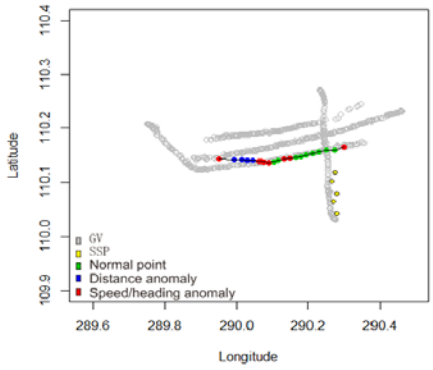

Figure 5. Abnormal trajectory.

E. *Evaluation of Trajectory Anomaly*

There are 15261 trajectory points in target data set, when sorted by MMSI, we get 103 trajectories. Each trajectory consist of certain number of trajectory points. By computing the percentage of anomalies in each trajectory, we can evaluate the trajectory anomalies.The threshold we set is 30%, if more than 30% points among the whole trajectory are anomalies,the trajectory will be marked as abnormal.In figure 5 an off course ship anomaly example is presented ,the trajectory consist of 30 points, in which there are 8 normal points and 22 distance abnormal points.

VI. CONCLUSION

In comparison with other trajectory data, AIS trajectory is more sparse and its distribution usually heavy-skewed. The clustering on AIS data usually get good result.In this paper, we present a straightforward and efficient distant-based anomaly detection approach based on cluster results. It extract GV and SSP from DBSCAN clustering results, reduce the vast amount of trajectory points to a far smaller data set.By measuring the geodetic distance, cluster relative distance and cluster angular distance between target points and GV/SSP, anomalies can be detected, including distance anomalies and speed/heading anomalies. The experiment on real AIS data set shows that this approach is effective.

This work is just a first step,many challenges lies ahead. The data set we used is still not big enough,the target data set only occupy a small spatial area and a short time period.In future works, as the data size increasing, we need to apply certain big data architecture such as hadoop or sparks.The method we proposed here is based on the calculation on AIS location data,and each vessel only have two status, moving or stop.Further research can be made on integrating trajectory semantics with anomaly detection.

REFERENCES

[1] Lane R,David N. Maritime anomaly detection and threaten assessment[J].Information Fusion,2011,18(4):1-8.

[2] Li XL, Han JW, Kim S, Gonzalez H. ROAM: Rule- and motif-based anomaly detection in massive moving object data sets[C]. Proc. of the SDM. Berlin.2007.

[3] Zhu J, Jiang W, Liu A, Liu GF, Zhao L. Time-Dependent popular routes based trajectory outlier detection. Proc. of the WISE.2015. 16−30.

[4] Wei LY, Zheng Y, Peng WC. Constructing popular routes from uncertain trajectories. In: Proc. of the KDD. 2012. 195−203.

[5] Chen C, Zhang DQ, Castro PS, Li N, Sun L, Li SJ. Real-Time detection of anomalous taxi trajectories from GPS traces. In: Proc. Of the MobiQuitous. 2011. 63−74.

[6] Ge Y, Xiong H, Zhou ZH, Ozdemir HT, Yu J, Lee KC. Top-Eye: Top-k evolving trajectory outlier detection. In: Proc. of the CIKM. 2010. 1733−1736.

[7] Knorr EM, Ng RT, Tucakov V. Distance-Based outliers: Algorithms and applications. VLDB Journal, 2000,8(3-4):237−253.

[8] Lee JG, Han JW, Li XL. Trajectory outlier detection: A partition-and-detect framework. In: Proc. of the ICDE. 2008. 140−149.

[9] Liu LX, Qiao SJ, Liu B, Le JJ, Tang CJ. Efficient trajectory outlier detection algorithm based on R-tree. Ruan Jian Xue Bao/Journal of Software, 2009,20(9):2426−2435.

[10] Zhou SG,Zhou AY,Cao J. A DBSCAN Clustering Algorithm based on Data Partitions[J].Journal of Computer Research and Development. 2000,37:1153-1159.