Archives of
Transport System Telematics

Volume 10

Issue 1

February 2017

Clusterization methods in detecting the restricted areas for sea transport

M. DRAMSKI
MARITIME UNIVERSITY OF SZCZECIN, Faculty of Navigation, Wały Chrobrego 1-2, 70-500 Szczecin, Poland
EMAIL: m.dramski@am.szczecin.pl

ABSTRACT
Clusterization is one of the data mining techniques which is responsible for classifying data. Selection of the proper parameters leads to some desired clusters behavior. This fact can be used in detecting the restricted areas for ships and other units. The allowed area can be marked as a data cluster and vice versa. The other advantage is the fact that each cluster consists of the set of points which can be used to find the shortest path in given area. In this paper the use of clusterization in detecting restricted areas is described. Few methods are analyzed and the conclusions presented.

KEYWORDS: data mining, clusterization, restricted area, shortest path

1. Introduction

Each point in the Euclidean metrics has it's coordinates [1]. If two points or more have the same coordinates, they are considered as the same point. From the other side it can be said that these points are different but similar. Human senses are able to orientate in 3 dimensions. In theory even n-dimensional space may be considered where . Two similar points are easy to imagine, but there is a question how to do it in n dimensions. Luckily, mathematics gives such tools. Two or more points are considered as similar, when they lie in the closest space from each other. In other words the metrics describing the similarity is the Euclidean distance:

$$d(x,y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2} \quad (1)$$

This metrics is common used in optimization problems where the distances are compared.

Clusterization is the task of grouping a set of points or objects in such a way that objects in the group are more similar to each other than to those in other groups. These groups are called clusters. In Euclidean metrics the geometric distance is used as a main criterion of grouping.

In this paper the use of clusterization in detecting the restricted areas for sea transport is described. The aim of navigation is to safely conduct the ship from the start point to it's departure. Usually it is solved by creating a graph of possible paths and projecting it on the electronic map. Then the shortest path algorithms such Dijkstra or A* are used. It is not a big deal in wide areas free of any obstacles (moving or stationary). The problem becomes more complicated, when the restricted areas are considered. The restricted area may contain some obstacles like shallows, rocks, reefs etc. natural or artificial origin. It is necessary also to take into consideration that our ship is not the only one moving object in the given area. Some other objects may appear in almost every moment. So, there is a need to update the situation from time to time. This time could be estimated e.g. by decision support system.

2. Clusterization

As mentioned in the introduction of this paper, clusterization is an automatic grouping of similar objects into sets. In this paper two methods were applied: k-means and mean-shift algorithms. In both the metrics is distance between the points. K-means is used the most common used approach, the clusters size is even. Mean-shift lets create more clusters and supports uneven cluster size.

K-means algorithm [2] aims to choose centroids that minimize the inertia or within-cluster sum of squared criterion:

$$\sum_{i=1}^{n} min_{\mu_j \in C} \left(\left\| x_j - \mu_j \right\|^2 \right) \quad (2)$$

This algorithm may have some problems if the incorrect number of clusters is set as the parameter. The distribution of data points is important too. The aim is to obtain the given number of clusters with equal (or almost equal) number of points in them.

In mean-shift algorithm [3] the bandwidth is given as the parameter. The number of clusters is determined during the calculations. Given a candidate for iteration, the candidate is updated according to the following equation:

$$x_i^{t+1} = x_i^t + m\left(x_i^t\right) \quad (3)$$

Where is the neighborhood of samples within a given distance around and the m is the mean-shift vector computed for each centroid according to the equation:

$$m(x_i) = \frac{\sum_{x_j \in N(x_i)} K(x_j - x_i)\, x_j}{\sum_{x_j \in N(x_i)} K(x_j - x_i)} \quad (4)$$

The mean-shift algorithm stops the search when the change of centroids is small.

Using clusterization methods makes possible easy decoding all the points from the digital map. In three-dimensional space there are three variables describing the coordinates.

3. The experiment

3.1. Theoretical description

As mentioned above, each point in three-dimensional space has three coordinates: , and . First two describe the position of the point in two-dimensional projection. In sea transport this is the visible aspect of coordinates. It is impossible to see the third dimension which means e.g. the depth but also can be used to store some other type of information. Sailing is possible if the depth in given area is enough to avoid hitting in the bottom of the sea or other obstacles which are located shallowly. Anyway, there is no need to store the precise information about the depth or other facts. The most important thing for the navigator is if sailing is possible or not. And this is the way of the experiment's construction. Three hypothetical areas were generated (Area 1, Area 2 and Area 3). Then on each area, the random set of points was created. The number of samples was: 100, 1200 and 2500. Each point had three coordinates. The third coordinate stored the information if the given point is available for sailing. Two methods of clusterization were applied, and the final result was to divide the area into two subsets: points allowed and points disallowed. Of course, the area can be divided into more parts if more detailed information is needed. All the experiments showed that both clusterization methods can be used in detecting the restricted areas.

3.2 Practical part

Fig. 1. illustrates the Area 1 and 100 generated samples. Clusterization results are shown in the bottom subplots. Detection of the restricted areas wasn't successful perfectly. This number of samples is not enough for this area size. Fig. 2. confirms this fact in the Area 2 and Fig. 3 in the Area 3.

Clusterization has sense if the total number of samples is ensured. If it's too low, there is a need to get more data.

Fig. 4, 5 and 6 show the experiment results in the case of 1200 random samples. Now it can be seen that due to the increasing the total number of samples, the clusterization is much more clearer. The shape of allowed and restricted areas is visible even without any calculations.

In the case of 2500 samples due to the pages limit of this paper, only the Area 2 is illustrated (Fig. 7).

All experiments were carried out using scikit-learn library for Python programming language.

Both clusterization methods gave almost the same result. In k-means method the target number of clusters was set to 2. Mean-shift algorithm calculated the number of clusters by itself and it was also set to 2.

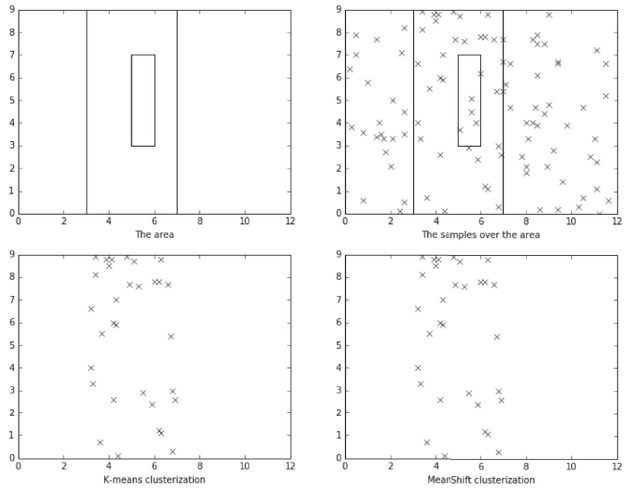

Fig. 1. Area 1, 100 samples [own study]

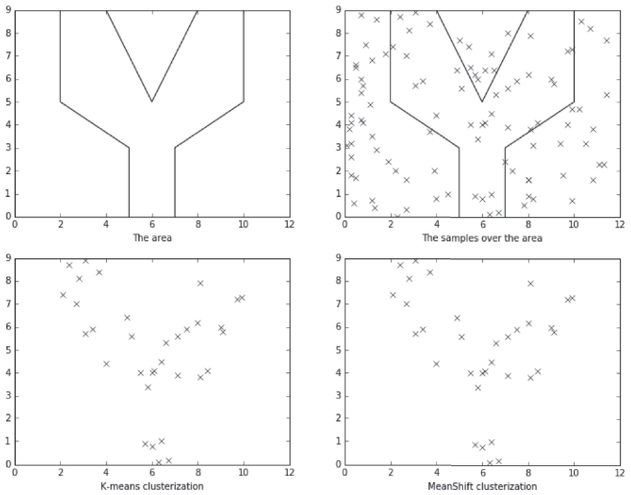

Fig. 2. Area 2, 100 samples [own study]

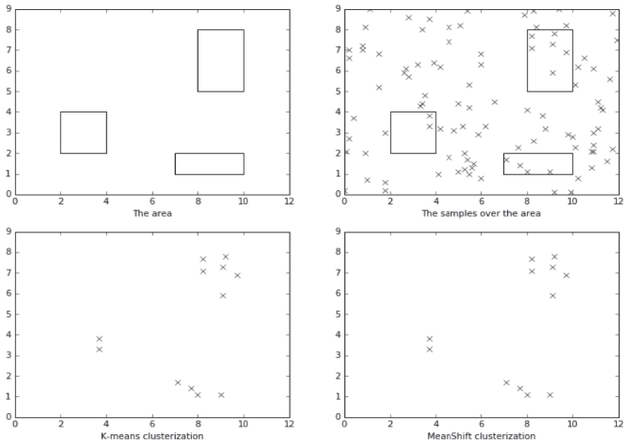

Fig. 3. Area 3, 100 samples [own study]

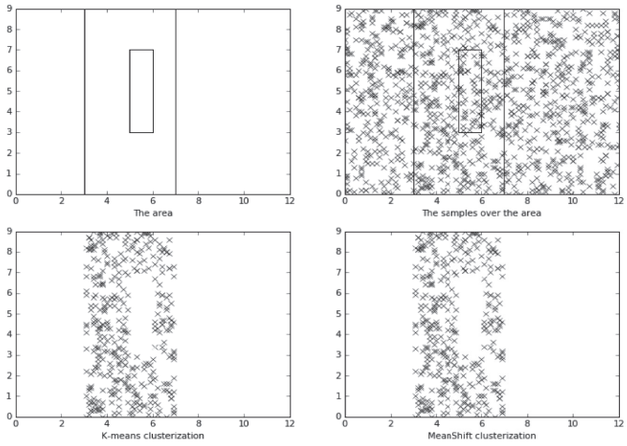

Fig. 4. Area 1, 1200 samples [own study]

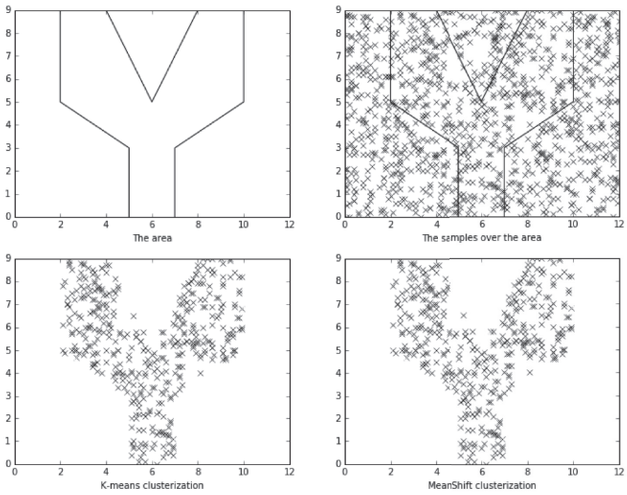

Fig. 5. Area 2, 1200 samples [own study]

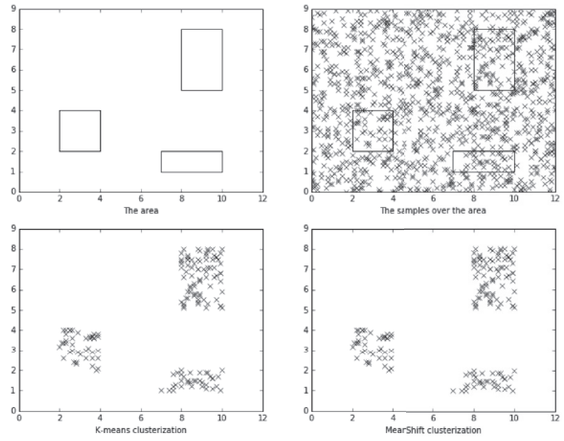

Fig. 6. Area 3, 1200 samples [own study]

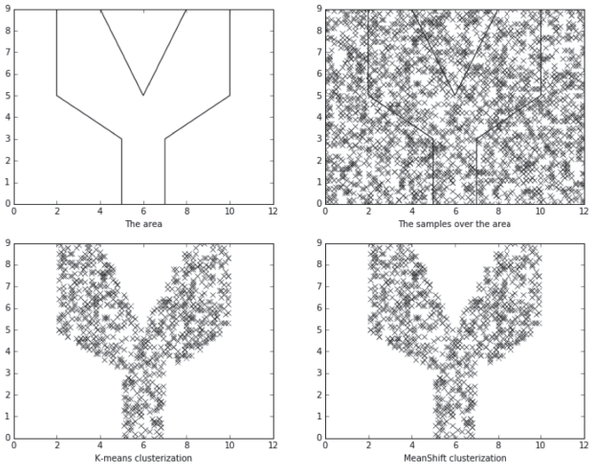

Fig. 7. Area 2, 2500 samples [own study]

The other experiments were carried out too. This time the average time of execution was measured (only the time of clusterization was taken – not time needed for drawing the plots). The results are illustrated in the Table 1.

Table 1. Average execution times [own study]

Number	Area	No. of samples	K-means [ms]	Means-shift [ms]
1	1	100	0.005	0.003
2	2	100	0.006	0.003
3	3	100	0.006	0.003
4	1	1200	0.007	0.011
5	2	1200	0.007	0.010
6	3	1200	0.006	0.010
7	1	2500	0.009	0.024
8	2	2500	0.009	0.023
9	3	2500	0.008	0.022
10	1	5000	0.013	0.048
11	2	5000	0.012	0.046
12	3	5000	0.012	0.045

4. Conclusion

It has been proved that clusterization algorithms can be applied to detection of the restricted areas. Two methods were used and the comparison between them was done. In all the cases the two final clusters were identical and it was consistent with the assumptions.

The number of samples has a big influence on the final result. It is necessary to cover with samples all the most important map's parts to avoid any wrong classification. Moreover, there is a possibility to store a lot of additional data in the coordinates.

Although both clusterization algorithms gave the same observable results, the execution times weren't similar. The higher number of the samples causes the increase of the average execution time, which fact is of course obvious, but k-means algorithm showed less complexity. In the case of 5000 samples was even four times faster than mean-shift. It leads to the conclusion that knowing the final number of clusters would be an advantage.

Clusters centers can be also used to create a graph which would make possible finding the shortest path in restricted area. The other approach could be only connect these centers and find the suboptimal path in this way. Anyway, further research should be done.

Bibliography

[1] DEZA E.: Encyclopedia of distances, Springer, p.94, 2009
[2] HARTIGAN J.A., WONG M.A: A K-means clustering algorithm, Applied Statistics Vol. 28, No. 1, p. 100-108, 1979
[3] CHENG Y.: Mean shift, mode seeking and clustering, IEEE Transactions on Pattern Analysis and Machine Intelligence 17(8) p. 790-799, 1995