Using Artificial Neural Network-Self-Organizing Map for data clustering of marine engine condition monitoring applications

Yiannis Raptodimos* and Iraklis Lazakis

Department of Naval Architecture, Ocean and Marine Engineering, University of Strathclyde, Glasgow, United Kingdom

* Corresponding author details:

Yiannis Raptodimos

Department of Naval Architecture, Ocean and Marine Engineering, University of Strathclyde

Henry Dyer Building, 100 Montrose Street, Glasgow G4 0LZ, Scotland UK

yiannis.raptodimos@strath.ac.uk

Other author details:

Iraklis Lazakis

Department of Naval Architecture, Ocean and Marine Engineering, University of Strathclyde

Henry Dyer Building, 100 Montrose Street, Glasgow G4 0LZ, Scotland UK

iraklis.lazakis@strath.ac.uk

Using Artificial Neural Network-Self-Organizing Map for data clustering of marine engine condition monitoring applications

Condition monitoring is the process of monitoring parameters expressing machinery condition, interpreting them for the identification of change which could be indicative of developing faults. Data pre-processing and post-processing is of great importance in a ship condition monitoring software tool, as misinterpretation of data can significantly affect the accuracy and performance of the predictions made. In this paper, data for key physical performance parameters for a PANAMAX container ship main engine cylinder are pre-processed and clustered using a two-stage approach. Initially, the data is clustered using the ANN (Artificial Neural Network)-Self-Organizing Map (SOM) and then the clusters created by the SOM are interclustered using the Euclidean distance metric into groups. A custom algorithm using a combination of logical operators and conditional statements is used to compare cluster distances and obtain neighbour clusters containing similar data. The case study results demonstrate the capability of the SOM to monitor the main engine condition by identifying clusters containing data which are diverse compared to data representing normal engine operating conditions. The results obtained from the clustering process can be further expanded for application in diagnostic purposes, identifying faults, their causes and effects to the main engine of a ship.

Keywords: ship condition monitoring; data clustering; ANN; self-organizing maps; ship machinery

1. Introduction

The increasing complexity of shipboard systems, heightened expectation and competitive requirements as to ship availability and efficiency and the influence of the data revolution on vessel operations, favour a properly structured Condition Based Maintenance (CBM) regime (Tinsley 2016). British Standard (2012), define CBM as the maintenance policy carried out in response to a significant deterioration in a machine as indicated by a change in a monitored parameter of the machine condition. Jardine et al. (2006) and Kobbacy and Murthy (2008) divide CBM into three main steps: data acquisition, data processing

and maintenance decision making. The heart of CBM is condition monitoring which aims at collecting data regarding equipment condition. Condition monitoring data are measurements related to the health condition of the system such as vibration signals, acoustic signals, temperature, pressure, oil and lubricant measurements amongst other, obtained using various sensors or techniques (Pascual 2015).

Maintenance tasks affect the reliability and availability levels of the shipping industry and are important factors in the lifecycle of a ship while it can minimize down-time and reduce operating costs which accounts for 20%-30% of a ship's operational expenses (Stopford 2009). According to a recent survey by Stephens (2017), vessel operating costs are expected to rise by 2.4% in 2018, with repairs, maintenance and spare parts being the cost categories which are projected to increase most significantly.

Although the maritime industry is responsible for the massive transportation of goods worldwide, it is only recently that new approaches investigating the enhancement of ship's reliability, availability and profitability have been considered (Lazakis and Ölçer 2015). In comparison to other industries, data pooling is not always possible since similar equipment in different conditions may have different failure patterns. Additionally, another issue is the constant appearance of new equipment, making historical records obsolete. Moreover, data is not collected in a standardised way so that it can lead to more informed and effective decision making, while technological advances, overburdened crew and high cost of ownership have resulted in considerable interest in advanced maintenance techniques (INCASS 2015). Moreover, Raza and Liyanage (2009) stated that an increasing demand exists for testing and implementing intelligent techniques as a subsidiary to existing condition monitoring programs. The question of how much data, which data, and how often this should be collected and how has also arisen, as although companies adopt CBM schemes, there seems to be an issue in processing, analysing and

utilising the recorded operational information. Nowadays, data collected by sensors used on ship machinery incorporate an enormous amount of measurement instrumentation, including temperature, pressure, flow, vibration and current sensors. To explore, extract, and generalize inherent patterns in spatiotemporal data sets, clustering algorithms are indispensable (Hagenauer and Helbich 2013).

The goal of clustering is to identify structure in an unlabelled sample or unordered data by objectively organising data into homogeneous groups (Yan 2014). Given a set of data objects, the objective of clustering is to partition them into a certain number of clusters in order to explore the underlying structure and provide useful insight for further analysis. However, there exists no universally agreed-upon and precise definition of the term cluster, partially due to the inherent subjectivity of clustering, which precludes an absolute judgment as to the relative efficacy of all clustering techniques. Data clustering definitions differ among researchers as these are dependent on the desired goal and the data properties. In this respect, Xu and Wunsch (2010) offer various interpretations of data clustering definitions such as: (1) A cluster is a set of data objects that are similar to each other, while data objects in different clusters are different from one another. (2) A cluster is a set of data objects such that the distance between an object in a cluster and the centroid of the cluster is less than the distance between this object and the centroids of any other clusters. (3) A cluster is a set of data objects such that the distance between any two objects in the cluster is less than the distance between any object in the cluster and any object not in it. (4) A cluster is a continuous region of data objects with a relatively high density, which is separated from other such dense regions by low-density regions. Thus, it can be observed that cluster analysis is the formal study of methods and algorithms for grouping or clustering objects according to measured or perceived intrinsic characteristics or similarity.

In this respect, Artificial Neural Network (ANN)-Self-Organizing Maps are employed in order to cluster data by similarity into meaningful classes. A Self-Organizing Map (SOM) is a type of ANN, trained through unsupervised learning for transforming an incoming signal patter of arbitrary dimension into a one- or two-dimensional discrete map, and to perform this transformation adaptively in a topologically ordered fashion. The data consists of key physical performance parameters of a ship main engine such as cylinder exhaust gas temperature, piston cooling oil outlet temperature and piston cool oil inlet pressure. The case study presented in this paper demonstrates the application of the SOM for clustering multidimensional monitored data of key performance parameters for a main engine cylinder. The present paper is organized as follows: Section 2 explains the background of the Self-Organizing Map followed by Section 3 demonstrating the methodology developed. Section 4 presents the case study through which the methodology is applied alongside the obtained results. Finally, discussion regarding the obtained results and concluding remarks are provided in Section 5.

2. Self-Organizing Maps

Namratha and Prajwala (2012) provided an overview of clustering techniques and compares the disadvantages and advantages of each technique. They concluded that each clustering technique depends on the scope of its application and that in order to overcome the disadvantages, optimisation techniques can be used for better performance when required. One of the most popular and simple clustering algorithms, K-means is still widely used today although it was first published over 50 years ago. This illustrates the difficulty in designing a general purpose clustering algorithm and the ill-posed problem of clustering (Jain 2010). Ultsch et al. (1995) demonstrated the capability of the SOM to classify a difficult artificially generated dataset using unsupervised learning, over other well-known statistical clustering methods such as k-means algorithm and hierarchical

clustering requiring previous information on the dataset. Further performance studies have demonstrated the advantages of the self-organizing map over other common clustering approaches. The SOM is a flexible, unsupervised neural network for data analysis and clustering (Hagenauer and Helbich 2013). It maps input data to neurons in such a way that the distance relationships between input signals are mostly preserved (Kohonen 2013). SOM projects input space on prototypes of a low-dimensional regular grid that can be effectively utilised to visualise and explore the properties of the data (Vesanto and Alhoniemi 2000).

Unsupervised learning is a type of machine learning algorithm used to draw inferences from datasets consisting of input data without labelled responses. The most common unsupervised learning method is cluster analysis, which is used for exploratory data analysis to find hidden patterns or grouping in data. In unsupervised learning, there are no predetermined classes, thus no labelled data are available and the goal of clustering is to separate a finite, unlabelled data set into a finite and discrete set of hidden data structures. A direct reason for unsupervised clustering is the need to explore the unknown nature of the data that are integrated with little or no prior information (Pascual 2015).

Self-Organizing Maps are a class of ANN with neurons arranged in a one or two dimensional structure and trained by an iterative unsupervised or self-organising procedure (Yan 2014). They find clusters in the data by evaluating neighbourhood measures and employing competitive strategies. These networks are based on competitive learning. The output neurons of the network compete among themselves to be activated or fired, with the results that only one output neuron, or one neuron per group is on at any one time. An output neuron that wins the competition is called a winner-takes-all neuron or winning neuron. Every data item is mapped into one point (node) in the map and the

distances of the items in the map reflect similarities between the items (Kohonen 1998) . Figure 1 shows the basic structure of the SOM.

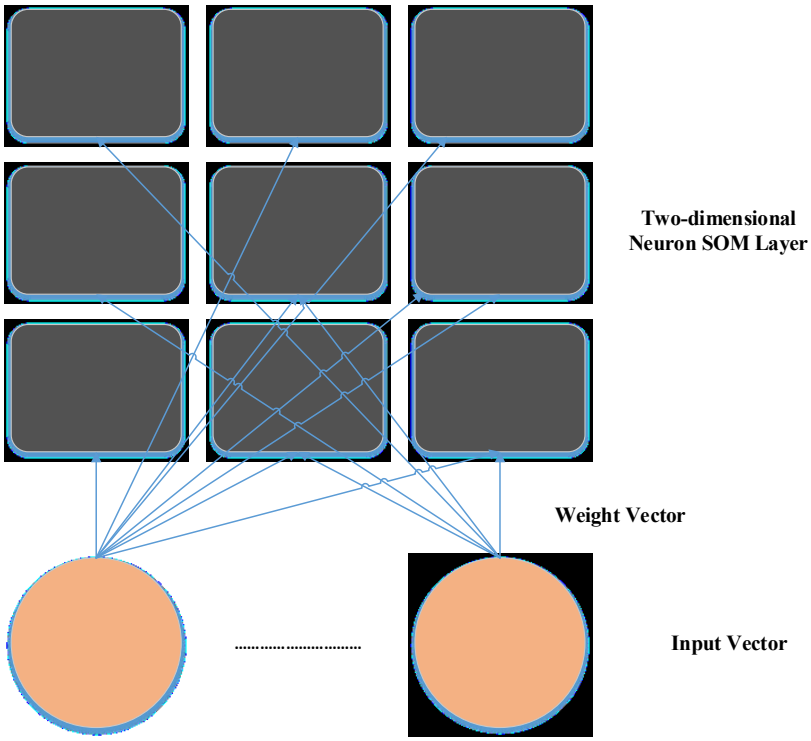

Figure 1 Self-Organizing Map Structure

The SOM consists of a grid of interconnected nodes, where each node is an N-dimensional vector of weights. In general, given a vector as input to the SOM, the node closest to it is found, and then its weights and weights of neighbouring nodes are updated so that they can approach that of the input vector. A self-organizing feature map network identifies a winning neuron using the same procedure as employed by a competitive layer. However, instead of updating only the winning neuron, all neurons within a certain neighbourhood of the winning neuron are updated, using the Kohonen rule as described further on (Curry and Morgan 2004). During training, the SOM forms an elastic net that folds towards the space formed by the input data. Data points lying near each other in the input space are mapped onto nearby map units. Thus, the SOM can be interpreted as a

topology for preserving mapping from input space onto the 2D grid of the map. The SOM is trained iteratively until no noticeable changes in the feature map are observed. At each training step, a sample vector x is randomly chosen from the input data set. There are three basic steps involved in the application of the algorithm after the initialisation stage: sampling, similarity matching and updating. These three steps are repeated until formation of the feature map has been completed based on the input data set. The algorithm is summarized as follows according to Haykin (1998):

(1) Initialisation: Choose random values for the initial weights w_j.
(2) Sampling: Draw a sample x from the input space with a certain probability; the vector x represents the activation pattern that is applied to the lattice.
(3) Similarity Matching: find the best matching (winning) neuron i(x) at time step n by using the minimum distance Euclidean criterion:

$$i(x) = \text{argmin}\left\|x(n) - w_j\right\|, \quad j=1, 2, \ldots, l \quad (1)$$

(4) Updating: adjust the synaptic weight vectors of all neurons by using the update formula:

$$w_j(n+1) = w_j(n) + \acute{\eta}(n)h_{j,i(x)}(n)(x(n) - w_j(n)) \quad (2)$$

Where $\acute{\eta}(n)$is the learning rate parameter and $h_{j,i(x)}(n)$ is the neighborhoood function centered around the winning neuron i(x); both $\acute{\eta}(n)$ and $h_{j,i(x)}(n)$ are varied dynamically during learning rate in order to obtain the best results.

(5) Continuation: Continue step 2 until no noticeable changes in the feature map are observed.

SOM main application areas include statistical methods at large, such as exploratory data analysis, statistical analysis and organisation of texts. Other application areas include industrial analyses, control and telecommunications, biomedical analyses and applications at large and financial applications (Kohonen 2013). Section 3, proposes

the SOM methodology which is a promising application in the maritime industry under the contexts of condition monitoring, diagnostics and maintenance.

3. Methodology

As in the case with other ANNs, the SOM operates in two modes. The first mode is the training phase in which the map is defined and shaped based on the input data, while the second phase automatically classifies new inputs into the clusters defined in the training stage (mapping). The SOM consists of an input and output layer. Inputs in the SOM are the input vector (1-dimensional data) or vectors (multidimensional data) containing data measurements of performance parameters. Additionally, the topology of the SOM defines the number of neurons (clusters) and a distance function in order to obtain distances between the neurons given their positions. The output of the algorithm is the number of neurons the input data has been assigned to. As shown in the methodology flow chart for the data clustering process in Figure 2, the data clustering stage comprises of a two-stage procedure. Initially, the input data is clustered in the SOM to produce the prototype clusters. Afterwards, the prototype clusters obtained from the SOM analysis are *interclustered* to separate or distinguish the prototype clusters into meaningful groups.

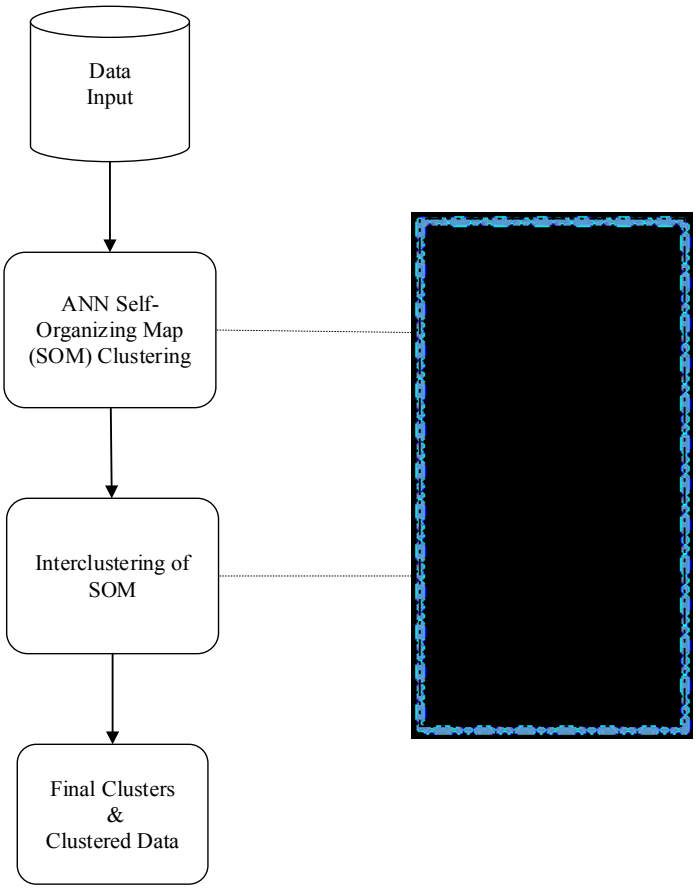

Figure 2 ANN-SOM Clustering Methodology Flow Chart

The neurons in the layer of a SOM are arranged in physical positions originally, according to a topology function, arranging the neurons either in a grid, hexagonal or random topology. The network is trained using the batch unsupervised weight/bias training algorithm. Batch training of a network proceeds by making weight and bias changes based on an entire set (batch) of input vectors. Incremental training changes are applied to the weights and biases of the network after presentation of each individual input vector.

Clustering strategies generally follow two fundamentally different strategies namely hierarchical or agglomerative clustering and point assignment clustering respectively. In hierarchical or agglomerative clustering, clusters can be combined based

on their "closeness", using a distance measure/metric (Vesanto and Alhoniemi 2000). As such, after obtaining the initial clusters from the SOM, the SOM clusters are *interclustered* based on the Euclidean distance metric in order to divide them into groups providing useful insight and information regarding the data. The centre of each SOM cluster is calculated and based on the hierarchical clustering principle, can be used in a Euclidean space for finding similar clusters. A Euclidean space allows the representation of a cluster by its centroid or average of the points in the cluster. Interclustering distances are defined by calculating the Euclidean distance (square roots of the sums of the squares of the differences between the coordinates of the points in each dimension) between the SOM clusters and selecting the clusters with the shortest distance. Cluster centres with small Euclidean distances between them could possibly contain similar data and could be confined under one cluster group. Stopping can be achieved by considering the number of clusters that should be in the data or when the best combination of existing clusters produces a cluster that is inadequate, pre-defined by the user or when the Euclidean distances exceed a threshold (Jung et al. 2003). The Euclidean distance between two points p and q is the length of the line segment connecting them. In Cartesian coordinates, if $p=(p_1,p_2,\ldots,p_n)$ and $q=(q_1,q_2,\ldots,q_n)$ are two points in Euclidean n-space, then the distance d from p to q or vice versa is given by the Pythagorean formula (Pascual 2015): Figure 3 demonstrates the Euclidean distance d_{12} of two points A_1 and A_2.

$$d(p,q) = d(q,p) = \sqrt{(q_1 - p_2)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2} = \sqrt{\sum_{j=1}^{n}(q_i - p_i)^2}$$

(3)

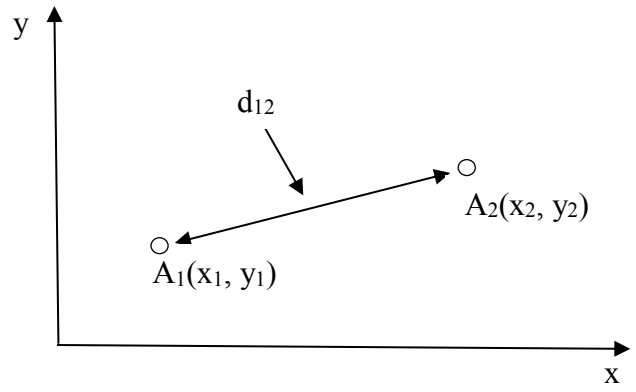

Figure 3 Euclidean distance d_{12} of two points A_1 & A_2

The Euclidean distances of the SOM cluster centres are imported into a custom algorithm using a combination of logical operators and conditional statements and expressions in order to compare cluster distances and search for the clusters that have the shortest distance between them. The algorithm searches for clusters containing similar data, starting from cluster centres that have the closest distance towards larger ones. The algorithm stops searching for close clusters if the distance is larger than 0.4, as in this case the cluster distances are large and in such cases the clusters do not contain similar data. Table 1 demonstrates the custom algorithm criteria used for the interclustering purposes.

Table 1 Criteria for obtaining similar clusters

Cluster Euclidean Distance Metric	Criteria
1st Possible Neighbour Cluster	Distance smaller or equal to 0.1
2nd Possible Neighbour Cluster	Distance smaller or equal to 0.2
3rd Possible Neighbour Cluster	Distance smaller or equal to 0.3
4th Possible Neighbour Cluster	Distance smaller or equal to 0.4

As seen in Table 1, the centres of the prototype clusters created by the SOM are used to calculate the Euclidean distance between each cluster. If the distance between two clusters in smaller or equal to 0.1, then these clusters are very close to each other, compared to clusters with larger distances. Thus, this is the starting point for searching

for similar clusters, starting from the smallest cluster distances towards the largest. If no clusters have a Euclidean distance smaller or equal to 0.1, then the algorithm searches for clusters with Euclidean distances smaller or equal to 0.2 and forth.

4. Case Study & Results

The methodology is applied on a two-stroke marine diesel engine of a PANAMAX container ship of which its main particulars are shown in Table 2. Specifically, data related to various parameters of a main engine cylinder are used for the ANN clustering case study and results. The parameters refer to the exhaust gas temperature outlet, piston cooling oil temperature outlet and piston cooling oil inlet pressure, which are all key performance parameters for examining the condition of the main engine cylinder.

Table 2 Ship Main Particulars

Container Ship Main Particulars	
DWT (Summer)	50829 tons
Length Overall	260.00 m
Beam	32.00 m
Depth Moulded	19.30 m
Draft (Summer)	12.60 m
Engine Particulars	
Main Engine	HSD 8K90MC-C
Maximum Continuous Rating (MCR)	49680 BHP @ 104 RPM
Cylinders	8
Bore	900 mm
Stroke	2300 mm

Data related to these parameters corresponds to a constant vessel speed of 14 knots as the main engine operates at 60 rpm. Specifically, 57 measurements per parameter were used as input for the training of the SOM. With the main engine operating at 60 rpm, the cylinder exhaust gas temperature ranged from 250 to 260 degrees Celsius, the piston cooling oil outlet temperature between 48 and 51 degrees Celsius and the piston cooling

oil inlet pressure readings were constant at 2.7-2.8 kg/cm^2. However, additional artificial data is used for the analysis which correspond to unusual measurements of the monitored parameters, representing abnormal engine behaviour affecting the performance of the main engine. These measurements are compared to those related to the aforementioned engine speed and rpm which represent the normal engine operating condition. As an additional functionality for the developed SOM, data exceeding alarm threshold levels from the Original Equipment Manufacturer (OEM) are also used for the analysis as these thresholds fulfil the requirements of the engine manufacturer and ensure the safe operation of the main engine. Table 3 displays the thresholds utilised to determine abnormal engine operation both for the scenario of the apparently vague parameter measurements and measurements exceeding the engine guide recommended thresholds. The abnormal state thresholds for the monitored parameters were defined based on discussions with experts such as senior and technical marine engineers, two ship operators and three Classification Societies.

Table 3 Alarm thresholds for the main engine monitored parameters

Measurable Parameter	**Normal Range**	**Abnormal State**	**OEM Alarm**
Cylinder exhaust gas temperature outlet ^oC	250-260	Lower than 200, Greater than 300	Greater than 450
Cylinder piston cooling oil temperature outlet ^oC	48-51	Greater than 65	Greater than 70
Piston cooling oil pressure inlet kg/cm^2	2.7-2.8	Lower than 1.8	Lower than 1.4

The SOM created for clustering the multidimensional input vectors consists of a 4-by-4 two-dimensional map of 16 neurons. The SOM topology prior to training the input data is shown in Figure 4.

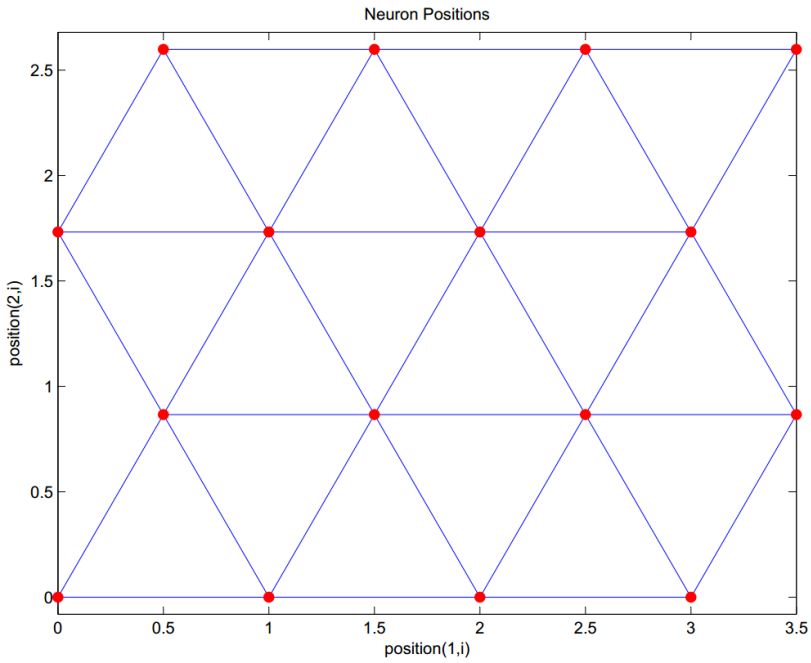

Figure 4 ANN-SOM 4x4 Topology

The SOM is trained for a maximum of 1000 epochs. Prior to training, the collected monitored data are normalized in order to standardise the range of independent variables of data. Once training is complete, the multidimensional input data vectors have been assigned into clusters. The SOM topology after training is shown in Figure 5, in which the green dots represent the input data for the cylinder main engine parameters and the red dots represent the SOM neurons-clusters assigned to the data points while the blue lines connect each node of the map. For example, cluster 15 represents data in which the cylinder piston cooling oil outlet temperature operates in an abnormal state while the piston cooling oil inlet pressure and exhaust gas temperature are operating normally.

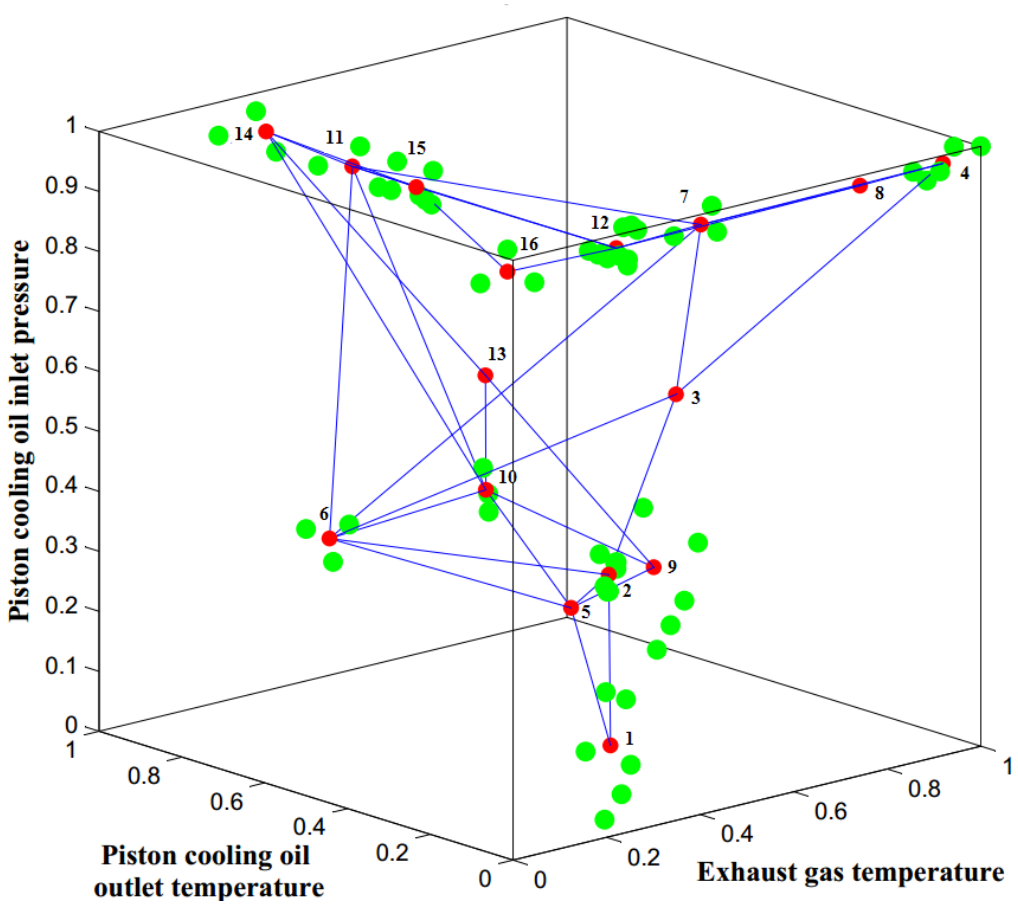

Figure 5 ANN-SOM Clusters

The data has been clustered into twelve clusters, specifically cluster numbers 1, 2, 4, 6, 7, 9, 10, 11, 12, 14, 15 and 16 as observed in Table 4. Each cluster the data has been assigned to has been labelled to provide informative insight to the user such as the ship operator, crew and on-shore operations department, regarding the condition of the monitored parameters and the engine.

Table 4 Description of clusters

Cluster	**Cluster Description**
12	No faults- Normal operating parameter values
16	Cylinder exhaust gas temperature outlet abnormal state, lower than 200 ^{o}C
7	Cylinder exhaust gas temperature outlet abnormal state, greater than 300 ^{o}C
4	Cylinder exhaust gas temperature outlet exceeding OEM alarm level
15	Piston cooling oil outlet temperature abnormal state
11 & 14	Piston cooling oil outlet temperature exceeding OEM alarm level
2	Piston cooling oil inlet pressure abnormal state
1	Piston cooling oil inlet pressure OEM alarm level
6 & 10	All monitored parameters operating in abnormal state
9	All monitored parameters exceeding OEM alarm levels

As observed in Table 4, the clusters produced by the SOM have clustered the multidimensional data related to the cylinder of the main engine and have been interpreted accordingly to provide useful data insight. Specifically, the data have been classified into 10 categories. Cluster 12 represents no faults, in which the monitored parameters are operating under their normal range as defined in Table 3. On the other hand, clusters 16 and 7 represent abnormal data indicating decreased or increased cylinder exhaust gas temperature respectively compared to the normal engine operating values at 60 rpm, while the cylinder piston cooling oil outlet temperature and inlet pressure are operating normally. Cluster 4 contains data related to the exhaust gas temperature exceeding the OEM alarm level. Additionally, data representing all parameters operating simultaneously in abnormal state have been clustered into cluster 6 and 10 and have been assigned under one group based on the interclustering approach. By identifying the cluster centre positions, the Euclidean distance between the clusters can be calculated in order to examine if any clusters can be interclustered based on the concept that clusters with the shortest distances between them share possible similarities in the data. Table 5 displays the Euclidean distances calculated for the sixteen cluster positions. The distance between a point and itself is 0.

Table 5 Cluster no.1-16 Euclidean Distances from their centre

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	0.00	0.28	0.58	1.10	0.38	0.57	0.85	0.99	1.02	0.61	1.04	0.83	0.88	1.19	0.95	0.86
2	0.28	0.00	0.34	0.90	0.38	0.48	0.58	0.77	1.05	0.53	0.83	0.54	0.78	1.01	0.72	0.59
3	0.58	0.34	0.00	0.59	0.44	0.60	0.32	0.45	0.95	0.47	0.65	0.39	0.62	0.84	0.53	0.55
4	1.10	0.90	0.59	0.00	0.93	1.16	0.52	0.17	1.09	0.92	0.95	0.71	0.89	1.10	0.85	0.93
5	0.38	0.38	0.44	0.93	0.00	0.45	0.76	0.84	0.69	0.25	0.79	0.78	0.51	0.91	0.74	0.86
6	0.57	0.48	0.60	1.16	0.45	0.00	0.80	1.02	1.02	0.39	0.61	0.73	0.63	0.72	0.58	0.69
7	0.85	0.58	0.32	0.52	0.76	0.80	0.00	0.35	1.21	0.74	0.65	0.19	0.83	0.86	0.50	0.41
8	0.99	0.77	0.45	0.17	0.84	1.02	0.35	0.00	1.10	0.82	0.82	0.53	0.83	0.99	0.71	0.76
9	1.02	1.05	0.95	1.09	0.69	1.02	1.21	1.10	0.00	0.65	1.10	1.31	0.55	1.10	1.12	1.44
10	0.61	0.53	0.47	0.92	0.25	0.39	0.74	0.82	0.65	0.00	0.59	0.76	0.28	0.67	0.58	0.84
11	1.04	0.83	0.65	0.95	0.79	0.61	0.65	0.82	1.10	0.59	0.00	0.63	0.56	0.22	0.16	0.66
12	0.83	0.54	0.39	0.71	0.78	0.73	0.19	0.53	1.31	0.76	0.63	0.00	0.89	0.85	0.47	0.23
13	0.88	0.78	0.62	0.89	0.51	0.63	0.83	0.83	0.55	0.28	0.56	0.89	0.00	0.57	0.60	0.99
14	1.19	1.01	0.84	1.10	0.91	0.72	0.86	0.99	1.10	0.67	0.22	0.85	0.57	0.00	0.38	0.87
15	0.95	0.72	0.53	0.85	0.74	0.58	0.50	0.71	1.12	0.58	0.16	0.47	0.60	0.38	0.00	0.51
16	0.86	0.59	0.55	0.93	0.86	0.69	0.41	0.76	1.44	0.84	0.66	0.23	0.99	0.87	0.51	0.00

In Table 5, it can be observed that for cluster 1, the distance between itself is zero, while the Euclidean distance between cluster 1 and cluster 2 is equal to 0.28, 0.58 to cluster 3 and so forth. The Euclidean distances are calculated based on the SOM feature map cluster centres as exhibited in Figure 5. The SOM clusters Euclidean distances are imported into the custom algorithm using a combination of logical operators and conditional statements and expressions in order to compare cluster distances and search for the clusters that have the shortest distance between them. Table 6 illustrates the results obtained from the algorithm for clusters 6 and 10 based on the distance criteria defined in Table 1.

Table 6 Similar clusters to cluster 6 and 10 based on Euclidean distance criteria

Euclidean Distance	**<0.1**	**<0.2**	**<0.3**	**<0.4**
Similar Clusters to Cluster 6	No	No	No	10
Similar Clusters to Cluster 10	No	No	5, 13	6

The results illustrate that the first close neighbouring cluster to cluster 6 is cluster 10 which also represents as previously mentioned data representing all parameters operating simultaneously in abnormal state. Additionally, clusters 5, 6 and 13 have been identified as neighbouring clusters related to cluster 10. During the SOM training process no data has been assigned to cluster 5 and 13. This is because the data in the training input vectors have been assigned to the other clusters, based on the SOM training process. On the other hand, cluster 6 is the first neighbouring cluster related to cluster 10 which also represents similar data. Therefore, cluster 6 and 10 can be interclustered and labelled under one group as they both represent similar data. In a similar manner, clusters 11 and 14 are assigned under one group.

After the network training phase, the network is saved to carry out additional simulations using new data as input. In order to validate the network performance in clustering data successfully, new input data is used to simulate the ANN-SOM model. The parameters are normalised for the simulation and the input data and the resulting cluster numbers are shown in Table 7.

Table 7 New Input Data and Assigned Clusters

Parameters	Healthy state			Increased exhaust gas temperature abnormal state			All parameters abnormal state		
Exhaust Gas Temperature	251	253	255	304	310	312	301	305	310
Piston Cooling Oil Outlet Temperature	49	48.5	50	50	48	49	65	66	67
Piston Cooling Oil Inlet Pressure	2.7	2.8	2.7	2.7	2.7	2.7	1.8	1.7	1.7
ANN-SOM Cluster	**12**	**12**	**12**	**7**	**7**	**7**	**10**	**10**	**10**

The results of the clustering process in the table above successfully demonstrate the ability of the trained ANN-SOM to cluster the input data. As observed, the first three sets of data represent healthy data and are assigned to the ANN-SOM cluster 12 which represents healthy data. The next three input vectors represent data for abnormal increased exhaust gas outlet temperature and this has been clustered accordingly in cluster 7. Finally, the last three input data display data in abnormal state compared to the normal engine operating condition for all three monitored parameters and the SOM has classified this data in cluster 10 effectively.

5. Discussion & Conclusions

The trained ANN-SOM can cluster new datasets for the cylinder case study and the interclustering strategy coupled with the custom algorithm can assist in identifying clusters containing similar data, which could be possibly grouped under one category.

The SOM was trained using 16 clusters in two dimensions with some clusters having no data assigned to them as also mentioned previously. This is because the data in the input vectors have been assigned to the other clusters, based on the SOM training process and similarity patterns contained in the data. Moreover, similar data would be assigned in the same cluster and by applying the interclustering technique, neighbour clusters can be assigned under one group when sharing data similarities. Furthermore, this implies that a SOM with different dimensions and thus number of neurons would also cluster the data in a similar manner, demonstrating the flexibility of the SOM for clustering applications and addressing the issue of finding the correct number of clusters which does not have a finite solution. Therefore, the ANN-SOM offers flexibility in terms of assigning the number of clusters to data, as it is an unsupervised learning process and can model the underlying structure or distribution in the data.

The ANN-SOM model monitors the condition of the main engine by clustering parameter measurements under normal conditions and conditions representing faults, which are characterized by parameter measurements exceeding alarm-threshold limits. The practicality of the developed ANN-SOM model can be demonstrated from the fact that the model clusters data both based on apparently vague input data described as abnormal affecting the engine performance and based on OEM alarm thresholds fulfilling the manufacturer requirements and ensuring the safe engine operation. Data are considered abnormal compared to the parameter measurements related to the steady engine rpm and vessel speed representing normal engine operating conditions. Moreover, the abnormal state thresholds for the monitored parameters were defined based on discussions with experts. Overall, the main aim is to obtain clusters displaying data which are diverse compared to those operating under steady ship speed and constant engine rpm. Thus, assuming a certain baseline is provided for the dataset to distinguish between healthy and faulty data, the SOM clustering methodology is applicable. The main practical advantage of the method is that the clustering technique can identify changes in the measurements of parameter while other parameters remain within steady limits.

As big data and the internet of things is becoming a reality and the shipping industry is endeavouring to advance, there is no clear definition of big data and it is still challenging to quantify the volume of data required for successful machine learning, data training and analysis. This paper presented a novel approach for clustering data containing measurements of physical parameters for a ship main engine cylinder using the ANN-SOM. In a similar manner, the approach can also be applied for the entire main engine of a ship. In such a case, it is the authors' opinion that the number of parameters to be used as input depends strongly of the availability of the data. As the number of input parameters increases, the SOM dimensions will have to increase also in order to provide

an adequate number of clusters to properly assign the data. The SOM dimensions and thus number of clusters will also depend on the number of distinguished data groups the user is trying to achieve in compliance with the ship operator, shipowner, technical department requirements. As a demonstrative example, to correctly cluster data for 10 main engine parameters and their fault states, a SOM of approximately 25 neurons or more should be adequate. However, attention should be taken to ensure that the data is correctly assigned to appropriate clusters. If the SOM performance is poor, then the SOM dimensions should be changed until a satisfactory level of clusters and performance is achieved.

The data was successfully assigned to clusters by the Self-Organizing Map. Future steps in this direction include investigating the SOM clustering capabilities with additional condition monitoring data of physical performance parameters of the main engine, training the SOM with larger data sets and validating the results to similar case studies and approaches. Finally, the results obtained from the clustering process can be further expanded for application in diagnostic purposes, identifying engine faults, their causes and their effects to the system.

Acknowledgments

The work in this paper is partially funded by INCASS project. INCASS has received research funding from the European Union's Seventh Framework Programme under grant agreement No. 605200. This publication reflects only the authors' views and European Union is not liable for any use that may be made of the information contained herein.

References

BS. 2012. Condition monitoring and diagnostics of machines. In: BS ISO 13372:2012 London, UK: BSI. p. 28.

Curry B, Morgan PH. 2004. Evaluating Kohonen's learning rule: An approach through genetic algorithms. European Journal of Operational Research.154:191-205.

Hagenauer J, Helbich M. 2013. Hierarchical self-organizing maps for clustering spatiotemporal data. International Journal of Geographical Information Science.27:2026-2042.
Haykin S. 1998. Neural Networks: A Comprehensive Foundation NJ, USA: Prentice Hall PTR.
INCASS. 2015. Deliverable D5.4 'Data exchange' UK: INCASS-Inspection Capabilities for Enhanced Ship Safety.
Jain AK. 2010. Data clustering: 50 years beyond K-means. Pattern Recognition Letters.31:651-666.
Jardine AKS, Lin D, Banjevic D. 2006. A review on machinery diagnostics and prognostics implementing condition-based maintenance. Mechanical Systems and Signal Processing.20:1483-1510.
Jung Y, Park H, Du D-Z, Drake BL. 2003. A Decision Criterion for the Optimal Number of Clusters in Hierarchical Clustering. Journal of Global Optimization. January 01;25:91-111.
Kobbacy KAH, Murthy DP. 2008. Complex system maintenance handbook: Springer Science & Business Media.
Kohonen T. 1998. The self-organizing map. Neurocomputing.21:1-6.
Kohonen T. 2013. Essentials of the self-organizing map. Neural Networks.37:52-65.
Lazakis I, Ölçer A. 2015. Selection of the best maintenance approach in the maritime industry under fuzzy multiple attributive group decision-making environment. Proceedings of the Institution of Mechanical Engineers, Part M: Journal of Engineering for the Maritime Environment.1-13.
Namratha M, Prajwala T. 2012. A comprehensive overview of clustering algorithms in pattern recognition. IOR Journal of Computer Engineering.4.
Pascual DG. 2015. Artificial Intelligence Tools: Decision Support Systems in Condition Monitoring and Diagnosis USA: Crc Press.
Raza J, Liyanage JP. 2009. Application of intelligent technique to identify hidden abnormalities in a system: A case study from oil export pumps from an offshore oil production facility. Journal of Quality in Maintenance Engineering.15:221-235.
Stephens M. 2017. Future operating costs report. London: MS LLP.
Stopford M. 2009. Maritime economics 3e New York, USA: Routledge.
Tinsley D. 2016. Dawning of new era in asset maintenance. Marine Power & Propulsion Supplement 2016 Sect. Section|:Start Page| (col. Column)|.
Ultsch A, Vetter C, Vetter C. 1995. Self-organizing-feature-maps versus statistical clustering methods: a benchmark Marburg, Germany: University of Marburg.
Vesanto J, Alhoniemi E. 2000. Clustering of the self-organizing map. IEEE Transactions on neural networks.11:586-600.
Xu R, Wunsch DC. 2010. Clustering algorithms in biomedical research: a review. IEEE Reviews in Biomedical Engineering.3:120-154.
Yan J. 2014. Machinery prognostics and prognosis oriented maintenance management: John Wiley & Sons.