Expert Systems with Applications 41 (2014) 2742–2753

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Knowledge discovery using genetic algorithm for maritime situational awareness

Chun-Hsien Chen [a,*], Li Pheng Khoo [a], Yih Tng Chong [b], Xiao Feng Yin [c]

[a] School of Mechanical and Aerospace Engineering, Nanyang Technological University, 50 Nanyang Avenue, Singapore 639798, Singapore
[b] Department of Industrial and Systems Engineering, National University of Singapore, Faculty of Engineering, 1 Engineering Drive 2, Singapore 117576, Singapore
[c] Computing Science Department, Institute of High Performance Computing, 1 Fusionopolis Way, #16-16 Connexis, Singapore 138632, Singapore

ARTICLE INFO

Keywords:
Genetic algorithm
Knowledge discovery
Machine learning
Defense
Maritime security
Decision support

ABSTRACT

Due to the large volume of data related to vessels, to manually pore through and to analyze the information in a bid to identify potential maritime threat is tedious, if at all possible. This study aims to enhance maritime situational awareness through the use of computational intelligence techniques in detecting anomalies. A knowledge discovery system based on genetic algorithm termed as GeMASS was proposed and investigated in this research. In the development of GeMASS, a machine learning approach was applied to discover knowledge that is applicable in characterizing maritime security threats. Such knowledge is often implicit in datasets and difficult to discover by human analysts. As the knowledge relevant to maritime security may vary from time to time, GeMASS was specified to learn from streaming data and to generate up-to-date knowledge in a dynamic fashion. Based on the knowledge discovered, the system functions to screen vessels for anomalies in real-time. Traditionally in maritime security studies, datasets that are applied as knowledge sources are related to vessels' geographical and movement information. This study investigated a novel leverage of multiple data sources, including Automatic Identification System, classification societies, and port management and security systems for the enhancement of maritime security. A prototype of GeMASS was developed and employed as a vehicle to study and demonstrate the functions of the proposed methodology.

1. Introduction

Maritime situational awareness, also known as maritime domain awareness, is an area of study that aims to use available data sources to create maximum awareness of activities in the maritime environment. There are currently various types of data that are relevant to maritime security, including Automatic Identification System (AIS) and data of classification societies.

The problem of identifying anomalous vessels involves analyzing multiple sets of data before a decision can be reached. In the case of having human analysts to study sets of information, qualitative reasoning techniques can be applied based on tacit knowledge within the analysts, usually derived from their experiences. In a decision-making process, such tacit knowledge serves to classify ships into different categories. As the information can be large in volume, disparate, and disorganized, it can be challenging for the analysts to pore through datasets, and to generate decisions. This problem is compounded if the time available for decision-making is limited.

Analogous to the tacit knowledge gained by domain experts through experiences, explicit knowledge can be computationally inducted from empirical data. In such approach, machine learning methodologies have been applied to knowledge acquisition. Machine learning techniques function to discover knowledge, for example, via inductive learning (Wong, Ziarko, & Li, 1986). Techniques such as decision tree learning (Quinlan, 1986), neural network learning (Fausett, 1994) and genetic algorithm-based learning (Goldberg, 1989) have also been developed by researchers to extract knowledge from historical datasets.

The general direction of this study is to enhance maritime security through a computational approach. It is known that intelligence that is applicable in detecting anomalous vessels potentially resides in data. Methodologies for extracting knowledge from multiple sets of maritime-related data are therefore studied. The next section briefly presents the state-of-the-art research in this domain.

2. A review of anomaly detection in maritime security

This section presents a review of decision support and anomaly detection systems in the field of maritime security. Fig. 1 shows an overview of the literature reviewed. As depicted in Fig. 1, there

* Corresponding author. Address: School of Mechanical & Aerospace Engineering, Nanyang Technological University, North Spine (N3), Level 2, 50 Nanyang Avenue, Singapore 639798, Singapore. Tel.: +65 6790 4888; fax: +65 6792 4062.
E-mail address: mchchen@ntu.edu.sg (C.-H. Chen).

0957-4174/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2013.09.042

have been mainly two approaches of anomaly detection in the maritime security domain – signature-based and norm-based.

2.1. Signature-based anomaly detection (expert-knowledge driven)

Signature-based methodologies rely on experts' inputs for knowledge on anomalous behaviors and characteristics. Typically, workshops with domain experts are conducted for knowledge elicitation (Roy, 2008; van Laere & Nilsson, 2009). Knowledge elicited includes categories and specific manifestations of a vessel that would constitute an anomaly or a security threat. For instances, frequent change of flags, the allegiance between vessel owners with known terrorist/criminal organizations, loitering maneuvers, cargo types that do not match the port of call, and the shutting down of the AIS.

An anomaly detection system known as Collaborative Knowledge Exploitation Framework (CKEF) was developed to improve maritime domain awareness (Roy, 2008). The rule-based system uses experts' concepts of anomalies to derive signatures as knowledge for applications. More recently an overall framework known as Rule-Based Expert System (RBES) with rules defined by experts and encoded by knowledge engineers was proposed (Roy, 2010). In another example of a rule-based expert system, basic spatial and kinematical relations between objects for the deduction of different situations, e.g. smuggling, hijacking and piloting was developed (Edlund, Grönkvist, Lingvall, & Sviestins, 2006; Laxhammar, 2008). Separately, Fooladvandi, Brax, Gustavsson, and Fredin (2009) investigated whether or not the Bayesian networks acquired from expert knowledge has the ability to detect activities based on a signature-based detection approach.

2.2. Norm-based anomaly detection (data-driven)

It is common to use classification algorithms via a data-driven approach to perform anomaly detection. However, in the case of maritime security research, conventional classification algorithms cannot be directly applied due to the lack of adequate samples and known cases that should be classified as anomalous (Riveiro, Falkman, & Ziemke, 2008). A challenge faced in this research project was therefore that adequate instances of maritime security threats are not readily available in historical data for system training purpose.

An alternate data-driven approach is to build *normal models* via discovering knowledge from historical data (Brax & Niklasson, 2009). In an example, Ristic, Scala, Morelande, and Gordon (2008) presented a statistical analysis of vessel motion patterns (using adaptive Kernel Density Estimation) in ports and waterways based on AIS data. To build normal models, other methods employed by researchers include clustering (Laxhammar, 2008) and probability models. Johansson and Falkman (2007) used Bayesian network (based on probability theory) to build models of normal ship movements. Similarly, Jakob, Vaněk, Urban, Benda, and Pěchouček (2010) built probability models of vessels' locations and trajectories for detecting anomalous vessels. There were also reports of using conformal prediction method for detecting maritime anomalies, again based on normal training data (Laxhammar & Falkman, 2010). A Markov model was applied to define normal models based on historical AIS data, with an aim of detecting abnormal movements (Tun, Chambers, Tan, & Ly, 2007). In BAE Systems' study of learning normal vessel movements, AIS data was used in methodologies, including ARTMAP neural network for pattern identification (Rhodes, Bomberger, Seibert, & Waxman, 2005). BAE Systems also reported the development of a system that fuses radar, AIS, and video-based information for more accurate geographical data of vessels (Seibert et al., 2006).

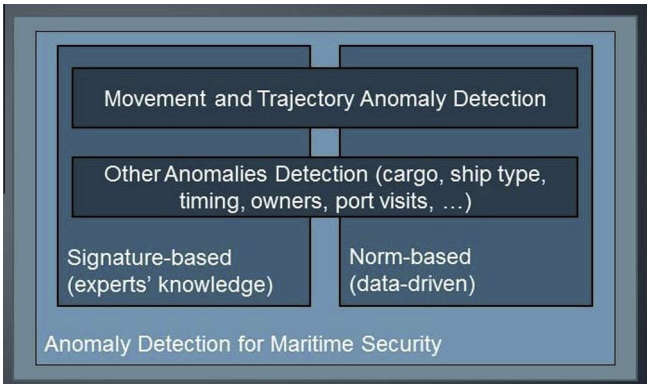

Fig. 1. Research of anomaly detection for maritime security.

3. Research problem and objective

There have been several key challenges for the signature-based anomaly detection approach. It is not straightforward for knowledge engineers to elicit, organize, and represent formal knowledge from experts. The process has to be on-going, as both real-world circumstances and experts' knowledge change with time. Signatures of illegal activities are open-ended, and may therefore be difficult for any system to obtain a robust coverage. Further, the boundaries between anomalous and normal behaviors are often difficult to be defined verbally during knowledge elicitation procedures. It is also a challenge to convert experts' verbatim into formal knowledge. As an alternative to the signature-based approach, there have been studies of building normal models, with the aim of detecting maritime anomalies.

Since anomaly can be defined as deviations from normality, normal models can be used to identify anomalies. This approach is akin to domain experts possessing knowledge of the norms derived from their experiences, and could therefore identify anomalies. As reflected in the reviewed literature, current studies of building normal models have been restricted to detecting anomalies in terms of vessels' physical movements, typically with AIS and/or radar-based datasets as inputs (Riveiro & Falkman, 2011). However, this study identified that information fields (e.g. cargo types, ship types, and port last visited) from AIS, port management systems, and classification societies' datasets potentially contain critical knowledge for identifying maritime anomalies. This research therefore investigates the extraction of knowledge from multiple datasets in supporting anomaly detection via a norm-based approach.

4. An approach of knowledge discovery using genetic algorithm

Knowledge discovery is a process of identifying patterns in a given training data set. These hidden patterns are useful as a basis for making decisions and predictions. Machine learning is an approach of knowledge discovery, whereby autonomous algorithms are prescribed for acquiring knowledge, and to improve the organization of the knowledge obtained (Tecuci & Kodratoff, 1995).

One important step in applying machine learning technique is to decide an effective representation scheme for both the training data and the knowledge extracted. Knowledge representation involves the modeling of knowledge in explicit schemes that facilitate the acquisition, learning, manipulation and application of knowledge. Schemes include mathematical expressions, predicate calculus, conceptual graphs, frames, scripts, objects, semantic networks and production rules (Chong, Chen, & Leong, 2009). In a

knowledge-based system (KBS), one or more types of knowledge representation schemes can be specified. Among the types of representation techniques, production rules are commonly employed due to its numerous advantages. For a KBS to employ production rules, in the forms of 'If–Then' structure, there are various advantages including (Morik, 1989):

- Rules are relatively simple to construct.
- It enables rapid prototyping, and tests can begin with just a few rules.
- It is a natural way to summarize human knowledge.

In this research, production rules are employed for knowledge representation.

In order to automate the process of identifying vessels of interest, explicit knowledge is required. To this end, this research addressed the problem by inducting knowledge from empirical data. A study of relevant data (including AIS data) in this project showed that the volume of training data can be large in practice. Computational effort to analyze the data within a limited time frame can be challenging, especially when the data is streamed in real-time. In handling large dataset, it is more computationally expensive to fully explore the entire solution space. A strategy is to heuristically search for solutions – production rules in this case.

This project employed genetic algorithm (Goldberg, 1989) as a method of inducting rules from large datasets. Genetic algorithms (GAs) are stochastic and evolutionary search techniques based on the principles of biological evolution, natural selection, and genetic recombination. They simulate the principle of 'survival of the fittest' in a population of potential solutions known as *chromosomes*. As shown in Fig. 2, the population evolves over time through a process of competition.

The algorithm begins with a population of chromosomes generated either randomly or from some set of known specimens. The population of chromosomes will be cycled through three steps, namely evaluation, selection, and reproduction. Each chromosome represents one plausible solution to the problem or a rule in a classification problem. It can be encoded using a binary string, or depending on the complexity and data of the problem domain, chromosomes can be encoded in integers or floating point variables. In the evaluation step, each string is evaluated according to a given performance criterion known as *fitness function*, and thereby assigned a *fitness score*. In the selection step, a decision is made according to the fitness score assigned to decide which of the individuals are permitted to produce offspring and with what probability. Finally, the reproduction step involves the creation of offspring chromosomes by two genetic operators, namely cross-over and mutation. This is a pivotal part of a genetic algorithm as these genetic operators have a significant impact on the performance of the algorithms. Upon the completion of the cross-over operation, mutation takes place. This step prevents the solution from converging prematurely. For a chromosome encoded in binary, genes are randomly selected to undergo mutation operation, where '1' is converted to '0' or vice versa. The three-step cycle continues for a predetermined number of *generations*, or until an acceptable performance level is achieved. In this research, genetic algorithm-based learning techniques were applied to discover production rules from large search spaces.

5. A proposed system – GeMASS

Based on the research problem identified in Section 3, a system termed as GeMASS (GEnetic algorithm knowledge discovery for MAritime Security System) was proposed and investigated in this research. The objective of GeMASS is to provide system users with alerts of vessels that bear anomalous characteristics embedded within datasets. System users or the port authority may therefore perform security checks on the highlighted vessels (see description in Section 5.1). GeMASS is a knowledge-based system that relies on real-world domain knowledge. The knowledge is computationally represented within the system in the form of production rules. GeMASS is prescribed to 'gain experiences' as time passes, so that the latest knowledge generated can be applied in real time. This learning ability of GeMASS is based on the machine learning approach. GeMASS is designed to learn from both real-time data stream and system users. The learning process is commonly known as knowledge (or rule) induction (or discovery). Rule induction in GeMASS is based on genetic algorithm (see Section 5.2).

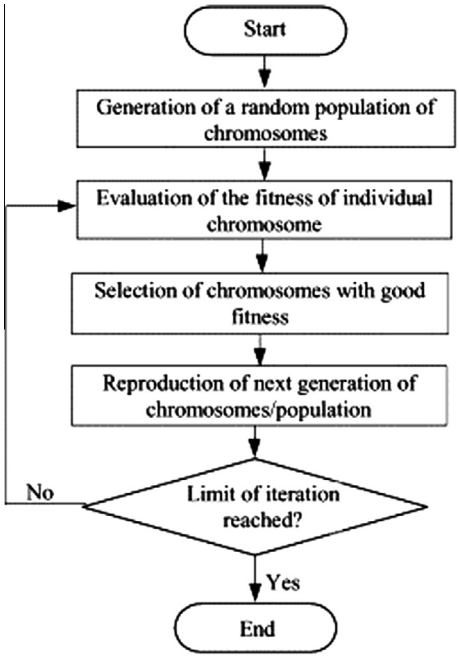

Fig. 2. Basic steps of a genetic algorithm.

5.1. An overview of GeMASS

Fig. 3 shows the concept of GeMASS, while Fig. 4 shows the flow of information within the system. The system receives data streams from AIS, vessels (e.g. via a port management or security system), and other sources. Classification societies' datasets are also applicable in the system. Real-time raw datasets are streamed into a data pre-processing function, as shown in Fig. 4. Preprocessing is required as the raw datasets acquired from a number of different sources can be heterogeneous and disparate. The preprocessed data forms an information table (with no classification and decision), and will be applied to knowledge inference in a real-time ship analysis function.

A knowledge inference engine serves to assess each incoming vessel based on production rules residing in the system. Outcome of the inference procedures will be post-processed and decoded into information that is meaningful for users' consumption. Based on the information provided to users, users may then decide on an appropriate follow-up action, for example to activate security forces to physically conduct checks on suspicious vessels. System

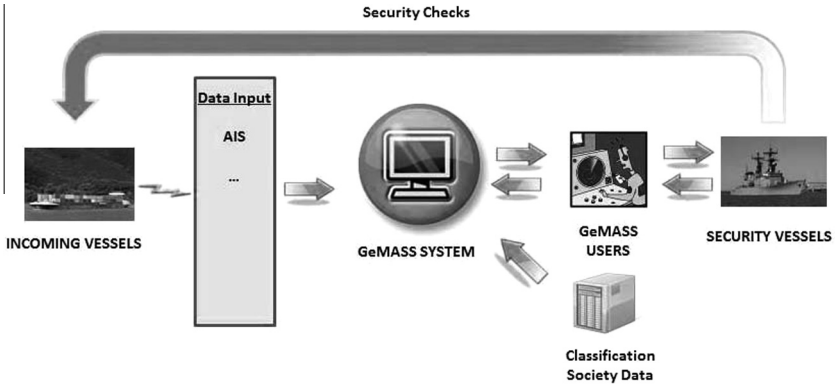

Fig. 3. The overall concept of GeMASS in operation context.

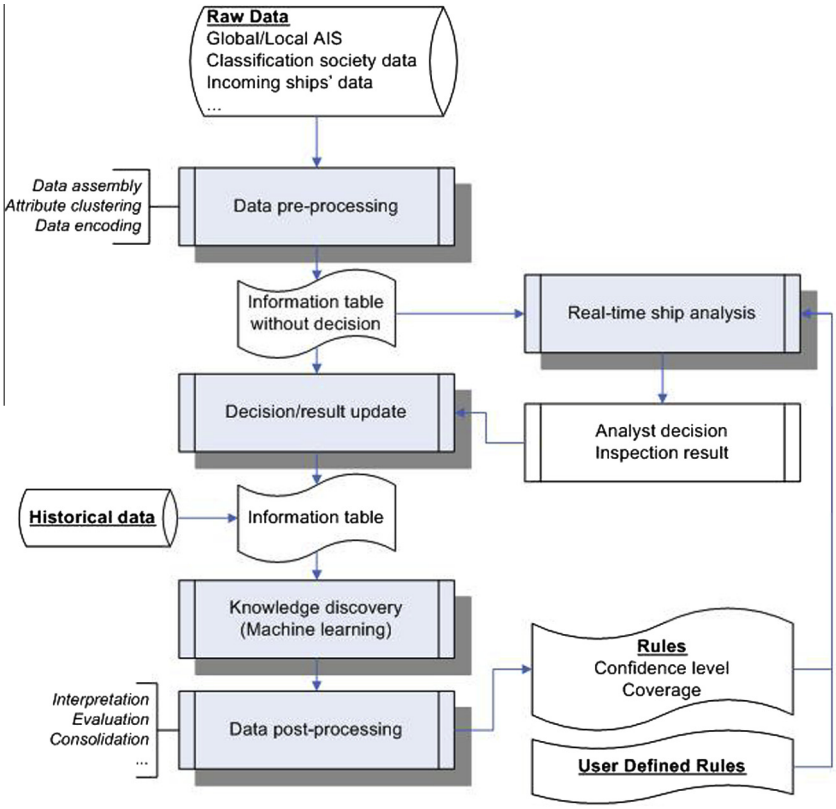

Fig. 4. Overall information flow in GeMASS.

users may feed GeMASS with the outcomes of such physical security checks, to support the generation of new knowledge. For example, during a random inspection, certain security breach may be discovered and can be updated in the system. In another example, an inspection prompted by GeMASS system, if later found positive, will validate the system's knowledge. This form of new found knowledge as a result of ship inspections can be transferred to the system. Such findings serve as decisions for the information table to be used for machine learning. Training datasets consist of an information table (with decisions) and any historical data accumulated. They are used for knowledge discovery via a machine learning process.

Knowledge in the form of production rules will be extracted by GeMASS and post-processed for accumulation within the system. Rules that are manually defined by users may be included as well. The latest set of rules will henceforth be applicable by the system for real-time knowledge inferences.

5.2. Knowledge discovery and inference

GeMASS consists of a knowledge discovery module based on genetic algorithm (GA). The chromosome of the GA employed in the system, which represents a production rule, is expressed as a string comprising of non-negative integers that represent the values of attributes. The length of the chromosome is fixed, which is equal to the number of attributes plus one (i.e. the decision). The *k*th gene of a chromosome represents the value of the *k*th attribute while the last gene represents the decision. For example, a chromosome "1-2-4-3-∗-1" is equivalent to the rule "if (Attribute 1 = 1 and Attribute 2 = 2 and Attribute 3 = 4 and Attribute 4 = 3 and Attribute 5 = any value), then (Decision = 1)". In this example, while the length of the chromosome is six, this research defines the *effective length* as five due to the 'any value' status of Attribute 5. This definition is for the purpose of introducing a unique fitness measurement for preventing knowledge over-generalization.

Partial crossover operation, which involves partial strings of two parent chromosomes, is used to exchange the ordering of chromosomes. Since these partial strings contain different types of genes, the offspring generated can be computationally illegal. Thus, a repair operation will be carried out immediately after each crossover operation in order to legalize the offspring. As the representation of a chromosome is context-dependent and the offspring generated needs to inherit the genetic traits of their respective parents, the partial string in each of the offspring must therefore have the same ordering as their parents. Otherwise, the genes in the partial string may refer to different operations. The procedure for the crossover operation is given as follows.

(1) Choose a partial string of same length in every parent randomly;
(2) Exchange the selected genes in the two parents; and
(3) Modify genes to legalize the offspring, that is, if a gene carries the value that is not valid for the attribute it represents, a valid value will be randomly selected and assigned to the gene.

The primary purpose of performing mutation is to inject variation into a population, to facilitate reviving some essential genetic traits, and to avoid pre-mature convergence caused by the existence of some super-chromosomes. In this work, a uniform mutation operation is introduced to replace the value of a randomly selected gene with a valid value of the attribute the gene represents.

The fitness or relative fitness of chromosomes needs to be evaluated in order to select suitable chromosome or chromosome pairs for genetic operations. In this work, the fitness value comprises of four measurements. They are confidence level, coverage, preferred minimal effective length and distance. The respective equations are as shown in (1)–(4).

$$C_i^f = n_i^{TP}/n_i \tag{1}$$

$$C_i^v = n_i^{TP}/n \tag{2}$$

$$C_i^l = \min(1, l_i/L) \tag{3}$$

$$C_i^d = \sum_{k=1}^{K} y_k/K \quad y_k = \begin{cases} 1 & \text{if } x_{ik} \in X_k^T \text{ or } x_{ik} = 0 \\ 0 & \text{else} \end{cases} \tag{4}$$

where C_i^f is the confidence level of a rule represented by chromosome i of a GA population, C_i^v is the coverage of the rule represented by chromosome i of a GA population, C_i^l is the preferred minimal effective length measurement of a rule represented by chromosome i of a GA population, C_i^d is the distance measurement of a rule represented by chromosome i of a GA population, n_i^{TP} is the number of records in a training dataset that are correctly classified by the rule represented by chromosome i of a GA population, n_i is the number of records in a training dataset that match the conditions of the rule represented by chromosome i of a GA population, n is the number of records in a training dataset, l_i is the length of a rule represented by chromosome i of a GA population, L is the preferred minimal effective length of a rule (set at 3 in the case study), K is the total number of attributes, x_{ik} is the value of attribute k of a rule represented by chromosome i of a GA population, and X_k^T is the list of values of attribute k that appear in a training dataset.

C_i^f is the confidence level of the rule represented by a chromosome. It indicates the percentage of the correctly classified records by applying the rule to a training dataset. C_i^v is a similar measurement as C_i^f. It represents how many records are covered and classified correctly by the rule in terms of percentage. C_i^l is used to reflect whether the number of attributes of the rule under examination is greater than or equals to the preferred minimal number that has been pre-defined. As in the GeMASS environment, it is necessary to have some specific rules that can pin-point detailed patterns instead of general rules that could possibly over generalize. The last measurement C_i^d is to account for the distance of a rule represented by a chromosome to the training dataset. It is able to improve the GA performance by attributing a score to the rules/chromosomes that cannot be measured and sorted by confidence level and coverage.

The fitness value of a chromosome is given by Eq. (5).

$$f_i = w_f C_i^f + w_v C_i^v + w_l C_i^l + w_d C_i^d \tag{5}$$

where f_i is the fitness value of chromosome i of a GA population, w_f is the weightage of confidence level, w_v is the weightage of coverage, w_l is the weightage of preferred minimal effective length measurement, and w_d is the weightage of distance measurement.

In general, w_f and w_v shall be given relatively higher values since confidence level and coverage are two key criteria that determine the quality of a generated rule. The weight w_l contributes to the penalty when the number of attributes in a rule is less than the preferred minimal effective length. It reduces the generation of shorter and over-generalizing rules. The weight w_d affects the speed of convergence of the GA population. As a reference, the weights 0.4, 0.4, 0.1, and 0.1 of w_f, w_v, w_l, and w_d respectively led to the generation of the most effective rule sets in our experiments.

In GA, the way in which a population is generated would affect the survival of fitter chromosomes. In this work, the roulette wheel approach (Goldberg, 1989) is used to select mating chromosomes within the population. Basically, the roulette wheel approach guarantees chromosomes with higher fitness values to occupy a larger slot-size in the roulette wheel. As a result, these chromosomes are more likely to be selected to form the next generation of chromosomes. Such an approach gives every chromosome a chance to propagate as it is based on the probability distribution of fitness values. Alternatively, the elitist selection scheme, which selects the fittest to form the next generation of chromosomes, can also be used together with the roulette wheel selection. It aims at preserving the fittest chromosomes and ensures their survival in the next generation.

Rules generated from the knowledge discovery engine are consolidated, as shown in Fig. 4. In the process of consolidation, duplicated rules will be removed. GeMASS then proceeds to perform rules pruning. Pruning of rules are performed to simplify a rule set without compromising its inference power. It improves the efficiency of knowledge inference by reducing the number of the rules in the knowledge base. It further helps to prevent the problem of over-fitting, which may lead to lower accuracy. The rule pruning mechanism performs two tasks: pruning and simplifying. It examines the rules extracted by the GA-based rule-induction engine and uses Boolean operators such as union and intersection to prune and simplify rules. During the pruning operation, redundant rules such as rules that are covered by a general rule associated with higher confidence level will be removed.

Sets of rules generated by the knowledge discovery engine are associated with two indices – the confidence and the coverage levels (as mentioned above). The confidence level reflects the accuracy of a rule against a training data set. On the other hand, coverage of a rule is defined as the number of specific instances covered by the rule in training datasets divided by thex total number of instances with the same decision as the rule. The post-processed rules, along with any user-defined rules will be

accumulated in the knowledge base of the system, for deployment in real-time ship analysis (i.e. knowledge inference).

The knowledge inference engine essentially serves to compare (i.e. match) the information table that represents incoming vessels with the set of rules that has been deployed within the knowledge base. For each vessel, the production rules that match its attributes will be noted by the system (see Fig. 5). A vessel that matches rules with higher coverage has attribute–value pairs (i.e. characteristics) that have been more commonly observed in history compared to a vessel that matches a rule with lower coverage. Therefore, matching with a rule that has high coverage suggests that the vessel is relatively common or normal with respect to historical data. On the other hand, matching with a rule that has lower coverage denotes that the vessel has less common characteristics. A vessel that matches rules with lower coverage value can be inferred to possess attribute–patterns that are less common. In effect, the system serves to highlight incoming vessels that bear uncommon characteristics. As such, the sorting of incoming vessels in terms of the coverage of matched rules (see Fig. 6) will allow system users to prioritise the vessels of interest.

A prototype system constructed during the investigation serves to sort and highlight incoming vessels. By clicking on the respective incoming vessel's record, analysts may proceed to view the details of the matched production rules (see Fig. 7). With the details of each matched rules as decision support, analysts may consider any follow-up actions.

6. A case study

To study the applicability of the proposed knowledge discovery and inference approach in the maritime domain, a prototype of GeMASS was constructed. A set of historical real-world test data was gathered for the study. The test was based on the following datasets:

(i) A classification society database with 95,990 vessel records.
(ii) A masked historical (six-month period) dataset from a vessel arrival notification system and a port movement system with 30,079 records. This dataset was applied as training dataset.
(iii) A masked historical (three-day period) dataset from the same ship arrival notification system and port movement system with 78 records (i.e. 78 vessels). This dataset was used as a testing dataset by simulating it as the data related to incoming vessels.
(iv) AIS data that can be used to trace the geographical paths of the 78 ships identified for system testing.
(v) A pseudo-blacklist with two records, as a typical list maintained by port security agencies or authorities.

The set of masked historical (6-month period) data from a ship arrival notification system and a port movement system with 30,079 records was fed to GeMASS for training. The parameters of GAs are very important to the performance of the algorithms.

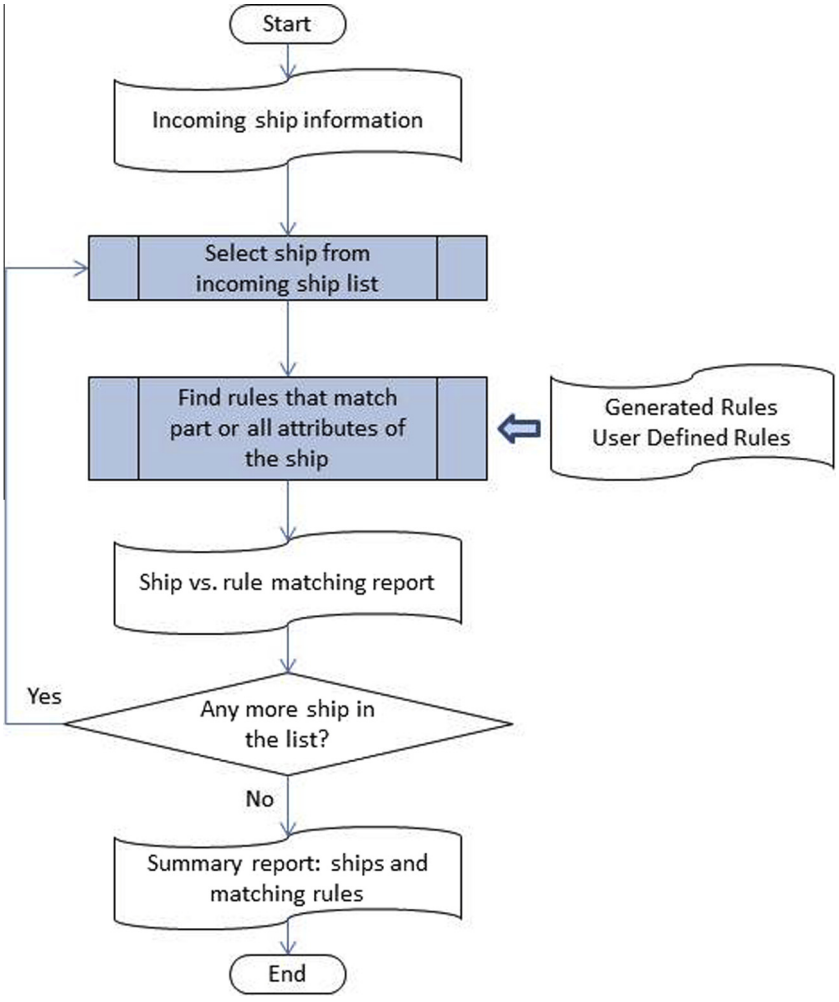

Fig. 5. Real-time ship analysis.

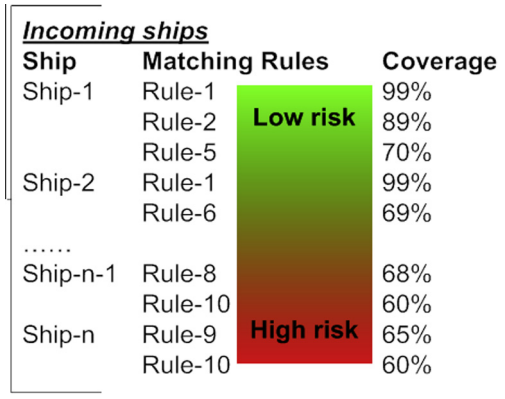

Fig. 6. Rules matching and sorting.

A set of carefully selected GA parameters improves the ability of the algorithm in reaching near global optimal solutions. However, as the mathematical foundation of GA is not rigorous, deterministic method for the selection of parameters does not exist. As such, different combinations of GA parameters were tested and the set with the highest performance as shown in Table 1a was applied in the case study.

Table 1b shows a set of typical rules generated by the knowledge discovery engine. For example, Rule 1 in Table 1b shows that 99.23% of the ships records were associated with the following pattern: (i) ship was not on the blacklist; and (ii) security related information, such as the approved security level, current security level onboard, accident and refugee information, was provided and accurate. However, the pattern of a ship not on the blacklist and its port movement being totally different from past visits covers a merely 2.66% of the ship records in the training dataset, as indicated by Rule 8. Comparing two ships, Ship A and Ship B that matches Rules 1 and 8 respectively, Ship B would warrant more attention than Ship A due to its matching with a rule of lower coverage – indicating a characteristic of lower occurrence in history.

Dataset from a three-day period that is disjoint from the training dataset was fed to the prototype system as testing data. The system performed screening and analysis of vessels based on the input information. Ten randomly selected vessels were highlighted for studies. The sorting of the 10 incoming vessels in terms of coverage of the matched rules in different stages of data assembly is as shown in Table 2. For the purpose of presentation and discussion, four ships (system identification (ID) numbers: 954,620, 1,003,170, 1,006,000, and 1,011,890) were singled out, as highlighted in Table 2.

A scenario of vessels progressively moving towards a port of interest was simulated. The progressive scenario was divided into three stages. In the first stage, vessels were assumed to be at a distance away where only AIS data was available and based upon for system analysis. In the second stage, data from a local ship arrival notification system was fed to the system. This is to simulate the reporting by vessels while approaching a port. Finally, port movement data when the vessels were nearer or within the port was made available to the system. The outputs and responses of the prototype system were studied and discussed.

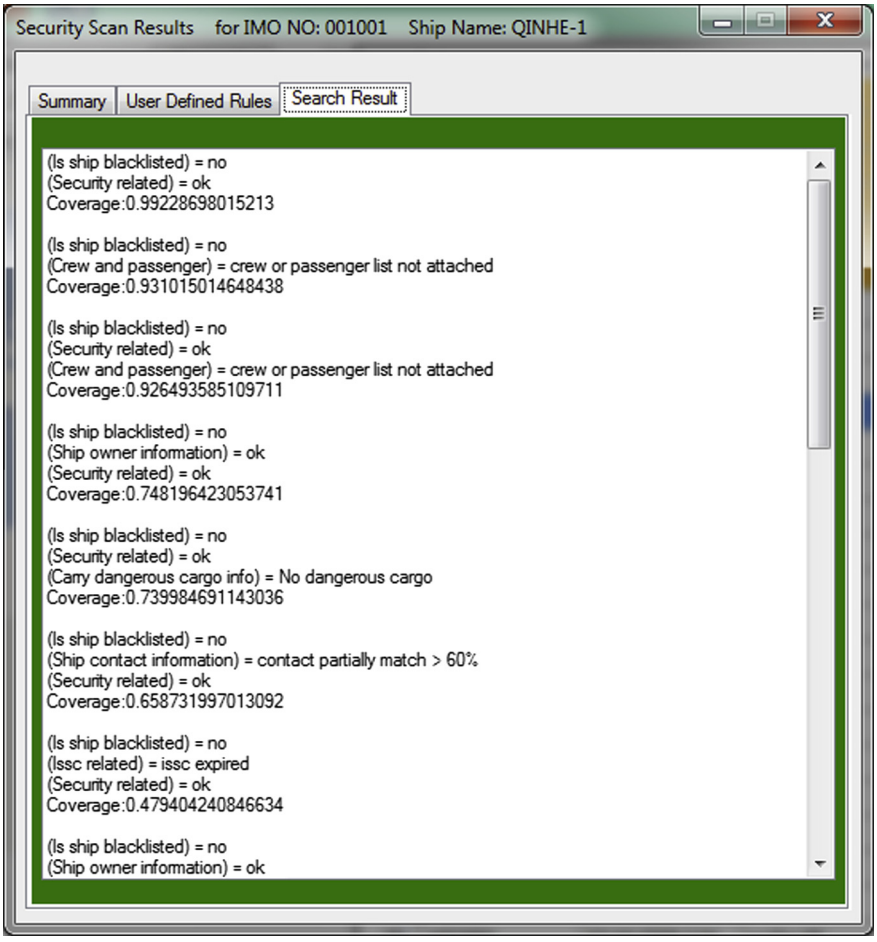

Fig. 7. Details of rules fired for a vessel.

Table 1a
Parameters defined for GeMASS rule extraction.

Parameters	Value	Description
GA_number_of_generation	1000	Total number of GA generations
GA_population_size	200	GA population size
GA_crossover_rate	0.9	GA crossover probability
GA_mutation_rate	0.1	GA mutation probability
GA_elitist_rate	0.08	Percentage for GA elitist strategy
F_weight_confidence	0.4	Weightage of confidence level
F_weight_coverage	0.4	Weightage of coverage
F_weight_minimal_effective_length	0.1	Weightage of minimal effective length measurement
F_weight_distance	0.1	Weightage of distance measurement

Table 1b
Sample rules generated by the prototype system.

No	Rule details	Coverage
1	(Is ship blacklisted) = no (Security related) = ok	0.9923
2	(Is ship blacklisted) = no (Port movement) = match	0.9734
3	(Is ship blacklisted) = no (Arrival information) = partially provided	0.9688
4	(Is ship blacklisted) = no (Crew and passenger) = crew or passenger list not attached	0.9310
5	(Is ship blacklisted) = no (Security related) = ok (Crew and passenger) = crew or passenger list not attached (Port movement) = match	0.9017
6	(Is ship blacklisted) = no (Arrival information) = partially provided (Security related) = ok (Crew and passenger) = crew or passenger list not attached	0.8980
7	(Is ship blacklisted) = no (Ship owner information) = ok (Cargo information) = not provided	0.1893
8	(Is ship blacklisted) = no (Port movement) = not match	0.0266
9	(Is ship blacklisted) = no (Crew and passenger) = crew or passenger list not attached (Port movement) = not match	0.0253

6.1. Stage with only AIS information

A screening of ships in the region around a particular port can be performed based on information gathered from the AIS. In practice, this is the stage where both ship arrival and port movement information were not available. During this stage, ship or vessel analysis was preliminary as there was only limited data, such as ship name, International Maritime Organisation (IMO) number and ship type gathered from the AIS system. Despite that, an early warning could be raised by identifying vessels of interest using data from the AIS. Table 2a shows a list of the ten selected vessels and the coverage values of the rules that matched the respective vessels. As shown, Ship 954,620 was associated with rules of the lowest coverage value at 24.07%, while Ship 1,006,000 with a rule of high coverage value at 99.23%. As such, the first few ships in the sorting list, such as Ships 954,620 and 1,003,170, might be of interest to system users, security agency, and/or the port authority. Details of the sample matching rules can be found in Table 3a. For example, Ship 954,620 matches Rule A-1 which had the lowest coverage due to that in this test, the ship was on the blacklist. Through further analysis of the two ships in question (Ships 954,620 and 1,003,170), it was found that there were no corresponding records in the classification societies' database. This results in the higher priority levels for Ships 954,620 and 1,003,170 comparing with Ships 1,011,890 and 1,006,000.

6.2. Stage with ship arrival information added

Before vessels calling the port of interest, relevant information has to be reported through a ship arrival notification system, as commonly practiced by port authorities. At this stage, GeMASS will be fed with more details of a particular ship, such as the cargos it carries, the expected date of arrival, and ship owner/company information. Consequently, a more detailed screening can be carried out by assembling AIS data with the arrival records. Table 2b shows the refreshed sorting of the ships when a ship analysis was conducted with both the ship arrival and AIS data. Ship 954,620 was still the ship on top of the priority list. By looking at the matched rules, it is apparent that the available information was not able to match any rule at that stage (see Table 2b). However, Ship 1,006,000 became second on the list, mainly due to the security concerns reported in the ship arrival system. This was indicated by the only matched rule (Rule B-1 in Table 3b), which has a lower coverage of 0.77%.

Ship 1,011,890 matched Rule B-2, which has a coverage value of 93.1%. Rule B-2 is a result of a common non-compliance in practice (i.e. not submitting crew and passenger list), as evident from the rule description shown by the system. The issue of *common non-compliance* will be raised in the discussion section.

6.3. Stage with port movement information added

With the addition of data registered by a port movement information system when a vessel is within the port's geographical limit, rules concerning port movement can be employed. The refreshed sorting list of vessels is as shown in Table 2c. Ship 954,620 matched with rules of higher coverage, such as Rules C-1 and C-2 (see Tables 3c-1 and 3c-2). In comparison, none of the rules was able to match with vessels' data during previous stages when the port movement information was not available.

7. Discussion

The prototype GeMASS was employed to test and demonstrate the functions of the proposed methodology. The test consists of three stages, according to the availability of datasets as vessels were simulated to move towards a port of interest as in real life cases. As vessels move along, data was simulated to be progressively received and considered in the series of real-time analysis. As a result, it was shown that the priority of vessels highlighted

Table 2a
Ship sorting result and rule coverage (with AIS data).

Ship Id	Coverage of matching rules
954,620	***0.2407***/0/0
100,3170	***0.9923***/0.931/0.9265/0.4794/0.2458/0.2441/0.2407/0.2407/0.2335/0.135/0.1314/0.0069
1,006,010	0.9923/0.931/0.9265/0.4794/0.2458/0.2441/0.2407/0.2407/0.2335/0.135/0.1314/0.0069
955,170	0.9923/0.931/0.9265/0.4794/0.2458/0.2441/0.2407/0.2407/0.2335/0.135/0.1314/0.0069
1,003,570	0.9923/0.931/0.9265/0.6587/0.4794/0.2441/0.2407/0.2407/0.2335/0.135/0.1346/0.1274
1,005,980	0.9923/0.931/0.9265/0.6587/0.4794/0.2441/0.2407/0.2407/0.2335/0.161/0.1346/0.1274
1,003,560	0.9923/0.931/0.9265/0.6587/0.4794/0.2441/0.2407/0.2407/0.2335/0.161/0.1346/0.1274
1,011,890	0.9923/0.931/0.9265/***0.6587***/0.4794/0.2441/0.2407/0.2407/0.2335/0.161/0.1346/0.1274
1,005,990	0.9923/0.931/0.9265/0.682/0.6587/0.6029/0.4794/0.3178/0.2441/0.2407/0.2407/0.2335/0.161
1,006,000	0.9923/0.931/0.9265/***0.682***/0.6587/0.6029/0.4794/0.3178/0.2441/0.2407/0.2407/0.2335/0.161

Table 2b
Ship sorting result and rule coverage (with ship arrival Information).

Ship Id	Coverage of matching rules
954,620	0/
1,006,000	***0.0077***/
1,003,560	0.931/0.1274/0.0077
1,011,890	***0.931***/0.1893/0.1274/0.0077/0.0066
1,003,570	0.9923/0.9688/0.9613/0.7516/0.7516/0.7482/0.7278/0.4794/0.3542/0.161/0.1346
955,170	0.9923/0.9688/0.9613/0.931/0.9265/0.898/0.2523/0.2458/0.2441/0.2407/0.2407/0.2388/0.2335/0.135/0.1314/0.0118/0.0069/0.0055
1,003,170	***0.9923***/0.9688/0.9613/0.931/0.9265/0.898/0.7516/0.7516/0.7278/0.501/0.2523/0.2458/0.2441/0.2388/0.135/0.1314/0.054/0.0055
1,005,990	0.9923/0.9688/0.9613/0.931/0.9265/0.898/0.7516/0.7516/0.7278/0.682/0.4794/0.3178/0.2523/0.2441/0.161
1,006,010	0.9923/0.9688/0.9613/0.931/0.9265/0.898/0.7516/0.7516/0.74/0.7278/0.4794/0.2458/0.2441/0.2388/0.135/0.1314
1,005,980	0.9923/0.9688/0.9613/0.931/0.9265/0.898/0.7516/0.7516/0.74/0.7278/0.501/0.2441/0.161/0.1346/0.1332/0.1274

Table 2c
Ship sorting result and rule coverage (with port movement Information).

Ship Id	Coverage of matching rules
954620	***0.9058***/0.6802/0.6664/0.6313/0.6172/0.2305/0.2232/0.1932/0.1579/0.1548/0.1495/0.0488/0.028/0.0255/0.02/0.0191/0.0177/0.0118/0.0072
1011890	***0.931***/0.1893/0.0266/0.0253/0.0223/0.0213/0.0187/0.0183/0.0183/0.0114/0.0077/0.0066/0.0026/0.0015/0.0005/0.0001
1006000	***0.9734***/0.7313/0.7209/0.7209/0.6802/0.6802/0.6664/0.6664/0.6115/0.487/0.475/0.475/0.2305/0.2305/0.1893/0.1817/0.1785/0.1579
1003560	0.9734/0.931/0.9058/0.9057/0.7209/0.7209/0.681/0.2421/0.2305/0.2305/0.2273/0.2232/0.2232/0.1989/0.1817/0.1597/0.1596/0.1486
1003570	0.9923/0.9734/0.9688/0.9662/0.9662/0.9613/0.9454/0.9453/0.9384/0.9383/0.7516/0.7516/0.7482/0.7429/0.737/0.7369/0.7313
1005980	0.9923/0.9734/0.9688/0.9662/0.9662/0.9613/0.9454/0.9453/0.9384/0.9383/0.931/0.9265/0.9058/0.9057/0.9017/0.9017/0.898
1003170	***0.9923***/0.9734/0.9688/0.9662/0.9662/0.9613/0.9454/0.9453/0.9384/0.9383/0.931/0.9265/0.9058/0.9057/0.9017/0.9017/0.898
1005990	0.9923/0.9734/0.9688/0.9662/0.9662/0.9613/0.9454/0.9453/0.9384/0.9383/0.931/0.9265/0.9058/0.9057/0.9017/0.9017/0.898
1006010	0.9923/0.9734/0.9688/0.9662/0.9662/0.9613/0.9454/0.9453/0.9384/0.9383/0.931/0.9265/0.9058/0.9057/0.9017/0.9017/0.898
955170	0.9923/0.9734/0.9688/0.9662/0.9662/0.9613/0.9454/0.9453/0.9384/0.9383/0.931/0.9265/0.9058/0.9057/0.9017/0.9017/0.898

to system users changes based on the anomalies detected (i.e. the coverage of the rules fired).

Genetic algorithm as a machine learning method allows GeMASS to learn the 'norms' of vessels' characteristics based on historical data. Knowledge generated, in the form of production rules, is characterized by a coverage index. As shown in Table 1b, rules discovered from the experimental data are described by a coverage index, ranging from 0.000 to 1.000. By definition of the coverage index, a rule with a low coverage is generated by an accordingly low number of training instances relative to the entire training dataset. Examples of rules with low coverage values can be seen in Table 1b. When an incoming vessel manifests characteristics that match a rule with low coverage (for examples Rules A-1 and B-1 in the test), the particular set of characteristics of the ship is accordingly of lower occurrence in history, based on the historical dataset used for system training. As demonstrated in the case study, this rationale provides the system with a means to detect and to sort vessels that manifest uncommon characteristics.

Table 3a
Sample ship information and matching rules detail when only AIS data was available (real data masked).

No	Ship record Id	IMO No	Ship name	Ship type	Ship blacklisted?	Found in classification DB?	Matching rule(s)/coverage
1	954,620	00000001	JASMINE	BULK CARRIER	Yes	Yes	Rule A-1: (Security related) = ok (Cargo information) = not provided Coverage:0.2407
2	1,003,170	00000002	EMERALD	CONTAINER SHIP	No	No	Rule A-2: (Is ship blacklisted) = no (Security related) = ok Coverage:0.9923 Rule A-3: (Is ship blacklisted) = no (Crew and passenger) = crew or passenger list not attached Coverage:0.9310
3	1,011,890	00000003	OCEAN	BARGE	No	Yes	Rule A-4: (Is ship blacklisted) = no (Ship contact information) = contact partially match >60% (Security related) = ok Coverage:0.6587
4	1,006,000	00000004	KUTAMI	GENERAL CARGO SHIP	No	Yes	Rule A-5: (Ship information 1) = ok (Is ship blacklisted) = no (Security related) = ok Coverage:0.6820

Table 3b
Sample ship information and matching rules detail when ship arrival data was available (real data masked).

No	Ship record Id	IMO No	Ship name	Ship type	Ship blacklisted	Found in classification DB	Security problem	Matching rule(s)/coverage
1	954,620	00000001	JASMINE	BULK CARRIER	Yes	Yes	No	
2	1,006,000	00000004	KUTAMI	GENERAL CARGO SHIP	No	Yes	Yes	Rule B-1: (Is ship blacklisted) = no (Security related) = some concern Coverage:0.0077
3	1,011,890	00000003	OCEAN	BARGE	No	Yes	Yes	Rule B-2: (Is ship blacklisted) = no (Crew and passenger) = crew/passenger list not attached Coverage:0.9310 Rule B-3: (Is ship blacklisted) = no (Ship owner information) = ok (Cargo information) = not provided Coverage:0.1893
4	1,003,170	00000002	EMERALD	CONTAINER SHIP	No	No	No	Rule B-4: (Is ship blacklisted) = no (Security related) = ok Coverage:0.9923 Rule B-5: (Is ship blacklisted) = no (Arrival information) = partially provided Coverage:0.9688

Table 3c-1
Sample ship information and matching rules detail when PTMS data was available (real data masked).

No	Ship record Id	IMO No	Ship name	Ship type	Ship blacklisted?	Found in classification DB?	Security problem	PTMS movement	Matching rule(s)/coverage
1	954,620	00000001	JASMINE	BULK CARRIER	Yes	Yes	No	Match	Rule C-1: (Crew and passenger) = crew or passenger list not attached (PTMS movement) = match Coverage:0.9058 Rule C-2: (Ship information 2) = ok (PTMS movement) = match Coverage:0.6802
2	1,011,890	00000003	OCEAN	BARGE	No	Yes	Yes	Not match	Rule C-3: (Is ship blacklisted) = no (Crew and passenger) = crew or passenger list not attached Coverage:0.9310 Rule C-4: (Is ship blacklisted) = no (Ship owner information) = ok (Cargo information) = not provided Coverage:0.1893 Rule C-5: (Is ship blacklisted) = no (PTMS movement) = not match Coverage:0.0266 Rule C-6: (Is ship blacklisted) = no (Crew and passenger) = crew or passenger list not attached (PTMS movement) = not match Coverage:0.0253

It should be mentioned that while in most cases rules with high coverage values (i.e. indicating common phenomenon) point to compliances, there may be cases of rules having high coverage values and yet representing non-compliances. For example, when majority of the vessels do not submit crew or passenger list (a common issue in practice), the coverage value of Rule 4 in Table 1b is high. Commonality or coverage value therefore typically but does not always indicate compliances. In many practical domains, what is common may not necessarily be in line with certain regulations. This phenomenon or behavior is generally true for all anomaly detection systems. In the third phase of the test, it is interesting to note that the commonness in non-compliance (i.e. non-submission of passenger and crew list) showed up as Rules C-1 and C-3, and matched with Ships 954,620 and 1,011,890

Table 3c-2
Sample ship information and matching rules detail when PTMS data was available (real data masked).

No	Ship record Id	IMO No	Ship name	Ship type	Ship blacklisted?	Found in classification DB?	Security problem	PTMS movement	Matching rule(s)/ coverage
3	1,006,000	00000004	KUTAMI	GENERAL CARGO SHIP	No	Yes	Yes	Match	Rule C-7: (Is ship blacklisted) = no (PTMS movement) = match Coverage:0.9734 Rule C-8: (Is ship blacklisted) = no (Ship owner information) = ok (PTMS movement) = match Coverage:0.7313
4	1,003,170	00000002	EMERALD	CONTAINER SHIP	No	No	No	Match	Rule C-9: (Is ship blacklisted) = no (Security related) = ok Coverage:0.9923 Rule C-10: (Is ship blacklisted) = no (PTMS movement) = match Coverage:0.9734

respectively. To address this aspect, GeMASS has the avenue of employing user-defined rules to pick out non-compliances that are considerably common in reality.

8. Conclusion

This study investigated the leverage of multiple data sources to increase situational awareness and security in the maritime environment. Due to the lack of adequate instances of maritime threats data found in history, this study approached the problem by learning the 'norms'. As the large volume of real-time streaming data related to maritime security poses challenges to port security operation in practice, a machine learning approach was proposed in this study. The research demonstrated that the proposed methodology functions to highlight incoming vessels that bear uncommon characteristics, showing promises in enhancing maritime situational awareness.

In previous studies of anomaly detection for maritime security, knowledge extraction and anomaly detection were based on vessels' movement data (e.g. radar-based datasets). However, there have been other untapped data sources that may hold critical knowledge for detecting security-related anomaly, such as ship classification societies and local port management systems. To that end, this study demonstrated a novel leverage of multiple data sources for maritime security. This work also innovate the use of distance and minimum effective length as fitness measurements, in addition to the coverage and confidence levels commonly applied in the GA methodology.

Future research includes the implementation, deployment and practical assessment of the proposed methodology, for example in conjunction with defense or port security exercises. Further investigations in the proposed fitness measurements of the GA approach will also be of interest. A The core techniques of knowledge extraction, norm models building, and anomaly detection as proposed in this research are expected to be applicable to other problem domains where similar practical challenges exist (e.g. large data volume and the lack of positive cases for signature-driven detection), such as in system network intrusion and terrorism detections.

References

Brax, C., & Niklasson, L. (2009). Enhanced situational awareness in the maritime domain: An agent-based approach for situation management. In Buford, J. F., Jakobson, G., Mott, S., & Mendenhall, M. J. (Eds.), *Intelligent sensing, situation management, impact assessment, and cyber-sensing, proceeding of SPIE* (Vol. 7352, pp. 1–8).

Chong, Y. T., Chen, C.-H., & Leong, K. F. (2009). A heuristic-based approach to conceptual design. *Research in Engineering Design, 20*(2), 97–116.

Edlund, J., Grónkvist, M., Lingvall, A., & Sviestins, E. (2006). Rule based situation assessment for sea-surveillance. In Dasarathy, B. V. (Eds), *Multisensor, multisource information fusion: Architectures, algorithms, and applications, proceedings of SPIE* (Vol. 6242).

Fausett, L. V. (1994). *Fundamentals of neural networks: Architectures, algorithms, and applications*. Englewood Cliffs, NJ: Prentice-Hall.

Fooladvandi, F., Brax, C., Gustavsson, P., & Fredin, M. (2009). Signature-based activity detection based on Bayesian networks acquired from expert knowledge. In *Paper presented at the 2009 12th international conference on information fusion, FUSION 2009* (pp. 436–443).

Goldberg, D. E. (1989). *Genetic algorithms in search, optimisation and machine learning. Reading, mass*. Addison-Wesley.

Jakob, M., Vaněk, O., Urban, Š., Benda, P., & Pěchouček, M. (2010). Employing agents to improve the security of international maritime transport. In *Proceedings of the 6th workshop on agents in traffic and, transportation (ATT2010), May 2010*.

Johansson, F., & Falkman, G. (2007). Detection of vessel anomalies – A Bayesian network approach. In *Paper presented at the proceedings of the 2007 international conference on intelligent sensors, sensor networks and information processing, ISSNIP* (pp. 395–400).

Laxhammar, R. (2008). Anomaly detection for sea surveillance. In *Paper presented at the proceedings of the 11th international conference on information fusion, FUSION 2008*.

Laxhammar, R., & Falkman, G. (2010). Conformal prediction for distribution-independent anomaly detection in streaming vessel data. In *Paper presented at the proceedings of the ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 47–55).

Morik, K. (Ed.). (1989). *Knowledge representation and organization in machine learning*. New York: Springer-Verlag.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning, 1*, 81–106.

Rhodes, B. J., Bomberger, N. A., Seibert, M., & Waxman, A. M. (2005). Maritime situation monitoring and awareness using learning mechanisms. In *Paper presented at the proceedings – IEEE military communications conference MILCOM*.

Ristic, B., Scala, B. L., Morelande, M., & Gordon, N. (2008) Statistical analysis of motion patterns in AIS data: Anomaly detection and motion prediction. In *11th International conference of information fusion*.

Riveiro, M., & Falkman, G. (2011). The role of visualization and interaction in maritime anomaly detection. In *Paper presented at the proceedings of SPIE – The international society for, optical engineering* (Vol. 7868).

Riveiro, M., Falkman, G., & Ziemke, T. (2008). Visual analytics for the detection of anomalous maritime behavior. In *Paper presented at the proceedings of the international conference on information visualisation* (pp. 273–279).

Roy, J. (2008). Anomaly detection in the maritime domain. In *Paper presented at the proceedings of SPIE – The international society for, optical engineering* (Vol. 6945).

Roy, J. (2010). Rule-based expert system for maritime anomaly detection. In *Paper presented at the proceedings of SPIE – The international society for, optical engineering* (Vol. 7666).

Seibert, M., Rhodes, B. J., Bomberger, N. A., Beane, P. O., Sroka, J. J., & Kogel, W., et al. (2006). SeeCoast port surveillance. In *Proceedings of SPIE, Photonics for port and harbor security II 18–19 April* (Vol. 6204). Orlando, FL, USA.

Tecuci, G., & Kodratoff, Y. (Eds.). (1995). *Machine learning and knowledge acquisition: Integrated approaches*. London: Academic Press.

Tun, M. H., Chambers, G. S., Tan, T., & Ly, T. (2007). Maritime port intelligence using AIS data. In *Proceedings of the 2007 RNSA security technology conference Melbourne* (pp. 33–43).

Van Laere, J., & Nilsson, M. (2009). Evaluation of a workshop to capture knowledge from subject matter experts in maritime surveillance. In *Paper presented at the 2009 12th international conference on information fusion, FUSION 2009* (pp. 171–178).

Wong, S. K. M., Ziarko, W., & Li, Y. R. (1986). Comparison of rough-set and statistical methods in inductive learning. *International Journal of Man-Machine Studies, 24*, 53–72.