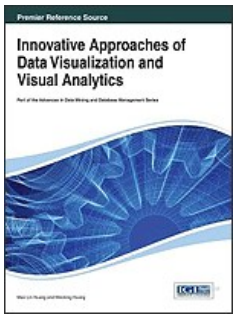

Innovative Approaches of Data Visualization and Visual Analytics

Chapter 7: The Importance of Visualization and Interaction in the Anomaly Detection Process

Maria Riveiro,
Informatics Research Centre, University of Skövde, Skövde
Sweden

ABSTRACT

Large volumes of heterogeneous data from multiple sources need to be analyzed during the surveillance of large sea, air, and land areas. Timely detection and identification of anomalous behavior or any threat activity is an important objective for enabling homeland security. While it is worth acknowledging that many existing mining applications support identification of anomalous behavior, autonomous anomaly detection systems for area surveillance are rarely used in the real world since these capabilities and applications present two critical challenges: they need to provide adequate user support and they need to involve the user in the underlying detection process. Visualization and interaction play a crucial role in providing adequate user support and involving the user in the detection process. Therefore, this chapter elaborates on the role of visualization and interaction in the anomaly detection process, using the surveillance of sea areas as a case study. After providing a brief description of how operators identify conflict traffic situations and anomalies, the anomaly detection problem is characterized from a data mining point of view, suggesting how operators may enhance the process through visualization and interaction.

INTRODUCTION

Exploring, analyzing and making decisions based on vast amounts data are complex tasks that are carried out in a daily basis. People, both in their business and private lives, walk the path from data to decision using diverse means of support. While purely automatic or purely visual analysis methods are used and continued to be developed, the complex nature of many real-world problems makes it indispensable to include humans in the data analysis process.

Automatic analysis methods cannot be applied on ill-defined problems. Furthermore, some real-world problems require dynamic adaptation of the analysis solution, which is very difficult to be handled by automatic means (Keim et al., 2009). Visual analysis methods exploit human creativity, knowledge, intuition and experience to solve problems at hand. While visualization approaches generally give very good results for small data sets, they fail when the required data for solving the problem is too large to be captured by a human analyst (Keim et al., 2009).

The surveillance of large sea areas normally requires the analysis of huge volumes of heterogeneous, multidimensional and dynamic data, in order to improve vessel traffic safety, efficiency and protect the environment (Kharchenko & Vasylyev, 2002). Human operators may be overwhelmed by the data, by the traditional manual methods of data analysis or by other factors, like time pressure, high stress, inconsistencies or the imperfect and uncertain nature of the information. In order to support the operator while monitoring large sea areas, the identification of anomalous behavior or situations that might need further investigation may reduce operator's cognitive load.

While it is worth acknowledging that many existing mining applications support identification of anomalous behavior, autonomous anomaly detection systems for area surveillance are rarely used in real world settings. We claim that anomaly detection systems present, among others, two key challenges: they need to provide adequate user support and they need to involve the user in the underlying detection process. Although these aspects cannot be considered independently, they present distinctive characteristics and demand different solutions. The first challenge concerns the necessity of providing adequate user support during the whole detection and identification of anomalous behavior process, allowing a true discourse with the information. This issue includes deepen our understanding of the human analytical and decision making processes. Due to the fact that anomaly detection is a complex and not a well-defined problem, user involvement is needed. The second challenge involves the study of adequate ways of interacting and visualizing the underlying data mining layers. Human expert knowledge is very valuable in these cases, as it can be used to guide the anomaly detection process, for example, reducing the search space, updating knowledge expert rules or refining normal models derived from the data. We believe that the visualization of the data and the data mining process, as well as the availability of interaction techniques play a crucial role in such involvement.

Thus, this chapter aims to: (1) review anomaly detection methods used in the maritime domain, with specific emphasis on the challenges they present from a user's perspective, (2) discuss the role that visualization and interaction plays in the anomaly detection process, (3) identify leverage points where the use of visualization and interaction could make a positive difference, and (4) present examples of how some of the challenges encountered have been tackled in current research carried out at our research center.

The remainder of the chapter is structured as follows: the following section briefly explores the use of visualization and interaction in data mining. The role of visualization and interaction in maritime anomaly detection is discussed afterwards. Then, a review of relevant anomaly detection approaches applied to the maritime anomaly detection problem is presented. Based on field work carried out at various maritime control centers, we provide a brief description of how maritime operators monitor traffic. Enhancements of the anomaly detection process using visualization/interaction and examples are introduced thereafter. Finally, conclusions are outlined.

THE ROLE OF VISUALIZATION AND INTERACTION IN DATA MINING

Data Mining (DM) is defined as the process of identifying or discovering useful and as yet undiscovered knowledge from the real-world data (Hand et al., 2001). Data mining is often placed in the broader context of Knowledge Discovery in Databases (KDD). KDD is an iterative process consisting of data preparation and cleaning, hypothesis generation (data mining is used basically in this phase) and interpretation and analysis. The CRISP-DM (CRoss Industry Standard Process for Data Mining) model (Shearer, 2000) describes the data mining process

in general, specifying the following phases and tasks: (1) business understanding (determine business objectives, situation assessment, determine data mining goal, produce project plan), (2) data understanding (collect initial data, describe data, explore data, verify data quality), (3) data preparation (data set description, select data, clean data, construct data, integrate data, format data), (4) modeling (select modeling technique, generate test design, build model, asses model), (5) evaluation (evaluate results, review process, determine next steps) and (6) deployment (plan deployment, plan monitoring and maintenance, produce final report). Here, the CRISP-DM is used as a framework to describe the anomaly detection process. Other descriptions of the data mining process can be found in the literature, such as the model presented by Harrison-John (1997), which describes the data mining process as a cyclic process of seven stages: problem definition, data extraction, data cleansing, data engineering, mining algorithm application and analysis of results, where the emphasis is on the data selection and parameter selection tasks.

The integration of DM and information visualization techniques has received a lot of attention in recent years, since automatic data mining approaches only work well for well-defined and specific problems (Kerren et al., 2007). Numerous authors (e.g. Keim [2002] and Fayyad et al. [2002]) recognize the need to tightly include the human in the exploration process.

Visualization can contribute to the data mining process in three ways: it can represent the results of complex computational algorithms, it can depict the data mining process and it can be used to discover complex patterns which cannot be detected automatically but by the powerful human visual system (visual data mining). Visual data mining focuses on integrating the user in the knowledge discovery process using effective and efficient visualization techniques and interaction capabilities. A classification of visual data mining methods regarding data type, visualization technique and the interaction/distortion technique can be found in Keim (2002). Additionally, significant examples of the use of data mining and data visualization can be found in Fayyad et al. (2002). In Meneses and Grinstein (2001), the authors present a description of the data mining process incorporating visualization as a component. Visualization allows users and analysts to interact with several entities involved in the data mining cycle.

Interaction is a core component of the analysis and knowledge discovery process. Users can interact with the data in many different ways (Fayyad et al., 2002): selecting sources of data, browsing, querying, sampling, selecting graphical representations, and so forth. But users may also interact with the underlying data mining process, selecting input parameters, selecting algorithms, validating models, modifying thresholds, and so forth. Nevertheless, examples of interactions between users and entities that are part of any data mining process are not common in the literature.

THE ROLE OF VISUALIZATION AND INTERACTION IN ANOMALY DETECTION

Anomaly detection methods have been used in multiple areas, like network security, video surveillance, human activity monitoring, etc. The majority of published work on anomaly detection focuses on the technological aspects: new and combinations of methods, additional improvements of existing methods, reduction of false alarms, correlations among alarms, etc. Publications regarding the use of anomaly detection methods in real environments or human factors studies regarding anomaly detection are scarce. Even if interaction, usability, cognitive task analysis or acceptability are not normally matters within anomaly detection research, visualization has received more attention.

The majority of the examples regarding the use of visualization to enhance anomaly detection are published in the area of network security. Even though an exhaustive review on the use of visualization for network security is out of the scope of this chapter, we outline here some examples where visualization has been used for enhancing the anomaly detection process.

Axelsson (2005) addresses the problem of false alarms within intrusion detection and proposes four different visualization approaches to aid the operator to correctly identify false (and true) alarms. Likewise, Mansmann (2008) devotes his dissertation to the use of visualization for monitoring, detecting and interpreting security threats. New scalable visualization metaphors for detailed analysis of large network time series are presented: a hierarchical map of the IP address space, graph-based approaches for tracking behavioral changes of hosts and higher-level network entities and the application of Self Organizing Maps (SOMs) to analyze both structured network protocol data and unstructured information, e.g., textual context of email messages. Other examples of novel visualization approaches for network traffic that support intrusion detection are presented in Onut et al. (2004), Teoh et al. (2004), Muelder et al. (2005), Livnat et al. (2005), and Cai and de M. Franco (2009).

Onut et al. (2004) present two types of graphical views for information extracted at the network layer: services behavior view (behavior of the internal/external hosts with respect to a certain set of services) and category view (hosts are sorted with respect to a particular relevant attribute, like number of IPs used). In Teoh et al. (2004), the authors describe an integration of visual and automated data mining methods for discovering and investigating anomalies in Internet routing. The analysis tool presents different components that complement each other, where visualization and interaction are key to support user involvement. Muelder et al. (2005) employ visualization to detect scans interactively, while Livnat et al. (2005) suggest a novel paradigm for visual correlation of network alerts from disparate logs, that facilitates and promotes situational awareness in complex network environments. This approach is based on the notion that an alert must possess three attributes, namely, what, when, and where. Cai and de M. Franco (2009) exploit both interaction and visualization to reveal real-time network anomalous events. Glyphs are defined with multiple network attributes and clustered with a recursive optimization algorithm for dimensional reduction. The user's visual latency time is incorporated into the recursive process so that it updates the display and the optimization model according to a human-based delay factor.

Despite the extensive number of examples of the application of visualization to anomaly detection in network security, few examples exist outside this domain. An exception is the work presented in Iwata and Saito (2004), where a new anomaly detection method that visualizes data in 2- or 3-dimensional space based on the probabilities of belonging to each component of the model and the probability of not belonging to any component, anomaly, is proposed. For evaluation purposes, the method is applied to an artificial time series.

ANOMALY DETECTION METHODS FOR MARITIME TRAFFIC

It is hard to clarify what exactly anomaly detection means. Anomaly is a many-sided concept and it is normally associated with terms like abnormal, unusual, irregular, rare, deviation, strange, illegal, threat, atypical, inconsistent, etc. Many data mining techniques analyze data in order to find behavioral anomalies. Behavioral anomalies are defined as deviations from the normal behavior. Here, an anomaly is defined

from a user (operator or organization) point of view, as events or situations that need to be detected and identified (see Riveiro et al. [2009] for a detailed discussion). A classification and examples of sea traffic anomalies from operators and practitioners point of view is provided in Roy (2008).

Most of the published work regarding anomaly detection, as previously shown, relates to intrusion detection applications for network traffic. Algorithms used in the detection of intrusions/attacks are traditionally classified in three main groups (Patcha & Park, 2007): anomaly (referring only to data-driven approaches), signature or hybrid. Systems based on anomaly detection schemes (data-driven approaches) look for abnormalities in the traffic, assuming that something that is abnormal is probably suspicious. Such detectors are based on what constitutes normal behavior and what percentage of the activity we want or are allowed/willing to flag as abnormal. Signature-based approaches look for predefined patterns in the data. Hybrid approaches combine data and knowledge driven approaches.

In the civil security domain, anomaly detection is not as mature as it is the network security arena. To the best of our knowledge, anomaly detection and behavioral analysis approaches applied to sea surveillance have not been covered in previously published anomaly detection reviews and, in particular, no review includes any analysis regarding human factors.

This section presents a review of anomaly detection approaches for sea surveillance. The objective is to analyze where human involvement is needed and how visualization and interaction might facilitate anomaly detection. The classification and description of each method includes information regarding: (1) detection method (data or knowledge driven), (2) nature of data analyzed, and (3) usage frequency (real-time continuous monitoring or periodic analysis). Moreover, for each method, we provide a brief description of its fundamentals and analyze the following aspects (if they apply): (1) input parameters, (2) normal model and rule set, (3) a description of the detection process, (4) output, and (5) explanation of the detections.

Data-Driven Anomaly Detection

In this category, approaches used within maritime anomaly detection can be classified as statistical (parametric and non-parametric) and machine learning based (e.g. Bayesian networks, neural networks or clustering techniques).

Statistical Parametric

Kraiman et al. (2002) present an anomaly detector processor, which exploits multisensor tracking and surveillance data to identify interesting events. The authors demonstrate the detector within a Vessel Traffic Service (VTS) environment, using input data regarding vessel type, speed, location, report time and heading, as well as environmental information such as tides, wind speed and direction (nonetheless, examples shown in the article are limited to position and speed values). The detection approach is a statistical parametric method, based on a combination of SOMs and Gaussian Mixture Models (GMM). The parameters of the Gaussian distributions (mean and covariance matrices) can be estimated from the available training data using SOMs. Each node of the grid is characterized by an N-dimensional Gaussian probability function, where the means are given by the final values of the nodes and the variances are given by the dispersion of the training data around each node. Therefore, the baseline profile or normal model is a multidimensional likelihood function that it is used to estimate the probability value of a new observation. Over the likelihood, Bayes' rule is applied to calculate the probability value of obtaining such observation. In order to do so, the user must introduce the percentage of the training data that is anomalous (an important input parameter).

The detector based on this approach presents a Graphical User Interface (GUI) to facilitate operator interaction. Even if the functionality of the GUI is not described in Kraiman et al. (2002), the following input parameters can be determined: attributes used during the training phase, weight of the attributes, characteristics of the SOM (number of nodes and training radius), percentage of training data that is anomalous, threshold for reported anomalies and width of the temporal window for cumulative probability calculation. The output consists of a plot of cumulative probability of anomaly versus time and a characterization of the anomaly, explanation (showing in percent how the different attributes have contributed to the anomaly). Anomalous vessels are displayed in red over the geographical area. The detector was trained over one week of traffic data, but no information regarding the performance of the detector is given.

The method described in Laxhammar (2008) is similar to Kraiman et al.'s approach, but in this case the normal model representing vessel behavior is built using a combination of a greedy version of the Expectation-Maximization (EM) algorithm and GMM. Here, EM is used to estimate the parameters, mean and covariance, needed to combine the Gaussian distributions. Since the classical EM algorithm is very sensitive to initialization (it may converge to a local optimal solution different from the global) a greedy version is proposed. Instead of starting randomly, the greedy EM builds the optimal mixture model adding new components one at a time (support for such initialization and components weights are input parameters). Another input parameter is the maximum number of mixture components. The method is tested over real maritime traffic data from Swedish waters, where position, speed and course are considered. Latitude and longitude are discretized. One week of data was used for training and one week for validation (EM requires a validation set during training).

Statistical Non-Parametric

Ristic et al. (2008) present a statistical non-parametric analysis of vessel motion patterns, in ports and waterways, using Automatic Identification System (AIS) data. The detection is carried out using adaptive Kernel Density Estimation (KDE). The variables used are position (two dimensions) and velocity (two dimensions). The suggested solution assumes that the AIS data has been preprocessed and patterns have been extracted (these patterns constitute the baseline used during the detection process).

The normal model is, thus, a collection of motion patterns extracted from historical AIS data. The necessary input parameters (even if they are not specifically pointed out in the paper) are type of kernel ('normal' is usually the default value), smoothing parameter (bandwidth of the kernel-smoothing window) and threshold value. Threshold determines the probability value of an alarm, establishing the border between two hypotheses (normal, H0, or abnormal vessel behavior, H1). The output is the outcome of the classification (H0 or H1).

The existing publications concerning this method do not contain information regarding how to create the normal model or baseline. It is problematic to define motion patters, since there are multiple origins, destinations and connections paths in maritime traffic data. Moreover,

non-parametric methods like KDE require large amounts of representative data of normal behavior, compared to traditional parameterized approaches.

Clustering and Outlier Detection

Vessel motion baseline profiles can also be built considering trajectories. Similar vessel trajectories are grouped thereby modeling regular traffic routes. Deviations from such routes are considered anomalous (Euclidean distances between clusters and trajectories may be used as metrics). An example application in the maritime domain is the work presented in Dahlbom and Niklasson (2007). The authors focus on the use of a trajectory clustering algorithm over maritime traffic in order to create normal sea lanes, not on the problem of detecting anomalies. Simulated radar readings of vessel traffic along the southern coast of Sweden are used in the experiments. The authors discuss the problems the clustering algorithm presents regarding matching incoming trajectories to clusters. The authors argue that prefix matching is not suitable for coastal surveillance and propose the use of splines.

Rhodes' research group (BAE Systems) has extensively studied the problem of learning normal vessel motion patterns (see Rhodes et al. (2005); Bomberger et al. (2006); Rhodes et al. (2007a)). The presented approaches are applied to harbor areas and both simulated and real AIS data are analyzed. Position (latitude and longitude) and velocity (course and speed) are considered. The discretization of both features (position and velocity) is necessary. The system takes real-time tracking information and uses continuous on-the-fly learning that enables concurrent recognition of patterns of current motion states. In Rhodes et al. (2005), the learning approach combines an unsupervised clustering algorithm (Fuzzy ARTMAP neural network) and a supervised mapping and labeling algorithm. Extensions of this approach can be found in Bomberger et al. (2006); Rhodes et al. (2007a). Even if the authors claim that operator intervention is not necessary, they agree that operators or analysts can help teaching the model via simple point and click actions, increasing the speed and performance of the learning phase.

Bayesian Inference

A Bayesian Network (BN) is a graphical model that encodes probabilistic relationships among variables of interest (Patcha & Park, 2007). The graphical model conveys information regarding causal relations and interdependencies between variables. A BN is a suitable approach to anomaly detection, since it can be used when there is a need to combine prior knowledge with data (Patcha & Park, 2007). Moreover, due to their transparency, human domain experts are able to validate and improve BNs.

An example of the application of BN to the maritime anomaly detection problem is provided in Johansson and Falkman (2007). Synthetic data is used during the experimental phase (simulated radar readings). The variables used are *x, y, heading, speed, Δheading, Δspeed and vesseltype*. The feature space is discretized. The BN represents the underlying probability distribution of the data, assuming that we can construct such representation. Based on the data, first the structure of the graph is built and then the conditional probabilities are estimated. Two important input parameters are, as in other approaches, the size of the window (number of most recent samples) that averages the probability value over time and the threshold used to flag an alarm (balance between the detection rate, recall, and the precision, false positives). The normal model is thus the BN built from data and the output is a joint probability value $P(x,y,heading,speed,\Delta speed,\Delta heading,vesseltype)$. When anomalous behavior is detected using this approach, no further information or explanation is provided, meaning that no feature or group of features are suggested as rationale behind alarms.

Knowledge-Driven Anomaly Detection

The majority of the few anomaly detection capabilities implemented in real maritime control centers are rule (signature or misuse) based systems. Such systems allow operators to create simple rules that will trigger an alarm (e.g. IF <vessel in shallow waters> TEHN <danger of grounding>).

Initial steps to more elaborate anomalous situation detector, i.e. combinations of events over time is presented in Edlund et al. (2006). Based on an agent framework and using an ontology geared toward sea surveillance, the authors described a rule-based situation assessment system that analyzes situations developing over time. Rules are created by experts using the rule editor agent. In order to create new rules, experts select known objects from a list, choose their relation (approaching, leaving, inharborarea) and connect them in time. The ontology is based on a previously published core ontology for situation awareness. No user interaction with the ontology is supported. The rule editor GUI and the detector, reasoner agent, are under development.

Hybrid Approaches

Hybrid approaches to anomaly detection combine both data-driven and knowledge-based methods, overcoming some of the drawbacks of each particular method (high false alarm rate in the data-driven case and the possibility of detecting only known patterns in the knowledge-based case). An example of a compound approach to the maritime anomaly detection problem is the detector implemented in *SeeCoast* (Seibert et al., 2006). The detector applies rule-based and learning-based pattern recognition algorithms to alert illegal, threatening and anomalous vessel activities. SeeCoast extends the detection capability of the learning-based pattern module described above (see Bomberger et al. (2006); Rhodes et al. (2007a,b)) using a rule-based track activity analysis. The rule-based component implements a three-stage approach to rule-building and matching: domain modeling, pattern definition and pattern matching. In the domain modeling stage, an ontology is built describing the data sources and the attributes of data reports (e.g., fused tracks as a data type, with velocity as a data field). In the pattern definition stage, operators use a GUI to create patterns based on the ontology (e.g. <any track whose location is within a restricted area>). A GUI allows operators to script patterns, walking the operator through a series of selections and questions that use information about the data environment to simplify the process. Operators can also create patterns from templates that only require specification of key inputs. In the pattern matching stage, operators select a set of patterns to be monitored for and the system then generates alerts for matching instances. A snapshot of the flagged vessel assists the operator deciding on further actions (offering thus, explanation capabilities).

SeeCoast is a complex and powerful port security and monitoring system that besides the anomalous detector module includes video processing to detect, classify and track vessels; multi-sensor track correlation of video track data with radar and AIS tracks; ship size

classification, display enhancements for improved situational awareness and forensic analysis.

HOW DO EXPERTS MONITOR MARITIME TRAFFIC?

In maritime transportation, traffic control is carried out by both coastal and port Vessel Traffic Services (VTS), whose centers aim to improve vessel traffic safety and efficiency, safeguard human life at sea, as well as protect the maritime environment, adjacent shore areas, work sites, and offshore installations from the possible adverse effects of marine traffic. Three maritime control centers were visited during our field work. Such centers offer their services 365 days/year and 24 hours/day. The essential sources of data used for monitoring maritime traffic are radar data, Automatic Identification System (AIS) messages, VHF radio, Closed Circuit TV cameras (CCTV), harbor planning and administrative information, data bases with historical information about the vessels, telephone and fax, and meteo/hydro equipment (weather reports and marine currents information). The VTS operators interviewed have lengthy maritime and seagoing experience and receive education in accordance with the International Association of Marine Aids to Navigation and Lighthouse Authorities guidelines.

VTS operators use various surveillance systems. The systems are customized for each center and display real-time radar and AIS data (sometimes referred to as ‘common operating maritime picture’) that serve as a basis for carrying out main tasks such as monitoring and information services. The visualization and interaction capabilities of the systems used are quite limited. The main visualization consists of a geographical map where vessels are displayed using different icons and colors. Speed vectors and navigational information are displayed over the background map. Other graphical representations are not provided (no abstract representations, links between entities, or 3D visualizations are available). Selection and zooming in/out are provided as interaction methods. The systems used at the VTS centers allow some manual identification of anomalies. For example, the operators can make queries that show all the vessels exceeding a certain speed value or crossing a particular borderline. These functionalities, which may be considered anomaly detectors, are rarely used, since they must be carried out manually (operators stated that they are time consuming procedures) and do not cover many of the situations the operators are interested in detecting (see Figure 1).

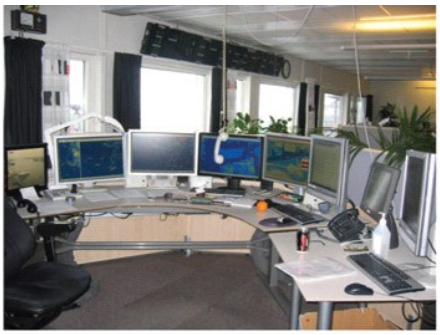

Figure 1: VTS Gothenburg, Sweden The figures depict two working areas in the control room, illustrating environment, tools, and systems used

VTS operators need the timely identification of possible traffic-conflict situations emerging in the surveyed area, and respond appropriately. Examples of such situations are vessel collisions and groundings in the port and entrance areas. Moreover, personnel interviewed in these centers would appreciate support in detecting vessels navigating through restricted zones, vessels not following the established sea lanes, vessels not following the normal route with regard to the reported destination, cargo of special interest, vessels carrying dangerous gods sailing close to passenger ships or protected areas, vessels with a history of being involved in illegal activities, suspicious flag or port, fishing or recreational craft approaching traffic separation zones, etc.

Despite slight differences among the three visited centers, the actual process of finding anomalous behavior and conflict situations can be summarized in five stages: (1) overview (monitor and explore): continuous control of the traffic in real-time, using radar, VHF radio, and AIS information; (2) if something is unusual or unfamiliar (operators normally base their judgment on their experience), detailed information must be obtained, like zooming into the area and starting VHF radio communication with vessel of interest; (3) waiting time: operators usually wait a reasonable period of time, observing how the situation develops. At this stage, operators might listen to VHF radio communications among vessels, to increase their understanding of the situation; (4) more detail (focus): if the situation has not become normal, they intensify the dialog with the vessel of interest or try to obtain more data using, for example, additional information stored in data bases; (5) taking action: if they

believe that an incident has occurred, they take action, alerting other organizations and reporting the situation.

This basic pattern, or loop, describes the typical overall process. Operators move back and forth between these stages, for example, between stage 3 (waiting time) and 4 (more detail). The stages vary in length and some stages include several sub-loops.

USING VISUALIZATION AND INTERACTION IN ANOMALY DETECTION

Considering the insights gained during our visits to maritime control centers and the review presented in one of the previous sections, the anomaly detection process can be divided in: on-line and off-line processing (see Figure 2). On-line processing refers to the analysis in real-time of the incoming data, whereas the off-line processing refers to the establishment or normal models from (training) data and rules that are used during the on-line detection process. Both processes resemble typical data mining cycles and are, obviously, interconnected.

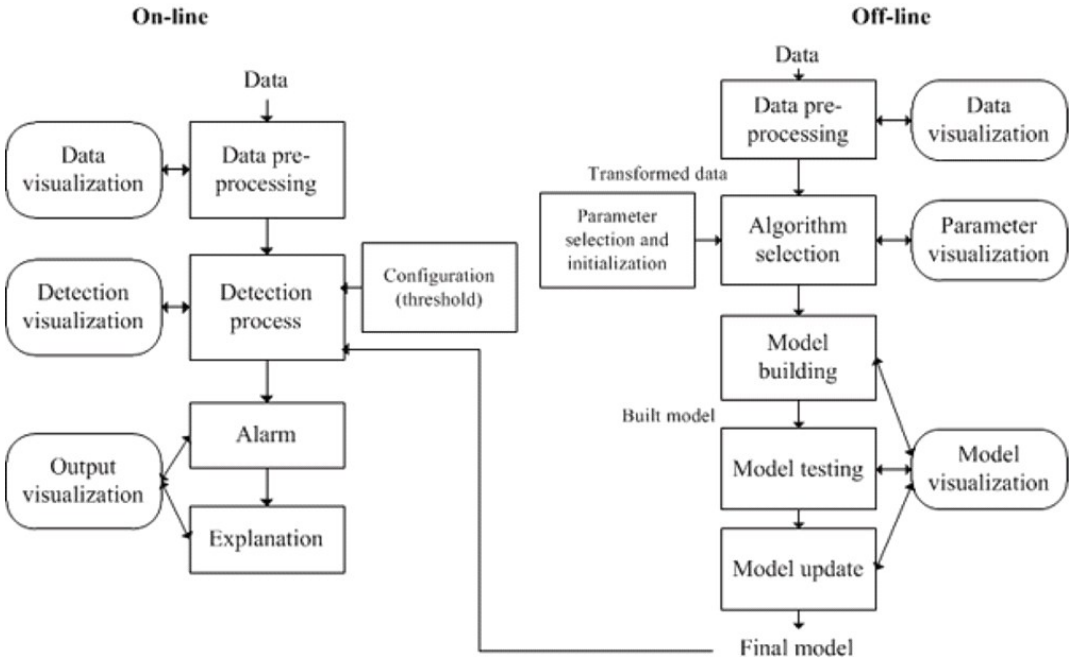

Figure 2: Using visualization and interaction to support user involvement in the anomaly detection process

We argue that visualization and interaction is key to improving anomaly detection performance in general, and in particular, visualization and interaction are key to perform an adequate analysis of the data, construct understandable normal models, update and validate such models and create useful and comprehensible output, that can not only generate suitable responses from operators but also improve the whole anomaly detection process. Figure 2 points where visualization and interaction could make a positive difference.

Data Visualization

Data visualization supports the understanding of the data and the interaction between the analyst/operator and the dataset during the preprocessing phase. There is a wide variety of techniques to visualize both low and multidimensional datasets (e.g. pixel-based techniques, scatter-plots, parallel coordinates, geometric projections and icon-based methods). Keim (2002) reviews and provides a classification of visualizations based on the data type to be visualized, the visualization technique and the interaction and distortion technique.

In order to select appropriate visualization techniques, it should be taken into account the spatial and temporal nature of the information in the maritime domain. An interesting example of the visualization of vessel tracks is the work presented in Willems et al. (2009). Analyzing AIS data, the authors present an overlay map that show where sea lanes, anchoring zones or slow moving vessels are located.

The visualization of the data may also support the analyst while cleaning, selecting and transforming the data. A common problematic phase that influences the detector performance in many of the anomaly detection methods reviewed (see Laxhammar (2008); Bomberger et al. (2006)) is the discretization and normalization of the feature space. Proper representations of the data regarding how the discretization affects the construction of the baseline behavior and how samples are distributed over the feature space are needed. Moreover, other aspects that should be considered in this case are, for example, how to represent inconsistencies in the data, uncertainty, quality, reliability, etc.

Parameter Visualization

Parameter visualization supports the interaction between the analyst/operator and the process of selecting, tuning and optimizing input values to the on-line and off-line processes involved. Parameter selection and tuning requires the exploration of several alternatives (Meneses & Grinstein, 2001) and it is a complex optimization problem.

Statistical anomaly detection methods require the selection and tuning of multiple parameters (e.g. learning rates, type of kernel function, smoothing values and number of Gaussian mixtures). The reviewed approaches do not make clear the correlation between domain features and parameter setting values, and parameters seem to be selected in a more or less ad hoc manner. One arduous task in all the reviewed

anomaly detection methods is tuning the anomaly threshold. The threshold value balances the detection rate (recall) and the number of false positives or false alarms (precision). Another delicate matter is the selection of the sliding window size that averages probability/likelihood values that are compared to the threshold. If the window size is too small, the system will be sensitive to data or sensor error, while a too large value may hide anomalies.

Visualization and interaction can be used to understand the parameter selection and tuning optimization process for a particular dataset, providing comprehensible views of the impact that these steps have in the final detection stages. Unfortunately, the visualization of parameter selection and tuning processes has been mainly overlooked by the anomaly detection research community (an exception is the work presented in Meneses and Grinstein [2001]).

Model Visualization

Model visualization supports the comprehension and interaction with normal models and rules embedded in the system. Such visual representations may support the creation, validation and update phases. The analyst/operator may be able to compare models, communicate them to colleagues and evaluate if they match his/her understanding of the world.

The representation of normal models built from data has hardly received any attention by the research community. An exception is Rheingans and desJardins (2000), where the authors describe a set of visualization methods that help users to understand and analyze the behavior of learned models (the article focuses on classification tasks using BN).

Knowledge-based approaches normally use a set of rules that represent situations that are of interest to an analyst/operator (unlike data-driven approaches, these signatures represent the ‘anomalous’ behavior). Visual and interactive representations of rules provide a natural way of understand, create, validate, update and prune them. In opposition to the lack of proposals regarding visualizations of induced models, extensive work has been done on the representation of rules (most of the publications refer to signatures extracted from large data sets). For example, a framework for mining and analyzing large rule sets through visualization is presented in Bruzzese and Davino (2008). During our review on knowledge-based approaches, another important matter related to the creation of rules that may benefit from the use of visualization and interaction is the necessity of constructing proper ontologies that represent objects, concepts, events and relationships (see Seibert et al. (2006)).

Detection Visualization

Detection visualization supports the understanding of the whole process, from data to alarms.

This process is a continuous hypothesis generation and testing cycle that involves all the aspects previously seen. An example of the visualization of the detection process can be seen in Kraiman et al. (2002). The GUI shows data, probability values vs. time, alarms and explanations.

Outcome Visualization

Outcome visualization refers to the representation of triggered alarms. Visual representations of alarms should support their analysis, in order to find, for example, correlations among them. Monitoring generated alarms is normally a challenging activity. In our visits to maritime centers, operators have highlighted the necessity of keeping interactive lists of alarms (ordered by importance). Operator response to the generated alerts (acknowledgment/rejection) may be used as a teaching signal to the detector, refining thus, its performance.

Explanations

Jensen et al. (1995) claim that decision support systems should have features for explaining how they have come up with their recommendations in order to support the decision maker as well as increase his/her confidence in the system. The ability of explaining the reasoning behind an alarm is of great importance in order for an operator to fully accept the advice the system provides. Despite this fact, the amount of research devoted to this subject is relatively sparse and most of the reviewed work does not tackle this issue. One of the reasons is that it might be difficult to point out which features have triggered the alarm or it might be difficult to construct or communicate the evidence. For example, the outcome of data-driven statistical approaches is normally a probability value per observation that represents P(features considered). It might not be possible to point out which feature or features are the cause of the alarm. In this case, we need additional methods that investigate further the outcomes generated.

Limitations

A relevant aspect that we have not discussed during the analysis of the anomaly detection process is the different users and roles that might use the anomaly detection capability. The daily operator (that monitors on-line traffic) might not be able to create or update normal models or rules, due to time constraints, policies or lack of background knowledge. On the other hand, analysts may be able to maintain and configure the system regarding their off-line workings, selection and tuning of parameters, update models, thresholds, etc.

EXAMPLES

In this section we illustrate some of the aspects discussed in the previous section with examples from our own research within the maritime domain. The objective of this section is to demonstrate how visualization can enhance the anomaly detection process, focusing on some of the steps of the process presented above.

The first example illustrates how parameter and model visualization can support the selection of methods that match the problem. Figure 3

(inspired by the study presented in Laxhammar et al. (2009)) shows how two different anomaly detection approaches model two parallel vessel trajectories. The left peak is calculated using GMM and the right peak is calculated using KDE. The GMM peak is unimodal, hiding the separation between the parallel sea lanes while the bimodal KDE satisfactorily captures the separation between them. Hence, we may conclude that KDE method models more accurate the vessel trajectories analyzed.

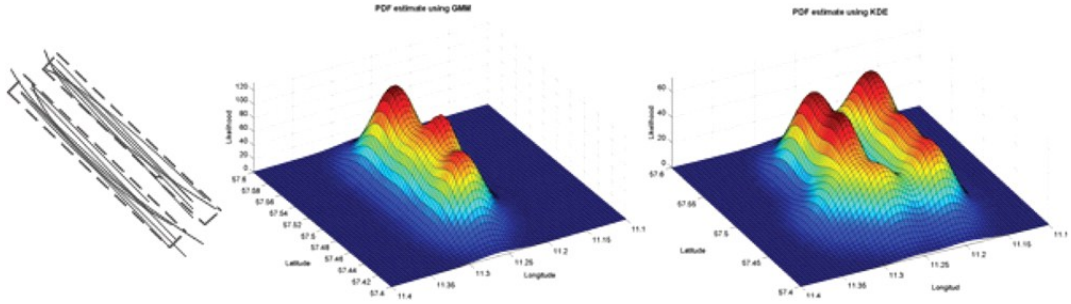

Figure 3: Two parallel vessel trajectories and their model estimations using GMM and KDE probability density functions. A full comparison between these two methods can be read in Laxhammar et al. (2009)

The second example concerns model visualization. Figure 4 presents a visualization of normal vessel behavioral models built from real AIS data. In order to build such model, we have used a statistical method that combines SOM and GMM. The biggest challenge we have faced while trying to represent the normal model was the fact that we would need an eight-dimensional space to represent such probability density function (we use eight vessel features: position, speed, course over ground, heading, length, width and draught). We projected the probability function over a 2 dimensional map. High values of probability are represented in red, while blue represents lower probability values. These visualizations allow comprehension of normal vessel behavior built from data, supporting validation and improvement of such models.

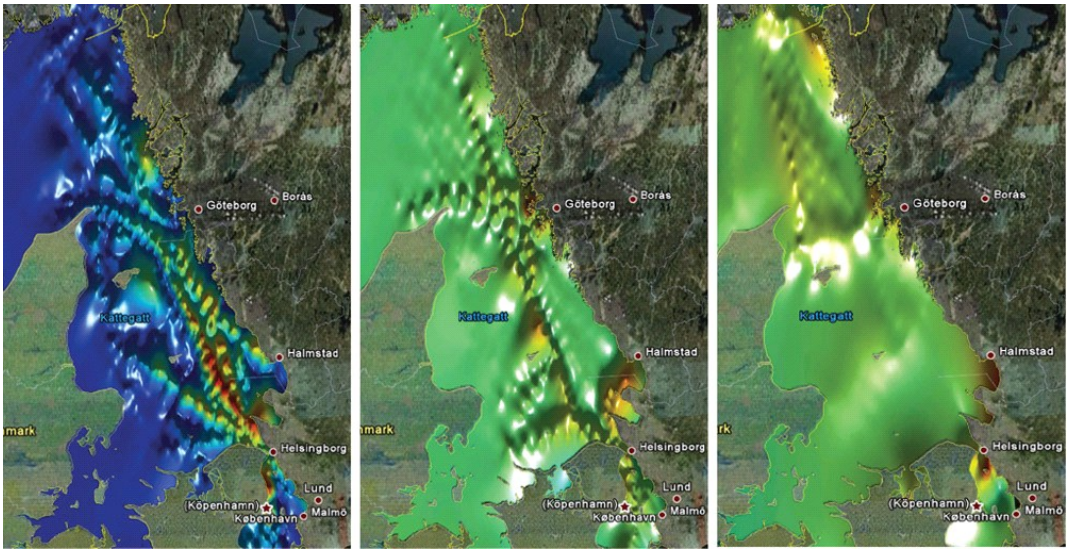

Figure 4: Visualizations of normal behavioral models for cargo (left), tanker (middle) and passenger (right) vessels. The models are calculated using a combination of SOMs and GMMs from real AIS data along the Swedish west coast. The following features are considered: position, speed, course over ground, heading, length, width and draught. The probability values are projected over a geographical map (Google Earth)

The last example, Figure 5, shows how explanations can be visualized using trees. In this case, BNs are used to find anomalous vessels hidden in AIS data. In order to generate comprehensible explanations from BNs outcomes, we have tested two algorithms, Explanation Tree and Causal Explanation Tree. The tree in Figure 5 shows which of the features are selected as best causes of anomaly when the BN pointed a vessel as suspicious (in this particular case, the abnormality hidden is a vessel speeding in a slow moving area). More details on the application of these algorithms can be found in Helldin and Riveiro (2009).

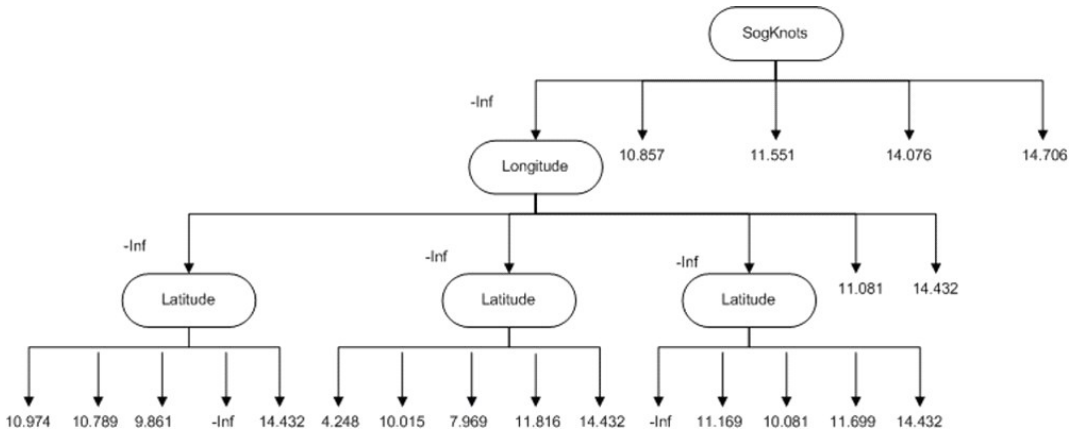

Figure 5: Explanation visualization: A Conditional Explanation Tree (CET) explains the inference made by a BN

CONCLUSION

Current anomaly detection capabilities and tools provide very limited possibilities to incorporate any expert knowledge or any user input at all. In our opinion, designers and developers underestimate the benefits of human involvement in the anomaly detection process. The necessity of such involvement can be seen from two perspectives. Firstly, anomaly detection systems for sea surveillance are not used autonomously in the real world. We need to provide adequate support for human decision makers, making transparent and trustworthy the anomaly detection process. Secondly, since the anomaly detection problem is hard to solve in an automatic manner (it normally generates high number of false alarms due to its complexity), we need to include expert knowledge in the loop in order to improve detector's performance.

Based on a review of anomaly detection methods applied to maritime traffic data, this chapter examines the anomaly detection process, highlighting where visualization and interaction can be used to support human involvement, thus, enhancing the process. The analysis presented here may inform the design of future anomaly detection systems when fully automatic approaches are not viable and human participation is needed. We would like to facilitate the design of interfaces that support human involvement and are properly integrated in the overall KDD process. The feedback that analyst/operator can provide to these processes can hardly be obtained by other means.

REFERENCES

Axelsson, S. (2005). *Understanding Intrusion Detection Through Visualization*. (Ph.D. thesis). Goteborg, Sweden: Chalmers University of Technology.

Bomberger, N., Rhodes, B., Seibert, M., & Waxman, A. (2006). Associative learning of vessel motion patterns for maritime situation awareness. In *Proceedings of 9th International Conference on Information Fusion*. New Brunswick, NJ: IEEE Press.

Bruzzese, D., & Davino, C. (2008). Visual mining of association rules . In *Visual Data Mining* (103–122). Berlin: Springer-Verlag. doi:10.1007/978-3-540-71080-6_8

Cai, Y. & de M. Franco, R. (2009). Interactive visualization of network anomalous events. In: *Computational Science, 5544*, 450–459. Berlin: Springer.

Dahlbom, A., & Niklasson, L. (2007). Trajectory clustering for coastal surveillance. In *Proceedings of the 10th International Conference on Information Fusion*. QC, Canada: IEEE Press.

Demšar, U. 2006. *Data Mining of Geospatial Data: Combining Visual and Automatic Methods*. (Ph.D. thesis). Stockholm, Royal Institute of Technology (KTH).

Edlund, J., Gronkvist, M., Lingvall, A., & Sviestins, E. (2006). Rule-based situation assessment for sea surveillance. In *Proceedings of SPIE Conference on Multisensor, Multisource Information Fusion: Architectures, Algorithms and Applications, 624,* 1–11. Bellingham, WA: SPIE Press.

Fayyad, U., Grinstein, G., & Wierse, A. (Eds.). (2002). *Information visualization in data mining and knowledge discovery*. San Francisco: Morgan Kaufmann Publishers Inc.

Hand, D. J., Mannila, H., & Smyth, P. (2001). *Principles of data mining. Adaptive computation and machine learning*. Cambridge, MA: The MIT Press.

Harrison-John, G. (1997). *Enhancements to the Data Mining Process*. (Ph.D. thesis). Stanford, CA, Stanford University.

Helldin, T., & Riveiro, M. (2009). Explanation methods for bayesian networks: review and application to a maritime scenario. In: *3rd Annual Skövde Workshop on Information Fusion Topic,* 11–16. New Brunswick, NJ: IEEE Press.

Iwata, T., & Saito, K. (2004). Visualization of anomaly using mixture model . In *Knowledge-Based Intelligent Information and Engineering System*, 624–631. Berlin: Springer. doi:10.1007/978-3-540-30133-2_82

Jensen, F., Aldenryd, S., & Jensen, K. (1995). Sensitivity analysis in bayesian networks . In *Symbolic and Quantitative Approaches to Reasoning and Uncertainty*, 243–250. Berlin: Springer. doi:10.1007/3-540-60112-0_28

Johansson, F., & Falkman, G. (2007). Detection of vessel anomalies–A bayesian network approach. In *Proceedings of the 3rd International Conference on Intelligent Sensors, Sensor Networks, and Information Processing.* New Brunswick, NJ: IEEE Press.

Keim, D. (2002). Information visualization and visual data mining. *IEEE Transactions on Visualization and Computer Graphics*, *7*(1), 1–8. doi:10.1109/2945.981847

Keim, D. A., Mansmann, F., & Thomas, J. (2009). Visual analytics: How much visualization and how much analytics. *SIGKDD Explorations*, *11*(2).

Kerren, A., Stasko, J., Fekete, J.-D., & North, C. (2007). Workshop report: Information visualization–human-centered issues in visual representation, interaction, and evaluation. *Information Visualization, 6,* 189–196.

Kharchenko, V., & Vasylyev, V. (2002). Application of the intellectual decision making system for vessel traffic control. In *Proceedings of 14th International Conference on Microwaves, Radar, and Wireless Communications, 2*, 639–642. New Brunswick, NJ: IEEE Press.

Kraiman, J. B., Arouh, S. L., & Webb, M. L. (2002). Automated anomaly detection processor. In Sisti & Trevisani (Eds.), *Proceedings of SPIE: Enabling Technologies for Simulation Science VI* (128–137). Bellingham, WA: SPIE Press.

Laxhammar, R. (2008). Anomaly detection for sea surveillance. In *Proceedings of the 11th International Conference on Information Fusion*, 47–54. Cologne, Germany: IEEE Press.

Laxhammar, R., Falkman, G., & Sviestins, E. (2009). Anomaly detection in sea traffic-A comparison of the gaussian mixture model and the kernel density estimator. In *Proceedings of the 12th International Conference on Information Fusion,* 756–763. New Brunswick, NJ: IEEE Press.

Livnat, Y., Agutter, J., Moon, S., Erbacher, R. F., & Foresti, S. (2005). A visual paradigm for network intrusion detection. In *Proceedings of the 2005 IEEE Workshop on Information Assurance and Security*, 92–99. New Brunswick, NJ: IEEE Press.

Mansmann, F. (2008). *Visual Analysis of Network Traffic: Interactive Monitoring, Detection, and Interpretation of Security Threats*. (Ph.D. thesis). Konstanz, Germany, Universität Konstanz.

Meneses, C. J., & Grinstein, G. G. (2001). Visualization for enhancing the data mining process. [Bellingham, WA: SPIE Press.]. *Proceedings of the Society for Photo-Instrumentation Engineers*, *4384*, 126–137. doi:10.1117/12.421066

Muelder, C., Ma, K.-L., & Bartoletti, T. (2005). Interactive visualization for network and port scan detection . In *Proceedings of 2005 Recent Advances in Intrusion Detection*, 1–20. New Brunswick, NJ: IEEE Press.

Onut, I. V., Zhu, B., & Ghorbani, A. A. (2004). A novel visualization technique for network anomaly detection. In *Proceedings of the 2nd Annual Conference on Privacy, Security, and Trust*, 167–174. New York: ACM Press.

Patcha, A., & Park, J.-M. (2007). An overview of anomaly detection techniques: Existing solutions and latest technological trends. *Computer Networks*, *51*(12), 3448–3470. doi:10.1016/j.comnet.2007.02.001

Rheingans, P., & desJardins, M. (2000). Visualizing high-dimensional predictive model quality. [New Brunswick, NJ: IEEE Press.]. *Proceedings of IEEE Visualization*, *2000*, 493–496.

Rhodes, B., Bomberger, N., Seibert, M., & Waxman, A. (2005). Maritime situation monitoring and awareness using learning mechanisms. *Military Communications Conference, 1,* 646–652. New Brunswick, NJ: IEEE Press.

Rhodes, B., Bomberger, N., & Zandipour, M. (2007a). Probabilistic associative learning of vessel motion patterns at multiple spatial scales for maritime situation awareness. In: *10th International Conference on Information Fusion*, 1–8.

Rhodes, B. J., Bomberger, N. A., Zandipour, M., Waxman, A. M., & Seibert, M. (2007b). Cognitively-inspired motion pattern learning & analysis algorithms for higher-level fusion and automated scene understanding. In *Military Communications Conference (MILCOM 2007)*, 1–6. New Brunswick, NJ: IEEE Press.

Ristic, B., Scala, B. L., Morelande, M., & Gordon, N. (2008). Statistical analysis of motion patterns in AIS data: Anomaly detection and motion prediction. In *Proceedings of 11th International Conference of Information Fusion*. New Brunswick, NJ: IEEE Press.

Riveiro, M., Falkman, G., & Ziemke, T. (2008). Improving maritime anomaly detection and situation awareness through interactive visualization. In *Proceedings of 11th International Conference on Information Fusion*, 47–54. New Brunswick, NJ: IEEE Press.

Riveiro, M., Falkman, G., Ziemke, T., & Kronhamn, T. (2009). Reasoning about anomalies: A study of the analytical process of detecting and identifying anomalous behavior in maritime traffic data . InTolone, & Ribarsky, (Eds.), *SPIE Defense, Security, and Sensing. Visual Analytics for Homeland Defense and Security*. *Volume 7346*. Orlando, FL: SPIE Press.

Roy, J. (2008). Anomaly detection in the maritime domain. In *Proceedings of SPIE, Volume 6945,* 69450W 1–14. Bellingham, WA: SPIE Press.

Seibert, M., Rhodes, B. J., Bomberger, N. A., Beane, P. O., Sroka, J. J., et al., & Tillson, R. (2006). SeeCoast port surveillance. In *Proceedings of SPIE, Volume 6204: Photonics for Port and Harbor Security II.* Orlando, FL: SPIE Press.

Shearer, C. (2000). The CRISP-DM model: The new blueprint for data mining. *Journal of Data Warehousing*, *5*(4), 13–22.

Teoh, S. T., Zhang, K., Tseng, S., Ma, K., & Wu, S. F. (2004). Combining visual and automated data mining for near-realtime anomaly detection and analysis in BGP. In *Proceedings of the 2004 ACM Workshop on Visualization and Data Mining for Computer Security*, 35–44. New York: ACM Press.

Thomas, J., & Cook, K. (Eds.). (2005). *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. Los Alametos, CA: IEEE Computer Society.

Willems, N., Wetering, H. V. D., & Wijk, J. J. V. (2009). Visualization of vessel movements. *Computer Graphics Forum*, *28*(3), 959–966. doi:10.1111/j.1467-8659.2009.01440.x

KEY TERMS AND DEFINITIONS

Anomaly Detection: Process of discovering anomalies in a data set. Such process normally compares the data of interest with a simplified description or model of the normality in order to find mismatches.

Anomaly: In this chapter an anomaly is defined from a user (operator or organization) point of view, as exceptional objects, events or situations that need to be detected and identified. We define the term anomalous as a property, meaning "not conforming to what might be expected because of the class or type to which it belongs or the laws that govern its existence, in a given situation or context".

Behavioral Anomaly: An anomaly that implies a deviation from the normal behavior.

Predictive Data Mining: Class or type of data mining processes used to predict some response of interest. Predictive data mining is employed to identify a model or a set of models from the data that can be used to predict, for example, the value of a particular attribute (Demšar, 2006). Statistical analysis, classification, and decision trees techniques are used to produce such outcomes. Predictive data mining techniques are used for anomaly detection.

Visual Analytics: Analytical reasoning supported by highly interactive visual interfaces (Thomas and Cook, 2005). Visual analytics strives to facilitate the analytical reasoning process by creating software that maximizes the human capacity to perceive, understand, and reason about complex, dynamic data and situations.