

Explanation Methods for Bayesian Networks: review and application to a maritime scenario

Tove Helldin and Maria Riveiro
Informatics Research Centre
University of Skövde, Sweden
Email: a03tovhe@student.his.se, maria.riveiro@his.se

Abstract—Surveillance systems analyze and present vast amounts of heterogeneous sensor data. In order to support operators while monitoring such systems, the identification of anomalous behavior or situations that might need further investigation may reduce operators' cognitive load. Bayesian networks can be used in order to detect anomalies in data. In order to understand the outcome generated from an anomaly detection application based on Bayesian networks, proper explanations must be given to operators.

This paper presents the findings of a literature analysis regarding what constitutes an explanation, which properties an explanation may have and a review of different explanation methods for Bayesian networks. Moreover, we present the empirical tests conducted with two of these methods in a maritime scenario. Findings from the survey and the experiments show that explanation methods for Bayesian networks can be used in order to provide operators with more detailed information to base their decisions on.

Keywords: anomaly detection, maritime situation awareness, Bayesian networks, explanation methods, Explanation Tree, Causal Explanation Tree.

I. INTRODUCTION

In pace with the globally expanding maritime industry, the maritime surveillance capacity must be further developed. A step in this direction is the development of methods and techniques that support the detection of anomalous behavior in maritime traffic, e.g. terrorist attacks, hazardous cargo transports or smuggling of humans and goods [1]. Bomberger et al. [2] claim that the detection of unusual vessel activity is an important homeland security issue.

Anomaly detection can be described as the process of detecting deviations from normality. According to [3], most anomaly detection methods build a model of normal behavior which is used as a template for detecting anomalous events in the incoming data. Anomaly detection is an area that has been investigated mainly within the domain of network security, though some of the methods developed within network security have been applied within the maritime domain.

The design and development of surveillance systems present many challenges related to the large amounts of heterogeneous data available. One problem associated with maritime surveillance of coastal regions is, according to [4] and [3], the massive amounts of data that have to be processed in order to detect anomalous events and objects. Riveiro et al. [3] further claim that monitoring these surveillance systems is

often a very challenging task due to not only the amount of information involved, but also factors such as time pressure, high stress and the uncertain nature of the information presented to an operator. As a result, the operators might have problems to achieve situation awareness, i.e. to understand what is happening in the environment that they are observing, something which is crucial for them in order to make high-quality decisions.

The situation awareness of the operators might improve if the surveillance system incorporates mature visualization techniques suitable for the specific system, as well as if the operators understand the underlying model that serves as a basis for the anomaly detection capability and its outcomes. If the operator understands how the system works and feel confident using the system, his or her trust in the system might increase. However, many such models are difficult to understand and interpret. One exception is the Bayesian network (BN) approach for detecting anomalous vessel behavior. According to [5], techniques based on BNs have two main advantages compared to opaque machine learning techniques: (1) it possible to include expert knowledge into the Bayesian model and (2) due to its graphical nature, it is easier for an operator to understand and interpret the model representing the situation.

According to [6], 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 or her confidence in the system. Lacave and Díez [7] claim that the ability of explaining the reasoning behind a decision is of great importance in order for an operator to fully accept the advice that the system proposes. The authors in [7] further claim that human-computer collaboration requires mutual understanding, which makes explanations in expert systems even more important. Despite this fact, the amount of research devoted to this subject is relatively sparse and none of the approaches that exist today are, according to [7], satisfactory for the end-users. In [7], the authors argue that many of the explanatory solutions identified so far have not been tested on practical examples, which limits their scope.

In order to improve operators' situation awareness and reduce their cognitive load, this paper tackles the problem of finding and building suitable explanations from BNs outcomes, when BNs are used for detecting anomalous behavior in maritime traffic.

This paper presents: (1) the findings of a literature analysis regarding what constitutes an explanation, which properties an explanation may have and a review of different explanation methods for BNs and (2) the empirical tests conducted with two of these methods in a maritime scenario. The main contribution of this paper is the analysis of the applicability and suitability of such explanation methods when BNs are used to detect anomalous behavior in maritime traffic data.

The remainder of this paper is structured as follows: section II presents relevant concepts and properties regarding explanatory methods for BNs and section III reviews relevant literature on explanation methods. Section IV describes the experiments conducted with real maritime traffic data and the results obtained. Finally, section V presents some conclusions and future work.

II. EXPLANATION PROPERTIES AND CONCEPTS

Lacave and Díez [7] argue that to explain something to someone is to subject the object of the explanation, the *explanandum*, in such a way that it is understandable for the receiver of the explanation, i.e. that he or she can improve his or her knowledge about the object. In [7], the authors further claim that an explanation generated from BNs can be characterized according to three features: *content*, *communication* and *adaption* (see table I). An explanation can, for example, be focused on describing the evidence, model or the reasoning behind the explanation. Furthermore, the generated explanation can either be presented at a micro or macro level of detail. The explanations generated can also have a causal or non-causal interpretation. How a generated explanation is presented to the user might also differ between different systems. The explanation can either gradually be presented to the user during the inference process, or after the calculations have been made. The explanations can also be presented with, for example, numbers, texts or graphics. Generated explanations can also be adapted to different kinds of users. It might be the case that, for example, novice users need more exhaustive explanations than expert users who have more knowledge of the current domain.

There are several methods that can be used in order to explain the outcome of a BN to a user. The difference between these methods can often be found in, for example, their treatment of the different variables of the network: some methods include all variables in the explanation, while others only include a subset of all possible variables. Nielsen, Pellet and Elisseeff [8] divide the different variables constituting an explanation into three different groups: the *observed variables*, the *explanatory variables* and the *explanandum* (see figure 1). The observed variables are those variables in the environment whose states are known. Though, it might be the case that only a subset of the variables is needed in order to give a reasonable explanation to a user. These variables are called the explanatory variables and can either be observed or unobserved. Existing explanation methods treat these explanatory variables differently: they are either included or excluded from the explanation, often depending on how much information

Table I
EXPLANATION PROPERTIES (ADAPTED FROM [7]).

Content	Focus Purpose Level Causality of the BN	evidence/ model/ reasoning description/ comprehension micro/ macro causal/ non-causal
Communication	User-system interaction Presentation Expressions of probability	menu/ predefined questions/ natural language dialog text/graphics/multimedia numeric/linguistic
Adaption	User's knowledge about the domain User's knowledge about the reasoning method Level of detail	no model/ scale/dynamic model no model/scale/dynamic model fixed/threshold/auto

they provide for the evidence to be explained. The state of the variable that is to be explained is called the *explanandum*. In an anomaly detection problem, an *explanandum* may consist of the variable “anomaly” having the value “yes”. In order to explain why this variable has taken on the specific value, an explanation considering the variables that can describe why the variable has taken on a positive value can be created. Such explanation can include both observed and unobserved variables. Unobserved variables are investigated in order to determine how much explanatory power they would add to the *explanandum*, if their their states were known.

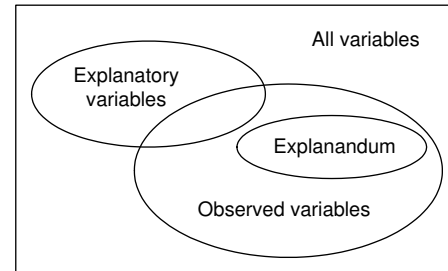


Figure 1. Description of different variables that can constitute an explanation (after [8]). Three different groups: the observed variables, the explanatory variables and the *explanandum*.

III. EXPLANATION METHODS

There are several different explanation methods that can be used in order to explain the inferences made by a BN. Three examples are *abductive inference*, the *Explanation Tree* (ET) method and the *Causal Explanation Tree* (CET) method. Flores [9] claims that the abductive inference methods aim to obtain the best configuration of the values for the explanatory variables that are consistent with the *explanandum* and can be assumed to predict it. Two of such methods are called *total abduction* and *partial abduction*. Both the total and partial abduction aim at finding the configuration of the variables that maximizes the posterior probability given the *explanandum*, though they differ in the way they treat the explanatory variables. The total abduction method, or the *Most Probable Explanation* (MPE) as it is also called,

includes all the variables in the explanation set, while the partial abduction method, or the *Maximum A Posteriori* (MAP) method, only includes a subset of the unobserved variables in the explanation set. This is one critique against the MPE method: since it includes all the variables in the explanation set, the method often produces a list of many different and uninformative explanations, which is defined as the “over specification problem” by Flores [9]. Moreover, as Nielsen et al. [8] claim, it is also difficult to distinguish between the often long explanations generated from the MPE method since they commonly resemble each other, and that their respective probabilities are low. This problem is addressed by the MAP method, since it reduces the number of variables included in the explanation set. The explanations resulting from the MAP method thus include fewer variables that do not contribute to the explanation given. Nielsen et al. [8] argue that the abductive methods do not distinguish between observing an explanatory variable X in a certain state x , and forcing it to have the value x . Thus, depending on the choice of explanatory variables, the most intuitive interpretation might not hold. Nielsen et al. [8] moreover claim that the MPE method, and to some extent also the MAP method, are not robust, that is, if changes occur in the network, this will often cause a change of the analysis, even if the changes occur in parts of the network that are largely independent of the explanandum.

The ET method and CET method present their explanations in a tree structure where every node in the tree represents a variable of the explanation set and every branch an instantiation of the variable to one of its possible states. The tree structure is, according to [8], a good way of presenting an explanation to the user, since many different variables can be presented in a compact and structured way.

When growing an explanation tree, two factors have to be considered: (1) which variable to split on and (2) when to stop growing the tree. According to the ET method, the variable that helps the most to determine the values of the other explanatory variables, given the explanandum, should be selected as the next node to split on [9]. In order to calculate this, one can use the mutual information measure or the Gini index [9]. Together with the CET method, the variable that shares the most causal information with the explanandum is the first variable chosen to be present in the explanation tree. For each possible value of the selected node, a branch is added to the root. A selected stopping criterion can be used in order to determine when to stop growing the tree. A change of the stopping criterion for the ET method affects how many levels of the explanation tree are presented, while a change of the stopping criterion together with the CET method affects how many variables should be presented in the generated tree. For example, if a stopping criterion 5 was used together with the ET method, 5 levels in the explanation tree would be generated, while the same stopping criterion together with the CET method would present a tree constituted of 5 variables.

Another difference between the two methods is that the CET algorithm assumes an underlying causal BN, while the ET algorithm makes no such assumption.

IV. EXPERIMENTS AND RESULTS

The purpose of the experimentation is to investigate if the ET and CET methods can be used in order to explain the cause of an anomaly to a user in a maritime scenario. The ET and CET methods were chosen for the experiments since the tree representation of different hypothesis is a good way of compactly presenting competing explanations to a user [8]. Moreover, due to their recent development, these methods have overcome many of the shortcomings of the abductive explanation methods (according to their founders [9] and [8]).

Several experiments were conducted in order to test the two explanation methods on real maritime AIS data¹, provided by Saab Microwave Systems. The data reflect five days of vessel traffic outside the coast of Gothenburg, and contain information about the different types of vessels, such as ID, position, heading and speed. The AIS data was divided into 6 different files containing messages from only one vessel type, such as cargo, passenger, tanker and pilot vessels. The AIS file was divided in this way in order to make it easier to investigate how the different types of vessels behave, which would help when detecting anomalies. From the AIS data, the columns “longitude”, “latitude”, “heading”, “speed” (speed measured in knots, “SogKnots”) and “course over ground” (“Cog”) were considered. An additional column, “anomaly”, was manually added to the data in order to make it easier to analyze the outcome of the explanation methods chosen for the experimentation.

The BNs used for the experiments were created using *Genie*², an application developed by the Decision Systems Laboratory in Pittsburgh which can be used in order to create BNs from data. An example of a network created in Genie using the AIS data as input is depicted in figure 3. The numeric features of the different variables in the AIS file were here first discretized into 5 different groups in order for the application to be able to create the network .

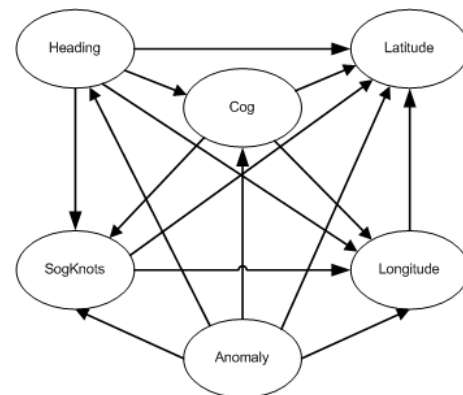


Figure 3. An example of a Bayesian network created in Genie from the AIS data

Three different types of anomalies were hidden in the AIS data: (1) one vessel presents high speed values (speeding

¹Automatic Identification System (AIS) broadcasted messages.

²For more information on Genie: <http://genie.sis.pitt.edu/>

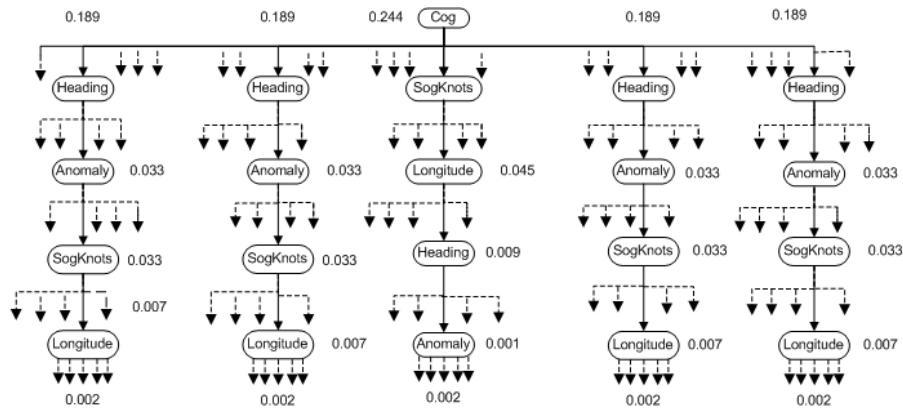


Figure 2. An ET explanation for abnormal position and speed of one vessel (stopping criterion 0).

situation), (2) one vessel is positioned far away from the other vessels in the data and (3) a combination of the two aforementioned anomalies. These anomalies were added by manually changing the speed variable and position variables for one vessel in the AIS data. Based on this data, different BNs were generated, which were later on used as a basis for the two explanation methods selected. When creating the BNs to be used together with the CET algorithm, the “anomaly” variable was excluded from the tests, to later be added manually to the network, in order to ensure the causal relationships between the explanandum and its parent variables. After having added the anomaly node to the network, the probabilities for the node were calculated based on the data in the column “anomaly” in the AIS data.

Figure 2 depicts the results from the tests conducted together with the ET method where the anomaly hidden in the AIS data consisted of one vessel speeding in an abnormal position. The different numbers in the figure represent how much information the variables reveal about the other variables in the network when using the ET method, and how much causal information the variable shares with the explanandum when using the CET method. The more information that the chosen variable reveals about the other variables/shares with the explanandum, the better the explanation. As can be read from the tree, the ET method chooses the “Course over ground” (Cog) variable as the first variable to split on since, based on the criteria of the algorithm, this is the variable that gives the most information about the other variables in the network. Thus, according to the ET method, “Cog” is the best explanation for the explanandum, i.e. that “anomaly” has the state “yes”. Due to the discretization of the different variables in the Genie application, one can furthermore see that the third category of the “Cog” variable constitutes a slightly better explanation than the other discretization categories (0.244 compared with 0.189), i.e. that the vessel has a medium course over ground value is the first best explanation for the state of the explanandum. The different discretization categories of the variables, i.e. the five different states of the variables presented in figure 2, should be interpreted as a scale of

the values from the AIS data, ranging from a low number of, for example, the “SogKnots” variable, i.e. slow speed, to a high number, i.e. high speed. Second, it chooses the variables “Heading” and “SogKnots”. Here, one can see that “SogKnots” gives slightly more information about the other variables in the network than the “Heading” variable (0.045 compared with 0.033), thus “SogKnots” is chosen to be the next variable in line that helps the most to determine the values of the other explanatory variables, given the explanandum. The algorithm continues to split the variables in the tree, though the information retrieval about the other variables in the network are decreased to a minimum, thus “Cog”, “SogKnots” and “Heading” are considered to be the best explanation(s) in this case, depending on the stopping criterion the operator has chosen for the ET algorithm.

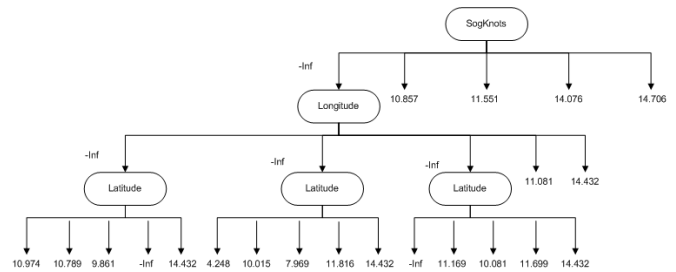


Figure 4. A CET explanation for abnormal position and speed of one vessel (stopping criterion 5).

The explanation tree presented in figure 4 reflects the CET algorithm’s choice of variables where one vessel is speeding in an abnormal position. The fifth discretization category of the speed variable, i.e. high speed of the vessel, is chosen as the first best explanation. Though, if the vessel is not speeding, the second best explanation can be explained by the “Longitude” variable, followed by the “Latitude” variable. The CET method thus manages to identify the three variables “SogKnots”, “Longitude” and “Latitude” as the variables that share the most causal information with the explanandum, though it does not manage, in this case, to identify the correct discretization categories of the variables.

V. CONCLUSIONS

Results from the experiments conducted show that the CET algorithm managed to identify the concealed causes of the hidden anomalies with better performance than the ET algorithm. Due to the ET algorithm's way of choosing the variables present in the explanation tree, the algorithm was not capable of generating explanations that reflected the hidden anomalies in the AIS data in an appropriate way.

Both the ET and the CET algorithms generate comprehensive explanations presented at a macro level, which makes it easy for an operator to understand the explanations. The tree structure also makes it easy for an operator to get an overview of the different variables presented. This may be a suitable solution for the scenario analyzed, since the variables are correlated and an explanation might involve several variables. Though, it might be difficult for an operator to understand what the numbers represent and no further description of the explanation is given. Based on their content, the only difference between the two methods is the causality characteristic. According to [7], a causal interpretation of an explanation is often more intuitive for the user, however, not all BNs can be considered to be causal, i.e. the CET method might not be applicable to all BNs.

More experiments can be conducted where several anomalies have been hidden in the AIS data, since only adding one anomaly per AIS file makes it difficult to calculate a proper Conditional Probability Table (CPT) for the anomaly node, which in turn makes it difficult to calculate the different relationships between the variables in the network. Thus the different variables in the BN are very strongly connected. Together with the ET algorithm, a strongly connected BN has a negative influence, since it is an indication to the algorithm that it must not stop the calculations before all the variables present in the BN have been covered in the generated explanation. This is most apparent together with the ET algorithm, since it affects how many levels of the generated explanation tree should present, and not how many variables should be present as together with the CET algorithm.

ACKNOWLEDGMENTS

We would like to thank Jean-Philippe Pellet (Data Analytics Group, IBM Zurich Research Lab) for his support during the experimental phase and his useful comments on Causal Explanation Trees and Explanation Trees. We are grateful to Fredrik Johansson for his feedback regarding Bayesian networks and Henrik Boström for his input during the whole project. Thanks also to Thomas Kronhamn, Martin Smedberg and Håkan Warston (Saab Microwave Systems) for providing the data.

REFERENCES

- [1] G. K. Høye, T. Eriksen, B. J. Meland, and B. T. Narheim, "Space-based ais for global maritime traffic monitoring," *Acta Astronautica*, vol. 62, no. 2-3, pp. 240–245, 2008.
- [2] N. Bomberger, B. Rhodes, M. Seibert, and A. Waxman, "Associative learning of vessel motion patterns for maritime situation awareness," in *Proceedings of the 9th International Conference on Information Fusion*, July 2006.
- [3] M. Riveiro, G. Falkman, and T. Ziemke, "Improving maritime anomaly detection and situation awareness through interactive visualization," in *11th International Conference on Information Fusion (ICIF 08)*, ISIF. IEEE, June-July 2008, pp. 47–54, Best Student Paper Award.
- [4] J. Roy, "Anomaly detection in the maritime domain," S. H. Craig, L. Daniel, and T. S. Theodore, Eds., vol. 6945. SPIE, 2008, pp. 69450W1–14, optics and Photonics in Global Homeland Security IV.
- [5] F. Johansson and G. Falkman, "Detection of vessel anomalies - a Bayesian network approach," in *Proceedings of the 3rd International Conference on Intelligent Sensors, Sensor Networks and Information Processing*, 2007.
- [6] F. Jensen, S. Aldenryd, and K. Jensen, "Sensitivity analysis in Bayesian networks," in *Symbolic and Quantitative Approaches to Reasoning and Uncertainty*, 1995, pp. 243–250.
- [7] C. Lacave and F. J. Díez, "A review of explanation methods for Bayesian networks," *Knowl. Eng. Rev.*, vol. 17, no. 2, pp. 107–127, 2002.
- [8] U. H. Nielsen, J.-P. Pellet, and A. Elisseeff, "Explanation trees for causal Bayesian networks," in *Proceedings of the 24th Annual Conference on Uncertainty in Artificial Intelligence (UAI-08)*, 2008, pp. 427–434.
- [9] M.-J. Flores, "Bayesian networks inference: Advanced algorithms for triangulation and partial abduction," Ph.D. dissertation, Departamento de Sistemas Informáticos, University of Castilla - La Mancha, Spain, November 2005.