HÖGSKOLAN SKÖVDE

Institutionen för kommunikation och information
Examensarbete i datavetenskap 30hp
Avancerad nivå
Vårterminen 2009

Explanation Methods for Bayesian Networks

Tove Helldin

Title

Submitted by Tove Helldin to the University of Skövde as a dissertation towards the degree of M.Sc. by examination and dissertation in the School of Humanities and Informatics.

Date

I hereby certify that all material in this dissertation which is not my own work has been identified and that no work is included for which a degree has already been conferred on me.

Signature: ____

Supervisor: Maria Riveiro

Explanation Methods for Bayesian Networks

Tove Helldin

Abstract

The international maritime industry is growing fast due to an increasing number of transportations over sea. In pace with this development, the maritime surveillance capacity must be expanded as well, in order to be able to handle the increasing numbers of hazardous cargo transports, attacks, piracy etc. In order to detect such events, anomaly detection methods and techniques can be used. Moreover, since surveillance systems process huge amounts of sensor data, anomaly detection techniques can be used to filter out or highlight interesting objects or situations to an operator. Making decisions upon large amounts of sensor data can be a challenging and demanding activity for the operator, not only due to the quantity of the data, but factors such as time pressure, high stress and uncertain information further aggravate the task. Bayesian networks can be used in order to detect anomalies in data and have, in contrast to many other opaque machine learning techniques, some important advantages. One of these advantages is the fact that it is possible for a user to understand and interpret the model, due to its graphical nature.

This thesis aims to investigate how the output from a Bayesian network can be explained to a user by first reviewing and presenting which methods exist and second, by making experiments. The experiments aim to investigate if two explanation methods can be used in order to give an explanation to the inferences made by a Bayesian network in order to support the operator's situation awareness and decision making process when deployed in an anomaly detection problem in the maritime domain.

Key words: Explanation methods, Explanation Tree, Causal Explanation Tree, Bayesian networks, anomaly detection, information fusion, maritime situation awareness

Contents

1 Introduction

The international maritime industry is expanding fast due to an increasing number of transportations over sea (Høye et al., 2008). Nevertheless, the maritime surveillance capacity has not been developed in the same pace, thus Høye et al. (2008) further claim that more effort must be put on investigating new ways of detecting abnormal vessel behavior. The development of new and better surveillance sensors and applications for anomaly detection within the maritime domain has increased the possibility for maritime organizations, both within the military and civilian domain, to detect anomalous vessel behavior. Though, with this development follows the need to fuse the massive amounts of data received from the sensors. Here is where techniques and methods used in information fusion can be of great support, since they integrate data from multiple sources (such as expert knowledge and sensors) in order to be able to make better inferences about possible vessel activities detected in the data.

Related to the concept of information fusion is the fact that much of the data collected from the different sources carries uncertainty. In order to deal with this uncertainty, probability theory can be used. One technique that uses probability theory in order to make inferences in data is Bayesian networks (BNs). According to Johansson and Falkman (2007) the Bayesian network approach for detecting anomalous behavior has several advantages compared to many other opaque data analysis techniques such as neural networks: BNs are good at handling incomplete data sets, experts can incorporate their knowledge into the model, as well as that it is possible for humans to understand the Bayesian model due to its graphical nature (Johansson and Falkman, 2007). Though, according to Chajewska and Halpern (1997), it is often difficult for humans to understand probabilistic inference, that is, it might be difficult for a user to understand the output of the anomaly detection system that is based on a Bayesian network. If the user does not understand how the system has come up with its recommendations, it might also be the case that the user does not trust the recommendations that the system provides when an anomaly has been detected, nor does the system help the user to obtain a better situation awareness of what is happening in the observed environment.

The aim of this thesis is thus to investigate which methods and techniques can be used in order to explain the output from a Bayesian network when used in the problem domain of anomaly detection. The thesis also investigates what an “explanation” is in this context, as well as what properties such an explanation may have. An analysis of the challenges related to using the explanation methods in order to depict an anomaly in an understandable way for an operator is also conducted.

Section 2 of this thesis presents the domain of information fusion, the concept of situation awareness, anomaly detection, Automatic Identification System (AIS), Bayesian networks, explanation properties and explanation methods for Bayesian networks. In section 3, the aim and objectives of this thesis are presented, whereas section 4 presents the methods that have been selected in order to achieve the aim. A description of the realization of the experiments conducted is the founding of section 5. In section 6, the results from the conducted experiments are presented. The last section of the thesis, section 7, presents the conclusions of the thesis together with some suggestions for future work.

2 Background

This section briefly presents important aspects related to this thesis: information fusion, situation awareness, anomaly detection, AIS-system, Bayesian networks, explanation properties and explanation methods for Bayesian networks. Subsection 2.1 presents an overview of the information fusion area where important concepts as well as the JDL model are presented. Subsection 2.2 considers the concept of situation awareness and gives a brief introduction to why this concept is important in relation to the work of the operators of complex systems. Subsection 2.3 presents an overview of the domain of anomaly detection, while subsection 2.4 gives a brief presentation of the AIS-system. Subsection 2.5 gives an introduction to Bayesian networks, while subsection 2.6 presents the results from the literature analysis conducted in this thesis where explanation properties and explanation methods for Bayesian networks are analyzed.

2.1 What is information fusion?

According to Dasarathy (2001), information fusion (sometimes referred to as data fusion) includes the theory, techniques and tools that can be used in order to make use of the synergy in the information obtained from multiple sources, for example sensors, databases and humans. Furthermore, Dasarathy (2001) claims that the aim of information fusion is to fuse data in order to produce decisions or actions that in some sense are better, either quantitatively or qualitatively, than would be possible if only one source of information had been used. Hall and Llinas (1997) contribute to this view and claim that through information fusion, we can get a more accurate picture of our environment and potential threats than if we had just used data from one source.

The concept of information fusion is not at all new. Elmenreich (2002) argues that both humans and animals fuse information in our everyday lives with the help of our senses. We use, for example, both our smell and taste senses in order to determine if the milk is drinkable or not, and even perhaps our vision to see if it got clumps in it. Hall and Llinas (1997), point out that information fusion methods historically were mostly influenced by the military domain. Automated threat recognition systems and battlefield surveillance systems are examples of applications that have emerged from the military view of the information fusion area. Much of the development of the fusion domain has its roots in the defense area, thus it is from here that the true engineering principle and standardized terminology stem. Though, according to Hall and Llinas (1997) one can now see a trend where information fusion applications originate from other domains as well, such as the medical and commercial areas. Regardless of the origins of the applications, the underlying techniques used in order to fuse the data are the same. Examples of fields from where these techniques have been collected are: artificial intelligence, digital signal processing and control theory (Hall and Llinas, 1997).

Hall and Llinas (1997) argue that the data fusion community for a long time lacked a unifying terminology, which hindered technology transfer within the area. Even within military applications different definitions were used for fundamental terms. In order to spread the technology and definitions in a unifying way, the Joint Directors of Laboratories (JDL) Data Fusion Working Group started with creating a terminology related to fusion. This work resulted in a data fusion process model called the JDL model. The JDL model is a conceptual model which describes the processes, functions, categories of techniques and specific techniques applicable to data fusion (Hall and Llinas, 1997).

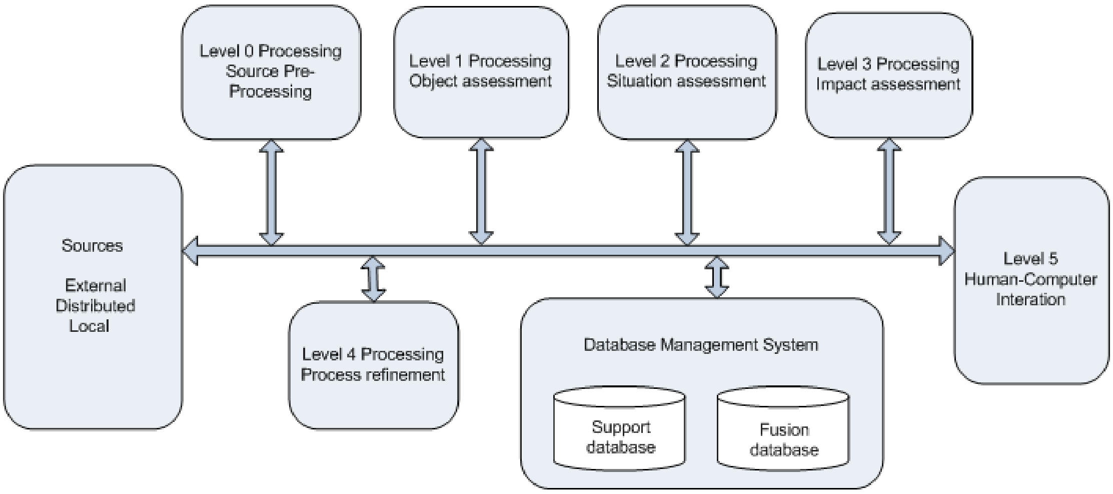


Figure 1 – The JDL process model (after Hall and Llinas, 1997)

Below, a short description of the different levels of the JDL model is presented.

- **Sources**

Elmenreich (2002) states that the data provided as input to the fusion system might originate from various different sources. Examples of such sources are, for example, local sensors, distributed sensors and data from external sources such as reference information or geographical information.

- **Level 0 - Source pre-processing**

At the source pre-processing step in the data fusion process, data is allocated to the appropriate processes (Elmenreich, 2002). Data and signals are “prepared”, i.e. are processed in such a way as to reduce irregularities in the data, for processing at other, higher levels in the process model. The source pre-processing step also forces the data fusion process to concentrate on the data that is the most relevant for the situation at hand (Hall and Llinas, 1997).

- **Level 1 – Object assessment**

At the object assessment level, the process aims at combining characteristics in the data in order to transform it into more accurate representations of objects and entities. According to Hall and Llinas (1997) data is here also assigned to objects so that the application can perform statistical operations on them in order to refine the estimation of an object’s identity or classification.

- **Level 2 – Situation assessment**

  The objects and events that were discovered during the former processing level are at this level, the situation assessment level, further investigated in order to establish relationships among them in the observed environment and also to categorize them. Most attention is here focused on the relational data in order to determine the meaning of the observed objects and entities (Hall and Llinas, 1997).

- **Level 3 – Impact assessment**

  At the third processing level, the impact assessment level, the data fusion system tries to "predict" the future, i.e. draw inferences about threats, opportunities and vulnerabilities (Elmenreich, 2002). Here, several predictions or hypothesis can be produced, since the future is uncertain (Hall and Llinas, 1997).

- **Level 4 – Process refinement**

  Elmenreich (2002) calls the process refinement level a meta-process, since processing on this level is concerned with improving the other fusion processes involved. Aims at this level are, for example, to identify which information is needed in order to make the fusion process better, from which sensors that data should be collected, which sensor adjustments should be performed in order to get higher quality data and also how to allocate the resources available in such a way that the fused data can help in achieving the goals of the mission.

- **Level 5 – Human-Computer Interaction**

  Human-computer interaction, or cognitive refinement, is the process of monitoring and improving the interaction between the human and the system. Riveiro (2007) claims that since humans often are involved in all the steps of the fusion process, as for example decision makers or information providers, it is of great importance that the information is presented in such as way that the user can understand and act on it.

- **Data base management**

  An important part of the information fusion process is to handle the data used for fusion. The database management system provides functions in order to retrieve, store, archive, compress, query and protect the data collected in the database (Hall and Llinas, 1997).

Bomberger et al. (2007) divides the JDL data fusion model into two levels. The processes at levels 0 and 1 are referred to as low-level fusion, while the processes at levels 2-5 are called high-level fusion. Hall and Llinas (1997) claim that the goal of the low-level fusion is to identify objects and their positions in the environment as well as assigning attributes to them, such as kinematic estimations, with the help of methods such as Kalman filters and alpha-beta filters. Bomberger et al. (2007) further describe the purpose of the higher-level fusion as the process of combining data, which has been processed by the lower-level fusion processes, with existing knowledge in order to achieve situation awareness (see subsection 2.2). Methods that have been used for this purpose are, for example, neural networks, fuzzy logic and Bayesian networks.

Over the years, the JDL model has been criticized for not paying enough attention to the operator using the fused information, nor giving sufficient guidance when it comes to developing a data fusion system. However, Hall and Llinas (1997) claim that the purpose of the JDL model is to give a general basis for common understanding and communication about information fusion.

2.2 Situation awareness

Situation awareness (SA) is, according to Endsley (1995), achieved when a user has perceived the elements in the environment, within a volume of time and space, has understood their meanings and can perceive their status in the near future. Though, Endsley (1995) claims that the increasingly complex systems and dynamic environments of today hinder the acquisition of situation awareness. In dynamic environments, operators are often required to make correct decisions within a limited time span and their tasks are often dependent on real-time analysis of the environment (Endsley, 1995). Since the observed environment is constantly changing, an operator's task is to obtain and maintain good situation awareness. It is not enough to just perceive the environment, but also understand the situation as a whole, in order to form a basis for decision making. Wallenius (2004) states that achieving situation awareness is a mental process that relies on the human mind and human senses, but that it can be enhanced by fusing data from several sources and combining it with stored knowledge.

Since situation awareness plays a great part of every operator's task, Endsley (1995) claims that it is important to incorporate the concept of situation awareness when designing the interface of the system (as suggested by the 5th level of the JDL model: Human-Computer Interaction, see figure 1 in subsection 2.1). Below, a model of the factors that affect the process of an operator acquiring situation awareness, as well as factors that in turn are affected by the operator's degree of SA, is depicted.

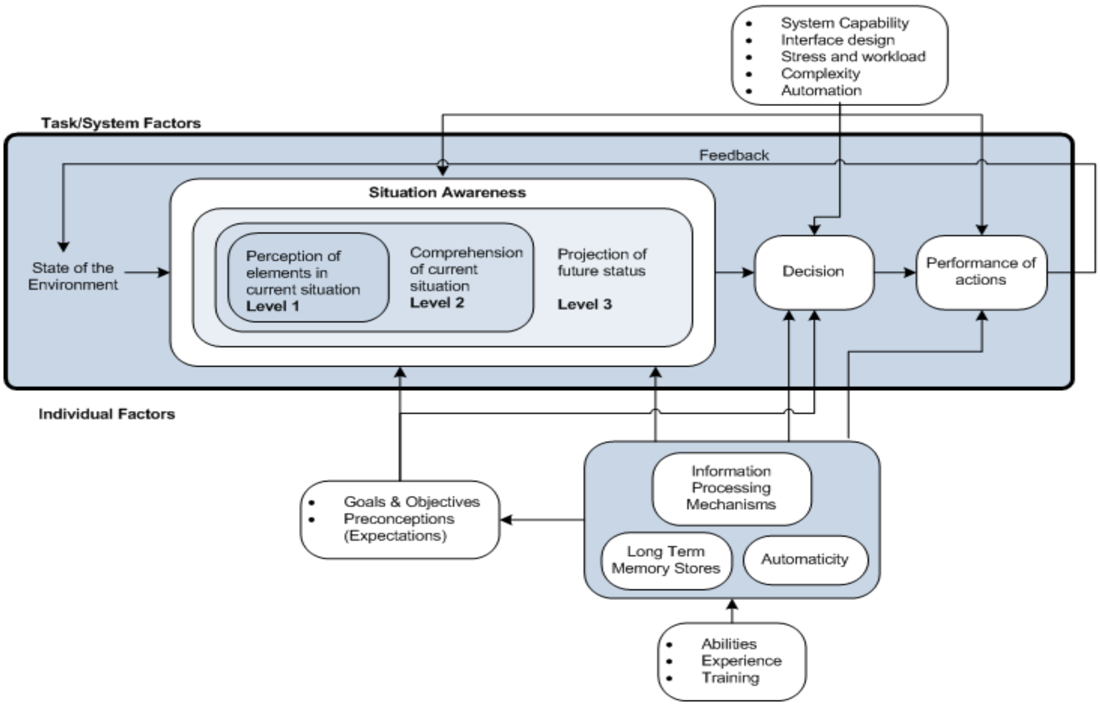


Figure 2 – A model of situation awareness (adapted from Endsley, 1995)

Endsley (1995) claims that there are three levels of situation awareness. The first level is called "perception". Endsley (1995) claims that the first step toward achieving good SA is to perceive the relevant elements in the environment together with their attributes. In a harbor scenario, this would for example imply that the operator has knowledge of where the vessels are situated, their directions and speeds.

The second level of SA, "comprehension", is about understanding the situation at hand. In order for an operator to achieve this level of SA, he or she must be able to unitize the perceived elements as well as understand their significance in relation to the goals of the tasks that the operator has. Only when this information has been fused by the operator can he or she create a holistic picture of the situation as well as comprehend the significance of objects and events (Endsley, 1995). In a harbor scenario, an operator monitoring an anomaly detection system can at this level of situation awareness detect a vessel that behaves abnormal according to some established criterion.

The third and last level of SA, "projection", is according to Endsley (1995) the process of analyzing the near future of the perceived elements and events. In a harbor scenario, this would for example imply that an operator can draw the conclusion that a boat approaching another boat with high speed will constitute a risk of collision. In light of this knowledge, the operator can decide to act in a way that meets his or hers objectives (Endsley, 1995).

As Lambert (2001) points out, the three levels of SA presented by Endsley can be compared to the levels 1-3 of the JDL model (object, situation and threat assessment). Though, Lambert (2001) further claims that the JDL model and situation awareness model do not depict the same concept: the JDL model should provide a technological basis in order to achieve situation awareness and communicate the results via a

Human-Computer Interface, while the model for situation awareness depicts the mental state of having achieved SA.

According to Endsley (1995), the operator's degree of situation awareness has great influence over the decision making process. The mental model that the operator has over the situation will directly affect the actions and problem-solving strategy that he or she will apply in order to meet his or hers objectives. Endsley (1995) further claims that there is an evident relationship between SA and performance. In general, it is expected that poor performance is a result of incomplete or inaccurate SA, when the correct actions are not known for the current situation or when time or other factors (such as stress, workload and complexity) limit an operator's ability to choose the correct action (Endsley, 1995). Thus, if an operator monitoring an anomaly detection system has a high degree of situation awareness, he or she should be better prepared to make high-quality decisions when it comes to acting upon the suggestions of the anomaly detection system. The cognitive load of the operator might as well decrease, which in turn might allow him/her to focus on the most interesting objects and events in the observed environment.

2.3 Anomaly detection

An anomaly, or outlier, can be described as a deviation from normality (Riveiro et al., 2008). The process of detecting such deviations is concerned with finding patterns in data that do not confirm to some expected behavior. Portnoy et al. (2001) claim that, typically, the approach for finding anomalies is to build a model of normal behavior and then attempt to detect deviations from the normal model. This can, for example, be used in order to detect intrusions in a network system, faults in a safety critical system, irregularities in the results of a medical system, as well as to detect anomalous events and objects in military surveillance systems. Thus, what can be defined as normal or anomalous is context and application dependent.

Despite knowing what can constitute an anomaly in a specific context, it might be difficult to identify anomalies: the difference between what constitutes normal behavior and anomalous behavior might not be obvious; the definition of "normal" might evolve over time in the specified environment and it might also be difficult to distinguish between noise in the data and anomalies (Roy, 2008).

Anomalies may be found for several reasons. Roy (2008) claims that anomalies can be detected as a result of erroneous or missing data (due to, for example, low-quality sensor readings, faulty sensors or communication channels), or malicious activities in the observed environment. Riveiro et al. (2008) claim that detecting anomalies can be seen as a classification problem, i.e. that an event or object is either classified as normal or anomalous. This classification problem can be divided into two parts: the construction of a normal model, learnt from training data, and a classification of new instances based on the learnt model. A challenge when detecting anomalies in the maritime domain is the huge amounts of data that has to be processed and analyzed in order to identify anomalies. Furthermore, the data typically comes in a streaming fashion, which requires on-line analysis (Roy, 2008). Roy (2008) also mentions the problem of the high false alarm rate, due to the large sized input. These factors are all contributing to making the process of identifying anomalies more difficult for the operators.

There are many different techniques that can be used in order to find anomalies. Most of the published work regarding anomaly detection can be found in the computer security area, though the techniques and methods used can now be found in other domains as well. Patcha and Park (2007) mention three different approaches for detecting anomalies in network traffic data: *statistical anomaly detection*, *machine learning based anomaly detection* and *data-mining based anomaly detection*[1]. In statistical anomaly detection, typically two profiles for each subject in the observed area are maintained: the current profile and the stored profile. In order to find anomalies, the current profile is compared with the stored profile, and an anomaly score is calculated. Depending on a threshold value for the anomaly score, the system generates an alarm (Patcha and Park, 2007). An anomaly detection system based on the machine learning approach is similar to the statistical approach for detecting anomalies. One difference between the two approaches is that the machine learning approach has the ability to learn and improve its performance when performing certain tasks, i.e. that the execution strategy of the system can be altered in order to better suit the problem at hand. Bayesian networks (described in subsection 2.5) are an example of this kind of anomaly detection approach. Data-mining based anomaly detection techniques have been increasingly investigated in order to eliminate the manual and ad hoc elements from the process of building an anomaly detection system. According to Patcha and Park (2007), data mining based anomaly detection is concerned with discovering patterns, associations, anomalies and statistically significant structures and events in data. An anomaly detection system based on this approach takes data as input and uses it in order to find patterns or deviations which might not be obvious at a first glance. Thus, data mining based anomaly detection techniques can help an operator to distinguish between normal and abnormal activities in data, based on patterns of normalcy.

2.4 AIS-system

The Automatic Identification System (AIS) is a maritime safety and vessel traffic system. The vessels using the system broadcast information such as vessel identity, position, heading, nature of cargo, destination, estimated time of arrival etc. every 2-10 seconds to other vessels or shores in the nearby environment (Høye et al., 2008). The system was first proposed by the International Association of Maritime Aids to Navigation and Lighthouse Authorities (IALA) in the early 1990s with the intention to have a means of identifying vessels on a radar screen (Eriksen et al., 2006).

When presented to the International Maritime Organization (IMO), requirements for the AIS system were developed. Such requirements specify that an AIS system shall: 1) automatically provide information about the vessel's identity, type, position, course, speed and other safety-related information to shore stations, other ships and aircraft information, 2) automatically receive such information from similarly fitted ships, 3) monitor and track ships and 4) exchange data with shore-based facilities. Vessels with built-in AIS shall have these AIS functions in operation at all times with the exception from where international agreements, rules or standards provide for the protection of navigational information (Eriksen et al., 2006).

[1] According to Patcha and Park (2007), anomaly-based detection methods only refer to data-driven approaches. In this thesis, the concept of an anomaly is broader, thus both rule-based methods, data-driven methods and a combination of the two are considered as anomaly-based detection methods.

The AIS has been mandatory on all new ships in international traffic since July 2002 and today all passenger ships, tankers and other ships of 300 tons as well as all ships over 500 tons or more in national voyages are affected by this mandatory rule (Eriksen et al., 2006).

For more information about AIS, see Eriksen et al. (2006) and Høye et al. (2008).

2.5 Bayesian networks

A Bayesian network (or belief network) is, according to Jensen (2000), a directed acyclic graph (DAG), consisting of nodes and edges. The nodes in the graph represent variables with a finite set of mutually exclusive and exhaustive states, while the edges represent probabilistic dependencies between the different variables in the graph (Jensen and Nielsen, 2007). A Bayesian network consists of two parts: one qualitative part and one quantitative part (Druzdzel and van der Gaag, 2000). The qualitative part is represented by the structure of the DAG, while the quantitative part of the model represents the beliefs about the states of the various variables (Jensen, 2000). These beliefs are encoded into the conditional probability tables (CPTs) related to each node. Nodes in the model have different names according to their relationships to the other nodes in the model. Nodes with edges directed to them are called "child" nodes, while nodes directed from them are called "parent" nodes. A node that has only edges directed from it is called a "root" node of the model.

Jensen (2000) claims that an important property concerning the qualitative part of the Bayesian network is the concept of conditional independence relations; relations that can be tested by making a directed graph separation (d-separation). Jensen (2000) gives the following definition of conditional independence:

> "Two variables A and B are independent if knowledge of A does not change the belief about B (and vice versa). A and B are conditionally independent given C if they are independent whenever the state of C is known." (Jensen, 2000, p.1-2).

Jensen (2000) claims that the structural part of the Bayesian network reveals if the nodes are conditionally independent or not, characterized by the different kinds of connections between the nodes in the network. There are three possible connections: serial, diverging and converging, as illustrated below.

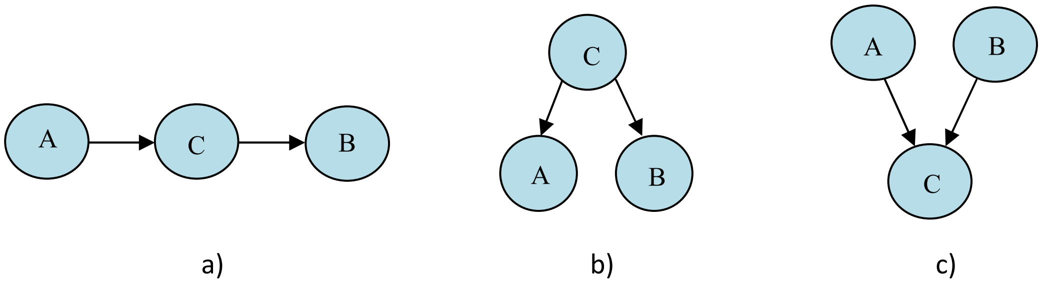


Figure 3 - Connections between nodes in a directed acyclic graph: a) serial connection, b) diverging connection and c) converging connection.

Jensen (2000) claims that two nodes, A and B, are independent given the evidence if for all paths between A and B, there is an intermediate node C such that either i) the connection is serial or diverging and the state of C is known ii) or the connection is converging and neither C nor any of its descendants have received evidence.

Figure 3a illustrates a serial connection between the nodes. According to Jensen's definition of conditional independence, the nodes A and B are only independent if evidence about the state of the intermediate node C is given. If not, the nodes are not conditionally independent, i.e. evidence about the state of the node A, will influence the evidence of the state of the node C, which in turn will influence the evidence of the state of the node B. The same rule of conditional independence applies to nodes in a diverging connection, as illustrated by figure 3b, i.e. if the state of the parent node C is known, then its children A and B are conditionally independent. For nodes in a converging connection, shown by the figure 3c, a different rule is used. Here, the parent nodes A and B are only conditionally independent if there is no evidence of the state of the child node C or its descendants.

Also associated with the conditional independence of nodes in a Bayesian network is the concept of the Markov blanket. The Markov blanket of a node contains the node's parents, children and children's parents (see figure 4). When predicting the behavior of a specific node in the network, the nodes that have to be considered for this prediction are the nodes belonging to the Markov blanket of the chosen node (Yap et al., 2008). Thus, the specific node is conditionally independent of the other nodes that do not belong to its Markov blanket, given that all the nodes in the current Markov blanket are instantiated.

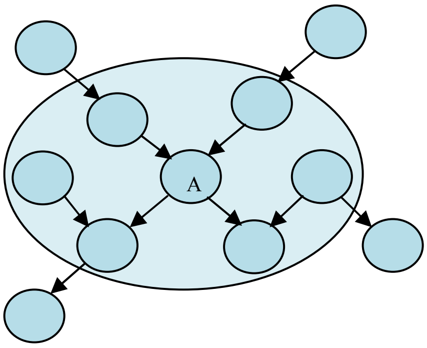


Figure 4 – The Markov blanket of the node A consists of the nodes in the greater circle.

Nodes with no parents have CPTs that are not influenced by the probabilities of the other nodes, while nodes with parents must take the probability values of the parents' different states into account when calculating or updating the probabilities of their own states.

The values of the CPTs can either be provided by experts of the specific domain, by statistical data collected from the observed environment, or a combination of the two (Johansson and Falkman, 2007). All nodes in the network have their own individual CPT. The values of the different nodes can either be discrete (with at least two states, like “true” or “false”) or continuous. Johansson and Falkman (2007) mention that the values of the different states of the nodes in the network together with the graph structure are then to be used in order to calculate the joint probability distribution of the network. The joint probability distribution can be described as a function assigning a number between [0,1] to each possible combination of states of the variables describing the domain (Johansson and Falkman, 2007). This can be done with the help of the so called chain rule of Bayesian networks (see Equation 1). The chain rule of Bayesian networks declares that a Bayesian network is a representation of a unique joint probability distribution over all the variables represented in the graph, from which marginal and conditional probabilities can be computed for each node in the network (Johansson and Falkman, 2007). If *U* is a universe of variables: *U* = {$X_1, X_2, \ldots, X_n$}, the joint probability of *U* becomes:

$$P(x_1, \ldots, x_n) = \prod_{i=1}^{n} P(x_i \,|pa(x_i\,))$$

Equation 1 – The chain rule of Bayesian networks

where Pa(x_i) denotes the node X_i’s parents. The Bayesian network below (figure 5), as described by Russel and Norvig (2003), illustrates the relationships between the variables *Burglary*, *Earthquake*, *Alarm*, *John*Calls and *Mary*Calls. A burglary and an earthquake can both trigger the alarm system to go off, which in turn can cause both John and Mary to call the police. The different variables presented can take on the states “true” or “false”, i.e. either a burglary has taken place or not. From the chain rule we get the joint probability: P(*Burglary*, *Earthquake*, *Alarm*, *JohnCalls*, *MaryCalls*)=P(*Burglary*)P(*Earthquake*)P(*Alarm*|*Burglary*,*Earthquake*)P(*JohnCalls*|*Alarm*)P(*MaryCalls*|*Alarm*).

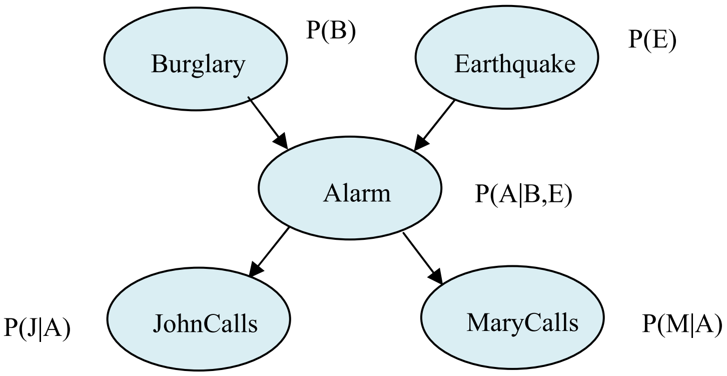


Figure 5 – An example of a Bayesian network

With the help of Bayes’ rule (see Equation 2), it is possible to update our beliefs about an event A, given that we receive new knowledge about another event B. Thus the term P(A) is called the prior probability of A, while the term P(A|B) is called the posterior probability of A given the new observation B (Jensen and Nielsen, 2007).

$$P(a|b) = \frac{P(a,b)}{P(b)} = \frac{P(b|a)P(a)}{P(b)}$$

Equation 2 – Bayes' rule

Druzdzel and van der Gaag (2000) claim that the process of creating a Bayesian network consists of three tasks. The first task is to identify the variables of importance in the specific domain, as well as to identify the possible states that the variables can take on. The second step of the process is to identify the relationships between the identified variables and to illustrate these in a graphical model. The third and last part of the process is concerned with obtaining the probabilities that are required for the CPTs of the different variables. Druzdzel and van der Gaag (2000) further claim that this is the most time-consuming and difficult part of the process of creating a Bayesian network.

Despite the fact that it is difficult to generate the probability values for the different variables in a network, Bayesian networks are often developed in order to assist operators making complex decisions. This might be due to the fact that Bayesian networks have some important strengths compared to many other techniques for data analysis. Johansson and Falkman (2007) mention that Bayesian networks are good at handling incomplete data sets, that it is possible for experts to incorporate their knowledge of the domain into the model, as well as it is easier for humans to understand and verify the Bayesian model, due to its graphical nature, compared to many other opaque models such as artificial neural networks. How the output generated from Bayesian network can be explained to an operator is the focus of the next subsection.

2.6 Explanation properties and methods

This subsection presents the results from the literature analysis conducted. Subsection 2.6.1 gives a brief presentation of what is meant by an "explanation" in the context of expert systems and Bayesian networks, as well as a depiction of the properties that such an explanation can have. Subsection 2.6.2 gives a short description of the methods and techniques identified that can be used in order to explain the outcome of a Bayesian network to a user.

2.6.1 Explanation properties

According to the online dictionary Merriam Webster[2], to explain something to someone is to make the object of the explanation plain or understandable for the recipient or to give reasons for or the cause of the object to be explained. Lacave and Díez (2002) elaborate on this definition and claim that to explain something to

[2] http://www.merriam-webster.com/dictionary/explain

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., so that he or she can improve his or her knowledge about the object. They further claim that the explanation must also be satisfactory so as to cover the receiver's expectations. Table 1 depicts some properties that, according to Lacave and Díez (2002), can be used in order to describe different characteristics of explanations obtained from Bayesian networks. They argue that an explanation can be characterized according to its content, how it communicates its results to the user, as well as if the explanation can be adapted so as to fit different categories of users, for example expert users and novice users. These explanation properties will be the focus of this subchapter.

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

Table 1 – Properties of explanations (after Lacave and Díez, 2002).

2.6.1.1 Content

Lacave and Díez (2002) claim that existing explanation methods for Bayesian networks can be classified into three groups with respect to the focus and content of the explanation. The first group of explanation methods is called "*explanation of evidence*" and is concerned with finding the most probable configuration of the unobserved variables in the Bayesian network in order to obtain the most probable explanation(s). Through this inference process, calculations are made in order to identify the variables that, according to some selected criterion for the algorithm, can explain the evidence and be considered possible causes of the evidence (Lacave and Díez, 2002). The second group of methods is called "*explanation of the reasoning process*". These methods are concerned with explaining how the system has come up with its results, which can make it easier for a user to improve the reasoning process of the system. This kind of method can also be used in order to explain to a user why the system did not produce a different conclusion expected by the user, i.e. which findings that oppose to the conclusion expected by the user. The third and last group of explanation methods is called "*explanation of the model*" and is concerned with showing the user the contents of the knowledge base that the system is based on. The objective of this kind of explanation is to permit a human expert to analyze the content of the system's knowledge base during the construction phase of the system or to offer a novice user some knowledge about the domain (Lacave and Díez, 2002). Nielsen et al. (2008) claim that this kind of explanation offers insight into the static components of the network such as independence relationships between variables, causal mechanisms, etc.

Lacave and Díez (2002) further categorize the different explanation methods as either *descriptive* or *comprehensive*, depending on the content of the explanation. A descriptive explanation consist of presenting the system's knowledge base to the user, providing details about the conclusion or displaying intermediate results of the reasoning process. A comprehensive explanation, on the other hand, tries to give the user a picture of the implications of the model, the conclusions that the system has come up with as well as how the different findings are related to the conclusions (Lacave and Díez, 2002). An explanation can also be presented with different levels of detail. Lacave and Díez (2002) present two such levels: micro level and macro level. An explanation at the micro level presents detailed information about the variations in the different nodes. A rule-based expert system presenting explanations at the micro level would consist of analyzing variables that are members of a certain rule, or the rules containing a specific variable (Lacave and Díez, 2002). An explanation at the macro level gives a presentation of the main lines of reasoning within the system, i.e. the paths that lead from the evidence to the specified conclusion. A rule-based expert system presenting explanations at this level would consist of one or several chains or rules (Lacave and Díez, 2002).

When explaining phenomenon with the help of Bayesian networks, the concept of causality is important. A Bayesian network can either be looked upon as causal or non-causal. A BN is said to be causal if all of the directed edges between the nodes in the network are causal. For example, the link A → B is a causal link when A is the cause of B (Lacave and Díez, 2002). In a non-causal Bayesian network, the links do not represent these dependencies. Though, it should be noted that not all situations can easily be described in terms of cause and effect relations. As Jensen and Nielsen (2007) claim, causal relations are not always obvious and the concept of causality is not well-understood (a detailed description of causality in Bayesian networks is out of the scope of this thesis, but more details can be found in (Halpern and Pearl, 2001)).

According to Lacave and Díez (2002) there are many reasons for using causal models in the context of probabilistic expert systems. One of the main reasons is that humans tend to interpret happenings in terms of cause and effect relations; thus a causal model makes it easier for humans to construct, modify and understand the model (Lacave and Díez, 2002). There is also a close relationship between causality and probability. If one knows how events and objects are related to each other, causality can provide a pattern of probabilistic dependencies between the nodes in the network, which also provides clues about causality in the network (Lacave and Díez, 2002). Bayesian networks are also suitable for representing causal relationships between nodes since their axiomatic properties (d-separation and the Markov property) are able to represent probabilistic dependencies and independencies that are a part of causal domains (Lacave and Díez, 2002).

2.6.1.2 Communication

Different expert systems have different ways of presenting the explanations generated. In some systems, the user must be an active component in the explanation process by, for example, posing questions to the system or by selecting some option from a menu that the application provides, while other systems present the results from the inference processes automatically. Furthermore, some systems are able to present the explanations generated during the ongoing inference process, while others are only able to present them after the inference process is finished (Lacave and Díez, 2002). The explanations can also be presented in different ways; either numerically with text and numbers, graphically with diagrams or through some multimedia medium such as videos, sounds and images. The probabilities of the different variables contained in the explanation can also be presented in different ways: they can either be numerical or quantitative, such as 0.47 or 78% or linguistic or qualitative, like expressions such as "seldom", "very likely" and "almost sure". Based on the literature review of this thesis, two different methods for communicating the explanations generated have been identified: (1) *text based explanations* and (2) *tree based explanations*, which are presented below.

Text based explanation

One way of presenting the explanations generated by an expert system, based on a Bayesian network, is through plain text that guides the user through the reasoning of the system. This kind of explanation is implemented in the Banter software, a tool developed in order to train users making medical diagnosis as well as choosing the most optimal diagnostic procedures (Haddawy et al., 1997). Haddawy et al. (1997) claim that this kind of explanation does not require the user to know anything about Bayesian networks in order to interact with it effectively, though it is helpful if the user has some knowledge of the particular domain as well as a basic understanding of probability theory.

Suppose that the system presents a patient with a number of different symptoms to a user, where one of the symptoms is the presence of gallstones. The user is then interested in diagnosing the presence of this symptom and wants the system to compute which tests that should be conducted in order to explain the symptom. The user might also want to have an explanation of how the system came up with these tests. In Banter, the user can press an "Explain" button, which generates an explanation based on known history and physical findings that influence the probability of the patient having gallstones. The explanation created by Banter, taken from the article by Haddawy et al. (1997), is presented below:

```
The best test to rule in GALLSTONES is CT.

The best test to rule out GALLSTONES is ULTRASOUND FOR GALLSTONES.


Before presenting any evidence, the probability of GALLSTONES being present
is 0.128.

The following pieces of evidence are considered important:

   o  Presence of GUARDING results in a posterior probability of 0.175 for
      GALLSTONES.
   o  AGE of 41 results in a posterior probability of 0.172 for GALLSTONES.

Their influence flows along the paths:
   o  GUARDING is caused by CHOLECYSTITIS, which is caused by GALLSTONES.
   o  AGE influences GALLSTONES.

Presentation of evidence results in a posterior probability of 0.227 for the
presence of GALLSTONES.


The best tests to rule in GALLSTONES (in order):

   o  A positive CT test results in a probability of 0.987 for GALLSTONES.
   o  A positive ULTRASOUND FOR GALLSTONES test results in a probability of
      0.601 for GALLSTONES.
      (…)

Their influence flows along the following paths:

   o  GALLSTONES are seen by CT.
   o  GALLSTONES are seen by ULTRASOUND FOR GALLSTONES.
      (…)


The best test to rule out GALLSTONES (in order):

   o  A negative ULTRASOUND FOR GALLSTONES test results in a probability of
      0.016 for GALLSTONES.
   o  A negative CT test results in a probability of 0.058 for GALLSTONES.
      (…)

Their influence flows along the following paths:

   o  GALLSTONES are seen by ULTRASOUND FOR GALLSTONES.
   o  GALLSTONES are seen by CT.
      (…)
```

Figure 6 – Example of a text based explanation (after Haddawy et al., 1997)

Another text based explanation is called "*Scenario based explanations*". Scenario based explanations, as described by Druzdzel and Henrion (1990), describe a sequence of events, i.e. the outcomes of all relevant variables in the scenario, which often forms a coherent, causal story that people easily can understand. The possible scenarios, or explanations, are then divided into two groups: those *compatible* with the hypothesis and those *incompatible* with the hypothesis, where those compatible with the hypothesis are those that include the outcome in questions. Thus, in order to generate an explanation in each case, the nodes in the network that are directly or indirectly relevant to the hypothesis of interest must be identified. The scenario based explanation method is then used in order to choose the most probable scenarios that together account for the most probable explanation of the hypothesis.

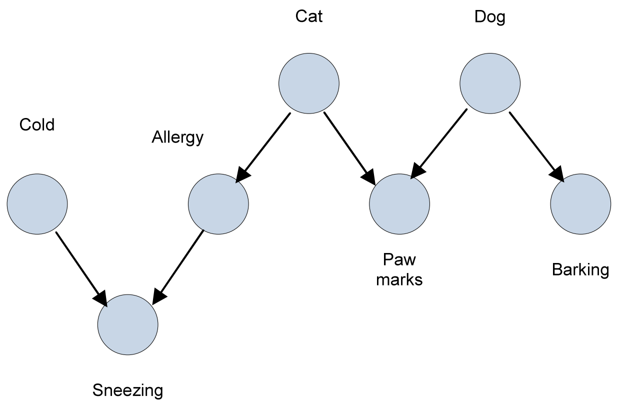


Figure 7 – A BN illustrating relationships between variables (after Druzdzel and Henrion, 1990).

Many variables can be excluded from the explanation with the help of the concept of d-separation. Variables can also be deleted directly. If one wants to explain why someone is sneezing, and if a cat has been observed, then the presence of paw marks, dog or barking are all made irrelevant, due to the structure of the network (see figure 7 above). This makes is possible to simplify the inference process (Druzdzel and Henrion, 1990). Below, a scenario based explanation is given. It explains the probability of a cold given that sneezing, paw marks and barking have been observed.

```
? (why ‘cold)

Given:

Sneezing must have been caused by cold or allergy.

Paw Marks could have been caused by cat or dog or another unknown cause.

Marking must have been caused by dog.

Scenario(s) compatible with cold:

        a. No cat, therefore no allergy, cold and therefore sneezing.     0.38
        b. Cat, therefore allergy, cold and therefore sneezing.           0.05
           Other less probable scenario(s)                                0.01
           Total probability of cold                                      0.44

Scenario(s) incompatible with cold:

        c.  No cold, cat, and therefore allergy, and therefore sneezing   0.56

Therefore cold is almost as likely as not (p=0.44).
```

Figure 8 – A scenario-based explanation (after Druzdzel and Henrion, 1990).

The explanation starts by listing the “Given”, i.e. the observed evidence and its relevance. Then, it gives two lists of scenarios: those compatible and those incompatible with the hypothesis.

Tree based explanation

Another way of presenting inference explanations to a user is to make use of the tree based structure of the underlying Bayesian network. One example of a tree based explanation can be found in the article by Nielsen et al. (2008), where the use of the Causal Explanation Tree model (CET) is demonstrated.

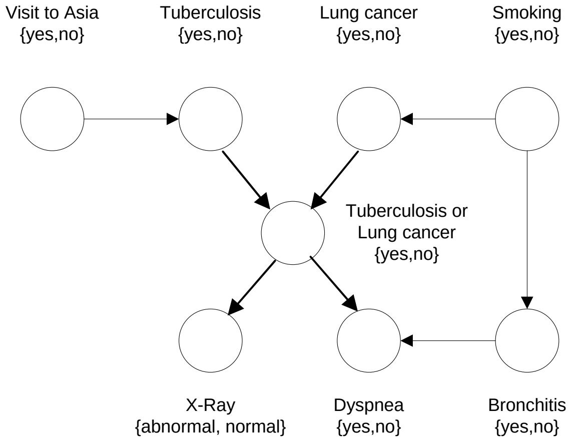


Figure 9 – A Bayesian network presenting the relations between diseases and symptoms (after Nielsen et al. 2008).

The Bayesian network presented above models the relationships between two indicators, “X-ray” results and “Dyspnea” (when someone has trouble breathing), of diseases that a patient can have. By analyzing the structure of the network, it becomes apparent that “Tuberculosis” is more likely to be a part of the explanation if the patient has visited Asia, and that it is more likely that the patient has lung cancer if he or she smokes. Both these variables increase the risk of having abnormal X-ray results and dyspnea. Bronchitis also increases the risk for dyspnea. (In this case, the node “Tuberculosis or Lung cancer” is just a modeling artifact and will not be considered in the explanation.)

Assume that we want to explain why a patient has dyspnea, given that the patient is a smoker. The tree based algorithm has selected smoker and bronchitis to be the largest contributing factors for why the patient has dyspnea (see figure 10 below). The next step in the explanation gives either the presence of lung cancer or the presence of tuberculosis as equal candidates for why the patient has dyspnea since, due to the numbers presented in the explanation tree, these candidates reveal the same amount of information for the state we want to explain, i.e., that the patient has dyspnea. (Together with the CET algorithm, the numbers present in the explanation tree represent how much causal information the variables share with the explanandum. A greater number indicates a more accurate and more probable explanation for the explanandum. The numbers presented in a generated explanation tree are to be compared only to the rest of the numbers within a specific tree in order to determine the influence each variable has on the generated explanation.) By illustrating the explanation with the help of the tree structure, the user can follow the reasoning of the expert system.

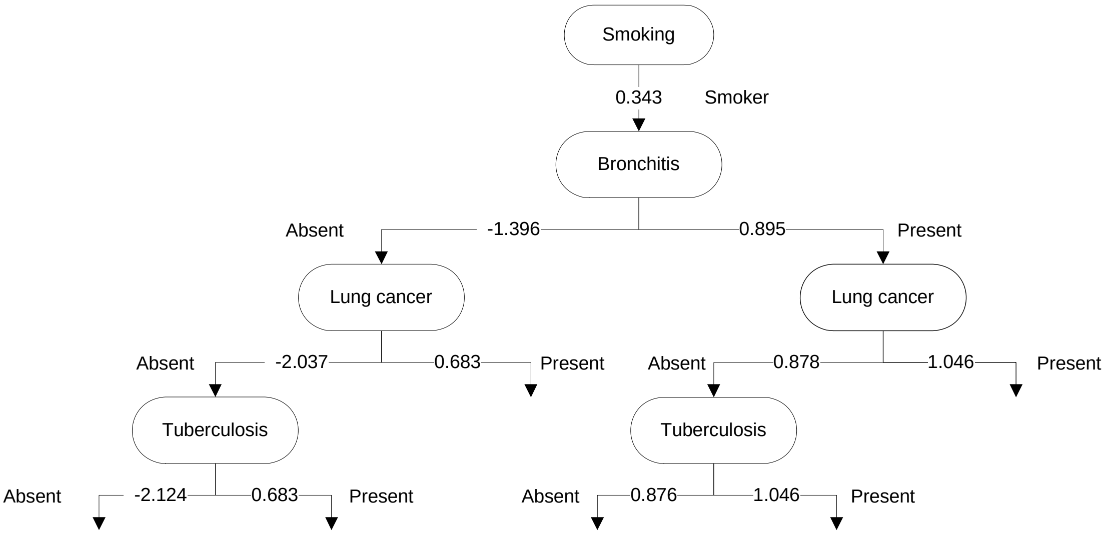


Figure 10– A CET illustrating the explanation to why the patient has dyspnea, given that he or she is a smoker (after Nielsen et al. 2008).

In order for a user to understand the conclusions that the tree based model has come up with, the user does not need to have any knowledge about the mathematics behind the model. Though, in order to briefly validate the results, the user should be able to analyze if the results are conceivable, for example, is it at all possible that smoking is a contributing factor for a patient to suffer from dyspnea?

2.6.1.3 Adaption

Different expert systems also differ in their ability to adapt the generated explanation to different kinds of users. It might be the case that a novice user needs a more exhaustive explanation than an expert user, who is familiar with the kind of problem or domain of interest. Lacave and Díez (2002) claim that one of the key features of an effective explanation is the ability to adapt the explanation to each user's specific needs and expectations. In order to identify the user's needs, one has to analyze how acquainted the user is with the domain as well as how familiar he or she is with the reasoning method behind the explanation. The result of the user analysis often ends up in a question of how much details should be included in the explanation or not.

Lacave and Díez (2002) further claim that an expert system can either have *static* or *dynamic* properties when it comes to adapting the explanation presented to the user. A static system is one that does not adapt itself to different kinds of users, while a dynamic system often possess a dynamic model for each user, that evolves over time and presents explanations accordingly. Though, it is also possible for an expert system to offer explanations according to different static modes, for example novice users, average users and expert users. The user is then requested to identify himself/herself with one of the modes and receives explanations accordingly.

2.6.2 Existing explanation methods and techniques

This subsection gives a brief introduction to the different terms and concepts used by researchers within the area as well as an overview of the existing methods and techniques that can be used in order to explain the reasoning outcome of a Bayesian network. The methods and techniques have been divided into two groups, based on the findings of the literature review: (1) *abductive methods* and (2) *tree based methods*, which are presented briefly in the sections below.

2.6.2.1 Explanation concepts

There are several methods that can be used in order to explain the outcome of a Bayesian network 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 the variables available in the explanation, while others only include a subset of all possible variables. Nielsen et al. (2008) divide the different variables constituting an explanation into three different groups: the *observed variables*, the *explanatory variables* and the *explanandum*. The observed variables are those variables in the environment whose states are known. Though, it might be the case that only a subset of these 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 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. The distinction between the different groups of variables in the network is useful in order to understand the differences and similarities between the different explanation methods presented in the following subchapters.

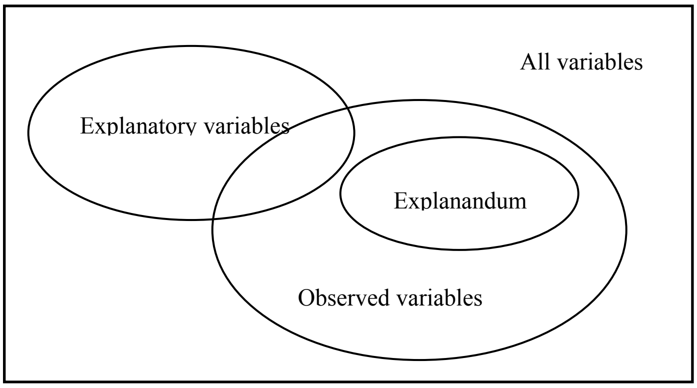


Figure 11 – Depiction of the different variables that constitute an explanation (after Nielsen et al., 2008).

2.6.2.2 Abductive inference: Total and Partial Abduction

Flores (2005) claims that abductive inference methods can be used in order to find and generate explanations to some observed facts. When using this kind of method, the best explanation for the given evidence is the configuration, or state of the world, that is the most probable given the evidence. Thus, given a set of observations or evidence, the abductive inference methods aim to obtain the best configurations of the values for the explanatory variables (the *explanation*) that are consistent with the explanandum and can be assumed to predict it (Flores, 2005). Campos et al. (2001) claim that finding plausible configurations, given the set of observed evidence, can be represented by the following inference rule:

$$\frac{\psi \rightarrow \omega, \omega}{\psi}$$

The rule depicts the relationship between two variables: ω and ψ. If we observe ω, and have the rule $\psi \rightarrow \omega$, then we can infer that ψ is a plausible hypothesis (or explanation) for the occurrence of ω (Campos et al., 2001).

It is often the case that there are several different hypotheses that can constitute valid explanations, and it might be necessary to choose among these in order to select the best explanation or the k best explanations. Thus, Campos et al. (2001) claim that the abductive way of finding the best explanations can be divided into two phases: *hypothesis generation* and *selection of hypothesis*.

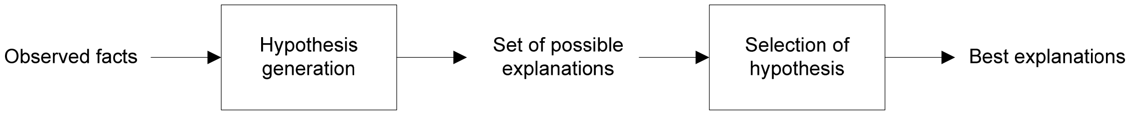


Figure 12 – The different phases of abductive inference (after Campos et al., 2001).

In order to select the best explanations from those generated, two kinds of criteria are used: the *metric based criteria* (such as probability, weight etc.) and the *simplicity criteria* (the best explanation is the simplest hypothesis available). According to Campos et al. (2001), the simplest explanation is, in this case, the hypothesis with the fewest number of variables involved.

Flores (2005) claims that two main abductive tasks can be identified: *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 by Flores (2005) is defined as the "over specification problem". Moreover, as Nielsen et al. (2008) 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. The decision of which variables to include in the explanation set or not can either be made manually by a user or via an automated analysis of the network. Though to decide which variables to include or not is not a nontrivial issue (Nielsen et al., 2008). Some researchers claim that all variables should be provided as input to the explanation set, as in the MPE method, while others claim that the Bayesian network should be considered to be a causal one and thus only include ancestors of the explanandum in the explanation set. Others claim that the principle of Occam's razor should be applied and thus only include those variables in the explanation set that have the most influence on the complete explanation (Flores, 2005).

Nielsen et al. (2008) claim that an additional drawback of the abductive inference methods is that they do not distinguish between the explanandum and the observations, which results in that additional state information that is not meant to be explained, is excluded from a possible explanation. For example, consider the Bayesian network illustrated in figure 9 in subsection 2.6.1.2, where the relationships between the variables "Dyspnea", "Lung cancer", "Smoking" etc. is depicted. Assume that the variables "Smoking = yes" and "Dyspnea = yes" have been observed and that one wants to explain the presence of "Dyspnea = yes". Together with an algorithm that does not distinguish between the observations and the explanandum, one can either input those observations to the algorithm or only the variable that one wants to have explained, in this case "Dyspnea = yes". Though, in the first case, the algorithm will try to explain the joint fact that "Smoking = yes" and "Dyspnea= yes", which is not what one wants to explain in this case. In the second case, where only the variable "Dyspnea = yes" is used as input, information is lost about the state of the network since one has discarded one observation that not only can change the importance of the smoking variable, but all of the other variables in the network as well which are (unconditionally) dependent on the smoking variable and have a directed causal path to the variable "Dyspnea". Thus, not making a difference between the observed variables and the explanandum reduces the power of the explanation algorithm, in that they cannot fully capture the information that one has about the network, according to Nielsen et al. (2008).

Nielsen et al. (2008) further claim 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. (2008) 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.

According to Nielsen et al. (2008), partial abduction is computationally more expensive than standard MPE, though the explanations generated are often more concise. There are several different versions of the abductive inference methods. One variant of the partial abduction method was developed by Campos et al. (2001) where the k most probable explanations are found and simplified according to their criteria of relevance and probability measure. Another variant was developed by Henrion and Druzdzel in the 1990's where partial assignments are allowed but only within a predefined tree that sets the limit of possible explanations (Nielsen et al., 2008). The explanation generated by Henrion and Druzdzel is called a scenario-based explanation, since the explanation is a path from the root of the tree to a leaf, denoting variable assignments for each branch in the tree, and where the best explanation is the branch with the highest probability (Nielsen et al., 2008).

2.6.2.3 Explanation Trees

Another method for explaining the outcome of a Bayesian network to a user is the *Explanation Tree* (ET) method, which aims at finding the best explanation(s) for the observed variables. The explanation generated with this method is represented in a tree structure, thereof the name of the method. In the tree, every inner node denotes a variable of the explanation set and every branch from these nodes indicates an instantiation of the variable to one of its possible states (Flores, 2005). Thus, a path from the root of the tree to one of its leaves is a series of assignments of the different nodes involved, which constitutes a full explanation (Nielsen et al., 2008). The values of the nodes are called their configurations, and the resulting explanation tree will present the probability of the possible configurations of the nodes, and their probabilities given the evidence. The set of explanations will be the set of configurations associated with the leaves of the explanation tree, ordered by their posterior probability given the evidence (Flores, 2005). Below, an explanation tree is presented (after the scenario depicted in subsection 2.6.1.2, figure 9).

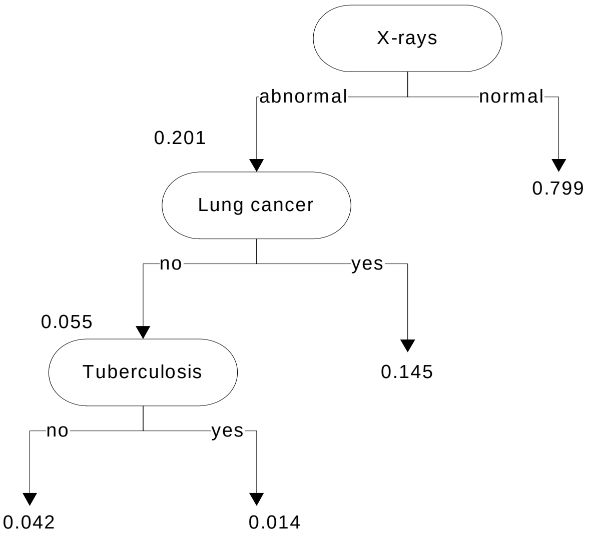


Figure 13 – An example of an explanation tree (after Nielsen et al., 2008)

The figure above depicts an explanation tree generated together with the ET algorithm when one wants to explain why a patient has dyspnea. The three variables “X-rays”, “Lung cancer” and “Tuberculosis” belong to the explanation set, since they are nodes in the tree (Flores, 2005). The first valid explanation for why a patient has dyspnea is, in this case, that the variable “X-rays” has the state “normal”. Though, if this is not the case, the variable “Lung cancer” can provide another explanation. Following the path, one can see that adding “Lung cancer = yes” to the explanation is also a valid explanation. Otherwise, “Lung cancer” takes on the value “no” and the node is expanded to “Tuberculosis”, whose configurations gives us two valid explanations.

Among all the explanations presented by the final tree, the best explanation is the one with the largest posterior probability (Nielsen et al., 2008).

When building the explanation tree, two factors have to be considered. The first factor deals with which variable to choose as the next variable to split on, and the second factor is concerned with when one should stop growing the tree (Flores, 2005). According to the Explanation Tree 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. In order to calculate this, two measures can be used: the *mutual information* and the *Gini index*[3]. Once a variable has been selected as the next node, a decision must be made whether to expand the current node or not. Again, one of the two measures can be used in order to calculate this (Flores, 2005). (For the interested reader, see the appendix of this thesis for a pseudo code presentation of the ET algorithm.)

Nielsen et al. (2008) claim that the Explanation Tree method is a good method to use when it comes to presenting different competing and mutually exclusive explanations in a compact and readable way, since it makes use of the tree structure to present the different explanations. Though, Nielsen et al. (2008) argue that the ET method has several drawbacks. The first drawback concerns the non-distinction between the constructed path and the explanandum. The variables are added to the explanation tree according to how much information they provide about the remaining variables in the set of explanatory variables. Though, according to Nielsen et al. (2008), this does not measure the information that these added variables share with the explanandum. Thus one cannot say that the variables chosen reduce the uncertainty of the explanandum (Nielsen et al., 2008). The second drawback, as can also be found together with the abductive inference methods, concerns the non-division of the variable(s) constituting the explanandum and the observed variables. Moreover, Nielsen et al. (2008) claim that the ET method fails at finding the best explanations to the given evidence. The ET algorithm chooses as the best explanation, given the evidence, the path that has the highest probability. Though, since several explanations can include many variables from the explanation set, and only one will be included in the explanation tree, the algorithm will often miss explanations which might seem more appropriate for the current situation, but did not cover as large a fraction of the variables in the explanation set (Nielsen et al., 2008). Furthermore, Nielsen et al. (2008) claim that an algorithm of this type should take causal dependencies and independencies into account, though the ET algorithm makes no distinction between ancestors and descendants of variables.

2.6.2.4 Causal Explanation Trees

In contrast to the Explanation Tree method, the *Causal Explanation Tree* (CET) method, as described by Nielsen et al. (2008), requires a causal Bayesian network in order to generate valid explanations. Here, the causality means that all the directed edges in the network depict cause-effects relationships between variables. An explanation is, according to Nielsen et al. (2008) an assignment of the explanatory variables that is compatible with the variables in both the explanatory set and the set of observed variables (see figure 11 in subsection 2.6.2.1).

[3] See Flores (2005) for more information about the two measures.

Nielsen et al. (2008) insist that there is a difference between the observed variables and the explanandum. The observations constitute all our knowledge about the current state of the system and this might not be what we want to have explained. An algorithm respecting this division should thus determine for each observed variable if its state is relevant when it comes to explaining the explanandum, and for each unobserved variable, find out if knowing its state adds "explanatory power" to the proposed explanation (Nielsen et al., 2008).

Like the Explanation Tree method, the Causal Explanation Tree method utilizes the tree representation in order to present the different explanations to the users. All the explanations in a path of the tree are causal, since the algorithm only selects those variables that causally influence the explanandum to be members of the explanation set (Nielsen et al., 2008). Not all Bayesian networks can be considered causal though, and before using the network together with the CET algorithm, the network must be checked for causality, since the algorithm assumes a causal network. The CET algorithm also assumes that its corresponding joint probability distribution is faithful and causally sufficient. Faithfulness of the distribution ensures that there is a unique graph whose arcs depict all conditional dependencies of the distribution, and only those (Nielsen et al., 2008). That the distribution is causally sufficient means that no hidden variables are permitted in the network, so that the arcs of the network built represent direct causation.

One important component of the CET algorithm is the *do-conditioning*. The expression P(X|Y=y) is an illustration of the probability of X, given the instantiation of Y. The difference between this expression and the P(X| do(Y=y) is that in the do case, the instantiation Y=y is imposed on the system, instead of being just an observation. As an example, consider the binary variables Raining (R) and Wet streets (W). The causal network associated with these variables can be illustrated by the expression R ⟶ W. If one observes that it is raining, then this will probably change the probability of the streets being wet. Similarly, if one sees that the streets are wet, this will probably change the probability values for the variable Rain. Though, if one forces the variable Rain to have the value "yes", i.e. one makes it rain somehow, this will also change the probability of the streets being wet since "Rain" has a causal effect on "Wet streets". On the other hand, if one forces the variable "Wet streets" to have the value "yes" (i.e. if one spills some water on the street), this will not influence the probability of the variable "Rain", since W has no causal effect on R. This is due to the causal direction of the arcs of the network.

The causal information flow measure is another important component of the CET algorithm. This is a measure that calculates how much shared causal information that two variables share with each other, and is illustrated by the following expression: I(X->Y). This measure is used when building the causal explanation tree in order to decide which variables that should be added to the tree, since it depicts how much causal contribution the variables give to the explanandum (Nielsen et al., 2008). The causal explanation tree is then built recursively, starting with selecting as root node the variable that has the maximum causal information flow to the state of the explanandum. For each possible value of the selected node, a branch is added to the root. (For the interested reader, see the appendix of this thesis for a pseudo code presentation of the CET algorithm.)

3 Problem

This section will describe the problem domain, followed by a more detailed description of the problem at hand, the aim of the project as well as the objectives identified to achieve the aim.

3.1 Problem domain

In pace with the globally expanding maritime industry, the maritime surveillance capacity must be further developed as well in order to detect anomalous behavior such as terrorist attacks, smuggling of humans and goods and hazardous cargo transports (Høye et al., 2008). Bomberger et al. (2006) confirm to this opinion and claim that the detection of unusual vessel activity is an important homeland security issue.

In order to identify these unusual events, anomaly detection techniques can be used. Anomaly detection can be described as the process of detecting deviations from normality. According to Riveiro et al. (2008) most anomaly detection methods first build a model of normal behavior which is then 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 its methods have also been introduced by researchers within the maritime domain.

One problem associated with maritime surveillance of coastal regions is, according to Roy (2008) and Riveiro et al. (2008), the massive amounts of data that has to be processed in order to detect anomalous events and objects. Riveiro et al. (2008) further claim that the large amounts of heterogeneous data from multiple sources also make the surveillance systems complex in both function and structure. Monitoring the system 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 uncertain information (Riveiro et al., 2008). As a result of this, 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 function of the system. If the operator understands how the system works and feel confident in using the system, his or her trust in the system might increase as well. However, many such models are difficult to understand and interpret. One exception though might be the Bayesian network approach for detecting anomalous vessel behavior. Johansson and Falkman (2007) claim that the Bayesian network technique has two main advantages compared to opaque machine learning techniques, namely that is it possible to include expert knowledge into the Bayesian model as well as for an operator to understand and interpret the model representing the situation, due to its graphical nature.

Jensen et al. (1995) state that decision support systems should have features for explaining how it has come up with its recommendations in order to support the decision maker as well as increase his or her confidence in the system. Lacave and Díez (2002) confirm to this opinion and 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. Lacave and Díez (2002) 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 Lacave and Díez (2002), satisfactory for the end-users according to the identified explanation characteristics (presented in subsection 5.1). Lacave and Díez (2002) further claim that many of the explanatory solutions identified so far have not been tested on practical examples, which limits their scope.

This thesis investigates if and how a Bayesian network approach for identifying anomalous vessel behavior can be used in order to support the operators in their decision making process, even in the presence of loads of data. If the situation awareness of the operators can be improved as well as their understanding of how the anomaly detection system works, the operators will be better prepared and qualified to improve the anomaly detection system itself by, for example, fine tuning the different threshold values used by the system in order to classify an event or object as anomalous or not. Thus, the input that the operator gives to the system may reduce the false alarm rate or increase the true alarm rate of the detector.

3.2 Problem description

The aim of this thesis is to investigate methods and techniques that can be used in order to comprehend and explain what has triggered the outcome of a Bayesian network when used in the problem domain of anomaly detection.

The objectives for achieving the aim are as follows:

- Investigate different explanation methods, what an “explanation” is in this context as well as investigate what properties an explanation may have.
- Identify which techniques and methods that are used today in order to comprehend and explain what has triggered the outcome of a Bayesian network.
- Study the challenges related to using these techniques in order to depict an anomaly in an understandable and satisfactory way according to an operator’s point of view.

4 Method

In this section, the methods for each objective presented in subsection 3.2 are presented as well as a motivation for why the specified methods have been chosen.

4.1 Summary of methods

In order to achieve the objectives specified in subsection 3.2 two methods have been identified. The first method used is a literature analysis. This method has been chosen in order to investigate what an “explanation” is in the specified context, which explanation methods are used for Bayesian networks, and also to investigate what properties an explanation may have. The literature analysis is also conducted in order to investigate which methods and techniques exist for explaining the outcome of a Bayesian network. Moreover, experiments are conducted in order to study the challenges that the chosen techniques imply in order to depict an anomaly in an understandable way.

The two methods presented above are further described in the following two subsections.

4.2 Literature analysis

A literature analysis is conducted in order to give an introduction to the research made within the area. The literature analysis aims at describing important concepts within the domain as well as providing a basis for the experimentation (see subsection 4.3). During the literature analysis, an investigation of what constitutes an explanation in this context as well as a description of the different explanation methods that exist today is made. A general overview of different methods and techniques for explaining the outcome of a Bayesian network is also conducted. This overview results in the identification and selection of two methods that can be applied in the maritime domain during the experimentation of this thesis.

The most relevant papers for this thesis are found from the international conferences on “Information Fusion” as well as from the conference papers from “Uncertainty in Artificial Intelligence”. Papers are selected according to their relevance and quality for the current topic. No papers from the fields of psychology or cognitive science regarding how explanations should be given to users are considered for this thesis. Furthermore, only introductory papers from the field of network security are considered in order to obtain a basic understanding of the research made on anomaly detection. The databases used for conducting the literature analysis of this thesis are the ACM Digital Library, CiteSeer, Science Direct and Springer Link. These databases have been chosen as they contain full text articles in the field of, for example, computer science and artificial intelligence. In addition to the article databases listed above, Google Scholar is also used since it provides a way of searching for reviewed articles, thesises, books etc. The top 20 search results based on their relevance for the key words used as input are considered. The reference lists from interesting articles for the thesis found are also used. The main key words used are: “anomaly detection”, “situation awareness”, “explanation methods + Bayesian networks”, “abductive inference methods”, “Explanation Trees” and “Causal

Explanation Trees”. Related articles are found by reading the adhering abstracts and key words.

4.3 Experimentation

In order to carry out the experimentation of this thesis, real AIS data, obtained from Saab Microwave Systems, is analyzed. The data reflects five days of vessel traffic outside the coast of Gothenburg, and contains information about the different vessels, such as ID, position, heading and speed. In this data, three different types of anomalies are hidden: one vessel is speeding, one vessel is positioned far away from the other vessels in the data, and a combination of the two anomalies, i.e. one vessel is speeding far away from the other vessels. Based on this data, different Bayesian networks are generated, which are later on used as a basis for the two selected explanation methods. The two explanation methods chosen are the Explanation Tree method and the Causal Explanation Tree method (presented in subsection 2.6.2). These two methods were chosen for the experimentation phase of this thesis in order to test their performance on real AIS data. The two methods are considered in this thesis since they have been developed quite recently and their originators claim that these methods overcome many of the difficulties that have been subjected to other previous explanation methods for Bayesian networks (Flores, 2005 and Nielsen et al., 2008).

The experimentation analyzes the two techniques chosen for explaining the outcome of a Bayesian network in a practical setting, as well as their advantages and disadvantages according to the specification of what should constitute an explanation in this context. The experimentation method was selected in order to analyze the performance of the selected explanation methods in a maritime setting.

5 Experimentation

This section presents an overview over the experimentation conducted together with the Explanation Tree method and the Causal Explanation Tree method (presented in subsection 2.6.2). Subsection 5.1 will give a short presentation of the purpose of the experimentation, while the remaining subsections will present the three parts of the experimentation: the preparation of the AIS data, the creation of the Bayesian network and the deployment of the Explanation Tree method and the Causal Explanation Tree method.

5.1 The purpose of the experimentation

The purpose of the experimentation is to investigate if the Explanation Tree method, developed by Flores (2005), and the Causal Explanation Tree method, developed by Nielsen et al. (2008), can be used in order to explain the cause of an anomaly alarm to a user. These methods were chosen for the experiment since Nielsen et al. (2008) claim that the tree representation of different hypothesis is a good way of compactly presenting competing explanations to a user. The methods were also chosen because of their recently development and since their founders claim that these two methods have overcome many of the shortcomings of the abductive explanation methods.

In order to test the two tree based methods, three types of anomalies were added to three different AIS files. The first anomaly hidden affects only one variable in the BN, namely speed. The speed of one vessel was raised above normal in order to test if the explanation methods could identify this anomaly. The second anomaly concerns two variables: longitude and latitude (i.e. the values for the position variables were raised above normal) while the third anomaly hidden affects three variables in the network: speed and position. The experimentation thus aims to analyze if the two methods can identify these variables as the cause of the anomaly.

5.2 Preparation of the AIS data

The AIS data used in the experiment was provided by Saab Microwave Systems and covers 5 days of vessel traffic outside the coast of Gothenburg. 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. The separated files were then sampled down by deleting messages from the different vessels contained in the file: instead of having reports from the vessels every 2-3 seconds, an update every sixth minute has been considered. The different files were then further sampled down by deleting messages from vessels that did not follow the normal routes at this particular coast area. This sampling down was made in order to make the AIS file smaller and thus easier to perform calculations on. At this point it must be stressed that the purpose of this thesis is not to investigate if Bayesian networks can be used in order to find anomalies in data (for this purpose, the reader is directed to, for example, the article by Johansson and Falkman (2007)), but to test if the Explanation Tree method and the Causal Explanation Tree method can be used in order to explain the reason(s) for an anomaly alarm. The data has been preprocessed in order to be able to make a better analysis when investigating the outcome of the explanation method. The resulting file used in the experiment contains 44200 lines of vessel messages and about 300 unique vessels.

From the AIS data, the columns *MMSI* (the unique identity of the vessel), *receive time* (the time the message was received), *longitude*, *latitude*, *heading*, *speed* (measured in knots and abbreviated *SogKnots*), *course over ground* (*cog)*, *dimensions*, *draught* and *length* of the vessel were considered for the experimentation. An additional column "anomaly" was manually added to the AIS data in order to make it easier to analyze the outcome of the explanation methods chosen for the experimentation of this thesis. Three different kinds of anomalies were hidden in the data: speeding of one vessel, deviating position of one vessel and 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.

5.3 Creation of the Bayesian network

The preprocessed AIS data was then used as input to the Genie[4] application: an application developed by the Decision Systems Laboratory in Pittsburgh which can be used in order to create Bayesian networks from data. 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. The variables containing manually added anomalies were discretized using the "Hierarchical" method available in the Genie application, while the other variables were discretized using the "Uniform Widths" method. These two methods were used for the discretization in order to receive a good distribution of the different variables.

When creating the Bayesian networks to be used together with the Causal Explanation Tree algorithm, only the variables "longitude", "latitude", "heading", "speed" and "course over ground" were considered. 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. When creating the Bayesian networks to be used together with the Explanation Tree algorithm, all of the dynamic variables listed above were considered, though tests were made both with and without the anomaly variable. In order to make a fair evaluation of the two methods, these tests were made so as to ensure that the underlying BNs in some cases were the same.

The dynamic variables used as input to the explanation algorithms were chosen in order to make it easier to understand the created networks. These variables were also chosen due to the fact that all anomalies hidden in the AIS data were a result of a change of some of these dynamic properties.

[4] http://genie.sis.pitt.edu/

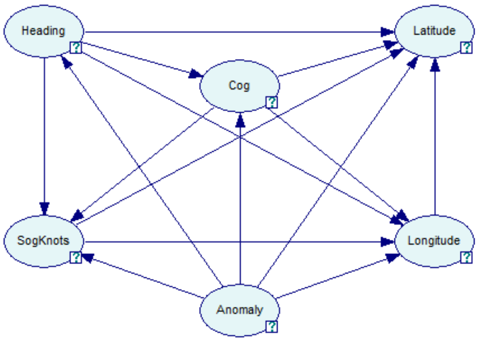


Figure 14 – An example of a Bayesian network created in Genie from the AIS data.

The network presented above depicts the relationships between the different variables when one of the vessels in the AIS data is marked as anomalous because of the speed of the vessel (the question marks present depict that the evidence for the different variables are not set). As the network shows, the dependencies between the variables are very strong, due to the many edges between the variables in the network. The network was created with the Essential Graph Search (EGS)[5] algorithm available in Genie and was later used together with the explanation tree algorithms. There are several algorithms that can be used in order to create a BN from data, and one of the algorithms available, in this case the EGS algorithm[6], was chosen for the experiment.

5.4 Deployment of explanation methods

After having created the Bayesian networks in Genie from the AIS data, the networks were used as input to the Matlab code, obtained through e-mail contact with Nielsen et al. (2008)[7] who had, for their own research, implemented the two tree based algorithms in order to analyze the quality of the explanations generated. Based on the BNs generated in Genie, the algorithms were used to calculate which variable(s) that, according to the specific criteria posed by the different algorithms, is/are to be present in the resulting explanation trees. The resulting explanation trees were later on visualized in the Graphviz application[8]; a free application developed by the AT&T Research Labs. Several different Bayesian networks were used as input to the Matlab code in order to test the quality of the two explanation tree methods.

[5] For more information about the EGS algorithm, see the article by Dash and Duzdzel (1999).

[6] Since the purpose of this thesis is not to build a Bayesian network from data, much time was not spent on analyzing and choosing one of the algorithms.

[7] Personal contact with the co-author of the article by Nielsen et al. (2008), Jean-Philippe Pellet, was made, which resulted in obtaining the Matlab code for the two tree based explanation methods.

[8] http://www.graphviz.org/

6 Results and discussion

This section presents a discussion about the results from the experiments made. Subsection 6.1 introduces the reader to how a generated explanation tree should be interpreted. The results from the experiments with the Explanation Tree method are presented in subsection 6.2, while the results from the Causal Explanation Tree method are presented in subsection 6.3. Furthermore, the two methods are evaluated according to the characteristics of the explanations generated as listed in subsection 5.1. The result of this evaluation is presented in subsection 6.4, while a summary of the results is given in subsection 6.5.

6.1 How to interpret the explanation trees?

In order to make it easier to understand and interpret the explanation trees presented in this section, an analysis of an explanation tree is given below.

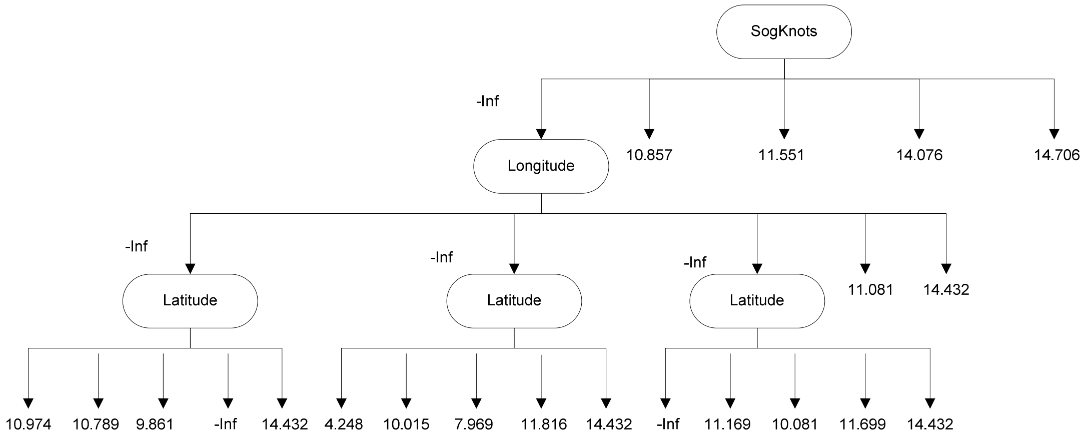


Figure 15 – An example of an explanation tree

As the tree above depicts, the variable that is the best explanation for the explanandum, i.e. that the variable “Anomaly” has the state “yes”, is “SogKnots” (the speed of the vessel measured in knots). The different numbers represent how much information the variable reveals 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. Due to the discretization of the different variables in the Genie application, one can furthermore see that the fifth discretization category of the “SogKnots” variable constitutes a slightly better explanation than the other discretization categories, i.e. that the vessel has a high speed 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 15, should be interpreted as a scale of the values from the AIS data, ranging from a low number for the “SogKnots” variable, i.e. slow speed, to a high number, i.e. high speed. (For example, the fifth discretization category of the speed variable encompasses speeds between 151.15 knots and more, in the case where one anomaly has been added to the AIS data, i.e. the speeding of one vessel. See the Appendix, table A1, for more information about the different discretization categories of the variables present where one anomaly has been hidden in the AIS data.) If the vessel is not speeding, the second best explanation for the state of the explanandum is constituted of the vessel having a high number for the "Longitude" variable.

The explanation trees generated with the ET algorithm with underlying non-causal networks have been sampled down in order to make them more readable, due to the discretization of the different variables in the network. The variables in the tree depicted in figure 16 originally have five different discretization categories, though with nearly the same numbers for all the categories. Common values of the different numbers presented have been selected from the generated explanation trees and are presented in the explanation tree in figure 16.

6.2 Results from the Explanation Tree method

Experiments have been made together with the Explanation Tree method in order to test its explanatory capacity in a maritime scenario. The experiments have shown that the critique against the method, as posed by Nielsen et al. (2008), is well founded since, compared to the Causal Explanation Tree method, the ET method generates less accurate explanations, especially in the cases where two or more anomalies were hidden in the data. Below, the explanation generated in the case where one vessel is speeding is presented. This explanation reflects the Bayesian network where the anomaly node was included in the Genie calculations (i.e. a non-causal network).

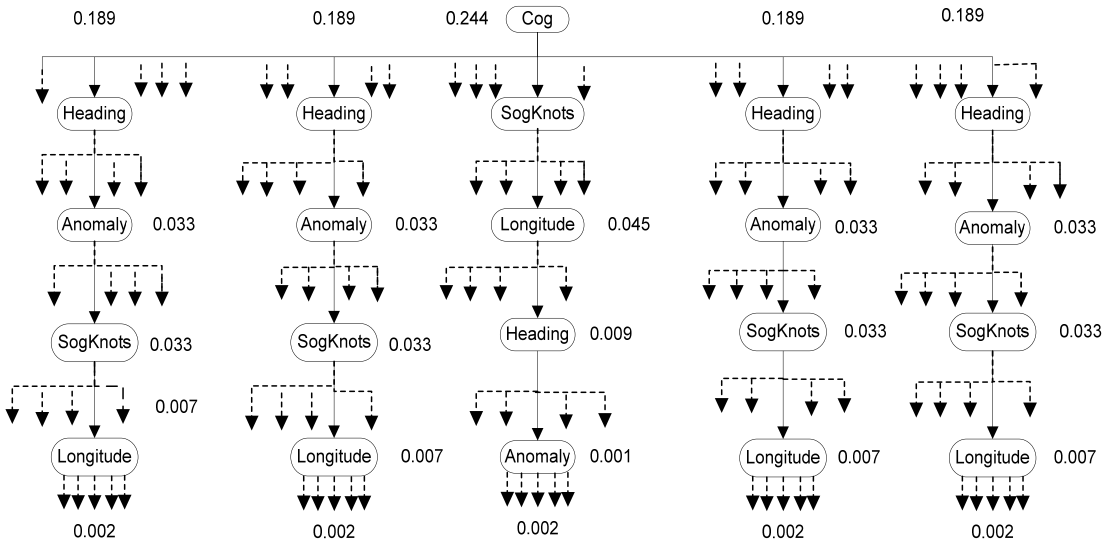


Figure 16 – An ET explanation for one vessel speeding (stopping criterion 0)

As can be read from the explanation tree above, 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”. Furthermore, it can be read from the tree that the third discretization category of the “Cog” variable reveals slightly more information than the other variables in the tree (0,244 compared with 0,189). 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 (a change of the stopping criterion for the ET method only affects how many levels of the explanation tree are presented).

The same network can be used in order to illustrate the explanation trees generated for the files containing two and three anomalies as well since, according to the ET algorithm, the variables present in the explanation tree presented in figure 16 above are still the ones that give the most information about the other variables in the network given the explanandum. There are some differences between the three different explanations generated, though the differences only concern some slight changes in the calculated values and are thus not considered to be of great importance for the experiment.

Below, an explanation tree generated together with the ET method where the anomaly node was manually added to the network is presented. The explanation depicts the anomaly where one vessel is speeding.

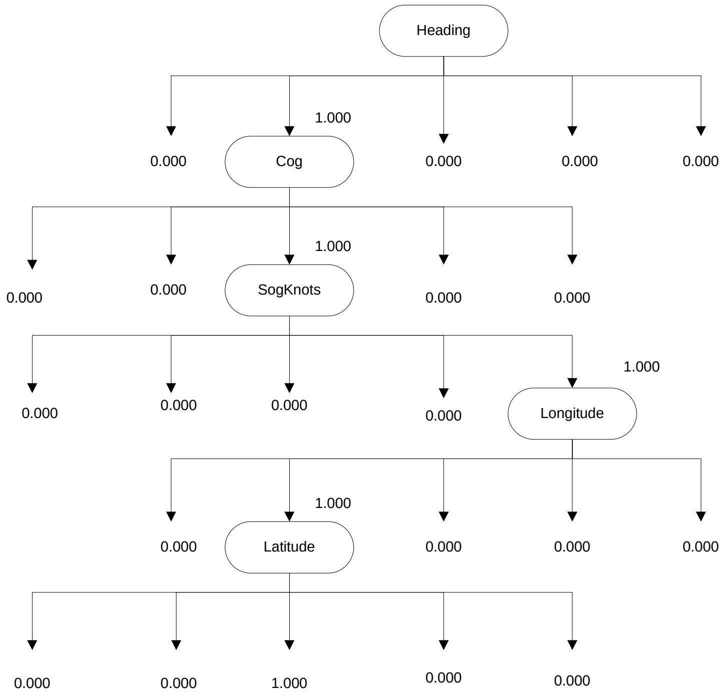


Figure 17 – An ET explanation for one vessel speeding with manually added anomaly node (stopping criterion 0)

With this underlying BN, the ET algorithm chooses the "Heading" variable as the variable that gives the most information about the other variables in the network. Thereafter it chooses the "Cog" (Course over ground), "SogKnots" (Speed measured in knots), "Longitude" and "Latitude" variables. As in the case with no manually added anomaly node, the same explanation presented above can also be used to depict the explanations generated for the cases where the position or the position and the speed of a vessel constitute the hidden anomaly, except from some slight changes in which of the five states of the variables that the algorithm chooses.

In comparison with the explanations generated where the anomaly node was included in the Genie calculations, as presented in figure 16, the explanations generated where the anomaly node was manually added to the network, generated worse explanations. Due to the static nature of the ET algorithm, the variables "Heading" and "Cog" are in all three anomaly cases chosen before the speed and position variables as the best explanation for the explanandum. The change of the underlying Bayesian network from a non-causal one to a causal one also makes the results of the ET algorithm more obvious (either 0 or 1 for the different discretization categories of the variables).

6.3 Results from the Causal Explanation Tree method

The CET method generally generates better explanations than the ET method and the cause and effect relationships make it easier to interpret the explanation generated. The generated explanations also seem to give better answers to the questions posed because of the cause and effect interpretation. The explanation trees below have been generated with stopping criterion 5 for the CET algorithm in order to make the generated trees more encapsulated and easier to interpret. Below, the explanation generated where one vessel is speeding is presented.

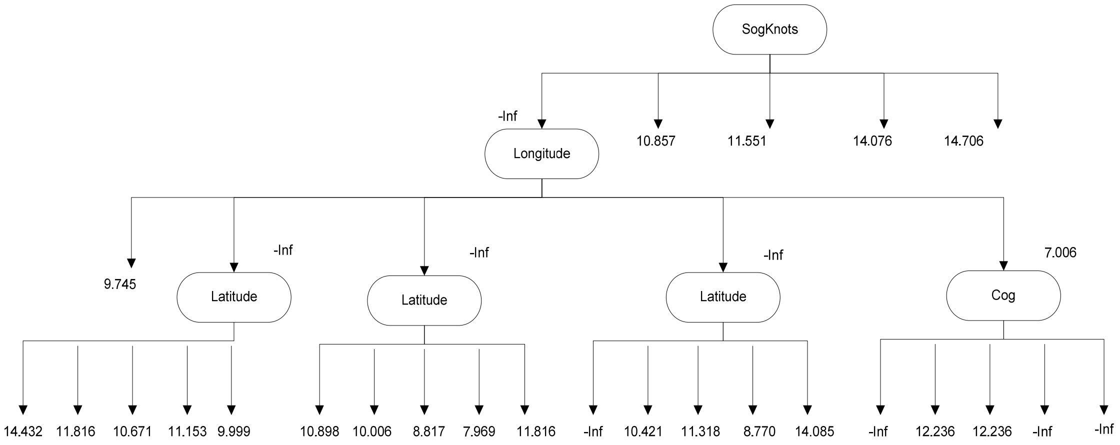


Figure 18 – A CET explanation for one vessel speeding (an anomaly based on one variable), (stopping criterion 5)

The CET method manages to identify the variable that is the manually hidden cause of the anomaly, namely "SogKnots" (the speed of the vessel measured in knots). The CET method chooses this variable since, according to the algorithm, it is the variable that shares the most causal information flow with the explananandum, i.e. "Anomaly

= yes”. One can also see from the tree that it is the fifth discretization category of the speed variable, i.e. very high speed, that constitutes the first best explanation. If it is not the case that the vessel is speeding, the “Longitude” variable is to be considered the most probable cause of the state of the explanandum.

Below, the causal explanation tree for the case where the position of one vessel is abnormal is presented.

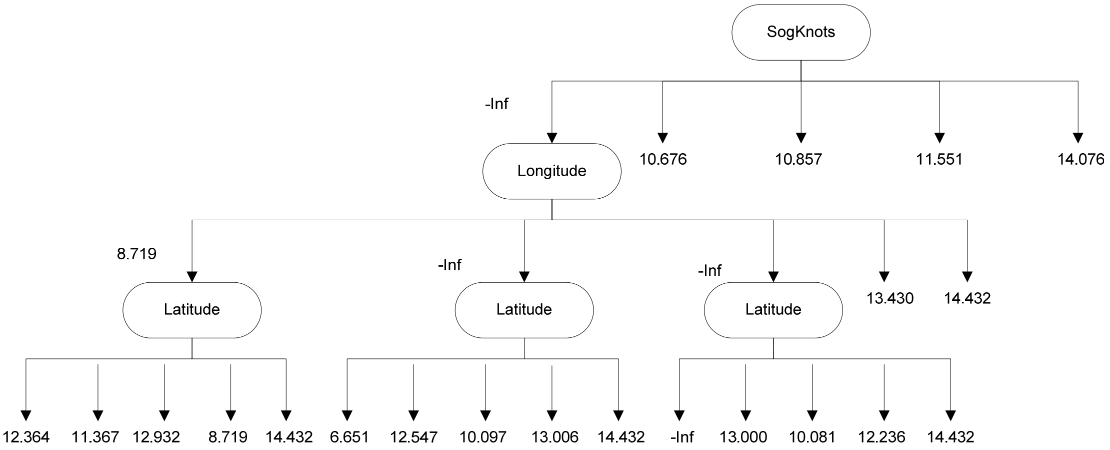


Figure 19- A CET explanation for abnormal position of one vessel (an anomaly based on two variables), (stopping criterion 5)

In the case where two anomalies have been hidden in the data, namely high values for the “Longitude” and “Latitude” variables, the CET method still performs quite well when generating the associated explanation tree. As in the case with one speeding vessel, the explanation tree for an abnormal position of one vessel first presents the “SogKnots” variable as the one that shares the most causal information flow with the explanandum. It can also be read from the tree that the highest discretization category of the speed variable, i.e. that the speed is high, is the first best explanation. If the vessel is not speeding, the next variable to consider is “Longitude” and thereafter “Latitude”. The tree generated representing an anomalous position of one vessel is thus not perfect, since it chooses the variable “SogKnots” before “Longitude” and “Latitude”, though the position variables appear as the second respectively the third best explanations for the explanandum.

Below, the causal explanation tree generated when a vessel is speeding in an abnormal position is presented.

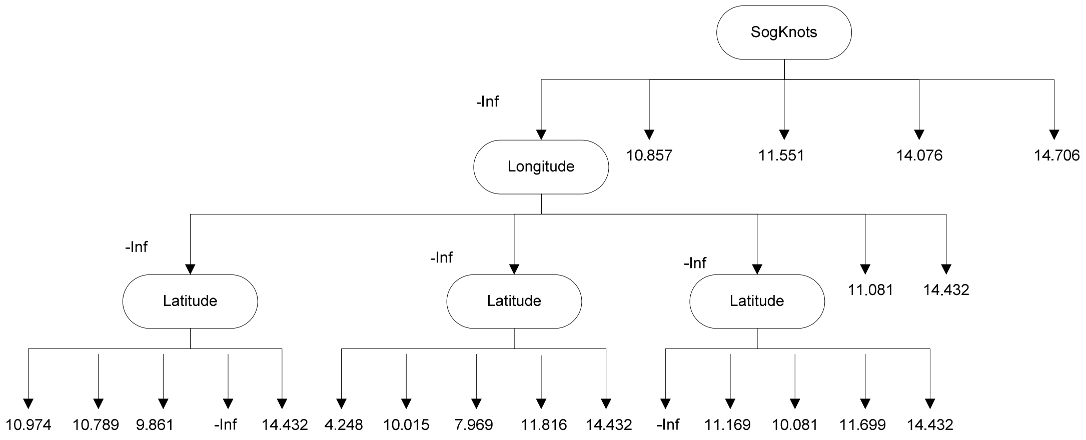


Figure 20 – A CET explanation for abnormal speed and position of one vessel (an anomaly based on three variables), (stopping criterion 5)

The explanation tree presented above 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.

6.4 Explanation characteristics

This subsection briefly analyzes the Explanation Tree method and the Causal Explanation Tree method according to their content, communication and adaption properties, as described in subsection 2.6.1. These properties are also evaluated according to their suitability for a maritime scenario.

6.4.1 Content

Both the ET and the CET method aim to explain the evidence of the reasoning process, i.e. to find the most probable configuration of the unobserved variables in order to obtain the most probable explanation(s). The purpose of the methods is thus to give a comprehensive overview over the conclusions that the two algorithms have come up with. These two properties are well suited for anomaly detection in a maritime scenario since it is the evidence behind the reasoning process that is of interest in order to reveal the most probable configurations of the variables that can explain the explanandum, together with receiving comprehensive explanations in order to evaluate the situation at hand.

The two methods also present their explanations at a macro level which, in a maritime scenario, gives a good overview over the explanations generated, since an operator monitoring an anomaly detection system should be presented with a good overview over the observed situation. Based on their content, the only difference between the two methods is the causality characteristic. As can be concluded from the experiments, the causal interpretation of the Bayesian networks generate explanations that are more intuitive for humans to understand, since the cause and effect relationships between variables in the network have a more intuitive interpretation (Lacave and Diez, 2002).

6.4.2 Communication

How the two methods present their conclusions does not differ; they both present the variables in a tree structure. Though the numbers present in the generated explanation trees have different interpretations depending on the choice of explanation method (see subsection 6.1. for more information about the different interpretations of the numbers presented in the explanation trees). The tree structure, compared to the text based structure, is in this case preferable since there might be many variables of interest in the AIS data, which would make a text based explanation long and uninformative at a quick glance. Though, to present the explanations generated in a tree structure might be misleading in some cases. For example, assume that the explanation algorithm chooses the speed variable to be the best explanation for the state of the explanandum. The correct interpretation of the explanation is thus that the vessel is speeding. If this is not the case, the operator should look for the second best explanation presented in the explanation tree. Though this interpretation of the explanation tree hinders the possibility that several variables together might constitute an anomaly.

The user must be somewhat active in the reasoning process in order to pose a question to the system, for instance: what explanation can be given to the fact that the variable anomaly has the state yes, given that a vessel has been seen speeding? Though, after having posed a question like this, the algorithms calculate, according to their splitting criterion, which variables should be present in the explanation tree. The probabilities of the different variables are later on presented numerically in a network structure, where the variable with the highest number is the best explanation. To present the probabilities in numerical form might be deceiving though. It might be the case that there is not a large difference between some variables in the explanation but because of one variable's slightly larger probability is chosen as the best explanation.

6.4.3 Adaption

In order for a user to understand the explanations generated by the two algorithms, no knowledge about the specifics of the algorithms is needed. Though, it is preferable if the user has some knowledge about anomalies in the maritime domain in order to identify strange relationships between variables, for example, that a variable X cannot be the cause of Y in a causal case.

The levels of detail of the explanations generated together with both the ET method and the CET method can be adjusted by changing the stopping criterion for the two algorithms. Together with the ET algorithm, the stopping criterion affects the number of levels that are presented in the resulting explanation tree. Together with the CET algorithm, the stopping criterion determines which variables should be present in the generated explanation. No further explanations in, for example textual form, are generated as the algorithms are currently implemented, which could be helpful for an inexperienced user to fully understand the explanations generated. Thus the explanations generated by the two methods are to be considered static.

6.5 Summary of results

This subsection gives a summary of the results obtained from the analysis of the literature analysis and the experimentation. The explanation methods are evaluated according to their performance when it comes to generating plausible explanations for the tests performed during the experimentation and also according to the explanation characteristics presented in subsection 2.6.1.

6.5.1 Results from experiments and literature analysis

In the tests made together with the Explanation Tree method, the algorithm did not succeed in identifying the hidden causes of the anomalies hidden in the AIS data as the best explanations in the given cases. Due to its way of choosing the variables present in the explanation tree, the explanations generated are static, that is, the algorithm was not capable of generating explanations that reflected the changes made in the AIS data in an appropriate way. In order to test if the poor explanations generated were due to the structure of the underlying Bayesian networks, tests were made with two types of BNs: one where the anomaly node was automatically added to the network in Genie and one where the anomaly node was manually added to the network. Due to this change, the algorithm produces different explanations for the anomalies hidden (as depicted in figures 16 and 17), though none of the generated explanations manage to reflect the anomalies hidden in the AIS data. In the presented cases, the explanations generated based on the Bayesian networks with the automatically added anomaly nodes are the best explanations generated together with this method, though none of the explanations manages to identify the causes of the anomalies. The generated explanations also reveals that, in the presented cases, the performance of the ET algorithm decreases as the number of features included in the hidden anomalies increases.

The explanations generated together with the Causal Explanation Tree method are of better quality than those generated by the ET algorithm, since the CET method managed to identify the speed variable as the first best explanation in the case with one and three anomalies hidden in the AIS data. In the case where the position constituted the hidden anomaly, the performance of the CET method decreased, choosing the position variables as the second and third best explanations. The CET method also generates explanations that are more intuitive and correct due to its cause and effect relationships between the variables as well as the fact that the algorithm makes a distinction between the observed variables and the explanandum. Despite this, the CET method also has its shortcomings. Critique against the CET method, as posed by Flores (2005), is that it assumes a causal Bayesian network, which might not always be possible to construct with certainty in a maritime scenario. In the experiments of this thesis, the causality of the Bayesian networks used has been imposed on the networks by manually adding an anomaly node to the networks. Depending on the selected stopping criterion, the performance of the CET method varies. With a high stopping criterion, the method performs quite well, even when it comes to explaining the presence of anomalies that involve more than one feature.

Though, with a low stopping criterion, tests have shown that the performance of the method decreases (see figures A2, A3 and A4 in the appendix of this thesis in order to analyze the results from the CET algorithm where stopping criterion 0 was used). This might be due to the fact that the different variables in the Bayesian network are strongly correlated.

In general, the two explanation algorithms are very dependent on the underlying AIS data and Bayesian networks. Different results should be obtained, for example, if several anomalies were added to the different AIS files. 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. This might be the reason for why the different variables in the Bayesian networks are strongly connected. Having a strongly connected BN has in the tests made together with the ET algorithm been proven to have a negative influence on the results, since it is an indication to the algorithm that is must not stop the calculations before all the variables present in the BN have been covered in the generated explanation. This result is most apparent together with the ET algorithm since the choice of stopping criterion for the ET algorithm only affects how many levels of the generated explanation tree should be present, and not how many variables should be present as together with the CET algorithm.

It should be mentioned that labeled data has been used during the experimental phase (by manually adding an anomaly node to the Bayesian network built from the data). Thus, the Bayesian network is used for classification purposes (where observations are classified in two classes: anomalous or not). Previous work regarding the use of Bayesian networks for maritime anomaly detection (see Johansson and Falkman (2007)) presents a slightly different approach, since the BN models only the normal behavior of the maritime data.

According to the different explanation characteristics presented in subsection 2.6.1, the ET method and the CET method are very similar. They both present comprehensive, macro-level explanations of the evidence generated in a tree based form. The operators must be equally active in the generation of the explanation trees as well as they do not need to have any knowledge about the algorithm behind the calculations. The main difference between the two methods lies in their causal/non-causal properties. According to Lacave and Díez (2002), as well as what can be analyzed from the explanation trees generated from the two methods, the trees generated with the CET method have a more natural interpretation than the non-causal trees generated by the ET method.

7 Conclusions

This section briefly summarizes the results from the work performed in this thesis. Potential ideas for future work related to the work performed in this thesis are also presented.

7.1 Investigation of explanation properties and methods

The literature analysis of this thesis has explored which explanation properties researchers within the domain believe that an explanation describing the output of a Bayesian network should have. These properties are divided into three categories: i) the content of the explanation, ii) how the explanation is communicated to the operator and iii) if and how the generated explanation can be adapted to different types of operators (novice operators, expert operators etc). When working together with the rich content of the AIS data, the tree representation as communication media has been identified as an appropriate way of representing the generated explanation, since the tree structure provides a compact way of presenting the variables contained in the explanation.

The literature analysis has also identified and analyzed existing methods for explaining the outcome of a Bayesian network as well as adhering concepts and relevant terminology. These methods are divided into three different categories in this thesis: i) abductive inference methods, ii) the Explanation Tree (ET) method and iii) the Causal Explanation Tree (CET) method. The ET method and the CET method are the most recently developed explanation methods for Bayesian networks and their founders argue that these methods have overcome many of the shortcomings of the abductive inference methods, thus the two methods were chosen for the experiment of this thesis.

7.2 Experimentation with explanation methods

The experimentation of this thesis has revealed that it is possible to use the two tested explanation methods in order to generate explanations for the presence of anomalies in the maritime scenarios used. It has been showed that the Causal Explanation Tree method performs better than the Explanation Tree method when it comes to identifying the underlying anomalies hidden in the AIS data. This might be due to the differences between the two methods when it comes to selecting the variables to be present in the explanation trees generated, as well as if the relations between the variables should be considered causal or not. Tests have revealed that the performance of the ET method decreases as the number of variables included in the hidden anomalies increases. The performance of the CET method does not decrease if the number of variables included in the anomaly increases, if the stopping criterion has been set high. It has also been discovered that the explanation algorithms are sensitive to changes in the underlying Bayesian network. Thus, tests have been made with the two explanation methods with the same BN in order to equally compare the performance of the methods. It has also been discussed that a change of the underlying AIS data and BNs might improve the explanations generated by the two explanation tree methods tested in this thesis.

7.3 Future work

In order to use the explanation methods presented in this thesis in a real maritime anomaly detection setting, it should be investigated how the generated explanations could be presented and visualized to the operators monitoring the system. The use of, for example, colors and sounds are two means of accentuating the findings of the explanation methods which might help the operators to understand what has happened in the observed environment. As a result of this, the operators might improve their capability of making correct and high-quality decisions. Thus, with the help of mature visualization techniques, the situation awareness of the operators might further be improved.

It could also be investigated what can be done in order to improve the performance of the two explanation methods presented in the experiment section of this thesis. It could here be analyzed what can be done with the AIS data or the Bayesian network in order for the two methods to perform better when it comes to, for example, detecting anomalies that involve multiple variables.

Acknowledgements

First of all, I would like to thank my supervisor Maria Riveiro for her excellent supervision and guidance through the work with this thesis.

I would also like to thank Fredrik Johansson and Henrik Boström for their useful comments and ideas in the field of Bayesian networks and report structuring.

I am also very thankful for the help I have received from Jean-Philippe Pellet who made it possible for me to test and fully understand the different explanation methods presented in this thesis.

Last but not least, I would like to thank the developers of Genie (the Decision Systems Laboratory), the developers of Graphviz (the AT&T Research) and Saab Microwave Systems for providing the AIS data used in the experiments of this thesis.

Thank you!

Tove Helldin

May 2009

References

Bomberger, N. A., Rhodes, B. J., Seibert, M., and Waxman, A. M., 2006. Associative Learning of Vessel Motion Patterns for Maritime Situation Awareness. In *Proceedings of the 9th International Conference on Information Fusion*, Florence, July, pp. 1-8.

Bomberger, N.A., Waxman, A.M., Rhodes, B.J., and Sheldon, N.A., 2007. A new approach to higher-level information fusion using associative learning in semantic networks of spiking neurons. *Information fusion*, 8(3), July, pp. 227-251.

Campos, L.M., Gámez, J.A., and Moral, S., 2001. Simplifying Explanations in Bayesian Belief Networks. In *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 9(4), pp. 461-489.

Chajewska, U., and Halpern J.Y., 1997. Defining explanation in probabilistic systems. In *Proceedings of the 13th Annual Conference on Uncertainty in Artificial Intelligence*, Providence, USA, August, pp.62-71.

Dasarathy, B. V., 2001. Information fusion – what, where, why, when, and how? *Information Fusion*, 2(2), 75-76.

Dash, D., and Druzdzel, M.J., 1999. A Hybrid Anytime Algorithm for the Construction of Causal Models from Sparse Data. In *Proceedings of the 15th Annual Conference on Uncertainty in Artificial Intelligence*, San Francisco, USA, pp.142-149.

Druzdzel, M.J., and Henrion, M., 1990. Using Scenarios to Explain Probabilistic Inference. In *Working notes of the AAAI-90 Workshop on Explanation*, Boston, USA, July, pp.133-141.

Druzdzel, M.J., and van der Gaag, L.C., 2000. Building Probabilistic Networks: Where Do the Numbers Come From? In *IEEE Transactions on Knowledge and Data Engineering*, 12(4), pp. 481-486.

Elmenreich, W., 2002. An Introduction to sensor fusion. Research Report 47, Institut fur Technische Informatik Vienna University of Technology, Austria.

Endsley, M.R., 1995. Toward a theory of situation awareness in dynamic systems. *Human Factors Journal*, 37(1), pp. 32-64.

Eriksen, T., Høye, G., Narheim, B and Meland, J. B., 2006. Maritime traffic monitoring using a space-based AIS receiver. *Acta Astronautica*, 58(10), May, pp. 537-549.

Flores Gallego, M.J. 2005. Bayesian networks inference: Advanced algorithms for triangulation and partial abduction. PhD. thesis. Departamento de Sistemas Informáticos, University of Castilla - La Mancha.

Haddawy, P., Jacobson, J., and Kahn, C.E Jr., 1997. BANTER: a Bayesian network tutoring shell. *Artificial Intelligence in Medicine* 10(2), pp. 177-200.

Hall, D. L., and Llinas, J., 1997. An Introduction to Multisensor Data Fusion. In *Proceedings of the IEEE*, 85(1), January, pp. 6-23.

Halpern, J.Y., and Pearl, J., 2001. Causes and Explanations: A Structural-Model Approach – Part II: Explanations. In *Proceedings of the 17th International Joint Conference on Artificial Intelligence*, San Fransisco, USA, June.

Høye, K. G., Eriksen, T., Meland, J. B., and Narheim, T. B., 2008. Space-based AIS for global maritime traffic monitoring. *Acta Astronautica*, 62(2-3), pp. 240-245.

Jensen, F.V. 2000. Bayesian Graphical Models. *Encyclopedia of Environmetrics*, Wiley, Sussex, UK.

Jensen, F.V., Aldenryd, S.H., and Jensen, K.B., 1995. *Sensitivity analysis in Bayesian networks*. Symbolic and Quantitative Approaches to Reasoning and Uncertainty, 946, pp. 243-250.

Jensen, F.V., and Nielsen, T.D. 2007. *Bayesian Networks and Decision Graphs*. 2nd edition, Springer, New York, USA.

Johansson, F., and 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*, Melbourne, Australia, December 3-6.

Lacave, C., and Díez, F., 2002. A Review of Explanation Methods for Bayesian Networks. *Knowledge Engineering Review*, 17(2), pp.107-127.

Lambert, D.A., 2001. Situations for Situation Awareness. In *Proceedings of the 4th International Conference on Information Fusion*, Montréal, Québec, Canada, August.

Merriam-Webster Online Dictionary, 2009 [online]. Available from: http://www.merriam-webster.com/dictionary/explain [Accessed 27 March 2009].

Nielsen, U.H., Pellet, J-P., and Elisseeff, A., 2008. Explanation Trees for Causal Bayesian Networks. In *Proceedings of the 24th Annual Conference on Uncertainty in Artificial Intelligence*, Helsinki, Finland, July, pp.427-434.

Patcha, A., and Park, J., 2007. An overview of anomaly detection techniques: Existing solutions and latest technological trends. *Computer Networks*, 51(12), August, pp. 3448-3470.

Portnoy, L., Eskin, E., and Stolfo, S., 2001. Intrusion detection with Unlabeled Data Using Clustering. *ACM Workshop on Data Mining Applied to Security* (DMSA-2001), USA, November.

Riveiro, M., 2007. Evaluation of Uncertainty Visualization Techniques for Information Fusion. In *Proceedings of the 10th International Conference on Information Fusion*, Québec, Canada, July, pp. 1-8.

Riveiro, M., Falkman, G., and Ziemke, T., 2008. Improving maritime anomaly detection and situation awareness through interactive visualization. In *Proceedings of the 11th IEEE International Conference on Information Fusion*, Cologne, Germany, June-July, pp. 47-54.

Roy, J., 2008. Anomaly detection in the maritime domain. *Optics and Photonics in Global Homeland Security IV*, 6945, pp. 69450W1-14.

Russel, S., and Norvig, P. 2003. Artificial Intelligence – A Modern Approach. 2nd edition, Pearson Education Inc., New Jersey, USA.

Wallenius, K., 2004. Support for Situation Awareness in Command and Control. In *Proceeding of the 7th International Conference on Information Fusion*, Stockholm, Sweden, June-July, pp.1117-1124.

Yap, G.E., Tan, A.H., and Pang, H.H., 2008. Explaining inferences in Bayesian Networks. *Applied Intelligence*, 29(3), pp. 263-278.

Appendix

The appendix presents the pseudo code for the Explanation Tree algorithm and the Causal Explanation Tree algorithm. Boldface capitals denote sets of random variables or nodes in a graph. Italicized capitals are random variables or nodes and elements of the set of all variables. Vectors are denoted boldface lowercase, scalars in italics (Nielsen et al., 2008). The appendix also presents an example of a discretization made together with the variables included in the AIS file where one anomaly has been manually hidden. Explanation trees generated together with the CET algorithm where stopping criterion 0 was used are also presented.

The Explanation Tree algorithm

```
function $T$ = EXPLANATION_TREE($\mathbf{H}, \mathbf{e}, \mathbf{p}; \alpha, \beta$)
  Input:    H: set of explanatory variables
            E=e: explanandum
            P=p: path of variable assignments
            α, β: stopping criteria
  Output:   T: an explanation tree
$X^* \leftarrow \arg\max_{X \in \boldsymbol{H}} \sum_{Y \in \boldsymbol{H}} \mathrm{INF}(X; Y | \mathbf{e}, \mathbf{p})$
if $\max_{Y \in \mathbf{H} \setminus X*} \mathrm{INF}(X; Y | \mathbf{e}, \mathbf{p}) < \alpha$ or $\mathrm{p}(\mathbf{p}|\mathbf{e}) < \beta$ then
return Ø
end if
$T \leftarrow$ new tree with root $X^*$
for each $x \in$ domain $(X^*)$ do
        $T' \leftarrow$ EXPLANATION_TREE($\mathbf{H} \setminus X^*, \mathbf{e}, \mathbf{p} \cup \{x\}$)
        add a branch to x to T with subtree T' and
        assign it the label p(p, x|e)
end for
return $T$
```

(The pseudo code is adapted after Nielsen et al., 2008)

The Causal Explanation Tree algorithm

```
function $T$ = CAUSAL_EXPLANATION_TREE($\mathbf{H}, \mathbf{o}, e, \hat{\boldsymbol{p}}; \alpha$)
  Input:    H: set of explanatory variables
            O=o: observation set
            E=e: explanandum
            p̂: path of interventions
            α: stopping criterion
Output:     T: a causal explanation tree
$X^* \leftarrow \arg\max_{X \in \boldsymbol{H}} I(X \rightarrow e|\mathbf{o}, \hat{\mathbf{p}})$
if $I(X^* \rightarrow e|\mathbf{o}, \hat{\mathbf{p}}) < \alpha$ then return $\emptyset$
$T \leftarrow$ new tree with root $X^*$
for each $x \in$ domain $(X^*)$ do
    $T' \leftarrow \text{CAUSAL\_EXPLANATION\_TREE}(\mathbf{H}\backslash\{X^*\}, \mathbf{o}, e, \hat{\mathbf{p}} \cup \{\hat{x}\})$
    add a branch x to T with subtree T' and
    assign it the contribution $\log(p(e|\mathbf{o}, \hat{\mathbf{p}}, \hat{x})/p(e|\mathbf{o}))$
end for
return $T$
```

(The pseudo code is adapted after Nielsen et al., 2008).

Discretization categories of AIS variables

The table below presents an example of a discretization made together with the AIS file where one anomaly has been manually hidden (high speed of one vessel). This table has been added to the appendix of this thesis in order to depict what the different discretization categories of the different variables represent. The table presents which discretization method, available in the Genie application, has been used on the specific variable as well as the numeric intervals the different discretization categories encompass (for example, the first discretization category of the speed variable (SogKnots) encompasses speeds between 0-12.25 knots).

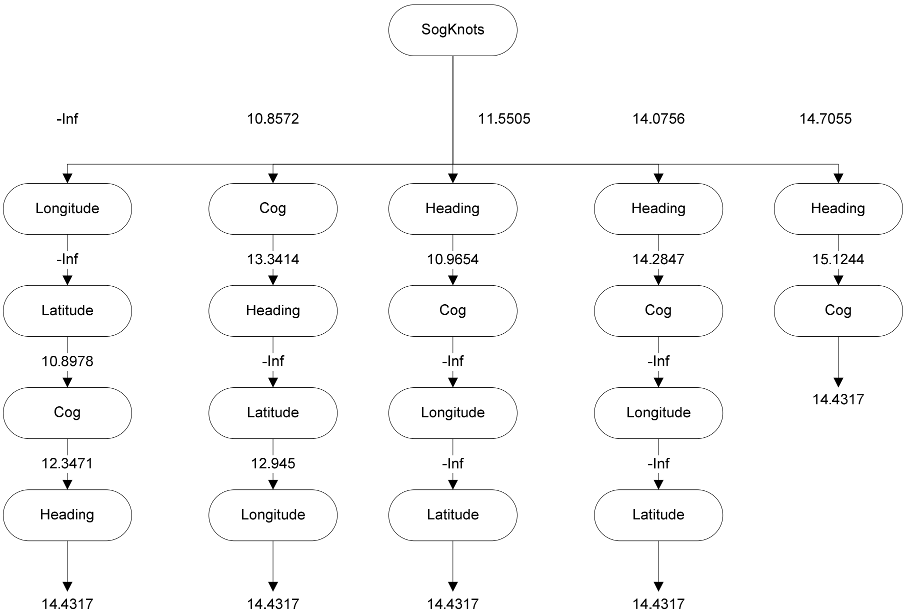
Variable	Discretization method	Discretization category	Numbers (from X to Y)	
Longitude	Uniform widths	1	0	56.1442
		2	56.1442	56.9107
		3	56.9107	57.6773
		4	57.6773	58.4442
		5	58.4442	-
Latitude	Uniform widths	1	0	10.278
		2	10.278	11.0892
		3	11.0892	11.9002
		4	11.9002	12.7112
		5	12.7112	-
SogKnots	Hierarchical	1	0	12.25
		2	12.25	16.35
		3	16.35	63.5
		4	63.5	151.15

		5	151.15	-
Cog	Uniform widths	1	0	720.5
		2	720.5	1440.5
		3	1440.5	2160.5
		4	2160.5	2880.5
		5	2880.5	-
Heading	Uniform widths	1	0	102.5
		2	102.5	204.5
		3	204.5	306.5
		4	306.5	435
		5	435	-

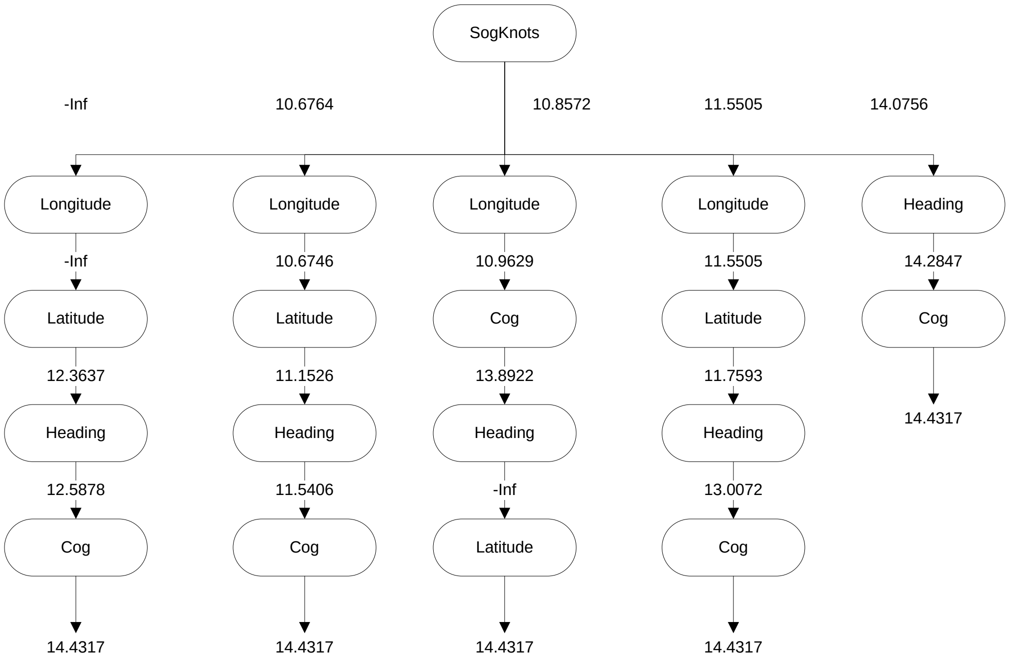
A1 – Discretization categories of variables where one anomaly has been hidden

CET explanations – Stopping criterion 0

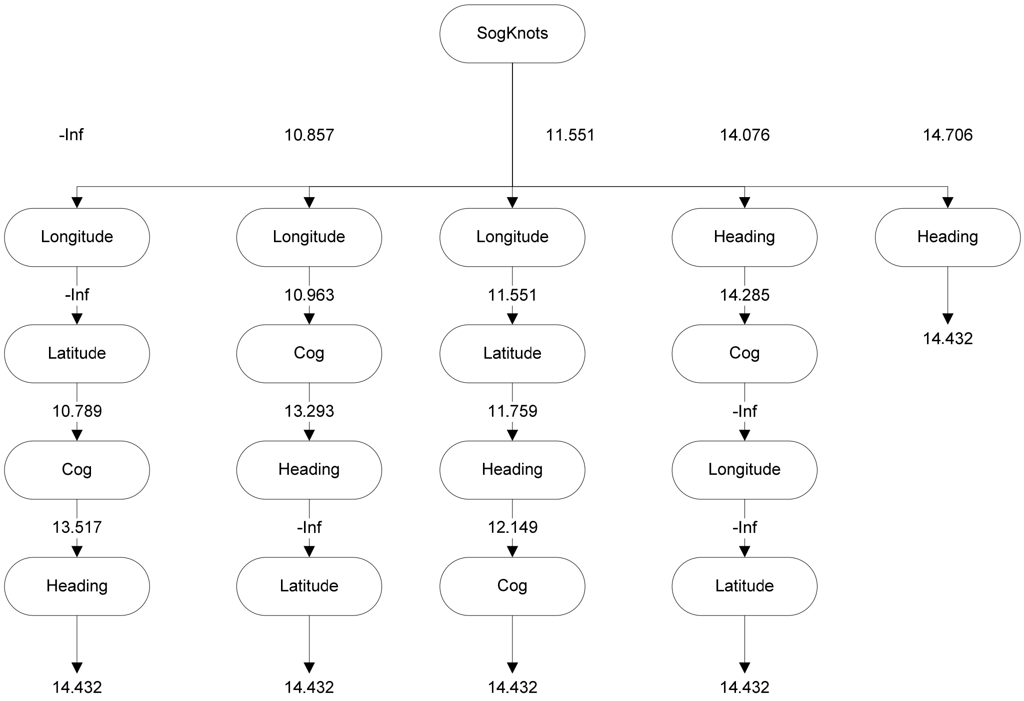
Three explanation trees generated together with the CET algorithm with stopping criterion 0 are presented below. These are included in the appendix of this thesis since the choice of stopping criterion strongly affects the performance of the CET method in the tested cases (presented in subsection 6.3). Together with the CET method, the stopping criterion affects how many variables should be present in the generated explanation tree. As can be read from the explanation trees presented in this appendix, the CET algorithm with stopping criterion 0 includes all the variables present in the underlying Bayesian networks. Thus, these explanations are more complex and more difficult to interpret than those with a higher stopping criterion (presented in subsection 6.3).



A2 – A CET explanation for one vessel speeding (an anomaly based on one variable)



A3- A CET explanation for abnormal position of one vessel (an anomaly based on two variables)



A4 – A CET explanation for abnormal speed and position of one vessel (an anomaly based on three variables)