A diagnostic system of an intelligent component based on Bayesian accurate inference networks

S I Kondratyev¹, A I Epikhin², S O Malakhov³

¹Admiral Ushakov Maritime State University, Department of Navigation, Lenin Ave., 93, Novorossiysk, 353924, Russia
²Admiral Ushakov Maritime State University, Department of Ship Mechanical Installations Exploitation, Lenin Ave., 93, Novorossiysk, 353924, Russian Federation
³Admiral Ushakov Maritime State University, Department of Radioelectronics and Information Technologies, Lenin Ave., 93, Novorossiysk, 353924, Russian Federation

E-mail: bsmbeton@mail.ru

Abstract. The article discusses the features of building an intelligent diagnostic system for technical systems based on Bayesian networks, which are a new method of probabilistic-statistical modeling based on a combination of graph theory, probability theory and methods of applied statistics. An example of a Bayesian network structure for diagnosing and predicting an accident in unmanned vessels is presented. Special attention is paid to the stages of building a Bayesian network for diagnosing technical systems using statistical data. Particular emphasis is placed on solving the problem of forming the Bayesian network output under conditions of decreasing uncertainty when new information arrives within the Bayesian network. The features of constructing an accurate probabilistic inference using the clustering algorithm are also considered. A number of experiments were carried out to compare the computational complexity of the clustering algorithm, which was used in the process of constructing the exact probabilistic inference in the Bayesian network, and the stochastic sampling algorithm to obtain an approximation. To simulate the structure of an intelligent system for technical diagnostics, it is proposed to divide the Bayesian network into subnets, the graphs of which will have a relatively small number of vertices. As an example, the following classes of diagnostic parameters of a technical system are proposed, for each of which it is further recommended to build a Bayesian network: probability of system uptime, system failure rate, average system (element) operation time to failure, average service life, gamma-percentage resource, average shelf life, etc.

1. Introduction

The volume of information in the world is growing rapidly every year; there is a saturation and overload of data. In addition, technical systems are rapidly becoming more complex, especially in the context of digitalization and the development of Industry 4.0. Obviously, in such conditions, the work of an analyst is transformed, who must study significant amounts of information in the process of solving the task of diagnosing the state, serviceability and reliability of technical systems and their components. In addition, in the context of the complexity and multi-element nature of modern equipment, “hidden” data is of great importance, requiring gigabytes and terabytes of information for storage, which a person is not able to explore on his own.

Obviously, to reveal "hidden" knowledge, it is necessary to use special methods of automatic
analysis, which allow extracting information from a huge amount of data. At the same time, attention should be paid to the fact that the level of automation of data processing, the use of objective mathematical models of technical systems at almost all levels of the management hierarchy is characterized, at the present time, by low indicators.

Bayesian networks (BN), which represent a new method of probabilistic-statistical modeling based on a combination of graph theory, probability theory and methods of applied statistics, are one of the promising tools of modern intellectualized information systems for the accumulation and processing of data, proper computer support for decision-making, diagnostics of complex technical systems.

The analysis of the methods of data mining used today showed that BN, in comparison with the well-known models of “black boxes”, make it possible to formulate clearer conclusions, and also imply a logical interpretation and modification of the structure of relations between the variables of the problem. In addition, BN also allow explicitly taking into account the existing a priori experience of experts [1].

The degree of success of the BN application in the process of modeling and the formation of statistical inference depends on the ability to correctly formulate the problem statement, select the process variables that adequately characterize its dynamics or statics, collect statistical data and use them to train the network, and also correctly form the result—the conclusion with the help of the constructed network, which in general is an important scientific and practical problem to the solution of which this article is devoted.

The development and implementation of intelligent diagnostic methods leads to increased attention to them from scientists and practitioners. For instance, such authors as Haghir Chehreghani Mostafa, Dridi Amna, Ciza Thomas, Maksimov A.G., Zavalishin A.D., Abramov M.V., Tulupyev A.L. are engaged in this problem.

Specificities and prospects of using BN in systems for classifying data of various nature, forecasting, automatic signal recognition are considered in the works of Patrocinio, A. C.; Schiabel, H.; Romero, R. A. F., Deng, Yu-Jing, Burakova D.P., Kozhomberdieva G.I., Dorozhko I.V., Osipova N.A.

However, despite the available publications, it should be noted that data mining is a multidisciplinary area of knowledge that generates a significant number of uncertainties inherent in individual variables and groups of variables, which in turn initiates a number of questions that require additional research and analysis. In particular, the problems of choosing the most optimal methods for calculating the probability of occurrence of events in the BN remain open. Also, special attention should be paid to the problem of constructing a quantitative criterion for assessing the degree of uncertainty, which can be described using BN.

Thus, taking into account the above, the purpose of the article is to consider the specificity of using data mining tools based on BN in technical diagnostics systems.

2. Materials and methods

2.1. Materials

Thomas Bayes is one of the first scientists to investigate the likelihood of certain events in the future using information from past tests. As a consequence, he developed a theorem that links posteriori and a priori probabilities of causes after observing the consequences.

Let $H_1, H_2, ..., H_n$ be pairwise incompatible events and their sum coincides with the entire selected space of events. Then, for any random event $X$, which can occur only if one of the events $H_1, H_2, ..., H_n$ appears, the following equality (1) is fulfilled:

$$p(H_k|X) = \frac{p(X|H_k)p(H_k)}{\sum_{i=1}^{n} p(X|H_i)p(H_i)}, \quad k = 1, n$$

(1)

where $H_k$ is any hypothesis out of $n$ possible ones. The probabilities $p(X|H_k)$ are set by experts a priori or calculated from training data. That is, they allow answering the question, “What will be the probability of a certain measurement, if it is known which hypothesis was realized?” The probabilities $p(X|H_k)$ are very useful in practice, because it is generally easier to find the probability of a sequence of cause-effect events than vice versa. The $p(H_k)$ values are prior probabilities, they define the initial probabilities for all hypotheses.
The indisputable advantage of BN for diagnosing the state of technical systems is that a priori probabilities can be refined (updated) in accordance with the actual specificities and results of the implementation of the process under study, the operation of a particular element and the system as a whole. This makes it possible to more accurately determine the probabilities of certain events when additional information is received.

At the next stage of the study, we will consider in more detail the algorithm for constructing a Bayesian network for diagnosing technical systems using statistical data.

### 2.2. Methods

**Input data.** There is a training sample (the superscript is the variable number, the subscript is the observation number) (2):

\[ D = \{d_1, ..., d_n\}, d_i = \{x^{(1)}_1, x^{(2)}_1, ..., x^{(N)}_1\} \]  

(2)

where \( n \) is the number of observations; \( N \) is the number of vertices (variables).

**First stage.** For all pairs of vertices, the values of mutual information are calculated: \( \text{Set}_{MI} = \{MI(x^i, x^j); \forall i,j\} \). After that, the elements of the \( \text{Set}_{MI} \) set are ordered in descending order (3):

\[ \text{Set}_{MI} = \{MI(x^{m_1}, x^{m_2}), MI(x^{m_3}, x^{m_4}), MI(x^{m_5}, x^{m_6}), ... \} \]  

(3)

**Second stage, step 1.** From the set of mutual information values \( \text{Set}_{MI} \), the first two maximum values are selected \( MI(x^{m_1}, x^{m_2}) \) and \( MI(x^{m_3}, x^{m_4}) \). Using the received value \( MI(x^{m_1}, x^{m_2}) \) and \( MI(x^{m_3}, x^{m_4}) \), a multitude of models \( G \) are built that have the following form (4):

\[
\begin{align*}
&\{(m_1 \rightarrow m_2; m_3 \rightarrow m_4), (m_1 \rightarrow m_2; m_3 \leftarrow m_4), (m_1 \leftarrow m_2; m_3 \rightarrow m_4), (m_1 \leftarrow m_2; m_3 \leftarrow m_4), \\
&(m_4 \leftarrow m_2; m_3 \text{ does not depend on } m_4), (m_4 \leftarrow m_2; m_3 \text{ does not depend on } m_4), \\
&(m_4 \leftarrow m_2; m_3 \text{ does not depend on } m_4)\}
\end{align*}
\]

(4)

Notation \( m_i \rightarrow m_j \) means that vertex \( x^{m_i} \) is a predecessor of vertex \( x^{m_j} \).

Step 2. Searches across multiple \( G \) models. Parameter \( g^* \) stores the optimal network structure. The optimal structure is the one with the smallest value of the function: \( L(g, x^n) \) is the description of the minimum length of the model structure for a given sequence (5) of \( n \) observations \( x^n = d_1, d_2, ..., d_n \).

\[ g^* \leftarrow g_0(\in G); \]

\[ \text{for } \forall g \in G \rightarrow \{g_0\}; \text{if } L(g, x^n) < L(g^*, x^n) \text{ then } g^* \leftarrow g; \]

(5)

на выходе \( g^* \) — искомое решение.

Step 3. After the optimal structure \( g^* \) of \( G \) has been found, using the set of mutual information values \( \text{Set}_{MI} \), the maximum value is selected: \( MI(x^{l_{\text{next}}}, x^{l_{\text{next}}}) \). From the obtained value \( MI(x^{l_{\text{next}}}, x^{l_{\text{next}}}) \) and the structure \( g^* \), a set of models \( G \) of the form \{\{(g^*; i_{\text{next}} \rightarrow j_{\text{next}}), (g^*; i_{\text{next}} \leftarrow j_{\text{next}}), (g^*; i_{\text{next}} \text{ does not depend on } j_{\text{next}})\}\}.

**Search end condition.** The heuristic search continues until a certain number of elements of the set or all \( \frac{N(N-1)}{2} \)-elements of the \( \text{Set}_{MI} \) set are analyzed. As practice shows, in most cases it makes no sense to analyze more than half (that is, \( \frac{N(N-1)}{4} \)) elements of the \( \text{Set}_{MI} \) set.

**Output:** optimal structure (structures) \( g^* \).

### 3. Results

Fig. 1 shows an example of a BN for diagnosing and predicting an accident in unmanned vessels.

The vertices of \( c_1, c_2, ..., c_m \) series of the net (navigation, seaworthiness, mechanical part, miscellaneous) shown in Fig. 1 at the second level mean hypothetical reasons (with a priori given probabilities), which individually or simultaneously several at once, with a certain probability, affect the occurrence of an emergency with an unmanned vessel, which is located at the first BN level.
Figure 1. Modular structure of interconnections between three levels of safety of an unmanned merchant ship [2].

According to Fig. 1, for example, a fire, by influencing the propulsion complex, can lead the vessel to ground. In turn, the cause for the fire may be different systems of an unmanned vessel.

It is also necessary to pay attention to the fact that the vertices-causes of BN for an unmanned ship (level 2, Fig. 1) contain means of communication with the user of the system using special computers and control algorithms. Such means, after receiving signals about the occurrence of certain events or phenomena and determining the probable causes of their occurrence, provide the user with messages regarding possible consequences and recommended actions [4-8].
The above set of edges, which is all the paths between some two vertices, corresponds to the conditional dependency between these vertices. On such a network, you can use Bayesian inference to calculate the probabilities of the consequences of events. The selection criterion for use cases is as follows: if there is no completely coinciding precedent, then the probability distribution is calculated for those features that do not coincide with the features of the current case. The precedent is chosen for which this probability is the greatest. Metric algorithms refer to use case reasoning techniques. Here we can surely talk about “reasoning”, since to the question “why the object X was assigned to the class Y?” the algorithm can give an understandable answer, “because there are precedents—objects similar to it, belonging to the class Y”, and present a list of these precedents. Basically, there are two ways to train Bayesian networks using use cases: refining the network parameters if the network structure is known, and choosing from a variety of models, applying the introduced metric to the entire use case.

In the process of diagnosing technical systems using BN, special attention should be paid, as noted earlier, to the possibility of correcting hypotheses and key parameters depending on new information that is collected during observation. In this regard, we will consider the features of the analysis of the influence of new information on the variable for which the hypothesis is formulated.

Let a BN contain variable $X = \{x_1, ..., x_n\}$ with respect to which the hypothesis is formulated, and information variable $E_i = \{e_{i1}, ..., e_{ik}\}$. When a new value $E_i = e_{i1}$ appears, the information is distributed over the network, $x_{ij}$ leads to a change in the distribution of $X$ by $p(x_i | e_{ij})$. The entropy of the variable $X$ becomes equal to $h(X|e_{ij})$. The direction of this entropy change depends on the following factors: (1) the value of the prior probabilities $p(x_i)$; (2) the level of knowledge accumulated in the BN in terms of conditional probabilities (these conditional probabilities play a key role in shaping the conclusion based on the BN).

In this case, the task of determining the effectiveness of the mechanism for generating the output of the BN under conditions of decreasing uncertainty as a result of the appearance of new information is actualized. This can be done by separating the influence of prior probabilities from the mechanism for forming the conclusion. One of the simple approaches to solving this problem is that it is assumed that there is no prior information about the key variable for which the hypothesis is formulated. This makes it possible to establish the degree of influence of new information on the initial uncertainty [2–7].

If there is no information about the key variable $X = \{x_1, ..., x_n\}$, then according to the principle of maximum entropy $p(x_i) = \frac{1}{n}$, $\forall i$, the prior entropy is $h(X) = \log n$. When information $E_1 = e_{1j}$ arrives, the entropy decreases by the value (6):

$$\Delta h_{e_{1j}} = \log n - h(X|E_1 = e_{1j})$$

The posterior probabilities $p(x_i|e_{1j})$ are calculated based on the prior probabilities $p(x_i) = \frac{1}{n}$, $\forall i$. Since $\log n$ is the maximum possible entropy, the above expression is always suitable. In this regard, it can be called a stage of uncertainty reduction due to information $E = e_j$. If $\{E_1, E_2, ..., E_m\}$ is new information, which is represented by several values $[e_{i1}, e_{2j}, ..., e_{mk}]$ obtained from different sources, then the degree of uncertainty reduction can be defined as (7):

$$\Delta h = \log n - h(X|E_1 = e_{1l}, E_2 = e_{2j}, ..., E_m = e_{mk})$$

Thus, the analysis of the problem of estimating the reduction of uncertainty when new information arrives within the Bayesian network leads to the following three cases:

1. Let $X$ be a root variable, and a hypothesis is formulated in relation to it. We assign this and other root variables a distribution that provides the same probability of occurrence of all possible values. After the network reaches equilibrium under such initial conditions, we assign one or more information variables to the corresponding value. The posterior probabilities $p(x_i|e_{1j}, e_{2j}, ..., e_{mk})$ are calculated after the network reaches equilibrium again. The degree of uncertainty reduction due to the emergence of new information can be calculated based on the prior and posterior probabilities.

2. If the variable $X$, in relation to which the hypothesis is formulated, is not a root one, then we remove its parent nodes (variables) from the network and implement the procedure described in the previous paragraph.

3. A situation is possible when new information affects $X$ through one of its parent nodes, for
example, $Y$. In this case, the removal of parent nodes will lead to a gap in the path of distribution of new information in $X$. Since new information affects $Y$ directly and only indirectly affects $X$, then we will consider $Y$ as the primary variable for which the hypothesis is formulated, and, accordingly, consider its prior probabilities instead of $X$. In other words, if $Y$ is a root variable, then we consider all its states as equiprobable (as in the first case), and if $Y$ is not a root variable, then we act as in the second case. Since $Y$ is the parent of $X$, setting the prior probabilities for $Y$ will automatically set the prior probabilities for $X$.

To construct an accurate probabilistic inference in Bayesian networks based on the results of diagnostics of technical systems, it seems appropriate to use the clustering algorithm.

The algorithm operates on merged trees, each vertex of which contains a set of variables and a table of conditional probabilities, this allows using the idea of probabilistic inference messaging based on Pearl's idea.

The algorithm for constructing a combined tree is presented by the block diagram in Fig. 2.

![Bayesian network](image)

**Figure 2.** Algorithm for constructing a combined tree in a BN.

To implement the clustering algorithm in a BN, the Hugin architecture is usually used.

The clustering algorithm for constructing accurate probabilistic inference in the BN guarantees the accuracy of the calculations, neglecting the computational complexity. It is one of the most efficient in its class.

Using Python 3.9 and its modules: matplotlib, numpy and pymc, a number of experiments were carried out to compare the computational complexity of the clustering algorithm, which was used in the process of constructing an accurate probabilistic inference in the BN, and the stochastic sampling algorithm to obtain an approximation [3].

Having built the BN model, for each algorithm, 10 tests were carried out, with a different number of vertices of the BN graph to find the learning time of the system. Tables 1 and 2 and Fig. 3. show the results of the experiment.

**Table 1.** System training time for clustering algorithms

| Number of vertices \ Learning time in the experiment, sec. | 1    | 2    | 3    | 4    | 5    | 6    | 7    | 8    | 9    | 10   |
|------------------------------------------------------------|------|------|------|------|------|------|------|------|------|------|
| 0–20                                                       | 0.35 | 0.41 | 0.35 | 0.51 | 0.38 | 0.42 | 0.37 | 0.48 | 0.47 | 0.48 |
| 21–40                                                      | 1.15 | 1.13 | 1.23 | 1.33 | 1.12 | 1.23 | 1.14 | 1.43 | 1.15 | 1.24 |
| 41–60                                                      | 2.20 | 2.12 | 2.21 | 2.14 | 2.57 | 2.45 | 2.25 | 2.35 | 2.46 | 2.21 |
| 61–80                                                      | 3.40 | 3.54 | 3.30 | 3.24 | 3.12 | 3.21 | 3.33 | 3.21 | 3.97 | 3.78 |
| 81–100                                                     | 4.10 | 4.31 | 5.01 | 3.93 | 4.74 | 4.65 | 4.25 | 4.12 | 4.21 | 4.15 |
Table 2. System learning time for stochastic sampling algorithms

| Number of vertices \ Learning time in the experiment, sec. | 1   | 2   | 3   | 4   | 5   | 6   | 7   | 8   | 9   | 10  |
|--------------------------------------------------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| 0–20                                                   | 0.15| 0.22| 0.23| 0.15| 0.14| 0.18| 0.23| 0.34| 0.24| 0.21|
| 21–40                                                  | 0.35| 0.37| 0.33| 0.37| 0.34| 0.34| 0.33| 0.32| 0.33| 0.34|
| 41–60                                                  | 0.67| 0.61| 0.58| 0.62| 0.64| 0.63| 0.57| 0.58| 0.69| 0.61|
| 61–80                                                  | 1.31| 1.31| 1.33| 1.31| 1.37| 1.32| 1.23| 1.22| 1.27|     |
| 81–100                                                 | 1.78| 1.87| 1.95| 1.99| 1.78| 1.87| 1.88| 1.75| 1.86| 1.91|

Figure 3. Dependence of the number of graph vertices on the learning time of the system for stochastic sampling and clusterization algorithms

As can be seen from Fig. 3, the clusterization algorithm has a high computational complexity, as evidenced by the exponential growth of time during learning after an increase in the number of vertices of the BN graph. In turn, the complexity of the approximation algorithm has a linear dependence on the number of vertices in the graph. However, with the number of vertices up to 30, the training time for both algorithms is almost the same. Therefore, for modeling the structure of an intelligent system for technical diagnostics of BN, it was shown that it was expedient to partition the BN into subnets, the graphs of which had a relatively small number of vertices.

4. Discussion

To model the structure of an intelligent system of technical diagnostics, BN can be divided into subnets, the graphs of which will have a relatively small number of vertices. As an example, the following classes of diagnostic parameters of a technical system can be distinguished, for each of which a BN can be built in the future: the probability of a failure-free operation of the system, the frequency of system failures, the average time of the system (element) to failure, the average service life, the gamma-percentage resource, average shelf life, etc.
5. Conclusion

Thus, the study carried out indicates that BNs are one of the effective tools for studying processes of various natures, and have a number of advantages over other intelligent methods of analysis and forecasting.

Rational use of BNs in the process of diagnostics of technical systems, their fast and reliable operation primarily depend on the model and algorithm for constructing output in the network.

References

[1] Puerta J M, Aledo J A, Gámez J A and Laborda J D 2021 Efficient and accurate structural fusion of Bayesian networks Inf. Fusion 66 155–69
[2] Wrobel K, Krata P, Montewka J and Hinz T 2016 Towards the Development of a Risk Model for Unmanned Vessels Design and Operations TransNav, Int. J. Mar. Navig. Saf. Sea Transp. 10
[3] Maksimov A G, Zavalishin A D, Abramov M V, Tulupyev A L 2020 Derivative graphs in global structures of algebraic Bayesian networks Computer tools in education 2 59–65
[4] Wang Z, Gao X, Yang Y, Tan X and Chen D 2021 Learning Bayesian networks based on order graph with ancestral constraints Knowledge-Based Syst. 211
[5] Senchenko V, Lopatina V, Studenikin D and Butsanets A 2021 Technical Automation Tools for High-Precision Navigating of Sea and River Ships
[6] Chernyi S G, Zheleznjak A A, Ovcharenko I K, Goryachev I S and Daragan P A 2020 Models and Information Technologies of Clustering and Location of Group Expert Assessments in Terms of Uncertainty Proceedings of the 2020 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering, EIConRus 2020
[7] Ivanchenko A A, Turkin V A, Konev G A and Karakayev A B 2019 State and perspective directions of development of CAD in shipbuilding Marine intellectual technologies 1-2 41–45
[8] NOVAC V, RUSU E, SCURTU I C 2019 Opportunities and Risks Related to Offshore Activities in the Western Black Sea Journal of Environmental Protection and Ecology 4 1698–1707