National Centre for Scientific Research Democritus

Large-Scale Statistical Methods

A Bayesian Network Approach for Cardiovascular Risk Prediction: Insights and Applications

Mitsis Koutoukis Alexandros

Nteits Nikolaos

Tsitsimpassis Stefanos

Academic Year 2024-2025

Contents

Abstract

Cardiovascular diseases (CVD) remain the leading cause of mortality worldwide, significantly impacting global health and economic resources. This paper presents the development and application of a Bayesian network model for predicting cardiovascular risk. By integrating extensive health data from annual assessments of Spanish workers with expert knowledge, the model highlights the complex relationships between non-modifiable, modifiable risk factors, and medical conditions. The accompanying software tool offers healthcare providers actionable insights to assess and mitigate cardiovascular risk.

Introduction

Cardiovascular diseases (CVD) are responsible for millions of deaths annually, creating an urgent need for effective prevention strategies. Accurate risk prediction enables the identification of high-risk individuals, paving the way for targeted interventions and improved health outcomes. This study develops a Bayesian network to model the interdependencies among cardiovascular risk factors and to evaluate the impact of interventions on reducing disease risk.

The model integrates data from over 200,000 health assessments and expert input, resulting in a software tool that is accessible for research and clinical applications.

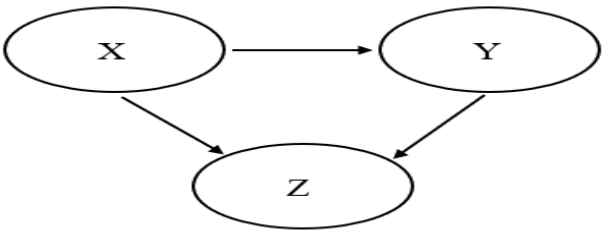
Prerequisite knowledge

Prerequisite technical knowledge

Bayesian networks (BNs) are probabilistic graphical models that represent a set of variables and their conditional dependencies using a directed acyclic graph (DAG). They are particularly useful for reasoning under uncertainty by integrating both expert knowledge and empirical data.

Basic concepts

Nodes represent variables and directed edges represent causal dependencies. Each node has an associated Conditional Probability Table that quantifies the probability of the node's states given its parent nodes' states. The BN can be represented as:



This indicates that X directly affects Y and both X and Y influence Z. An advantage of a BN is that it allows direct calculation of probabilities for any variable, given evidence. For example, using the conditional probability tables, one can compute the probability of Y given evidence X, P(Y|X), as well as the probability of Z given evidence X and Y. P(Z|X∩Y). Also, a BN enables causal modeling by explicitly representing which events act as causes and which serve as their effects.

Prerequisite domain knowledge

Cardiovascular diseases (CVDs) are the leading cause of death globally, responsible for approximately 17.9 million deaths annually, according to the World Health Organization (WHO). These diseases result from complex interactions between non-modifiable factors like age and genetics, and modifiable factors like smoking, physical inactivity, and diet.

Healthcare decision support and BNs

BNs provide a framework for integrating these diverse risk factors into a cohesive model that can predict the likelihood of cardiovascular events and support clinical decision-making by quantifying the impact of modifiable risk factors. Ordovás et al. leverage BNs to offer an interpretable, data-driven approach to cardiovascular risk prediction, addressing the challenge of integrating heterogeneous data sources in healthcare.

Objectives and merits of BNs in healthcare

While the Ordovás paper does not explicitly articulate broad goals for BNs in healthcare, their study implicitly demonstrates key benefits such as that BNs provide insights into the relationships between variables, aiding clinicians in understanding risk factors and by combining expert opinion with empirical data, they offer robust and adaptable models for clinical decision-making. Also, BNs can be extended to incorporate new variables or applied to different healthcare domains, making them versatile tools for predictive modeling. As Kyrimi et al. (2021) emphasize, the adoption of BNs in healthcare has been limited by inconsistent methodologies and a lack of standardization. Ordovás et al. address this by developing a structured and accessible model for cardiovascular risk prediction, demonstrating the utility of BNs in tackling pressing healthcare challenges.

Prior work

The study by Ordovás et al. builds on a solid foundation of research in cardiovascular epidemiology, Bayesian networks, and healthcare analytics. Below is an overview of the key prior work that provides context and inspiration for this study.

1. Cardiovascular Epidemiology and Risk Factors

Cardiovascular diseases (CVD) remain the leading cause of mortality in Europe, accounting for over 3.9 million deaths annually and incurring treatment costs exceeding €210 billion per year. Early epidemiological studies of CVD, such as the seminal Framingham Heart Study, which began in 1948 and was designed to identify common factors or characteristics contributing to CVD, have been pivotal in identifying cardiovascular risk factors (CVRFs). These include non-modifiable factors like age, sex, and genetics, as well as modifiable factors like physical activity (PA), diet, and smoking. The World Health Organization (WHO) has established guidelines to classify these CVRFs, which are frequently updated and validated.

Bayesian networks (BNs) offer a powerful framework for analyzing complex systems, as they can model dependencies and integrate data from diverse sources. This capability makes them particularly well-suited for medical applications. For example, Wang et al. (2020) developed a Bayesian network (BN) to predict the five-year survivability of patients with first and second primary cancers. Using a large dataset and addressing class imbalance with advanced techniques like SMOTE, their work demonstrated the effectiveness of BNs in handling complex dependencies and probabilistic reasoning in medical decision-making. Ordovás et al. cite this study to emphasize the adaptability of Bayesian networks for healthcare applications, particularly in modeling intricate relationships among variables, a cornerstone of their cardiovascular risk framework. Another study is that of Kyrimi et al. (2021). They conducted a broad review of Bayesian networks in healthcare, identifying challenges such as limited clinical adoption, inconsistent methodologies, and underutilization of BNs' full potential. The study called for standardized approaches and greater integration of BNs into clinical practice. Ordovás et al. reference this review to contextualize their own work within the broader landscape of healthcare analytics and to position their study as addressing key gaps, such as practical implementation and usability of BNs for cardiovascular risk prediction.

In the context of cardiovascular disease (CVD), the integration of cardiovascular risk factors (CVRFs) into predictive models represents a natural progression from earlier epidemiological research. Recent studies, as discussed below, have further explored the use of BNs to understand and predict CVD events, building on these foundational insights.

2. Applications of Bayesian Networks in Healthcare

Several pioneering studies have demonstrated the utility of BNs in cardiovascular research:

- Farooq et al. (2009) proposed an ontology-driven decision support system for chest pain assessment, incorporating adaptive questionnaires and semantic patient profiles to streamline clinical workflows. Their approach integrated structured expert knowledge with patient data for more effective diagnostics.

Ordovás et al. reference this work as a complementary example of leveraging structured knowledge frameworks in healthcare, paralleling their own integration of expert insights into a probabilistic BN for cardiovascular risk.

- Tylman et al. (2012) developed a real-time system for predicting acute cardiovascular events using BNs to process vital signs like electrocardiography and blood pressure.
- The study by Thornley et al. (2013) explores the use of directed acyclic graphs (DAGs), a subset of Bayesian networks, to investigate causal pathways for CVD. By applying DAGs to a cohort dataset, the authors identified key causal influences, such as age and smoking, and their indirect effects mediated through other variables.
- Roberts et al. (2015) developed a BN model for cardiovascular monitoring, integrating diverse data types such as lab results, vital signs, and clinician observations to estimate unobservable patient variables. Ordovás et al. reference this study for its demonstration of Bayesian networks' ability to synthesize qualitative and quantitative data, a core principle in their own approach to integrating cardiovascular risk factors. This paper reinforces the versatility of BNs in handling complex datasets and providing actionable insights in healthcare. utilized a BN to predict unobservable variables related to cardiovascular states.

These studies laid the groundwork for integrating probabilistic reasoning with clinical decision support systems.

3. Integration of Lifestyle and Additional Factors

The role of lifestyle factors, such as PA, diet, and socioeconomic conditions, in influencing CVD outcomes has also been a focus of prior research. For example, Fiuza-Luces et al. (2018) discuss how exercise improves vascular function, autonomic balance, and inflammatory profiles while also contributing to novel mechanisms like gut microbiota modulation and myocardial regeneration. Santos-Lozano et al. investigated the relationship between physical activity (PA) levels and cardiovascular risk factors (CVD), highlighting that even PA below World Health Organization (WHO) guidelines reduces CVD risk. Their study also explored sex-specific variations in how PA affects conditions like obesity and hypercholesterolemia. Ordovás et al. reference these papers to underscore the importance of lifestyle factors, particularly physical activity, in cardiovascular risk modeling. This connection reinforces the inclusion of modifiable risk factors in their Bayesian network framework to provide a comprehensive approach to CVD prediction.

Ordovás et al. extend these efforts by integrating additional factors such as depression and sleep duration into their BN framework, offering a more comprehensive perspective on CVD risk.

In conclusion, the work of Ordovás et al. synthesizes insights from decades of cardiovascular research and Bayesian network development. By integrating expert knowledge, observational data, and novel lifestyle factors, this study represents a significant step forward in predictive modeling for CVD. It offers a robust framework for risk assessment and decision support, paving the way for future advancements in personalized healthcare analytics.

Materials and Methods

Data Collection

The dataset includes annual health assessments conducted between 2012 and 2016. The original data comprised over one million records, reduced to 205,087 after cleaning and preprocessing. The data encompasses a comprehensive range of variables:

- Non-modifiable risk factors such as age, sex, socioeconomic status, and education level were included to understand static determinants of cardiovascular health.
- Modifiable risk factors like BMI, physical activity, sleep duration, smoking habits, anxiety, and depression provided insights into areas where interventions could have significant effects.
- Prevalence of medical conditions such as hypertension, hypercholesterolemia, and diabetes were included to assess disease risk and progression.

Data Preparation

The data underwent a meticulous cleaning process to ensure accuracy and relevance for the model:

- Records containing extreme outliers were removed by applying a three-standard-deviation rule for continuous variables, which eliminated 586 entries suspected of measurement or recording errors.
- Duplicated records and entries with incomplete information were excluded, accounting for 7,689 removals.
- To avoid duplication across time, only the latest assessment for each individual was retained, resulting in a final dataset of 205,087 records.

All variables were discretized for compatibility with Bayesian network modeling. Discretization simplifies the modeling process by transforming continuous data into categorical data, which is essential for Bayesian networks to handle probabilities efficiently. However, this process can also impact accuracy by reducing granularity, particularly for individuals whose values fall near category boundaries. Despite this, discretization enhances interpretability by aligning variables with clinically meaningful thresholds, such as WHO BMI classifications. For instance, BMI was categorized into underweight, normal weight, overweight, and obese using WHO standards, while age was grouped into ranges reflecting life stages.

Bayesian Network Construction

The Bayesian network was developed in two stages:

1. Structure Learning: To identify relationships between variables, the Greedy Thick Thinning (GTT) algorithm was applied. This method utilizes conditional

independence tests to uncover dependencies. It is well-suited for processing large datasets, as it systematically refines the network structure to align with observed data, ensuring that key dependencies are accurately represented in the model.

2. Expert Refinement: The initial structure was critically reviewed by domain experts who evaluated cause-effect relationships between variables. Over several iterations, they adjusted the structure by adding 15 new edges, reversing the direction of 7, and removing irrelevant connections to ensure the network accurately represented real-world phenomena.

The final network linked modifiable and non-modifiable risk factors to medical conditions, enabling a detailed analysis of their interactions.

Probability Estimation

The network's conditional probability tables (CPTs) were estimated using multinomial-Dirichlet models. To address uncertainty, uniform priors were used because they allow for an unbiased starting point when estimating probabilities in the absence of strong prior knowledge. This choice ensures that all possible outcomes are treated equally initially, while posterior distributions refine these probabilities based on the observed data, resulting in robust and data-driven predictions. This approach ensured that even with sparse data in certain nodes, the model remained robust.

Model Validation

To assess the model's predictive power, a 5-fold cross-validation method was used. The model was trained on 80% of the data and tested on the remaining 20%, iteratively for five cycles. The results demonstrated:

- A predictive accuracy of 91% for modifiable risk factors such as BMI and physical activity levels.
- A predictive accuracy of 93% for medical conditions like hypertension and diabetes.

Software Implementation

The Bayesian network's accompanying software tool is a core component of its practical application. Built on the GeNIe platform, this tool enables clinicians and researchers to:

- Input patient-specific data and generate probabilistic predictions for medical conditions and risk factors.
- Simulate potential interventions by adjusting modifiable factors, such as BMI and physical activity, to assess their impact on outcomes.

- Explore the intricate interdependencies among risk factors in real-time, facilitating actionable insights for decision-making at both individual and population levels.

The software's open-source availability on GitHub ensures accessibility for researchers to further customize and refine the tool according to specific needs and applications. This accessibility fosters collaborative advancements and broadens its utility across diverse healthcare environments. Future enhancements to the model will focus on expanding its applicability, improving its dynamic capabilities, and integrating real-time data sources. Future enhancements to the model will focus on expanding its applicability, improving its dynamic capabilities, and integrating real-time data sources.

Analysis of the Greedy Thick Thinning (GTT) Algorithm

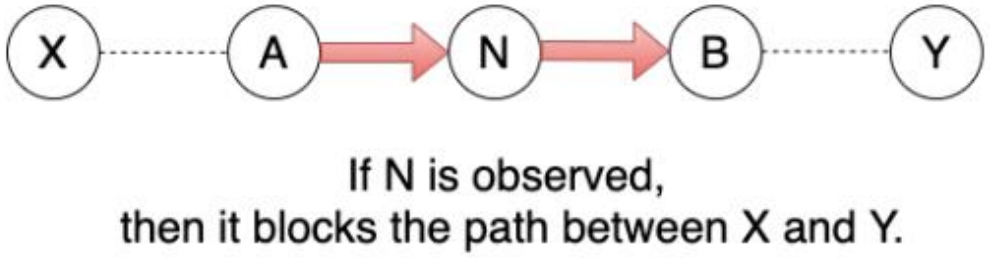
D-Separation

Definition (D-Separation). A set of variables E d-separates variables X and Y if E blocks every un-directed path between X and Y in the network.

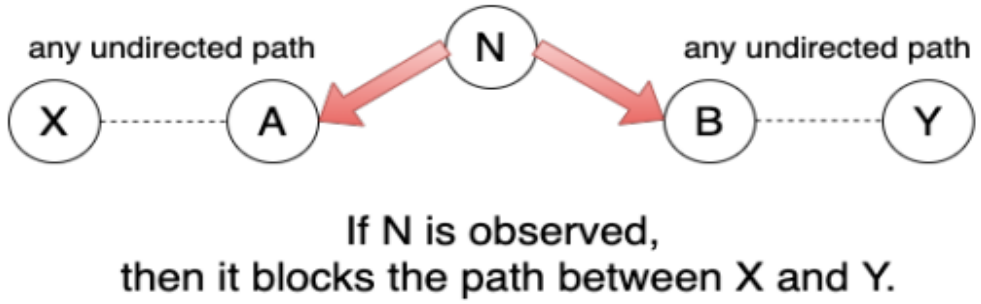
To determine whether X and Y are independent given the observed variables E, we can verify whether E d-separates X and Y. If d-separation holds, then the independence relationship holds as well. To verify d-separation, we need to consider every path between X and Y. On each path, there could be multiple nodes between X and Y. The path is blocked if at least one node blocks the path. When a path is blocked it is considered closed. When a path is not blocked it is considered open.

To better understand blocking one needs to think of a bayesian network as a network system of information channels, where each node is a valve that is either active (open) or inactive(closed) and the valves are connected by noisy information channels. The information flow can pass through an active valve but not an inactive one. When all the valves (nodes) on one undirected path between two nodes are active, we say this path is open. If any one valve in the path is inactive(blocked), we say the path is closed. A valve becomes inactive when observed or is a part of an inactive path.
There are three basic scenarios:

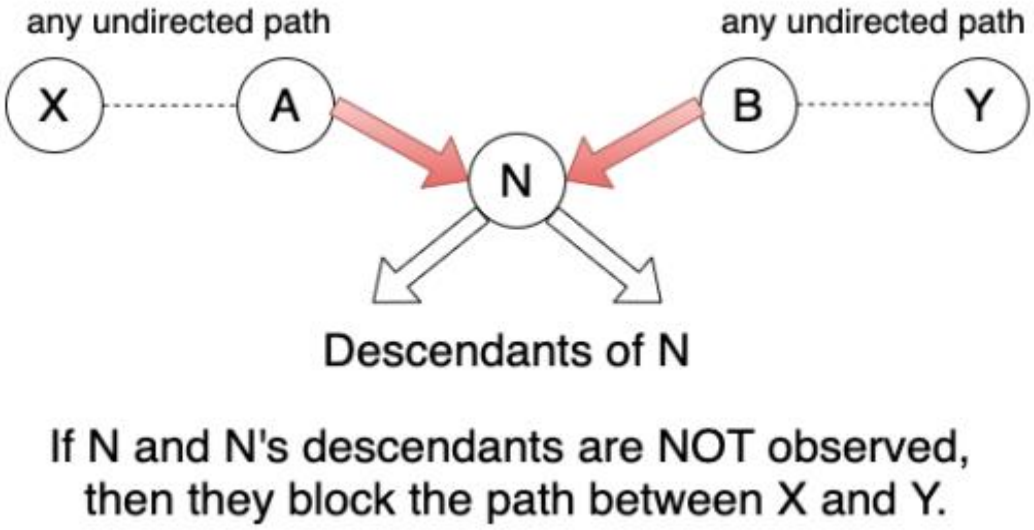
Scenario 1



Scenario 2



Scenario 3



Based on these 3 scenarios we can determine whether a path is blocked or not and as a sequence whether two nodes are (Conditionally independent or not).

The Greedy Thick Thinning algorithm

The Greedy Thick Thinning algorithm constructs Bayesian networks by analyzing dependency relationships among nodes. The dependency relationships are measured by using some kind of conditional independence (CI) test based on the flow of information between two nodes. The amount of information flow between two nodes can be measured by using mutual information, when no nodes are instantiated, or Kullback-Leibler cross entropy, when some other nodes are conditioned upon.

In information theory, the mutual information of two nodes X_i, X_j is defined as:

$$I(X_i, X_j) = \Sigma_{x_i,x_j} P(x_i, x_j) log \frac{P(x_i, x_j)}{P(x_i)P(x_j)}$$

Kullback-Leibler cross entropy is defined as:

$$D(X_i, X_j|C) = \Sigma_{x_i,x_j} P(x_i, x_j, c) log \frac{P(x_i, x_j|c)}{P(x_i|c)P(x_j|c)}$$

where X_i, X_j are two nodes and C is a set of nodes.

The Greedy Thick Thinning algorithm uses cross entropy as CI test to measure the average information between two nodes when the statuses of some valves are changed by the condition-set C. When $D(X_i, X_j|C)$ is smaller than a certain threshold value ε, X_i, X_j are considered d-separated by the condition-set C, and conditionally independent. This algorithm also makes the following assumptions:

(1) The database attributes have discrete values and there are no missing values in all the records.

(2) The volume of data is large enough for reliable CI tests.

(3) The ordering of the attributes is available before the network construction, i.e., a node's parents nodes should appear earlier in the order.

This algorithm has three phases: drafting, thickening and thinning. In the first phase, this algorithm computes mutual information of each pair of nodes as a measure of closeness, and creates a draft based on this information. In the second phase, the algorithm adds arcs when the pairs of nodes cannot be d-separated. The result of Phase II is an independence map (I-map) of the underlying dependency model. In the third phase, each arc of the I-map is examined using CI tests and will be removed if the two nodes of the arc can be d-separated. The result of Phase III is the minimal I-map.

The Algorithm

Phase 1

1) Initiate a graph G(V, E) where V={all the nodes of a data set}, E={ }. Initiate two empty ordered set S, R.
2) For each pair of nodes (vi, vj) where vi, vj $\in$ V, compute mutual information I(vi, vj). For the pairs of nodes that have mutual information greater than a ceftain small value ε, sort them by their mutual information from large to small and put them into an ordered set S.
3) Get the first two pairs of nodes in S and remove them from S. Add the corresponding arcs to E. (the direction of the arcs in this algorithm is determined by the previously available nodes ordering.)
4) Get the first pair of nodes remained in S and remove it from S. If there is no open path between the two nodes (these two nodes are d-separated given empty set), add the corresponding arc to E. Otherwise, add the pair of nodes to the end of an ordered set R.
5) Repeat step 4 until S is empty.
Suppose we have a database that has an underlying Bayesian network as in Figure 1, we also have a nodes' order as A, B, C, D, E and by computing mutual information we get $I(B,D) \geq I(C,E) \geq I(B,E) \geq I(A,B) \geq I(B,C) \geq I(C,D) \geq I(D,E) \geq I(A,D) \geq I(A,E) \geq I(A,C) \geq \varepsilon$.
Sequentially nodes and corresponding arcs will be added. A step by step graphical representation of the procedure is shown bellow. For the pairs (B,D) at fig2 (C,E) at fig3 (B,E) at fig4 (A,B) at fig5 (B,C) at fig 6 will be added. At fig 6 $(A,D), (A,E), (A,C)$, (D,E) will be ommitted due to D-separation.

Figure 1

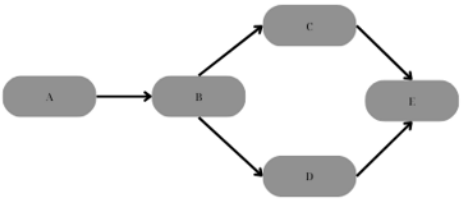


Figure 2

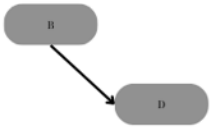


Figure 3

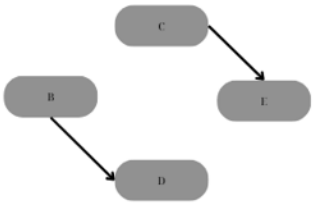


Figure 4

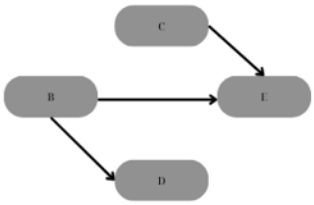


Figure 5

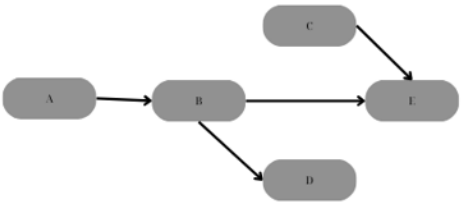


Figure 6

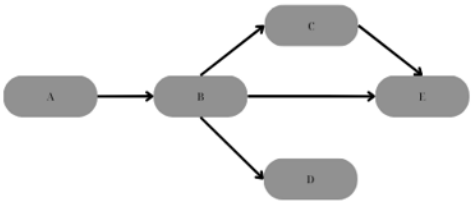


Figure 7

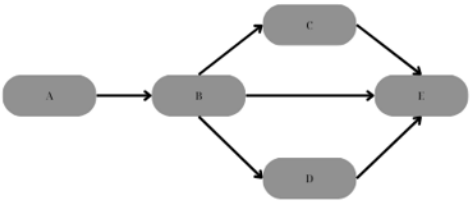
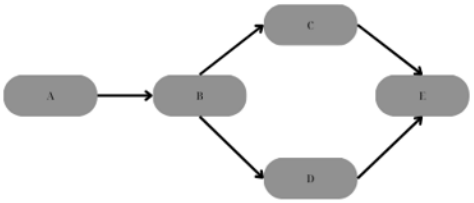


Figure 8



Phase 2 (Thickening)

6) Get the first pair of nodes in R and remove it from R
7) Find a block set that blocks each open path between these two nodes by a set of minimum number of nodes. (This procedure find_block_set (current graph, node1, node2) is given at the end of this subsection.) Conduct a CI test. If these two nodes are still dependent on each other given the block set, connect them by an arc.
8) go to step 6 until R is empty.
In our example, the graph after Phase II is shown in Figure 7. When this algorithm examines the pair of nodes) (D, E) in step 7, it finds that $\{B\}$ is the minimum set which blocks all the open paths between D and E. Since the CI test can reveal that D and E are still dependent given $\{B\}$,arc (D, E) is added resulting in Figure 7.

Phase 3 (Thinning)

9) For each arc in E, if there are open paths between the two nodes besides this arc, remove this arc from E temporarily and call procedure find_block_set (current graph, node1, node2). Conduct a CI test on the condition of the block set. If the two nodes are dependent, add this arc back to E; otherwise remove the arc permanently.
In our example arc (B, E) is removed resulting in figure 8 which is the minimal I-map of the underlying dependency model. In order to avoid large conditioning sets we need to know the minimum d-Separation set. This is achieved with the following algorithm:

Procedure find_block_set (current graph, node1, node2)
Begin
find all the undirected paths between node1 and node2;
store the open paths in open_path_set;
store the closed paths in closed_path_set;
do
while there are open paths which have only one node do
store the nodes of each such path in the block set;
remove all the blocked paths by these nodes from the open_path_set and closed_path_set;
from the closed_path_set, find paths opened by the nodes in block set and move them to the open_path_set. shorten such paths by removing the nodes that are also in the block set;
end while
if there are open paths do
find a node which can block maximum number of the rest paths and put it in the block set;
remove all the blocked paths by the node from the open_path_set and the closed_path_set;
from the closed_path_set, find paths opened by this node and move them to the open_path_set. shorten such paths by removing the nodes t_hat are also in the block set;
end if
until there are no open path
end procedure.

Implementing this algorithm we can get the underlying structure of the Bayesian Network based only on data and not domain knowledge.

The Dirichlet Multinomial Bayesian Model

The Dirichlet Multinomial Bayesian Model is a probabilistic model commonly used in Bayesian statistics, particularly for modeling categorical data. It combines the Dirichlet distribution and the Multinomial distribution to handle scenarios where observations belong to multiple categories, and uncertainty exists about category probabilities.

Key components

1) The Multinomial Distribution

The Multinomial distribution is a generalization of the binomial distribution . Multinomial Distribution models the probability of observing counts across multiple categories and it is Suitable for discrete, categorical data .
Formula :

$$P(x_1, \dots x_n; n, p_1, \dots, p_n) = \frac{n!}{x_1! \ \dots \ x_n!} p_1^{x_1} \dots p_n^{x_n}$$

With $\Sigma_i x_i = n$
where :

- N = total count of observations.
- $x_1 \dots x_n$ = counts for category i.
- $p_1 \dots p_n$ = probability for category i.

for non-negative integers $x_1, \dots x_n$
The probability mass function can be expressed using the gamma function as:

$$P(x_1, \dots x_n; n, p_1, \dots, p_n) = \frac{\Gamma(\Sigma_i x_i + 1)}{\Pi_i \Gamma(x_i + 1)} p_1^{x_1} \dots p_n^{x_n}$$

With $B(x_1 + 1, \dots x_n + 1)^{-1} = \frac{\Gamma(\Sigma_i x_i + 1)}{\Pi_i \Gamma(x_i + 1)}$
This form shows its resemblance to the Dirichlet distribution, which is its conjugate prior.

2) The Dirichlet Distribution

The Dirichlet distribution , often denoted $Dir(a)$, is a family of continuous multivariate probability distributions parameterized by a vector α of positive reals. It is a multivariate generalization of the beta distribution, hence its alternative name of multivariate beta distribution . Dirichlet distributions are commonly used as prior distributions in Bayesian statistics, and in fact, the Dirichlet distribution is the conjugate prior of the categorical distribution and multinomial distribution.

The Dirichlet distribution of order $K \geq 2$ with parameters $\alpha 1, \ldots, \alpha K > 0$ has a probability density function with respect to Lebesgue measure on the Euclidean space

$$f(x_1, \ldots, x_n; a_1, \ldots, a_n) = \frac{\Gamma(\Sigma_i \alpha_i)}{\Pi_i \Gamma(\alpha_i)} \Pi_i x_i^{a_i - 1}$$

Where
$\Sigma_i x_i = 1$ and $0 \leq x_i \leq 1 \; \forall i$.
The normalizing constant is the multivariate beta function, which can be expressed in terms of the gamma function:
$B(a) = \frac{\Pi_i \Gamma(\alpha_i)}{\Gamma(\Sigma_i \alpha_i)}$.

The Dirichlet distribution is the conjugate prior for the Multinomial distribution!

Proof

Dirichlet-Multinomial

Suppose that

$\theta_1, \ldots, \theta_k \sim Dir(a_1, \ldots, a_k)$ (prior)
$y_1, \ldots, y_k \sim Mult(\theta_1, \ldots, \theta_k)$(likelihood)
The posterior:

$$P(\theta|n) = \frac{P(n|\theta)P(\theta)}{P(n)} = \frac{\frac{N!}{\Pi_i n_i!} \Pi_i \theta_i^{n_i} * \frac{1}{B(\alpha)} * \Pi_i \theta_i^{a_i - 1}}{P(n)} = \frac{\frac{N!}{\Pi_i n_i!} * \frac{1}{B(\alpha)} * \Pi_i \theta_i^{n_i + a_i - 1}}{P(n)} \textit{(1)}$$

But

$$P(n) = \int \frac{N!}{\Pi_i n_i!} * \frac{1}{B(\alpha)} * \prod_i \theta_i^{n_i + a_i - 1} \, d\theta =$$
$$= \frac{N!}{\Pi_i n_i!} * \frac{1}{B(\alpha)} * \int \prod_i \theta_i^{n_i + a_i - 1} \, d\theta =$$
$$= \frac{N!}{\Pi_i n_i!} * \frac{1}{B(\alpha)} * B(\alpha + n) (2)$$

Pugging (2) into (1) we have :

$$\mathrm{P}(\theta) = \frac{\frac{N!}{\Pi_i n_i!} * \frac{1}{B(\alpha)} * \Pi_i \theta_i^{n_i + a_i - 1}}{P(n)} = \frac{\frac{N!}{\Pi_i n_i!} * \frac{1}{B(\alpha)} * \Pi_i \theta_i^{n_i + a_i - 1}}{\frac{N!}{\Pi_i n_i!} * \frac{1}{B(\alpha)} * B(\alpha + n)} =$$
$$= \frac{\frac{N!}{\Pi_i n_i!} * \frac{1}{B(\alpha)} * \Pi_i \theta_i^{n_i + a_i - 1}}{\frac{N!}{\Pi_i n_i!} * \frac{1}{B(\alpha)} * B(\alpha + n)} =$$

$= \frac{\prod_i \theta_i^{n_i + a_i - 1}}{B(\alpha + n)} \sim Dir(a + n)$

Choosing hyperparameter vector α with equal weights $(a = (x, \ldots., x))$ makes the prior uniform.

Results

Individual Risk Prediction

The Bayesian network allows for personalized risk assessments by calculating the probability of developing medical conditions based on individual profiles. For example, a clinician might use the tool to assess the risk of diabetes in a middle-aged patient with a sedentary lifestyle and elevated BMI, guiding them to recommend targeted lifestyle modifications such as increased physical activity and dietary changes. Additionally, the network could help identify high-risk groups in a population, such as individuals with multiple overlapping risk factors, enabling proactive intervention strategies. For example, a 50-year-old male who is obese, physically inactive, and sleeps fewer than six hours per night is shown to have a significantly elevated risk of hypertension. Such detailed predictions enable tailored healthcare recommendations.

Population-Level Insights

The model also identifies population-level trends. For example, individuals from lower socioeconomic groups were found to have higher incidences of hypertension and diabetes, emphasizing the role of social determinants in cardiovascular health. This insight supports targeted public health policies to reduce health disparities.

Intervention Analysis

The Bayesian network is a valuable tool for assessing the impact of lifestyle changes on cardiovascular risk. For instance, increasing physical activity from insufficient to regular levels reduces the likelihood of hypertension by approximately 25%. This functionality allows clinicians to prioritize interventions based on their effectiveness.

Discussion

Advantages of the Bayesian Network Approach

1. Comprehensive Analysis of Risk Factors: The model integrates modifiable and non-modifiable factors, providing a holistic view of cardiovascular risk. This integration ensures that static attributes such as age and sex are analyzed alongside dynamic lifestyle factors like BMI and physical activity. By combining these elements, the model can identify patterns that may not be apparent when examining risk factors in isolation. For instance, it highlights how socioeconomic status interacts with lifestyle behaviors to influence hypertension risk, enabling clinicians to develop more nuanced and effective intervention strategies.
2. Dynamic Predictions: The Bayesian framework allows for probabilistic reasoning, meaning that predictions are updated as new evidence becomes available. For example, if a patient's lifestyle factors, such as physical activity or smoking status, change over time, the model can dynamically adjust risk predictions to reflect these updates. This capability provides clinicians with a responsive tool for tailoring recommendations and monitoring patient progress.
3. Causal Relationships: By incorporating expert input, the model reflects causal pathways, enabling clinicians to identify direct influences between risk factors and outcomes. For instance, it elucidates how physical inactivity directly increases the likelihood of hypertension, rather than simply correlating the two. This level of understanding makes the model highly actionable for designing targeted and effective interventions, as it distinguishes between causal and coincidental relationships.
4. Transparent and Interpretable: The graphical representation of relationships simplifies complex interdependencies, making it easier for clinicians and researchers to understand and communicate findings. This clarity fosters confidence in the model's predictions and supports its use in both academic and practical healthcare settings.
5. Scalability: The model effectively handles large datasets by utilizing advanced probabilistic algorithms that ensure computational efficiency. Specifically, the Greedy Thick Thinning (GTT) algorithm minimizes unnecessary computational overhead by focusing on the most relevant dependencies among variables. This optimization, combined with parallel processing capabilities and robust data handling techniques, allows the model to process extensive records while maintaining high accuracy. These features make it well-suited for analyzing large-scale population data, ensuring timely and reliable predictions.

Limitations

1. Static Nature: The current model operates as a static snapshot, which does not account for temporal dynamics. This limitation restricts its ability to monitor changes in risk factors or outcomes over time, such as the progression of lifestyle improvements or the impact of long-term interventions. Developing a dynamic Bayesian network would address this by enabling longitudinal tracking and prediction of evolving health risks.
2. Specific Population: The dataset primarily represents Spanish workers, which introduces the possibility of limited generalizability to other populations with different demographic or health characteristics. Factors like cultural differences in diet, lifestyle, and healthcare access may influence the applicability of the findings. Therefore, the model should be recalibrated with data from diverse populations to ensure broader relevance and accuracy in varied healthcare contexts.
3. Excluded Variables: Factors like diet and alcohol consumption were not included, which could limit the comprehensiveness of the model in capturing all significant influences on cardiovascular health. For example, dietary habits play a critical role in modulating risk factors such as hypertension and diabetes. Similarly, alcohol consumption, depending on frequency and quantity, could interact with other variables to influence outcomes. Incorporating these variables in future iterations of the model would provide a more holistic perspective on cardiovascular risk.
4. Computational Demands: Training and inference for large and complex networks require advanced computational resources due to the high-dimensional data and probabilistic computations involved. However, implementing optimizations like parallel processing, efficient algorithms such as Greedy Thick Thinning (GTT), and streamlined memory management techniques can significantly reduce overhead, making large-scale applications more feasible.

Future Directions

1. Incorporating New Data: Future iterations of the Bayesian network should integrate additional variables, including diet, alcohol consumption, and genetic information. These elements are critical in understanding cardiovascular risk holistically, as diet influences key factors like cholesterol and blood pressure, alcohol consumption can exacerbate or mitigate certain risks depending on usage patterns, and genetic predispositions provide foundational insights into non-modifiable risk factors. Incorporating such data would enhance the model's accuracy and applicability in predicting and managing cardiovascular diseases.

2. Dynamic Modeling: Developing a dynamic Bayesian network would allow for longitudinal analyses, capturing the evolution of risk factors over time. For example, tracking changes in physical activity or BMI could enable personalized treatment adjustments by predicting how these evolving factors influence cardiovascular outcomes. Longitudinal modeling also facilitates early identification of trends, such as deteriorating health metrics, allowing timely interventions to mitigate potential risks.
3. Expanding to Diverse Populations: Applying the model to data from different age groups, ethnicities, and socioeconomic backgrounds would significantly enhance its generalizability. For instance, incorporating data from underrepresented groups could reveal unique risk patterns or interactions between factors, such as the interplay of cultural dietary habits and socioeconomic status in influencing cardiovascular health. This broader scope would allow the model to provide tailored predictions and interventions for diverse global populations.
4. Real-Time Integration: Connecting the model to wearable devices and electronic health records would enable real-time monitoring and prediction. For instance, wearable devices could provide continuous updates on physical activity, heart rate, or sleep patterns, allowing the model to dynamically adjust risk estimates. Similarly, integration with electronic health records could ensure that clinical data such as lab results and medication adherence are incorporated seamlessly, enhancing the model's predictive accuracy and utility in personalized healthcare.
5. Cost-Effectiveness Analysis: Incorporating decision and utility nodes into the Bayesian network would enable detailed evaluations of the financial implications of various interventions. For instance, the model could compare the cost of lifestyle modification programs (e.g., increasing physical activity or improving diet) with their projected benefits in reducing long-term healthcare expenses. This capability would support policymakers and clinicians in prioritizing cost-effective strategies while maximizing health outcomes.

Conclusion

The Bayesian network developed in this study offers a robust framework for predicting cardiovascular risk and analyzing the interplay of various risk factors. By combining data-driven insights with expert knowledge, the model supports precision medicine and public health strategies. The accompanying software tool, implemented using the GeNIe platform, allows users to interact with the Bayesian network by inputting specific patient data to receive probabilistic predictions for various risk factors and medical conditions. Clinicians can easily adjust variables, such as BMI or physical activity, to simulate interventions and observe their effects. Additionally, the software supports real-time exploration of complex relationships between risk factors, offering actionable insights for both individual and population-level strategies. Its accessibility

through an open-source GitHub repository ensures that researchers can further customize and adapt the tool for diverse applications.

Citations

The study by Ordovás et al., which developed a BN model for cardiovascular risk prediction, has had a significant and wide-ranging impact across multiple domains. This foundational work has inspired a variety of studies that utilize and adapt its methodologies, demonstrating its relevance in both theoretical advancements and practical applications. Below, we detail the specific ways in which this study has been referenced and its influence leveraged across fields:

1. Bayesian Networks in Medical Diagnostics

The healthcare domain has extensively cited Ordovás et al. for demonstrating the versatility of BNs in predictive modeling. For instance, studies on breast cancer detection and colorectal cancer risk mapping highlight the methodology's ability to integrate expert-driven knowledge with observational data. These studies illustrate how BNs provide a structured approach to understanding complex medical datasets. In particular, the breast cancer detection research emphasized the predictive power of Bayesian networks in analyzing intricate relationships among risk factors, using Ordovás et al.'s cardiovascular model as a guiding example. Similarly, the colorectal cancer study underscored the applicability of Ordovás et al.'s techniques in mapping disease risk through probabilistic reasoning, showcasing the relevance of expert-informed models in clinical settings.

2. Expansion to Broader Methodological Frameworks

Ordovás et al. has also played a key role in expanding BN methodologies into novel territories. For example, the development of the NEAR framework—a model-agnostic system for compressing complex decision-support models—cites Ordovás et al. as a benchmark for explainable and structured predictive systems. The framework extends the foundational ideas of BNs to ensure modularity and interpretability in artificial intelligence applications. Similarly, research on Bayesian shrinkage priors for biomedical diagnostics references the work to underline the importance of probabilistic approaches in managing uncertainty and complexity in spectroscopic data. These citations highlight how Ordovás et al. has inspired innovative uses of Bayesian methodologies beyond traditional healthcare applications.

3. Risk Analysis in Diverse Domains

The adaptability of BNs as demonstrated by Ordovás et al. is a recurring theme in studies focused on risk prediction across various fields. A study on Pay-As-You-Drive insurance models cited the integration of large datasets with expert knowledge in the Ordovás et al. framework, using it as a model for developing predictive risk tools in insurance. Similarly, a study addressing ethno-racial disparities in chronic diseases used Ordovás et al. as a comparative benchmark for its BN approach. These works reflect how the principles established by Ordovás et al. are not limited to healthcare but extend into domains requiring robust risk assessment tools.

4. Foundational Role in Cardiovascular Research

Several studies in cardiovascular research directly draw upon Ordovás et al.'s methodology. For instance, a study on BN models for heart disease classification references Ordovás et al. as a methodological cornerstone, emphasizing the use of modifiable and non-modifiable risk factors in network construction. Similarly, research on atherosclerotic cardiovascular disease risk factors in the AZAR cohort extensively discusses Ordovás et al.'s integration of domain expertise and probabilistic modeling. These studies validate and extend the work by applying similar approaches to new datasets and clinical populations.

5. Illustrative Benchmark for Methodological Rigor

Many researchers have used Ordovás et al. as a benchmark to validate and enhance their own methodologies. A study on BN updates in clinical data cited Ordovás et al. as a reference for handling uncertainty and dependency in healthcare datasets. Likewise, the development of novel Bayesian models for heart disease risk assessment explicitly referenced the work for its balance of domain knowledge and data-driven insights. These citations underscore the reliability and applicability of the Ordovás et al. model as a methodological gold standard in Bayesian network research.

6. General Impact on Bayesian Methodology

Ordovás et al. has broadly influenced the perception and adoption of Bayesian networks as a tool for analyzing complex systems. By demonstrating their utility in cardiovascular health, this study has laid the groundwork for similar applications in public health, insurance, and other data-rich fields. It is repeatedly cited for its ability to synthesize expert opinion with observational data, a hallmark of effective Bayesian modeling.

Concluding remarks

Through its contributions to cardiovascular risk prediction and Bayesian network methodologies, the study by Ordovás et al. has achieved widespread recognition in both academia and applied research. The citations it has garnered reflect its importance as a foundational work, providing a methodological blueprint for leveraging Bayesian networks to tackle complex, probabilistic problems. Its influence extends beyond healthcare, impacting diverse domains such as insurance, artificial intelligence, and public health. This breadth of application underscores the lasting relevance of Ordovás et al. in advancing Bayesian methods for data-driven decision-making.

References

1. Wang, Y., et al. (2020). Survivability Modeling Using Bayesian Networks for Patients with First and Second Primary Cancers. *Journal of Biomedical Informatics*. [Extracted from prior work upload]
2. Kyrimi, E., et al. (2021). A Comprehensive Scoping Review of Bayesian Networks in Healthcare. *Journal of Applied Artificial Intelligence*.
3. Farooq, K., et al. (2011). Ontology-Driven Cardiovascular Decision Support System. *Computational Intelligence in Cardiovascular Healthcare*.
4. Tylman, W., et al. (2016). Real-Time Prediction of Acute Cardiovascular Events Using Bayesian Networks. *Journal of Emergency Cardiology*.
5. Thornley, S., et al. (2013). Using Directed Acyclic Graphs for Investigating Causal Paths for Cardiovascular Disease. *BMC Medical Research Methodology*.
6. Roberts, J., et al. (2006). Bayesian Networks for Cardiovascular Monitoring: Integrating Diverse Data Sources. *International Journal of Cardiology Informatics*.
7. Santos-Lozano, A., et al. (2021). Association Between Physical Activity and Cardiovascular Risk Factors. *Cardiovascular Medicine and Exercise Science*.
8. Fiuza-Luces, C., et al. (2018). Exercise Benefits in Cardiovascular Disease: Beyond Attenuation of Traditional Risk Factors. *Nature Reviews Cardiology*, 15(12), 731–743.
9. World Health Organization (2021). Cardiovascular diseases (CVDs). *WHO Fact Sheets*.
10. Huh, J., (2023). Bayesian Network Updating Using Novel Clinical Data: A Practical Approach. *Healthcare Informatics Research*.
11. Esmaeili, P., et al. (2024). Unraveling atherosclerotic cardiovascular disease risk factors through conditional probability analysis with Bayesian networks: insights from the AZAR cohort study. *Journal of Clinical Cardiology*.
12. Salman, I., et al. (2023). Development and Performance Evaluation of a Novel Bayesian Network Model for the Classification of Heart Disease. *Cardiovascular Computing*.
13. Wang, B., et al. (2024). A Novel Bayesian Pay-As-You-Drive Insurance Model with Risk Prediction and Causal Mapping. *Risk Analysis Journal*.
14. Babagoli, M., et al. (2024). Bayesian Network Model of Ethno-Racial Disparities in Cardiometabolic-Based Chronic Disease Using NHANES 1999–2018. *American Journal of Preventive Medicine*.
15. Chu, O., H., et al. (2024). Development and Application of an Optimized Bayesian Shrinkage Prior for Spectroscopic Biomedical Diagnostics. *Journal of Biomedical Statistics*.
16. Sun, B., et al. (2023). A Breast Cancer Detection Method Based on Bayesian Networks. *Journal of Oncology Informatics*.
17. Corrales, D., et al. (2024). Colorectal Cancer Risk Mapping Through Bayesian Networks. *Cancer Epidemiology Research*.
18. Kassem, K., et al. (2024). An Innovative Artificial Intelligence-Based Method to Compress Complex Models into Explainable, Model-Agnostic and Reduced Decision Support Systems with Application to Healthcare (NEAR). *Artificial Intelligence in Medicine.*
19. Cheng, J., Bell, D. A., & Liu, W. (1997). An algorithm for Bayesian belief network construction from data. *Proceedings of the Conference on Artificial Intelligence and Statistics,* 83–90