Concept of Usage of Bayesian Networks in Clinical Decision Support Module

M. STRAWA
marcin.strawa@gmail.com

Institute of Computer and Information Systems
Faculty of Cybernetics, Military University of Technology
Kaliskiego Str. 2, 00-908 Warsaw, Poland

Concept of decision support module utilizing a repository of clinical pathways has been presented in this paper: the definition of Bayesian networks and its major concepts, description of chosen inference algorithm and an example of diagnosis.

Keywords: Bayesian networks, belief networks, clinical decision support system.

1. Introduction

Bayes' theorem expresses the conditional probability of hypothesis H (given evidence E) in terms of the prior probability of H, the prior probability of E, and the conditional probability of E given H.

$$P(H \mid E) = \frac{P(H)P(E \mid H)}{P(E)} \tag{1}$$

Bayesian network is a tool that is based on this theorem. In this paper, the concept of usage of Bayesian networks in a clinical decision support module is presented. The purpose of the module is to cooperate with clinical pathways repository and taking decisions, which are located in decision nodes of appropriate pathways, as well as the selection of the proper pathway to follow. It is illustrated in an example, where preliminary diagnosis is taken and basing on this, the clinical pathway is chosen. Also, the algorithm is presented for making decisions in a single decision node of the pathway. The disease chosen for the example is chronic myeloid leukemia.

The paper begins with the definition of Bayesian networks and its major concepts. Next, the short description of chronic myeloid leukemia takes place.

In the following section, based on the description of the diagnostic process, the simple clinical pathway has been constructed as an example. It is a basis for the illustration of the reasoning procedure, which is shown further. The next section shows the reasoning algorithm and its execution for the exemplary Bayesian network.

2. Bayesian Networks

The Bayesian network (other name: *belief network*) is a probabilistic graphical model representing a set of random variables and its conditional dependencies as a acyclic directed graph. Vertices of a Bayesian network represent all attributes defined in the problem's domain, while edges can be interpreted as a representation of the direct causal dependency between them.

There are a few formal definitions of the Bayesian network. For all the definitions given below let us assume that $G = (V, E)$ is an acyclic directed graph, and $X = (X_v)_{v \in V}$ is a set of random variables indexed by V.

1. X is Bayesian network with respect to G, if its joint probability distribution can be defined as a product of conditional probabilities of all the nodes:

$$P(x) = \prod_{v \in V} P\left(x_v \middle| x_{pa}(v)\right) \tag{2}$$

where $pa(v)$ is a set of all parents of node v.

2. X is Bayesian network with respect to G, if it satisfies the *local Markov property*: each node is conditionally independent of its non-descendants given values of its parent nodes:

$$\begin{aligned} &P\left(X_v = x_v \middle| \forall j \in V - \{v\} - S_v : X_j = x_j\right) = \\ &= P\left(X_v = x_v \middle| \forall j \in U_v : X_j = x_j\right) \end{aligned} \tag{3}$$

where:

S_v – set of numbers of all (direct or indirect) descendants of node v

U_v – set of numbers of all parents of node v

The main advantage of Bayesian networks (having a proper structure) is the ability of representing indirectly the joint probability distribution of all variables in an efficient way. To represent such a distribution with the Bayesian network, for each node v it is required to know conditional probabilities of its values, given the values of its parent nodes. It is sufficient then to store Vk^{u+1} probability values (where: $k = \max_{v\in V}|X_v|$, $u = \max_{v\in V}|U_v|$), while the direct representation of the joint probability distribution would require to store the following number of probability values: $\prod_{v=1}^{V}|X_v| \le k^n$.

Example[1]:

House alarm systems react to burglaries as well as earthquakes. Neighbours Mary and John are agreed to call the owner when they hear the alarm. John always calls, but sometimes takes the ringing phone for an alarm signal and calls then, too. Mary likes loud music and then sometimes misses the alarm. Given the evidence who has or has not called we want to estimate the probability of a burglary. The Bayesian network for this example is presented below:

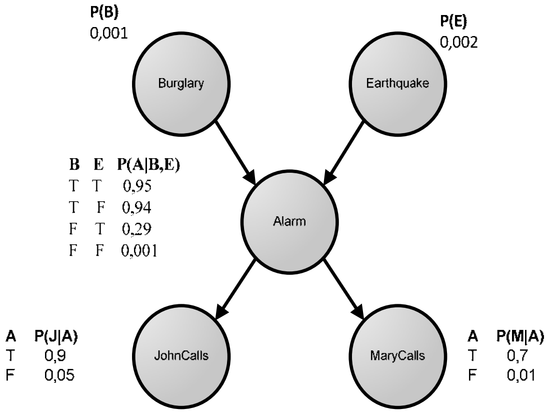

Fig. 1. Example Bayesian network

Let us say we want to calculate the probability of the alarm when there was no burglary and earthquake, given that both John and Mary called:

$$P(J \wedge M \wedge A \wedge \neg B \wedge \neg E) =$$
$$= P(J \mid A)P(M \mid A)P(A \mid \neg B \wedge \neg E)P(\neg B)P(\neg E) =$$
$$= 0.90 \cdot 0.70 \cdot 0.001 \cdot 0.999 \cdot 0.998 = 0.00062$$

Knowledge of joint probability distribution for all variables makes it possible to carry out probabilistic reasoning for values of every combination of variables, given the values of other variables.

Let $V_0(x_q)$ – set of numbers of variables having known values;

$V_h(x_q) - V - V_0(x_q)$ – set of numbers of variables, which values are not known for some example $x_q \in X$. We look for probability distribution of variables numbered by $V_h(x_q)$, having given values of variables numbered by $V_0(x_q)$. To calculate it, the following formula can be used:

$$P(\forall i \in V_h(x_q): X_i = x_i | \forall i \in V_0(x_q): X_i = x_i) =$$
$$= \frac{P(\forall i \in V: X_i = x_i)}{P(\forall i \in V_0(x_q): X_i = x_i)}$$

for all values of $x_i \in X_i$ if $i \in V_h(x_q)$ and for known values of $x_i = x_q$ if $i \in V_0(x_q)$.

The answer for every query can be obtained by calculating, with the usage of the network, joint probability distribution and applying it for subsequent calculations. Unfortunately, this approach means giving up one of the main advantages of representing joint probability distribution as a Bayesian network – efficiency. Due to this fact, other algorithms are used for answering such queries. In general, reasoning in Bayesian networks is NP-hard, so approximation algorithms are mainly utilized in problem solving. There is also one type of Bayesian network for which the reasoning problem is much simpler, so that effective exact algorithms can be applied. These are networks, where only one undirected path exists between any pair of nodes.

Simple examples of four reasoning patterns, for which Bayesian networks can be utilized are shown on graph 2. *E* stands for observable attribute (evidence variable), *Q* – for attribute, which is a subject of a query (question variable).

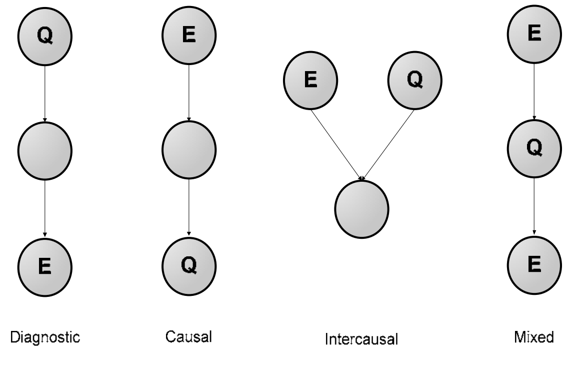

Fig. 2. Reasoning patterns that can be handled by a Bayesian network

[1] Example taken from [2]

3. Chronic Myeloid Leukemia

Leukemia[2] is a malicious tumor of hematopoietic cells being formed as a consequence of systemic, scattered and autonomous growth of one leucocyte clone and the spreading of cancer-altered, immature blast cells from bone marrow into the blood.

Main forms of leukemia can be divided into four categories:

- acute myeloid leukemia (AML)
- chronic myeloid leukemia (CML)
- acute lymphocytic leukemia (ALL)
- chronic lymphocytic leukemia (CLL)

In contrary to acute leukemia, progress of chronic myeloid leukemia (CML) is long lasting and relatively slow.

Despite CML it is quite a frequent category of leukemia, its occurrence is rare, taking into account global population. Most patients are adults. Children are only 2–4% of the cases.

Chronic myeloid leukemia is caused by changes in the genetic code of some cells in the bone marrow. In these cells, a part of chromosome 9 becomes a place of part of chromosome 22 – a process called translocation. Abnormal chromosome called the Philadelphia chromosome is formed. Abnormal chromosome stimulates the overproduction of white blood cells in the bone marrow.

Chronic myeloid leukemia generally proceeds in three phases. Most patients are diagnosed in the initial phase called chronic. Over time it transforms into the acceleration phase – the disease accelerates, and finally into the blastic phase – the most malicious and of a course similar to acute leukemia.

Chronic phase is a first phase of disease and lasts much longer than others. There is a larger number of white blood cells in blood and bone marrow, but most of them are mature cells that function properly. Most patients (80%) remain in the stable phase for at least 5 years. Symptoms of the chronic phase of CML depend on what kind of white blood cells are present in the blood of a given patient. Typically, the symptoms are scarce, and the disease is diagnosed by routine blood tests

Symptoms may include:

- fatigue
- headache
- pain or feeling of fullness in the left mid-abdomen (caused by an enlarged spleen).

Acceleration phase. At this stage there is an increasing number of immature cells (blasts) in the blood, bone marrow, liver and spleen. Blasts cannot fight infections like normal white blood cells. In the past, the length of acceleration was usually one to six months before progressing to the blastic phase. Depending on the treatment, this phase can be extended to more than 1 year.

Signs of accelerated phase are more intensive and include:

- fever
- night sweats
- weight loss
- pale skin, easy fatigue, shortness of breath (deficiency of red blood cells, or anemia).

Blastic phase. In this phase comes the rapid progression of the disease and the creation of huge numbers of malignant cells in the blood. The result is an increasing number of blasts from the marrow and the displacement of normal blood cells in the blood – red cells, white blood cells and thrombocytes. Patients often report problems with infections, easy bruising and bleeding. The course of the disease resembles acute myeloid leukemia, or in rare cases, acute lymphoblastic leukemia.

In order to recognize chronic myeloid leukemia and to assess the progress of the disease, a puncture is carried out (bone marrow picking).

4. Clinical Pathway

Based on the description above, a sample fragment of the clinical pathway for CML, enclosing the diagnosis stage, can be constructed. It will be useful further in this paper as an input to the decision support module utilizing Bayesian networks.

[2] Description taken from [5] and [4]

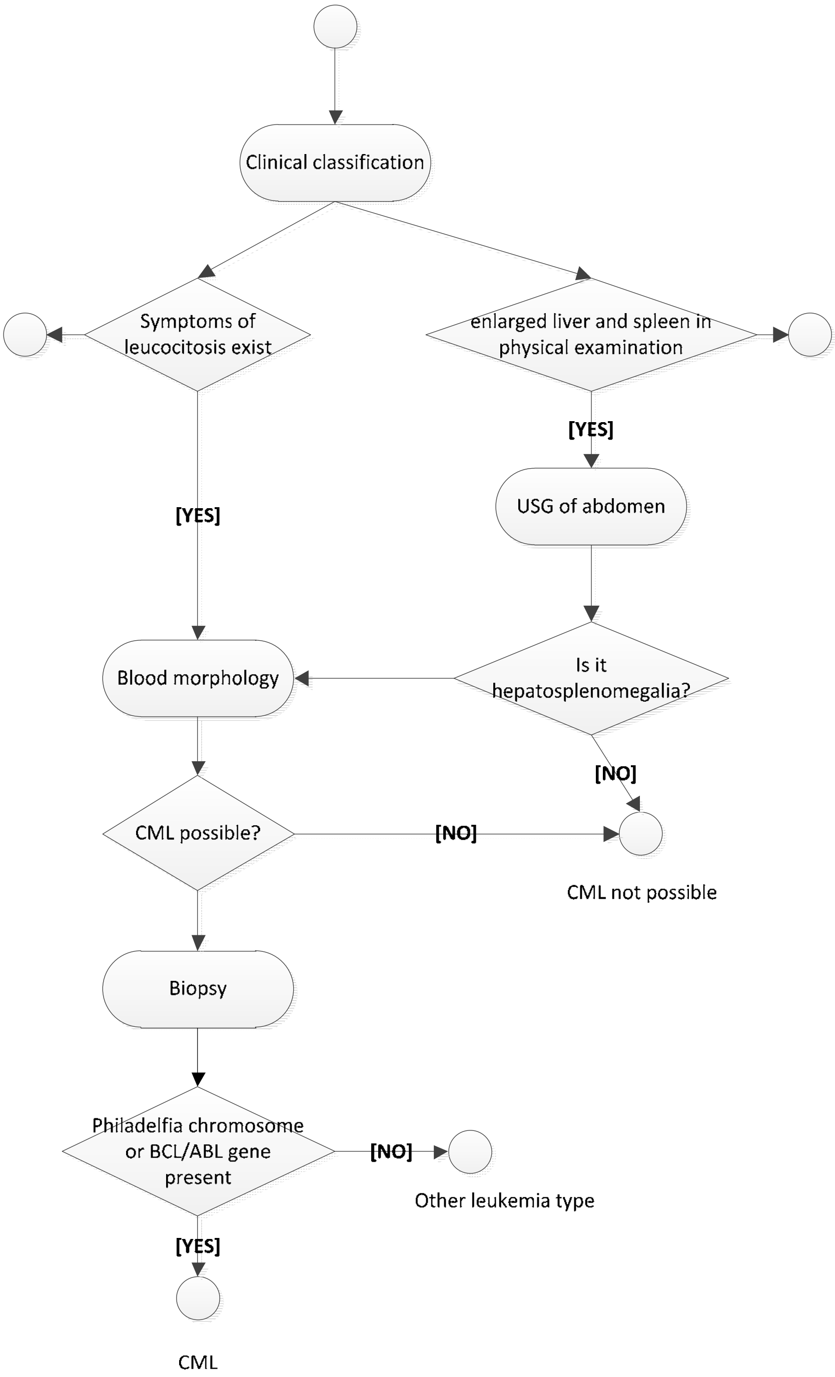
Clinical classification
Symptoms of leucocitosis exist
enlarged liver and spleen in physical examination
[YES]
USG of abdomen
[YES]
Blood morphology
Is it hepatosplenomegalia?
[NO]
CML possible?
[NO]
CML not possible
Biopsy
Philadelfia chromosome or BCL/ABL gene present
[NO]
Other leukemia type
[YES]
CML

Fig. 3. Sample fragment of the clinical pathway for CML

5. Example of diagnosis with the usage of bayesian network

In order to show a mechanism of reasoning, sample bayesian network for CML has been constructed. Probability values have been taken arbitrary, only to illustrate the example. In working system they must be determined with help of domain experts as well as by network learning (sample methods are presented in further part of this paper).

The purpose of the constructed network is to conduct diagnostic reasoning. For evidence variables stand observable disease symptoms (through patient interview or examinations results), for question variables – diseases. Continuous variables have been transformed into discrete variables by dividing their set of values into ranges bounded with values having medical significance (limits of various norms for a healthy adult person, typical values for analyzed diseases, etc.).

Algorithm. For every iteration of the reasoning process, the disease having highest occurrence probability, in context of known symptoms, is searched. Then the clinical pathway, which is most suitable for found disease, is selected. If the probability value is not equal 1 or a defined level for proper diagnosis, the decision will be taken to make examinations defined on a selected pathway required to verify the current diagnosis. Results of examinations extend the set of evidence variables and a new iteration takes place. Results of examinations defined on the clinical pathway, which results are known, are marked as done to avoid multiple executions. There are two stop conditions:

1. diagnosis has been found, there is no other one with a higher possibility value,
2. all examinations defined on executed pathways have been done.

The reasoning method applied in each algorithm iteration. As it was stated before, the reasoning problem in Bayesian networks is NP-hard and accurate, effective algorithms exist only for networks having a polytree structure. The network constructed for the analyzed diagnostic problem does not have a mentioned structure, as there can be found at least one pair of nodes having more than one undirected path between them. Usage of approximation algorithms is required then.

Three classes of reasoning algorithms for Bayesian networks are known:

- Clustering methods – the network is transformed into probabilistically equivalent (but topologically different) polytrees by merging nodes. Then known accurate algorithms can be applied
- Conditioning methods – variables in networks are substituted with particular values. Every possible substitution is evaluated
- Stochastic simulation (Monte – Carlo) – big number of samples (networks with defined values of attributes) is generated, for which conditional probabilities in nodes are consistent with the ones in the analyzed network. Distribution of results is an approximation of exact evaluation.

Interesting effectiveness comparison of a few most popular algorithms can be found in [6]. For the analyzed example Monte-Carlo-class algorithm will be used. Its name is *Likelihood Weighting*.

Every iteration of simulation using this algorithm looks like the following:

1. generate values of variables for all root nodes with probability distribution defined in nodes,
2. for each following node:
   a. if the node is not an evidence variable: generate the variable's value according to its conditional probabilities table, assuming known values of conditions,
   b. if the node is an evidence variable: find the probability value in its conditional probabilities table assuming the known value of the observed variable and known values of conditions. The found probability will be the weight of whole simulation step.

After finishing the simulation step, the probability of reaching some value by the question variable, under the condition of evidence variables, is known.

The simulation result is the quotient of the sum of probabilities of interesting events' occurrences by the sum of all probabilities attained in simulation steps.

```
function LikelihoodWeighting(X, e, n, N)
returns estimation P(X|e)
local variables: W, vector of weights of values
in X

for j = 1 to N do
   x, w := WeightedSample(n, e)
   W[x] := W[x] + w where x is value of X in x
return Normalize(W[X])
```

```
function WeightedSample(n, e) returns event
and weight

x := n-element event;
w := 1
for i = 1 to n do
   if $X_i$ has value $x_i$ in e
      then w := w × $P(X_i = x_i \mid Parents(X_i))$
      else $x_i$ := random value with $P(X_i \mid Parents(X_i))$
return x, w
```

Bayesian network for chronic myeloid leukemia. Nodes represent chosen symptoms and causes of the disease. The node representing leukemia is node **CML**. There is also a few other diseases present, which can be potential causes of the symptoms.

Inference for the exemplary network. Let us provide patient complains about seeing disorders, night sweats and during the physical examination where the enlarged spleen was noted. So the specified symptoms are evidence variables in our exemplary network. Node CML represents the question variable.

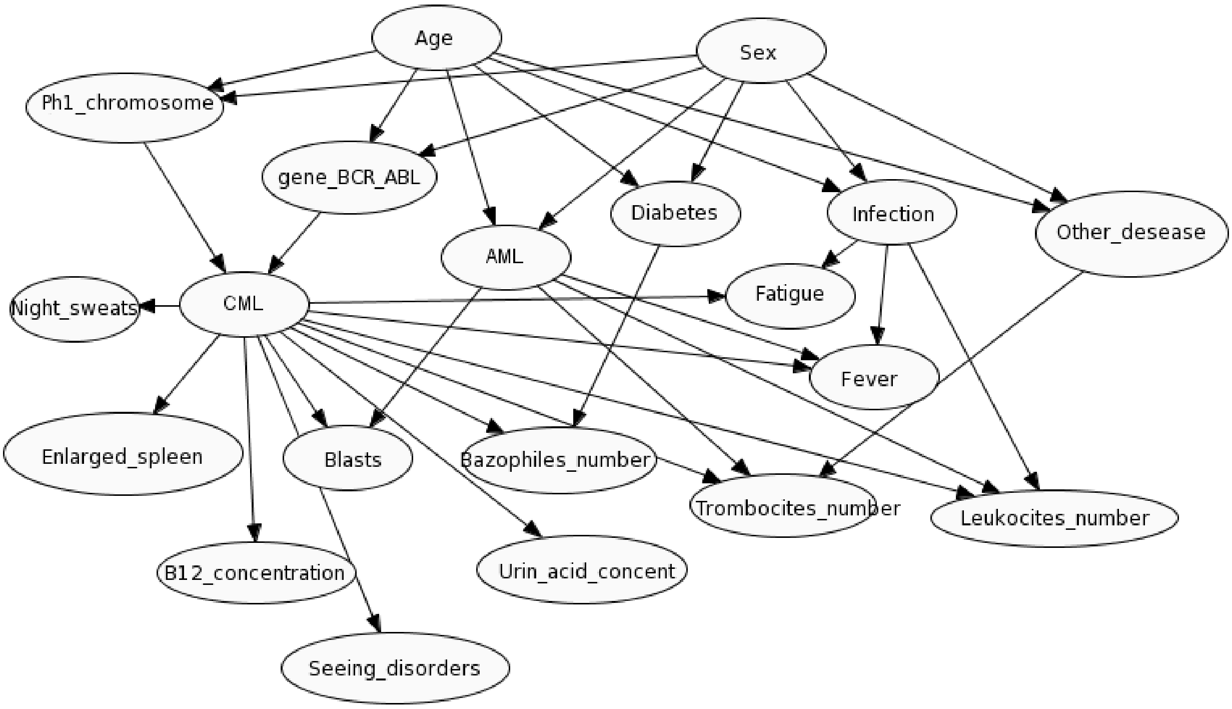

Fig. 4. Sample Bayesian network for chronic myeloid leukemia

In one iteration of the likelihood weighting algorithm, nodes (due to limited space, proceeding for only the chosen ones is presented) will take on the values as the following:

1. Take the weight of the iteration *w* = 1.
2. For node **Age** draw a value according to its probability table. Let us provide value is *Age* = 40 – 60.
3. For node **Sex** draw a value according to its probability table. Let us provide that the value is *Sex* = *F*.

   For node **Ph1_chromosome** draw a value according to its conditional probability table. For values of predecessors: *Age* = 40 – 60 and *Sex* = *F* probability values are: 0.000002, that *Ph*1_ *chromosome* = *True* and 0.999998, that *Ph*1_ *chromosome* = *False*. Generation takes place according to these values. Let us provide that the value drawn is *Ph1_ chromosome* = *False.*
4. For node **gene_BCR_ABL** draw a value according to its conditional probability table. For values of predecessors: *Age* = 40–60 and Sex = *F* probability values are: 0.000012, that *gene_BCR_ABL* = *True* and 0.999988, that *gene_BCR_ABL* = *False*. Generation takes place according to these values. Let us provide the value drawn *gene_BCR_ABL* = *True.*
5. For node **CML** draw a value according to its conditional probability table. For values of predecessors: *ph1_chromosoe* = *False* and *gene_BCR_ABL* = *True* probability values are 0.99999, that *CML* = *True* and 0.00001, that *CML* = *False*. Let us provide the value drawn *CML* = *True.*
6. **Night sweats** node is an evidence variable as it contains symptoms found during examination. We know that its value equals *True*, and the predecessor's value: *CML* = *True*. So the weight of iteration *w* must by modified according to node's conditional probabilities table. *w* := *w* * *P*(*Night_sweats* = *T* | *CML* = *T*) = *w* * 0,1 = 0,1.
7. **Enlarged_spleen** node is an evidence variable. We know that its value equals *True*, and predecessor's value: *CML* = *True*. So the weight of iteration *w* must by modified according to node's conditional probabilities table. *w*:= *w* * *P*(*Enlarged_spleen* = *T* | *CML* = *T*) = 0,1 * 0.35 = 0.035.
8. **Seeing_disorders** node is an evidence variable. We know that its value equals
9. *True*, and the predecessor value: *CML* = *True*. So the weight of iteration *w* must by modified according to node's conditional probabilities table. *w*: = *w* * *P*(*Seeing_disorders* = *T* | *CML*= *T*) = 0.035 * 0.1 = 0.0035.
10. Return set of all nodes' values together with weight of iteration *w* = 0.035.

After conducting the necessary number of simulations, following the steps above, the probabilities for interesting nodes must be determined. In the analyzed case the interesting nodes are: CML, AML, diabetes, infection, other_disease and more specific – probability that their value equals *True*. Thus, for each one the following value must be calculated:

$$P(X = True \setminus e) = \frac{\sum_{i \in W : X = True} W_i}{\sum_{i=1}^{N} W_i},$$

where:
X – interesting node,
N – number of simulation steps,
e – observable attributes,
W – vector of simulation results.

Let us provide that the calculated probability values are as the following:

- *P*(*CML* = *T*) = 0.2
- *P*(*AML* = *T*) = 0.01
- *P*(*Diabetes* = *T*) = 0.007
- *P*(*Infection* = *T*) = 0.08
- *P*(*Other_disease* = *T*) = 0.13

So the decision on the diagnosis cannot be taken, but the node having the highest probability value is CML. The pathway for this disease must be then proposed. As the first examination showed a suspicion of spleen enlargement, the system will advise USG of the abdomen in this decision node. Next, after the results of new examinations are presented in the network as a evidence variables, the reasoning process will be repeated.

The major problem in the application of the presented method is the necessity of performing a big number of simulations (which means a large amount of time) to obtain precise probability values for the least probable events. The time required to reach a particular precision level is reversely proportional to the probability of the event.

6. Learning Bayesian Networks Basing on Examples

Although it was provided that the network would be constructed using knowledge of the domain experts, the ability to automatically construct one may significantly increase its usability. It can be achieved using methods of Bayesian networks learning with the usage of training data.

There are two criteria by which we can group problems of learning:

- knowledge of the network's structure or the lack of it
- all or only part of the attributes are observable in the training data.

For the proposed reasoning module, we have a problem where the network's structure is well defined but not all values of the attributes for the training data are known. The problem can be transformed to the calculation of conditional probability tables for the network with a defined structure and some training set T. The goal of learning is to find a hypothesis h, which is most consistent with training data. It means maximization of probability P(T|h). Descriptions of algorithms used to achieve this goal can be found, for example, in [1], [2], [8].

Many researches are currently made in this area. Interesting methods can be found i.e. in [9].

7. Summary

The concept of the decision support module utilizing the repository of clinical pathways has been presented in this paper. The utility used in this module is the Bayesian network. It has already been successfully applied in supporting medical diagnostic processes in the country, as well as abroad. Examples of these systems are HEPAR and MUNIN. The concept of graphical network presentation makes it easier to construct one in cooperation with domain experts, because it helps to understand causal dependencies between variables.

The problem of inference in Bayesian networks is NP-hard, but the number of effective algorithms producing approximate results of good quality was invented. One of them is, described here, Likelihood Weighting algorithm, based on the Monte-Carlo approach. Its major weakness is the precision of computing probability values of slightly probable events. However, modifications exist, which allow reducing this problem. There are also researches on effective reasoning methods.

The other developing domains are algorithms of learning Bayesian networks, which are very useful for constructing networks with the usage of training data.

Implementation of the described module would allow performing few experiments regarding effectiveness and performance of various learning and inference algorithms, in cooperation with various clinical pathways. It might lead to the formulation of a few research problems.

8. Bibliography

[1] P. Cichosz, *Systemy uczące się*, WNT, Warszawa, 2007.

[2] S.J. Russel, P. Norvig, *Artificial Intelligence: A Modern Approach*, Prentice Hall, Englewood Cliffs, New Jersey, 1995.

[3] A. Ameljańczyk, „Analiza wpływu przyjętej koncepcji modelowania systemu wspomagania decyzji medycznych na sposób generowania ścieżek klinicznych", *Raport z realizacji zadania 2. projektu POIG.01.03.01-00-145/08*, WAT, Warszawa, 2009.

[4] B. Kowalczyk, „Przewlekła białaczka szpikowa", *Encyklopedia zdrowia MediWeb.pl*, http://mediweb.pl/diseases/wyswietl_d.php?id=92

[5] *abc Białaczka.pl*, http://abcbialaczka.pl

[6] R. Liu, R. Soetjipto, *Analysis of Three Bayesian Network Inference Algorithms: Variable Elimination, Likelihood Weighting, and Gibbs Sampling*, Berkeley, 2004.

[7] P. Długosz, „Opracowanie koncepcji modułu wspomagania podejmowania decyzji klinicznych w modelu repozytorium z wykorzystaniem metod teorii zbiorów przybliżonych", *Raport z realizacji zadania 3. projektu POIG.01.03.01-00-145/08*, WAT, Warszawa, 2009.

[8] D. Heckerman, *A Tutorial on Learning with Bayesian Networks*, Microsoft Corporation, Redmond, 1995.

[9] R. Niculescu, T. Mitchell, R. Rao, "Bayesian Network Learning with Parameter Constraints", *Journal of Machine Learning Research 7*, 1357–1383, MIT, Boston, 2006.

[10] A. Oniśko i inni, *HEPAR I HEPAR II – komputerowe systemy wspomagania diagnozowania chorób wątroby*, XII Konferencja Biocybernetyki i Inżynierii Biomedycznej, Warszawa, 2001.

[11] S. Andreassen, "MUNIN – An Expert EMG Assistant", *Computer-Aided Electromyography and Expert Systems*, Vol. 21, Elsevier Science Publishers, Amsterdam, 1989.

[12] "Chronic Myelogenous Leukemia", *NCCN Practice Guidelines in Oncology*, http://www.nccn.org/professionals/physician_gls/PDF/cml.pdf

[13] "Acute Myeloid Leukemia", *NCCN Practice Guidelines in Oncology*, http://www.nccn.org/professionals/physician_gls/PDF/aml.pdf

[14] „Ścieżki kliniczne jako dynamiczne środowisko dostępu do informacji medycznej pacjenta", *wersja 0.8 Zintegrowany System Informacji Medycznej o Pacjencie*, Bielsko-Biała – Kraków, 2008.

[15] „Przewlekła białaczka szpikowa", *Wikipedia*, http://pl.wikipedia.org/wiki/Przewlek%C5%82a_bia%C5%82aczka_szpikowa

Koncepcja wykorzystania sieci bayesowskich w module wspomagania decyzji medycznych

M. STRAWA

W artykule przedstawiono koncepcję budowy modułu wspomagania decyzji medycznych, współpracującego z repozytorium ścieżek klinicznych. Składają się na nią: definicja sieci bayesowskich oraz najważniejszych pojęć z nimi związanych, opis wybranego mechanizmu wnioskowania oraz przykład generowania diagnozy w module.

Słowa kluczowe: sieci bayesowskie, sieci przekonań, system wspomagania decyzji medycznych.