MEDINFO 2013
C.U. Lehmann et al. (Eds.)

doi:10.3233/978-1-61499-289-9-461

A System for Automated General Medical Diagnosis using Bayesian Networks

Adam Zagorecki[a], Piotr Orzechowski[a], Katarzyna Hołownia[a]

[a] *Infermedica,Wrocław, Poland*

Abstract

In this paper we present a computer-assisted diagnostic system for general medical diagnosis developed using Bayesian network methodology and a medical data base created by experts. The system is intended for the general public as a self-diagnostic tool and is available online free of charge (currently only in Polish, with an English version to be released soon). It serves as an educational self-diagnostic tool intended to encourage the user to visit a doctor if the system so suggests, as is most often the case. In this paper we discuss the underlying modeling principles: assumptions behind Bayesian network architecture, solutions to scalability challenges, and computation performance. The distributed software architecture is presented, and finally, initial results based on over 97,000 diagnoses are discussed. The results suggest that the most common health problems for the young generation in Poland (typical user profile) are those resulting from stress and an unhealthy lifestyle.

Keywords:

Computer-Assisted Diagnosis, Expert Systems, Bayesian networks.

Introduction

The idea of a computer being able to perform a medical diagnosis is not new [1], with the theoretical foundations of the modern Medical Diagnosis Decision Support (MDDS) having been laid as early as the 1950s [2]. MDDS based on probabilistic principles that are used in the system presented here were considered as early as the 1960s [3]. In the 1990s MDDS based on Bayesian networks (Bayesian belief networks) began to dominate [4] the field of MDDS based on probabilistic approaches. Apart from probabilistic methods other methods for producing MDDSs have been tried, including logic and rule-based systems, fuzzy logic, neural networks, pattern matching, various methods that exploit wide bodies of medical texts to generate models, etc. Recently IBM announced that the underlying technology behind the Watson computer, which gained public attention by winning the *Jeopardy* TV quiz show, is intended to be used as a general medical diagnostic tool, putting MDDS into the spotlight and generating interest among the wider public [5].

Diagnosis in the medical context is the process of assigning a label to an illness or other problem by examining observations and symptoms. There are several challenges related to this process. Most important are the uncertainties related to observations and symptoms: rarely can the presence or absence of a single observation or symptom lead to a diagnosis, especially in the initial stages of diagnosis. The second aspect is that observations and symptoms are linked to multiple illnesses, and often the presence or absence of a symptom does not directly indicate or exclude a given illness. In order to develop a reliable model for medical diagnosis, one must account for these two aspects. Other aspects that should be accounted for are the prevalence of the diseases and risk factors such as age and sex that influence both the structure of dependencies between symptoms and illnesses and the related likelihoods.

In this paper we present a medical diagnostic system for self-diagnosis that is intended to inform and educate users about likely disorders or diseases they may suffer from and to encourage them to visit a doctor. The system in its current form is intended for adult users only.

Our system is intended as a general medical diagnostic tool, with the assumption that the user is suffering from a new, undiagnosed condition and has not yet been examined by a doctor, so that no medical test results are available at the time of the diagnosis. The assumptions behind the system are intended to capture a typical situation in which a patient contacts his/her primary care physician (general practitioner) about an undiagnosed health concern. We assume that only basic information is available prior to the diagnostic process, including age, sex, body mass, and four aspects of the patient's medical history: high blood pressure, increased cholesterol, diabetes, and a history of cancer. In practice there are no technical obstacles to including medical test results in the system's medical database and the reasoning algorithm, but it was not our intention to do so. The above assumptions affect the system's ability to diagnose certain conditions, and should be taken into account when interpreting the system's performance.

Based on a literature review, we realized that the way the information is entered into the system is one of the key challenges for MDDS. Because our system is intended for the general public, with practically no means of providing any training on its usage, it was especially important to ensure that information would be entered into the system in an intuitive and time-efficient manner. A second, related aspect is the system's response time – since the diagnostic process typically includes 10 to 20 unique questions, we assumed that a diagnostic step (which in practice involves a complete diagnostic reasoning) should take less than one second to provide a satisfactory user experience. To achieve that, we had to split the domain model into several Bayesian networks that included overlapping variables.

The system has been made available in Polish through a WWW site (www.doktor-medi.pl). While an English version (www.symptomate.com) of the system is in an advanced stage of development, the results presented in this paper pertain to the Polish version of the system. At the time of the paper's submission, the system had performed over 97,000 unique diagnoses.

The rest of the paper is composed as follows: we briefly introduce Bayesian networks and the concepts necessary to understand our approach. Then we discuss the modeling approach taken, including building of the medical data base and the reasoning engine, and outline the diagnostic process. Next, the

software architecture is briefly presented. Initial results obtained from the first period of the system's use – over 2 months – are presented and discussed. We conclude the paper with discussion and future directions.

Methods

Bayesian Networks

The Bayesian network (BN) [6] is a very popular modeling tool, especially for domains involving uncertainty [3]. BNs have been applied to a wide range of domains, including credit risk modeling, user modeling, disease outbreak detection, and many more [7]. The most successful applications of BN are related to hardware diagnosis [8]; in this context a large number of observations (on-board monitoring systems, human observations and checks, etc.) are used to identify faulty components, which are defined in terms of *line replacement units* – the smallest replaceable parts of the engineering system. Such systems based on BNs have been proven in commercial settings, such as diagnostic systems for Hewlett-Packard printers [9]. Other examples include locomotives and aircraft [8].

A BN is a directed acyclic graph in which nodes represent random variables and arcs indicate direct probabilistic dependencies. One or more probability distributions are associated with each node in a BN model. Single probability distributions (prior probabilities) are associated with nodes that have no incoming arcs (no parents) in the graphical part, while the remaining nodes have multiple conditional probability distributions associated with them, typically stored in conditional probability tables (CPTs). A BN is in fact a compact representation of the joint probability distribution (JPD) over the variables included in the model; by exploiting independencies between variables by means of graphical representation it greatly reduces the number of probability distributions required to specify the JPD. BN allows for efficient answering of queries related to arbitrary conditional probabilities involving the variables in the model, such as what the probability of a given disease is assuming a set of symptoms has been observed.

One of the particular weaknesses of BNs is the size of CPTs; in general the number of probabilities required to specify a CPT grows exponentially in the number of parent nodes, making specification of the CPT impractical. To address the problem of large CPTs, a number of approaches have been proposed [10]. One approach is particularly useful for diagnostic applications: the noisy-OR model [11]. The noisy-OR can be viewed as a probabilistic version of the logical OR. Let us assume that an effect node E (symptom) has several possible causes $\mathbf{C} = C_1,\ldots,C_n$ (diseases). The noisy-OR assumes that each of the causes is capable of producing the effect with some probability p_i, but while absent it is unable to produce the effect with certainty. Additionally, the *leak* probability p_l is introduced and corresponds to the probability of the effect being present given that none of the causes is present. Any arbitrary distribution in the CPT can be derived as follows:

$$P(E = true \mid \mathbf{C}) = 1 - \frac{\prod_{i \in \mathbf{C}^+} (1 - p_i)}{1 - p_l} \quad (1)$$

Where $\mathbf{C}^+$ stands for a subset of the set of causes $\mathbf{C}$ that are in the *present* state. The noisy-OR reduces the number of probability distributions required to specify a CPT from exponential to linear in the number of parents. Elicitation of the required probabilities for the noisy-OR is very intuitive: the parameter p_i is the probability of the cause i producing the effect, assuming that all other causes are absent. An additional benefit of the noisy-OR is that it reduces the complexity of the reasoning algorithms [10].

Model and medical data elicitation

In our approach we assumed that the BN model is a bipartite graph: a two layer network with disease nodes as the top layer, the observations (symptoms) placed in the lower layer, and the arcs coming from the disease nodes to observations. All observations are assumed to be noisy-OR models. This type of model is not a new concept and is known in the literature as BN2O [12].

We used several medical experts, mostly practicing physicians whose names are given in the Acknowledgments, to elicit the medical data required for the model. Two of the doctors were awarded a scholarship to contribute to the development of the knowledge base; one of the requirements was that they be active academics with a Ph.D. The remaining doctors were selected according to their specializations and paid for their service. The participating doctors were carefully screened for their ability to understand the task and their commitment to contributing quality input, and provided with the necessary training. They were encouraged to use available medical literature rather than their experience in the process of creating the model. Three distinct types of information are needed to create BN2O:

- Connectivity between diseases and related observations – the experts' role was to identify relevant symptoms for the given diseases, but limited to those suitable for a self-diagnosis by a patient
- Prior probabilities over diseases – such data can be obtained from medical literature and databases, but we sometimes needed to adjust them to more precisely reflect the expected user population (for example dynamics of flu outbreaks)
- Conditional probability of a symptom occurring given the disease

Additionally we had to include some additional constraints: information such as the fact that pregnancy is impossible for men, and risk factors such cancer history, age, etc. This is achieved using the soft evidence approach [12].

Such an approach has particular weaknesses, such as the assumption that diseases are statistically independent of each other (with no observations present), which is obviously incorrect, as some diseases often co-occur. Modeling of risk-factors that influence the prior probabilities of diseases is very limited at this point as well.

To facilitate the knowledge elicitation involving multiple medical experts we developed a specialized elicitation and collaboration software tool. A separate tool was designed to automatically create the BN model from the elicited data.

Network dissection

The version of the medical database in use when the paper was written included over 150 diseases and over 600 symptoms. Because of the network's dense connectivity, which is one of the key performance factors for inference in BNs, inference using exact and approximate algorithms is either intractable or takes too long to provide reliable results (for approximate algorithms). Since the goal we set is to answer an arbitrary query to the system in less than 1 second, we were forced to search for solutions to address this problem. Initially we intended to use a single, densely connected BN, but at some point even the use of relevance reasoning [13] and noisy-OR decompositions [10] would not guarantee the required performance. We decided to split the original model into a set of smaller models, for which cumulative query time would be

within the assumed time frame. Intuitively, it should be possible to use medical specialties such as cardiology, psychiatry, etc. as templates for sectioning the domain into multiple BN models. However, we decided to automate the process of identification of the sub-models, so that the number of sub-models, and consequently performance, could be controlled.

The problem with splitting the model into unconnected sub-models is that it removes information on dependencies between nodes that are in different models. Ideally, we would prefer to have the number of sub-models as small as possible. In order to reduce this effect we did the following:

- Preferred a smaller number of sub-models
- Allowed for repetitive observation nodes across the different sub-models
- Attempted to place disease nodes sharing the same symptoms into the same sub-model. We used a specially developed criterion to measure interdependency between two disease nodes based on the number of common symptoms and the noisy-OR parameters of the shared symptoms.

We developed a special hierarchical aggregation algorithm that serves the purpose of identifying an optimal split of the original model into a subset of models. The algorithm starts with a set of BN models and then iteratively merges some of the models into larger models. This step is done in the *fuse_models* procedure. The procedure takes as an input a set of models V_{in} and outputs a set of models V_{out} with smaller cardinality (number of elements).

At the beginning of the process we initialize V_{in} with n BN models, where n is the number of diseases in the database. Each of the initial BN models consists of a unique single disease node and its children nodes that correspond to all observations for this disease.

The algorithm iteratively calls the procedure *fuse_models* until the resulting models fail to pass the required computational efficiency criteria. Operation *fuse_models* is defined as follows:

The function *distance*(u,v) is basically a similarity measure between two BN models u and v. In this context the most obvious similarity measure is the number of observation nodes shared between the two models u and v. More elaborate measures of similarity are possible as well, for example based on strengths of shared observations, which would require taking into account the noisy-OR parameters. The *combine*(u,v) procedure fuses two BN models into one. Since both models are BN2O networks, the fusion amounts to structurally combining two models, which is a straightforward task, and updating the noisy-OR leak parameters. The algorithm terminates when a performance criterion based on combined diagnosis time is reached.

Diagnostic process

The diagnostic process starts with asking the user several initial questions such as his/her age, sex, weight and height (to determine the BMI index), after which the user is asked to specify his/her primary symptom, and optionally to select the parts of the body that are a source of concern or discomfort.

The collected data is entered as evidence into all sub-models. At this point the iterative procedure starts and a ranked list of suspected diseases is formed. In order to determine the next question to be posed to the user, a score based on the value of information and cross-entropy is used and the most informative observation based on this score is determined. Because the same observation can appear in several sub-models, and in each of these sub-models it can have a different score, we assumed that the highest value of the score should be taken. The question related to the observation with the highest score is presented to the user. The user always has the option to skip the question (when he or she is not sure of the answer), and in such cases the question with the second highest score is presented, and so on. Every answer to a question is added to the evidence set and the process is repeated. The process stops when the probability of the most likely disease is greater than 70% and at least 9 questions have been asked, or a fixed number of steps has been reached. In fact, the stop condition is set empirically to improve user experience: we determined that cases of successful diagnosis (70% reached) are distributed normally, and the 95th percentile of this distribution is used as a threshold to present the user with information that the system is not able to make a diagnosis at that point and that further investigation could potentially lead to an invalid diagnosis.

If the diagnostic process determines that the most likely diagnosis meets the positive stop condition, the user is presented with the ranking of suspected diseases. In most cases the list includes one to three diseases.

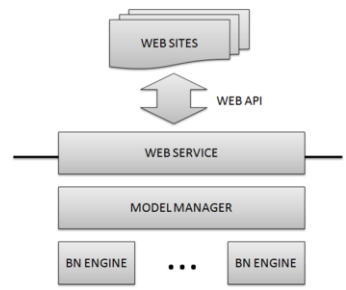

Figure 1 – Software system architecture

System Architecture

The software system responsible for providing the diagnosis is implemented as a web-service. The outline of the design is presented in Figure 1. At the heart of the system lies a distributed, parallel system of multiple *BN engines* that are responsible for performing queries to individual BNs. Each of these engines is based on SMILE general purpose BN software (http://genie.sis.pitt.edu). In order to achieve scalability and reliability, the BN engines are stateless – all necessary information is encoded in query and result data.

The distributed system is designed as a two-layer system with the first layer responsible for distributing incoming queries between n servers on the hardware level, which is implemented as a round-robin. The second layer splits each query between multiple BN engines. This is done by the *Model Manager*, a piece of software that is responsible for distributing the load of queries efficiently and integrating the results. We developed an original solution rather than relying on off-the-shelf solutions. It keeps in memory a number of preloaded BN models in the form of instances of SMILE engines.

The diagnostic functionality is exposed through a web API which is made available to developers. The websites (such as http://doktor-medi.pl) are independent entities and communicate with the diagnostic system through web API calls.

Results

Period of evaluation and collected data

At the time of the paper's submission, the system has been available to the public for over 2 months and over 97,000 diagnoses have been made. These 97,000+ diagnoses included some cases in which the users were simply testing the system, but because we are unable to identify them, they are included in the results. During that period the system has undergone several updates: some of the changes were responses to the initial feedback. These included identifying the upper threshold on the number of questions, beyond which the system was unlikely to produce a reasonably confident diagnosis. Additionally, the system's medical database has been under constant development and regular updates are performed.

Results

A list of the most common diagnoses is presented in Table 1. The percentages correspond to the percentage of all diagnoses presented to the users; as the system can diagnose multiple disorders at the same time, in many cases more than one suspected disorder was presented. The average number of suspected disorders presented to users was 1.66, with the vast majority between 1 and 3. The results shown in Table 1 can be better interpreted by viewing the profile of the users (Figure 2) – most of the diagnoses are made for users aged 25-39 (52.1% female and 47.9% male), which may explain symptoms such as depression, tiredness, tension headaches, or anxiety disorders, all of which are characteristic of the modern lifestyle.

Table 1 – Leading diagnoses for all users

	Diagnosed Problem	%
1	Discopathy	4.06
2	Anxiety disorders	3.90
3	Premenstrual syndrome	3.29
4	Tension headache	3.24
5	Tiredness	2.44
6	Pregnancy	2.40
7	Irritable bowel syndrome	2.35
8	Migraine	2.16
9	Depression	2.16
10	Reflux	2.08

A review of the initial symptoms and locations (the user has the option to select multiple parts of the body subject to pain or discomfort on images of human body) is particularly interesting. The most frequently selected locations are the head (6.8%), genitals (4.4%) and lower abdomen (3.9%). Although we have no hard evidence, these may suggest that the system is often used to self-diagnose problems related to sexual health and other conditions users may feel uncomfortable going to a doctor about. The anus, for example, was indicated in 2.2% of cases, which we find unusually high (frequency similar to that of chest pain, sleepiness, or tiredness).

We further investigated the most common diagnoses that were produced for the age groups 55-70 and 70+. As can be seen in Table 2, these are different than for the user population as a whole (which is dominated by younger users). Only three of the symptoms from Table 1 are included in Table 2: *discopathy*, *acid reflux*, and *tension headache*. The remaining leading diagnoses for patients from the older age groups are consistently age-related problems, such as *osteoarthritis*, *joint or bone trauma*, *ischemic heart disease*, and *gallstones*.

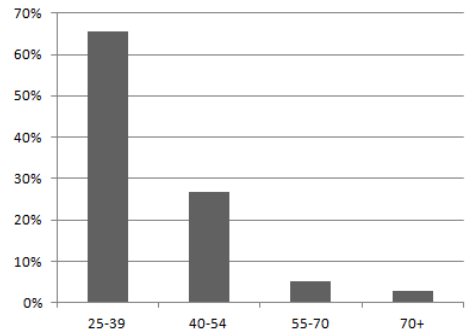

Figure 2 – Age profile of users

Table 2 – Most typical diagnoses for age 55 and above

	Diagnosed Problem	%
1	Discopathy	7.68
2	Osteoarthritis	3.64
3	Joint or bone trauma	2.99
4	Sleep apnea	2.80
5	Ankylosing spondylitis	2.60
6	Acid reflux	2.54
7	Hypertension	2.47
8	Tension headache	2.47
9	Ischemic heart disease	2.41
10	Gallstones	2.41

Discussion

In this paper we presented a general medical self-diagnostic system – its underlying BN model, software implementation, and analysis of initial results based on 97,000+ diagnoses made during the first few weeks after deployment of the system.

It is difficult to evaluate the actual performance of the system because we have no access to users and no way to perform follow-ups. Theoretically, it would be possible to design a study in which the system would first be used to assess the patient before he/she was seen by a primary care physician, and then the system's performance would be validated against an actual doctor's diagnosis. It would be even possible to conduct a follow-up study. We are currently investigating such an option.

An interesting observation can be made regarding the nature of the problems with which users visit the website. We can speculate that the most common diagnoses are those related to stress and an unhealthy lifestyle, especially among younger users. Diagnoses such as *tension headaches*, *tiredness*, and *anxiety disorders* are typically induced by stress. The leading disease, irrespective of the age group, is *discopathy*. This may not be surprising in the older population, but among young people who spend a lot of time in front of a computer (who are likely users of the system) it is known to be a problem as well.

It should be noted that in some cases users try to play with the system without having an actual health problem; in such cases the performance of the system is quite poor, as it tries to investigate several possible diagnoses to fit random answers.

One of the dominating aspects is sexual heath or other problems related to the genitals. While some of these cases were obviously joke entries (often identified by the system), in the majority of the remaining cases there was no strong reason to doubt their validity. Taking into account the seriousness of

some of the initial symptom entries (e.g. *much enlarged testicle*), which were typed in by the user rather than selected from a list of options, we can see an alarming trend of users to avoiding and/or postponing a visit to a doctor even when symptoms are serious and/or advanced. This is even more troubling when the relatively young age of the users is taken into account.

Conclusions

One of the concerns with this type of system is its influence on the doctor-patient relationship. The ease of access to medical information (of various types and quality) via the Internet has resulted in patients trying to diagnose themselves by finding information online, which typically leads to the false impression that they are as knowledgeable about the problem as a doctor. This may apply to the tools such as the one presented here. We claim that the solution described can actually be a better option than one in which users find information online using a search engine and stick with the first plausible option, typically falling into the confirmation bias trap. Our tool is intended to provide an alternative to self-diagnosis based on information extracted from the Internet. Finally, the tool always sends the user to a doctor for a real diagnosis.

The paper reports on the initial results obtained after the system was made available to the general public. The system is constantly under development as we continually improve the medical data base, revising the existing data and adding new diseases and symptoms. An English version of the system is in the final stages of development and should be released very soon. Users have indicated interest in a pediatric version of the system, which we may consider, although it would mean effectively implementing a separate version of the system.

Acknowledgments

The authors would like to thank Dorota Frydecka, Mateusz Palczewski, Anna Rogozińska, Katarzyna Trybucka, and Marcin Zawadzki, who contributed to the creation of the medical knowledge base. The SMILE inference engine developed at the Decision Systems Laboratory, University of Pittsburgh, is used to perform diagnostic inference (http://genie.sis.pitt.edu). The creation of the system would have not been possible without funding provided by the Wrocław Research Center EIT+.

References

[1] Miller RA. Medical diagnostic decision support systems -- past, present, and future: a threaded bibliography and brief commentary. J Am Med Inform Assoc 1994 Jan-Feb; 1(1): 8-27.

[2] Ledley RS and Lusted LB. Reasoning foundations of medical diagnosis; symbolic logic, probability, and value theory aid our understanding of how physicians reason. Science 1959 Jul 3: 130(3366): 9-21.

[3] Warner HR, Toronto AF, Veasey LG, and Stephenson R. A mathematical approach to medical diagnosis. Application to congenital heart disease. JAMA 1961 Jul 22: 177-183.

[4] Shwe M, Middleton B, Heckerman DE, Henrion M, Horvitz EJ, Lehmann HP, and Cooper GF. Probabilistic diagnosis using a reformulation of the INTERNIST-1/QMR knowledge base. Methods Inf Med 1991 Oct: 30(4): 241-55.

[5] Strickland E, and Guy E. Watson goes to med school. IEEE Spectrum 2013 Jan: 50(1): 42-45.

[6] Pearl J. Probabilistic reasoning in intelligent systems: networks of plausible inference. San Francisco, CA: Morgan Kaufmann, 1988.

[7] Olivier P, Naïm P, and Marcot B, editors. Bayesian networks: a practical guide to applications. England: Wiley, 2008.

[8] Przytula KW, and Thompson D. Construction of Bayesian networks for diagnostics. In: IEEE Aerospace Conference Proc.; vol. 5. Piscataway, NJ: IEEE; 2000. pp. 193-200.

[9] Skaanning C, Jensen FV, Kjærulff U. Printer troubleshooting using Bayesian networks. Lecture Notes in Computer Science, vol. 1821. Berlin, Germany: Springer-Verlag; 2000. pp. 367-380.

[10] Zagorecki A, Voortman M, and Druzdzel MJ. Decomposing local probability distributions in Bayesian networks for improved inference and parameter learning. In: Proc. 19th International FLAIRS Conf. Menlo Park, CA: AAAI; 2006. pp. 860-865.

[11] Diez FJ. Parameter adjustment in Bayes networks. The generalized noisy OR-gate. In: Proc. 9th Conf. on Uncertainty in Artificial Intelligence. San Francisco, CA: Morgan Kaufmann; 1993. pp. 99-105.

[12] D'Ambrosio B. Symbolic probabilistic inference in large BN2O networks. In: Proc. 10th Conf. on Uncertainty in Artificial Intelligence. San Francisco, CA: Morgan Kaufmann; 1994. pp. 128-135.

[13] Yan L, Druzdzel MJ. Computational advantages of relevance reasoning in Bayesian belief networks. In: Proc. 13th Conf. on Uncertainty in Artificial Intelligence. San Francisco, CA: Morgan Kaufmann; 1997. pp. 342-350.

Address for correspondence

Adam Zagorecki, Infermedica, ul. Plac Solny 14/3, 50-062 Wrocław, Poland, email: adam.zagorecki@infermedica.com.