In *Intelligent Information Systems*, M. Klopotek, M. Michalewicz, S.T. Wierzchon (eds.), pp. 303-313, Advances in Soft Computing Series, Physica-Verlag (A Springer-Verlag Company), Heidelberg, 2000

Extension of the HEPAR II Model to Multiple-Disorder Diagnosis

Agnieszka Oniśko[1], Marek J. Druzdzel[2], and Hanna Wasyluk[3]

[1] Institute of Computer Science, Białystok University of Technology, Ul. Wiejska 45-A, 15–351 Białystok, Poland, aonisko@ii.pb.bialystok.pl
[2] Decision Systems Laboratory, School of Information Sciences, Intelligent Systems Program, and Center for Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA 15260, USA, marek@sis.pitt.edu
[3] The Medical Center of Postgraduate Education, and Institute of Biocybernetics and Biomedical Engineering, Polish Academy of Sciences, Marymoncka 99, 01–813 Warsaw, Poland, hwasyluk@cmkp.edu.pl

Abstract. The HEPAR II system is based on a Bayesian network model of a subset of the domain of hepatology in which the structure of the network is elicited from an expert diagnostician and the parameters are learned from a database of medical cases. The model follows the assumption made in the database that each patient case is diagnosed with a single disorder, i.e., disorders are mutually exclusive.

In this paper, we describe an extension of the HEPAR II system to multiple-disorder diagnosis. We show that our network transforms readily to a network that can perform multiple-disorder diagnosis with some benefits to the quality of numerical parameters learned from the database. We demonstrate empirically that the diagnostic performance in terms of single-disorder diagnosis improves under this transformation. The new model is more realistic and we expect that it will be of higher value in clinical practice.

1 Introduction

Decision analysis has had a major influence on computer-based diagnostic systems. The field of Uncertainty in Artificial Intelligence, through which this influence was funneled, has developed practical modeling tools based on probabilistic graphical models, such as Bayesian networks [9] (also called belief networks or causal networks) and influence diagrams [4] (also called relevance diagrams or decision networks). Bayesian networks are directed acyclic graphs modeling probabilistic dependencies among variables. The graphical part of a Bayesian network reflects the structure of a problem, while local interactions among neighboring variables are quantified by conditional probability distributions. One of the main advantages of Bayesian networks over other schemes for reasoning under uncertainty is that they readily combine existing frequency data with expert judgment within the probabilistic framework. Often, for example, hospitals and clinics collect patient data, which over time allow for discovering statistical dependencies and potentially improving the overall quality of diagnosis. When incorporated into a model,

they can provide a valuable enhancement to the subjective knowledge obtained from an expert. Bayesian networks have been employed in practice in a variety of fields, including engineering, science, and medicine (for examples of successful real world applications of Bayesian networks, see March 1995 special issue of the journal *Communications of the ACM*) with some models reaching the size of hundreds of variables.

Bayesian networks can be extremely valuable in medical diagnosis. A major advantage of Bayesian networks, compared to other modeling tools, is that they readily model simultaneous presence of multiple disorders. Many approaches, such as those based on classification methods, assume that in each diagnostic case only one disorder is possible, i.e., various disorders are mutually exclusive. This is often an unnecessarily restrictive assumption. It happens fairly often that a patient suffers from multiple disorders and a single disorder may not account for all observed symptoms. Worse even, a situation can arise that a single disorder offers a better explanation for all observations than any other single disorder, while the true diagnosis consists of, for example, two other disorders appearing simultaneously.

In this paper we focus on multiple-disorder diagnosis in the context of the HEPAR II system [6–8]. Our work on the HEPAR II system is continuation of the HEPAR project [1,10], conducted in the Institute of Biocybernetics and Biomedical Engineering of the Polish Academy of Sciences in collaboration with physicians at the Medical Center of Postgraduate Education in Warsaw. The HEPAR system was designed for gathering and processing of clinical data on patients with liver disorders and aimed at reducing the need for hepatic biopsy by modern computer-based diagnostic tools. An integral part of the HEPAR system is its database, created in 1990 and thoroughly maintained since then at the Gastroentorogical Clinic of the Institute of Food and Feeding in Warsaw. The current database contains over 800 patient records and its size is steadily growing. Each hepatological case is described by over 200 different medical findings, such as patient self-reported data, results of physical examination, laboratory tests, and finally a histopathologically verified diagnosis. One of the assumptions made in the database that was available to us is that every patient case is ultimately diagnosed with only one disorder. This assumption, while imposed on us by the data, is not necessary — in reality a patient can be suffering from multiple disorders at the same time and a diagnostic system should consider this possibility.

We describe an extension of our Bayesian network model that relaxes this assumption. We have modified the structure of the network using expert knowledge and subsequently learned the parameters of the new network from the database. While we had to make some assumptions about the data (please note that the data assumed mutual exclusivity of disorders), we show that the diagnostic performance of the modified model on the single-disorder diagnosis is better than that of the original model. The new model is more realistic and we expect that it will be of higher value in clinical practice.

The remainder of this paper is structured as follows. Section 2 summarizes the single-disorder version of the HEPAR II model. Section 3 describes the structural modifications that we performed on the model in order to be able to perform multiple-disorder diagnosis. Section 4 describes the details of learning of the conditional probability distributions of the enhanced model from the database and compares them in terms of their complexity and reliability. Section 5 compares the single-disorder and the multiple-disorder models in terms of their diagnostic accuracy. Finally, Section 6 discusses general issues related to the performed study and directions for further work.

2 The Single-Disorder Diagnosis Version of the HEPAR II Model

The HEPAR II project aims at applying decision-theoretic techniques to diagnosis of liver disorders. Its main component is a Bayesian network model involving a subset of variables included in the HEPAR database. The version of the database used in our project consists of 570 patient cases described by 119 medical findings and classified into 16 different classes (15 disorder classes and one class that represents the hepatologically normal state). One limitation of the data set that we have been using is that it assumes that all disorders are mutually exclusive, i.e., each diagnosed patient suffers from at most one disorder. This limitation led us to the original single-disorder diagnosis model. We selected from this database 94 variables that we judged to be the most important in diagnosis and built a causal Bayesian network. We elicited the structure of the model, i.e., dependencies among the variables, based on medical literature and conversations with our domain expert, a hepatologist Dr. Hanna Wasyluk (third author) and two American experts, a pathologist, Dr. Daniel Schwartz, and a specialist in infections diseases, Dr. John N. Dowling from the University of Pittsburgh. We estimate that elicitation of the structure took approximately 40 hours with the experts, of which roughly 30 hours were spent with Dr. Wasyluk and roughly 10 hours spent with Drs. Schwartz and Dowling. This includes model refinement sessions, where previously elicited structure was reevaluated in a group setting.

The numerical parameters of the model, i.e., the prior and conditional probability distributions, were extracted from the HEPAR database. Prior probability distributions are simply relative counts of various outcomes for each of the variables in question. Conditional probability distributions are relative counts of various outcomes in those data records that fulfill the conditions described by every combination of the outcomes of the predecessors. While prior probabilities can be learned reasonably accurately from a database of consisting of a few hundred records, conditional probabilities present more of a challenge. In cases where there are several variables directly preceding a variable in question, individual combinations of their values may be very unlikely to the point of being absent from the data file. In such cases,

we made an arbitrary assumption that the distribution is uniform, i.e., the combination is completely uninformative. The restructuring effort described in this paper has as one of its long-term goals addressing this problem.

Given a patient's case, i.e., values of some of the modeled variables, such as symptoms or test results, the model derives the posterior probability distribution over the possible liver disorders. This probability distribution can be directly used in diagnostic decision support. We measured the performance of our model by how well it can predict the disorder in each of the available patient cases. To this effect, we applied the standard leave-one-out approach [5], i.e., using repeatedly all but one record in the database to learn the parameters and then using the remaining record to test the prediction. We were interested in both (1) whether the most probable diagnosis indicated by the model is indeed the correct diagnosis, and (2) whether the set of k most probable diagnoses contains the correct diagnosis for small values of k (we chose a "window" of k=1, 2, 3, and 4). Results were approximately 34%, 47%, 56%, and 67% for k=1, 2, 3, and 4 respectively. In other words, the most likely diagnosis indicated by the model was the correct diagnosis in 34% of the cases. The correct diagnosis was among the four most probable diagnoses as indicated by the model in 67% of the cases. Our experts considered this performance to be in the right ballpark given the inherent difficulty of the problem, small size of the data set, and many missing values. Please note that given 16 states of the disorder node, mean performance based on random guessing would barely exceed 6%. More details on the tests performed can be found in [8].

3 Structural Changes to the HEPAR II Model

We have identified several problems with the HEPAR II model. The first problem is that all disorders in the network were modeled as distinct states of one node. This is equivalent to the assumption of mutual exclusivity of disorders. As we mentioned in the previous section, this structure was implied by the data set available to us that had one final diagnosis for each of the patient cases. This assumption is not very realistic in medicine. Presence of a disorder often weakens a patient's immune system and as a result the patient may develop multiple disorders. Since one of the applications of our model is training novice diagnosticians, we would like to model the interactions between disorders and symptoms correctly.

The second problem is still suboptimal diagnostic performance of the HEPAR II network. We believe that the diagnostic performance of the model can be further improved by improving both the structure and the quality of the numerical parameters. Our long-term plans are to use parametric probability distributions, such as Noisy-OR gates [2,3,9], to enhance the quality of the conditional probability distributions learned from data. In order to be able to apply parametric probability distributions, we had to restructure the

network in such a way that various nodes express either propositions or various grades of intensity of some quantity. The disorder node in the HEPAR II model is a categorical variable with 16 outcomes that is not suitable for a parametric probability distribution. One way of preparing the structure for these distributions is by breaking the disorder node into separate nodes for each of the disorders. This modification takes care of two problems: it relaxes the assumption of mutual exclusivity of disorders and it makes the nodes more amenable to parametric quantification.

We have concentrated the structural changes on the disorders. In our initial approach, we reduced the number of disorders modeled from 15 to 9. The six disorders excluded were either represented by very few records in the data base (*Acute hepatitis*, *HBV*, *Alcoholic cirrhosis*) or were later stages of other disorders (*Fibrosis hepatis*, *Carcinoma*). We plan to add the excluded disorders to our model in the future. The 9 modeled disorders were five binary nodes (*Toxic hepatitis*, *Reactive hepatitis*, *Steatosis*, *Hyperbilirubinemia*, *PBC*) and two nodes with three outcomes each (*Chronic hepatits*, *Cirrhosis*). The nodes that we originally modeled as causes/effects of the liver disorder variable were broken down into several groups, specific to each of the 9 disorders. In order to be able to compare the performance of the single-disorder to the multiple-disorder versions of the model, we created a single-disorder version of the original HEPAR II model consisting of the same 9 disorders and precisely the same feature variables as the newly developed multiple-disorder model. As a result, we worked with 66 features and 505 records (65 records of the 570 available to us belonged to the omitted disorder classes) in the database. The resulting models consisted of 67 nodes (66 feature nodes and one disorder node in the single-disorder model) and 73 nodes (66 feature nodes and 7 disorder nodes in the multiple-disorder model) respectively.

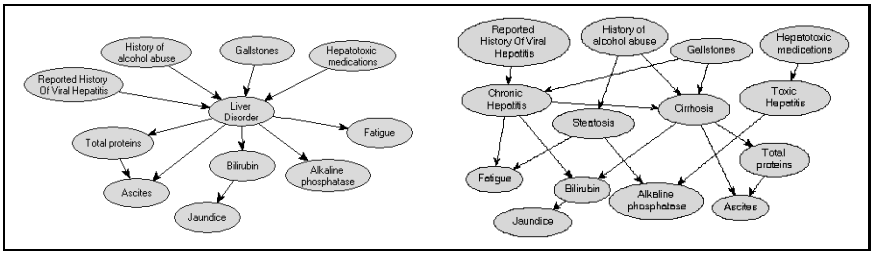

Fig. 1. A simplified fragment of the HEPAR II network: single-disorder diagnosis (left) and multiple-disorder diagnosis (right) version

Fig. 1 shows a simplified fragment of both models and gives an idea of the structural changes performed in the transition from the single-disorder to the multiple-disorder versions of the model. In particular, the models share each of the four risk factors (*Reported history of viral hepatitis*, *History of alcohol*

abuse, *Gallstones*, and *Hepatotoxic medications*) and six symptoms and test results (*Fatigue*, *Jaundice*, *Bilirubin*, *Alkaline phosphatase*, *Ascites* and *Total proteins*). The single *Liver disorder* node is replaced by four disorder nodes (*Chronic hepatitis*, *Steatosis*, *Cirrhosis* and *Toxic hepatitis*). The main difference between the models is that some of the four new disorder nodes are not connected with the risk factors and symptoms. This leads to a significant reduction in the number of numerical parameters necessary to quantify the network.

4 Parameters of the HEPAR II Model

In order to compare the single-disorder to the multiple-disorder versions of the model, we used the same data to extract the numerical parameters (i.e., still each patient was described by only one disorder). The data set contained 505 patient records classified in 9 different disorder classes. A side-effect of our structural changes is that they have decreased the number of numerical parameters in the model. We have mentioned in Section 2 that it is quite common in learning the conditional probability distributions from data that there are too few records corresponding to a given combination of parents of a node. Breaking the original disorder node into several nodes representing individual disorders decreases the size of conditional probability tables and, hence, increases the average number of records for each combination of parents in a conditional probability distribution table. Indeed, the multiple-disorder version of the model required only 1,488 parameters (we counted $\mu = 87.8$ data records per conditional probability distribution) compared to the 3,714 ($\mu = 16.8$ data records per conditional probability distribution) parameters needed for the single-disorder version of the model. With an increase in the average number of records per conditional probability distribution, the quality of the model parameters improves.

Fig. 2 shows the distribution over the number of data records per parent combination for the single-disorder and the multiple-disorder models. We can see that over 50% of the conditional probability distributions in the single-disorder model contained zero records. In the multiple-disorder model this number is dramatically smaller — only 0.5% of all cases involved zero records and there is quite a high proportion of conditional probability distributions for which tens of records were available.

The fact that we used a data set in which each patient record had a single-disorder diagnosis placed us before a difficulty in assessing conditional probabilities of nodes that had several disorder nodes as parents — there were no records in the database for conditions involving combinations of various disorders. We applied a simple solution, in which we included in the calculation all records that described the disorders present in the condition. For example (see Fig. 1), when computing the conditional probability distribution of node *Fatigue* given presence of both *Chronic hepatitis* and *Steatosis*,

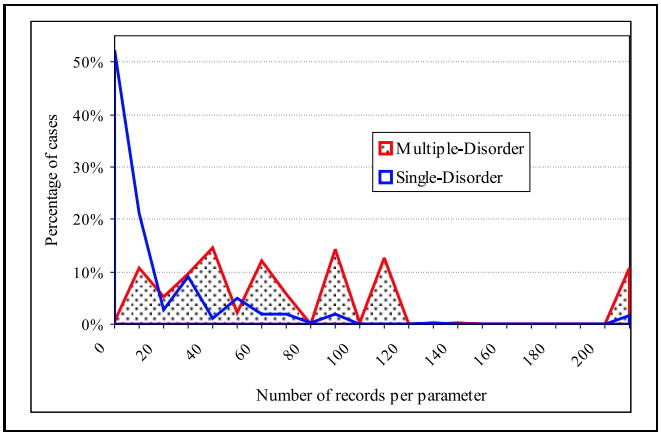

Fig. 2. Distribution over the number of data records per parent combination for the single-disorder and the multiple-disorder models

we used both: records that were diagnosed as *Chronic hepatitis* and records that were diagnosed as *Steatosis*. This amounted to averaging the effect of various disorders. We also tried taking the maximum effect of all disorders present in the condition with a very modest improvement in performance. Another limitation of the HEPAR data that has a serious implication on our work is that mutual exclusivity of disorders did not allow us to extract dependencies among disorders. Hepatology often deals with disorders that are consequences of the previous disorders, e.g., a chronic liver disorder implies *Fibrosis hepatis* which can further cause *Cirrhosis*. In the future we plan to model and quantify these dependencies by combining data with expert judgment.

5 Diagnostic Accuracy of the Multiple-Disorder Model

Our first empirical test focused on the overall performance of the model in terms of classification accuracy (each of the disorders was viewed as a separate class that the program predicted based on the values of all the other variables). This test is very conservative towards the multiple-disorder model, as this is the task for which the single-disorder version of the model was designed. We applied again the leave-one-out approach. Essentially, given n=505 data records, we used $n-1$ of them for learning model parameters and the remaining one record to test the model. This procedure was repeated n times, each time with a different data record.

One of the assumptions that we used in learning the model parameters was that missing values for discrete finding variables corresponded to state *absent* (e.g., a missing value for *Jaundice* was interpreted as *absent*). In case of continuous variables, a missing value corresponded to a normal value, elicited

from the expert (e.g., a missing value for *Bilirubin* was interpreted as being in the range of 0–5), which included the typical value for a healthy patient. In our tests, we used as observations only those findings that have actually been reported in the data (i.e., we did not use the values that were missing, even though we used their assumed values in learning). Similarly to the tests performed on the original HEPAR II model, we used window sizes of k=1, 2, 3, and 4. Results (pictured graphically in Fig. 3) were for the multiple-disorder version of the model approximately 44% (compared to 42% for the single-disorder version), 59% (57%), 68% (68%), and 77% (78%) for k=1, 2, 3, and 4 respectively. In other words, the most likely diagnosis indicated by the model was the correct diagnosis in 44% of the cases. The correct diagnosis was among the four most probable diagnoses as indicated by the model in 77% of the cases. The performance of both versions of the model was very similar, with the multiple-disorder version being slightly more accurate. Please note that the diagnostic accuracy of the single-disorder model in our test is significantly higher than the accuracy reported for the original Hepar II model. This is an effect of the fact that the new version of the model had fewer disorders and most disorders considered were well represented in the database.

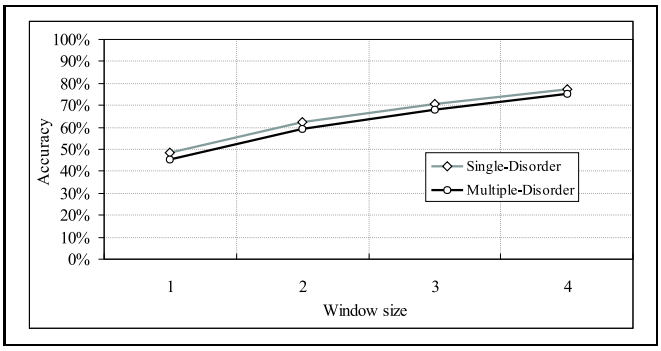

Fig. 3. Diagnostic accuracy of the single-disorder and the multiple-disorder models

Performance for each of the 9 disorders individually is pictured graphically in Fig. 4. We can see that in case of some of the disorders, the multiple-disorder version of the model performed significantly better than the single-disorder version. In order to gain some insight into when multiple-disorder version of the model is better, we focused our second test on the relationship between the number of records in the database for each class and the diagnostic accuracy within that class.

Fig. 5 shows the relationship between the number of records for a particular disorder and the system accuracy in diagnosing this disorder for windows of size 1 (i.e., the most likely disorder) and 4 (the true diagnosis is among the four most likely diagnoses). It is clear that accuracy increases signifi-

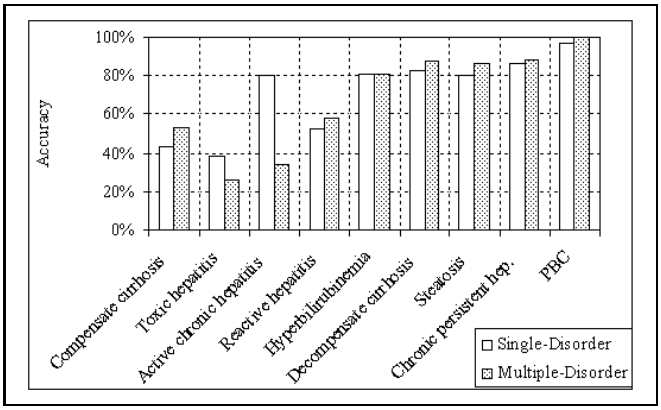

Fig. 4. Diagnostic accuracy per disorder of the single-disorder and the multiple-disorder models

cantly with the number of data records. Disorders with more than 50 records present in the database showed quite high diagnostic accuracy. Another interesting result is that the multiple-disorder model performed often better than the single-disorder model for those disorders that have many records. This promises a higher diagnostic value of our approach when the available data set is sufficiently large, i.e., when the quality of parameters is high.

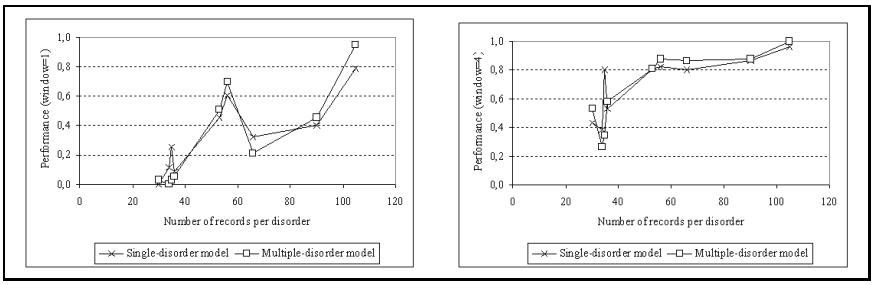

Fig. 5. Diagnostic accuracy as a function of the number of disorder cases in the database (class size) of the single-disorder and the multiple-disorder models for the one-disorder and four-disorder window cases

6 Discussion

The exercise that we went through shows that Bayesian network models readily accommodate multiple-disorder diagnoses. It was relatively easy to derive the multiple-disorder version of the model from the existing single-disorder version. We estimate that the total time spent with the expert was

less than 10 hours. Of course, some of the reduction in time, compared to the original model, can be explained by our increased modeling proficiency.

While the performance of the multiple-disorder diagnosis version of the model is only slightly better than the single-disorder diagnosis version, we should keep in mind that the two models were compared on a task for which the latter is specialized. Furthermore, the data that we learned our parameters from were single-disorder data.

We believe that pure diagnostic performance, in terms of the percentage of correct diagnoses, is in itself not an adequate measure of quality of a medical decision support system. In the domain of medicine, the physician user carries the ultimate responsibility for the patient and he or she will be unwilling to accept a system's advice without understanding it. The effort described in this paper is a further step towards making our model mimic the causal structure of the domain. While a causal model may perform worse in numerical terms than a regression-based model (if it does at all; this remains an empirical question), it offers three important advantages: (1) its intuitive and meaningful graphical structure can be examined by the user, (2) the system can automatically generate explanations of its advice that will follow the model structure and will be reasonably understandable, and (3) the model can be enhanced with expert opinion; interactions absent from the database can be added based on knowledge of local causal interactions with the existing parts and can be parameterized by expert judgment.

Our future work includes expert verification of the probability distributions of those nodes that have several disorder nodes as parents. As we mentioned above, these parameters cannot be learned from our data and the arbitrary assumptions that we made in the learning process may have had a negative effect on diagnostic performance of the system. At a later stage, we plan to replace most of the interactions by parametric probability distributions, such as Noisy-OR gates. We expect that this will increase the model performance even further. We also plan to elaborate on the disorder-to-disorder dependencies. This information is lacking from the database, so here again we will have to rely on expert judgment.

Acknowledgments

This research was supported by the Air Force Office of Scientific Research, grants F49620–97–1–0225 and F49620–00–1–0112, by the National Science Foundation under Faculty Early Career Development (CAREER) Program, grant IRI–9624629, by the Polish Committee for Scientific Research, grant 8T11E02917, by the Medical Centre of Postgraduate Education of Poland grant 501–2–1–02–14/99, and by the Institute of Biocybernetics and Biomedical Engineering Polish Academy of Sciences, grant 16/ST/99. The HEPAR II model was created and tested using SMILE, an inference engine, and GeNIe, a development environment for reasoning in graphical probabilistic mod-

els, both developed at the Decision Systems Laboratory and available at `http://www2.sis.pitt.edu/~genie`.

References

1. Leon Bobrowski. HEPAR: Computer system for diagnosis support and data analysis. Prace IBIB 31, Institute of Biocybernetics and Biomedical Engineering, Polish Academy of Sciences, Warsaw, Poland, 1992.
2. F. Javier Diez. Parameter adjustment in Bayes networks. the generalized Noisy-OR gate. In *Proceedings of the Ninth Annual Conference on Uncertainty in Artificial Intelligence (UAI–93)*, pages 99–105, Washington, D.C., 1993.
3. Max Henrion. Some practical issues in constructing belief networks. In L.N. Kanal, T.S. Levitt, and J.F. Lemmer, editors, *Uncertainty in Artificial Intelligence 3*, pages 161–173. Elsevier Science Publishers B.V., North Holland, 1989.
4. Ronald A. Howard and James E. Matheson. Influence diagrams. In Ronald A. Howard and James E. Matheson, editors, *The Principles and Applications of Decision Analysis*, pages 719–762. Strategic Decisions Group, Menlo Park, CA, 1984.
5. A.W. Moore and M.S. Lee. Efficient algorithms for minimizing cross validation error. In *Proceedings of the 11th International Conference on Machine Learning*, San Francisco, 1994. Morgan Kaufmann.
6. Agnieszka Oniśko, Marek J. Druzdzel, and Hanna Wasyluk. Application of Bayesian belief networks to diagnosis of liver disorders. In *Proceedings of the Third Conference on Neural Networks and Their Applications*, pages 730–736, Kule, Poland, October 1997.
7. Agnieszka Oniśko, Marek J. Druzdzel, and Hanna Wasyluk. A probabilistic causal model for diagnosis of liver disorders. In *Proceedings of the Seventh International Symposium on Intelligent Information Systems (IIS–98)*, pages 379–387, Malbork, Poland, June 15–19 1998.
8. Agnieszka Oniśko, Marek J. Druzdzel, and Hanna Wasyluk. A Bayesian network model for diagnosis of liver disorders. In *Proceedings of the Eleventh Conference on Biocybernetics and Biomedical Engineering*, volume 2, pages 842–846, Warszawa, Poland, December 2–4 1999.
9. Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers, Inc., San Mateo, CA, 1988.
10. Hanna Wasyluk. The four year's experience with HEPAR-computer assisted diagnostic program. In *Proceedings of the Eighth World Congress on Medical Informatics (MEDINFO–95)*, pages 1033–1034, Vancouver, BC, July 23–27 1995.