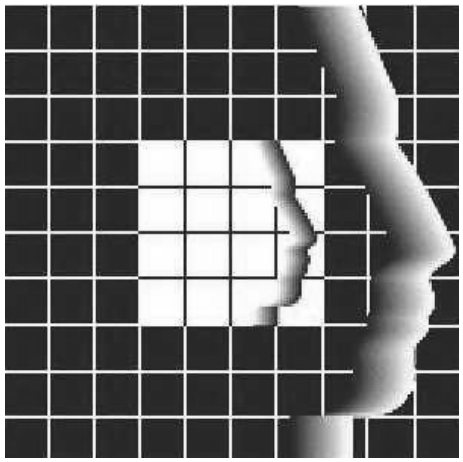
Bayesian Models in Medicine

Working notes of the workshop held during
The *European Conference on Artificial Intelligence in Medicine, AIME'01*
Cascais, Portugal, 1st July, 2001



Peter Lucas*, *Linda C. van der Gaag*, *Ameen Abu-Hanna
(Editors)

Preface

These are the working notes of the workshop on Bayesian Models in Medicine, which was held during the *European Conference on Artificial Intelligence in Medicine, AIME'01*, on 1st July, 2001, in Cascais, Portugal. The workshop brought together various theoretical and practical approaches to using Bayesian models in tackling biomedical and health-care problems.

Bayesian networks with their associated methods have now been around in medicine for more than a decade. They have become increasingly popular for representing and handling uncertain knowledge in medicine, for example to assist in the diagnosis of disorders, or to predict the natural course of a disorder or the most likely outcome after treatment. Almost simultaneously, the use of *Bayesian statistics* has increased in popularity in medicine, for example to study spatial distributions of disease. The advantage of Bayesian methods offered here is that knowledge of a background population can be taken as a starting point of a study. Currently, interest is also emerging within the field of *bioinformatics* to use Bayesian methods for building models of various kinds, for example based on the analysis of gene and protein data. With the increase in research activities in Bayesian models, it was considered timely to organise a workshop to enable researchers in the field to assess the current state of the art, to identify obstacles to progress and to determine future research directions. Of course, the opportunity to actually meet each other to exchange views and to explore possibilities for collaboration was also considered most valuable.

The contributions included in these workshop notes cover a wide range of topics, from learning to modelling, and from theory to the use of software tools to develop biomedical applications. We hope that the reader will be left with the feeling that developing Bayesian models in medicine as a research subject is very much alive and thriving.

We are grateful to our colleagues who served on the programme committee of the workshop on Bayesian Models in Medicine (A. Abu-Hanna (co-chair) K.-P. Adlassnig, R. Bellazzi, C. Berzuini, G.F. Cooper, R.G. Cowell, F.J. Díez, M.J. Druzdzel, L.C. van der Gaag (co-chair), P. Haddawy, D. Hand, I.S. Kohane, P. Larrañaga, A. Lawson, L. Leibovici, T.Y. Leong, P.J.F. Lucas (co-chair), S. Monti, L. Ohno-Machado, K.G. Olesen, M. Paul, M. Ramoni, A. Riva, P. Sebastiani, G. Tusch, J. Wyatt, B. Zupan). They carefully read and reviewed each submission. Each paper was reviewed by at least two members, and in most cases by three members. Thanks are further due to Andrew Lawson, Marco Ramoni, and Paola Sebastiani for accepting our invitation to give an invited talk at the workshop. Last but not least, the participants of the workshop made the effort we put into organising the workshop worthwhile. To them also we would like to express our gratitude.

The Editors:
Peter Lucas, Department of Computing Science, University of Aberdeen
Linda van der Gaag, Institute of Information and Computing Sciences, Utrecht University
Ameen Abu-Hanna, Department of Medical Informatics, University of Amsterdam

20th June, 2001

Contents

Invited Talks

Bayesian Bioinformatics (Invited Talk)

Marco Ramoni
Harvard Medical School
and Children's Hospital
Boston, MA 02115, USA
E-mail: marco_ramoni@harvard.edu

Paola Sebastiani
Department of Mathematics and Statistics
University of Massachusetts
Amherst, MA 01003, USA
E-mail: sebas@math.umass.edu

Abstract

Bioinformatics, the computational challenger of the genome, offers unparalleled opportunities to machine learning research in general and to Bayesian learning methods in particular. In this talk, we outline some of the opportunities and the challenges and we describe where the effort of "cracking the code of life" can most benefit of a Bayesian approach.

1 Introduction

The recent completion of a first draft of the human genome has brought several surprises to many and changed our views about many aspects of the genetic code. What has not been shattered is the humongous size of the task of decoding the genome by find a meaning, i.e. a function, for each part of it. Since the beginning of the Human Genome Project, the international effort to characterize the genomes of human and selected model organisms through complete mapping and sequencing of their DNA, it was clear that the management and analysis of the vast amount of information gathered insofar would have been impossible without the massive use and development of appropriate computational and analytical techniques. The reward of these efforts is even greater than the task itself: a new understanding of the basis of life and a new array of medicines and cures to make it healthier and longer.

The convergence of computational and biomedical sciences is leading to a radical change on both sides. Computer science is importing in biological sciences a new quantitative awareness, while post-genomic data are prompting the development of novel computational methods able to handle their peculiar characters. This talk will describe how Bayesian methods can offer interesting and unique solutions to some of the critical problems involved in the decoding of the semantics of the genome.

2 Functional Genomics

The aim of functional genomics is to understand the function of genes as parts of the entire human genome. Current research is mainly focused on the understanding of gene expression mechanisms, i.e. the processes inducing a particular gene to be transcribed and ultimately to code for a protein. The identification of the genes expressed, say, in a cancer cell line or in a dystrophic muscle, can cast a new light on the genetic basis of a disease and lead to potential remedies. The long term promise is to "reverse engineer" the regulatory mechanisms underlying the genomic controls of an organism and their interaction with external conditions, pathogenic agents, and pharmaceutical products.

2.1 Differential Expression Analysis

Current technology allows for the simultaneous expression analysis of thousands of genes using devices known as *microarrays*. These massively parallel methods to study the simultaneous expression of large number of genes are based on hybridization either to cDNA [16] or to synthetic oligonucleotides [11]. Despite some substantial technical differences, both approaches rely on high-resolution arrays measuring the expression level of each gene as a function of the gene transcript abundance [12]. This abundance is in turn measured by the emission intensity of the region where the gene transcript is located in the scanned image of the microarray, and the signal is filtered to remove noise generated by the microarray background and non-specific expression. Figure 1 shows the picture of an oligonucleotide microarray.

The simplest functional genomic study we can conduct is based on test vs. control experiments identifying which genes are up- or down-regulated by a particular agent or condition. For instance, we can compare the gene expression levels of a cancer cell line line against a healthy cell line and identify which genes are differentially expressed in the two cell lines. Early analyses of these array data identify differentially expressed genes by taking the ratio of the intensities and choosing an arbitrary cut-off factor above (below) which the genes will be taken to be differentially expressed [16; 6]. More sophisticated statistical techniques take into account the distribution of the intensities in the whole microarray. In the first statistical analysis of these data, Chen *et al.* [5] propose a method to identify statistically significant changes between the two samples, under several distributional assumptions, including normality, null intercept, and constant variation coefficient. Bayesian approaches to this problem have been emerging in the past few months. Newton *et al* [13] offer

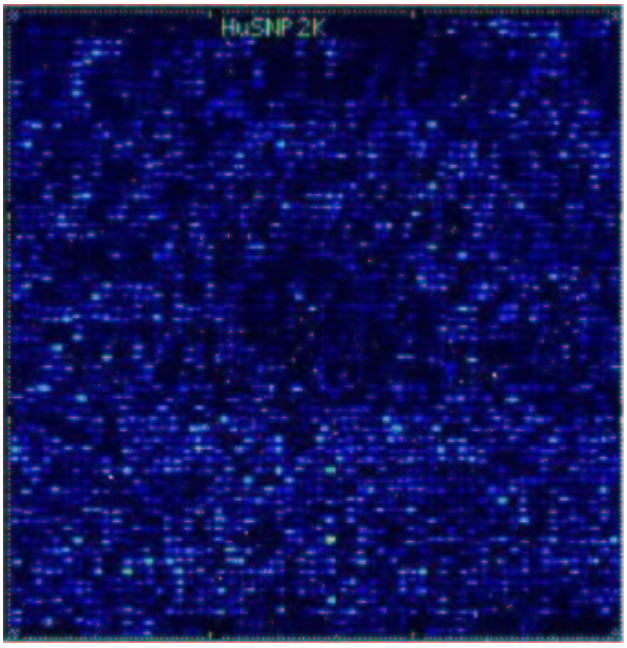


Figure 1: An oligonucleotide microarray.

.

a Bayesian approach to the problem using a hierarchical model and identify differentially expressed genes on the basis of the posterior odds of change. A similar approach has been proposed by Baldi and Long [2] by modeling expression values as independent log-normal distributions, parameterized by corresponding means and variances with hierarchical prior distributions. Probably because of their recent appearance, these approaches have not been used yet in main-stream biomedical research and they remain in the realm of the interesting computational techniques.

2.2 Functional Clustering

A more ambitious approach to functional genomics tackles to challenge of portraying a functional picture of the genome of an entire organism. The chief tool of this quest has been correlation-based hierarchical clustering [7; 17]. Given a set of expression values measured for a set of genes under different experimental conditions, this approach recursively clusters pairs of genes according to the correlation of their measurements under the same experimental conditions. The intuition behind this approach is that correlated genes are acting together since they belong to the same functional categories.

The critical point of this approach is that it always ends up generating generates a single tree and leaves the burden of identifying the actual functional clusters by relying on the available domain knowledge. We have developed a Bayesian clustering method able to identify the set of most probable processes responsible for sequences of observations [15]. The idea underpinning this Bayesian approach is that the observed data are generated by processes. The aim of the algorithm is to find the set of processes most likely, a posteriori, to have generated the sequences of ob-

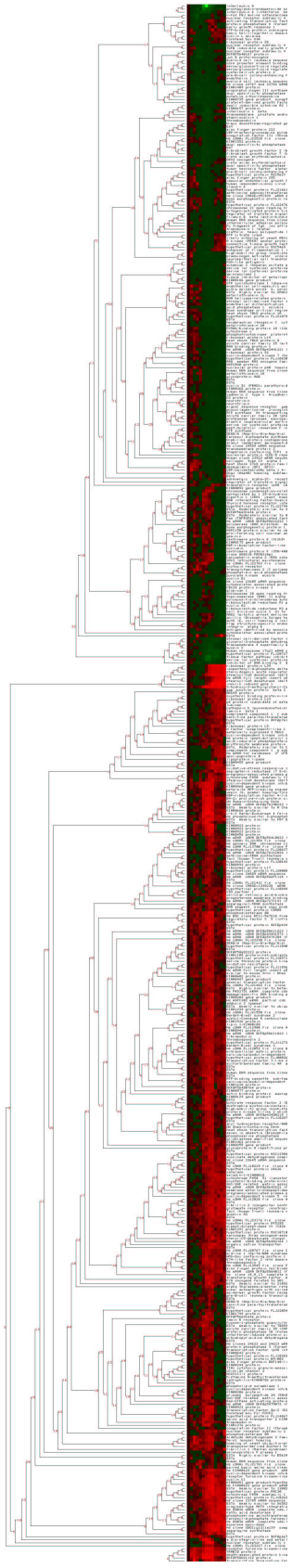


Figure 2: Bayesian clustering of gene expression data.

.

servations in the database.

We have applied this method to cluster observations on 517 genes in a study of the response of human fibroblasts to serum. The data were collected using competitive cDNA microarrays. These microarrays measure the expression level of a gene simultaneously in a basal or control condition and in an experimental condition. The overall expression induced by the experimental condition is measured as the ratio of the two intensity levels, and these are the data used as input by clustering algorithms. Figure 2 shows the clustering obtained on these data by our Bayesian clustering method. The tree on the left represents the steps of the clustering algorithm and reports the four clusters identified by the algorithm. The squares in the center of the picture represents the gene expression measurements. Each row displays the expression levels of a gene in each experimental condition, represented by the columns. Green cells and red cells represents higher than one and lower than one expression levels, respectively. The intensity of the color represents the distance between the measurement and one.

2.3 The Circuitry of the Cell

BBNs are not new to genetic research. As a matter of facts, networks based on directed acyclic graphs actually originated from the genetics studies by Sewall Wright [18], who developed a method called *Path Analysis* [19; 20], a recognized ancestor of BBNs. The application of BBNs to functional genomics is, on the other hand, very recent. BBNs hold the promise of answering very interesting questions in functional genomics and, in principle, they seem to be the right technology to take advantage of the massively parallel analysis of whole-genome data to discover how interact, control each other, and align themselves in pathways of activation. BBNs offer an alternative view to the more popular clustering algorithms currently used for the analysis of massively parallel gene expression data [7; 17]. While these algorithms attempt to locate groups of genes that have similar expression patterns over a set of experiments to discover genes that are co-regulated, BBNs dive into the regulatory circuitry of genetic expression to discover the web of dependencies among genes.

The promise of BBNs in functional genomics goes even further, as intensive research efforts have been addressed, during the past decade, to define conditions under which BBNs actually uncover the *causal* model underlying the data [14; 10]. The most ambitious question is therefore: given a set of microarrays data, can we discover a causal model of interaction among different genes? The challenge is the common problem of sound statistical methods when faced with microarray data: a large number of variables with a small number of measurements. In the context of BBNs, this situation results in the inability to discriminate among the set of possible models as the small amount of data is not sufficient to identify a single most probable model.

Friedman *et al.* [8] address these problems using partial models of BBNs and a measure of confidence in a learned model. The strategy they follow is to search a space of under-specified models, each comprising a set of BBNs, and select a class of models rather than a single one. They also adopt a measure of confidence based on bootstrapping to evaluate the reliability of each discovered dependency in the database in order to avoid the risk of ascribing a causal role to a gene when no enough information is actually available to support the claim. Hartemink *et al.* [9] tackle the under-determination problem by turning the unsupervised search of the most probable network structure. They leverage on established biological knowledge to select a small number of networks and then limit their comparisons to these networks only.

We have taken a slightly different approach, adopting the strategy used in differential gene expression analysis and converting the ratio measures generated by cDNA microarrays into discrete variables by thresholding the measures at 2 folds up and 2 folds down, the same used by the authors of the original paper. Figure Figure 3 shows a BBN generated by Bayesware Discoverer[1] from the fibroblasts response to serum dataset used for the functional clustering displayed in Figure 2.

Although the use of BBNs in functional genomics is still in its infancy, the simple comparison of the network in Figure 3 with the clusters in Figure 2 displays the potential of BBNs to improve the results already obtained by clustering methods by to dissecting the inner structure of the regulatory circuitry of life.

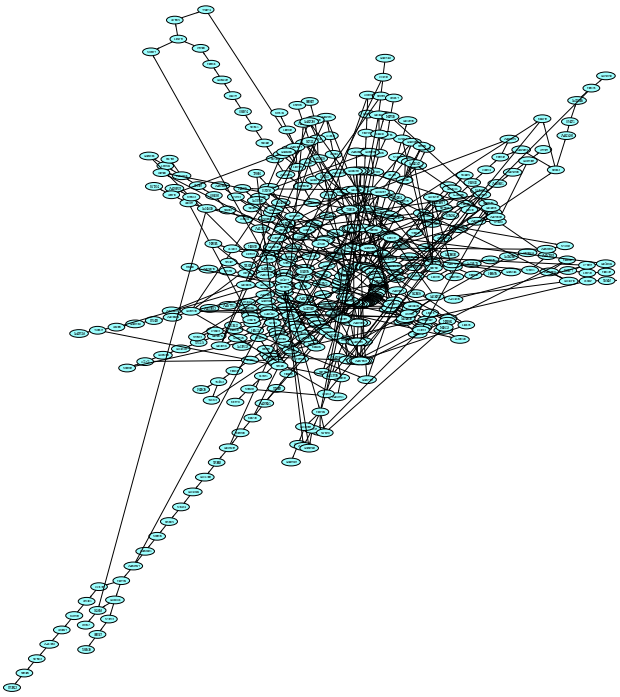


Figure 3: A Bayesian network for functional genomics.

[1] Information about Bayesware Discoverer is available from http://bayesware.com.

3 The New Population Genetics

A further side effect of the completion of a first draft of the human genome has been to infuse new life and provide new perspectives to the more traditional fields of population genetics. One of the surprising statistics emerged from this first draft may hold the key to unlock the code: on average, the genomes of any two human individuals are identical at 99.9% of all nucleotides. While this high degree of identity is striking, the enormous size of the genome (over 3 x 109 base pairs) means that a 0.1% rate of divergence is still equivalent to over 3 million differences between any two people, which translates, on the average, into one difference every 1000 bases. These subtle variations, called *polymorphisms*, have been proven to be invaluable tools to relate genetic code to phenotypes. Albeit small, these variations of the genome have a major impact on how humans respond to disease; environmental insults such as bacteria, viruses, toxins, and chemicals; and drugs and other therapies. But most of all, they have the potential to reveal, in various ways, how the genetic code relates to genetic expression and ultimately to phenotype [3]. Compared to functional genomics, population genetics offers a way to related higher level characters, such as genetic diseases and observable individual features, to their genetic basis.

The most frequent type of polymorphism are Single nucleotide polymorphisms (SNPs). SNPs (pronounced "snips" or "S N Ps") are DNA sequence variations that occur when a single nucleotide (A, T, C, or G) in the genome sequence is altered [3]. For example a SNP might change the DNA sequence AAGGCTAA to ATGGCTAA. Two of every three SNPs involve the replacement of cytosine (C) with thymine (T). Technically, SNPs are single base pair positions in genomic DNA at which different sequence alternatives (alleles) exist in normal individuals in some population(s). The most precious property of SNPs is to be markers on the genetic code [4]. A second desirable property of SNPs is that their frequencies are also evolutionary stable — not changing much from generation to generation — making them easier to follow in population studies.

Original studies on identification of the genetic basis of phenotypes relied on so-called association studies. This kind of studies, typically performed in case-control settings, use measures of correlation between genetic regions and phenotype of interest in order to identify which region is linked to the phenotype. These studies have been shown to be prone to false positives because of spurious associations arising from stratification in the studied population (population admixture). To avoid this situation, pedigree studies based on family members of affected individuals have been introduced and appropriate tests for this setting have been developed, most notably the Transmission Disequilibrium Test (TDT) and its extensions. These tests control for the stratification problem, but they can best be used to test for the transmission of a single genetic region at a time. Therefore, their application to the analysis of complex traits maybe problematic, as this analysis requires the simultaneous test of the transmission of multiple (co-occurring) genetic variations, or for the interacting effect of genetic variations and environmental conditions on phenotypes. Furthermore, as family members are sometime not available for study, pedigree studies are not always an option and case-control association studies must be undertaken. Finally, as both study types are based on traditional frequentist approaches to hypothesis testing, they suffer from the well-known problem of multiple comparisons, whereby the same data are used to test multiple hypotheses, thus confusing the statistical significance of the correlations measured.

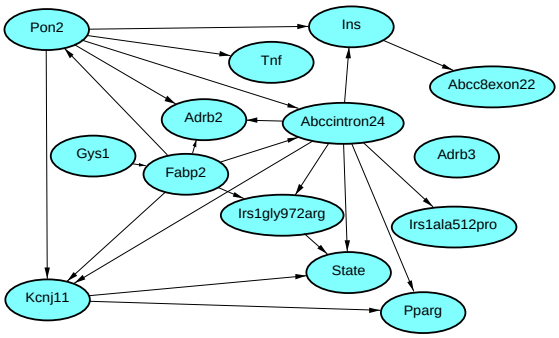


Figure 4: A Bayesian network for SNPs analysis.

BBNs are an ideal tool to tackle the analysis of these data. BBNs are not restricted to pair-wise models of interactions, but they can describe, and therefore help assessing, models where more than one variable is responsible for changes in others. SNPs, environmental conditions and observable characters are represented as stochastic variables, thus allowing for the seamless integration of the relevant information. Patterns of inheritance and interactions among traits, SNPs and environmental conditions, are modeled by means of probabilistic functions, thus accounting for the non-deterministic (stochastic) nature of these interactions. Furthermore, the Bayesian metric used to learn BNs from data avoids the multiple comparisons problem by selecting the dependencies on the basis of their probability rather than the probability of error. The metric used to assess the existence of a dependency between a set of SNPs and phenotypes of interest is based on a competitive measure of probability, which scores multiple models at once and allows for the simultaneous evaluation of multiple correlations among SNPs and phenotypes. In so doing, this technology holds the promise to be a viable exploratory tool to mine the databases produced by the genotyping of the participants to large-scale epidemiological studies collecting dozens of clinical phenotypes. The availability of unsupervised methods like the one proposed here will enable the fast exploration of the potential dependencies between SNPs and phenotypes. Finally, BNs provide a global model of the dependencies among all the genetic regions under consideration, a global picture of transmissions and dependencies that can be used to assess endemic correlations and help to identify potential stratifications.

Figure 4 shows the network generated from a single-sided pedigree study to identify the genetic region respon-

sible for Insulin Dependent Diabetes Mellitus (IDDM)[1]. The study collects data on the 13 SNPs identified so far in literature as associated with IDDM. Our preliminary analysis shows that the pathological status (State) is directly affected by three different SNPs. The high connectivity of the network also helps explain why other SNPs were found to be related to IDDM when measured by conventional tests of pair-wise correlations: if one variable/SNP is removed from the network, a weaker, yet non-negligible, dependency might be manifested between the phenotype and a more remotely linked SNP (for example, the SNP Kcnj11 renders State conditionally independent of the SNPs Fabp2 and Pon2. Removing Kcnj11 would result in State to become dependent on the two ancestors SNPs Fabp2 and Pon2). The model shown was assessed as 54 times more likely than the next best competing model by our Bayesian metric. In other words, the observed data were 54 times more likely to have been generated by this model than by the next best one, and the phenotype of interest was found to be several thousands times more likely to have been generated by the interaction of thse genes than from any single gene alone

4 Conclusions

Since its origins, Artificial Intelligence has always had a special relationship with biomedical sciences: the first Bayes classifier, the first expert system, the first efforts to develop Bayesian networks — just to mention only a few — all emerged from medical applications. The opportunity today is to turn Artificial Intelligence into an integral part of the new biomedical sciences and to join in a single venture two of the major endeavors of the century: the reproduction of intelligence and the understanding of life.

References

[1] D. Altshuler, J.N. Hirschhorn, M. Klannemark, C.M. Lindgren, M-C. Vohl, J. Nemesh, C.R. Lane, S.F. Schaffner, S. Bolk, C. Brewer, T.Tuomi, D. Gaudet, T.J. Hudson, M. Daly, L. Groop, and E.S. Lander. The common PPAR Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes. *Nature Genetics*, 26:76–80, 2000.

[2] P. Baldi and A.D. Long. A Bayesian framework for the analysis of microarray expression data: regularized t -test and statistical inferences of gene changes. *Bioinformatics*, 17(6):509–19, 2001.

[3] A. J. Brookes. The essence of SNPs. *Gene*, 234:177–186, 1999.

[4] A. Chakravarti. Population genetics — making sense out of sequence. *Nature Genetics Suppl.*, 21:56–60, 1999.

[5] Y. Chen, E.R. Dougherty, and M.L. Bittner. Ratio-based decisions and the quantitative analysis of cdna microarray images. *Biomedical Optics*, 2:364–374, 1997.

[6] J. DeRisi, L. Penland, P.O. Brown, M.L. Bittner, P.S. Meltzer, M. Ray, Y. Chen, Y. A. Su, and J.M. Trent. Use of a cdna microarray to analyse gene expression patterns in human cancer. *Nature Genetics*, pages 457–60, 1996.

[7] M. Eisen, P. Spellman, P. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proc. Nat. Acad. Sci. USA*, 95:1486314868, 1998.

[8] N. Friedman, M. Linial, I. Nachman, and D. Pe'er. Using bayesian network to analyze expression data. *Journal of Computational Biology*, 7:601–620, 2000.

[9] A. J. Hartemink, D. K. Gifford, T. S. Jaakkola, and R. A. Young. Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks. In *Proceedings of the Pacific Symposium on Bioinformatics (PBS-01)*, pages 422–433, 2001.

[10] David Heckerman. Bayesian networks for data mining. *Data Mining and Knowledge Discovery*, 1(1):79–119, 1997.

[11] D.J. Lockhart, H. Dong, M.C. Byrne, M.T. Follettie, M.V. Gallo, M.S. Chee, M. Mittmann, C. Wang, M. Kobayashi, H. Horton, and E.L. Brown. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnology*, 14:1675–1680, 1996.

[12] D.J. Lockhart and E.A. Winzeler. Genomics, gene expression and dna arrays. *Nature*, 405:827–836, 2000.

[13] M.N. Newton, C.M. Kendziorski, C.S. Richmond, F.R. Blattner, and K.W. Tsui. On differential variability of expression ratios: Improving statistical inference about gene expression changes from microarray data. *Journal of Computational Biology*, 2001.

[14] J. Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–710, 1995.

[15] M. Ramoni, P. Sebastiani, and P. R. Cohen. Bayesian clustering by dynamics. *Machine Learning*, 2001. To appear.

[16] M. Schena, D. Shalon, R. W. Davis, and P. O. Brown. Quantitative monitoring of gene expression patterns with a complementary dna microarray [see comments]. *Science*, 270(5235):467–70, 1995.

[17] P. Spellman, G. Sherlock, M. Zhang, V. Iyer, K. Anders, Eisen, Brown M., D. P., Botstein, and B. Futcher. Comprehensive identification of cell cycle-regulated genes of the yeast sacccharomyces cerevisiae by microarray hybridization. *Molecular Biology of the Cell*, 9:32733297, 1998.

[18] S. Wright. Correlation and causation. *Journal of Agricultural Research*, 20:557–585, 1921.

[19] S. Wright. The theory of path coefficients: a reply to niles' criticism. *Genetics*, 8:239–255, 1923.

[20] S. Wright. The method of path coefficients. *Annals of Mathematical Statistics*, 5:161–215, 1934.

Bayesian Hierarchical Modelling for Spatial Disease Surveillance

Andrew B. Lawson
University of Aberdeen,
UK

Abstract

The analysis of the geographical distribution of disease incidence is of growing importance within public health (PH). Much attention has been paid to modelling of disease maps within a Bayesian hierachical modelling framework. However little attention has been paid to the analysis of such spatial data in real time. Similarities between spatio-temporal/dynamic analysis within image restoration and disease surveillance are apparent while there is clear relations also with areas of Data mining.

This paper outlines and reviews the main issues found in the area of spatio-temporal disease surveillance and examines a specific application of global surveillance within a Bayesian hierarchical framework.

1 Introduction

Disease mapping is now an important tool for public health surveillance. The analysis of the geographical distribution of disease incidence has important applications in environmental risk assessment, health resource allocation and in epidemiological research. Although spatial epidemiology is a broad field with various foci, a major theme of recent work has been the application of Bayesian hierarchical models to the analysis of the geographical distribution of disease, including correlated and uncorrelated heterogeneities[16],[14],[13].

2 Disease mapping, Data Mining and Surveillance

Surveillance includes a set of statistical paradigms for the analysis of disease incidence either in time, or space or in space-time

These methods should be designed to detect health anomalies in the relevant domain of interest. Often this task has a multiple focus: in PH it may be useful to be able to detect clusters, change points or trends (in space or time) in the incidence. They may also be implemented in *real time* i.e. sequential detection of anomalies could be important (e.g. emerging adverse health indicators around putative pollution sources). Data Mining can be viewed in a similar light: '*Relatively little of statistics is concerned with real time analysis, though data mining problems often require this.*' '*.....in pattern detection, one is seeking to identify small departures from the norm, to detect unusual patterns of behaviour. To many, it is this ..exercise which is the essence of 'data mining'- an attempt to locate nuggets of value amongst the dross*' David Hand (1999) 'Statistics and Data Mining: Intersecting Disciplines' [9].

3 Spatial Models

Here we consider a set of p arbitrary non-overlapping regions within which counts of disease are observed: $\{n_i\}$, $i = 1, ..., p$. In addition, we usually also observe a set of expected rates within the regions of interest: $\{\ e_i\ \}$. The expected rates are included so that spatial variation in risk can be correctly modelled. The usual assumption within Bayesian Hierarchical models for these counts is that, conditional on known and unobserved effects, the counts have independent Poisson distributions with expectation $E(n_i) = e_i.\theta_i$. The θ_i are the relative risk for the disease of interest in the i th region, and is the focus of most models for risk variation. The likelihood is given by

$$l = \sum_{i=1}^{p} \{\ n_i \log e_i.\theta_i - e_i.\theta_i \}.$$

It is often the case that unobserved effects could be thought to exist within the observed data and that these effects should also be included within the analysis. These effects are often termed *random* effects, and their analysis has provided a large literature both in statistical methodology and in epidemiological applications (see e.g. [17],[15],[3],[5]). Within the literature on disease mapping, there has been a considerable growth in recent years in modelling random effects of various kinds, within a hierarchical framework. In the mapping context, a random effect could take a variety of forms. In its simplest form, a random effect

is an extra quantity of variation (or variance component) which is estimable within the map and which can be ascribed a defined probabilistic structure. This component can affect individuals or can be associated with tracts or covariables. For example, individuals vary in susceptibility to disease and hence individuals who become cases could have a random component relating to different susceptibility. This is sometimes known as frailty. Another example is the interpolation of a spatial covariable to the locations of case events or tract centroids. In that case, some error will be included in the interpolation process, and could be included within the resulting analysis of case or count events. Also, the locations of case-events might not be precisely known or subject to some random shift, which may be related to uncertain residential exposure. First, a form of independent and spatially uncorrelated extra variation can be assumed. This is often called *uncorrelated heterogeneity (*UH*)* (see e.g. [2]). Another form of random effect is that which arises from a model where it is thought that the spatial unit (such as regions) is correlated with neighbouring spatial units. This is often termed *correlated heterogeneity (*CH*).* Essentially, this form of extra variation implies that there exists spatial autocorrelation between spatial units. This can usually be modelled at the next level of the hierarchy with prior distributions representing the UH and CH. The model for the log relative risk is often assumed:

$$\log \theta_i = t_i + v_i + u_i$$

where the t_i is a spatial trend component, the v_i is the UH and u_i is the CH. Suitable prior distributions for these components are specified and the posterior distribution of θ_i is constructed. Full posterior inference for Bayesian models has only recently become feasible, largely because of the increased use of McMC methods of posterior sampling. This has been facilitated by the availability of general Gibbs sampling packages such as BUGS (GeoBugs and WinBugs). More recently, Metropolis-Hastings algorithms have been applied in comparison to approximate MAP estimation by Lawson and coworkers [15] and Diggle and coworkers[6], and hybrid Gibbs-Metropolis samplers have been applied to space-time problems by Waller and coworkers[18]. Figure 1 shows the relative risk estimates from a UH-CH model fit for lip cancer incidence in Eastern Germany (1980-1989). On this map the relative risks display some residual variation: a north-south trend in incidence is apparent while a cluster of *low* incidence appears in the south of the map.

4 Temporal Models

The variation of disease incidence in time has been addressed by a range of workers, an a variety of models similar to the spatial case can be defined. For the purposes of surveillance, we wish to identify in

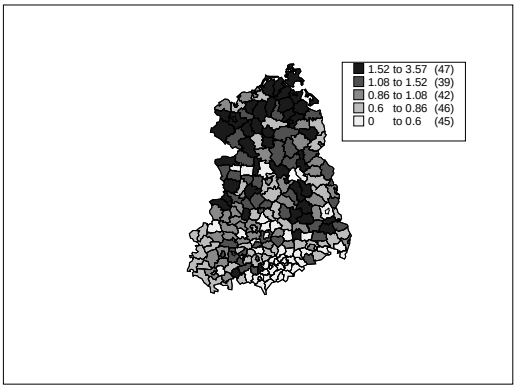


Figure 1:

real time particular forms of process change. In figure 2 the relative risk (log scale) is considered to behave like a gaussian random variable which can be monitored using upper and lower control warning limits (UCL, LCL). In statistical process control (SPC) applications, these warning limits are used to detect changes in mean level (usually). Alternative methods are available for changepoint detection (e.g. cusums). Often in the temporal literature the methods of process control are adopted and the use of control limits are advocated. These have a number of drawbacks within disease surveillance, not least of which is the fact that a time dependent baseline risk is found. This display serves to highlight the features which might be thought to be important in detection of 'adverse' disease risk situations. These can be summarised as *changepoints* (A)(mean level, variance), *clusters (B)* and *overall process change (C).*

Some or all of these features may be of interest in a disease surveillance system. Of course a temporal surveillance system would operate in real time, and so it may be difficult to identify some or all of these features quickly. For example, a cluster of disease (B) could not be identified until sometime after its peak incidence were found.

5 Optimal Surveillance

Likelihood-based surveillance can be developed for the time-domain where an alarm function is constructed from a likelihood ratio evaluated as each time measurement point. With a simple change in parameter an alarm function can be defined as:

$$p(x_s) = \sum_{k=1}^{s} \pi_k \left[\prod_{u=k}^{s} \frac{L(x(u)|\mu^1)}{L(x(u)|\mu^0)} \right] / \sum_{k=1}^{s} \pi_k$$

where:

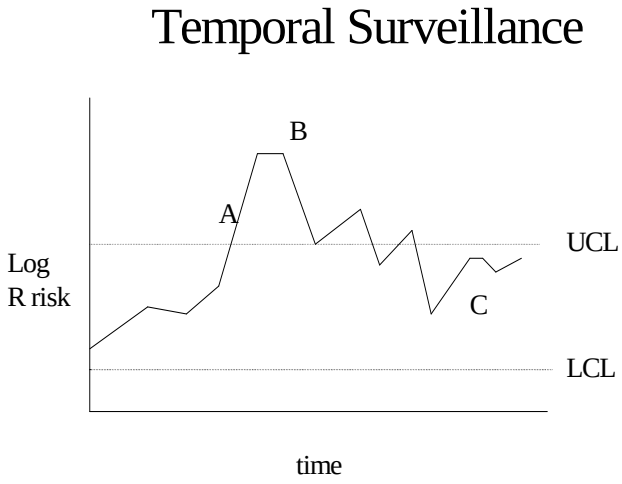


Figure 2:

$p(.)$ is the alarm function
π_k is the prior probability of a change from μ_0 to μ_1 at k
$L(.|\mu)$ is the likelihood given the parameter μ
$x(u)$ is the data at time u
s is a chosen time point

This approach allows the analysis of important measures of the potential delay in alarm based on e.g. conditional expected delay, the expected delay, and the probability of alarm at a given time. This can be extended to multivariate forms and for Bayesian models, but is *global* in nature in that it examines simple changes in a global parameter or parameters[7],[8],[12],[19],[10]. The Bayesian hierarchical model extension of the LR method above defines there to be a posterior distribution for any parameter of interest, and a sample from this distribution must be monitored. Clearly simple changes in parameter can be assessed but must be summarised to represent the mean or variance or other functional of the posterior sample. Note that for complex hierarchical models which require computationally intensive sampling algorithms (MCMC or RjMCMC) then it may be useful to examine resampling methods (particle filters) (see e.g. [11],[1]).

6 Spatio-Temporal Models

The classic spatial surveillance situation would be where spatio-temporal disease incidence observations are available, possibly within fixed time periods. This form of data consists of sequences of maps of disease, and analysis may proceed as in the case of imaging, where a general model for the spatio-temporal process can be conceived and thence and monitoring procedure implemented. There are clearly many complications with surveillance in space-time, not least of which is the wealth of potential interactions between spatial structure and temporal structure.

A Possible Simple Model (fixed time period and spatial frame) assumes we want to examine a change within a fixed time and spatial unit:

$$n_{it} \sim Poisson(e_{it}.\theta_{it}) \tag{1}$$

where n_{it} is the observed number of cases in the i th region at time t

e_{it} is the expected number of cases in the i th region at time t

θ_{it} is the relative risk in the i th region at time t.

A simple model can be assumed for the relative risk:

$$\begin{aligned} \ln \theta_{it} &= \lambda_{it} = \varphi_t + \phi_i + \nu_i + \gamma_{it}, \\ \varphi_t|\varphi_{t-1} &\sim N(\rho\varphi_{t-1}, K_1.\sigma_t^2) && (2) \\ \phi|\phi_{-i} &\sim N(\overline{\phi}_{\delta i}, K_2.\frac{\sigma_\phi^2}{m_i}) && (3) \\ \nu_i &\sim N(0, K_3.\sigma_\nu^2) && (4) \end{aligned}$$

where m_i is the number of spatial neighbours of the i th region, and

$$\gamma_{it} \sim N(0, K_4.\sigma_{st}^2) \tag{5}$$

The idea is that we can monitor a variety of the changes to the process by examining changes in K_1, K_2, K_3. If the process is in control then $K_1 = K_2 = K_3 = 1$. If $K_1 > 1$ then a sharp jump in the risk occurs in time, $K_2 > 1$ is a change in the global spatial correlation structure, if $K_3 > 1$ then a change in global variability occurs, while $K_4 > 1$ is a global change in the risk inplying changes at specific space-time locations.

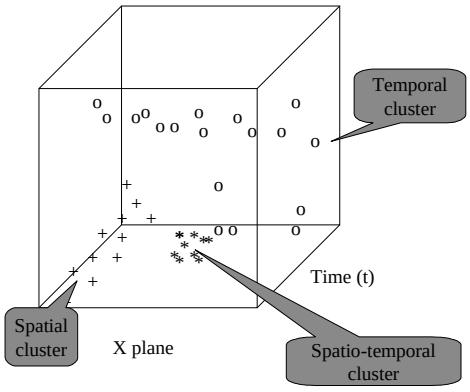


Figure 3:

Another simple model might examine the monitoring of clustering (object recognition) via a cluster model.

Assume that we have a spatio-temporal cluster model of the form:

$$E(n_{it}) = e_{it}.\{1 + \alpha_1 \sum_{i=1}^{ns} K(\mathbf{x}_i - \mathbf{y}_{1i}; \kappa_{1i}) + \alpha_2 \sum_{i=1}^{nt} K(t - y_{2i}; \kappa_{2i}) + \alpha_3 \sum_{i=1}^{nst} K((\mathbf{x}_i, t) - \mathbf{y}_{3i}; \kappa_{3i})$$

where y_{1i} are the spatial cluster centres, y_{2i} are the temporal cluster centres, y_{3i} are the spatio-temporal cluster centres, and $(\mathbf{x}_i, t)$ is the spatial and temporal location of the region.. The 'weights' α_1, α_2 and α_3 represent the increase in relative risk invoked by the cluster centres, and the $\{\kappa_{*i}\}$are cluster variances which are allowed to vary with the cluster label. A similar dynamic model formulation can be specified for this situation.

Figure 3 displays the form of clusters found in space-time:

7 Local versus Global Surveillance

Optimal surveillance methods and most monitoring procedures focus on global parameter surveillance, which is the simplest form of surveillance possible. However in spatial problems there is a need for other forms of surveillance related to localised changes in the structure of the problem. For example, a disease can spread spatially within a time period but only within a small number of regions, and it may be important to report such a spread quickly, especially if it were a highly contagious disease (e.g. FMD). A global model for a map usually doesn't contain measures of localised behaviour and what is needed is a form of locational surveillance which examines localise increases in risk. We call this distinction *Locational* versus *Parameter surveillance*. It is in the area of locational surveillance that most work is needed. The obvious application of this form of monitoring is the assessment of clusters of disease and aggregated areas of adverse risk.

8 Example

In this report a simple example of a spatial disease map analysis will be presented where there is a multi-focus on global parameter changes and cluster assessment. A special surveillance model will be considered for this application.

The surveillance model chosen was that described in 1, 2, 3, 4, and 5 above. We examine the Ohio lung cancer data set described by [4] (amongst others). This data set consists of lung cancer counts for 88 counties of Ohio for 21 year time periods (1968-1988). The surveillance of the 21 years of the map sequence will be examined. The model was fitted using Metropolis Hastings steps with a moving window to reduce the computational time (see e.g. [1]). A fuller discussion of the modelling issues will be presented, and the multiple time series of $K_1, K_2,\ K_3,\ K_4$ will be discussed.

References

[1] C. Berzuini, N. Best, W. Gilks, and C. Larissa. Dynamic conditional independence models and markov chain monte carlo methods. *Journal of the American Statistical association*, 92:1403–1412, 1997.

[2] J. Besag, J. York, and A. Mollié. Bayesian image restoration with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, 43:1–59, 1991.

[3] N. Breslow and D.G. Clayton. Approximate inference in generalised linear mixed models. *Journal of the American Statistical Association*, 88:9–25, 1993.

[4] B. P. Carlin and T. A. Louis. *Bayes and Empirical Bayes Methods for Data Analysis*. Chapman and Hall, London, 1996.

[5] D. G. Clayton. A Monte Carlo method for Bayesian inference in frailty models. *Biometrics*, 47:467–485, 1991.

[6] P. Diggle, J. Tawn, and R. Moyeed. Model-based Geostatistics. *Journal of the Royal Statistical Society C*, 47:299–350, 1998.

[7] M. Frisen. Evaluations of methods for statistical surveillance. *Statistics in Medicine*, 11:1489–1502, 1992.

[8] M. Frisen and J. De Mare. Optimal surveillance. *Biometrika*, 78:271–280, 1991.

[9] D. J. Hand. Statistics and data mining: Intersecting disciplines. *ACM SIGKDD*, 1:16–19, 1999.

[10] E. Jarpe. Surveillance of spatial patterns. Technical Report 3, Department of Statistics, Goteborg University, Sweden, 1998.

[11] A. Kong, J. Lai, and W. Wong. Sequential imputations and bayesian missing data problems. *Journal of the American Statistical Association*, 89:278–288, 1994.

[12] T. L. Lai. Sequential changepoint detection in quality control and dynamical systems. *Journal of the Royal Statistical Society*, 57:613–658, 1995.

[13] A. B. Lawson. *Statistical Methods in Spatial Epidemiology*. Wiley, New York, 2001.

[14] A. B. Lawson, A. Biggeri, D. Boehning, E. Lesaffre, J.-F. Viel, A. Clark, P. Schlattmann, and F. Divino. Disease mapping models: An empirical evaluation. *Statistics in Medicine*, 19:2217–2242, 2000. special issue: Disease Mapping with emphasis on evaluation of methods.

[15] A. B. Lawson, A. Biggeri, and C. Lagazio. Modelling heterogeneity in discrete spatial data models via MAP and MCMC methods. In A. Forcina, G. Marchetti, R. Hatzinger, and G.Galmacci, editors, *Proceedings of the 11th International Workshop on Statistical Modelling*, pages 240–250. Graphos,Citta di Castello, 1996.

[16] A. B. Lawson, D. Böhning, A. Biggeri, E. Lesaffre, and J.-F. Viel. Disease mapping and its uses. In A. B. Lawson, D. B'ohning, E. Lesaffre, A. Biggeri, J.-F. Viel, and R. Bertollini, editors, *Disease Mapping and Risk Assessment for Public Health*, chapter 1. Wiley, 1999.

[17] R. Marshall. Mapping disease and mortality rates using empirical Bayes estimators. *Applied Statistics*, 40:283–294, 1991.

[18] L. Waller, B. Carlin, H. Xia, and A. Gelfand. Hierarchical spatio-temporal mapping of disease rates. *Journal of the American Statistical Association*, 92:607–617, 1997.

[19] P. Wessman. Some principles for surveillance adopted for multivariate processes with a common change point. *Communications in Statitsics, Theory and Methods*, 27, 1998.

Papers

Representing Health-Care Systems Using Bayesian Networks. An Experimental Comparison of Learning Algorithms

Silvia Acid and **Luis M. de Campos**
Dpto. de Ciencias de la Computación e I.A.
Universidad de Granada, 18071-Granada, Spain
Susana Rodríguez and **José María Rodríguez** and **José Luis Salcedo**
Hospital Universitario Virgen de las Nieves
Granada, Spain

1 Introduction

Health-care systems are complex and depend on organizational, economical and structural factors. The availability of appropriate tools for their representation would allow to study and understand the interactions among the different elements that determine their behaviour, as well as to analyze some alternatives to improve their performance. As many of the factors that influence on the performance of a health-care system are of a uncertain nature, Bayesian networks could play an important role in their study. In this paper we introduce some representation models, based on Bayesian networks, applied to the specific case of an emergency medical service. These models have been obtained, from real data recorded at the hospital "Virgen de las Nieves", using algorithms for learning Bayesian networks.

The paper is structured as follows: In Section 2 we describe the problem we are going to study, the available data and the preprocessing steps (discretization, variable selection,...) used to get a database appropriate for the learning algorithms. In Section 3 we briefly comment on the different learning algorithms we have considered for our experiments. Section 4 describes the networks obtained for the different algorithms. In Section 5 we summarize the results of several experiments, which try to assess the quality of the networks from different points of view. Finally, Section 6 discusses the conclusions of this work.

2 The Problem

As we have already commented, we want to model some aspects of the health-care system for patients that arrive to the emergency department of a hospital. Our first aim is simply to better understand the interactions between some of the factors that shape this system, and obtain a model that describes reasonably well the nature of the system. Afterwards, this model could be used to make predictions about some variables of interest, or even to make decisions about the configuration of the system itself. Our approach is management oriented, and tries to assist to the hospital manager in organizational and economical questions (for example, possible redistribution or reinforcement of personnel and/or infrastructure) instead of clinical problems[1].

2.1 The Data Set

From the set of variables which are collected when a patient enters in the emergency department, the variables displayed in Table 1 were initially selected. In this table we also show either the number of possible values or the range for each variable. For the experiments we had at our disposal a database containing 31937 records (corresponding to all the arrivals to the emergency departments of the hospital "Virgen de las Nieves" at Granada, from 01/01/2001 to 02/20/2001).

Variable	Possible values
Financing	11
Date of Admission	date
Time of Admission	0:01-24:00
Cause of Admission	8
Pathology	7
P10	2
Identification	6
Date of Discharge	date
Time of Discharge	0:01-24:00
Cause of Discharge	11
Medical Service	36

Table 1: Variables initially considered.

Financing represents the type of entity that supports the expenses (Social Security, Insurance companies, International agreements, ...). *Cause of Admission* codifies 8 different values (considered as confidential by the hospital staff). *Pathology* includes Common Disease, Common Accident, Industrial Accident, Traffic Accident, Aggression, Self-inflicted Lesion and Other. *P10* represents whether the patient was sent to the emergency medical service by a family doctor. *Identification* codifies the type of identification document of the patient (Identity Card, Social Security Card, Passport, Oral Identification, Other

[1] Although a better use of the available resources would also imply an improvement of the medical care.

and Unidentified). *Cause of Discharge* represents several reasons (Return to duty, Death, Hospitalization, Transfer to another hospital, ...). *Medical Service* includes all the different emergency units at the hospital. All the described variables were used just as they are, but for the remaining four variables in Table 1, some additional treatment was necessary.

2.2 Preprocessing of Data

We have discretized some variables as follows:

- Date of Admission: We discretized it into 7 values, corresponding to the days of the week. From now on we call this variable *Day*.
- Time of Admission: We discretized it into 3 values corresponding to the three different horary periods of the day: morning (8:01-15:00), evening (15:01-22:00) and night (22:01-8:00). From now on we call this variable *Shift*.

We also defined new variables, which were considered relevant:

- *Duration*: The lenght of time (in hours) that the patient stayed in the emergency department. This value is calculated from the values of Date and Time of Admission and Date and Time of Discharge. Moreover, this new variable was discretized into 3 values (from 0 to 8 hours, from 8 to 72 hours, and more than 72 hours) which were considered meaningful by the physicians[2].
- *Centre*: The hospital has three different emergency departments corresponding to the three centres that compose it (Maternity hospital, Orthopedic Surgery and General hospital).

The variables Date and Time of Discharge were considered irrelevant to our purposes, because the truly relevant information is the Duration of the stay. Therefore, these two variables were removed. So, we have considered a total of 11 variables.

3 The Learning Algorithms

We have applied to our problem several algorithms for learning the structure of a Bayesian network. On one hand, we aim to compare their performance on a real problem; on the other hand, the arcs appearing in all the learned networks could be considered as being the "core" for this representation model. Any consensus Bayesian network should be built from this shared structure.

The algorithms that we have used are the following:

- PC [SGS93], an algorithm based on independence tests. This type of algorithms carries out a qualitative study of the dependence and independence relationships among the variables in the domain (using conditional independence tests), and tries to find a network that represents these relationships as far as possible.
- Another algorithm, BN Power Constructor (BNPC), that uses independence tests and cross entropy [CBL98].
- A scoring-based algorithm, that uses local search (LS) in the space of dags (directed acyclic graphs) [HGC95]. The algorithms based on a scoring metric try to find a graph with the minimum number of links that "best" represents the data, according to a specific metric. All of them use a metric in combination with a search method to measure the goodness of each explored structure. Each one of these algorithms is characterized by the specific metric and search procedure used. In our case, the local search used is based on the classical operators of addition, deletion and reversal of arcs (and an initial empty graph); the (Bayesian) metric used is the K2 metric [CH92].
- A version of the BENEDICT[3] (BE) algorithm [AC01b]. This algorithm, which searches in the space of equivalence classes of dags[4], is based on a hybrid methodology [Aci99], that shares with the methods based on scoring metrics the use of heuristic search methods to build a model and then evaluate it using a scoring metric. At the same time, the method has some similarities with those based on independence tests: it explicitly uses the conditional independencies embodied in the topology of the network to elaborate the scoring metric, and carries out conditional independence tests to limit the search effort.

  The basic idea of this algorithm is to measure the discrepancies between the conditional independences represented in any given candidate network G (d-separation statements) and those displayed by the database (probabilistic conditional independences). The lesser these discrepancies are, the better the network fits the data. The aggregation of all these local discrepancies results in a measure of global discrepancy between the network and the database. The local discrepancies are measured using the Kullback-Leibler cross entropy, $Dep(x, y|Z)$, which measures the degree of dependence between two variables x and y, given that we know the values of the subset of variables Z. To evaluate a network G, only the values $Dep(x, y|Z)$ for pairs of non-adjacent variables in G, given a d-separating set of minimum size, Z, of x and y in G [AC96], are calculated. The main search process is greedy and only addition of arcs is permitted, although a final refining process (reinsertion of discarded arcs and pruning of inserted arcs) mitigates the irrevocable character of the whole search method.

[2] They correspond, respectively, to "normal", "complicated" and "anomalous" cases.

[3] Acronym of BElief NEtwork DIscovery using Cut-set Techniques.

[4] Other versions of BENEDICT, that search in the space of dags with a given ordering of the variables, and use a slightly different metric, can be found in [AC01a].

The experiments we are going to describe have been carried out by means of our own implementations for the cases of PC, LS, and BE. For BNPC, we used the software package available at `http://www.cs.ualberta.ca/~jcheng/bnsoft.htm`.

To compute the conditional (or marginal) probability distributions stored at each node in the network, thus obtaining a complete Bayesian network, we used a maximum likelihood estimator (frequency counts).

4 Results

After running the four learning algorithms we obtained four different networks. Due to space limitations, we cannot show all of them. Figure 1 displays the edges in common to all the networks (four arcs and four links (excluding the directionality)).

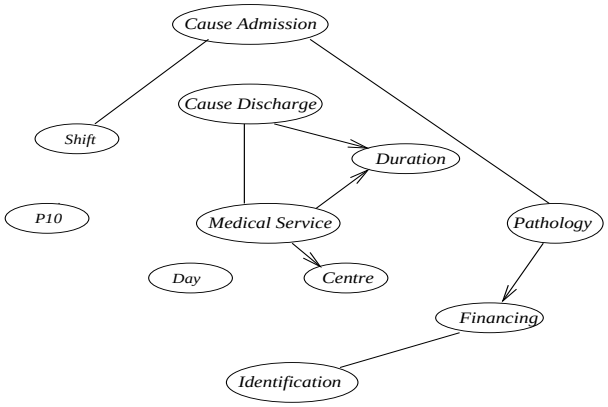


Figure 1: The incomplete structure shared by all the networks.

We do not assume a causal interpretation of the arcs in the networks (although in some cases this could be reasonable). Instead, we interpret the arcs as direct dependence relationships between the linked variables, and the absence of arcs means the existence of independence relationships.

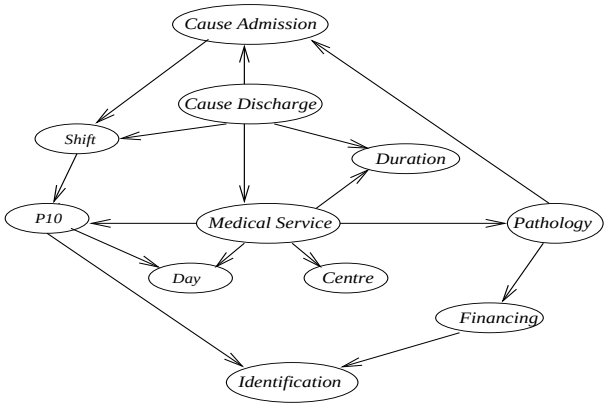


Figure 2: Structure obtained by the BE algorithm.

The strong relation between *Pathology* and *Financing* is explained because the expenses are charged to different entities depending on the type of pathology (traffic accident, industrial accident,...). *Financing* also depends on *Identification* (obviously the expenses will be charged to some entity or company only if the patient can be identified as belonging to this entity). The connection between *Pathology* and *Cause of Admission* seems us quite obvious. The relation between *Cause of Admission* and *Shift* may be due to the fact that the reason to go to the emergency department is not homogeneous across the different hours (Shifts). The arc going from *Medical Service* to *Centre* is justified because Centre is a variable functionally dependent on Medical Service (each Centre has its own emergency medical units). The *Duration* of the stay at the emergency department essentially depends only on the medical unit that tended the patient and the *Cause of Discharge* (the seriousness of the diseases and the degree of congestion of the service, which are strongly related with the duration of the stay, probably vary from one unit to another). In turn, these two variables are highly correlated: For example, a decease as being the cause of discharge is much more unlikely for some medical units than for others.

For illustrative purposes we also show a complete network, corresponding to the BE algorithm (see Figure 2). From this network, we can obtain (using d-separation) a number of conditional independence relationships: For example, *Pathology* and *Cause of Discharge* are independent when we know *Medical Service*; also, *Financing* and *Duration* are conditionally independent given *Medical Service*.

To give an idea of the resemblance between models, Table 2 shows, for each pair of algorithms, the two numbers $l\,/\,a$, where l is the number of common links (in either direction) and a is the number of common arcs between the networks learned by these algorithms[5].

	PC	LS	BE	BNPC
PC	11/11	9/8	9/7	8/4
LS	-	18/18	10/9	9/5
BE	-	-	16/16	10/6
BNPC	-	-	-	13/13

Table 2: Number of common links and arcs, $l\,/\,a$, between pairs of learned networks.

5 Experiments

The information we have collected about the experiments with the different learning algorithms is the following:

[5]The main diagonal in this table represents the number of arcs contained in each network.

- The Kullback-Leibler (KL) distance (cross-entropy) between the probability distribution, P, associated to the database (the empirical frecuency distribution) and the probability distribution associated to the learned network, P_G. In this way we try to assess the performance of the algorithm from the perspective of how closely the probability distribution learned approximates the empirical frequency distribution. Actually, we have calculated a decreasing monotonic linear transformation of the Kullback-Leibler distance, because this one has exponential complexity and the transformation can be computed in a very efficient form [Cam98]. The interpretation of our transformation of the Kullback-Leibler distance is: the higher this value is the better the network fits the data. However, this measure should be handled cautiously, because a high KL value may also indicate overfitting (a network with many edges probably will have a high KL value).

- The values of the K2 metric [CH92] (log version) and the BIC (Bayesian Information Criterion) metric [Swa78] for the learned network. These values give an idea of the quality of the learned network from different points of view.

- The learned networks can also be used with predictive purposes, by using the inference methods (propagation of evidence) available for Bayesian networks. More precisely, from the perspective of a classification problem, we want to use the networks to predict the values of some variable of interest given some evidence, and compare the predictions obtained with the true values of this variable, thus obtaining the corresponding percentages of success. We have considered three different situations:

  (a) Predicting the values of *Duration*, given evidence about the values of all the other variables, except *Cause of Discharge*. In this way, we try to determine the most probable duration of the stay at the emergency department before the patient is effectively discharged.

  (b) Predicting the values of *Medical Service*, given evidence relative to all the remaining variables, except *Pathology*, *Cause of Discharge* and *Duration*, which would be unknown at the arrival time of the patient. If accurate, this prediction could serve to direct the arriving patient to the appropriate emergency unit.

  (c) Predicting the value of each one of the eleven variables, given evidence about all the ten remaining variables. In this way we try to test the behaviour of the network models for different problems. This experiment could serve to assess the robustness of the networks as general classifiers (as opposed to have to manage a different model to classify each variable of interest).

For all the classification problems, we have learned the different networks using a training subset containing 21309 cases and the success percentages have been calculated using a test set containing the remaining 10628 cases.

We have also computed the performance measures corresponding to the empty network ($\emptyset_{\text{em}}$), which is obviously a quite poor model (no interaction among the variables). These values may serve as a kind of scale.

The values of the different metrics for all the networks considered are displayed in Table 3[6].

Algorithm	KL	K2	BIC
BE	2.4473	*-101016*	**-243420**
PC	2.1518	-104834	*-249509*
LS	**2.5180**	**-99896**	-315000
BNPC	*2.4850*	-101308	-258123
$\emptyset_{\text{em}}$	0	-133315	-306937

Table 3: Performance measures for the different networks.

The LS algorithm performs quite well with respect to both the KL and K2 metrics, although its BIC value is the worst (including the empty network!). Moreover, LS is the algorithm that obtains the most dense network (18 arcs), as was to be expected taking into account the previous values. On the other hand, PC produces the sparsest network (11 arcs) and gets bad KL and K2 values. The BE algorithm obtains a network quite balanced with respect to all the metrics and an intermediate number of arcs. As a whole, BNPC seems to behave worse than BE and LS and better than PC. Anyway, from the point of view of the KL and K2 metrics, there are not important differences (less than 3%) among the learning algorithms, with the exception of PC[7]. With respect to the BIC metric the differences are greater.

Algorithm	Duration %	M. Service %
BE	**89.8**	**76.0**
PC	**89.8**	**76.0**
LS	**89.8**	**76.0**
BNPC	**89.8**	**75.9**
$\emptyset_{\text{em}}$	85.9	31.0

Table 4: Success percentages of classification for *Duration* and *Medical Service*.

Table 4 displays the percentages of success of the different networks for the first two classification prob-

[6]The best value for each metric is written in bold, the second best value in italic and the worst value (excluding the empty network) in small font.

[7]For the K2 metric these differences become quite large when we consider the natural probability space instead of the log space.

lems considered[8]. In both cases all the learned networks perform equally. With respect to predicting the duration of the stay, note that the percentage of improvement obtained with respect to the prediction of the empty network is rather small, 3.9% (although this value amounts for 415 patients). The reason is that the distribution of the duration of the stay is quite biased towards its first value (from 0 to 8 hours) and therefore the default rule that assigns to all the cases the 'a priori' most probable class gets a high percentage of correct classifications[9]. For the problem of predicting the medical service involved, the results outperform remarkably the prediction of the empty network.

Table 5 displays the percentages of success of the different networks for the other eleven classification problems. In this case LS performs slightly better than BE and BNPC, which in turn are also a bit better than PC. All the learned network significantly outperform the empty network in most cases.

	BE	PC	LS	BNPC	$\emptyset_{em}$
CoA%	**90.0**	**90.0**	**90.1**	**89.9**	**90.0**
CoD%	**76.6**	**76.1**	**76.8**	**76.6**	60.8
Cen%	**100**	**100**	**100**	**100**	39.3
Day%	18.1	17.8	**19.3**	18.1	17.8
Dur%	**90.0**	**89.8**	**90.2**	**90.0**	85.9
Fin%	**98.0**	**98.0**	**98.1**	**98.0**	93.9
Ide%	**86.0**	**86.6**	**86.4**	**86.5**	**86.0**
MS%	**83.3**	**82.8**	**82.9**	**83.0**	31.0
P10%	**94.6**	**94.6**	**94.6**	**94.6**	**94.6**
Pat%	**86.2**	80.7	**85.6**	**86.2**	79.2
Shi%	47.5	47.5	**49.9**	46.6	46.6

Table 5: Success percentages of classification for the eleven variables.

Taking into account the results shown in Tables 3, 4 and 5 for the different learned networks, we can conclude that all the models seem to be very competitive.

6 Concluding Remarks

The complexity of the health-care systems requires appropriate tools for their representation, study and optimization. Bayesian networks constitute a very attractive formalism for representing uncertain knowledge (which is the result of the sinergy of statistical methods for data analysis and Artificial Intelligence tools), that has been successfully applied in different fields. However, Bayesian networks have been used in medicine essentially to assist in the diagnosis of disorders and to predict the natural course of disease after treatment (prognosis). A novelty of this work is the application of Bayesian networks to other, more management oriented, medical problems.

For future works, we plan to extend and refine our model (using consensus networks), including more variables (e.g., seasonal variables), validate it taking into account expert knowledge and use it as a tool to assist to the manager hospital. We also plan to apply Bayesian networks to other management medical problems (as waiting-lists).

[8]In this table and also in Table 5, all the values which do not exhibit significant differences (at a 95% confidence level) with respect to the best value are written in bold font.

[9]Probably, a finer discretization of the variable *Duration* would give rise to much greater differences.

References

[Aci99] S. Acid. *Métodos de aprendizaje de redes de creencia. Aplicación a la clasificación.* PhD thesis, Universidad de Granada. Spain, 1999 (in Spanish).

[AC96] S. Acid, L.M. de Campos. An algorithm for finding minimum d-separating sets in belief networks. in: E. Horvitz, F. Jensen (Eds.), *Proceedings of the Twelfth Conference on Uncertainty in Artificial Intelligence*, Morgan Kaufmann, San Mateo, 1996, pp. 3–10.

[AC01a] S. Acid and L.M. de Campos. A hybrid methodology for learning belief networks: Benedict. To appear in *International Journal of Approximate Reasoning*, 2001.

[AC01b] S. Acid and L.M. de Campos. An algorithm for learning probabilistic belief networks using minimum d-separating sets. Submitted to Journal of Artificial Intelligence Research.

[Cam98] L.M. de Campos. Independency relationships and learning algorithms for singly connected networks. *Journal of Experimental and Theoretical Artificial Intelligence*, 10(4):511–549, 1998.

[CBL98] J. Cheng, D.A. Bell, W. Liu. Learning Bayesian networks from data: An efficient approach based on information theory. Technical Report, University of Alberta, 1998.

[CH92] G.F. Cooper and E. Herskovits. A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9(4):309–348, 1992.

[HGC95] D. Heckerman, D. Geiger, D.M. Chickering. Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning* 20 (1995) 197–243.

[Pea88] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* Morgan Kaufmann, San Mateo, 1988.

[SGS93] P. Spirtes, C. Glymour, R. Scheines. *Causation, Prediction and Search.* Lecture Notes in Statistics 81. Springer Verlag, New York, 1993.

[Swa78] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6:461–464, 1978.

Ovarian cancer classification with rejection by Bayesian Belief Networks

Peter Antal **Geert Fannes** **Frank De Smet** **Joos Vandewalle** **Bart De Moor**
Electrical Eng. Dept. ESAT/SISTA
Katholieke Universiteit Leuven
Kasteelpark Arenberg 10, B-3001 Heverlee (Leuven), Belgium

Abstract

Belief Networks in the Bayesian approach provide a well-established methodology to fuse prior knowledge and statistical observations for an enriched decision support. In this paper we investigate one of the advantages of the Bayesian approach - the provided additional uncertainty information for predictions - in a medical classification problem. We perform a Bayesian analysis using Belief Network models to discriminate between benign and malignant ovarian masses. The performance of such Bayesian Belief Network models are reported when the exclusion of some data points is allowed based on various uncertainty measures of the prediction.

1 Introduction

The Bayesian approach is becoming more attractive for the machine learning community because it can cope with the valuable subjective prior information in a principled way and it provides more detailed information for decision support. These properties are particularly attractive in medical applications, since detailed uncertainty information can be vital in a medical decision and frequently abundant prior domain knowledge is available beside the statistical data. Under certain conditions Belief Networks are especially suitable for Bayesian modeling, that is to formalize the prior domain knowledge, to update it by observations and to perform inference in a Bayesian way [4]. In the paper we investigate a Belief Network model from the Bayesian perspective to discriminate between benign and malignant ovarian masses.

The paper is organized as follows: Section 2 reviews the Bayesian approach in classification problems. Section 3 recapitulates the medical problem which will serve as a test case, introduces the data and defines relevant performance measures. In Section 4 we discuss the applied Belief Network model and the algorithms used to approximate the Bayesian performance measures. Section 5 presents the performance of the model using thresholds based on various uncertainty measures of the prediction to exclude some data points. In Section 6 we summarize our findings about having a detailed Bayesianist prediction in this medical problem.

2 Bayesian Classification

Starting with a prior distribution expressing the initial beliefs concerning the parameter values of the model, we can use the observations to transform this into the posterior distribution for the model parameters expressing the beliefs after observing the data. Using this posterior distribution over the model parameters, useful random variables can be defined for functions depending on the model parameters, like predictions and error measures.

In a binary classification task this rationale means the following. We are primarily interested in the correct classification of an observation $\boldsymbol{x} \in \mathbb{R}^l$. This can be achieved by constructing a *binary decision function* $g(\boldsymbol{x}, \boldsymbol{\omega}) \in \{0, 1\}$ where $\boldsymbol{\omega} \in \Theta$ are the model parameters. A more informative predictive model provides not only a class label, but also the *class probabilities*, though it is a more complex task both from a statistical and computational point of view. As a further step in improving the decision support, *uncertainty information* can be provided for the class probabilities, for example the posterior distribution of class probabilities in the Bayesian framework.

In this paper we follow the Bayesian approach to solve the classification problem for two main reasons: to incorporate prior background information in a general and principled way and to provide detailed information with clear semantics for decision support. For a *probabilistic regression* model $P(T = 1|\boldsymbol{x}, \boldsymbol{\omega}) = f(\boldsymbol{x}, \boldsymbol{\omega}) \in [0, 1]$ it means there is a prior distribution $p_{\boldsymbol{\Omega}}(.)$ over the model parameters $\boldsymbol{\omega} \in \Theta$. $F_{\boldsymbol{\Omega}|\underline{\boldsymbol{d}}}$ denotes the random variable for the predicted posterior class probability (as a scalar in the $[0, 1]$ interval).

We assume the existence of a labeled training set $\underline{\boldsymbol{d}} = \{\boldsymbol{x}_k, t_k\}_{k=1}^n$, $(\boldsymbol{x}_k, t_k) \in \mathbb{R}^l \times \{0, 1\}$, where $\boldsymbol{x}$ is a real valued l-dimensional input vector and t is the corresponding class label. In the paper we use capitals for random variables, bold indicates a vector and a bold underline indicates a matrix.

Using the observed data $\underline{\boldsymbol{d}}$ and applying Bayes' rule, the prior distribution can be transformed to the posterior distribution $p_{\boldsymbol{\Omega}}(\boldsymbol{\omega}|\underline{\boldsymbol{d}})$ given by

$$\frac{p_{\boldsymbol{T}}(t_1,\ldots,t_n|\boldsymbol{\omega},\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n)p_{\boldsymbol{\Omega}}(\boldsymbol{\omega}|\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n)}{p_{\underline{\boldsymbol{D}}}(\underline{\boldsymbol{d}})}$$

that is, by

$$L(\boldsymbol{\omega}|\underline{\boldsymbol{d}})p_{\boldsymbol{\Omega}}(\boldsymbol{\omega})$$

where $L(\boldsymbol{\omega}|\underline{\boldsymbol{d}})$ denotes the probability of the data given the parameters.

Once we have this posterior distribution for the model parameters, we can define random variables related to predictions, performance, etc. In classification problems for example, we are interested, for a given $\boldsymbol{x}$, in the random variable $f(\boldsymbol{x},\boldsymbol{\Omega})$ where $\boldsymbol{\Omega}$ is a random parameter vector. In this way we have uncertainty information about the predicted class probability.

We can simplify this result to scalar values for the class probabilities $P(T=1|\boldsymbol{x},\underline{\boldsymbol{d}})$. The optimal step back depends on the cost function attached to the reported scalar value. Assuming the L_2 loss function, the optimal strategy is to report the expectation of the class probability in the posterior parameter probability space $f(\boldsymbol{x}) := E_{\boldsymbol{\Omega}|\underline{\boldsymbol{d}}}[f(\boldsymbol{x},\boldsymbol{\omega})]$. A further simplification is to discretize this scalar value using some user specified threshold λ, deriving a binary decision function

$$g_\lambda(\boldsymbol{x}) := \begin{cases} 1 \text{ if } E_{\boldsymbol{\Omega}|\underline{\boldsymbol{d}}}[f(\boldsymbol{x},\boldsymbol{\omega})] \geq \lambda \\ 0 \text{ else.} \end{cases}$$

These three distinct levels, the distribution of the class probabilities $(f(\boldsymbol{x},\boldsymbol{\Omega}))$, the class probabilities $(f(\boldsymbol{x}))$ and the class labels $(g_\lambda(\boldsymbol{x}))$ provide diminishing possibilities for decision support, though the burdening statistical and computational complexity should be considered too.

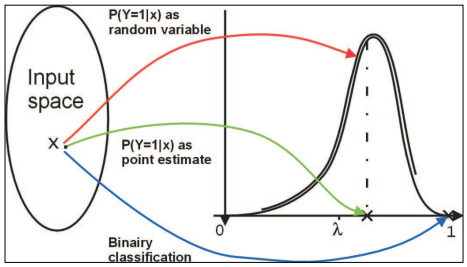


Figure 1: Three levels of predictions.

A more refined scheme allows rejection based on the uncertainty of the prediction of the class probability $F_{\boldsymbol{\Omega}|\underline{\boldsymbol{d}}}$

$$g_{\lambda,\sigma}(\boldsymbol{x}) := \begin{cases} \text{"}rejected\text{" if } \delta[F_{\boldsymbol{\Omega}|\underline{\boldsymbol{d}}}] \geq \sigma \\ 1 \text{ if } E_{\boldsymbol{\Omega}|\underline{\boldsymbol{d}}}[f(\boldsymbol{x},\boldsymbol{\omega})] \geq \lambda \\ 0 \text{ else.} \end{cases}$$

We are using the following uncertainty measures derived by the transformations of the random variable $F_{\boldsymbol{\Omega}|\underline{\boldsymbol{d}}}$ into a scalar δ:

$$\begin{aligned} \delta_{L_1}[F_{\boldsymbol{\Omega}|\underline{\boldsymbol{d}}}] &= min(E_{F_{\boldsymbol{\Omega}|\underline{\boldsymbol{d}}}}[f], 1 - E_{F_{\boldsymbol{\Omega}|\underline{\boldsymbol{d}}}}[f]) \\ \delta_{Var}[F_{\boldsymbol{\Omega}|\underline{\boldsymbol{d}}}] &= Var_{F_{\boldsymbol{\Omega}|\underline{\boldsymbol{d}}}}[f] \\ \delta_{L_1,Var}[F_{\boldsymbol{\Omega}|\underline{\boldsymbol{d}}}] &= \delta_{Var}[F_{\boldsymbol{\Omega}|\underline{\boldsymbol{d}}}] - \delta_{L_1}[F_{\boldsymbol{\Omega}|\underline{\boldsymbol{d}}}] \\ \delta_H[F_{\boldsymbol{\Omega}|\underline{\boldsymbol{d}}}] &= H(F_{\boldsymbol{\Omega}|\underline{\boldsymbol{d}}}) \\ \delta_{Bayes}[F_{\boldsymbol{\Omega}|\underline{\boldsymbol{d}}}] &= min(\int_0^{1/2} dF_{\boldsymbol{\Omega}|\underline{\boldsymbol{d}}}, \int_{1/2}^{1} dF_{\boldsymbol{\Omega}|\underline{\boldsymbol{d}}}) \end{aligned} \quad (1)$$

δ_{L_1} is a non-Bayesian uncertainty measure, the distance of the point-value prediction from a decision threshold. δ_{Var} is the variance of the Bayesian prediction of the class probability. $\delta_{L_1,Var}$ is the distance of the threshold from the credible region $[\delta_{L_1} - 1/2\,\delta_{Var}, \delta_{L_1} - 1/2\,\delta_{Var}]$. δ_H is the entropy of the class probability distribution and δ_{Bayes} is the belief in the less probable class (i.e. the belief in the unavoidable minimal error).

3 Classification of Ovarian Masses

Ovarian malignancies represent the greatest challenge among gynaecologic cancers. A reliable preoperative prediction in terms of benign and malignant ovarian tumors would be of considerable help to clinicians selecting an appropriate treatment. There are two sources of information to construct such predictive models: prior knowledge and data.

The available relevant medical literature and expert knowledge is abundant and very diverse (for an overview, see [5]). In addition to the prior background information, data were collected prospectively from 300 consecutive patients who were referred to a single institution (University Hospitals Leuven, Belgium) from August 1994 until June 1997. The data collection protocol ensure that the patients had an apparent persistent extrauterine pelvic mass and excludes other causes that may have similar symptoms such as infection or pregnancy, so the primary aim is differentiation between benign and malignant masses (for a detailed description, see [5]). Univariate statistics of data set are presented in Table 1. Since the data set is mostly complete with respect to the used model in the paper we used only this subset after certain statistically or medically motivated discretizations (e.g. CA 125="< 35", "$35-65$","$65 \leq$").

Standard statistical studies indicate that a multimodal approach – the combination of various types of variables – is necessary for the discrimination between benign and malignant tumors. Therefore Logistic Regression models, Multilayer Perceptrons and Belief Networks were previously applied [5; 1]. These models predicted the scalar class probabilities and they were developed and tested in the classical statistical framework.

	Age	Parity	CA 125	Color score
$\widehat{E}[.\|Benign]$	47.77	1.50	110.3	1.98
$\widehat{E}[.\|Malignant]$	58.62	1.57	1222.2	3.20
$\widehat{Std}[.\|Benign]$	15.60	1.40	976.5	0.84
$\widehat{Std}[.\|Malignant]$	15.18	1.73	3779.6	0.95

Table 1: Univariate statistics for the benign and malignant subpopulation in the ovarian cancer data set.

A natural step to provide more detailed information for medical decision support is to apply the Bayesian approach to provide the distribution of class probabilities. We can use the classical statistical performance measures, such as Misclassification Rate (MR) and the area under the Receiver Operator Characteristics curve (ROC), for the evaluation of the models in the Bayesian framework, since any performance measure is a function of the model parameters (for fixed observations/test data). These performance measures then become random variables which provide more information than a point estimate. For the definition and interpretations of the ROC curve, see e.g.[6].

4 Bayesian Belief Networks

A Belief Network represents a joint probability distribution over a set of variables (see e.g. [2]). We assume that these are discrete variables, partitioned into three sets $\mathbf{X}$, $Y\ in\ \{c_0,\ c_1\}$, $\mathbf{Z}$: set of input, output, and intermediate variables respectively. The model consists of a qualitative part (a directed graph) and quantitative parts (dependency models). The vertices of the graph represent the variables and the edges define the qualitative dependency-independency relations among the variables. There is a dependency model for every vertex (i.e., for the corresponding variable) to describe its probabilistic dependency on the parents (i.e., on the corresponding variables).

Assuming parameter independence we use Dirichlet distributions as dependency models (see e.g. [4; 3]). In this case the prior background knowledge is formalized as a fixed Belief Network structure and the prior distribution $p_{\mathbf{\Omega}}(.)$ over the model parameters $\boldsymbol{\omega} \in \Theta$, where the hyper parameters of the Dirichlet distributions, N_{ijk}, can be interpreted as the number of previously seen corresponding examples.

Using such Dirichlet distributions, an expert can express his belief in parametrizations and for complete samples the posterior distribution $p_{\mathbf{\Omega}}(\boldsymbol{\omega}|\underline{\boldsymbol{d}})$ has the same analytic formula with updated hyper parameters [4; 3].

We built the Belief Network from the available prior knowledge from expert and literature in a "heterogeneous" way incorporating biological models of the underlying mechanism quantifiable by the literature, parts quantified by a medical expert and parts quantified by previously published studies [1]. The structure of the Belief Network model is shown on Fig. 2.

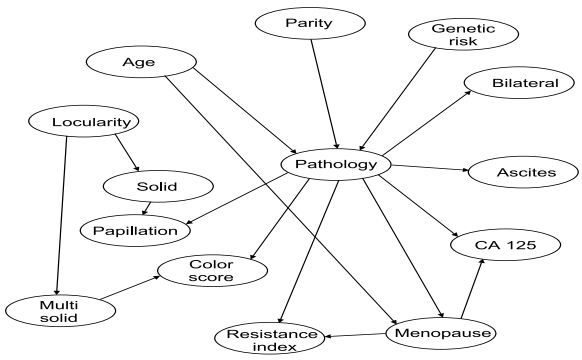


Figure 2: The BN model structures.

The target random variables to be estimated are hierarchical: the inference $P(T = 1|\mathbf{\Omega}, \boldsymbol{x}^{obs}, \boldsymbol{z}, \underline{\boldsymbol{d}})$ and the performance related $MR(\mathbf{\Omega}, \underline{\boldsymbol{d}})$. We sample the posterior distribution $p_{\mathbf{\Omega}}(\boldsymbol{\omega}|\underline{\boldsymbol{d}})$ by direct sampling from the updated Dirichlet distributions and compute the conditional probabilities of malignancy for the drawn parametrizations by an exact inference algorithm using a join tree (see [2]). Based on these predictions the corresponding MR and AUC values can be computed.

5 Results

We investigated the advantages of having a more detailed probabilistic prediction in the Bayesian framework. At first we manually evaluated the Bayesian predictions of the Belief Network model from a medical point of view. We noticed that the predictions for misclassified cases are more uncertain, e.g. they have higher variances $Var_{\mathbf{\Omega}|\underline{\boldsymbol{d}}}[f(\boldsymbol{x}, \boldsymbol{\omega})]$ which is one measure for the 'uncertainty'. Generally spoken, the cases with a high value for $Var_{\mathbf{\Omega}|\underline{\boldsymbol{d}}}[f(\boldsymbol{x}, \boldsymbol{\omega})]$ were also hard to classify by a medical professional, in contrast with cases with a low value for $Var_{\mathbf{\Omega}|\underline{\boldsymbol{d}}}[f(\boldsymbol{x}, \boldsymbol{\omega})]$, that were almost always straightforward to predict.

To identify automatically these medically hard cases, we tried to quantify the uncertainty of the prediction by the δ-measures introduced in section 2. Fig. 3 and 4 show the correlation between the δ_{Var}, δ_E and δ_H measures. Correct classified samples are denoted with "*" and incorrect ones with a "o".

One promising possibility of having a quantification for the uncertainty of the prediction is to allow the rejection of the most uncertain cases, which in practice can mean referring such a patient to an expert or further examinations. To investigate the efficiency of the identification of hard cases, we computed the area under the ROC curve and the misclassification rate when various proportions of the most uncertain cases are rejected. Fig. 5 shows the misclassification

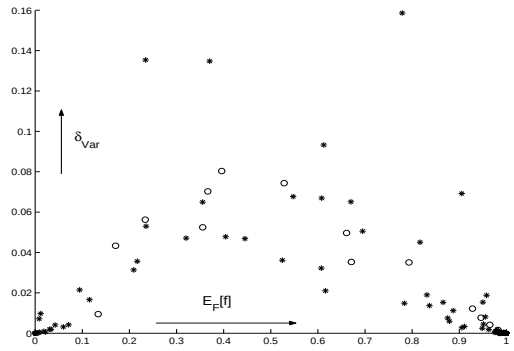


Figure 3: Correlation between $E_{F_{\Omega|\underline{d}}}[f]$ and δ_{Var}. Correct classified samples are denoted with "*" and incorrect ones with "o".

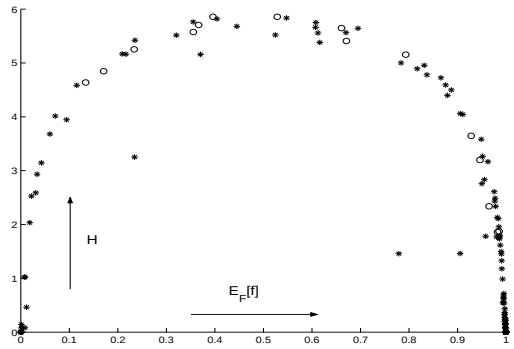


Figure 4: Correlation between $E_{F_{\Omega|\underline{d}}}[f]$ and δ_H. Correct classified samples are denoted with "*" and incorrect ones with "o".

rates after excluding various proportions of the data set based on δ-measures (δ_{L_1}, δ_{Var} and δ_H) as defined in Section 2, Fig. 6 shows the same for the rejected partition. In these experiments, we partitioned the data set described in Section 3 randomly to a test (50%) and training (50%) set, this was repeated 30 times to eliminate dependency on separation. The reported results are based on the test set.

Tables 2 and 3 show the misclassification rates that are achieved for 'non-rejected' respectively 'rejected' samples for varying uncertainty measures defined by Eq. 1

6 Discussion

Since the Bayesian approach is becoming more and more popular as an efficient inductive method for integrating prior knowledge and statistical data, the question arises how we can use other potential advantages of this framework. One attractive candidate is the detailed Bayesian prediction of class probabilities, since

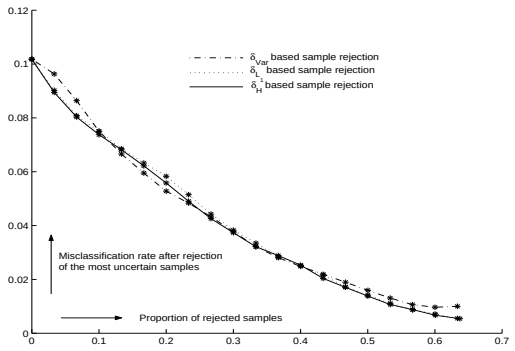


Figure 5: The misclassification rate on the test set after rejecting varying proportions.

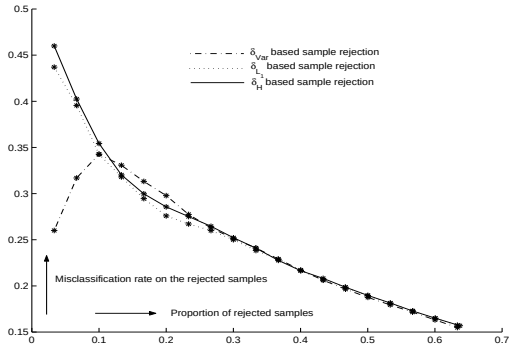


Figure 6: The misclassification rate on the rejected data points for varying proportions.

it may allow automatic identification of the uncertain cases for special treatment. To test this idea, we experimented with representing the uncertainty of a prediction by a scalar and investigated the classification performance when cases with uncertainty above a given threshold are 'rejected' as defined in Section 2.

The first manual evaluations by a medical doctor show that the derived uncertainty measures from the Bayesian prediction are in close correspondence with the subjective uncertainty of a human decision maker, though the quantitative evaluation needs further experiments. To evaluate the efficiency of the rejection methods based on these measures we investigated their effect on the classification performance. As Table 2 shows without rejection the misclassification rate is 10.2% while in the rejected sets it can be above 40% for small rejected sets and in the most interesting region it is still between $20 - 30\%$. For example, if we set our rejection threshold to exclude 20% of the cases, the misclassification rate drops to 5%. In practice, this means that a decision support system can be specified to classify 80% of the cases with a low misclassifica-

% Reject.	δ_{L_1}	δ_{Var}	$\delta_{L_1,Var}$	δ_H	δ_{Bayes}
0	10.1	10.1	10.1	10.1	10.1
6.66	8.08	8.64	8.15	8.03	8.09
13.3	6.85	6.65	6.81	6.81	6.73
20	5.82	5.27	5.67	5.58	5.30
26.6	4.42	4.32	4.38	4.25	*
33.3	3.34	3.21	3.25	3.21	*
40	2.51	2.48	2.46	2.52	*
46.6	1.71	1.89	1.82	1.70	*
53.3	1.11	1.30	1.25	1.06	*
60	0.716	0.966	0.700	0.667	*

Table 2: The misclassification rate on the test set after rejecting varying proportions.

% Reject.	δ_{L_1}	δ_{Var}	$\delta_{L_1,Var}$	δ_H	δ_{Bayes}
0	*	*	*	*	*
6.66	39.5	31.7	38.5	40.2	39.2
13.3	31.8	33.0	32.0	32.0	32.4
20	27.5	29.7	28.2	28.5	29.1
26.6	25.9	26.2	26.1	26.4	*
33.3	23.8	24.1	24.0	24.0	*
40	21.6	21.7	21.7	21.6	*
46.6	19.8	19.6	19.7	19.8	*
53.3	18.1	17.9	17.9	18.1	*
60	16.4	16.3	16.4	16.5	*

Table 3: The misclassification rate on the rejected samples.

tion rate and identify the remaining 20% as hard cases that need special considerations.

As Table 2 and Fig. 6 illustrates, the effect of various rejection methods based on different δ-measures are similar and it also holds for the δ_{L1}, which is a non-Bayesian uncertainty measure. However, they have slightly different characteristics which can be interesting for various decision support strategies or problems.

7 Conclusions

In this paper we investigated one of the advantages of the Bayesian approach - the provided additional uncertainty information for predictions - in a medical classification problem. We performed a Bayesian analysis using Belief Network models to discriminate between benign and malignant ovarian masses allowing the exclusion of some data points.

We introduced various uncertainty measures for characterizing the confidence in the prediction. Preliminary medical evaluations show that these uncertainty measures are promising tools for identifying hard cases. We demonstrated that a classifier with 'rejection' based on these measures can efficiently exclude certain subset to improve its performance on the remaining cases significantly. In practice, this may result in a decision support method where the difficult cases are classified as 'rejected' requiring special investigations, while the MR is lower on the remaining subset than on the overall set. Though the examined uncertainty measures behave similarly for this modeling method and problem, their slightly different characteristics can be utilized in various decision support strategies or problems. In general, their comparison needs further investigation.

Acknowledgements

Joos Vandewalle is Full Professor at the K.U.Leuven. Bart De Moor is Full Professor at the K.U.Leuven. Peter Antal is a Research Assistant with the K.U.Leuven. Geert Fannes is a Research Assistant with the F.W.O. Vlaanderen. Frank De Smet is a research assistent with the K.U.Leuven. This work was carried out at the ESAT laboratory and supported by grants and projects from the Flemish Government: Concerted Research Action GOA-MEFISTO-666 (Mathematical Engineering for Information and Communication Systems Technology) and F.W.O. project G.0262.97: Learning and Optimization: an Interdisciplinary Approach and the F.W.O. Research Communities: ICCoS (Identification and Control of Complex Systems) and Advanced Numerical Methods for Mathematical Modelling and Bilaterale Wetenschappelijke en Technologische Samenwerking Flanders-Hungary, BIL2000/19; from National Fund for Scientific Research (OTKA) under contract number T030586; from National Fund for Scientific Research (OTKA) under contract number F-030763; from the IDO/99/03 project (K.U.Leuven) "Predictive computer models for medical classification problems using patient data and expert knowledge", of the FWO grants G.0326.98 and G.0360.98, the FWO project G.0200.00. and Cult. Affairs-Interuniversity Poles of Attraction Programme (IUAP P4-02 (1997-2001): Modeling, Identification and Control of Complex Systems. The scientific responsibility is assumed by its authors.

References

[1] P. Antal, H. Verrelst, D. Timmerman, Y. Moreau, S. Van Huffel, B. De Moor, and I. Vergote, *Bayesian networks in ovarian cancer diagnosis: Potential and limitations*, Proc. of the 13th IEEE Symp. on Comp.-Based Med.Sys., 2000, Houston, pp. 103–109.

[2] E. Castillo, J. M. Gutiérrez, and A. S. Hadi, *Expert systems and probabilistic network models*, Springer, 1997.

[3] D. Heckerman et al., *Learning bayesian networks: The combination of knowledge and statistical data*, Machine Learning **20** (1995), 197–243.

[4] D. J. Spiegelhalter et al., *Bayesian analysis in expert systems*, Statistical Science **8** (1993), no. 3, 219–283.

[5] D. Timmerman et al., *Artificial neural network models for the pre-operative discrimination between malignant and benign adnexal masses.*, Ultrasound Obstet. Gynecol. **13** (1999), 17–25.

[6] J. A. Hanley et al., *The meaning and use of the area under receiver operating characteristic (roc) curve.*, Radiology **143** (1982), 29–36.

Selection of Highly Accurate Genes for Cancer Classification by Estimation of Distribution Algorithms

Rosa Blanco, Pedro Larrañaga, Iñaki Inza, Basilio Sierra
Computer Science and Artificial Intelligence Department
University of the Basque Country
P.O. Box 649, 20080 San Sebastián – Donostia
{rosa, ccplamup, ccbincai, ccpsiarb}@si.ehu.es

Abstract

In spite of cancer classification is considerably improved, nowadays a general method that classifies known types of cancer has not been yet developed. In this work, we propose the use of supervised classification techniques, coupled with feature subset selection algorithms, to automatically perform this classification in gene-expression datasets. Due to the huge number of features of gene-expression datasets, the search of a highly accurate combination of features is done by means of the new Estimation of Distribution Algorithms paradigm. In order to asses the accuracy level of the proposed approach the *naïve-Bayes* classification algorithm is employed in a wrapper form. Promising results are achieved, in addition to a considerable reduction in the number of genes. Stating the optimal selection of genes as a search task, an automatic and robust choice in the finally selected genes is performed, in contrast to previous works in the same type of problems.

1 INTRODUCTION

Cancer classification is based basically on morphological appearance of the tumor. However, tumors with similar appearance present different responses to therapy. This fact makes very important a correct cancer classification. The gene expression data can be used to learn classification models to aid cancer classification. Taking into account that one pattern only belongs to one class (or type of cancer), the probabilistic approach to the supervised classification problem reduces to find c^* such as:

$$c^* = \arg\max_c p(C = c \mid X_1 = x_1, \ldots, X_n = x_n)$$

where C is the cancer class variable and X_i $(i = 1, 2, \ldots, n)$ is the variable related with the i-th gene expression data. Nevertheless, depending on the model and the number of features (and their values) of the data set, the solution of the previous problem might require a huge number of instances in order to reliably estimate the parameters needed to learn the joint probability distribution.

On the other hand, the cases of gene expression datasets usually have a great number of variables. Thus, the question is whether all variables are "useful" to correctly classify new instances. The Feature Subset Selection problem (FSS) tries to answer this question, searching the best subset of features for a data set and a learning algorithm [2; 9].

Obviously, FSS has several advantages. Some of them are: improvement of the comprehensibility of the final classification model, a faster induction of it, and an improvement in the classification accuracy.

Several classification algorithms can be chosen to solve the supervised classification problem. The *naïve-Bayes* [5] is a paradigm based on the conditional independence of the predictive variables given the class. Thus, the number of parameters to estimate the joint probability distribution is considerably reduced.

The aim of this work, i.e. a feature subset selection to maximize the classification model accuracy, can be expressed in the form of a search problem [10]. In our work, the search engine are the novel Estimation of Distribution Algorithms (EDAs) [12]. EDAs have been successfully used in similar FSS problems [7]. However, due to the huge number of genomic features, the initialization of the variables' probabilities is a crucial point: four types of initializations are proposed, three of them based on the results of a classic greedy search algorithm. To guide the search, a wrapper approach over *naïve-Bayes* is used. As previous works [3; 6; 15] in this kind of problems are not based on a search task, they perform a somewhat arbitrary choice in the finally selected number of genes, with the use of a search technique, an automatic and robust choice is performed.

Two different well-known, gene expression datasets are used to test the proposed approach. The first dataset, related with colon cancer, has 62 instances involving 2000 predictive variables (gene expression length) and the class indicates whether the patient suffers cancer or not. The second dataset is related with leukemia: 72 instances containing 7129 predictive variables (gene expression length) are presented and the class shows the kind of leukemia suffered: AML or ALL. The experimental results suggest that the accuracy of *naïve-Bayes* classifier is improved (better than 90%) with a significant reduction in the number of variables involved in the learning (less than 20 in all runs).

The work is organized as follows: the next section presents the wrapper approach, *naïve-Bayes* supervised paradigm and EDAs. Section 3 presents the integration of these elements to carry out the FSS, employing three different initialization methods. Section 4 shows the experimental results. We finish with conclusions and future work.

2 WRAPPER, NAÏVE-BAYES AND EDA PARADIGMS

2.1 THE WRAPPER APPROACH

Irrelevant features on the data set can degrade the predictive accuracy of learning algorithms. Features which information contribution is overlapped or repeated can act in the same way. Algorithms such as *naïve-Bayes* are robust with respect to irrelevant features but very sensitive to correlated features.

This lack of accuracy can be improved if the learning algorithm only uses the adequated features [10]. For this purpose, a feature selection process is required. FSS can be used to find a feature subset that maximizes the predictive accuracy of the classification model built over this subset. From this point of view, FSS can be faced as a search problem where each point of the search space represents a feature subset [10].

The aim of the search is to maximize the performance of the classifier. Some evaluation functions carry out this goal by looking only at the intrinsic characteristics of the data and measuring the power to discriminate among the classes of the problem. These kind of evaluation functions are known as *filter* functions [9] report that when the goal is to maximize the accuracy of the classification model, the FSS should depend not only on the features and the concept to learn, but also on the characteristics of the classifier. This allows the development of the *wrapper* approach: when a feature subset is selected by the search algorithm, its predictive accuracy is estimated with respect to the supervised classification algorithm proposed to generate the final model.

2.2 THE NAÏVE-BAYES PARADIGM

The goal of a supervised classification algorithm is to build a classification model using a data set. This model is used to predict the class of new instances. From a probabilistic perspective, the class chosen, c^*, for a given new instance will be the class with the highest a posteriori probability, given the values of the predictive features:

$$c^* = \arg\max_c p(C = c \mid X_1 = x_1, \ldots, X_n = x_n).$$

The cost of the estimation of the class depends on the complexity of the model and the assumptions over the data.

The *naïve-Bayes* is a supervised classification algorithm built over the assumption of conditional independence of the predictive variables given the class. Although this assumption is violated in numerous occasions, this fact does not degrade the performance of the paradigm in many situations [5]. Under this assumption, the prediction of the class for an unseen instance is simplified.

When the predictive features are discrete the predicted class for an unseen $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ test instance is as follows:

$$c^* = \arg\max_c p(C = c) \prod_{i=1}^{n} p_{X_i \mid C=c}(x_i \mid c)$$

where $p_{X_i \mid C=c}(x_i \mid c)$ represents the conditional probability of $X_i = x_i$ given that $C = c$.

In the case that the predictive features are continuous:

$$c^* = \arg\max_c p(C = c) \prod_{i=1}^{n} f_{X_i \mid C=c}(x_i \mid c)$$

where $f_{X_i \mid C=c}(x_i \mid c)$ represents the density function of the i-th feature conditioned on $C = c$.In this work, we assume that the previous density conditioned functions follow a normal distribution. That is, for all $i = 1, \ldots, n$ and $c = 0, 1$:

$$f_{X_i \mid C=c}(x_i \mid c) \rightsquigarrow \mathcal{N}(\mu_i^c, \sigma_i^c).$$

In both cases –with predictive variables either discretes or continuous– the parameters are estimated by means of their maximum likelihood estimates.

2.3 ESTIMATION OF DISTRIBUTIONS ALGORITHMS

A new approach in the evolutionary computation to solve optimization problems is the Estimation of Distribution Algorithms (EDAs) [12; 11; 14]. Its birth is motivated by the difficulty to choose

1. $D_0 \leftarrow$ Generate M individuals randomly (the initial population)

2. Repeat for $l = 1, 2, \ldots$ until the stopping criterion is met:

2.1. $D_{l-1}^{Se} \leftarrow$ Select $N < M$ individuals from D_{l-1} according to the selection method

2.2. $p_l(\mathbf{x}) = p(\mathbf{x}|D_{l-1}^{Se}) \leftarrow$ Estimate the probability distribution of selected individuals

2.3. $D_l \leftarrow$ Sample M individuals from $p_l(\mathbf{x})$ (the new population)

Figure 1: Pseudo-code for EDA Approach.

the optimal parameters in Genetic Algorithms and the impossibility to predict the movements of the populations in the search space [11].

Although they are based on populations, there are neither crossover nor mutations operators in EDAs. Instead, the new population of individuals is sampled from a probability distribution, which is learned from some selected individuals at each generation.

Figure 1 shows the basic scheme of the EDA paradigm. In the first step M individuals are generated at random, for example, from a uniform distribution for each variable. These M individuals constitute the initial population, D_0, and each of them is evaluated. In an iterative process until the stopping criterion is met we repeat the following steps: first, a number N ($N < M$) of individuals are selected usually those with the best objective function values. Second, a n-dimensional probability distribution is learned from the selected individuals. Finally, M new individuals (the new population) are obtained from sampling the probability distribution learned in the previous step.

The estimation of the joint probability distribution associated to the selected individuals is the bottleneck of EDAs. Different ways to estimate this joint probability exist, with different assumptions on the interrelations between the variables.

The most simple assumption that can be made over the variables is their independence. In this way, the new individuals can be generated by sampling from the univariate probability distribution of each variable. The Univariate Marginal Distribution Algorithm (UMDA) [13] works in this way. It estimates the joint probability distribution of the selected individuals at each generation, $p_l(\boldsymbol{x})$ in the following form:

$$p_l(\boldsymbol{x}) = p(\boldsymbol{x} \mid D_{l-1}^{Se}) = \prod_{i=1}^{n} p_l(x_i|D_{l-1}^{Se}) =$$

$$\prod_{i=1}^{n} \frac{\sum_{j=1}^{N} \delta_j(X_i = x_i \,|D_{l-1}^{Se})}{N}$$

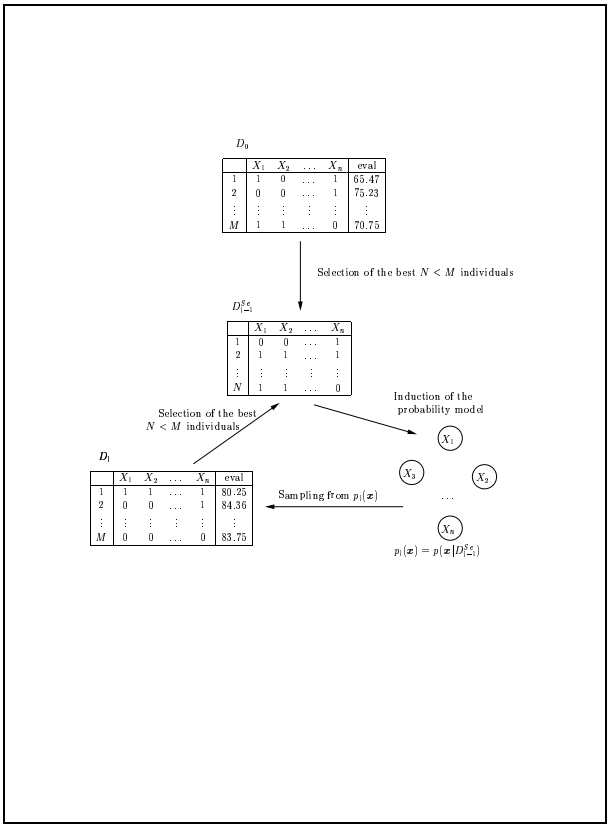


Figure 2: FSS by the EDA (UMDA) Approach.

where

$$\delta_j(X_i = x_i \mid D_{l-1}^{Se}) = \begin{cases} 1 & \text{if in the } j\text{-th case of } D_{l-1}^{Se}, \\ & X_i = x_i \\ 0 & \text{otherwise.} \end{cases}$$

That is, the joint probability distribution of the selected individuals at each generation, $p_l(\boldsymbol{x})$, is factorized as a product of independent univariate marginal distributions. Each univariate marginal distribution is estimated from marginal frequencies.

Due to the huge dimension of our genomic databases, the use of an EDA approach covering the interrelations of order two or superior among the variables of the problem is discarded [11]. Moreover, the number of parameters to estimate these multivariate relations might be also huge.

Figure 2 overviews the proposed approach to select features by means of the EDA (UMDA) algorithm.

3 PROPOSED APPROACH

Taking into account the huge dimension of the problem, an appropriate initialization of the search can save much computation time [1]. In this work, the search initialization is based on the simulation of a probability distribution for each variable. We compare four different initializations, three of them

based in the results of a FSS greedy algorithm and the *naïve-Bayes*.

Greedy algorithms are deterministic algorithms, that is, over a fixed dataset and with the same initial conditions they always give the same solution. Sequential Forward Selection (SFS) [8] is a classic greedy search algorithm which starts from an empty subset of features and sequentially selects features until no improvement is achieved in the evaluation function value.

Based on the feature subset obtained by SFS, three initializations for EDAs are proposed. First, init-A, assigns to all variables the same probability resulting in a Bernouilli distribution with parameter $p = \frac{nvcg}{tnv}$ where *nvcg* is the number of variables chosen by the greedy algorithm and *tnv* is the total number of variables of the data set. Thirdly, init-B, assigns to each variable a probability proportional to the accuracy estimated when the classifier is only built with this variable. Finally, the third initialization, init-C, assigns to each variable a probability proportional to the accuracy estimation increment when the classifier is built with this variable and the ones selected before by SFS. It must be noted that for three initialization methods, the expected number of selected features in each individual of the first population is the number of variables finally selected by SFS. The init-0 initicalization is not depending on the feature subset obtained by SFS. In this initialization each variable is chosen with probability 0.5.

In the proposed EDA approach the population size is fixed to 100 individuals, and 50 individuals are selected to learn the probability distribution.

Each solution is evaluated to measure the accuracy of the built model by means of *leave-one-out*. If we denote the number of instances as n_c, this kind of cross-validation builds a model with $n_c - 1$ instances of the dataset and tests it with the remaining instance, leaving as a test set one different instance the n_c times. The accuracy of the classification model built with the n_c instances is estimated by the percentage of correctly classified instances obtained with the n_c models induced with $n_c - 1$ instances.

4 EXPERIMENTATION IN ONCOLOGY

The proposed approach has been carried out over two well-known biological data sets. The first was presented by [4]. This data set is composed with 62 instances of colon cancer patients. Each instance is characterized by 2000 predictive variables, each one related with the numeric expression of a certain gene. The task to be predicted is whether patients suffer colon cancer disease.

The second data set was proposed by [6]. It contains 72 instances of leukemia patients involving 7129 variables, each one related with the numeric expression of a certain gene. The class to be predicted is the specific type of leukemia: AML or ALL.

For he discrete *naïve-Bayes* models each variable is discretized into two values taking into account its corresponding mediane.

Table 1 shows the results of *leave-one-out* with all the variables and SFS. The results support that not all the variables are relevant to learn the classification model or the existence of redundant features.

Table 1: Results of *Leave-One-Out* with All the Variables and SFS.

DATA	TYPE	ALL FEATURES		SFS	
		accuracy	no.var.	accuracy	no.var.
Colon	disc	70.97	2000	91.93	5
	cont	53.23	2000	95.83	3
Leukemia	disc	63.89	7129	98.61	6
	cont	84.72	7129	87.09	2

These results follow the discoveries of [6] and [15], relating the low number of features needed to improve the accuracy of the whole feature set.

For each dataset and initialization method 10 EDAs independent runs have been executed. Table 2 shows the estimated accuracy of *naïve-Bayes* and the number of selected features for the best run of each initialization method. Table 3 shows the estimated average accuracy, the number of selected features for the 10 executions of each initializations method and the average generation where the best solution on the execution is shown.

Table 2: Best Estimated Accuracy and Corresponding Number of Features.

DATA	TYPE	INIT.	ACC.	VAR.
Colon	disc	init-0	67.74	985
		init-A	95.16	13
		init-B	95.16	13
		init-C	91.93	5
	cont	init-0	74.19	1069
		init-A	98.39	6
		init-B	98.39	10
		init-C	95.16	3
Leukemia	disc	init-0	45.8	3402
		init-A	100	8
		init-B	98.61	15
		init-C	98.61	6
	cont	init-0	76.39	3587
		init-A	100	10
		init-B	100	11
		init-C	98.61	4

Although EDAs in the continuous model do not report a significant accuracy improvement with respect to SFS in the Colon dataset, the opposite behavior,

Table 3: Average Results: Estimated Accuracy and Number of Features. Average Generation Where the Best Solution of the Run Appears. Standard-deviation of Averages is also Reported.

DATA	TYPE	ACC.	VAR.	GENER.
Colon	0	64.5 ± 0.2	987 ± 39.1	29.0 ± 6.9
	disc A	91.9 ± 0.1	11.9 ± 4.1	13.0 ± 4.0
	B	91.2 ± 0.2	11.8 ± 3.2	11.8 ± 3.2
	C	90.9 ± 0.1	6.3 ± 1.6	3.9 ± 1.6
	0	64.9 ± 10.5	1035 ± 52.4	19.14 ± 8.7
	cont A	95.0 ± 2.3	7.1 ± 2.1	15.2 ± 4.6
	B	94.7 ± 2.9	7.2 ± 2.4	12.7 ± 6.9
	C	93.4 ± 1.6	6.0 ± 1.9	12.8 ± 5.0
Leukem	0	44.0 ± 0.1	3476 ± 57.0	18.2 ± 6.7
	disc A	97.2 ± 0.1	14.6 ± 3.6	14.2 ± 4.2
	B	96.9 ± 0.1	14.8 ± 3.6	12.9 ± 4.7
	C	98.6 ± 0.0	8.1 ± 1.8	3.3 ± 1.2
	0	75.9 ± 0.8	3561 ± 35.9	9.3 ± 1.5
	cont A	98.8 ± 1.8	11.0 ± 3.6	18.1 ± 5.7
	B	98.8 ± 1.5	11.8 ± 3.2	16.3 ± 3.6
	C	96.3 ± 1.1	3.7 ± 1.1	5.9 ± 5.0

obtaining a significant accuracy improvement by EDA techniques, is shown in Leukemia domain. However, the use of an extremely low number of features is not recommended in previous works [6]: this is because the use of a so small number of genes ([6] fix 10) may produce a classification model which depends too heavily in any gene, producing spuriously high prediction strengths.

Previous works in this type of problems [6; 15] alert us about their somewhat arbitrary choice in the finally selected number of genes. Thus, stating the problem as a search task and waiting until the convergence, a robust an automatic criteria is adopted to carry out this selection, obtaining competitive results with previous cited works.

Table 4: p-values when Comparing A, B, and C Initializations

DATA	INIT	ACCURACY	NO.VAR.	NO.GEN.
Colon	disc	$p = 0.440$	$p = 0.002$	$p < 0.001$
	cont	$p = 0.232$	$p = 0.446$	$p = 0.187$
Leukem	disc	$p = 0.004$	$p = 0.001$	$p < 0.001$
	cont	$p = 0.003$	$p < 0.001$	$p < 0.001$

We carry out the Kruskall-Wallis test over the results of the A, B, and C initializations. Table 4 reports the test outcome.

In the Colon database, we only obtained differences significants ($p < 0.05$) in the discrete model in relation with the number of variables and the number of generations needed to convergence. In the Leukemia database, the test showed that the differences in all criteria in respect with the three initializations are statistically significants in the two models.

Table 5 shows the results obtained when applying the Mann-Whitney test in order to compare the behaviour between the discrete and continuous *naïve-Bayes* models.

Table 5: p-values when Comparing Discrete Versus Continuous Models

DATA	INIT	DISCRETE vs CONTINUOUS		
		accuracy	no.var.	no.gen.
Colon	0	$p = 0.47$	$p = 0.042$	$p = 0.174$
	A	$p = 0.007$	$p = 0.009$	$p = 0.353$
	B	$p = 0.043$	$p = 0.004$	$p = 0.631$
	C	$p = 0.001$	$p = 0.631$	$p < 0.001$
Leukem	0	$p = 0.017$	$p = 0.017$	$p = 0.067$
	A	$p = 0.063$	$p = 0.063$	$p = 0.075$
	B	$p = 0.089$	$p = 0.075$	$p = 0.063$
	C	$p < 0.001$	$p < 0.001$	$p = 0.218$

In the Colon database we found differences statistically significants in relation with the accuracy of the model for the initializations A, B, and C obtaining the best results in the case of continuous *naïve-Bayes*. With respect to the number of variables selected by the EDA, we obtain that in the initializations A and B the continuous *naïve-Bayes* model needs significantly more variables than its corresponding discrete model. Finally and regarding the number of generations needed until convergence is reached the differences are statistically significants for initialization C where the discrete *naïve-Bayes* need a bigger number of generations.

In the Leukemia database we only found that the differences are statistically significants in the case of initialization C and with respect to the accuracy of the model –better result for the discrete *naïve-Bayes*– and the number of variables –less variables for the continuous *naïve-Bayes*.

5 CONCLUSION AND FUTURE WORK

An application of the EDA approach (by its UMDA algorithm) to select a highly accurate combination of genes in two high-dimensional, well-known genomic datasets is carried out. The selection of genes is performed within a wrapper approach with respect to the *naïve-Bayes* supervised classification algorithm. Four different approaches, three of them inspired in a sequential selector, are compared to initialize the EDA search.

The discoveries of previous works on the same datasets are confirmed [6; 15], noting that with a low number of genes the accuracy level of the whole feature set is significantly improved. In contrast of these works, stating the selection of genes as a search task, an automatic and robust selection of the final number of genes is performed.

As future work, apart from UMDA, we plan the use

of other EDA univariate approaches. We also envision the use of other supervised classifiers that extend the univariate scheme of *naïve-Bayes*, involving relations among two or more variables. Finally, the discretization task should be improved using a clever heuristic approach.

Acknowledgments

This work is support by grant 9/UPV/EHU 00140.226-12084/2000 of the University of the Basque Country.

References

[1] Aha, D. and Bankert, R. (1994). Feature selection for case-based classification of cloud types: An empirical comparison. In *Proceedings of the AAAI-94*, pages 106–112.

[2] Almuallim, H. and Dietterich, T. G. (1991). Learning with many irrelevant features. In *Proceedings of Ninth National Conference on Artificial Intelligence*, pages 547–552. MIT Press.

[3] Beibel, M. (2000). Selection of informative genes in gene expression based diagnosis: A nonparametric approach. In *Proceedings of the First International Symposium on Medical Data Analysis*, pages 300–306, New York. Springer-Verlang.

[4] Ben-Dor, A., Bruhn, L., Friedman, N., Nachman, I., Schummer, M., and Yakhini, Z. (2000). Tissue classification with gene expression profiles. *Journal of Computational Biology*, 7(3–4):559–584.

[5] Domingos, P. and Pazzani, M. (1997). On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning*, 29(2–3):103–130.

[6] Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caliguri, M. A., Bloomfield, C. D., and Lander, E. S. (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286:531–537.

[7] Inza, I., Larrañaga, P., Etxeberria, R., and Sierra, B. (2000). Feature subset selection by Bayesian network-based optimization. *Artificial Intelligence*, 123:157–184.

[8] Kittler, J. (1978). Feature set search algorithms. In Chen, C. H., editor, *Pattern Recognition and Signal Processing*, pages 41–60. Sijthoff and Noordhoff.

[9] Kohavi, R. and John, G. H. (1997). Wrappers for feature subset selection. *Artificial Intelligence*, 97(1–2):273–324.

[10] Langley, P. and Sage, S. (1994). Induction of selective bayesian classifiers. In *Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence*, pages 399–406. Morgan Kaufmann.

[11] Larrañaga, P., Etxeberria, R., Lozano, J. A., and Peña, J. M. (2000). Combinatorial optimization by learning and simulation of Bayesian networks. In Boutilier, C. and Goldszmidt, M., editors, *Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence*, pages 343–352. Morgan Kaufmann.

[12] Larrañaga, P. and Lozano, J. A., editors (2001). *Estimation of Distribution Algorithms. A New Tool for Evolutionary Computation.* Kluwer Academic Publishers, Boston.

[13] Mühlenbein, H. (1997). The equation for response to selection and its use for prediction. *Evolutionary Computation*, 5(3):303–346.

[14] Mühlenbein, H. and Paaß, G. (1996). From recombination of genes to the estimation of distributions I. Binary parameters. *Lecture Notes in Computer Science 1411: Parallel Problem Solving from Nature-PPSN IV*, pages 178–187.

[15] Xing, E. P., Jordan, M. I., and Karp, R. M. (2001). Feature selection for high-dimensional genomic microarray data. In *International Conference on Machine Learning.*

Medical Knowledge Acquisition and Automated Generation of Bayesian Networks

Joachim Horn
Siemens AG, Corporate Technology, Information and Communications
CT IC 4, D–81730 München, Germany

Thomas Birkhölzer
Siemens AG, Medical Solutions, Software Components and Workstations
MED SW S, D–91050 Erlangen, Germany

Oliver Hogl
Bavarian Research Center for Knowledge–Based Systems, Knowledge Acquisition Group
Am Weichselgarten 9, D–91058 Erlangen, Germany

Marco Pellegrino
Siemens AG, Corporate Technology, Information and Communications
CT IC 4, D–81730 München, Germany

Ruxandra Lupas Scheiterer
Siemens AG, Corporate Technology, Information and Communications
CT IC 4, D–81730 München, Germany

Kai–Uwe Schmidt
Siemens AG, Medical Solutions
MED GT 2, D–91050 Erlangen, Germany

Volker Tresp
Siemens AG, Corporate Technology, Information and Communications
CT IC 4, D–81730 München, Germany

Abstract

A novel approach for knowledge acquisition and generation of the Bayesian network of a medical domain is presented. First, the knowledge is acquired in a structured representation using the software tool MedKnow that allows the expert to specify diseases and findings, their interconnections, and specific marginal and conditional probabilities. Next, the software tool KnowledgeCompiler generates the Bayesian network, embracing both the graph of the network and the conditional probability tables. For calculation of the conditional probability tables, the acquired probabilities first get transformed to yield those probabilities that describe the respective logical gate. The resulting Bayesian network can be used by the HealthMan® Dialogue and Advisory System.

1 Introduction

Probabilistic models such as Bayesian networks are well suited for medical decision support and are the basis of many successful applications [1; 3; 4; 7; 9; 10; 11]. Bayesian networks [6] provide a rigorous and efficient framework for inference, i.e. for calculating the probability of each stochastic variable given a set of observations. Nevertheless, knowledge acquisition and generation of the network are still demanding tasks when large medical domains have to be modelled.

Here, a novel approach for knowledge acquisition and generation of the network is presented. The approach was developed as part of the HealthMan® project [2]. First, the knowledge is collected and put into a structured representation using a software tool (MedKnow) tailored for the medical domain. MedKnow allows the expert to specify diseases and findings, their interconnections, and specific marginal and conditional probabilities. Next, another tool (Knowl-

edgeCompiler) generates the Bayesian network using the structured knowledge. The resulting network can be used by the HealthMan® Dialogue and Advisory System.

An overview of the HealthMan® Dialogue and Advisory System is given in section 2. Knowledge acquisition using MedKnow is discussed in section 3. Automated generation of Bayesian networks using KnowledgeCompiler is presented in section 4.

2 Overview of the HealthMan® Dialogue and Advisory System

Figure 1: HealthMan® Dialogue and Advisory System

Siemens Medical Solutions has started a major effort to introduce intelligent communication in medical care by using modern information technologies with the goal of achieving a more efficient, faster, better and more affordable patient health care system. It is well known that the most important source of information for a physician is his dialogue with the patient - despite all advanced measurement devices. Therefore, the key component of such an intelligent communication will be a generally applicable medical dialogue and advisory system - the HealthMan® Dialogue and Advisory System. To give a few examples, HealthMan® will provide self diagnosis capabilities servicing the most common health challenges for the family, it will be the health advisor for patients with chronic health problems supporting their disease management process, or it will allow a physician to focus on the real issues by automating necessary but tedious routine questionnaires.

HealthMan® emulates the anamnese process of the physician, i.e. an interactive process which is dynamically driven by medical knowledge analyzing the information already at hand (history of the dialog, data from the integrated personal health record). Bayesian networks are an appropriate technology, because they allow knowledge acquisition in the medically relevant direction i.e. from diseases to symptoms by being able to take into account prior disposition for diseases. Furthermore, Bayesian networks provide a formally correct calculus for the inherent uncertainty. The HUGIN Bayesian network library is used for inference.

As an experimental domain, the scenario 'initial assessment of the severity of common child diseases' was chosen. In collaboration with several pediatricians, networks for several subdomains (e.g. infections, respiratory system, skin, abdomen, eyes, ears) were developed. The system was tested by a professional usability lab and was generally well received by the users (mothers of young children) as well as by the physicians.

3 Knowledge Acquisition Using MedKnow

There were two main goals in developing MedKnow: First, MedKnow allows medical experts to formulate their medical knowledge, without requiring intimate knowledge of Bayesian networks or probability theory. Second, MedKnow makes sure that the acquired knowledge is complete such that the Bayesian network can be generated automatically.

MedKnow uses two classes of stochastic variables: diseases and findings. A finding may play the role of a symptom or the role of an enhancing or inhibiting factor of a disease. An example of knowledge representation using MedKnow is shown in figure 2. In the left part of the window, all defined diseases and findings are listed. In the main part of the window, the selected disease or finding is presented. Here, the medical domain of infections is modeled, and the disease 'measles' is selected.

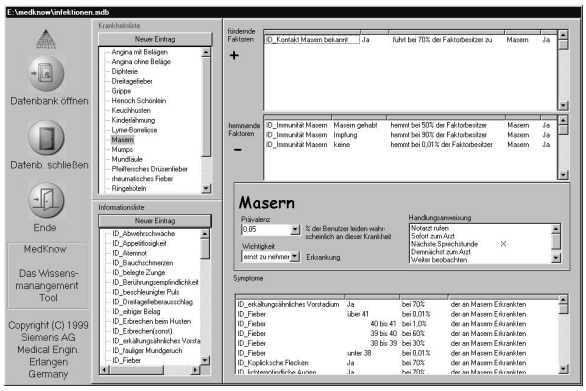


Figure 2: Specification of measles using MedKnow

The upper part of the main window shows the enhancing and inhibiting factors of the disease, here contact to infected persons and immunity. Furthermore, conditional probabilities have to be specified to quantify the effect of the enhancing and inhibiting factors. The meaning of these conditional probabilities and underlying assumptions will be discussed in detail in section 4. The central part of the main window shows

the selected disease, its marginal probability, and additional information used in HealthMan®, e.g. the urgency to see a doctor. The lower part of the main window shows the symptoms of the disease together with the conditional probability that the disease will cause the symptom.

A similar display is presented when a finding has been selected.

4 Automated Generation of Bayesian Networks Using KnowledgeCompiler

Generating the Bayesian network using the acquired knowledge may be divided into two subtasks: generating the graph and calculating the conditional probability tables.

Generating the graph is straightforward: each disease and finding is represented by a node and additional nodes are created for – separately – collecting enhancing factors and for collecting inhibiting factors of each single disease. Directed edges are drawn from diseases to the respective symptoms, from enhancing factors to the respective collecting nodes, from inhibiting factors to the respective collecting nodes, and from collecting nodes to the respective diseases. Figure 3 shows the graph of the Bayesian Network for infections generated by KnowledgeCompiler.

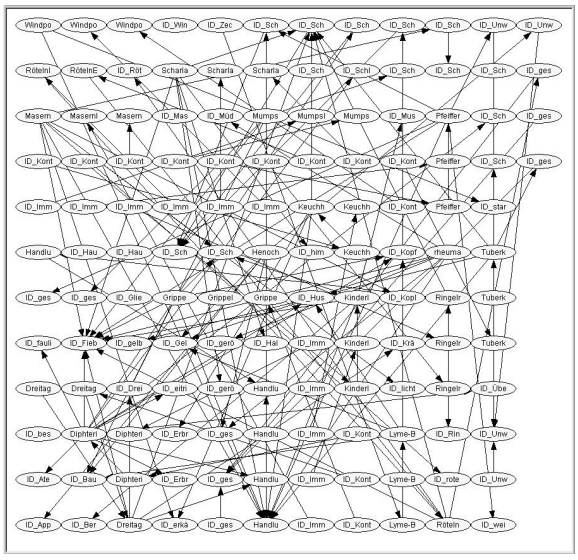


Figure 3: Bayesian Network for infections generated by KnowledgeCompiler

Calculation of the conditional probability tables of the Bayesian network is based on the specified probabilities and the selected type of gate. For findings, gates like NoisyOR [6], NoisyMAX, and NoisyELENI [8] are used. Diseases are modelled as an enhance–inhibit–gate [5].

The expert specifies those probabilities that are most convenient and well–known. Considering the

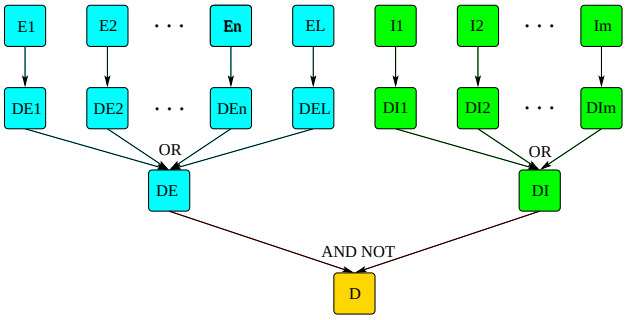


Figure 4: Model for enhancing and inhibiting factors of a disease

enhance-inhibit–gate, typically the marginal probability $P(D)$ of disease D being present, the conditional probabilities $P(D|E_i)$, $i = 1,\ldots,n$, of disease D being present given that the enhancing factor E_i is present, and the conditional probabilities $P(D|I_i)$, $i = 1,\ldots,m$, of disease D being present given that the inhibiting factor I_i is present, are specified. Nevertheless, these probabilities are not explicitly used to specify the enhance–inhibit–gate. Instead, the leak probability $P(DE_L)$ and the conditional probabilities $P(DE_i|E_i)$, $i = 1,\ldots,n$, and $P(DI_i|I_i)$, $i = 1,\ldots,m$, are required. Thus, the probabilities specified by the medical expert have to be transformed to yield the probabilities that are used for the gate. The transformation equations are derived in the sequel. For the sake of simplicity, the case of binary variables is discussed.

4.1 Problem Formulation

Disease D is influenced by enhancing factors E_i, $i = 1,\ldots,n$, and inhibiting factors I_i, $i = 1,\ldots,m$, according to

$$D{=}(\text{NoisyOR}\,(E_1,\ldots,E_n,E_L))\ \text{AND NOT}\ (\text{NoisyOR}\,(I_1,\ldots,I_m))\ , \tag{1}$$

where E_L denotes the leak enhancing factor which is not specified by the expert. Thus, the probability $P(D)$ of disease D being present is given by

$$P(D) = \left[1-(1-P(DE_L))\prod_{i=1}^{n}(1-P(DE_i|E_i)\,P(E_i))\right] \prod_{i=1}^{m}(1-P(DI_i|I_i)\,P(I_i))\ . \tag{2}$$

Therefore, the probabilities $P(DE_L)$, $P(DE_i|E_i)$, $i = 1,\ldots,n$, $P(DI_i|I_i)$, $i = 1,\ldots,m$, $P(E_i)$, $i = 1,\ldots,n$, and $P(I_i)$, $i = 1,\ldots,m$, have to be known. It is assumed that the factors are indepent.

The medical expert specifies the following probabilities:

$$P(DE_i|E_i)\ \text{ or }\ P(D|E_i)\ \text{ or }\ P(E_i|D)\,,\ i = 1,\ldots,n\ ,$$

$P(DI_i|I_i)$ or $P(D|I_i)$ or $P(I_i|D)$, $i = 1, \ldots, m$,
$P(E_i)$, $i = 1, \ldots, n$,
$P(I_i)$, $i = 1, \ldots, m$,
$P(D)$.

Using the specified probabilities, the required probabilities have to calculated.

4.2 Case 1: Specification of $P(DE_i|E_i)$, $i = 1, \ldots, n$, and $P(DI_i|I_i)$, $i = 1, \ldots, m$

Here, only the leak enhancing probability $P(DE_L)$ has to be calculated. Solving equation (2) yields

$$P(DE_L) = 1 - \left[1 - P(D) \Big/ \prod_{i=1}^{m} (1 - P(DI_i|I_i)\, P(I_i))\right] \Big/ \left[\prod_{i=1}^{n} (1 - P(DE_i|E_i)\, P(E_i))\right] \quad . \tag{3}$$

4.3 Case 2: Specification of $P(D|E_i)$, $1 \le i \le n$, and $P(D|I_i)$, $1 \le i \le m$

In the sequel, it is shown that case 2 can be transformed into case 1.

Specification of $P(D|I_i)$, $1 \le i \le m$
Here it is shown that $P(DI_i|I_i)$, $1 \le i \le m$, can be calculated using $P(D|I_i)$ and the marginal probabilities $P(D)$ and $P(I_i)$. For $1 \le i \le m$, we have

$$P(D) = P(D|I_i)\; P(I_i) + P\left(D|\bar{H}_i\right)\; (1 - P(I_i)) \tag{4}$$

with

$$P\left(D|\bar{H}_i\right) = \left[1 - (1 - P(DE_L)) \prod_{k=1}^{n} (1 - P(DE_k|E_k)\, P(E_k))\right] \prod_{\substack{k=1 \\ k \neq i}}^{m} (1 - P(DI_k|I_k)\, P(I_k)) \quad . \tag{5}$$

This yields

$$\left[1 - (1 - P(DE_L)) \prod_{k=1}^{n} (1 - P(DE_k|E_k)\, P(E_k))\right] \prod_{\substack{k=1 \\ k \neq i}}^{m} (1 - P(DI_k|I_k)\, P(I_k)) = (P(D) - P(D|I_i)\; P(I_i)) \,/\, (1 - P(I_i)) \quad . \tag{6}$$

Inserting into equation (2) gives

$$P(D) = (1 - P(DI_i|I_i)\, P(I_i)) \; (P(D) - P(D|I_i)\; P(I_i)) \,/\, (1 - P(I_i)) \quad . \tag{7}$$

Solving for $P(DI_i|I_i)$ finally yields

$$P(DI_i|I_i) = \left[1 - P(D)\; (1 - P(I_i)) \,/\, (P(D) - P(D|I_i)\; P(I_i))\right] \Big/ P(I_i) \quad . \tag{8}$$

Specification of $P(D|E_i)$, $1 \le i \le n$
Here it is shown that $P(DE_i|\bar{E}_i)$, $1 \le i \le n$, can be calculated using $P(D|E_i)$, $P(DI_k|I_k)$, $k = 1, \ldots, m$, and the marginal probabilities $P(D)$, $P(E_i)$, and $P(I_k)$, $k = 1, \ldots, m$. For $1 \le i \le n$, we have

$$P(D) = P(D|E_i)\; P(E_i) \; + \; P\left(D|\bar{F}_i\right)\; (1 - P(E_i)) \tag{9}$$

with

$$P\left(D|\bar{F}_i\right) = \left[1 - (1 - P(DE_L)) \prod_{\substack{k=1 \\ k \neq i}}^{n} (1 - P(DE_k|E_k)\, P(E_k))\right] \prod_{k=1}^{m} (1 - P(DI_k|I_k)\, P(I_k)) \quad . \tag{10}$$

This yields

$$(1 - P(DE_L)) \prod_{\substack{k=1 \\ k \neq i}}^{n} (1 - P(DE_k|E_k)\, P(E_k)) = 1 - (P(D) - P(D|E_i)\; P(E_i)) \Big/ \left((1 - P(E_i)) \prod_{k=1}^{m} (1 - P(DI_k|I_k)\, P(I_k))\right) \quad . \tag{11}$$

Inserting into equation (2) gives

$$P(D) = \left[1 - (1 - P(DE_i|E_i)\, P(E_i)) \left[1 - (P(D) - P(D|E_i)\; P(E_i)) \Big/ \left((1 - P(E_i)) \prod_{k=1}^{m} (1 - P(DI_k|I_k)\, P(I_k))\right)\right]\right] \prod_{k=1}^{m} (1 - P(DI_k|I_k)\, P(I_k)) \quad . \tag{12}$$

Solving for $P(DE_i|E_i)$ finally yields

$$P(DE_i|E_i) = \left\{1 - \left[1 - P(D) \Big/ \prod_{k=1}^{m} (1 - P(DI_k|I_k)\, P(I_k))\right] \Big/ \left[1 - (P(D) - P(D|E_i)\; P(E_i)) \Big/ \left((1 - P(E_i)) \prod_{k=1}^{m} (1 - P(DI_k|I_k)\, P(I_k))\right)\right]\right\} \Big/ P(E_i) \quad . \tag{13}$$

4.4 Case 3: Specification of $P(E_i|D)$, $1 \le i \le n$, and $P(I_i|D)$, $1 \le i \le m$

Applying Bayes' law yields

$$P(D|E_i) = P(E_i|D)\; P(D) \,/ P(E_i), \; 1 \le i \le n \tag{14}$$

and

$$P(D|I_i) = P(I_i|D)\ P(D) \big/ P(I_i)\,,\ 1 \leq i \leq m\ . \tag{15}$$

Thus, case 3 can be transformed into case 2.

4.5 Summary of Calculation of the Required Probabilities

The influence of enhancing factors E_i, $i = 1, \ldots, n$, and inhibiting factors I_i, $i = 1, \ldots, m$, of a disease D is modeled according to equation (1). Thus, the conditional probabilities $P(DE_i|E_i)$, $i = 1, \ldots, n$, the leak enhancing probability $P(DE_L)$, and the conditional probabilities $P(DI_i|I_i)$, $i = 1, \ldots, m$, have to be known. The medical expert specifies the marginal probabilities $P(E_i)$, $i = 1, \ldots, n$, $P(I_i)$, $i = 1, \ldots, m$, and $P(D)$, as well as $P(DE_i|E_i)$, $P(D|E_i)$, or $P(E_i|D)$ for each enhancing factor and $P(DI_i|I_i)$, $P(D|I_i)$, or $P(I_i|D)$ for each inhibiting factor. The required probabilities are calculated according to the following scheme:

- Step 1: If $P(I_i|D)$, $1 \leq i \leq m$, has been specified, $P(D|I_i)$ is calculated using equation (15).
- Step 2: If $P(E_i|D)$, $1 \leq i \leq n$, has been specified, $P(D|E_i)$ is calculated using equation (14).
- Step 3: If $P(D|I_i)$, $1 \leq i \leq m$, has been specified or calculated in step 1, $\overline{P}(DI_i|I_i)$ is calculated using equation (8).
- Step 4: If $P(D|E_i)$, $1 \leq i \leq n$, has been specified or calculated in step 2, $\overline{P}(DE_i|E_i)$ is calculated using equation (13).
- Step 5: $P(DE_L)$ is calculated using equation (3).

5 Conclusions

An approach for knowledge acquisition and automated generation of Bayesian networks has been presented. First, the knowledge is acquired using the software tool MedKnow that allows the expert to specify diseases and findings, their interconnections, and specific marginal and conditional probabilities. Next, the software tool KnowledgeCompiler generates the Bayesian network in HUGIN format [6] using the acquired knowledge. The probabilities specified by the expert get transformed into the conditional probability tables of the specific gate according to [5; 8]. As an example, the equations for the enhance–inhibit–gate have been derived. The resulting Bayesian network can then be used by the HealthMan® Dialogue and Advisory System.

References

[1] S. Andreassen, M. Woldbye, B. Falck, S. K. Andersen: "MUNIN – A Causal Probabilistic Network for Interpretation of Electromyographic Findings". *Proceedings of the Tenth International Joint Conference on Artificial Intelligence*, Milan, Italy, August 1987, pp. 366–372.

[2] T. Birkhölzer, M. Haft, R. Hofmann, J. Horn, M. Pellegrino, V. Tresp: "Intelligent Communication in Medical Care". *Proceedings of the Joint European Conference on Artificial Intelligence in Medicine and Medical Decision Making (AIMDM 99)*, Aalborg, Denmark, June 1999, p. 4.

[3] D. E. Heckerman, E. J. Horvitz, B. N. Nathwani: "Toward Normative Expert Systems: Part I. The Pathfinder Project". *Methods of Information in Medicine*, Vol. 31, 1992, pp. 90–105.

[4] D. E. Heckerman, B. N. Nathwani: "Toward Normative Expert Systems: Part II. Probability–Based Representations for Efficient Knowledge Acquisition and Inference". *Methods of Information in Medicine*, Vol. 31, 1992, pp. 106–116.

[5] J. Horn: *HealthMan Bayesian Network Description: Enhancing and Inhibiting Factors of Diseases.* Siemens AG, ZT IK 4, Internal Report, 1999.

[6] F. V. Jensen: *An Introduction to Bayesian Networks.* UCL Press, 1996.

[7] P. J. F. Lucas, H. Boot, B. Taal: "A Decision–Theoretic Network Approach to Treatment Management and Prognosis". *Knowledge–Based Systems*, Vol. 11, 1998, pp. 321–330.

[8] R. Lupas Scheiterer: *HealthMan Bayesian Network Description: Disease to Symptom Layer.* Siemens AG, ZT IK 4, Internal Report, 1999.

[9] B. Middleton, M. A. Shwe, D. E. Heckerman, M. Henrion, E. J. Horvitz, H. P. Lehmann, G. F. Cooper: "Probabilistic Diagnosis Using a Reformulation of the INTERNIST–1/QMR Knowledge Base. II. Evaluation of Diagnostic Performance". *Methods of Information in Medicine*, Vol. 30, 1991, pp. 256–267.

[10] K. G. Olesen, U. Kjaerulff, F. Jensen, F. V. Jensen, B. Flack, S. Andreassen, S. K. Andersen: "A MUNIN Network for the Median Nerve – A Case Study on Loops". *Applied Artificial Intelligence*, Vol. 3, 1989, pp. 385–403.

[11] M. A. Shwe, B. Middleton, D. E. Heckerman, M. Henrion, E. J. Horvitz, H. P. Lehmann, G. F. Cooper: "Probabilistic Diagnosis Using a Reformulation of the INTERNIST–1/QMR Knowledge Base. I. The Probabilistic Model and Inference Algorithms". *Methods of Information in Medicine*, Vol. 30, 1991, pp. 241–250.

Bayesian Modeling of Cerebral Information Processing

Vincent LABATUT **Josette PASTOR**
INSERM U455
Fédération de Neurologie, CHU Purpan
F31059 TOULOUSE Cedex
France
{Vincent.Labatut, Josette.Pastor}@purpan.inserm.fr

Abstract

Modeling explicitly the links between cognitive functions and networks of cerebral areas is necessitated both by the understanding of the clinical outcomes of brain lesions and by the interpretation of activation data provided by functional neuroimaging techniques. At this global level of representation, the human brain can be best modeled by a probabilistic functional causal network. Our modeling approach is based on the anatomical connection pattern, the information processing within cerebral areas and the causal influences that connected regions exert on each other. The information processing within a region is implemented by a causal network of functional primitives that are the interpretation of integrated biological properties. This explicit modeling approach allows the formulation and the simulation of functional and physiological assumptions.

1 Introduction

In Neurology and Neuropsychology, the understanding and the prediction of the clinical outcomes of focal or degenerative cerebral lesions, as well as the assessment of rehabilitation procedures, necessitate knowing the cerebral substratum of cognitive or sensorimotor functions. Human brain mapping is performed through activation studies, where subjects are asked to perform a specific task while data of their brain functioning are obtained through functional neuroimaging techniques. Such studies, as well as animal experiments, have shown that sensorimotor or cognitive functions are the offspring of the activity of large-scale networks of anatomically connected cerebral regions [Bressler, 1995]. However, a one to one correspondence between activated networks and functions cannot be found in all cases [Démonet *et al.*, 1994]. Understanding such incongruent results is crucial for the care of cerebral lesions.

Neuroimaging techniques and their traditional interpretation methods only address the following topics:

(1) *Visualization of activated areas* (tomographic techniques) and *times of specific cerebral events* (surface electromagnetic techniques);
(2) *What areas could participate in the same function* ("functional connectivity" [Herbster *et al.*, 1996]) and *what is the role of anatomical links on the activation* ("effective connectivity" [Büchel and Friston, 1997]).

Clearly, if the "where" and "when" (1), and the "what" and "how" (2) are answered, the "why", i.e. how the activation of large-scale networks derives from cerebral information processing mechanisms, is missing. Our goal is the understanding of that "why", which only can explain apparently conflicting activation data. Our research is twofold: providing plausible models, at the level of large-scale networks, of cerebral information processing mechanisms in humans and building a flexible simulator, allowing a quick implementation of the models, for a better interpretation of cerebral functional images.

Connectionist methods are the dominant approach in the modeling of the cerebral functional structure. However, they focus on functions emerging from a networked architecture of *populations of undifferentiated neuronal cells* [Grossberg *et al.*, 1997]. Modeling explicitly the role of *networks of regions* on information processing requires departing from this dominant viewpoint for at least two reasons. On one hand, we aim at modeling the function that emerges from the activity of *networks of differentiated cerebral areas*. On the other hand, the information processed by a cerebral area can be considered as the abstraction of the global signal emitted by the region's neurons, representing both the pattern of firing neurons as well as their average firing rate, and can therefore hardly be modeled by a single numerical value. Moreover, since the cerebral response to a given stimulus may vary, the brain can be considered as a probabilistic information processor. In the next paragraphs, we will demonstrate that these constraints are in favor of a Bayesian approach, and more especially of causal functional networks.

2 Biological constraints

Modeling is constrained both by the necessity of a certain biological plausibility and by the purpose of the model building, that is allowing neurologists to express *explicitly*, as *cause-effect relationships*, their knowledge and hypotheses about the human brain.

A networked architecture

The nodes of a large-scale cerebral network are functionally homogeneous, anatomically well-defined, *cerebral regions*, connected by *oriented anatomical links* (axon bundles), which are the network's edges [Pastor *et al.*, 2000]. Each region can itself be considered as a functional network of *processor*s, such as *information processors* that are specific neuronal populations (e.g. GABA neurons) implementing *functional primitives* (e.g. inhibition).

Causality and temporality

Every function (primitive or cognitive/sensorimotor function) is, in the brain, the outcome of the activation of an oriented network (called hereafter cerebral network), whose nodes are neurons or neuronal populations, and oriented links are axons or axon bundles. Information propagation results from a *cascade of causal events*, since the signal or information emitted by the firing of a node provokes the activation of its downstream nodes. The brain can therefore be considered as *a causal network*. According to Hume, A is the cause of B if they are contiguous, if A precedes B and if the relationship is regular. Our definition of causality, which extends Hume's one, is based on contiguity, probabilistic regularity and temporal consistency (ie. the beginning of A must precede the beginning of B).

Since anatomical links, which convey information with very short transmission delays, connect physically nodes in a cerebral network, the nodes are spatially and temporally adjacent and the condition of *contiguity* is strictly met.

Either at a large or small scale level, the response of a neuronal population to a given stimulus or information is not deterministic. The relationship between two nodes in a cerebral network has therefore a *probabilistic regularity*.

Physiologically, the temporal consistency is met, that is there is an order of activation for the cerebral areas. But the interpretation of activation data requires *representing explicitly time* in the models, and this representation has to be consistent with both the sampling time of neuroimaging techniques and the cerebral processing time. Depending on the temporal granularity chosen in the model, a cause-node and an effect-node could fire within the same time unit. This could lead to cycles in the network and hence to a loss of the causality. Imposing the model's network not to be cyclic will be necessary to keep the model causal.

A two-dimensioned representation of information

Cerebral information can be considered as the abstraction, at the level of a neuronal population, of the integrated activity of the individual cells. Any piece of information is defined as a couple of an *energy* and a *category*, where category stands for which neurons react to a specific stimulus, and energy determines how they respond [Pastor *et al.*, 2000].

The cerebral *energy* reflects roughly the number of firing neurons and their firing rates. The energy of a stimulus can be extracted from its physical parameters (e.g. the intensity for a sound). It has a numerical representation.

The *category* of a stimulus summarizes the minimal set of physical properties that characterizes the information (e.g. the frequency for a tone). This "external" category is consistent with the "internal" category, that corresponds to the general pattern of neurons excited by the information. Information categorization is reflected in the "topic" organization of primary cortices and other areas [Alexander *et al.*, 1992]. For example, the auditory cortex can be decomposed in subareas reacting to precise frequency intervals. The category has a symbolic representation.

The pattern and the number of activated fibers of an axon bundle [Leiner and Leiner, 1997], which correspond to the pattern and the number of activated neurons in the emitting cerebral node, represent the category and the energy that is transmitted between two nodes.

Uncertainty and imprecision

Uncertainty arises from the probabilistic regularity of cerebral events.

Furthermore, in humans, the only external evidences of energy values are provided by neuroimaging techniques and are therefore very imprecise. For example, the metabolic activity (tomographic signal) is an indirect measure of the neuronal activity.

Conditions and non-linearity

The relationships between cerebral nodes (neurons or areas) are intrinsically non-linear. Moreover, the presence of conditions on information propagation increases the non-linearity of the brain processing. These conditions may go from very simple (firing thresholds) to very complex (role of special areas on the propagation between other regions).

Habituation and learning

Both are related to the brain's adaptability. Habituation is a transient decrease of the activation that occurs when a neuronal population receives consecutively, several times, the same stimulus and that disappears when a new stimulus is presented. This kind of "energy saving" phenomenon may happen as soon as the second presentation of a stimulus [Miller *et al.*, 1991].

Learning is a permanent change of the brain state that occurs when a neuronal population receives regularly the same information pattern. The population's response becomes more efficient, that is fewer neurons fire and they become specialized in the processing of that information. The population is supposed to create a new information category, which represents the information pattern.

3 A new formalism for cerebral modeling

So-called "causal networks" should meet the two first

constraints of §2. However, all do not cope with the other requirements and all do not deserve to be called "causal". The pros and cons of different causal formalisms are described hereafter, and our arguments in favor of probabilistic functional causal networks are given.

3.1 Causal Qualitative Networks

Causal Qualitative Networks (CQNs) have initially been designed to model physical devices and they are largely inspired by process control. CQNs are oriented graphs, whose nodes are qualitative variables, generally state variables, and edges are cause-effect relationships, generally influences between the state variables.

Causality is based here on three requirements [de Kleer, 1979]: locality (the cause acts only on its direct neighbors), precedence and regularity. Locality is weaker than contiguity, the neighborhood being not precisely defined. Precedence and regularity are stronger constraints than temporal consistency and probabilistic regularity. Therefore, CQNs do not meet our definition of causality

Qualitative algebras are at the core of CQNs. They take imprecision into account implicitly, by representing numerical values by some qualitative properties: signs, orders of magnitude or real intervals centered on the values. CQNs do not support uncertainty and, since imprecision is implicitly represented, it is not measurable or controllable.

An interval-based CQN, with an explicit discrete time representation, has been used in a previous tentative modeling of large-scale networks [Lafon *et al.*, 1999; Pastor *et al.*, 2000]. In order to meet biological constraints, the basic formalism was augmented by a limited non-linearity (piecewise linearity) and uncertainty (multivalued logic). However, it suffers drawbacks: a classical flaw of interval calculus [Struss, 1990] makes the range of intervals increase dramatically at each simulation step and uncertainty and imprecision are defined by different formalisms. Moreover, all the causes to a node are processed independently and then combined. This is the opposite of what happens in formal neural networks when the node processes the weighted sum of inputs. Whether the combination precedes the processing or not is still an open question in neuroscience. Those drawbacks restrict considerably the applicability of the system to cerebral modeling and has moved the research effort to Bayesian approaches.

3.2 Dynamic bayesian networks

Among the different dynamic Bayesian networks [Dean and Kanawaza, 1989], State Space Models (SSMs) [Ghahramani, 1998], an extension of Hidden Markov Models (HMMs), seem to be the most interesting formalism for our cerebral modeling approach. Relationships are defined by the probabilities of the current response variables conditionally to the current hidden state variables, and the expression of every current hidden state variable as a linear function of the past values of the hidden state variables, plus a random variable.

SSMs meet our definition of causality. The respect of temporal consistency, contiguity and probabilistic regularity is derived from the definition of the oriented, autonomous and stable relationships [Pearl 2000]. Other constraints are respected: the explicit and discrete representation of time, the possible handling of the numerical (energy) and symbolic (category) parts of cerebral information, the expression of conditions in the relationships' deterministic part, a straightforward measure of uncertainty and learning mechanisms implemented by probability revisions.

However, two major requirements are not satisfied: non-linearity cannot be represented in the deterministic part of hidden state variables and no instantaneous relationship can be defined.

3.3 Causal functional networks

Causal Functional Networks (CFNs) [Pearl, 2000] are based on structural equations. Basic structural equations are asymmetric linear relationships, that is the equality symbol in each equation should be replaced by an affectation symbol (:= or <=) [Druzdzel and Simon 93]. Therefore, they are causal relationships.

However, in most applications of Structural Equation Modeling (SEM), relationships are symmetric and the equations system is identified globally, by fitting the theoretical covariance matrix to the observed one. This non-causal version is used in the "effective connectivity" image interpretation approach [Büchel and Friston, 1997].

CFNs [Pearl 2000] extend *causal* SEM in different aspects: variables can be numerical or symbolic, non-linear functions are used to model relationships and time can have an explicit, discrete representation. In fact, like SSMs, CFNs respect, temporality, uncertainty, conditioning, learning and cerebral information representation constraints and, in addition, they allow non-linearity and instantaneous relationships.

Their main drawback concerns the representation of imprecision: probability theory can directly measure only uncertainty, while imprecision can be only estimated by an average value and a dispersion value. A direct measure of both imprecision and uncertainty could be obtained by the use of the possibility theory [Dubois and Prades, 1994]. However, three points are in favor of the probability calculus: it has a well developed mathematical theory, neuroimaging data are statistical summaries, and overall, brain processing is mostly probabilistic.

CFNs seem to be the best paradigm for cerebral modeling. Moreover, like all causal Bayesian models, they can answer clinical questions such as: "What happens in area A when area B is activated?" (observation) and "What happens when area A is damaged?" (intervention). In addition, they have the specific ability of answering "What would happen if area A was activated, knowing that it is not activated in reality ?" (counterfactual).

4 A Tentative Model of a Cerebral Area

Network

CFNs seem to be the most adapted formalism to model cerebral mechanisms. The adaptation, in terms of a CFN, of the model described by Pastor *et al*. [2000], is given. This model aimed at explaining results from Fox and Raichle's experiment [1984] that study focused on the modulation of the activation of the striate cortex by the presentation rate of visual stimuli.

4.1 The causal network

The hypothesis is that the experimental results can be explained by the interactions between the striate cortex and the thalamus [Pastor *et al*., 2000]. The "large-scale" network is a simple anatomical loop, the cortex and the thalamus being connected by opposite oriented axon bundles. The global functional network is the connection of the two functional networks representing the striate cortex and the thalamus (*Figure 1*), plus an additional node standing for the stimulus. Since delays are associated to the links in the network, at a given time, the network is an acyclic oriented graph.

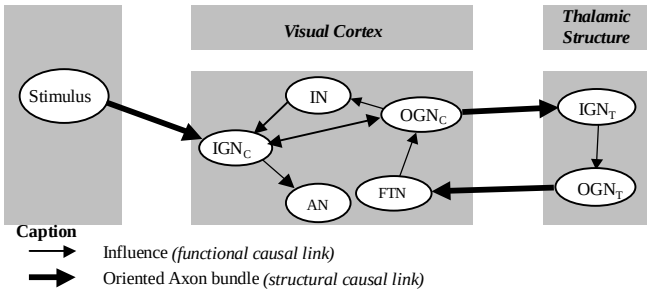


Figure 1. *The structural and functional network*

4.2 Modeling Approach

The cerebral information, or part of it, is processed at each node. It is therefore a flowing entity, while nodes are processing entities and links are propagating entities.

Information Representation

The flowing entity is characterized by the values of its *Magnitude* (the representation of the information energy) and its *Type* (the representation of the category). Its state is, functionally (at each node of the causal network, *after* it has been processed by the corresponding information processor) and temporally (at each discretized instant t), represented by a two-dimension random variable $X(t)=(X_M(t), X_T(t))$, attached to the node. X_M, the magnitude, is a real variable. X_T, the type, takes a multiple symbolic value $\{s_1, \ldots, s_n\}$. and a probability $P(s_i|X)$ is associated to each symbol s_i. A symbol s_i represents a pure type (something theoretical), and the associated probability stands for the proportion of energy (i.e. X_M) emitted by the s_i-typed neuronal population of the X node. At the metabolic processor nodes, the information representation is limited to its magnitude part.

Propagation and Processing

A relationship is a couple of two functions dedicated, respectively to the magnitude and the type. $X(t)$ is updated at each instant t of the simulation, according to the values of its causes, previously computed.

$$X_C(t) = f_{X_C}(PA_X(t), U_{X_C}), X_M(t) = f_{X_M}(PA_X(t), U_{X_M})$$

In the equations, $PA_X(t)$ stands for the parents of $X(t)$, and includes generally $X(t-1)$. The $U_{X_\bullet}$ are error variables that do not depend on time.

For each region R, a Type Preference Table (TPT) contains the region's sensitivity to pure types. It is represented by the set of $P(A|R, s_i)$, where A stands for "Activation" and $P(A|R, s_i)$ represents the chance for R to be activated, given that the received stimulus' category is of the s_i type.

The conditions are expressed by logical expressions that are included in the functions. These conditions take probabilistic values. Currently, to simplify the computation, we only calculate an expression according to the most probable value (true or false) of the corresponding condition (i.e. we do not care about the other case).

4.3 An example

Two processors exist both in the cortex and the thalamus. The Input Gating Node (IGN) expresses the area's neuronal reactivity to the stimulus. It may be considered as the abstraction, in terms of pattern and average firing rate, of the activation of the area's pyramidal cells' somas.

$$\begin{cases} IGNc_M(t) = \underbrace{[OGNc_M(t-1) > a] \cdot [b \times f_{TPT}(STIM_T(t-2), IGNc_{TPT}(t-1)) \times STIM_M(t-2)]}_{M_1} + \\ \qquad \underbrace{c \times IGNc_M(t-1)}_{M_2} + d \times IN_M(t-1) + u_{IGNc} \\ IGNc_T(t) = \frac{M_1}{M_1+M_2} \times STIM_T(t-2) + \frac{M_2}{M_1+M_2} \times IGNc_T(t-1) \end{cases}$$

The Output Gating Node (OGN) sends information to the downstream areas. It represents, more or less, the integrated activity at the junction between the cells' somas and axons.

$$\begin{cases} OGNc_M(t) = [IGNc_M(t-1) > FTN_M(t-1) \wedge FTN_M(t-1) > 0] \cdot a \times IGNc_M(t-1) \\ \qquad + b \times OGNc_M(t-1) + u_{OGNc} \end{cases}$$

Three other processors are specific to the cortex. The Activation Node (AN) reflects the level of the cortex's metabolic activity, linked to the neuronal energy demand. The inhibitory node (IN) represents the integrated behavior of the GABA-neurons. The dynamic Firing Threshold Node (FTN) is modulated by the thalamus that can lower it.

$$\{FTN_M(t) = c - (b \times (c - FTN_M(t-1)) + a \times OGNt_M(t-1)) + u_{FTN}$$

In the visual cortex, as soon as the energy, at IGN, is greater than FTN, OGN transmits information to the thalamus. The two points are illustrated by the definition of the cortical input node *IGNc* (*Figure 1*).

Simulation results

In the reference experiment [Fox and Raichle, 1984], the stimuli are orange square-waves pulses of constant intensity and duration (5ms) that are presented during 40s scans (PET) at rates of 1, 3.9, 7.8, 15.5, 33.1 and 61 Hz. For the simulation, we suppose that the stimulus is deterministic

with a magnitude of 1 and the type "orange". The results are measures of the metabolic activation, i.e. measures, for each 40s-scan, of the regional cerebral blood flow variations (ΔrCBF%) in the visual cortex measure.

In the model, the time unit is 1ms. The summation over 40s of all the AN values is a measure of ΔrCBF%, once the brain's average activation level is set in the model, at its experimental value. *Figure 2* shows slightly better results for our model than for the CQN model [Pastor *et al.*, 2000], the main advantage being a better control of the divergence.

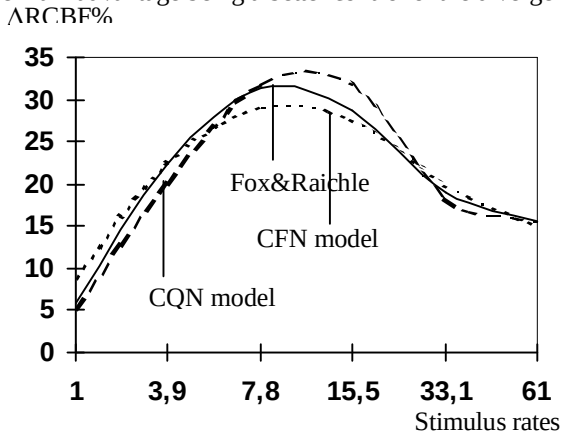


Figure 2. *Results of the simulations (mean values)*

5 Conclusion

Modeling large-scale cerebral networks, so that new evidences can be incorporated in the model and hypotheses can be assessed, is still a challenge. In the paper, we went through two major steps. The most important result is that causal functional networks are the best approach to cerebral modeling, since they fulfill theoretically all the requirements. Moreover, in the brief description of our modeling approach, we showed the flexibility and the adaptability of the formalism. Then, we proved that this formalism is really applicable, describing an example of cerebral model. With this model, we managed to approach experimental data, and furthermore we obtained (slightly) better results than with our previous CQN formalism.

The next steps will be on one hand to deepen the theoretical aspects of our modeling approach, and on another hand to assess the model by comparing simulation results to new experiments, involving more complex large-scale networks and a better temporal definition.

References

[Alexander *et al.*, 1992] G.E. Alexander, M.R. DeLong, M.D. Crutcher, Do cortical and basal ganglionic motor areas use "motor programs" to control movement?, *Behavioral and Brain Sciences*, 15:656-665, 1992.

[Bressler, 1995] S.L. Bressler, Large-scale cortical networks and cognition, *Brain Res. Rev.*, 20:288-304, 1995.

[Büchel and Friston, 97] C. Büchel, K.J. Friston, Modulation of connectivity in visual pathways by attention: cortical interactions evaluated with structural equation modeling and fMRI, *Cerebral cortex*, 7:768-778, 1997.

[Dean and Kanawaza, 1989] T. Dean and K. Kanawaza, A model for reasoning about persistence and causation, *Comput. Intell.*, 5:142-150, 1989.

[de Kleer, 1977] J. de Kleer. Multiple representations of knowledge in a mechanics problem-solver, In *Proceedings of IJCAI'77*, 299-304, 1977.

[Druzdzel and Simon, 1993] M.J. Druzdzel and H.A. Simon, Causality in bayesian belief networks, In *Proceedings of UAI'93*, 1993.

[Dubois and Prades, 1994] D. Dubois, H. Prade, Ensembles flous et théorie des possibilités : notions de base, In *Logique Floue*, Ed. Masson, 1994.

[Fox and Raichle, 1984] P.T. Fox, M.E. Raichle, Stimulus rate dependence of regional cerebral blood flow in human striate cortex, demonstrated by positron emission tomography, *J. Neurophysiol.*, 51(5): 1109-1120, 1984.

[Ghahramani, 1998] Z. Ghahramani, Learning dynamic bayesian networks, In *C.L. Giles and M. Gori (eds.)*, Adaptive Processing of Sequences and Data Structures. *Lecture Notes in Artificial Intelligence*, Springer-Verlag, Berlin, 168-197, 1998.

[Grossberg *et al.*, 1997] S. Grossberg, K. Roberts, M. Aguilar, D. Bullock, A neural model of multimodal adaptative saccadic eye movement control by superior colliculus, *J Neurosci*, 17(24):9706-9725, 1997.

[Herbster *et al.*, 1996] A.N. Herbster, T. Nichols, M.B. Wiseman, M.A. Mintun, S.T. DeKosky, J.T. Becker, Functional connectivity in auditory-verbal short-term memory in Alzheimer's disease, *Neuroimage* 4:67-77, 1996.

[Hume, 1740] D. Hume, Treatise of Human Nature, 1740.

[Lafon *et al.*, 1999] M. Lafon, L. Travé-Massuyès and J. Pastor, Hierarchical causal modeling of cerebral information propagation mechanisms, In *Workshop KRR-4*, IJCAI, 1999.

[Leiner and Leiner, 1997] H.C. Leiner and A.L. Leiner, How fibers subserve computing capabilities : similarities between brains and machines, I*nternational Review Of Neurobiology*, 41:535-553, 1997.

[Pastor *et. al*, 2000] J.Pastor, M.Lafon, L.Travé-Massuyès, J.-F.Démonet, B.Doyon and P.Celsis, Information processing in large-scale cerebral networks : the causal connectivity approach, *Biol. Cyb.*, 82:49-59, 2000.

[Pearl, 2000] J. Pearl, *Causality*, Cambridge University Press, 2000.

[Struss, 1990] P. Struss, Problems of interval-Based Qualitative Reasoning, in *Qualitative Reasoning about physical systems*, Eds Weld & de Kleer, 288-305, 1990.

Debugging medical Bayesian networks with Elvira's explanation facility

Carmen Lacave
Dept. Computer Science
University of Castilla-La Mancha
13071 Ciudad Real, Spain
E-mail: clacave@inf-cr.uclm.es

Agnieszka Oniśko
Institute of Computer Science
Białystok University of Technology
Białystok, 15-351, Poland
E-mail: aonisko@ii.pb.bialystok.pl

Francisco J. Díez
Dept. Artificial Intelligence
UNED
28040 Madrid, Spain
E-mail: fjdiez@dia.uned.es

Abstract

When the structure or the parameters of a Bayesian network are obtained from subjective estimates, debugging the network is one of the essential phases of knowledge elicitation. However, given that probabilistic inference is quite different from human reasoning, the identification of the elements that need to be fixed is a very difficult task. In this paper we show that the use of an explanation facility can significantly contribute to the process of refining a Bayesian network.

1 Introduction

1.1 Bayesian networks

Bayesian Networks (BNs) provide a way to build expert systems by using probability as a measure of uncertainty. A Bayesian network consists of an acyclic directed graph (ADG), whose nodes represent random variables, together with a probability distribution over its variables that satisfies the d-separation property [5]. This property implies that the joint probability distribution can be factored as the product of the probability of each node conditioned on its parents:

$$P(x_1, ..., x_n) = \prod_i P(x_i|pa(x_i)) \qquad (1)$$

In case of discrete variables, $P(x_i|pa(x_i))$ takes the form of a conditional probability table (CPT).

A *finding* is a piece of information that states with certainty the value of a random variable. A finding may be, for example, the assertion that a patient is a male; other findings might be that he is 54 year old, that he presents with fever, that he does not usually have headaches, etc. Each set of findings constitutes an *evidence case*. In medicine, an evidence case is typically a set of symptoms, signs, complementary test, antecedents, etc. collected for a certain patient at a given moment of the consultation. Diagnosis by probabilistic reasoning consists in computing the posterior probability of the unobserved variables given the available evidence.

1.2 Building Bayesian networks

There are basically two ways of building a Bayesian network. The *automatic* process consists in taking a database and applying one of the many algorithms that yield both the structure and the conditional probabilities. The *manual* process consists of two stages: (1) building the structure of the network with the help of a human expert, by selecting the variables and drawing causal links among nodes, and then (2) introducing the corresponding conditional probabilities—CPTs in the case of discrete variables. Ideally, those CPTs should be obtained from objective data, such as databases or epidemiological studies. However, in practice the lack of objective data often forces the knowledge engineer to obtain the CPTs from human experts' estimations. This task is difficult, time consuming and prone to errors and biases. For this reason, debugging is an intrinsic phase of probability elicitation. Debugging is also necessary when the parameters are obtained from a database, not only in order to refine the structure of the network, but also for detecting wrong parameters, which can be due to several reasons: scarcity of cases in the database for certain conditional probabilities, missing data and selection biases.

In this debugging process it is specially difficult to identify the model parameters that must be adjusted in order to attain the correct diagnoses. This task requires an explanation facility that helps both the expert and the knowledge engineer trace and understand the propagation of evidence. Explanation of reasoning is also crucial for the acceptance of medical expert systems, but we are not going to address that issue in the current paper.

1.3 Explanation in Bayesian networks

There are several features that characterize an explanation facility. First of all, explanations may be *verbal* or *graphical*. It is possible to differentiate between *explanation of the model*, also called *static explanation*, which consists in showing the user in an intuitive fashion the information contained in the Bayesian network, and *explanation of the inference*, also called *dynamic explanation*, which consists in showing how

the evidence has led to the posterior probabilities and, consequently, to certain diagnoses. Additionally, there are two levels of explanation, *micro* and *macro*; the former tries to justify the variation of the probability of a certain node; in contrast, explanation at the macro level analyzes the main lines of reasoning leading to the conclusions. (See [2] for a detailed study of those features and a review of explanation methods for Bayesian networks.)

1.4 Elvira

Elvira is an environment for the edition and evaluation of Bayesian networks and influence diagrams, developed as a research project of several Spanish universities. The software package includes a parser for reading networks in its own Elvira format and in HUGIN format. It also contains a graphical interface for editing networks, with specific options for canonical models (OR, AND, MAX, etc.), exact and approximate algorithms for both discrete and continuous variables, explanation facilities, learning methods for building networks from databases, algorithms for fusing networks, etc. Although some of the algorithms work with both discrete and continuous variables, the interface and the explanation capability assume that all the variables are discrete. Elvira is implemented in Java, so that it can run under different platforms. In the near future the program, with its source code, will be made publicly available on Internet.

Next section offers an overview of Elvira's explanation facility and the following one describes how to use it for debugging medical Bayesian networks, by using HEPAR II as an example.

2 Explanation in Elvira

Elvira has three main modes: *edition* (for "manually" building and modifying Bayesian networks and influence diagrams), *learning* (for building networks from databases) and *inference* (for propagating evidence). Most of explanation capabilities are offered in this mode. The explanation capability of Elvira is based on a system of windows and menus. It offers verbal and graphical explanations at the micro level, such as information about specific nodes or links and it is capable also to give a verbal explanation of the model, although in this paper we only discuss graphical explanations.

2.1 Sign of influences

One of Elvira's explanation options, available in both edition and inference modes, consists in automatically coloring the links of the network, in order to offer qualitative insight about the conditional probability tables. More specifically, given two ordinal discrete variables A and C such that there is a link $A \to C$, this link is said to be **positive** if higher values of A lead to higher values of C for any configuration of B, where B represents the set of other parents of C:

$$a_i > a_j \Rightarrow Dist(C|a_i, b) > Dist(C|a_j, b) \qquad (2)$$

The comparison of probability distributions is defined by

$$\begin{aligned} &Dist(C|a_i, b) > Dist(C|a_j, b) \\ &\quad \Longleftrightarrow \{[\forall c,\ P(C \geq c|a_i, b) \geq P(C \geq c|a_j, b)] \\ &\qquad \wedge [\exists c,\ P(C \geq c|a_i, b) > P(C \geq c|a_j, b)]\} \end{aligned} \qquad (3)$$

i.e., the probability distribution of C given a_i is higher than that given a_j if the cummulative probability is greater at least for a certain value c and not smaller for the other values of C. The definition of **negative link** and **null link** are analogous. When the influence is not positive nor negative or null, then it is said to be **unknown**. In Elvira these four kinds of links are colored in red, blue, black or purple, respectively.

Typical orderings of values of a variable are $+a > \neg a$, *present* > *absent*, *severe* > *moderate* > *mild* > *absent*, *positive* > *negative*, etc. If A and C are binary variables, the above definition implies that link $A \to C$ is positive if and only if $P(+c|+a, b) > P(+c|\neg a, b)$. If variable A represents a cause or a risk factor for C, or C is a test that detects A, then influence $A \to C$ is in general positive. In causal networks, most of the links are positive.

2.2 Management of cases

Elvira differs from other Bayesian-network tools in its ability to simultaneously handling several evidence cases. At each moment there is one *current case*, such that all the findings introduced by the user are added to this case, until he decides to generate a new case or return to one of the previous cases.

Figure 1 presents the HEPAR II network in Elvira. There were three evidence cases entered into the model: the first one had no evidence; the second one contained two findings, itching and hepatomegaly (enlarged liver); and the third one, which was the current case when the screen was captured, contained one more observation, an increased value of cholesterol. These three nodes are expanded (in the next section we explain that expanded nodes display a probability bar for each value and each evidence case) in order to show that the probability of itching and hepatomegaly is 1 in the second and third evidence cases, and also the probability of a certain increased value of cholesterol is 1 in the third case.

The second toolbar in Elvira's main window is the *explanation bar*. A text field on this bar displays the name of the current case ("Cholesterol", in this example). The four buttons around it allow the user to navigate across the set of evidence cases. This bar also contains widgets for setting the *expansion threshold* (see next section), saving evidence cases in files, generating new cases, expanding or contracting the selected nodes, modifying the inference options, etc. One of the buttons opens a *monitor of cases* that allows the user to select the cases to be displayed (by means of a checkbox), to add or remove evidence cases, to assign names and colors to cases, etc. In the same way, another button opens the *editor of cases*, which

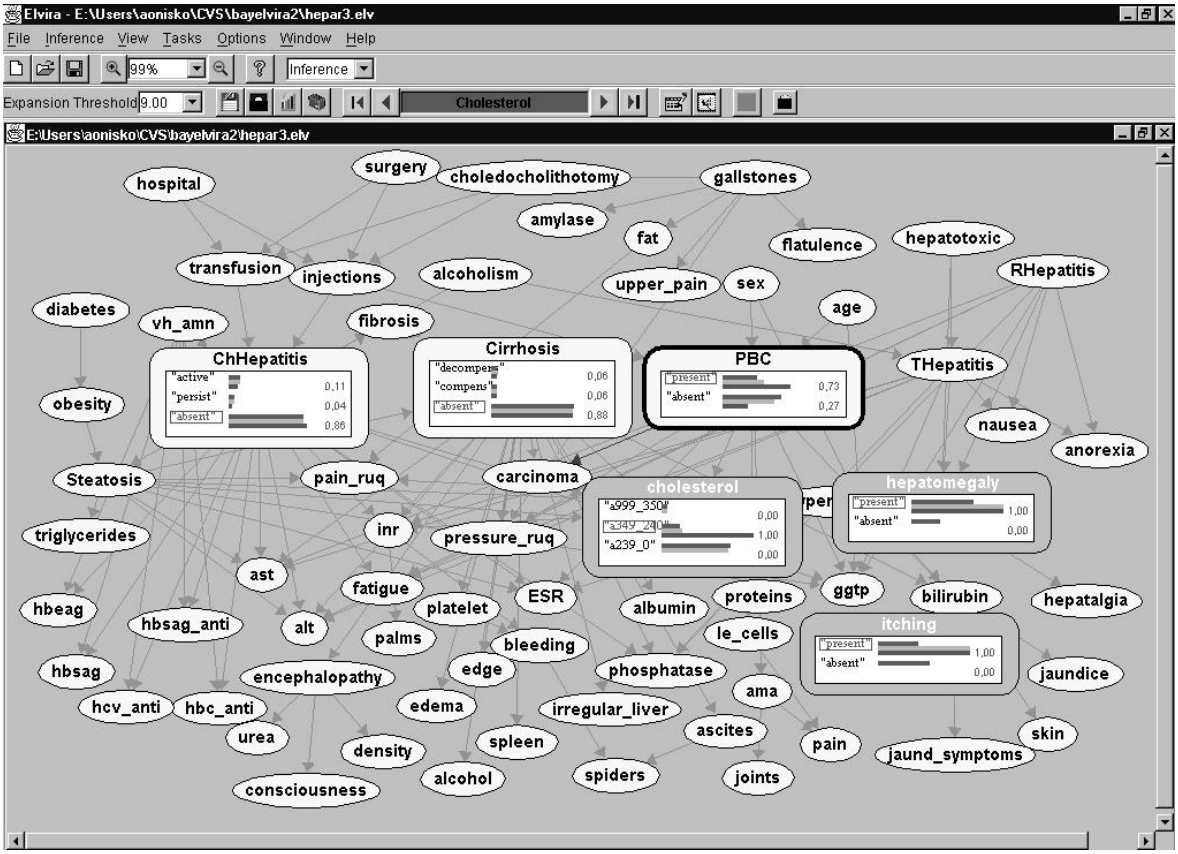


Figure 1: Elvira main window in inference mode

permits to introduce or remove findings from an evidence case. The main utility of this editor is to enter or remove evidence when the number of variables is so high that it becomes impossible or cumbersome to display the whole network on a screen. A more detailed description of the monitor and editor of cases can be found in [1].

2.3 Graphical display of probabilities

In Elvira, each node has an *importance factor* subjectively assigned by the human expert when defining the node properties. The importance factor for the main diseases is 10; intermediate anomalies and findings have lower assignments; its default value is 7.0. When switching from edition to inference mode, the nodes whose importance factor is greater or equal than the *expansion threshold* are automatically expanded. In Figure 1, only 6 of the 71 nodes are expanded; three of them represent the findings of the current evidence case (itching, hepatomegaly and cholesterol) and the other three capture liver disorders (chronic hepatitis, cirrhosis and PBC).

Expanded nodes contain a line for each value/state, which displays its name, a bar proportional to its probability and the numerical value of its probability. In this sense, Elvira is similar to other tools, such as HUGIN and Netica, but differs from them in its ability to displaying several probability bars, one for each evidence case, although only the numerical probability corresponding to the current case is displayed. Additionally, the most probable value (for the current case) is highlighted by a surrounding rectangle. In Figure 1, there are three bars per value/state, corresponding to the three cases mentioned above; since the first case contains no evidence, the upper bar represents the prior probability.

A finding (an observation) can be introduced by double-clicking on the corresponding line of the expanded node or by a contextual menu if the node is contracted. The background color of observed nodes automatically changes to gray, so that the user can easily identify the evidence of the current case.

2.4 Changes in posterior probability

When evaluating the impact of evidence, it is useful to know if the probability of a certain node has increased or decreased. In Elvira, a node whose probability has not changed keeps the yellow color of edition mode; if the probability has increased, according to the definition given in (3), the node is colored in red; if it has decreased, in blue; and if the change is neither positive nor negative, the node is colored in purple.

(Please note that the sign of influences only depends on the CPTs, while the changes in probability also depend on the propagation of evidence; therefore, the coloring of nodes only makes sense in inference mode, while the coloring of links can be in both edition and inference mode.)

The explanation-options window allows the user to decide whether the current probability distribution of a node is compared to the prior probability, the probability of the previous case or the probability of a fixed case. When doing hypothetical reasoning, for instance, when trying to determine how different levels of cholesterol will affect the posterior probabilities of other variables, the user should create an evidence case for each possible finding (each level of cholesterol) and compare the posterior probability of each case to the prior probability. In contrast, if the user wishes to observe how the probability distributions evolve when new findings are entered, he should select the option "compare to the previous case" and generate a new case for each finding.

3 Debugging the HEPAR II model

3.1 The HEPAR project

HEPAR II [3] is an expert system for the diagnosis of liver disorders, based on a Bayesian network that models a portion of the domain of hepatology. The structure of the model (i.e., the nodes of the graph along with links among them) was built with knowledge obtained from the medical literature and conversations with three domain experts. The most recent version of the model consists of 71 nodes (see Fig. 1): 9 disorder nodes (representing 11 different liver diseases), 18 risk factors, and 44 symptoms, signs, and laboratory tests results.

The numerical parameters of the model, i.e., the prior and conditional probability distributions, were extracted from HEPAR, a clinical database created in 1990. The data used to extract the numerical parameters contained 699 patient records. Simultaneously, we have built another version of HEPAR II based on the same structure but in which the numerical parameters were elicited directly from our expert.

A quantitative evaluation of HEPAR II [3; 4], previous to the use of Elvira, showed that there was still room for improving the diagnostic accuracy of both versions of the network. For this reason we decided to debug the Bayesian network by using Elvira's explanation facilities. The following sections describe our experience in the debugging of the database version of HEPAR II.

3.2 Analysis of negative influences

As mentioned above, in causal Bayesian networks most of the influences should be positive. However, the introduction of the database version of HEPAR II into Elvira showed that there was a high number of negative links. Therefore, the first step of our debugging process concentrated on analyzing one by one those influences in order to determine which of them were justified and which should be corrected.

For instance, although the expert asserted that *encephalopathy* is a symptom of *cirrhosis*, the link was negative in our Bayesian network. When looking at the data set more carefully, we realized that about 70% of patients having *encephalopathy* suffered from *PBC* (a type of liver cirrhosis) and only 13.5% of those patients presented with *cirrhosis*. The conclusion of this analysis was that we should draw a link from *PBC* to *encephalopathy*.

Similarly, the expert asserted that *injections* was a risk factor of *chronic hepatitis* and a crucial finding for its diagnosis. However, it turned out that around 63% of patients who had injections suffered from *PBC* and only 10% of patients presented with a *chronic hepatitis*. So, according to the database, we should also consider modelling *injections* as a risk factor of *PBC*.

We have also identified some negative influences in case of those variables whose lower values represent higher degrees of anomaly, such as *INR*, one of the laboratory tests. This intrinsic property of *INR* justifies the fact that the link from that node to symptom *liver palms* is negative.

Another negative link was found from *history of viral hepatitis*, that represents patient self-reported finding, to *presence of HCV antibodies in blood*. The explanation suggested by our expert is that patients who suffer from an asymptomatic viral hepatitis seldom know about the presence of the virus. Therefore, when using clinical data we must keep in mind that patient self-reported data may be unreliable.

In the network we model *sex* and *age* variables which are risk factors for some liver diseases. For example, 90% of patients suffering from *PBC* are middle-aged women. Patients presenting with *functional hyperbilirubinemia* are typically young men. However, since *sex* is not an ordinal variable, the sign of an influence relative to this variable is arbitrary, and should not be taken into account. In fact, in future versions of Elvira we will explicitly declare ordinal variables, so that links involving non-ordinal variables will not be colored, in order to avoid confusing the user.

The above-mentioned analysis refers us to the version of the HEPAR II in which the numerical parameters were learnt from the database. We have conducted a similar analysis for a model whose probabilities were elicited from the expert. However, in this network most of the links were positive. This can be explained by consistency in expert judgement, i.e., the structure was compatible with the elicited probabilities.

3.3 Analysis of patient cases

Further interactions with Elvira focused on the analysis of patient cases selected from the HEPAR database. We introduced the data into HEPAR II by splitting

the patient data into several evidence cases. The first evidence case consisted of self-reported data, such as symptoms and history of diseases. The second evidence case included the previous findings and those gathered by the doctor at the physical examination. Finally, the third evidence case added the results of laboratory tests.

Elvira's explanation facilities allowed us to compare the evolution of probability for each node after the addition of new findings (see Secs. 2.2 to 2.4). Our expert found the coloring of nodes especially useful for observing which variables were influenced positively or negatively by each set of findings, and for assessing the relevance of a certain variable by checking whether it has any impact on a particular diagnostic situation or not.

In particular, we noticed that in some situations a certain diagnosis was already apparent after observing a group of symptoms and signs and the laboratory tests did not change the first diagnosis.

4 Conclusions

The use of Elvira for debugging HEPAR II has shown that its explanation capability can offer significant insight for detecting the inconsistencies and inaccuracies of a Bayesian network.

With respect to *static explanation* (aka explanation of the model), the coloring of each link according to the sign of influence allows the user to observe at a glance the qualitative properties of the network. In causal models, most of the links are positive (red), because usually the presence of the cause increases the probability of the presence of the effect. For this reason, blue and purple links make the knowledge engineer suspect that the parameters introduced for a certain CPT may be wrong. In our analysis of HEPAR II with Elvira we realized that it was possible to find a justification for some of the negative links, while in other cases a reexamination of the data set, led us to modifying the structure and parameters of the network.

With respect to *dynamic explanation* (aka explanation of inference), in our evaluation of HEPAR II it has been very useful the facility of saving evidence cases in files, the possibility of working with several cases simultaneously and the fact that observed nodes are clearly identifiable for each case. By comparing the posterior probabilities resulting from different cases, it is possible to perform hypothetical reasoning, sometimes called "what-if" in the literature on expert systems, which allows human experts to predict the impact that would have observing each of the possible values of a certain variable. Analogously, given a certain evidence, it is possible to assess the diagnostic value of its components by assigning different cases for each finding or set of findings, either individually or incrementally.

A question that remains to be studied is how the refinements introduced in the Bayesian network after thorough debugging with Elvira contribute to improving the diagnostic accuracy of HEPAR II.

Finally, we would like to mention that the coloring of links and nodes is inspired in Wellman's work on *qualitative probabilistic networks* [6]. However, the fact that our networks contain numerical probabilities and that propagation of evidence is done by quantitative algorithms allows us to determine the sign of probability changes in many cases in which Wellman's algorithms would lead to unknown signs.

Acknowledgments

We would like to thank Dr. Diego Rodríguez Leal, from Hospital Alarcos in Ciudad Real, for his useful comments. We would also like to thank all the researchers participating in the Elvira and HEPAR II projects, specially Andrés Cano for his assistance with Elvira.

This work has been partially financed by the Spanish CICYT under project TIC-97-1135-C04. Our collaboration was enhanced by travel support from NATO Collaborative Linkage Grant number PST.CLG.976167.

References

[1] C. Lacave, R. Atienza, and F. J. Díez. Graphical explanations in Bayesian networks. In *Proceedings of the First International Symposium on Medical Data Analysis, (ISMDA-2000)*, pages 122–129, Frankfurt, Germany, 2000. Springer-Verlag, Heidelberg.

[2] C. Lacave and F. J. Díez. A review of explanation methods for Bayesian networks. Technical Report IA-00-01, Dpto. Inteligencia Artificial, UNED, Madrid, 2000.

[3] Agnieszka Oniśko, Marek J. Druzdzel, and Hanna Wasyluk. Extension of the Hepar II model to multiple-disorder diagnosis. In S.T. Wierzchoń M. Kłopotek, M. Michalewicz, editor, *Intelligent Information Systems,* Advances in Soft Computing Series, pages 303–313. Springer-Verlag, Heidelberg, 2000.

[4] Agnieszka Oniśko, Marek J. Druzdzel, and Hanna Wasyluk. Learning Bayesian network parameters from small data sets: Application of Noisy-OR gates. *International Journal of Approximate Reasoning*, 2001. To appear.

[5] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* Morgan Kaufmann Publishers, San Mateo, CA, 1988.

[6] M. Wellman. Fundamental concepts of qualitative probabilistic networks. *Artificial Intelligence*, 44:257–303, 1990.

A Bayesian Belief Network for Lower Back Pain Diagnosis

K.R. McNaught, S.L. Clifford and M.L. Vaughn,
Cranfield University, RMCS Shrivenham, Swindon, UK.
E-mail: K.R.McNaught@rmcs.cranfield.ac.uk

A.J.B. Fogg and M.A. Foy,
Princess Margaret Hospital, Swindon, UK.

Abstract

In this paper we discuss some ongoing work on the development of a Bayesian Belief Network to classify patients into one of three categories associated with lower back pain. A BBN with only thirteen nodes has been found to give classification accuracies in the region of 76% for new cases. This is comparable to the performance obtained by a neural network on the same data. However, the BBN requires less detailed information to be entered for each patient, a consideration if such a system is to be of use to GPs. The network was developed based on an analysis of 100 cases, and by utilising expert knowledge provided by two consultant orthopaedic surgeons. The network was then tested on a holdout sample.

1 Introduction

Low back pain (LBP) is a common ailment, estimated to affect in the region of 70% of the population at some stage of their lives. Most recover in a matter of weeks, with only around 10% requiring investigation [Clinical Standards Advisory Group, 1994; Waddell, 1992]. This 10%, however, constitute a significant cost to health services in terms of resources, to industry in terms of lost productivity, and to society in terms of social security benefits. There is also the cost to the individuals themselves, of course, in terms of suffering and impaired quality of life. UK government statistics have estimated the number of working days lost as a result of LBP problems to be in the region of 100 million [Clinical Standards Advisory Group, 1994]. The financial costs run into billions of pounds.

There is little doubt about the magnitude of the LBP problem. Yet successful diagnosis remains difficult, as pointed out by several authors [Nachemson, 1992; Bigos et al, 1994]. This was the backdrop to the application of artificial neural networks for LBP diagnosis. Bounds et al [1988, 1990] developed a neural network for LBP with a higher diagnostic accuracy than that achieved by clinicians.

Later, Vaughn et al [1999] also developed a multi-layer perceptron neural network for this purpose, collecting a dedicated data set for the study. It is that data set which is described and employed in this paper.

Vaughn et al [1999] were not only concerned with diagnostic accuracy, but with explanation of the neural network's outputs. However, rather than attempting to induce general rules from the neural network, as described by Fu [1994], for example, their approach is concerned with explaining the neural network's outputs on a case by case basis. Recent work is described in Vaughn et al [2000, 2001]. While such work is valuable, it is to be expected that concerns will remain in some quarters about utilising output from neural networks in clinical settings. That clinicians may be reluctant to employ methods they do not fully understand or which do not leave a clear audit trail is quite understandable and has been alluded to elsewhere [de Dombal et al, 1997]. The impact of the trend towards increasing litigation in health services also remains to be seen. The more methods which are available to doctors, however, the greater is the chance that they will find one which they feel comfortable with. The purpose of this study is therefore to investigate the potential effectiveness of BBNs in this domain, since their explanations are arguably more transparent, and their reasoning is based on classical probability theory which some clinicians may be familiar with. Jensen [1996] provides a good introduction.

2 Classification of LBP Patients

Although there are many more possible clinical classifications of LBP patients, the study by Vaughn et al [1999] employed three mutually exclusive classes: Simple Lower Back Pain (SLBP), Root Pain (ROOTP) and Abnormal Illness Behaviour (AIB). Since this study is

using the same data set, the same classes of patient are used here.

SLBP refers to mechanical lower back pain, minor scoliosis and old spinal fractures. ROOTP involves nerve root compression due to either disc, bony entrapment or adhesions. AIB also features mechanical low back pain, degenerative disc or bony changes, with symptoms magnified, possibly as a result of distress in response to chronic pain. AIB has rather more psychological symptoms associated with it than the other two classes.

3 The Data

Most of the data were collected by a research physiotherapist**,** with the remaining attributes being provided by the patient's orthopaedic consultant, including the patient's classification as SLBP, ROOTP or AIB. The attributes collected from the patients by the physiotherapist included the following: Age, Gender, Duration of Pain (acute, recurring or chronic), Pain Began (suddenly or gradually), Any Leg Pain?, Which Pain is Worse? (back, leg or equal)**,** What is the Pain Aggravated by?(e.g. coughing, standing, sitting, bending), Is the Pain Worse in the Morning?, Is there Night Pain?, Any Weight Loss?, Any Previous Spinal Operations?, Use of Walking Aids?, Does the Case Involve Litigation?, Smoker?, Unemployed for more than two years?, Claiming Invalidity Benefit?

The attributes collected by the orthopaedic consultant included: Lumbar Flexion (<30deg, 30-45deg**,** >45deg), Lumbar Extension (<5deg, 5-15deg , >15deg), Catch on Extension, Straight Leg Raise (<45deg, 45-70deg , >70deg) for each leg, Raise Limited by (back pain, leg pain, hamstrings, not limited) for each leg, Any Cross-Leg Pain?, Any Neurological Signs? (Motor Loss, Sensory Loss, Loss of Reflexes), Any Nerve Involvement? (1 or 2 nerve roots, multiple nerve roots, none), Neurone Pattern (typical UMN, typical LMN, not typical), Inappropriate Signs, ODI, MSPQ, Zung and DRAM.

Some of the latter attributes are derived from questionnaires completed by the patient, e.g. the ODI(Oswestry Disability Index) and the Zung score. It was noticed that the Zung and DRAM scores were in one-to-one correspondence, and so the DRAM attribute was considered redundant. These attributes are related to the patient's psychological state. Inappropriate Signs refers to observations made by the consultant about the patient's condition. If the patient answers certain questions or responds to certain prompts in a contradictory way, often exaggerating their condition, then they are said to exhibit inappropriate signs. The number of these is recorded by the consultant.

All of the attributes are represented by discrete variables, and many of these are binary. For example, 'Age' is split into 'Over 55' and 'Under 55'. This follows the format of the data employed in the neural network study by Vaughn et al [1999].

4 Development of the BBN

The BBN was developed by examining the data for attributes which appeared to discriminate between the different possible classes of patient, and then arranging these in a logical structure, consistent with the perceived characteristics of LBP. Potentially discriminating variables were identified by a simple inspection of their conditional distributions. It was not felt necessary to apply formal statistical tests for association since the final list of variables and their relationships would be subject to the domain experts' validation in any case, and indeed some variables were rejected by the consultants as they did not believe them to be generally useful indicators. Some attributes were discarded for having insufficient cases associated with them. Some others, e.g. 'Unemployed for more than two years?' , 'Does the Case Involve Litigation?' and 'Claiming Invalidity Benefit?' were considered to be not only potentially offensive to new patients, but questionable as predictive attributes for LBP which might be included in a medical decision support tool. While there is certainly evidence from the data set that social factors such as these are more associated with some conditions than others, we were not comfortable with them being used to aid a medical diagnosis, even if their inclusion might have raised the classification accuracy achieved. Should a doctor conducting a diagnosis believe that some social aspects of a particular patient's background might be relevant to their condition, they could, of course, take due account of that information separately.

The structure of the LBP BBN is close to that of a naive Bayesian classifier [Friedman et al, 1997; Domingos and Pazzani, 1997]. The hypothesis node 'Category' consists of the three states SLBP, ROOTP and AIB, and most of the other nodes depend only on it. However, 'Use of Walking Aids?' depends on both the LBP classification and 'Age'. It was initially thought likely that some of the variables would remain dependent even when conditioned on the hypothesis node, e.g. some of the variables linked to the patient's psychological state. Interestingly, however, this was not the case and even these variables were conditionally independent given the LBP classification, i.e. the joint distribution of the two variables conditioned on LBP classification is virtually identical to the product of the two individually conditioned distributions.

The 'Straight Leg Raise' node is a deterministic node, representing the maximum raise over the right and left legs, which are dependent. The conditional probability tables and the prior distribution for the hypothesis node were obtained from the training set of 100 cases, and

verified as reasonable by the orthopaedic consultants. In most cases, we have simply used the relative frequencies in the training sample as estimates of the required probabilities. The only exceptions have been to avoid values of zero which are inevitable with small samples. In such cases, we have substituted an appropriately small value such as 0.05, and adjusted the other probabilities in the table accordingly.

After the construction of several candidate networks, the one that was found to give the highest classification

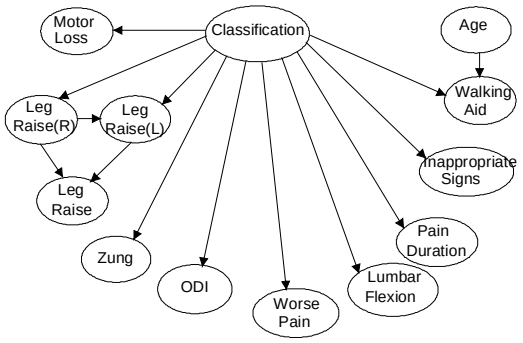


Figure 1. A BBN for Lower Back Pain Classification.

accuracy with the training set is shown in Figure 1. Some of the other networks considered were only one or two percent poorer at classifying the training set. Variables used in these networks but not in the one above, included MSPQ and Nerve Involvement. Adding these variables to the final network did not improve the classification accuracy, however, and so they were omitted.

5 Results

The above BBN correctly classified 75 new cases out of 99 in the test set. This is similar to the classification rate achieved in the training set. The confusion matrix is shown in Table 1.

		Predicted Class		
		SLBP	ROOTP	AIB
True Class	SLBP	29	4	0
	ROOTP	8	30	6
	AIB	3	3	16

Table 1. Confusion matrix for the current BBN.

As the confusion matrix shows, the performance is not too uneven even over the three classes, with the network correctly classifying 88% of SLBP cases, 68% of ROOTP cases and 73% of AIB cases. The quadratic loss score is 0.3814. Furthermore, the probability estimates of the network are well calibrated, i.e. the model's output probabilities are consistent with observation. This lends credibility to its outputs and should give confidence to it's users that the probabilities generated are meaningful. It is also an indication that important dependencies have not been omitted. As Nikovski [2000] has pointed out, failing to represent explicit dependencies between variables is likely to result in over-confident model outputs. Poor calibration between predicted and realised probabilities would then follow.

The BBN in Figure 1 requires discrete inputs for each of the following variables for each case: Age, Straight Leg Raise (Right), Straight Leg Raise (Left), Which Pain is Worse?, Pain Duration, Lumbar Flexion, Uses Walking Aid?, ODI Score, Inappropriate Signs, Zung Score and Neurological Signs (Motor Loss). Age, ODI Score, Uses Walking Aid?, Inappropriate Signs and Motor Loss are all represented as binary variables. Straight Leg Raise (Right), Straight Leg Raise (Left), Lumbar Flexion, Pain Duration, Which Pain is Worse? and Zung Score are all represented as ternary variables. Pain Duration has the three states – 'Acute', 'Recurring' and 'Chronic', while Which Pain is Worse? has the states 'Back', 'Leg' and 'Equal'.

6 Conclusion

The 'true classifications' are themselves not clear cut. Indeed, some of the original cases were reclassified in the light of the results from the neural network study. Most of the reclassifications are now considered AIB cases, a category noted as difficult to diagnose by other studies [Waddell et al, 1984]. This highlights the difficulty of accurate diagnosis in this area, and strongly supports the claims of AI and related approaches to be able to provide a useful second opinion to consultants at low cost. This is one very useful role which such systems can provide. Another is as a 'first opinion' at health centres or General Practitioners' surgeries, where expert human diagnosis is unlikely to be available. A diagnostic tool such as this could act as a filter, separating the more clear cut cases from the more difficult, thus helping to ensure that each patient is referred to the correct or most appropriate specialist for their likely condition. Any mis-diagnosis should obviously be picked up by the specialists at this point. However, such an initial screening should help to speed up the whole process of referring patients to the appropriate consultants, providing benefits both to patients and to hard-pressed health service resources since there should be less 'passing on' of patients between consultants.

The simple BBN developed in this study has demonstrated a relatively high classification accuracy for

this domain. Although around 5% poorer than an MLP neural network developed on the same data, less information is required by the BBN. In particular, no intrusive questions about the patients' social backgrounds have to be posed. Furthermore, the output probabilities from a BBN are more naturally interpretable than the output activations usually are from a neural network. Another advantage of the BBN representation is that it is easily incorporated within an influence diagram [Oliver and Smith, 1990] to permit decision analytic modelling. This facilitates an expected utility approach, within which misclassification costs and risk can be incorporated. In turn, this provides the doctor and patient with a more sophisticated support tool to help in determining the most appropriate treatment regime for a patient. Such modelling is becoming increasingly popular in the medical domain.

This work is very much in its infancy, and we hope in time to explore the available data further, and find ways of improving the model. We also hope that more data will be collected in order to allow further development and testing. We have not yet attempted to create a model directly from the data, as described by Heckerman [1997], for example, but that may provide an interesting comparison with the current model and generate further ideas for improvement.

We have also yet to conduct a sensitivity analysis of the network. Furthermore, while we have noted that the neural network developed from the same data uses more information in the course of achieving a slightly higher classification accuracy, we do not yet know how well a neural network might perform with the same subset of data employed in the BBN, or with another subset of a similar size. There remains much scope for comparison between the two approaches and possibly for the development of a hybrid classifier.

References

[Bigos et al., 1994] Bigos, S., Bowyer, O. and Braen, G. Acute low back problems in adults. Clinical Practice Guideline 14, AHCPR Publication 95-0642, U.S. DHHS, 1994.

[Bounds et al., 1988] Bounds, D.G., Lloyd, P.J., Mathew, B.G. and Waddell, G. A multi-layer perceptron network for the diagnosis of low back pain. Proc. IEEE conference on Neural networks, San Diego, pp 481-489, 1988.

[Bounds et al., 1990] Bounds, D.G., Lloyd, P.J. and Mathew, B. A comparison of neural networks and other pattern recognition approaches to the diagnosis of low back disorders. *Neural Networks* **3**, 583-591, 1990.

[Clinical Standards Advisory Group, 1994] Clinical Standards Advisory Group: 'Report on Back Pain'. HMSO, London, 1994.

[de Dombal et al., 1997] de Dombal, F.T., Clamp, S.E. and Chan, M. Bayes' theorem and decision support in 'front line' clinical medicine. In: French and Smith(eds), *The Practice of Bayesian Analysis*. Arnold, London, 1997.

[Domingos and Pazzani, 1997] Domingos, P. and Pazzani, M. On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning* **29**, 103-130, 1997.

[Friedman et al., 1997] Friedman, N.I.R., Geiger, D. and Goldszmidt, M. Bayesian network classifiers. *Machine Learning* **29**, 131-163, 1997.

[Fu, 1994] Fu, L.M. Rule generation from neural networks. *IEEE Trans. on Systems, Man. And Cybernetics* **24** (8), 1114-1124, 1994.

[Heckerman, 1997] Heckerman, D. Bayesian networks for data mining. *Data Mining and Knowledge Discovery* **1**, 79-119, 1997.

[Jensen, 1996] Jensen, F.V. *An Introduction to Bayesian Networks*. UCL Press, London, 1996.

[Nachemson, 1992] Nachemson, A. Newest knowledge of low back pain - a critical look. *Clin. Orth. Rel. Res.* **279**, 8-19, 1992.

[Nikovski, 2000] Nikovski, D. Constructing Bayesian networks for medical diagnosis from incomplete and partially correct statistics. *IEEE Trans. on Knowledge and Data Engineering* **12**(4), 509-516, 2000.

[Oliver and Smith, 1990] Oliver, R.M. and Smith, J.Q.(eds). *Influence Diagrams, Belief Nets and Decision Analysis*. Wiley, Chichester, 1990.

[Vaughn et al., 1999] Vaughn, M.L., Cavill, S.J., Taylor, S.J., Foy, M.A. and Fogg, A.J.B. Direct knowledge discovery and interpretation from a multilayer perceptron network which performs low-back-pain classification. In: Bramer(ed.), *Knowledge Discovery and Data Mining*. IEE Press, pp 160-179, 1999.

[Vaughn et al., 2000] Vaughn, M.L., Cavill, S.J., Taylor, S.J., Foy, M.A. and Fogg, A.J.B. Direct explanations and knowledge extraction from a multilayer perceptron network that performs low back pain classification. In: Wermter and Sun(eds), *Hybrid Neural Systems (Lecture Notes in Computer Science; 1778: Lecture notes in*

artificial intelligence). Springer, Berlin, pp 270-283, 2000.

[Vaughn et al., 2001] Vaughn, M.L., Cavill, S.J., Taylor, S.J., Foy, M.A. and Fogg, A.J.B., 2001. Direct explanations for the development and use of a multi-layer perceptron network that classifies low-back-pain patients. *International Journal of Neural Systems* (awaiting publication).

[Waddell et al., 1984] Waddell, G., Bircher, M., Finlayson, D. and Main, C. Symptoms and signs: physical disease or illness behaviour. *BMJ* **289**, 739-741, 1984.

[Waddell, 1992] Waddell, G. Biophysical analysis of low-back pain, *Baillieres Clin. Rheu***. 6** (3), 523-557, 1992.

Evaluation of the HEPAR II System for Diagnosis of Liver Disorders

Agnieszka Oniśko
Białystok University of Technology
Institute of Computer Science
Białystok, 15–351, Poland
aonisko@ii.pb.bialystok.pl

Abstract

The last decade has seen a number of practical decision support systems based on the normative principles of probability theory and decision theory. While, on theoretical grounds, such systems can be expected to perform well and be useful in practice, there is still little empirical data that would validate this expectation. In fact, a skeptic might doubt the practical value of the normative approach on two grounds: (1) possibly inferior performance of normative models compared to humans, and (2) because of its formal reasoning approach and possibly counterintuitive results, users might reject a system's advice, even if it is correct.

This paper describes a study conducted in order to validate a normative decision support system in a practical setting. The study compares performance of HEPAR II, a medical system for diagnosis of liver disorders, with the performance of general practitioners on ten randomly selected patient cases with histopathologically confirmed diagnosis. The results are encouraging: HEPAR II's diagnostic accuracy was 40% higher than the best of the physicians'. The study also tests the effect of system's suggestions on the ultimate diagnosis indicated by the physicians. Here the results are encouraging as well: system's advice doubled the accuracy of physicians.

1 Introduction

The last decade has seen considerable progress in the field of decision-theoretic systems, including several practical decision support systems based on the normative principles of probability theory and decision theory. Decision support is one area where computer-based systems can make a tangible difference. In the field of medicine, for example, where costs of making errors are high, decision support plays a particularly important role. For example, in the domain of hepatology, inexperienced clinicians have been found to make a correct diagnosis in jaundiced patients in less than 45% of the cases [15]. Computer-based decision support systems have the potential for improving the quality of diagnosis and, effectively saving lives.

Critics of the normative approach might doubt its value on two grounds: (1) possibly inferior performance of normative models compared to humans, and (2) because of its formal reasoning approach and possibly counterintuitive results, users might reject a system's advice, even if it is correct. While several studies have shown excellent performance of normative systems, it is not unusual to encounter criticism. And so, in the earlier joint work on the HEPAR II system we were criticized for poor performance of our diagnostic model of liver disorders based on Bayesian networks (the diagnostic accuracy of the HEPAR II system was around 49%). As far as user acceptance of a systems advice, it is theoretically possible and, to my knowledge, never tested empirically that the combined performance of a user and a decision support system becomes poorer than the user's unaided performance, even if the system by itself performs significantly better than an unaided user [7]. Although numerous medical decision support systems have been developed to date, the clinical use of such systems has been limited. They have seldom undergone a clinical evaluation. Whether a computer system is useful in practice, even if it performs well overall, remains often an open issue.

There have been several evaluation studies of medical systems performed in the past (e.g., [1; 2; 6; 8; 16; 19]). The most well known study concerned the evaluation of MYCIN [20], a system for assisting the diagnosis and treatment of patients with infectious diseases. The evaluation of MYCIN focused on a quantitative measure of its diagnostic accuracy and its qualitative impact on decisions made by physicians. It was a blinded evaluation in which the performance of MYCIN and the performance of clinicians were assessed by independent experts who did not know the identity of the prescribers. A study on INTERNIST-I [10], a computer program assisting diagnosis in internal medicine, compared diagnoses made by the system to those made by human ex-

perts (hospital clinicians and discussants) as well as the evaluation of capabilities of the system. The results were encouraging: INTERNIST-I performed at clinicians level and only slightly worse than the discussants.

This paper describes an evaluation of a computer-based system for diagnosis of liver disorders, HEPAR II [11; 12]. The HEPAR II system is based on a Bayesian network model of a subset of the domain of hepatology. The structure of the network has been elicited from an expert diagnostician and the parameters have been learned from a clinical database. The most recent version of HEPAR II models 11 liver disorders, 18 risk factors, and 44 symptoms and laboratory tests results. A quantitative evaluation of the HEPAR II model, consisting of testing its diagnostic accuracy, has been performed previously [11; 12]. In those studies, the answers of the system were compared to the final diagnosis captured in the HEPAR database. This paper focuses on comparing the performance of HEPAR II to the diagnostic accuracy of physicians. Additionally, it also addresses the issue of the system's impact on physicians decisions and the resulting combined performance of the system-user team.

The remainder of this paper is structured as follows. Section 2 describes briefly the HEPAR II model. Section 3 presents the design details of the study. Section 4 reports experimental results of the evaluation. Finally, Section 5 discusses general issues related to the performed study and directions for further work.

2 The HEPAR II Model

Support of a diagnosis in the management of liver disorders has been the focus of a number of research projects in Artificial Intelligence (e.g., [1; 9; 14]). The work on the HEPAR II model is a continuation of the HEPAR project [3; 18]. The HEPAR system was designed for gathering and processing clinical data on patients with liver disorders and aimed at reducing the need for hepatic biopsy. An integral part of the HEPAR system is its database, created in 1990 and thoroughly maintained since then at the Gastroenterological 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 patient 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. The HEPAR II project, which is a collaboration between Białystok University of Technology, Medical Center of Postgraduate Education, and the University of Pittsburgh, is an attempt to address the same problem using decision-theoretic methods. The modeling tool chosen for the HEPAR II project are Bayesian networks.

The HEPAR II model, a simplified fragment of which is shown in Figure 1, is a Bayesian network consisting of 71 nodes. The structure of the model, (i.e., the nodes of the graph along with arcs among them) was built based on medical literature and conversations with our domain expert, a hepatologist Dr. Hanna Wasyluk and two American experts, a pathologist, Dr. Daniel Schwartz, and a specialist in infectious diseases, Dr. John N. Dowling, from the University of Pittsburgh. The elicitation of the structure took approximately 50 hours of interviews with the experts, of which roughly 40 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 most recent version of the model, consists of 9 disorder nodes representing 11 different liver diseases and 62 feature nodes encoding medical findings such as patient self-reported data, signs, symptoms and laboratory tests results.

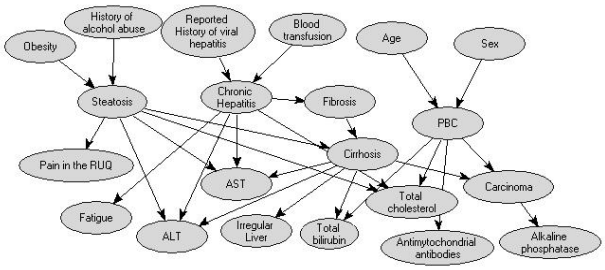


Figure 1: A simplified fragment of the HEPAR II network

The numerical parameters of the model, i.e., the prior and conditional probability distributions, were extracted from the HEPAR database. The data used to extract the numerical parameters contained 699 patient records. All continuous variables were discretized by our expert. In dealing with missing values, we followed the suggestions reported by Peot and Shachter [13] that missing values in medical data sets are not missing at random and are either indications of normal or less severe symptoms. In other words, if a symptom is absent, there is a high chance that it is not reported, i.e., missing from the patient record. And conversely, a missing value suggests that the symptom was absent. In learning the model parameters, missing values for discrete finding variables were assigned to state *absent* (e.g., a missing value for *Jaundice* was interpreted as *absent*). In case of continuous variables, a missing value was assigned a normal value, elicited from the expert as the typical value for a healthy patient (e.g., a missing value for *Bilirubin* was interpreted as being in the range of 0–1 *mg/dl*).

Given a patient case, i.e., values of some of the modeled variables, such as symptoms or test results, the system computes the posterior probability distribution over the possible liver disorders. This probability distribution can be directly used in diagnostic decision support.

3 Experimental design

The experiment was conducted in the Center for Medical Postgraduate Education, Warsaw, Poland.

Participants: 19 internists and pediatricians from various medical centers in Poland. 13 physicians were beginning practitioners with clinical experience ranging between one and two years, and 6 physicians were pediatricians with the clinical experience ranging between 10 and 30 years. All participants were general medicine fellows (primary health-care speciality). None of them had participated in the development of the HEPAR II system.

Patient cases: 10 patient cases selected randomly without replacement from the HEPAR database.

Measurement: Diagnostic accuracy.

Study details: The physicians participating in the experiment received printouts with descriptions of patient cases (every physician received the same 10 patient descriptions). Each patient case was outlined in a story-like-form that reported various findings such as symptoms, signs, and laboratory tests results with the biopsy data removed. The summary included an alphabetically ordered list of 11 hepatological diagnoses. The physicians were asked to indicate the four most likely diagnoses based on the observed findings. In case they did not find an appropriate diagnosis on the list, they were asked to write in additional diagnoses.

The second stage of the experiment focused on the impact of the HEPAR II model on its users. 15 of the 19 physicians participated in this experiment. After they had completed the first stage (i.e., diagnosing the ten patient cases), they attended a presentation of the HEPAR II system and were given the opportunity to interact with HEPAR II on a personal computer. Then they received the answers of the HEPAR II model, i.e., printouts consisting of lists of rated diagnoses for each patient case with a value of posterior probability for each diagnosis. After the physicians had seen the answer of the system they were given an opportunity to change their original answers. The whole experiment took approximately 1.5 hours.

4 Results

The experiment yielded a total of 187 diagnoses. This number captures 10 patient cases evaluated by 19 physicians (three of the diagnoses were accidentally missed by the subjects).

Diagnostic accuracy is defined as the proportion of patient cases that were diagnosed correctly among all patient cases. By the correct diagnosis is meant the histopathologically verified diagnosis included in the HEPAR database. There are two aspects of diagnostic accuracy that are of interest: (1) whether the most probable diagnosis indicated by the user or the HEPAR II model was indeed the correct diagnosis, and (2) whether the set of k most probable diagnoses contains the correct diagnosis for small values of k (following the previous evaluations of HEPAR II model, I chose a "window" of k=1, 2, 3, and 4). The design of the experiment and the instructions to the subjects asked for a rank-ordered list of four most likely diagnoses. However, in roughly 80% of the cases (150 diagnoses), the subjects indicated only one or two most probable diseases. The results presented in this paper are only for $window = 1$, i.e., for the most likely diagnosis.

Table 1: The diagnoses of the physicians and the HEPAR II system for each patient case

	1	2	3	4	5	6	7	8	9	10	%
1	–	–	–	–	–	C	–	–	–	C	20.0
2	–	C	C	C	–	C	–	–	–	–	40.0
3	C	–	C	C	–	–	–	–	–	–	30.0
4	C	–	C	C	–	–	–	–	–	–	30.0
5	C	–	C	C	–	–	C	–	–	–	40.0
6	–	C	C	C	–	–	C	C	–	–	50.0
7	C	C	C	C	–	*	–	–	–	–	44.4
8	C	–	C	C	–	–	–	–	–	–	30.0
9	C	C	C	C	–	C	–	–	–	–	50.0
10	–	C	–	C	–	C	C	–	–	–	40.0
11	–	C	C	C	–	–	–	C	–	–	40.0
12	C	–	C	–	–	*	–	C	–	–	33.3
13	C	–	C	C	–	C	–	–	–	–	40.0
14	–	–	C	–	C	–	C	–	–	–	30.0
15	–	–	C	–	–	–	–	–	–	–	10.0
16	*	–	C	–	–	–	–	–	–	–	11.1
17	C	C	C	–	–	–	C	–	–	–	40.0
18	–	–	C	–	–	–	–	–	–	–	10.0
19	–	–	C	–	–	–	–	C	–	–	20.0
%	55	37	89	58	5	29	26	21	0	5	32.1
H	C	–	C	C	C	–	C	C	–	C	70.0

4.1 Comparison of diagnostic accuracy

There was a significant difference between the diagnostic accuracy of the physicians and the accuracy of the HEPAR II system. The average accuracy over all patient cases was 32.1% and 70%[1] for the physicians and the system respectively. The hypothesis that the accuracy of physicians is the same as or better than HEPAR II's was rejected at $p < 0.02$ level (paired, one-sided Student-t test). Table 1 captures the results of the experiment. It presents correct and incorrect diagnoses made by the physicians and the HEPAR II system. Each row of this table corresponds to one of the 19 physicians and each column corresponds to one of the 10 patient cases. The symbols C, $-$, and $*$ stand for correct, incorrect, and missing diagnosis respectively. The next to last row presents the percentage of decisions that matched the correct diagnosis made by the physicians for each of 10 patient cases (i.e., the average diagnostic accuracy of the physicians per case). The last row captures the answer of the HEPAR II system. Finally, the last column presents the overall accuracy for each physician. None of the physician's accuracy was higher than the system's accuracy. The

[1] The overall diagnostic accuracy of the HEPAR II model on the selected 10 patient cases was higher than the accuracy for the set of 699 patients (70% compared to 49%). This can be attributed to a relatively small sample size (10 out of 699 cases). A larger sample size, which we considered, would have put an unacceptable burden on the physicians time

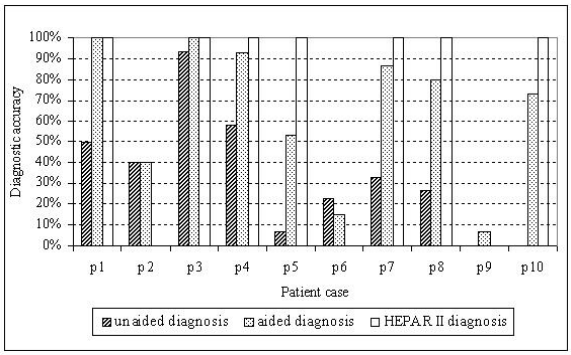


Figure 2: Diagnostic accuracy for each patient case for unaided, aided, and the HEPAR II diagnosis

highest accuracy reached by the physicians was 50%, which means that HEPAR II's diagnostic performance was 40% higher than that of the best physician.

4.2 Impact of the HEPAR II system on physician behavior

Figure 2 shows the diagnostic accuracy for each of the ten patient cases. The picture captures the performance of the physicians for both unaided and aided diagnosis and the performance for the HEPAR II diagnosis. By 'unaided diagnosis' is meant the diagnosis made by a physician and 'aided diagnosis' indicates the diagnosis made by a doctor after he or she saw the answer of the HEPAR II system. The overall diagnostic accuracy roughly doubled, increasing from 32.6% (unaided diagnosis) to 66% (aided diagnosis), which approaches the system's accuracy. This increase was significant at $p < 0.003$ level (paired, one-sided Student-t test).

Table 2 presents the number of correct and incorrect diagnoses made by the physicians and HEPAR II. Table 3 shows corresponding results, but it captures the situation when the physicians had seen HEPAR II's result and were given the opportunity to change their diagnostic decisions. The columns of these tables correspond to the user decisions and the rows capture the system answers. When analyzing the Tables 2 and 3 we can observe 'a migration' of the diagnoses. In those cases where the HEPAR II system proposed the correct diagnoses (the second row of the tables), the users typically changed their decisions, i.e., 49 incorrect diagnoses made by the physicians (33% of all diagnoses and 49% of incorrect diagnoses made by the physicians) were changed to the correct diagnosis suggested by the system. The number in the third row and the third column of Table 3, marked by an asterisk, is broken into two groups: 24 decisions were the same incorrect diagnoses provided by the system and the users and 10 were different, incorrect diagnoses.

Table 2: Before: Correct and incorrect diagnoses of the users and the HEPAR II system

system / user	correct	incorrect	total
correct	39	65	104
incorrect	9	34	43
total	48(32.6%)	99(67.4%)	147(100%)

Table 3: After: Correct and incorrect diagnoses of the users and the HEPAR II system

system / user	correct	incorrect	total
correct	88	16	104
incorrect	9	(24+10)*	43
total	97(66%)	50(34%)	147(100%)

5 Discussion

This paper addressed two problems: (1) an empirical comparison of the diagnostic performance of a normative diagnostic system based on a Bayesian network to the performance of physicians and (2) the impact that a normative system has on its user. In both cases, the conclusions from the performed empirical study are encouraging toward normative systems.

The first result of this study is that diagnosis of liver disorders is far from trivial. System performance of 49% correct diagnoses among all cases recorded in the database, subject of criticism on the part of reviewers of our earlier papers on HEPAR II, seems quite better than average performance of general practitioners with limited clinical practice. On a subset of 10 cases randomly drawn from the database, HEPAR II was twice as accurate as the physicians (70% vs. 32.1% accuracy) and 40% better than the most accurate physician. The most straightforward explanation of the system's apparently low performance of 49% is that the problem of diagnosing a liver disorder is hard (until biopsy is performed).

The second result of this study is that a diagnostic system like HEPAR II can be quite beneficial to its users – interaction with the system more than doubled the users' accuracy (from 32.6% before to 66% after seeing the system's suggestion).

In the experiment conducted here, the answers of HEPAR II did not include any explanation on how the diagnoses were reached, i.e., the physicians received simply a list of ordered posteriors for each diagnosis. Some authors (e.g., [5; 17]) report that an explanation module can significantly increase insight into system recommendations and effectively increase the quality of the ultimate diagnosis made by the physicians using the system. An explanation facility might be more effective in convincing the doctors to modify their wrong diagnoses when the system is right. Quite likely such module would increase the benefit of using the system even more.

Our future research plans include conducting a similar evaluation study with the experts in the domain of hepatology. A question that is worth pursuing is whether the system performs at the expert level.

Acknowledgments

I would like to thank my collaborators in the HEPAR II project, Drs. Hanna Wasyluk and Marek Druzdzel, for their continuous support and valuable suggestions. Drs. Marek Druzdzel's and Javier Díez's comments on the paper improved the presentation. Thanks also go to Gosia Kretowska for her useful suggestions. Support for this research was provided by the Białystok University of Technology grant W/II/1/00, and by NATO Collaborative Linkage Grant PST.CLG.976167.

References

[1] K. P. Adlassnig and W. Horak. Development and retrospective evaluation of HEPAXPERT-I: A routinely-used expert system for interpretive analysis of hepatitis A and B serologic findings. *Artificial Intelligence in Medicine*, 7:1–24, 1995.

[2] R. A. Bankowitz and M. A. McNeil. A computer-assisted medical diagnostic consultation service. Implementation and prospective evaluation of a prototype. *Annals of Internal Medicine*, 110(10):824–832, 1989.

[3] 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.

[4] Bruce G. Buchanan and Edward H. Shortliffe, editors. *Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project.* Addison-Wesley, Reading, MA, 1984.

[5] William J. Clancey. Use of MYCIN's rules for tutoring. In Buchanan and Shortliffe [4], chapter 26, pages 464–489.

[6] B. Kaplan and H. P. Lundsgaarde. Toward an evaluation of an integrated clinical imaging system: Identifying clinical benefits. *Methods of Information in Medicine*, 35:221–229, 1996.

[7] Paul E. Lehner, Theresa M. Mullin, and Marvin S. Cohen. A probability analysis of the usefulness of decision aids. In *Uncertainty in Artificial Intelligence 5*, pages 427–436, New York, N. Y., 1989. Elsevier Science Publishing Company, Inc.

[8] P. Lucas and A. R. Janssens. Second evaluation of HEPAR, an expert system for the diagnosis of disorders of the liver and biliary tract. *Liver*, 11:340–346, 1991.

[9] P. J. F. Lucas, R. W. Segaar, and A. R. Janssens. HEPAR: an expert system for diagnosis of disorders of the liver and biliary tract. *Liver*, 9:266–275, 1989.

[10] Randolph A. Miller, Harry E. Pople, Jr., and Jack D. Myers. INTERNIST–I, an experimental computer-based diagnostic consultant for general internal medicine. *New England Journal of Medicine*, 307(8):468–476, August 1982.

[11] Agnieszka Oniśko, Marek J. Druzdzel, and Hanna Wasyluk. Extension of the Hepar II model to multiple-disorder diagnosis. In S.T. Wierzchoń M. Kłopotek, M. Michalewicz, editor, *Intelligent Information Systems,* Advances in Soft Computing *Series*, pages 303–313, Heidelberg, 2000. Physica-Verlag (A Springer-Verlag Company).

[12] Agnieszka Oniśko, Marek J. Druzdzel, and Hanna Wasyluk. Learning Bayesian network parameters from small data sets: Application of Noisy-OR gates. *International Journal of Approximate Reasoning*, 2001. To appear.

[13] Mark Peot and Ross Shachter. Learning from what you don't observe. In *Proceedings of the Fourteenth Annual Conference on Uncertainty in Artificial Intelligence (UAI–98)*, pages 439–446, San Francisco, CA, 1998. Morgan Kaufmann Publishers.

[14] S. Shiomi, T. Kuroki, H. Jomura, T. Ueda, N. Ikeoka, K. Kobayashi, H. Ikeda, and H. Ochi. Diagnosis of chronic liver disease from liver scintiscans by fuzzy reasoning. *Journal of Nuclear Medicine*, 36:593–598, 1995.

[15] A. Theodossi, D. J. Spiegelhalter, and B. Portman. The value of clinical, biochemical, ultrasound and liver biopsy data assessing patients with liver disease. *Liver*, 3:315–326, 1983.

[16] J. van der Lei and M. Musen. Comparison of computer-aided and human review of general practitioners' management of hypertension. *The Lancet*, 338:1504–1508, 1991.

[17] Jerold W. Wallis and Edward H. Shortliffe. Customized explanations using causal knowledge. In Buchanan and Shortliffe [4], chapter 20, pages 371–388.

[18] 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.

[19] Jeremy Wyatt. Lessons learnt from the field trial of ACORN, an expert system to advise on chest pain. In *Proceedings of the MEDINFO-89*, pages 111–115, 1989.

[20] L. Yu and L. M. Fagan. Antimicrobial selection by a computer. A blinded evaluation by infectious diseases experts. *JAMA*, 242:1279–1282, 1979.

The notion of diagnosis in decision-theoretic planning

Niels Peek
Department of Medical Informatics, University of Amsterdam,
P.O. Box 22700, 1100 DE Amsterdam, The Netherlands
E-mail: n.b.peek@amc.uva.nl

Abstract

Decision-theoretic formalisms such as influence diagrams and POMDPs can be used to solve complex decision problems in clinical medicine. These formalisms then help to construct a decision policy that prescribes the best clinical actions given patient-specific findings. The notion of *diagnosis* and its role in clinical reasoning is however left implicit. This paper shows how diagnoses can be made explicit in multivariate POMDPs that are used for clinical problem solving. The aim is to facilitate the user's understanding of the recommendations of a decision policy. Three types of differential diagnoses, each with its own perspective on the importance of diagnostic hypotheses, are presented.

Keywords: diagnosis, POMDPs, decision-theoretic planning

1 Introduction

The notion of diagnosis is central to everyday clinical reasoning and decision making. The first thing a doctor will usually ask himself when confronted with a patient is, 'What is wrong with this patient?'. If this is not clear from the presented symptoms and findings, considerable effort may be taken to ascertain the patient's disorder by conducting diagnostic investigations. After having established the diagnosis with sufficient certainty, the patient is often treated according to a guideline or protocol that is based on classifications of disease.

It is not surprising, then, that diagnosis has received more attention in formalisations of medical reasoning than any other aspect of clinical problem solving. A large number of formal approaches to diagnosis have been developed and applied, rooted in such diverse fields as Bayesian probability theory [1], heuristic reasoning [2], qualitative simulation [3], and logical abduction [4].

Recent advances in computational techniques for probabilistic inference have yielded renewed interest in the application of *decision theory* to complex, real-world problems. Among the decision-theoretic formalisms that are now most widely studied are *influence diagrams* (IDs) and *partially-observable Markov decision processes* (POMDPs). Clinical medicine seems to be a natural area of application for decision theory; both IDs and POMDPs have been applied to medical decision problems [5; 6; 7].

Decision-theoretic formalisms support the construction of *decision policies*. Such a policy provides a *decision maker*, i.e. a person facing a decision problem, with the preferred action choice given specific observations on the problem; the preferred choice then has the property of giving the best (expected) prospects for the future. However, for a doctor that is treating a patient, this type of decision support will often be too shallow as it lacks any reference to the patient's perceived disorder. The doctor is therefore not supported in the type of diagnosis-based reasoning he feels comfortable with.

In this paper, we show how the notion of diagnosis can be made explicit in decision-theoretic reasoning. As a formal framework we will employ multivariate POMDPs; the notations on POMDPs to be used in this paper are introduced in Section 2. We will first formulate a Bayesian type of differential diagnosis within this framework in Section 3. Then, in Section 4, we present two other types of diagnoses, one that stresses the patient's prognosis, and one that stresses the next decision to be made. The paper is concluded with a discussion in Section 5 and conclusions in Section 6.

2 POMDPs

POMDPs [8; 9] are models for action planning under uncertainty with partial information, or *decision-theoretic planning* for short. The underlying concept can be described as follows. At a specified point in time, a decision maker observes the state of a dynamic system. Based on this observation, he chooses an action. The action choice produces two results: the decision maker receives an immediate reward, and the system evolves to a new state at a subsequent point in time according to an effect determined by the action

choice. At this subsequent time point, the decision maker faces a similar problem, but now the system may be in a different state. The decision maker's objective is to develop a decision- making policy that maximises the expected total reward over a predefined period of time.

This section introduces the notations on multivariate POMDPs that will be used throughout the paper. Let $T = \{0, 1, \ldots, N\}$ be a set of time points, and let X be a set of finite random variables. We will refer to a subset $T' \subseteq T$ of subsequent elements in T as a *time segment*. The set X is taken to jointly describe a *dynamic system*; we use Ω_X to denote the set of all possible joint value assignments to the variables of X. An element $\mathrm{c}_X \in \Omega_X$ is called a *configuration*, or equally *state*, of X.

Definition 1 *A* multivariate POMDP *over X and T is a tuple $M = (A, \gamma, o, r)$, where A is a finite set of available* actions, *$\gamma : \Omega_X \times A \times \Omega_X \to [0, 1]$ is a* transition probability function, *$o : A \to \wp(X)$ is an* observation function, *and $r : \Omega_X \times A \to \mathbb{R}$ is a* reward function.

The time points in T denote moments where the decision maker is expected to select an action $a \in A$ to influence and/or observe the current state of the dynamic system. The action effects are modelled as follows. When configuration $\mathrm{c}_X \in \Omega_X$ characterises the state at time point $t \in T$, selection of action $a \in A$ will result in a transition to state $\mathrm{c}'_X \in \Omega_X$ at time point $t + 1$ with probability $\gamma(\mathrm{c}_X, a, \mathrm{c}'_X)$. Furthermore, the decision maker is able to observe the values of variables from the set $o(a) \subseteq X$ at time point t; the observed values are used to optimise the decisions at future time points. No decision is made at the final time point $t = N$; this moment is included for evaluation of the final state only.

Because we are interested in changes in the dynamic system over time, we define the *state function* s_x associated with random variable $x \in X$ over time segment T' as a random function $s_x : T' \to \mathrm{dom}(x)$, where $\mathrm{dom}(x)$ denotes the value domain of variable x. The set $S_X = \{s_x \mid x \in X\}$ of all state functions over time segment T' is called the *joint state function* over T'. We will sometimes think of S_X as a function of time and write $S_X(t) = \mathrm{c}_X$ to indicate that $\mathrm{c}_X \in \Omega_X$ is the configuration of X obtained by parallel application of all state functions in S_X to time point $t \in T'$.

Now, let P_0 be a joint probability distribution on the set X at time point $t = 0$, reflecting the decision maker's prior beliefs on the initial state. Given P_0 and a sequence $\alpha = a_0, \ldots, a_{t-1}$ of action choices up to time point $t \in T$, a probability distribution P_t^α on the joint state function S_X over time segment $\{0, \ldots, t\}$ is constructed as follows:

$$P_t^\alpha(S_X) = P_0(S_X(0)) \cdot \prod_{i=0}^{t-1} \gamma(S_X(i), a_i, S_X(i+1)). \quad (1)$$

The distribution P_t^α represents the decision maker's prior beliefs on S_X once he is certain to choose action sequence α (or has already done so); it does not yet take into account the observations that are made over time. If ξ denotes the *evidence*, i.e. the collected observations, up to time point t then the conditional distribution $P_t^\alpha(S_X(t) \mid \xi)$ represents the decision maker's beliefs with respect to the current state. For instance, $P_t^\alpha(s_x(t) = v \mid \xi)$ is the probability that variable $x \in X$ has value $v \in \mathrm{dom}(x)$ at time point t after choosing action sequence α, and given evidence ξ.

The decision-making processes is guided by the objective to maximise expected *utility*, which is defined as the sum of rewards that are received at subsequent time points. A reward $r(\mathrm{c}_X, a)$ is received when c_X is the system's state and the decision maker chooses action $a \in A$. A special reward function $r_N : \Omega_X \to \mathbb{R}$ is used for the final time point $t = N$ where no action is chosen. Temporal risk preferences can be incorporated by employing an exponentially increasing discount factor. The solution to a given POMDP consists of a *decision policy* $\pi = \pi_0, \ldots, \pi_{N-1}$, where each π_t is a function that provides the decision maker with the preferred action choice on the basis of given evidence. For instance, $\pi_t(\xi) = a$ indicates that action a is preferred at time point t if ξ has been observed.

3 Bayesian diagnosis in POMDPs

Many problems of patient management in clinical medicine require temporal action planning with uncertain and incomplete information. POMDPs have therefore been suggested as a suitable framework to study these problems [6; 7; 10]. Specific value of applying POMDPs is to be expected when diagnostic and therapeutic decisions interact, when patient management extends over a significant period of time, or when a careful tradeoff between short-term and long-term risks is required.

In medical applications of POMDPs, configurations of the set X usually describe *clinical conditions* of a patient, e.g. 'healthy', 'diseased without clinical signs', or 'diseased with clinical signs'. The respective variables of X then represent *attributes* of these conditions such as the patient's disorder, or specific signs or test results. The set T is chosen to cover a sufficiently large time span to model patient management problems in the domain in question, and the set A comprises all clinical actions that require distinction in the problem under consideration. From a conceptual point of view, it is often convenient to separate *test actions* (i.e. examinations and diagnostic procedures) from *treatment actions* (i.e. therapy and interventions). We note, though, that it is sometimes difficult to make a formal distinction between these action types (e.g. surgery generally yields a wealth of diagnostic information). Finally, the reward $r(\mathrm{c}_X, a)$ is generally the patient's *life expectancy* over time interval $[t, t + 1)$ in the given circumstances, and the

reward $r_N(c_X)$ is the patient's future life expectancy in condition c_X.

After a specific patient management problem thus has been modelled, the expected results of following different decision policies, possibly for different types of patients, can compared. For problems that are relatively small in size, it is also possible to compute an optimal decision policy, i.e. a decision policy that provides for the best action choice in all perceivable situations during patient management. As the problem of computing optimal decision policies is PSPACE-complete [11], this is not possible for larger problems. In either case, the type of support provided by these models is based on decision policies, i.e. mappings from evidence to actions. The patient's perceived disorder within these decision policies is left implicit; the rationale for prescribing a particular action is therefore obscured. Below, we show how to take out the notion of diagnosis in clinical POMDPs.

We will assume that $d \in X$ is a random variable that represents the patient's disorder. For convenience, we assume that $\mathrm{dom}(d) = \{1, 2, \ldots, k\}$; that is, there are k possible, mutually-exclusive disorders. It is the value of variable d that we are after in diagnosing the patient; we will refer to an expression of the form $s_d(t) = j$ as a *diagnostic hypothesis* at time point t. The variable d is assumed to be hidden from observation, i.e. there is no action $a \in A$ such that $d \in o(a)$. Now, let as before α be the sequence of actions conducted over time segment $\{0, \ldots, t-1\}$, and let ξ be the collected evidence so far. The decision maker faces the decision at time point t.

Definition 2 *Diagnostic hypothesis $s_d(t) = j$ is called an* explanation *of the evidence ξ at time t when*

$$P_t^\alpha(s_d(t) = j, \xi) \quad > \quad 0. \tag{2}$$

The set of all explanations of ξ at time t is denoted by $expl_t^\alpha(\xi)$.

A hypothesis $s_d(t) = j$ is considered to be a possible explanation of given evidence ξ if they are non-contradictory, i.e. if $P_t^\alpha(s_d(t) = j, \xi) > 0$. Otherwise, the hypothesis is not comprised in $expl_t^\alpha(\xi)$, and said to be *rejected*.

When uncertainty abounds in the domain of application, the set $expl_t^\alpha(\xi)$ will often be too large to be of practical value. The common solution in clinical medicine is to order the set $expl_t^\alpha(\xi)$ by posterior probability. We thus obtain a list of diagnoses that are ranked from most to least probable given the available findings with respect to the patient's condition; we will speak of a *Bayesian differential diagnosis*.

Definition 3 *A* Bayesian differential diagnosis *is a list $(h_1, p_1), \ldots, (h_m, p_m)$ ordered by decreasing values of p_i, where $h_1, \ldots, h_m \in expl_t^\alpha(\xi)$ and*

$$p_i \quad = \quad P_t^\alpha(h_i \mid \xi), \tag{3}$$

for all $i = 1, \ldots, m$.

A Bayesian differential diagnosis provides a concise picture of the decision maker's beliefs with the respect to the patient's disorder. The more probable a hypothesis is, the higher its ranking will be; the most probable hypothesis is considered to be the most important one. From the distribution of p_i values one can estimate the *uncertainty* in the diagnosis, using an information measure (e.g. Shannon entropy).

A Bayesian diagnosis focuses on events (findings) in the past but ignores the patient's prospects for the future. For a doctor, a differential diagnosis will usually be a starting point for further clinical action; this may be diagnostic testing to gather more information on the patient's condition, or may be therapy aimed to improve that condition. In both cases, the objective is to reach a better prognosis for the patient: indirectly through better opportunities to treat the patient with the information gathered, or directly through reaching a hopefully better health status. A Bayesian differential diagnosis does however not consider the patient's prognosis in ranking the hypotheses, and may therefore fall short in providing directions for further management. For instance, an unlikely but highly threatening disorder may guide the decision to conduct a specific diagnostic test. In the next section we present two other types of diagnoses that focus on the patient's prognosis.

4 Prognostic and decisional diagnoses

We will now assume that a decision-making policy $\pi = \pi_{t+1}, \ldots, \pi_{N-1}$ for time points $t+1, \ldots, N-1$ is available. Let $\tilde{u}_t^\alpha(a, \pi \mid \xi)$ be the expected utility of choosing action $a \in A$ at time point t and following policy π thereafter, where, as before, action sequence α has been carried out so far and ξ is the available evidence. Clearly,

$$a^* \quad = \quad \arg\max_{a \in A} \{\tilde{u}_t^\alpha(a, \pi \mid \xi)\} \tag{4}$$

is the optimal action choice in the current situation. In a *prognostic diagnosis*, we use the effects of the various diagnostic hypotheses on expected utility to quantify their importance. Let therefore $\tilde{u}_t^\alpha(a, \pi \mid h, \xi)$ be a similar type of utility as above, where now diagnostic hypothesis $h \in expl_t^\alpha(\xi)$ is known to be the patient's true disorder.

Definition 4 *A* prognostic differential diagnosis *is a list $(h_1, u_1), \ldots, (h_m, u_m)$ ordered by* increasing *values of u_i, where $h_1, \ldots, h_m \in expl_t^\alpha(\xi)$, and*

$$u_i \quad = \quad \tilde{u}_t^\alpha(a^*, \pi \mid h_i, \xi) \tag{5}$$

for all $i = 1, \ldots, m$.

In a prognostic differential diagnosis, the hypotheses are ranked by increasing life-expectancy, under the assumption that action a^* is chosen next and policy π is followed thereafter. Hence, the hypothesis that presents the highest risk to the patient, under the given decision policy, is now considered as most important. A high value of u_i indicates that the disorder h_i is harmless, that a^* is an effective cure for h_i (if a^* is a treatment action), or that a^* is effective in discriminating h_i from other hypotheses (if a^* is a

test action). Furthermore, from the distribution of u_i values, it is possible to see how different the various future scenarios are.

The Bayesian and prognostic types of differential diagnosis can be combined to obtain a *utility-theoretic diagnosis*; the expected utility u_i associated with hypothesis h_i is then weighed with its posterior probability p_i. This will descrease the rank of risky but unlikely disorders, and increase the rank of moderately dangerous disorders with higher probability. Note, however, that the quantity $u_i \cdot p_i$ does not have an intuitive semantics. It is the (absolute) contribution of hypothesis h_i to the patient's life-expectancy $\tilde{u}_t^\alpha(a^*, \pi \mid \xi)$ under action choice a^*, as

$$\sum_{h_i \in expl_t^\alpha(\xi)} u_i \cdot p_i \quad = \quad \tilde{u}_t^\alpha(a^*, \pi \mid \xi). \tag{6}$$

The quantity $u_i \cdot p_i$ itself is however not a life-expectancy. It does seem favourable, therefore, provide the probabilities and utilities separately to facilitate a decision maker's understanding of the situation.

The prognostic and utility-theoretic differential diagnoses focus on the results of choosing the optimal action a^*. As such, they assume that the decision at time point t is already made and neglects the fact that the decision maker is facing a choice. The third type of diagnosis we present therefore considers the impact of varying decisions on given hypotheses; it is called the *decisional differential diagnosis*. Let

$$\tilde{u}_i^{\max} \quad = \quad \max\{\tilde{u}_t^\alpha(a, \pi \mid h_i, \xi) \mid a \in A\} \tag{7}$$

and

$$\tilde{u}_i^{\min} \quad = \quad \min\{\tilde{u}_t^\alpha(a, \pi \mid h_i, \xi) \mid a \in A\} \tag{8}$$

be the maximum and minimum expected utility, respectively, when hypothesis $h_i \in expl_t^\alpha(\xi)$ is the patient's true disorder.

Definition 5 *A* decisional differential diagnosis *is a list* $(h_1, q_1), \ldots, (h_m, q_m)$ *ordered by decreasing values of* q_i, *where* $h_1, \ldots, h_m \in expl_t^\alpha(\xi)$, *and*

$$q_i \quad = \quad \frac{\tilde{u}_i^{\max} - \tilde{u}_i^{\min}}{\tilde{u}_t^\alpha(a^*, \pi \mid \xi)} \tag{9}$$

for all $i = 1, \ldots, m$.

In a decisional differential diagnosis, the value q_i is the maximum *loss* in expected utility when hypothesis h_i is true, relative to what is to be expected when the 'normal', optimal decision-theoretic choice a^* is made. The loss here would be due to making the wrong decision at time point t: $\tilde{u}_i^{\max}$ represents the scenario where, considering only hypothesis h_i, the best decision is made, and $\tilde{u}_i^{\min}$ represents the scenario where the worst decision is made with respect to h_i. Note though that these decisions may have rather different properties when considering all hypotheses and their respective probabilities.

If q_i is high, then h_i is probably the hypothesis to focus on when making the decision at time point t because of its potential impact on the patient prognosis. It is possible that $q_i > 1$; this occurs when h_i represents a rather improbable but well-treatable illness in an otherwise unfavourable situation. The hypothesis h_i is then a last straw to catch in treating the patient. We will usually find that the maximum loss $\tilde{u}_i^{\max} - \tilde{u}_i^{\min}$ is relatively small compared to $\tilde{u}_t^\alpha(a^*, \pi \mid \xi)$, and therefore $q_i \ll 1$.

If q_i is low, then our decision at time point t will hardly influence the patient's prognosis if hypothesis h_i is true; it is therefore reasonable to more or less neglect h_i when making the decision. It is possible that $q_i = 0$, in which case we can completely forget about h_i in our choice at moment t. Note, though, that this does not preclude h_i from being quite probable, or even being the correct hypothesis.

5 Discussion

Decision-theoretic formalisms such as POMDPs incorporate both models of action-driven state change and action-driven information acquisition. As such, these formalisms can capture the interplay between diagnostic and therapeutic reasoning in clinical patient management. Yet, the notion of diagnosis itself is not made explicit: diagnostic reasoning is only covered in the sense that information-gathering actions are prescribed in some situations. The approach presented here tries to make the notion of diagnosis explicit in POMDPs.

The objective of our work is to facilitate the user's understanding of the recommendations made by a decision-support system, by supplementing prescribed action choices with differential diagnoses. A diagnosis that is based on posterior probability may however fall short in providing the rationale for a decision; we have therefore presented additional prognostic and decisional types of diagnoses. Each of these diagnosis types incorporates a different perspective on what makes a diagnostic hypothesis important. They are therefore, to a large extent, orthogonal; it is possible though to create rankings that use a mixture of perspectives. The three diagnosis types were presented within the POMDP framework, but they can be employed in any model that incorporates notions of prognosis (utility) and choice. Others have earlier argued that probabilistic diagnoses should be *evaluated* in terms of utility [12], but the prognostic and decisional types of diagnosis presented here have not been applied before in decision-theoretic reasoning formalisms.

It is often complained that diagnostic programs produce moderately long lists of diagnoses, containing many diagnoses that a knowledgeable physician would regard as completely irrelevant [13]. When the evidence ξ is little symptomatic (i.e. does not exclude many explanations from the differential diagnosis), our approach carries the risk of meeting a similar complaint. To anticipate this situation, one may lift the requirement that all explanations of given evidence be contained in a differential diagnosis. One could, for instance, omit explanations that are improbable, pose

little or modest risks to the patients, and have low decisional importance. We do note that after omitting explanations from the differential diagnosis, however improbable they are, the correct hypothesis may no longer be included.

We have restricted ourselves to considering a single diagnostic variable in the multivariate POMDP. This implies that we assume (i) only a single disorder to be present, (ii) that all diagnostic hypotheses are mutually exclusive, and (iii) that given findings are always explainable by some diagnostic hypothesis. These assumptions may however prove unrealistic in real-world domains. The POMDP framework does allow a more general approach where these assumptions are lifted; we have investigated this approach in [14].

6 Conclusions

Decision-theoretic reasoning is characterised by the fact that each situation of choice is ultimately reduced to a utility-theoretic tradeoff. Yet from a conceptual point of view, these choice situations may be very different. In medical decision making, differences between choice situations typically stem from the perceived diagnosis of the patient; the notion of diagnosis is however left implicit in many decision-theoretic formalisms. In this paper we have shown how the notion of diagnosis can be made explicit in decision-theoretic planning problems, without losing the underlying utility-theoretic tradeoff. We believe that this will facilitate the user's understanding, and hence, acceptance, of the recommendations made by a decision-support system.

We have compared a Bayesian notion of diagnosis with two new notions of diagnosis that focus on the patient's future prospects. The prognostic notion of diagnosis highlights the differences in expected life time associated with different diagnostic hypotheses; the decisional notion of diagnosis highlights the maximum losses in expected utility under different diagnostic hypotheses. The three notions of diagnosis jointly provide a clear picture of the role of diagnosis in clinical decision making.

In the future we plan to evaluate the presented notions of diagnosis on a POMDP model that is developed to support paediatric cardiologists in the management of children with a ventricular septal defect; this model is described in [7]. Furthermore, we intend to develop a heuristic solution method for POMDP problems based on the notions described here. Most heuristic problem-solving knowledge from clinicians is centred on the patient's diagnosis. Once the notion of diagnosis is made explicit in a decision-theoretic formalism, this knowledge is therefore easier to incorporate in the associated reasoning methods. For instance, most physicians are able to provide fairly reliable treatment rules for given disorders. If a particular disorder stands out in a differential diagnosis, application of the associated rule may be a reasonable option.

Acknowledgements

The author wishes to thank Peter Lucas and John-Jules Meyer for their valuable comments on the ideas presented.

References

[1] F.T. de Dombal, D.J. Leaper, J.R. Staniland et al. Computer-aided diagnosis of acute abdominal pain. *BMJ*, 2:9–13, 1972.

[2] W.J. Clancey. Heuristic classification. *Artif. Intell.*, 27:289–350, 1985.

[3] I. Bratko, I. Mozetič, and N. Lavrač. *KARDIO: A study in deep and qualitative knowledge for expert systems.* MIT Press, Cambridge, MA, 1989.

[4] J. de Kleer, A.K. Mackworth, and R. Reiter. Characterizing diagnoses and systems. *Artif. Intell.*, 52:197–222, 1992.

[5] R. Bellazzi and S. Quaglini. Reusable influence diagrams. *Artif. Intell. Med.*, 6:483–500, 1994.

[6] M. Hauskrecht and H. Fraser. Planning medical therapy using partially observable Markov decision processes. In *Proc. 9th Int. WS Principles of Diagnosis (DX–98)*, pp. 182–189, 1998.

[7] N.B. Peek. Explicit temporal models for decision-theoretic planning of clinical management. *Artif. Intell. Med.*, 15(2):135–154, 1999.

[8] G.E. Monahan. A survey of partially observed Markov decision processes: theory, models, and algorithms. *Manage. Sci.*, 28(1):1–16, 1982.

[9] M.L. Littman. *Algorithms for Sequential Decision Making.* Ph.D. thesis, Dept. of Computer Science, Brown University, 1996.

[10] G. Tusch. Optimal sequential decisions in liver transplantation based on a POMDP model. In W. Horn, editor, *ECAI 2000: Proc. 14th Europ. Conf. Artificial Intelligence*, pp. 186–190. IOS Press, 2000.

[11] C.H. Papadimitriou and J.N. Tsitsiklis. The complexity of Markov decision processes. *Math. Oper. Res.*, 12(3):441–450, 1987.

[12] J.D.F. Habbema and J. Hilden. The measurement of performance in probabilistic diagnosis. IV. Utility considerations in therapeutics and prognosis. *Methods Inf. Med.*, 20:80–96, 1981.

[13] E.S. Berner, G.D. Webster, A.A. Shugerman et al. Performance of four computer-based diagnostic systems. *NEJM*, 330(25):1792–1796, 1994.

[14] N.B. Peek. *Decision-Theoretic Planning of Clinical Patient Management.* Ph.D. thesis, Institute of Information and Computing Sciences, Utrecht University, 2000.

Probability Assessment with Maximum Entropy in Bayesian Networks

Wim Wiegerinck **Tom Heskes**
SNN, University of Nijmegen,
Geert Grooteplein 21, 6525 EZ, Nijmegen, The Netherlands
wimw@mbfys.kun.nl

Abstract

Bayesian networks are widely accepted as tools for probabilistic modeling in the medical domain. In modeling Bayesian networks in collaboration with domain experts, the definition of the network structure is relatively easy. The assessment of the conditional probability tables (CPT) is often a much more difficult task, even though there is a lot of statistical information available in the medical literature. The problem is twofold. In the first place it is usually not possible to use this information directly to fill in the CPTs. In the second place, the information is usually insufficient for a unique definition of the CPTs. A standard approach to define a probabilistic model on the basis of insufficient statistical information is to apply the Maximum Entropy Method (MaxEnt). MaxEnt searches for the unique model that maximizes the entropy under the constraints that it satisfies the given statistical information. In standard applications of MaxEnt for models defined by one joint probability table, these constraints are linear in the table entries. However, if MaxEnt is applied to a Bayesian network, i.e. the joint distribution is factorized into a product of CPTs, these constraints are typically nonlinear in the CPTs. In this paper we show how these nonlinear constraints can be dealt with, and we describe an algorithm that (locally) maximizes entropy under constraints in Bayesian networks. The method is illustrated by an example.

1 Introduction

Computer-based diagnostic decision support systems will play an increasingly important role in health care. They may improve the quality of the diagnostic process in accuracy and efficiency, while costs and burden of patients may be reduced. In addition, they can play an invaluable role in medical education. Potential users include general internists, super specialists, residents in internal medicine, and medical students.

The modern view is that decision support systems should be based on a probabilistic model. This approach has the advantage that it can deal with uncertainty in a consistent and mathematically correct way. In particular Bayesian networks[5; 3] provide a powerful and conceptually transparent formalism for probabilistic modeling.

Modeling of a Bayesian network consists of two parts, a qualitative and a quantitative part. The qualitative part is the determination of the structure of the network. If the network is build in collaboration with domain experts, the determination of the structure is often considered as a relatively easy task, since this task usually fits well with knowledge that medical experts often have about causal relationships between variables. The quantitative part consists of quantifying the conditional probability tables (CPTs) in the network. This part is often considered by medical experts as a much harder or even impossible task [2]. The reason is that medical domain experts themselves often have no idea about these probabilities. In most medical domains some statistical information $\mathcal{I}$ is provided in the literature. In such a case, one may try to choose the CPTs in the network such that network fits with $\mathcal{I}$. Unfortunately, $\mathcal{I}$ often does not translate directly into network CPTs, that is to say, it is often not clear to the experts how $\mathcal{I}$ should be translated into quantitative CPTs in the Bayesian network. Typically, $\mathcal{I}$ consists of conditional probabilities in the 'wrong direction', from 'effect' to 'cause'. In addition, these 'reversed' CPTs are often insufficient to uniquely define the desired CPTs in the network. The toy problem in the last section in the paper is an example where $\mathcal{I}$ has wrong direction and is insufficient for unique determination of the model. Often $\mathcal{I}$ can be formulated as linear probabilistic constraints, i.e., constraints of the form $\sum_{\{x\}} p(x) f_\alpha(x) = 0$, and/or $\sum_{\{x\}} p(x) g_\beta(x) \leq 0$, where $p(x)$ is the (joint) probability distribution and $f_\alpha(x)$ and $g_\beta(x)$ are functions of the state space $\{x\} = \{x_1, \ldots, x_n\}$. A typical example is a constraint on the conditional probability $p(x_1 = a | x_2 = b) = c$ which can be expressed as

$\sum_x p(x)(\delta_{x_1 a}\delta_{x_2 b} - c\delta_{x_2 b}) = 0$, where we used the Kronecker delta ($\delta_{xy} = 1$ if $x = y$ and $\delta_{xy} = 0$ if $x \neq y$).

In this paper, it is assumed that $\mathcal{I}$ is consistent, i.e. that there is at least one parameter setting of the distribution that satisfies the constraints. However, since $\mathcal{I}$ is in general insufficient for a unique determination of the model p, a whole set of distributions will satisfy the constraints. A standard way to proceed is to select a representative of this set of distributions by applying the Maximum Entropy Method (MaxEnt) [4]. MaxEnt searches the distribution that maximizes entropy under the given constraints. Roughly spoken, it selects the distribution p that satisfies the constraints without introducing any additional information.

In this paper, we apply MaxEnt to a Bayesian network with a given structure $p(x) = \prod_i p(x_i|\pi_i)$ to quantify its CPTs. The difference with MaxEnt applied to a general model $p(x)$ is that MaxEnt applied to a Bayesian network has to deal with a set of constraints *and* a set of independency statements. One approach could be to try to formulate the independency statements as as additional constraints to a general model $p(x)$ and apply standard MaxEnt to p. The way we proceed is, however, to keep the factorization into CPTs, and try to find the CPTs that maximizes the entropy of the joint distribution. As a consequence, a technical difference with standard MaxEnt is that the constraints, which are linear in the joint probability $p(x)$, are *non-linear* in the CPTs $p(x_i|\pi_i)$. This causes some complications in the optimization scheme of MaxEnt. However, one can effectively deal with these complications.

This workshop paper is organized as follows. In section 2 standard MaxEnt is shortly reviewed. In section 3, we show how the method applies to Bayesian networks. In section 4, the method is applied to a toy problem. We end the paper with a short discussion in section 5.

2 Maximum Entropy (MaxEnt)

In this section, we shortly review the standard Maximum Entropy (MaxEnt) method with linear probabilistic constraints [4]. We consider probability distributions $p(x)$ on a set of discrete variables $x = x_1, \ldots, x_n$ with a finite domain, $x_i \in \{1, \ldots, n_i\}$. If a set of linear constraints on p,

$$\sum_{\{x\}} f_\alpha(x)p(x) = 0 \qquad \alpha = 1 \ldots k \tag{1}$$

$$\sum_{\{x\}} f_\alpha(x)p(x) \geq 0 \qquad \alpha = k+1 \ldots m \tag{2}$$

is given, MaxEnt tries to find the probability distribution $p(x)$ that maximizes the entropy

$$H(p) = -\sum_{\{x\}} p(x) \log p(x) \tag{3}$$

under these constraints.

Introducing Lagrange multipliers $\lambda = \{\lambda_\alpha\}$, and a Lagrange multiplier γ to ensure normalization of p, we can formulate the optimization problem by Lagrangian

$$L(p, \lambda, \gamma) = H(p) + \sum_\alpha \sum_{\{x\}} \lambda_\alpha f_\alpha(x)p(x) + \gamma(\sum_{\{x\}} p(x) - 1) \tag{4}$$

which should be maximized with respect to p and minimized with respect to the Lagrange multipliers λ (within the domain $\lambda_\alpha \leq 0$ for $\alpha > k$) and γ. Taking the gradient of L with respect to $p(x)$, setting it to zero, and eliminating γ, we can solve p as explicitely as a function of λ. The solution p^* has the well known exponential form

$$p^*(x) = \frac{1}{Z} \exp \sum_\alpha \lambda_\alpha f_\alpha(x) \tag{5}$$

where Z is a proper normalization constant resulting from elimination of γ. Now we substitute the solution p^* into the Lagrangian L, which now becomes a function of λ only,

$$F(\lambda) = H(p^*) + \sum_\alpha \sum_{\{x\}} \lambda_\alpha f_\alpha(x)p^*(x) \tag{6}$$

which has to be optimized numerically , leading to the solution λ^*. According to the theory of Lagrange multipliers, the constrained optimization problem is now solved by the distribution p^* at λ^*.

3 MaxEnt in Bayesian networks

In this section, we show how the MaxEnt method under linear probabilistic constraints operates for a Bayesian network

$$p(x) = \prod_i p(x_i|\pi_i) \tag{7}$$

Again we want the maximize the entropy

$$H(p) = -\sum_{\{x\}} p(x) \log p(x) \tag{8}$$

under a set of linear constraints in p

$$\sum_{\{x\}} f_\alpha(x_\alpha)p(x) = 0 \qquad \alpha = 1 \ldots k \tag{9}$$

$$\sum_{\{x\}} f_\alpha(x_\alpha)p(x) \geq 0 \qquad \alpha = k+1 \ldots m \tag{10}$$

In which $f_\alpha(x_\alpha)$ is a function that depends on a subset of variables x_α. Introducing Lagrange multipliers λ_α for these constraints, and γ_i for normalization of the CPTs we can can formulate the optimization problem with the Lagrangian

$$L(\{p_i\}, \lambda, \gamma) = H(p) + \sum_\alpha \sum_{\{x\}} \lambda_\alpha f_\alpha(x_\alpha)p(x) + \gamma_i(\pi_i)(\sum_{\{x\}} p_i(x_i|\pi_i) - 1) \tag{11}$$

which should be maximized with respect to the CPTs $\{p_i\}$ and minimized with respect to the Lagrange multipliers. Now we cannot solve $\{p_i\}$ directly by taking the gradient of (11) with respect to all parameters, since this would only lead to a set of coupled non-linear equations for the CPTs.

What we can do, however, is taking the gradient of L with respect to a single CPT $p_i(x_i|\pi_i)$, for fixed λ and remaining CPTs$\{p_j\}_{j\neq i}$. Setting the gradient to zero, and eliminating $\gamma_i(\pi_i)$, we again get an *explicit* solution of $p_i(x_i|\pi_i)$, as a function of λ and the remaining CPTs $\{p_j\}_{j\neq i}$:

$$p_i^*(x_i|\pi_i) = \frac{1}{Z_i(\pi_i)} \exp\left\langle \sum_{\alpha\in C_i} \lambda_\alpha f_\alpha(x_\alpha) - \sum_{j\in D_i} \log p_j(x_j|\pi_j) \right\rangle_{x_i,\pi_i} \quad (12)$$

The average $\langle \ldots \rangle_{x_i,\pi_i}$ is taken with respect to the conditional distribution $p(x|x_i,\pi_i)$ (which only depend on the CPTs $\{p_j\}_{j\neq i}$). In (12), C_i is the subset of the constraints α such that the distribution of x_α depends on the state of x_i. In other words, $\alpha \notin C_i$ implies $p(x_\alpha|x_i) = p(x_\alpha)$. In a similar way, D_i is the subset of the nodes $j \neq i$, such that the child-parent combinations $\{x_j, \pi_j\}$ depends on the state of x_i. Again, in other words, $j \notin D_i$ implies $p(x_j, \pi_j|x_i) = p(x_j, \pi_j)$. Finally, $Z_i(\pi_i)$ are normalization constants for the CPTs.

Since the solution (12) is unique, it corresponds to the global maximum of L given that the other CPTs (and the Lagrange multipliers) are fixed. This means that in a sequence where at each step different CPTs are selected and updated (while keeping λ fixed) , the Lagrangian increases at each step (or remain constant). Since the Lagrangian is bounded for fixed λ we conclude that this iteration over all clusters of CPTs leads to a local maximum of L.

To find saddle points of L we propose the following two-step gradient descent procedure.

Initialization

- Initialize with random λ and random CPTs $\{p_i\}$.
- Fix λ and iterate (12) sequentially until a local maximum of $\{p_i\}$ is obtained.

λ **- step** Fix the CPTs and take a λ step into the direction of the negative gradient of the Lagrangian (some components, related to inequality constraints, will be set equal to zero if this step would push them outside their domain).

p **-step** Fix λ and iterate (12) sequentially, with CPTs initialized at their previous values, until convergence is reached.

In this way, we minimize with respect to λ in its domain, while remaining on a ridge $\partial L/\partial p_i(x_i|\pi_i) = 0$. If we converge, we obtain a local maximum of the entropy under the required constraints.

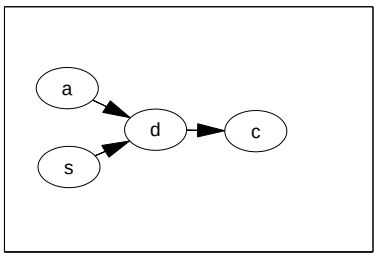


Figure 1: Structure of the network for coronary hart disease with four variables: *age* (a), *sex* (s), *hart-disease* (d), and *chest-pain* (c)

sex	age	asympt	non-AP	atyp-AP	typ-AP
m	30-39	1.9	5.2	21.8	67.7
m	40-49	5.5	14.1	46.1	87.3
m	50-59	9.7	21.5	58.9	92.0
m	60-69	12.3	28.1	67.1	94.3
f	30-39	0.3	0.8	4.2	25.8
f	40-49	1.0	2.8	13.3	55.2
f	50-59	3.2	8.4	32.4	79.4
f	60-69	7.5	18.6	54.4	90.6

Table 1: Conditional probabilities (percentages) of hart disease given age, sex and type of chest-pain (asymptomatic, non-AP pain, atypical AP-pain, typical AP-pain). This table is taken from literature and served as a constraint for the probability model in figure 1

4 An example: coronary hart disease

We illustrate the method by example involving the diagnosis of coronary hart disease, taken from [1]. In this example, we have four variables: *age* (a), *sex* (s), *hart-disease* (d), and *chest-pain* (c) . Following the example, *age* has four states (30-39, 40-49, 50-59, 60-69), *sex* has two states (*male*, *female*), *hart-disease* has two states (*true*, *false*), and *chest-pain* has four states (asymptomatic, non-AP pain, atypical AP-pain, typical AP-pain). We build a graphical structure according to figure 1.

The information that we have is a probability table $q(d|a,s,c)$ with conditional probabilities for all states of d, a, s, c, tabulated in table 1. Furthermore, there is no information, but we assume that we have the additional information that s and a are homogeneously distributed. The constraints $p(d|a,s,c) = q(d|a,s,c)$, $p(a) = 0.25$, $p(s) = 0.5$ are insufficient to uniquely specify the CPTs $p(d|a,s)$ and $p(c|d)$. We have applied MaxEnt to this problem. The CPTs that we obtained in this way are given in tables 2 and 3.

5 Conclusion and future work

If direct quantitative assessment of CPTs is to difficult for domain experts, and if other statistical information about the domain is available, but in the 'wrong direction' and insufficient to uniquely define the desired CPTs in the network, then MaxEnt for Bayesian

age	male	female
30-39	19	4
40-49	42	12
50-59	55	29
60-69	64	51

Table 2: Conditional probabilities (percentages) of hart disease (d = true) conditioned on age and sex. These CPTs are obtained by MaxEnt.

d	asympt.	non AP	atypical AP	typical AP
true	3	7	35	55
false	31	33	30	6

Table 3: Conditional probabilities (percentages) of having a certain state of chest pain (asymptomatic, non-AP pain, atypical AP-pain, typical AP-pain), given the state of hart disease (true or false). This CPTs is obtained by MaxEnt.

networks may provide a useful method for assessment of the quantitative CPTs. MaxEnt is not the only method for quantitative assessment of CPTs. Other methods have been proposed previously [2]. One of the features of MaxEnt for Bayesian networks is that the optimization procedure requires only local computations (if the constraints are local, i.e. involve only a few variables). This feature is crucial for application to large scale models.

Currently we collaborate with domain experts to study the feasibility of the construction of large scale Bayesian networks for medical diagnosis. Typically these networks will consists of several hunderds of nodes. One of the bottlenecks is the quantitative assessment of CPTs in these network, for reasons described in this paper. Our future work will include the study of the practical usefulness of the MaxEnt method for quantitative assessment of CPTs in such models.

Acknowledgements

This project is funded by the Dutch Technology Foundation STW. Jan Neijt is thanked for pointing us at the medical example.

References

[1] *Diagnostisch Kompas.* 1997.

[2] Marek J. Druzdzel and Linda C. van der Gaag. Elicitation of probabilities for belief networks: Combining qualitative and quantitative information. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, pages 141–148, 1995.

[3] F.V. Jensen. *An introduction to Bayesian networks.* UCL Press, 1996.

[4] R. Levine and M. Tribus, editors. *The Maximum Entropy Formalism.* 1979.

[5] J. Pearl. *Probabilistic Reasoning in Intelligent systems: Networks of Plausible Inference.* Morgan Kaufmann Publishers, Inc., 1988.