Interpretation of Hidden Node Methodology with Network Accuracy

Jung-Wook Bang*[†] Alexandros Pappas* Duncan Gillies
Dept. of Computing, Imperial College London, London SW7 2AZ, UK
{jbang, dfg}@doc.ic.ac.uk

Abstract

Bayesian networks are constructed under a conditional independency assumption. This assumption however does not necessarily hold in practice and may lead to loss of accuracy. We previously proposed a hidden node methodology whereby Bayesian networks are adapted by the addition of hidden nodes to model the data dependencies more accurately. Empirical results in a computer vision application to classify and count the neural cell automatically showed that a modified network with two hidden nodes achieved significantly better performance with an average prediction accuracy of 83.9% compared to 59.31% achieved by the original network. In this paper we justify the improvement of performance by examining the changes in network accuracy using four network accuracy measurements; the Euclidean accuracy, the Cosine accuracy, the Jensen-Shannon accuracy and the MDL score. Our results consistently show that the network accuracy improves by introducing hidden nodes. Consequently, we were able to verify that the hidden node methodology helps to improve network accuracy and contribute to the improvement of prediction accuracy.

1 Introduction

Bayesian Networks employ both probabilistic reasoning and graphical modelling to represent the relationships of variables in a given domain based on the assumption of conditional independence [Pearl, 1988]. However, in practice the variables may contain a certain degree of dependence and as a result the validity of a network can be questioned.

Pearl proposed a star-structure methodology to overcome the dependency problem by introducing a hidden node when any two nodes have strong conditional dependency given a common parent [Pearl, 1986][Verma and Pearl, 1991]. Pearl's idea was to simulate the common cause between two nodes by introducing a hidden node, though he did not provide a mechanism for determining the parameters of a discrete node. In some cases hidden node can be introduced subjectively through expert knowledge. However, it is not usual to have information about common causes that result in variables being partially correlated. It is therefore necessary, in many cases to use an objective method to introduce a hidden node into a network and estimate the number of states and the link matrices statistically. In neural networks, hidden layers have been widely used to discover symmetries or replicated structures. In particular, Boltzmann machine learning and backward propagation training have been proposed to determine hidden nodes [Ackley and Hinton, 1985].

Friedman proposed a technique called the Model Selection Expectation-Maximization (MS-EM) to update a network by discovering a hidden node. This approach, however, required defining the size of the hidden node prior to certain processes being carried out [Friedman, 1998].

Bang and Gillies extended Kwoh and Gillies' idea [Kwoh and Gillies, 1996] by proposing a diagonal propagation method to form a symmetric propagation scheme (*Symmetric Hidden Node Method*: SHNM) that compensated for the weakness of forward propagation in the gradient descent process [Bang and Gillies, 2002a]. This method utilized gradient descent to update the conditional probabilities of the matrices linking a hidden node to its parent's and children. Experiments in neural cell morphology showed significant improvement in performance [Bang and Gillies, 2002b]. The results showed that a modified network with two hidden nodes achieved 41.4% improvement in performance.

In this paper, we examine the hidden node methodology in terms of the network accuracy, in order to justify the performance improvement; in particular, we show that the introduction of a hidden node results in the improve-

* authors in alphabetical order
[†] contact author

ment of network accuracy and thus the improvement of prediction accuracy.

2 Hidden Node Methodology

2.1 General Concepts

Hidden nodes are introduced to a network (BN_H) by first identifying a triple (A, B, C in Figure 2.1) where the child nodes have high conditional dependency given some states of the parent node in the original network (BN_O). Once the hidden node is introduced into the network, its states and conditional probabilities are set to make B and C conditionally independent given A (BN_H). This requires the use of a representative data set with values for A, B and C.

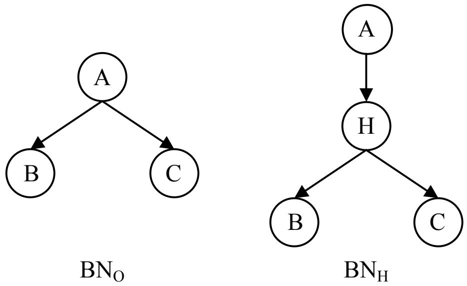

Figure 2.1 Introducing a hidden node in a Bayesian network

Having inserted the hidden node H, three conditional probability matrices (CPTs) linked to the hidden node are also created. Empirical results showed that the optimal number of states of a hidden node lies between the largest numbers of states among the other nodes (A, B and C) and two times the largest states [Bang and Gillies, 2002a].

To obtain the CPTs, we compute the derivative of the error cost function E with respect to each component of the vector $\vec{p}$ containing all the conditional probabilities. The vector derivative, $\nabla E(\vec{p})$, is called the gradient of E with respect to $\vec{p}$ and denoted as

$$\nabla E(\vec{p}) \equiv \left[\frac{\partial E}{\partial p_1}, \frac{\partial E}{\partial p_2}, \ldots \frac{\partial E}{\partial p_n}\right]$$

The training rule of gradient descent is given as

$$\vec{p}_i \leftarrow \vec{p}_i + \Delta\vec{p}_i$$

where $\Delta\vec{p}_i$ is $-\mu\nabla E$, and μ is a positive constant called the step size (or a learning rate) that determines how fast the process converges. For individual probabilities the rule is further expanded to

$$p_i \leftarrow p_i - \mu\left[\frac{\partial E}{\partial p_i}\right]$$

The objective of gradient descent is to determine iteratively the minimum error:

$$E(\vec{p}) = E_{\min}$$

or equivalently

$$E'(\vec{p}) = 0 .$$

In our case, using backward propagation the error function can be written as

$$E(\vec{p}) = \sum_{data}\sum_{x=1}^{|A|}\left[D(a_x) - P'(a_x)\right]^2$$

where |A| is the number of values of A, a_x is the x^{th} value, and the vector $\vec{p}$ contains, as its elements, all the unknown conditional probabilities in the link matrices. $P'(a_x)$ is the posterior probability of the parent node A and is calculated by instantiating the children and propagating these values through the hidden node. $D(a_x)$ is the desired value of the parent node originally from the data.

An exact gradient solution is only available in the linear cases. We, therefore, need to expand the equations to derive discrete operating equations.

2.2 Operating Equations for Gradient Descent in Bayesian Networks

The operating equations for gradient descent are derived using the chain rule to differentiate the error function. The equations for diagonal propagation are summarized.

In right-to-left propagation we instantiate root node A and child node C simultaneously. The information from the instantiated nodes propagates through hidden node H until it reaches node B. We need to determine the derivative of the error cost function $E(p)$ with respect to the three link matrix elements. For example consider $\partial E(p)/\partial P(b_j \mid h_t)$. The derivative is expanded using a chain rule as

$$\frac{\partial E(p)}{\partial P(b_j \mid h_t)} = \sum_{y=1}^{|B|}\left[\frac{\partial E(p)}{\partial P'(b_y)}\frac{\partial P'(b_y)}{\partial \pi(b_j)}\frac{\partial \pi(b_j)}{\partial P(b_j \mid h_t)}\right]$$

The first term on the right side of the above equation is the derivative of the sum of square error cost function $E(p)$ with respect to $P'(b_y)$. Differentiating $E(p)$ with respected to $P'(b_y)$ yields

$$\frac{\partial E(p)}{\partial P'(b_y)} = \sum_{y=1}^{|B|} -2\left[D(b_y) - P'(b_y)\right].$$

The second term of the equation is the derivative of the posterior probabilities of a target node $P'(b_y)$ with respect to $\pi(b_j)$. Initially the posterior probabilities are denoted as the product of the evidence of the hidden node H and the prior probability distribution of target node B, respectively.

$$\begin{aligned} P'(b_y) &= \beta\lambda(b_y)\pi(b_y) \\ &= \beta\pi(b_y) \end{aligned}$$

where the normalization factor β is $1\Big/\sum_{y=1}^{|B|}\pi(b_y)$ and $\lambda(b_j)$ has unit values. In the denominator of β the sum is taken over the states of target node *B*. The derivation of the second term yields

$$\frac{\partial P'(b_y)}{\partial \pi(b_j)} = \beta \frac{\partial \pi(b_y)}{\partial \pi(b_j)} + \pi(b_y) \frac{\partial \beta}{\partial \pi(b_j)}$$

where $\frac{\partial \beta}{\partial \pi(b_j)} = \frac{1}{\left[\sum_{y=1}^{|B|} \pi(b_y) \right]^2} = -\beta^2$.

The second term is, furthermore, extended with respect to $\pi(b_j)$ for two cases; $j = y$ and $j \neq y$.

$$\beta\delta(j,y) - \pi(b_y)\beta^2 = \beta\left[\delta(j,y) - \beta\pi(b_y)\right]$$

where $\delta(j,y) = 1$ for $j = y$, 0 otherwise.

The last term is the derivative of $\pi(b_j)$ with respect to $P(b_j \mid h_t)$. Initially we have

$$\begin{aligned} \pi(b_y) &= \sum_{s=1}^{|H|} P(b_y \mid h_s)\pi_b(h_s) \\ &= \sum_{s=1}^{|H|} P(b_y \mid h_s)\lambda(h_s)\pi(h_s) . \end{aligned}$$

Then the derivation yields

$$\frac{\partial \pi(b_j)}{\partial P(b_j \mid h_t)} = \lambda(h_t)\pi(h_t) .$$

After combing the three terms, we have

$$\frac{\partial E(p)}{\partial P(b_j \mid h_t)} = \sum_{y=1}^{|B|} \left(\begin{array}{l} \sum_{data} -2\left[D(b_y) - P'(b_y)\right] \\ \cdot \sum_{data} \beta\left[\delta(j,y) - \beta\pi(b_y)\right] \cdot \lambda(h_t)\pi(h_t) \end{array} \right) .$$

Other elements are derived similarly as follows

$$\frac{\partial E(p)}{\partial P(c_k \mid h_t)} = \sum_{y=1}^{|B|} \left[\frac{\partial E(p)}{\partial P'(b_y)} \frac{\partial P'(b_y)}{\partial \varepsilon(h_t)} \frac{\partial \varepsilon(h_t)}{\partial \lambda(h_t)} \frac{\partial \lambda(h_t)}{\partial P(c_k \mid h_t)} \right]$$

$$= \sum_{y=1}^{|B|} \left(\sum_{data} -2\left[D(b_y) - P'(b_y)\right] \cdot \beta P(b_y \mid h_t)^2 \cdot \pi(h_t) \cdot P'(c_k) \right)$$

where ε is a posterior probability of hidden node H.

$$\frac{\partial E(p)}{\partial P(h_t \mid a_i)} = \sum_{y=1}^{|B|} \left[\frac{\partial E(p)}{\partial P'(b_y)} \frac{\partial P'(b_y)}{\partial \varepsilon(h_t)} \frac{\partial \varepsilon(h_t)}{\partial \pi(h_t)} \frac{\partial \pi(h_t)}{\partial P(h_t \mid a_i)} \right]$$

$$= \sum_{y=1}^{|B|} \left(\sum_{data} -2\left[D(b_y) - P'(b_y)\right] \cdot \beta P(b_y \mid h_t)^2 \cdot \pi(h_t) \cdot P'(c_k) \right)$$

The operating equations for right-to-left propagation are found simply by swapping b and c in the above equations. Further details on the formalism and the updating process can be found in Bang and Gillies [Bang, 2002].

3 Network Accuracy in Hidden Node Methodology

3.1 Network Accuracy

In essence, a Bayesian network is a construct that represents a joint probability distribution, and can be used to model the distribution specified by a given data set. In such a case, an important characteristic of a Bayesian network is the degree to which the network models the distribution specified by the given data set accurately; the accuracy of a Bayesian network with respect to a data set. Apparently, the prediction accuracy of a Bayesian network is influenced by the network accuracy.

The accuracy of a Bayesian network can be determined precisely by evaluating the degree to which the distribution represented by the Bayesian network matches the distribution specified by the data set.

Recent work [Pappas, 2003] employs the Euclidean distance, the Cosine distance and the Jensen-Shannon divergence as measures of distributional similarity to derive different models for the accuracy of a Bayesian network.

The Euclidean inaccuracy is the geometrical distance between the points in multi-dimensional space corresponding to the distribution represented by the Bayesian network and the distribution specified by the data set.

$$Euclidean = \sqrt{\sum (P_{BN} - P_D)^2}$$

The Cosine inaccuracy is the angular separation of the vectors in multi-dimensional space corresponding to the distribution represented by the Bayesian network and the distribution specified by the data set.

$$Cosine = 1 - \frac{\sum P_{BN} * P_D}{\sqrt{\sum P_{BN}{}^2} * \sqrt{\sum P_D{}^2}}$$

The Jensen-Shannon inaccuracy of a Bayesian network is the divergence of the average of the information of the distribution represented by the Bayesian network and the information of the distribution specified by the data set over the information of their average distribution.

$$\begin{aligned} Jensen\text{–}Shannon = \frac{1}{2} \sum \left(P_{BN} * \log_2 \frac{P_{BN}}{\frac{P_{BN} + P_D}{2}} \right) \\ \text{et} + \frac{1}{2} \sum \left(P_D * \log_2 \frac{P_D}{\frac{P_{BN} + P_D}{2}} \right) . \end{aligned}$$

The accuracy of a Bayesian network can also be determined indirectly by examining alternative characteristics of the network that reflect the accuracy.

Such a model is the Minimum Description Length formalism, which models the accuracy of a Bayesian network as the likelihood of the data set given the Bayesian network, and provides the MDL score – ignoring the complexity term – as a precise measure of accuracy [Rissanen, 1978][Grunwald 1998].

$$MDL = -\sum_{d \in D} \log_2 \left[P_{BN}(d) \right]$$

3.2 Network Accuracy with a Hidden Node

The introduction of a hidden node attempts to amend the structure of the Bayesian network, so that the network no longer makes unrealistic assumptions, and thus models the dependencies accurately.

In essence, the introduction of a hidden node in the structure of a Bayesian network aims to increase the network accuracy by withdrawing the implied conditional independencies that violate the independence assumption.

In reality, the introduction of a hidden node does indeed remove inaccurate conditional independencies, but also asserts superfluous conditional independencies in connection with the introduced hidden node.

Let us consider Figure 2.1. The original structure (BN_O) implies the conditional independence of the children nodes B and C given the parent node A, indicated as BC|A. Thus, the network accuracy depends on the accuracy of that conditional independence; whether the children nodes are indeed independent given the parent node, according to the data set.

The introduction of a hidden node remove the implied conditional independence BC|A, and results in a modified structure (BN_H) that no longer implies that the children nodes B and C are independent given the parent node A.

However, the modified structure asserts a set of conditional independencies in connection with the introduced hidden node; in particular, the modified structure implies the conditional independencies AB|H, AC|H and BC|H.

The introduction of a hidden node does not necessarily result in a Bayesian network that is more accurate; the training of the hidden node and the assignment of values for the conditional probability matrices H|A, B|H and C|H should be done in such a way as to minimize the inaccuracy of the new conditional independencies implied by the modified network structure.

The accuracy for both the original Bayesian network (BN_O) and the modified Bayesian network (BN_H) can be determined precisely, by employing one of the models of accuracy mentioned in the previous section.

Since the accuracy is determined with respect to the distribution of the data set, which includes only the variables A, B, and C, the hidden node is not considered in the calculation of the network accuracy.

Therefore, the accuracy of the original Bayesian network (BN_O) is determined using the distribution of the data set (P_D) and the distribution of the original network (P_{BN0}), while the accuracy of the modified Bayesian network (BN_H) is determined using the distribution of the data set (P_D) and the distribution of the modified network over the non-hidden variables A, B and C (P'_{BNH}).

The distribution specified by the data set is:

$$P_D \equiv P_D(A,B,C) = P_D(A)P_D(B \mid A)P_D(C \mid A,B)$$

The distribution represented by the original Bayesian network is:

$$P_{BN_O} \equiv P_{BN_O}(A,B,C) = P_D(A)P_D(B \mid A)P_D(C \mid A)$$

The distribution represented by the modified Bayesian network is:

$$P_{BN_H} \equiv P_{BN_H}(A,B,C,H) = P_D(A)P(H \mid A)P(B \mid H)P(C \mid H)$$

The distribution – over variables A, B and C – represented by the modified Bayesian network is:

$$P'_{BN_H} \equiv P_{BN_H}(A,B,C) = \sum_H P_{BN_H}(A,B,C,H)$$

$$= \sum_H P_D(A)P(H \mid A)P(B \mid H)P(C \mid H)$$

4 Case Study: Neural Cell Morphology

Developmental biologists are frequently interested in classifying the development of cells in culture. In this way they can determine the effects of pollutants (or other reagents) on growth. Oligodendrocytes are a class of cell that is frequently studied. They provide the myelin sheath needed for nervous impulse conduction. Failure of these cells to develop leads to the disease multiple sclerosis. In studies, biologists view culture dishes under a microscope and attempt to count the cells using a small number of classes, for example, progenitors, immature type 1, immature type 2 and differentiated. This is a difficult, inaccurate and subjective method that could be greatly improved by using computer vision.

Our data was taken from studies in which the cultures were photographed using a Photonic Science microscope camera. Biologists classified the cells in the pictures into four developmental classes. One data set had 12 progenitor cells, 24 immature type 1, 15 immature type 2 and 9 fully differentiated cells. The images were then processed to extract several features, of which five proved to have good discriminant properties [Kim and Gillies, 1998]. These were called the Scholl coefficient [Sholl, 1953], the fractal dimension [Flook, 1978], the 2nd moment [Wechsler, 1990], the total length and the profile count.

We conducted a series of tests using the cell class (index no 6: neuron type) as a hypothesis node, and the five measured features (index no 1: Sholl coefficient, index no 2: Fractal dimension, index no 3: profile count, index no 4: Total length and index no 5: 2nd Moment) as variables. Our results with two hidden nodes (case *156* and *236*) showed significantly better performance with an average prediction accuracy of 83.9% compared to 59.31% achieved by the original network.

In addition to the prediction accuracy, the Euclidean, the Cosine and the Jensen-Shannon inaccuracy, along with the MDL score are determined for each of the Bayesian networks employed in the experiments as shown in Figure 4.1. Figure 4.1 shows the improvement ratio of prediction accuracy (far left of each case) and the improvement ratio of four network accuracy measures for five single hidden node cases. For example, case *126* represents a hidden node is introduced between node index 1 and 2 given root node 6.

Subsequently, the improvement in the network accuracy achieved due to the introduction of a hidden node is also determined. The experimental results demonstrate that the introduction of a hidden node to a Bayesian network consistently improves the network accuracy. This is due to the proper training of the hidden node, which results in a modified Bayesian network that does not violate the independence assumptions to such an extreme degree as the original Bayesian network.

The experimental demonstration of the improvement of network accuracy due to the introduction of a hidden node confirms the previous theoretical claims, and illustrates the potential benefits of the hidden node methodology in terms of the network accuracy.

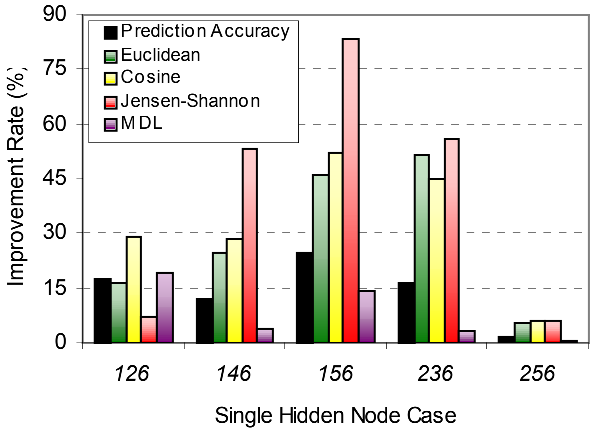

Figure 4.1 Comparison between improvement of prediction accuracy and network accuracy measures in single hidden node cases.

5 Discussion and Conclusion

In this paper, we have provided a theoretical rationale to the effects of the introduction of a hidden node within the structure of a Bayesian network. In particular, we have clarified the effects of such an action with regards to the network accuracy.

The introduction of a hidden node amends the set of conditional independencies implied by the structure of the Bayesian network. This is done an attempt to improve the network accuracy by withdrawing the implied conditional independencies that violate the independence assumption.

In computational complexity aspect, for example of naïve Bayesian networks, there are ${}^{n-1}C_2$ places that a single hidden node could be added. A second hidden node could be added at ${}^{n-2}C_2$ different places. A tree structure has fewer possibilities than the naïve case for the same number of nodes. To avoid exhaustive tests of unnecessary cases we can use the conditional dependency measure, together with the results on adding single nodes, to decide where to add further hidden nodes.

The experimental results demonstrate the improvement of network accuracy due to the introduction of a hidden node and its proper training. Furthermore, the experimental results demonstrate that the improvement of network accuracy results in the improvement of prediction accuracy.

Therefore, we have provided a theoretical and experimental justification to the empirically observed fact that the prediction accuracy improves when employing the hidden node methodology.

In our previous work, we were able to verify correlation between the improvement ratio of prediction accuracy and the degree of conditional dependency [Bang and Gillies, 2002b]. Our current results, however, show less correlation between the improvement ratio of prediction accuracy and the improvement ratio of network accuracy. This may due to the small number of tests or qualify of the network accuracy measures. We will extend our study further in investigating the relationships between prediction accuracy and network accuracy with hidden node in the future.

Other immediate study plan for a real world domain is related to bioinformatics. Lately in bioinformatics, there have been several attempts to model metabolic pathway [Angelopoulos and Muggleton, 2002]. Metabolic pathway represents the functionality of biochemical reactions within the organism and helps to understand others such as predictive toxicology. Examples of metabolic pathway can be found in KEGG (http://www.genome.ad.jp) for bioinformaticians to allow cross-reference knowledge such as the location and sequence of known genes, protein products and ligands with known reaction pathway in metabolism. One example is the aromatic amino acid pathway of yeast [Bryant *et al*., 2001]. However even one of the simplest pathways contains incomplete and incorrect information and as a result causes uncertainty. In addition metabolite(s) and enzyme(s) given a generated metabolite(s) tends to be strongly correlated and thus strongly conditionally dependent. Since each pathway is series of metabolite(s) and enzyme(s), the prediction accuracy of a network can be questionable due to the violation of conditional independence assumption.

Our hidden node methodology can be a suitable candidate to directly apply to deal with possible conditional dependency problems in metabolic networks. In addition, once a hidden node learned, it could be compared with non-counted variables to identify any unknown intermediate variable by mapping methods. Our future work will examine the possibility of the modeling of metabolic networks with introduction of hidden node methodology in Bayesian networks and identifying any unknown intermediate states.

Acknowledgments

The authors are grateful to Stephen Muggleton. This paper was partially supported by DTI project (Metalog Project), UK.

References

[Ackley and Hinton, 1985] Ackley, D.H., Hinton, G.E., and Sejnowski, T.J.: A Learning algorithm for Boltzmann machine. *Cognitive Science*, 9;147-169, 1985.

[Angelopoulos and Muggleton, 2002] N. Angelopoulos and S. Muggleton. Machine Learning metabolic Pathway Descriptions using a Probabilistic Relational Representation. *Machine Intelligence* 19, 2002.

[Bang and Gillies, 2002a] J-W Bang and D. Gillies. Estimating Hidden nodes in Bayesian Networks. *Proceedings of Int'l Conference on Machine Learning and Applications*, Las Vegas, USA, 2002.

[Bang and Gillies, 2002b] J-W Bang and D. Gillies. Using Bayesian Networks with Hidden Nodes to Recognize Neural cell Morphology. *In Proceedings of the Seventh Pacific Rim Int'l Conference in Artificial Intelligence*, LNAI, Springer-Verlag, Tokyo, Japan, 2002.

[Bang, 2002c] J-W Bang. Hidden Nodes in Bayesian Networks and their application to Prognostic Analysis of Hepatitis C. PhD Thesis, Dept. of Computing, Imperial College, London, UK, 2002.

[Bryant *et al*., 2001] C. Bryant and S. Muggleton, S. Oliver, D. Kell, P. Raiser and R. King. Combining Inductive Logic Programming, Active learning and Robotics to Discover the Function of genes. *Electronic Transaction in Artificial Intelligence*, 6(012), 2001.

[Friedman, 1998] N. Friedman. The Bayesian Structural EM Algorithm. *In Fourteenth Conference on Uncertainty in Artificial Intelligence*, 1998.

[Flook, 1978] A. Flook. The use of Dilation Logic on the Quantimet to Achieve Fractal Dimension Characterization of Textured and Structured Profiles. Powder Technology, 21, 195-198, 1978.

[Grunwald, 1998] P. Grunwald. The Minimum Description Length Principle and Reasoning under Uncertainty. PhD Thesis, ILLC Dissertation DS-1998-03, CWI, 1998.

[Kim and Gillies, 1998] J. Kim and D. Gillies. Automatic Morphometric analysis of neural cells. *Machine Graphics & Vision*, Vol. 7, No. 4, 693-709, 1988.

[Kwoh and Gillies, 1996] C-K. Kwoh and D. Gillies. Using hidden nodes in Bayesian networks. *Artificial Intelligence*, 88:1-38, 1996.

[Pappas and Gillies, 2002] A. Pappas and D. Gillies. A New Measure for the Accuracy of a Bayesian Network, MICAI 2002.

[Pappas, 2003] A. Pappas. The Accuracy of a Bayesian Network. PhD Thesis, Department of Computing, Imperial College, 2003.

[Pearl, 1986] Pearl, J. Fusion, propagation, and structuring in belief networks. *Artificial Intelligence*. 29: 241-288, 1986.

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

[Rissanen, 1978] J. Rissanen. Modeling by Shortest Data Description. Automatica, 14:465-471,1978.

[Sholl, 1953] D. Sholl. Dendritic Organization in the Neurons of the Visual and Motor cortices of the Cat. *Journal of Anatomy*, 87, 387-406, 1953.

[Verma and Pearl, 1991] Verma, T.S. and Pearl J.: Equivalence and Synthesis of Causal Models. *Uncertainty in Artificial Intelligence*, 6, Cambridge, MA, Elsevier Science Publishers, 220-22, 1991.

[Wechsler, 1990] H. Wechsler. *Computational Vision*. Academic Press Inc. 1990.