

# Qualitative Discovery in Medical Databases

David A. Maluf

Research Institute for Advanced Computer Science, Computational Science Division,  
NASA Ames. maluf@ptolemy.arc.nasa.gov

**Abstract.** Implication rules have been used in uncertainty reasoning systems to confirm and draw hypotheses or conclusions. However a major bottleneck in developing such systems lies in the elicitation of these rules. This paper empirically examines the performance of evidential inferring with implication networks generated using a rule induction tool called KAT. KAT utilizes an algorithm for the statistical analysis of empirical case data, and hence reduces the knowledge engineering efforts and biases in subjective implication certainty assignment. The paper describes several experiments in which real-world diagnostic problems were investigated; namely, medical diagnostics. In particular, it attempts to show that (1) with a limited number of case samples, KAT is capable of inducing implication networks useful for making evidential inferences based on partial observations, and (2) observation driven by a network entropy optimization mechanism is effective in reducing the uncertainty of predicted events.

**Key Words**— implication rules, data mining, medical databases, diagnostic, entropy search.

## 1 Introduction

One of the important aspects of using expert systems technology to solve real-world problems lies in the management of domain-knowledge uncertainty. Several methods of reasoning under uncertainty have been proposed in the past [1] [4] [13] [15] [17]. All these approaches require a representation of domain knowledge. Generally speaking, constructing a valid knowledge representation is a time-consuming task and often subject to opinion biases or semantics invalidity if it is built purely based on human heuristics. To overcome the difficulties in knowledge acquisition, several investigations have been carried out in recent years to explore the effectiveness and validity of automated means such as algorithms to perform this task.

Pitas *et al.* [16] have proposed a method of learning general rules from specific instances based on a minimal entropy criterion. Geiger [9] has formulated a learning algorithm for uncovering a Bayesian conditional dependence tree. This algorithm combines entropy optimization with Heckerman's similarity networks modeling scheme [10]. Cooper and Herskovits [2] have developed an algorithmic method of empirically inducing probabilistic networks, which utilizes a Bayesian framework to assess the probability of a network topology given a distribution of cases. A heuristic technique is provided to optimize the search for probable

topologies. Simulation results have shown that a small 37-node, 46-link network can be derived with 3,000 cases.

In this paper, we present a new rule-learning algorithm for inducing implication relations based on a small number of empirical data samples. The major difference between Cooper and Herskovits' approach and ours is that their approach focuses on topological induction accuracy while ours is concerned with the accuracy of inferences based on an induced network, without regards to the topological uniqueness.

Our approach to implication network induction has been implemented in a tool box called KAT, which contains several modules; namely empirical data acquisition interface, implication rule elicitation module, network validation module, optimal observation determination module, and embedded diagnostic inferring engine which implements uncertainty reasoning schemes.

Our approach to implication induction draws on the previous work on empirical construction of inference networks [5]. The present study further extends the earlier work by augmenting the implications with certainty measures. Another related work is the development of a *prediction logic* based on a contingency-table of probabilities, as proposed by Hildebrand *et al.* [11]. In their work, the emphasis was on the definition and computation of precision and accuracy of propositions represented. An analogy was made between contingency table based prediction logic and formal proposition logic.

To validate the implication networks generated from KAT, we have conducted a series of empirical experiments to examine the performance of evidential inferring with the induced networks. The chosen problem domain is *medical diagnosis*; this task shares many commonalities with other real-world problems as described in [1] [8] and has been in part inspired by earlier studies on *knowledge space theory* (KST) by Doignon and Falmagne [6]. The KST presents an interesting set-theory interpretation of knowledge states as well as its mathematical foundations. In our present framework, unlike the one by Doignon and Falmagne, the interdependencies among knowledge units are the closures under union and intersection, which can be correctly represented with a directed inference network. Hence, our implication networks representation (i.e., an instance of implication networks) can be viewed as a proper subset of the knowledge space representation.

In this paper, we examine the effectiveness and exactness of inferences with statistically induced networks. Our claim is that the proposed network induction method is capable of generating logically and empirically sound implication-based domain representations useful in predicting unobserved events upon receiving certain partial information. While validating the networks in several real-world task domains, we attempt to demonstrate the generality of the algorithmic rule induction and reasoning approach in solving problems where a complete set of events is too difficult to observe or the diagnostic judgments are subject to human errors.

## 2 Implication Network Induction

In the present work, we refer the term *implication network* to a directed acyclic graph in which the nodes represent individual event variables or hypotheses, and the arcs signify the existence of direct implication (e.g., influence) among the nodes. The value taken on by one event variable is dependent on the values taken on by all variables that influence it. Each value indicates the likelihood of an unobserved event. The value is updated every time new information is obtained (e.g., some symptom is observed). The strengths of the event interdependencies are quantified by functions (e.g., belief functions), as weights associated with the arcs.

Formally, an implication network can be represented as an ordered quadruple:

$$Net = \langle \mathcal{N}, \mathcal{I}, \alpha_c, p_{min} \rangle, \quad (1)$$

where  $\mathcal{N}$  is a finite set of nodes and  $\mathcal{I}$  is a finite set of arcs.  $\alpha_c$  is the network induction error and  $p_{min}$  is the minimal conditional probability to be estimated in the arcs. Furthermore, each induced implication rule can be specified by the following quadruple:

$$Imp = \langle N_{ant}, N_{concl}, W_I, \bar{W}_I \rangle, \quad (2)$$

where  $W_I$  and  $\bar{W}_I$  are weight functions that map the pairs of antecedent-consequent nodes, i.e.,  $N_{ant}$  and  $N_{concl}$ , and their negations to a real number between 0 and 1, respectively. That is,

$$W_I : N_{ant} \times N_{concl} \rightarrow [0, 1]. \quad (3)$$

$$\bar{W}_I : \neg N_{concl} \times \neg N_{ant} \rightarrow [0, 1]. \quad (4)$$

		$B$		$\neg B$	
		$A$	$N_{A \wedge B}$	$N_{A \wedge \neg B}$	$\neg A$

Fig. 1. contingency table where cells indicate the number of co-occurrences.

### 2.1 The Rule-Elicitation Algorithm

The basic idea behind the *empirical construction* is that in an ideal case, if there is an implication relation  $A \Rightarrow B$ , then we would never expect to find the co-occurrences as in Figure 1 that event  $A$  is true but not event  $B$ , from the empirical data samples. This translates into the following two conditions:

$$P(B|A) = 1 \quad (5)$$

$$P(\neg A|\neg B) = 1 \quad (6)$$

In reality, however, due to noise such as sampling errors, we have to relax Conditions 5 and 6. KE takes into account the imprecise/inexact nature of implications and verifies the above conditions by computing the lower bound of a  $(1 - \alpha_{\text{error}})$  confidence interval around the measured conditional probabilities. If the verification succeeds, an implication relation between the two events is asserted. Two weights are associated with the relation<sup>1</sup>, which correspond to the relations' conditional probabilities  $P(B|A)$  and  $P(\neg A|\neg B)$ . In fact, these weights together express the degree of certainty in the implication. Once an implication relation can be determined, another logical operator " $\Leftrightarrow$ " is readily defined as follows:

$$(A \Rightarrow B) \Rightarrow ((B \Rightarrow A) \Rightarrow (B \Leftrightarrow A)) \quad (7)$$

The elicitation of dependences among the nodes requires considering the existence (or nonexistence) of direct relationships between pairs of random variables in a domain model. In theory, there exist six possible types of implications between any two nodes or events.

The implication rule elicitation algorithm can be stated as follows:

### The Rule-Elicitation Algorithm

**Begin**

set an arbitrary level  $\alpha_c$  and a minimal conditional probability  $p_{\text{min}}$  (this test can be repeated for different  $\alpha_c$  and  $p_{\text{min}}$ . An example is  $\alpha_c = 0.05$  and  $p_{\text{min}} = 0.5$ ).

for  $\text{node}_i, i \in [0, n_{\text{max}} - 1]$  and  $\text{node}_j, j \in [i + 1, n_{\text{max}}]$

for all empirical case samples

compute a contingency table  $T_{ij} = \begin{bmatrix} N_{11} & N_{12} \\ N_{21} & N_{22} \end{bmatrix}$  where  $N_{11}, N_{12}, N_{21}, N_{22}$  are the

numbers of occurrences with respect to the following combinations:

$$\begin{aligned} N_{11} &: \text{node}_i = \text{TRUE} \wedge \text{node}_j = \text{TRUE} \\ N_{12} &: \text{node}_i = \text{TRUE} \wedge \text{node}_j = \text{FALSE} \\ N_{21} &: \text{node}_i = \text{FALSE} \wedge \text{node}_j = \text{TRUE} \\ N_{22} &: \text{node}_i = \text{FALSE} \wedge \text{node}_j = \text{FALSE} \end{aligned}$$

for each rule type  $k$  out of the six possible cases.

test the following inequality:

$$P(x \leq N_{\text{error\_cell}}) < \alpha_c \quad (8)$$

based on the two lower tails of binomial distributions  $\text{Bin}(N, p_{\text{min}})$  and  $\text{Bin}(\bar{N}, p_{\text{min}})$ , where  $N$  and  $\bar{N}$  denote the occurrences of antecedent satisfactions in the two inferences using a type  $k$  implication rule, i.e., in *modus ponens* and *modus tollens*, respectively.  $\alpha_c$  is the alpha error (or significance level) of the conditional probability test.

if the test succeeds

return a type  $k$  implication rule.

<sup>1</sup> With respect to the two directions of the inference, i.e. *modus ponens* vs. *modus tollens*.

```

endif
endfor
endfor

endif
End

```

Here it is assumed that the conditional probability is  $p$  in each sample, and all  $n$  samples are independent. If  $X$  is the frequency of the occurrence, then  $X$  satisfies a binomial distribution, i.e.,  $X \sim Bin(n, p)$ , whose probability function  $p_X(k)$  and distribution function  $F_X(k)$  are given below:

$$p_X(k) = \binom{n}{k} p^k q^{n-k} \quad (9)$$

$$F_X(k) = p(X \leq k) = \sum_{j=0}^k \binom{n}{j} p^j q^{n-j} \quad (10)$$

where  $p = 1 - q$ .

$A_1 \wedge B_1 \Rightarrow C_2 \vee C_3$	<table style="border-collapse: collapse; text-align: center;"> <tr> <td></td> <td style="padding: 0 5px;"><math>C_1</math></td> <td style="padding: 0 5px;"><math>C_2</math></td> <td style="padding: 0 5px;"><math>C_3</math></td> </tr> <tr> <td style="padding: 2px 5px;"><math>A_1 \wedge B_1</math></td> <td style="padding: 2px 5px;">■</td> <td style="padding: 2px 5px;">□</td> <td style="padding: 2px 5px;">□</td> </tr> <tr> <td style="padding: 2px 5px;"><math>A_1 \wedge B_2</math></td> <td style="padding: 2px 5px;">□</td> <td style="padding: 2px 5px;">■</td> <td style="padding: 2px 5px;">□</td> </tr> <tr> <td style="padding: 2px 5px;"><math>A_2 \wedge B_1</math></td> <td style="padding: 2px 5px;">□</td> <td style="padding: 2px 5px;">□</td> <td style="padding: 2px 5px;">□</td> </tr> <tr> <td style="padding: 2px 5px;"><math>A_2 \wedge B_2</math></td> <td style="padding: 2px 5px;">■</td> <td style="padding: 2px 5px;">□</td> <td style="padding: 2px 5px;">■</td> </tr> </table>		$C_1$	$C_2$	$C_3$	$A_1 \wedge B_1$	■	□	□	$A_1 \wedge B_2$	□	■	□	$A_2 \wedge B_1$	□	□	□	$A_2 \wedge B_2$	■	□	■
		$C_1$	$C_2$	$C_3$																	
$A_1 \wedge B_1$		■	□	□																	
$A_1 \wedge B_2$		□	■	□																	
$A_2 \wedge B_1$		□	□	□																	
$A_2 \wedge B_2$	■	□	■																		
$A_1 \wedge B_2 \Rightarrow C_1 \vee C_3$																					
$A_2 \wedge B_1 \Rightarrow C_1 \vee C_2 \vee C_3$																					
$A_2 \wedge B_2 \Rightarrow C_2$																					

Fig. 2. A contingency table where cells indicate the number of co-occurrences in the case of multivariate implications.

Thus, the test of hypothesis for  $A \Rightarrow B$  can be obtained by computing by a lower tail confidence interval over a binomial function:

$$p(X \leq N_{A \wedge \neg B}) = \sum_{i=0}^{N_{A \wedge \neg B}} \binom{n}{i} p^{n-i} (1-p)^i \quad (11)$$

where  $n$  has the same definition as above, and where  $p$  is set to the desired minimal conditional probability. This formula represents the probability that as small a number as  $X$  of unpredicted results would be observed if the true probability of a predicted result were exactly  $p$ . The smaller the probability given by the formula is, the less likely it is that the true probability of a predicted result is *less than*  $p$ .

From a theoretical point of view, we could increase the dimensionality of the distribution to incorporate all variables relevant to the problem in question and

allow the variables to be multivariate as illustrated in Figure 2. In such a case, the probability function to be considered becomes:

$$p_{X_1, \dots, X_r}(k_1, \dots, k_r) = \frac{n!}{k_1! \dots k_r!} p_1^{k_1} \dots p_r^{k_r} \quad (12)$$

From a practical point of view, this would also introduce exponential computational complexity. In the present study, we concentrate on *bivariate* variables *pairwise*, which reduces the scope of problem for which probabilities have to be elicited. Often this is known as *naive Bayes*.

## 2.2 An Example of Positive Implication Induction

The following section illustrates how the above algorithm is used to verify the existence of a *positive* implication rule:  $A \Rightarrow B$ .

In the first step of positive implication rule induction, a two-dimensional contingency table for variables  $A$  and  $B$  is compiled. As computed from an empirical data set, the cells in the contingency table contain the observed joint occurrences for the respective four possible combinations of values. Table 1 shows an example of the contingency table with respective co-occurrences of variables  $A$  and  $B$  in a hypothetical data set.

	$B$	$\neg B$
$A$	20 ( $N_{A \wedge B}$ )	1 ( $N_{A \wedge \neg B}$ )
$\neg A$	8 ( $N_{\neg A \wedge B}$ )	1 ( $N_{\neg A \wedge \neg B}$ )

Table 1. Distribution of observed occurrences

where  $N_{..}$  denotes the occurrences of the respective situations. The total numbers of  $A$  and  $\neg B$  can be derived accordingly as follows:

$$\begin{aligned} N_A &= N_{A \wedge B} + N_{A \wedge \neg B} = 21 \\ N_{\neg B} &= N_{A \wedge \neg B} + N_{\neg A \wedge \neg B} = 2 \end{aligned}$$

### Statistical Tests for Implication Existence

The second step of our induction method consists of an assessment of the numerical constraints imposed by  $A \Rightarrow B$ . More specifically, the assessment is based on the lower tails of binomial distributions  $Bin(N_A, p_{min})$  and  $Bin(N_{\neg B}, p_{min})$  to test measured conditional probabilities  $P(B | A)$  and  $P(\neg A | \neg B)$ , where  $N_A = N_{A \wedge B} + N_{A \wedge \neg B}$ ,  $N_{\neg B} = N_{A \wedge \neg B} + N_{\neg A \wedge \neg B}$ , and  $p_{min}$  is an arbitrary number chosen as the *minimal conditional probability* for an implication relation. For each of the two binomial distributions, we check to see whether Inequality 8 can be satisfied.

Suppose that in this example,  $p_{min} = 0.85$ ;  $\alpha_c = 0.20$ . Accordingly the binomial distribution for testing  $P(B | A)$  can be written as:  $Bin(21, 0.85)$ . The computation of the lower bound proceeds as follows:

$$\begin{aligned} P(x \leq N_{A \wedge \neg B}) &= P(x \leq 1) \\ &= P(x = 0) + P(x = 1) \\ &= \binom{21}{0} 0.85^{21} 0.15^0 + \binom{21}{1} 0.85^{20} 0.15^1 \\ &= 0.155 \end{aligned}$$

hence

$$P(x \leq N_{A \wedge \neg B}) < \alpha_c$$

where symbol  $\binom{j}{k}$  represents the number of combinations of  $k$  in  $j$ . The inference with  $A \Rightarrow B$  in the *modus ponens* direction is significant with confidence level  $(1 - \alpha_c)$ . In a similar way, given  $Bin(2, 0.85)$ , the test for  $P(\neg A | \neg B)$  yields:

$$P(x \leq N_{A \wedge \neg B}) = \binom{2}{0} 0.85^2 0.15^0 + \binom{2}{1} 0.85^1 0.15^1 = 0.98$$

hence,

$$P(x \leq N_{A \wedge \neg B}) \not< \alpha_c$$

Since Inequality 8 for the test of  $P(\neg A | \neg B)$  is not satisfied,  $A \Rightarrow B$  cannot be used for *modus tollens* inference. Hence, the positive implication rule  $A \Rightarrow B$  is rejected. The overall, worst-case time complexity of inducing an implication network with the above algorithm is  $O(n_{max}^2)$  where  $n_{max}$  is the number of nodes for modeling the domain.

### 3 Empirical Cases

This section describes the empirical data used in a series of experiments aimed to investigate the effectiveness and exactness of induced implication networks in diagnostic reasoning. The selected task domain is medical diagnosis.

In the current study, we model the different possible knowledge states by a *partial order*. Although this formalism could not fully represent all possible knowledge states, it captures a large part of the constraints on the ordering among KU and can be used for the purpose of automatic knowledge assessment [3], [7]. The data used to induce implication networks for medical diagnosis consists of a set of attributes which are continuous variables. In order to build a network, these attributes were first transformed into bivariate (i.e., binary) values using thresholds.

### 3.1 Cancer Diagnosis

The medical diagnostic method developed in this work was first validated using the empirical cancer data samples collected from 69 healthy people and 31 cancer patients. Each sample contains the information on 22 chemical residues (i.e., attributes) found in a biopsy. In order to build the network, we first transformed the ordered continuous variables, i.e., trace element concentrations, into two-valued Boolean variables, by means of thresholding.

Zn $\Rightarrow$ Mg	0.7826 0.7959	Cd $\Rightarrow$ Zn	0.7096 0.8333	Mg $\Rightarrow$ Ca	0.8823 0.8775
Zn $\Rightarrow$ Ca	0.8695 0.8775	Cd $\Rightarrow$ Ni	0.8064 0.8846	Mg $\Rightarrow$ Cu	0.7058 0.7272
Zn $\Rightarrow$ Cu	0.6956 0.7454	Cd $\Rightarrow$ Co	0.7096 0.8571	V $\Rightarrow$ Ni	0.7058 0.8076
Co $\Rightarrow$ Ni	0.7297 0.8076	Cd $\Rightarrow$ Cu	0.8064 0.8909	Cu $\Rightarrow$ Ca	0.7555 0.7755

Table 2. The original trace concentration data samples.

The derived data set was used to induce the network. Tables 3 and 4 show a few examples of the original and the derived data set samples, respectively. Table 3.1 presents a subset of the induced implication network in the form of pairwise gradation relations.

Zn	Pb	Ni	Co	Cd	Mn	Cr	Mg	V	Al	Ca	Cu	Ti	Se	Categ.
237.84	8.50	1.532	1.045	0.590	1.953	1.717	223.62	1.698	0.010	1806.75	8.71	0.732	0.001	1
203.15	12.70	2.362	1.707	0.898	1.347	1.204	46.33	0.811	4.189	405.20	13.92	0.689	0.001	1
266.34	4.44	0.085	1.013	0.382	2.151	0.340	47.73	0.010	13.137	367.92	17.10	2.896	0.003	2

Table 3. The transformed trace concentration data samples (subset).

Zn	Pb	Ni	Co	Cd	Mn	Cr	Mg	V	Al	Ca	Cu	Ti	Se	Category
01.00	01.00	01.000	01.000	01.000	1.000	1.000	01.00	1.000	0.000	01.00	00.00	1.000	0.000	1
01.00	01.00	01.000	01.000	01.000	1.000	1.000	00.00	1.000	0.000	00.00	01.00	0.000	0.000	1
01.00	01.00	00.000	01.000	01.000	1.000	1.000	00.00	0.000	1.000	00.00	01.00	1.000	0.000	2

Table 4. Examples of the induced positive implication rules (subset).

## 4 Evidential Inferences

To validate the accuracy of the evidential inferences generated from implication networks, we have conducted a series of experiments in simulated diagnostic task settings. In particular, we used constructed implication networks as the basis for evidential inferences. Each simulation run consisted of selecting a portion of a subject's sample data and propagating evidential supports throughout the network.



## 4.1 Experimental Method

There exist various interpretations of the imprecision measure associated with an implication rule [13]. Each interpretation dictates the way in which inferences are to be performed. Bayesian inference is based on the mapping of an implication relation into conditional probabilities [15]. Taking an implication  $A \Rightarrow B$  for example, updating the probability would be based upon  $P(B | A)$ , which should approach 1.0 if the implication is strong. The difficulty with this scheme stems from the fact that if further observation of  $C$  is obtained and if there is a relation  $C \Rightarrow B$ , then there is a need to update the value of  $B$  based upon  $P(B | A, C)$ , and so on. As more observations occurs, the conditional probabilities become practically impossible to estimate, whether subjectively or from sample data. To address this difficulty in a Bayesian belief network, the assumption of independence is made between individual implication relations. In the present work, we have applied the Dempster-Shafer (D-S) method of evidential reasoning to propagate supports (whether confirming or disconfirming) throughout the implication network. The D-S inferencing scheme may be regarded as a complex theoretical deviation from the Bayesian theory. According to the D-S scheme, the set of possible outcomes of a node is called the *frame of discernment*, denoted by  $\Theta$ . If the antecedents of a rule confirm a conclusion with degree  $m$ , the rule's effect on belief in the subsets of  $\Theta$  can be represented by so-called probability masses. In our bivariate case of knowledge assessment, there are only two possible outcomes for each node,  $q_i$ , that is,  $\Theta = \{known, -known\}$ .

The D-S scheme provides a means for combining beliefs from distinct sources, known as *Dempster's rule of combination*. This rule states that two assignments, corresponding to two independent sources of evidence, may be combined to yield a new one, that is,

$$m(X) = k \sum_{X_i \cap X_j = X} m_1(X_i) m_2(X_j) \quad (13)$$

where  $k$  is a normalization factor. Another evidential inference methodology, called *Certainty Factors* (CF) as previously implemented in MYCIN [1], was also applied in this study. This approach may be viewed as a special case of the D-S evidential reasoning. The two approaches differ from each other only in combining two opposite beliefs (i.e., one confirming and the other disconfirming).

## 4.2 Results in Medical Diagnosis

This section presents the empirical results of evidential inferences using the databases of cancer diagnosis instances as mentioned in Section 3. In each of the two experiments, the numeric-valued attributes were first discretized into binary values which were then used for both network induction and inferencing validation.

In the case of cancer diagnosis, 40 patient samples were compiled to induce the implication network with  $p_{min} > 0.5$  and  $\alpha_c < 0.30$ . The generated network

contains 87 implication relations. Another set of 60 patient samples was used to validate the evidential inferencing.

During the validation, a certain percentage of attributes in each test cases were randomly sampled, and the rest of the attributes were inferred from the implications. Upon the completion of inferencing, a pair of thresholds  $(u, v)$  (i.e., bi-directional thresholds) was defined to filter the numeric-valued weights. That is, if a specific node has a weight  $w > v$ , then the node is believed to be TRUE. On the other hand, if  $w < u$ , the node is believed to be FALSE (i.e., the corresponding attribute does not exist). The resulting filtered predictions were compared with the actual values in the test samples.

#### 4.3 Experiment E-5 Cancer Diagnosis

Globally speaking, given the distributions of evidentially predicted *weights* and initial *weights* with respect to various bi-directional thresholds, it can be observed that in the guessing case, both the correctly predicted nodes and the errors were almost the linear functions of the observation rate. However, in the evidential inferencing case, the shapes of these two rate profiles were changed, indicating that as the observation increased, additional nodes were added to both the correct predictions and the errors. It should also be noted that the error rates in the inferencing case were quickly stabilized after the amount of observation exceeded a certain percentage.

To further compare the results of inference-based prediction and initial weight-based guessing, a pair of bi-directional thresholds was picked up from each of the two figures such that the selected two cases would have similar error rates. At 0% sampling, the inferencing case predicted about 45% due to its conservative thresholding. However, as the observation increased, correct predictions were quickly added along with some wrong predictions. The evidential inferencing resulted in a consistently better performance in evaluating the unobserved nodes when the observation sampling exceeded 18%, as compared to the pure initial weight based guessing. For instance, at 45% observation, the inferencing method correctly predicted 4% more attributes than the guessing method.

### 5 Entropy-Driven Diagnosis Based on Induced Networks

In diagnostic reasoning, various rules may be applied to determine which node is to be observed next. One approach is to randomly choose symptom nodes from a complete symptom set that spans all the symptoms in the diagnostic structure, as studied in the previous section. Another approach is to apply entropy optimization and choose the most informative node. This section investigates the performance of entropy-driven evidential inferences based on the induced implication networks.

## 5.1 Experimental Method

In the following experiments, the expected information yield of each individual node over all the possible outcomes is computed and weighted by the likelihood of each outcome. The node that has the maximum expected information yield is chosen as the potentially most informative one, which is to be observed next. Formally, the expected information yield of an observation is defined as follows:

$$\begin{aligned}\Delta I_i &= E_{cur}(net) - E_{exp}(net) \\ &= E_{cur}(net) - [p_i E(net | node_i = TRUE) + (1 - p_i) E(net | node_i = FALSE)] \\ &= p_i \sum_{k=1}^{n_{max}} (p'_k \log p'_k + p''_k \log p''_k) - \sum_{k=1}^{n_{max}} (p_k \log p_k + p''_k \log p''_k)\end{aligned}$$

where  $E_{current}(net)$  denotes the current network entropy.  $E(net | \bullet)$  denotes the updated network entropy having observed  $node_i$ .  $p_i$  is the current probability of  $node_i = TRUE$ .  $p'_k$  and  $p''_k$  are the updated probabilities of a network  $node_k$ , having observed that  $node_i = TRUE$  and  $node_i = FALSE$ , respectively.

In what follows, we examine the diagnostic performance at the level of individual nodes. The performance is analyzed with respect to three observation modes, which are:

- (I) inferences based on the entropy-driven observation: nodes are given initial probabilities (i.e., averaged weights). If a node is observed to be TRUE, it is assigned 0.9 and 0.1 otherwise, taking into account the random error in the original data. Inference propagation is performed based on that observed node;
- (II) inferences based on random observation (as in the previous section): same as (I) but nodes are chosen at random; and,
- (III) no inference condition (or guessing): same as (II) but no inference propagation is performed.

Since the comparison between the D-S and Certainty-Factors approaches, as presented in the preceding section, does not reveal any significant performance difference, here we shall focus on the methods of observation *with the D-S evidential inferencing only*.

## 5.2 Experiment E-11 Cancer Diagnosis

This section examines the performance of evidential inferences under entropy-driven observation mode, using the empirical cancer database. For the sake of comparison, the networks to be tested in the following two experiments are the same as the ones used in the random mode observation as described in Section 4.2. During the validation, the inferred attributes were accepted based upon a pair of thresholds  $\langle u, v \rangle$  for filtering the numeric-valued weights, and then compared with the actual discretized attribute values in the samples.

Unlike the distribution mentioned before in experiment E-5, the distributions of the *weight-based-guessing-only mode* have become non-linear to the observation rate. This indicates that the entropy-driven observation tends to pick up the nodes with relatively higher uncertainty. At the same time, the inferences with entropy-driven observation added more information than the purely weight-based guessing with the same observation mode, revealing that the selection of the nodes was not based purely on the present weights of the nodes but also their connectivities in the network.

A main result may be stated that if the entropy-driven observation sampling is more than 13%, the performance of inferencing is consistently better than that of guessing. For instance, at 45% observation, the inferencing scheme produces 11% additional correct predictions as compared to the pure guesses.

It is worth mentioning that for evidential inferencing with two different observation modes, i.e., entropy-driven vs. random, the results are significantly different. In the former observation mode, the correctly predicted nodes at 45% observation can reach up to 87%, whereas the latter produces around 81% given the same amount of observation. In the random mode, it requires at least 18% observation in order for the inferencing scheme to show better performance. In the present entropy-driven mode, this percentage is further lowered to 13%.

## 6 Conclusion

In this paper, we have described a series of empirical validation experiments which examined the performance of evidential inferences based on implication networks that were algorithmically induced by a rule learning tool (KAT). In the experiments, building implication networks for evidential inferencing in various real-world diagnostic task domains (as shown in the experiments, some may have less strong implications than the others) is translated into the task of statistically induction, *from a small number of individual instances or empirical data samples* (e.g., the sizes of the samples for the experiments are respectively 47, 20, 40, and 153). Generally speaking, evidential inferencing with such induced networks is effective in generating valid predictions about unobserved events such as knowledge units and diagnostic attribute values.

This study also explored an entropy-driven diagnostic method and compared its performance with a random sampling method. The result of comparisons has shown that while both the random and the minimum-entropy-based methods are desirable, the latter is in general far better for reducing uncertainties, especially when the observation rate is more than 13% (e.g., as shown in Experiments 7, 11, and 14).

As validated in the cancer experiments, the binary representation of diagnostic attributes enables the induction of valid implication networks, which are useful not only in the predictions of unobserved attributes but also in patient diagnostic classification. The conducted experiments also reveal that the implication network is less sensitive to the particular inferencing scheme performed. In addition to the D-S and Certainty Factors schemes of evidential inferenc-

ing, we have also implemented and applied other schemes such as the Bayesian approach, with very little variation in the performance.

## References

1. B. G. Buchanan and E. H. Shortliffe, *Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project*, Addison-Wesley, Reading, MA, 1984.
2. G. Cooper and E. Herskovits, A Bayesian method for the induction of probabilistic networks from data, *Machine Learning*, 9 1992, pp. 309-347.
3. E. Degreef, J.-P. Doignon, A. Ducamp, and J.-C. Falmagne, Languages for the assessment of knowledge, *Journal of Mathematical Psychology*, 30 1986, pp. 243-256.
4. A. P. Dempster, A generalization of Bayesian inference, *Journal of the Royal Statistical Society*, 30 1960, pp. 205-247.
5. M. C. Desmarais and J. Liu, Experimental results on user knowledge assesment with an evidential reasoning methodology, in *Proceedings of International Workshop on Intelligent User Interfaces*, Orlando, FL 1993.
6. J.-P. Doignon and J.-C. Falmagne, Spaces for the assessment of knowledge, In *ternational Journal of Man-Machine Studies*, 23 1985, pp. 175-196.
7. J.-C. Falmagne, J.-P. Doignon, M. Kippen, M. Villano, and L. Johannesen, Introduction to knowledge spaces: how to build, test and search them, *Psychological Review*, 97 1990 pp. 201-224.
8. T. D. Garvey, J. D. Lowrance, and M. A. Fischler, An inference technique for integrating knowledge from disparate sources, in *Proceedings of IJCAI*, 1981, pp. 319-325.
9. D. Geiger, An entropy-based learning algorithm of bayesian conditional trees, in *Proceedings of the Eighth Conference on Uncertainty in Artificial Intelligence*, Stanford University, July 17-19 1992, pp. 92-97.
10. D. Heckerman, *Probabilistic Similarity Networks*, The MIT Press, Cambridge, MA, 1991.
11. D. K. Hildebrand, J. D. Laing, and H. Rosenthal, *Prediction Analysis of Cross Classifications*, John Wiley and Sons, New York, NY, 1977.
12. K. G. Joreskog and H. Wold, *Systems Under Indirect Observation*, North Holland, Amsterdam, 1982.
13. H. E. Kyburg, Bayesian and non-bayesian evidential updating, *Artificial Intelligence*, 1987, pp. 271-293.
14. S. K. Murthy, S. Kasif, and S. Salzberg, A system for induction of oblique decision trees, *Journal of Artificial Intelligence Research*, 2 1994, pp. 1-32.
15. J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufmann, San Mateo, CA, 1988.
16. I. Pitas, E. Milios, and A. N. Venetsanopoulos, A minimum entropy approach to rule learning from examples, *IEEE Transactions on Systems, Man, and Cybernetics*, pp. 621-635.
17. G. Shafer, *A Mathematical Theory of Evidence*, Princeton University Press, Princeton, N. J., 1976.