Exploring A New Best Information Algorithm for Iliad

Di Guo, MS (1), Michael J. Lincoln, MD (4, 1, 2), Peter J. Haug, MD (1),
Charles W. Turner, PhD (3, 1, 4), Homer R. Warner, MD, PhD (1)

University of Utah: Department of Medical Informatics (1),
Department of Internal Medicine (2), Department of Psychology (3),
Salt Lake City Veteran's Administration Medical Center (4)

Abstract

Iliad is a diagnostic expert system for internal medicine. One important feature that Iliad offers is the ability to analyze a particular patient case and to determine the most cost-effective method for pursuing the work-up. Iliad's current "best information" algorithm has not been previously validated and compared to other potential algorithms. Therefore, this paper presents a comparison of four new algorithms to the current algorithm. The basis for this comparison was eighteen "vignette" cases derived from real patient cases from the University of Utah Medical Center. The results indicated that the current algorithm can be significantly improved. More promising algorithms are suggested for future investigation.

Introduction

Description of the Current Best Information Algorithm in Iliad

Iliad is a diagnostic expert system for internal medicine, which represents the culmination of over two decades of expert systems research at the University of Utah. The system recognizes over 5000 medical findings and provides accurate medical decisions for over 1,150 diagnoses in internal medicine. Iliad uses Bayesean and Boolean knowledge frames to describe diseases encountered in internal medicine. These frames permit use of sensitivities, specificities (in Bayesean frames) and rules (in Boolean frames) to describe the relationship of a disease to its manifestations and provide a basis for explaining Iliad's conclusions [3,13].

One of the primary skills that Iliad teaches is how to best pursue a cost-effective medical work-up. Students can ask Iliad to recommend the most useful history, physical exam, or test to pursue next in their patient work-up. Iliad evaluates alternative work-up strategies by means of a "best information algorithm". This algorithm combines an information content calculation together with a cost factor. The calculations then provide a rank-ordering of the alternative patient findings according to cost-effectiveness. The information utility in the current version of Iliad is,

$$\text{Utility (F)} = \frac{\text{Prob} \times \text{Information Gain}}{\text{Cost}} \quad (1)$$

$$\text{Information Gain} = \max\{x \times (\frac{\text{sen}}{1 - \text{spec}}), y \times (\frac{1 - \text{sen}}{\text{spec}})\} \quad (2)$$

Where:

Prob = the prior probability of the frame being true (i.e., before obtaining the item of information).

Sen = true positive rate, P(F|D), spec = true negative rate, P(nF|nD).

X = degree of closeness to being true of a finding or a cluster in a disease frame, y = degree of closeness to being false of a finding or a cluster in a disease frame. For a normal finding in a disease frame, x and y are either 0 or 1. The definition of x and y is discussed more fully in appendix A.

Cost = the dollar cost to acquire the finding, the cost used in most cases is the actual charge at the University of Utah medical center. History items are set to cost $1 and physical exam items $2 [3, 10, 13].

This algorithm selects the finding with the maximum Utility (F) and identifies the corresponding hypothesis for which the finding is relevant. The current algorithm was developed and refined to provide adequate results with reasonable computational speed. Currently, disagreements exist between Iliad and our medical experts concerning optimal strategies for data collection. In this paper we investigate several new algorithms for determining information gain. The research is intended to improve the current best information algorithm and thereby ensure that students who use Iliad receive accurate training.

In this paper, we introduce four new best information algorithms. Two algorithms are derived from recent developments in information theory, and two are from reasonable intuitions, which will be called "quasi-utilities". The current algorithm and these new algorithms consider findings only within the context of each individual disease frame. They do not consider the information that a single finding could provide across several diseases. Future experiments will examine algorithms that sum information across multiple hypotheses.

0195-4210/91/$5.00

Derivation of the Four New Information Content Models

Two Models from Information Theory

The mathematical concept of information was initially developed by Shannon [2, 9, 12]. One key assumption of information theory is that a message is not significant by itself. Rather the information in a message depends on the extent to which it resolves uncertainty. Another key assumption is that the information conveyed by a series of messages is additive. One plausible equation for measuring the uncertainty, H in "bits" is given by:

$$H = -\log_2 P \quad (3)$$

Johnson [6] used -H to represent the deficit of information, or uncertainty. If the initial probability of a disease is P, the associated uncertainty is -H bits of information. In order to conclude a disease, most of those -H bits of uncertainty have to be removed. The information provided by a diagnostic observation or test, ΔH, is the difference between the diagnostic uncertainty on hand after the finding is known and the uncertainty before the finding is known. This difference can be expressed as:

$$\Delta H = \log_2 P(D|F) - \log_2 P(D) \quad (4)$$

where P(D|F) is the posterior probability of the disease after the test, and P(D) is the prior probability of the disease before the test is performed. If $I(D|F_K)$ represents the information contributed by a known result F_K, either positive or negative,

$$I(D|F_K) = \text{abs}(\Delta H) \quad (5)$$

Our first model compares the information gain contributed by either the positive or the negative finding. The model chooses the greater one, which can be expressed as:

$$I(D|F) = \max\{I(D|F+), I(D|F-)\} \quad (6)$$

Eqn (6) represents the "$\log P_2 - \log P_1$" model, and the algorithm derived from it will be called the $\log P_2 - \log P_1$ algorithm. This model calculates the information contribution from the posterior probability and the prior probability of a disease. The model also considers both positive and negative results and chooses the larger value.

Another model from the information theory can be derived by using the average amount of information or "entropy" defined by Shannon. Suppose there are a set of mutually exclusive probabilistic events (such as the presence and absence of a disease, or a group of mutually exclusive diseases). If the events are denoted as D_i and the prior probability by $P(D_i)$, the uncertainty can be described by the equation [1, 5, 8]:

$$H(D) = -\sum_i P(D_i) \log_2 P(D_i) \quad (7)$$

where H(D) represents the uncertainty or entropy and D is the set of D_i's. In our limited version of this algorithm, D_i represents either the presence or absence of a disease. The information conveyed about D by F_k (either positive or negative):

$$I(D|F_k) = \text{abs}(H(D) - H(D|F_k)) \quad (8)$$

where H(D) is the entropy before the finding is known, and $H(D|F_k)$ is the entropy after the finding is known.

There are some problems with Shannon's approach. Asch, et al. [1] noted that standard information theory fails to capture reasonable intuitions about the quantity of information provided by a diagnostic test. For instance, when the prior and the posterior probabilities are complementary (e.g. the prior is 0.1, the posterior is 0.9), the finding has provided no change in uncertainty, thus no information was conveyed. This result is counter-intuitive. To overcome the limitations of eqn (8), we have followed Pitkeathly's approach by taking account of any intervening maximum value [7]. We know that the maximum uncertainty in a system exists when the probabilities of each possible event are equal. If the function H(D) passes through a maximum when the hypothesis moves from the prior state to the posterior state, the information contributed by finding F_k (either positive or negative) can be measured by:

$$I(D|F_k) = (H_{max} - H(D)) + (H_{max} - H(D|F_k)) \quad (9)$$

where H_{max} represents the value of H(D) at the maximum between the two states. Let D represent the mutually exclusive probabilistic events: presence and absence of a disease. Substituting disease status P(D) = 0.5 and no-disease status P(nD) = 0.5 in eqn (7), we obtain $H_{max} = 1$ (bit). If the function H(D) does not pass through the maximum H_{max} due to the change of disease status (e.g. the prior and the posterior of the disease are both either greater than 0.5 or less than 0.5), $I(D|F_k)$ still follows eqn (8). Eqn (8) and (9) serve as the second model of measuring information resulted from a finding F_k (either a positive or negative finding). We used the following expression to implement eqn (8) and eqn (9),

$$I(D|F) = \max\{I(D|F+), I(D|F-)\} \quad (10)$$

I(D|F) represents the information gain resulting from a finding F to the disease D. This approach is the "Shannon" model, and the corresponding algorithm will be called the Shannon algorithm.

Two Models from "Quasi-utilities"

The model called "linear information theory" measures information by linear change of the probability of disease [1],

$$I(D|F_k) = \text{abs}(P(D|F_k) - P(D)) \quad (11)$$

where F_k is either positive or negative. The value of uncertainty of a disease D here is just the probability of the disease. The third model for measuring the information gain of a finding F is:

$$I(D|F) = \max\{\text{abs}(P(D|F+) - P(D)), \text{abs}(P(D|F-) - P(D))\} \quad (12)$$

We will call this algorithm the "$P_2 - P_1$" model, and derive the corresponding $P_2 - P_1$ algorithm.

Another model which can be characterized as a "quasi-utility" is called "weight of evidence" [4, 11]. Medical experts do not always seek clinical data based on a global

view of the differential diagnostic set. Under certain circumstances, they focus on a single diagnostic possibility and choose which information to seek in this context. Weight of evidence considers information in terms of its effect on the likelihood of a specific disease. The weight of evidence contributed by a known finding F_k (either a positive or negative) for a disease D is:

$$W(D|F_k) = \log_2 \frac{P(F_k \mid D)}{P(F_k \mid \bar{D})} \quad (13)$$

where $W(D|F_k)$ measures the contribution of a finding F_k to the diagnosis of a specific disease D as opposed to the alternate condition $\bar{D}$. This algorithm can serve as a measure of information gain contributed by a known finding for a disease. In our implementation, we used the following expression to represent the information gain by a finding to the disease D,

$$W(D \mid F) = \max \{ W(D|F+), W(D|F-) \} \quad (14)$$

We call our implementation of weight of evidence "LogLR" model, where LR stands for the likelihood ratio. When the status of a finding is unknown, the expected weight of evidence of finding F is:

$$W(D|F) = \sum_i P(F_i \mid D)\, W(D \mid F_i) \quad (15)$$

$$= \sum_i P(F_i \mid D) \log_2 \frac{P(F_i \mid D)}{P(F_i \mid \bar{D})}$$

Implementation of the Four Proposed Best Information Algorithms

The best information option allows an Iliad user to find out which item of information to obtain in order to reach a diagnosis by the most cost-effective pathway. The best information algorithm divides the information content by the cost of obtaining the data. The four proposed models were implemented in Iliad using the same function for cost as for the current algorithm but different ways of calculating information gain.

Method

Subjects Six physicians specializing in internal medicine served as the subject-judges in the experiment. These physicians were all faculty members of the University of Utah School of Medicine. They were experienced in the use of computerized expert systems for medical decision making.

Experimental Design The experiment is a 3 x 5 x 3 (Case x Algorithm x Work-up Stage) factorial design. All independent variables are within subjects factors. The two dependent variables were measures (1) of the probability of being chosen as the best algorithm and (2) of the expert's judgments about the appropriateness of the findings selected by the algorithms.

Procedure The four proposed algorithms were integrated into different versions of Iliad. We chose to compare the sequential work-up decisions of these versions to the decisions made by a group of expert internists. Three real pulmonary disease cases were selected by a pulmonary expert to provide the case material. Each case was divided into six diagnostically interesting stages. Thus, there were 18 vignettes. Each version of Iliad suggested the best data elements to seek next in each vignette. We found that the algorithms often pursued different hypotheses from each vignette's differential diagnosis. The different strategies occurred because each algorithm provided different evaluation for information content of the alternative findings. Each expert rated the suggestions from each version of Iliad and then picked the best overall combination of the hypothesis and suggested work-up. The findings were rated on a scale of 1 to 5 (1 = least cost-effective, 5 = most cost-effective). Ratings reflected cost-effectiveness of the findings proposed by an algorithm for that algorithm's hypothesis. The ratings were not based on the appropriateness of the algorithm's hypothesis. The choice of the best overall algorithm was based on the combination of the disease to be pursued and the cost-effectiveness of the question for the disease. Several algorithms might be regarded as being equally effective and might be simultaneously chosen as the best algorithm.

Results and Discussion

There are six vignettes for each pulmonary case. Those vignettes were divided into three interesting diagnostic stages: the early stage (vignette 1 and 2) denotes the preliminary steps in the case work-up, when the diagnostic certainty was less and there were many diagnostic competitors. The middle stage (vignette 3 and 4) and the late stage (vignette 5 and 6) considered later steps in the work-up when major diagnostic competitors were considered and finally eliminated. There were two outcome variables generated for each algorithm and each vignette.

The first outcome variable was the probability of being chosen as the best algorithm, which represents the proportion of experts who chose the result of that algorithm as the best of the diagnostic approaches. If three of the six physicians indicated that the work-up plan suggested by the first algorithm for a vignette was the best, then this algorithm was assigned a score of 50% for the vignette. For the statistical analysis, the scores for two vignettes at each work-up stage were averaged as explained above. There were frequent ties where two or more of the algorithms proposed the same work-up plan for the same hypothesis. In these cases, the tied algorithms were given the same score. The judges' ratings of the Best Algorithm dependent variable were analyzed using a 3 x 5 x 3 (Case x Algorithm x Work-up Stage) factorial analysis of variance. The results indicated that the main effects of the Work-up Stage [$F(2,425) = 3.08$, $p < .01$] and the Algorithm [$F(4,85) = 3.86$, $p < .006$] were statistically significant. The primary hypothesis was that the Algorithm x Work-up Stage

interaction would be significant and results supported this hypothesis, $F(8,425) = 2.23$, $p < .002$. All reported statistically significant comparisons among means were based on the Newman-Keuls multiple range procedures.

Figure 1 shows the overall probability of each algorithm being chosen as the best at different stages of the work-up (early, middle, and late). Shannon and P_2 - P_1 algorithms both performed better than the current algorithm in all three stages. The logLR algorithm and the current algorithm did not differ significantly and were chosen as the best overall algorithm equally often. The $logP_2$-$logP_1$ algorithm is better than the current algorithm in the early stage but is worse than the current algorithm in the middle and late stages. Figure 2 shows the overall (across stages) percentage of time each algorithm is chosen as the best. Because of the frequent ties the total is greater than 1.0. Newman-Keuls procedures demonstrated that the Shannon and P_2-P_1 algorithms were not significantly different in performance, and both of them were significantly better than the current algorithm. The performance of the $logP_2$-$logP_1$, logLR, and the current algorithm were not significantly different.

The second outcome variable was the "finding scores" (scale from 1 to 5, 1 = least cost-effective, 5 = most cost-effective) given by the experts for each algorithm. A 3 x 5 x 3 (Case x Algorithm x Work-up Stage) factorial analysis of variance was performed on the expert judges' ratings of the Findings Scores. The results indicated that the Work-up stage main effect [$F(2,425) = 9.98$, $p < .001$] and the Work-up Stage x Algorithm interaction [$F(8,425) = 1.74$, $p < .025$] were statistically significant. The algorithm main effect was not statistically significant, $F = 1.69$. Figure 3 shows average scores given by the experts for the work-up, independent of the hypothesis proposed by each algorithm. These scores represented the appropriateness of the finding to the hypothesis that the algorithm pursued; the scores were unaffected by how good the suggested working hypothesis was. Again, Newman-Keuls procedures indicated that Shannon and P_2-P_1 algorithms were significantly better than the current algorithm at each work-up stage (early, middle, and late). The logLR algorithm was also significantly better than the current algorithm, but the $logP_2$-$logP_1$ algorithm shows unstable performance compared to the current algorithm.

In summary, the Shannon and P_2-P_1 algorithms performed better than the current algorithm on the cases in this study. The logLR algorithm could be considered to be better than the current algorithm, but it was not as good as the Shannon and P_2-P_1 algorithms. The logLR algorithm might have the advantage of better computational efficiency. The $logP_2$-$logP_1$ algorithm performed well initially, but worsened in the middle and late work-up stages. Further study with additional cases is needed to test the generality of the results found in this study.

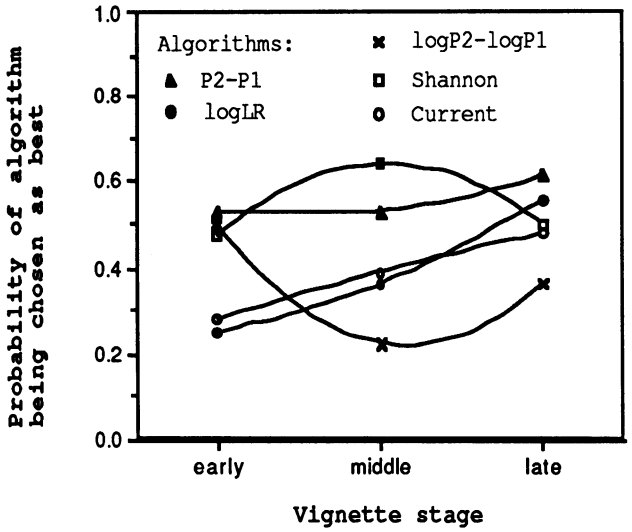

Figure 1. Probability of each algorithm being chosen by experts as the best in different stage of work-up.

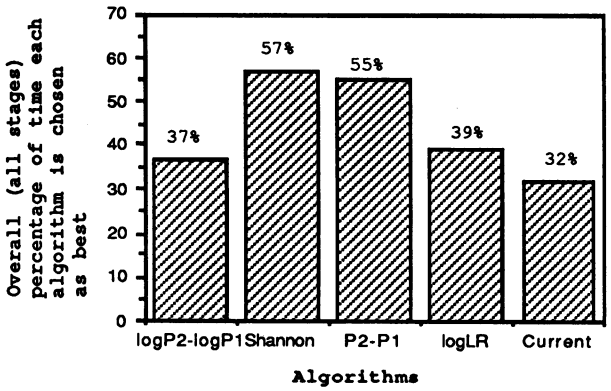

Figure 2. Overall (all stages) percentage of time each algorithm being chosen as the best by experts.

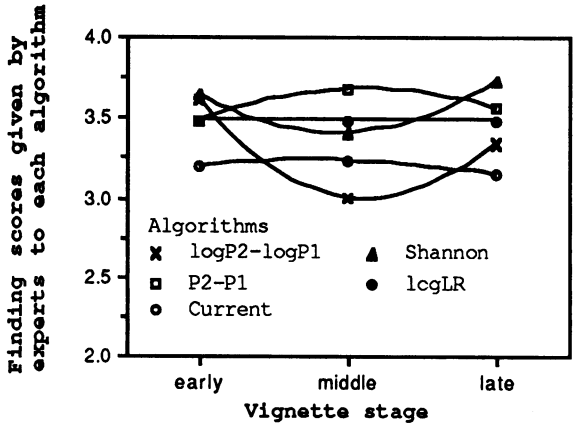

Figure 3. Average finding scores given by experts to each of five algorithms at different stage of work-up.

During the experiment, it was commonly observed that experts wanted to perform certain laboratory tests, (e.g., chest X-rays and spirometry) earlier than Iliad suggested. Iliad tended to continue to suggest history and physical questions when physicians were ready to perform the laboratory tests. This difference in work-up might have resulted from the fact that we did not include factors other than direct hospital charge in our best information algorithms. The doctor's time, the patient's length of stay, and the risk of testing were "costs" which were not considered in Iliad's best information algorithm. In medical practice, patient history and physical findings are usually acquired in a systematic fashion, and several laboratory tests may be ordered at the same time in order to shorten the hospital stay. Further study is needed to produce the algorithm that can best mimic medical expert behavior, and to evaluate the use of such a best information algorithm in teaching medical decision making skills.

Acknowledgements

The research reported in this paper was supported in part by grant number 5R01-LM-046043 from the National Library of Medicine. Special thanks to Dr. Attilio Renzetti, Dr. David Bjorkman, Dr. Frank Tyler, Dr. James Williams, Dr. Larry Reimer, and Dr. James Caldwell for their participation in the experiment. Chinli Fan and Dr. Robert Cundick also made valuable contributions to this research.

References

[1] Asch, DA, Patton, JP, Hershey, JC. Knowing for the Sake of Knowing: The Value of Prognostic Information. Medical Decision Making, 47 - 57, Vol 10/No 1, Jan - Mar 1990.

[2] Bharath, Ramachandran. Information Theory. BYTE, 291 - 298, December 1987.

[3] Cundick, R, Turner, CW, Lincoln, MJ, Buchanan, JP, Anderson, C, Warner, HR Jr., Bouhaddou, O. Iliad as a Patient Case Simulator To Teach Medical Problem Solving. Proceedings of the Symposium on Computer Applications in Medical Care, 13, Washington, DC: IEEE Computer Society Press, 902-906, 1989.

[4] Good, IJ, Card, WI. The Diagnostic Process with Special Reference to Errors. Meth. Inform. Med., 176 - 188, Vol 10/No 3, 1971.

[5] Heckerling, PS. Information Content of Diagnostic Tests in the Medical Literature. Meth. Inform. Med., 61 - 66, Vol 29/No 1, 1990.

[6] Johnson, HA. Diagnosis by the Bit: A Method for Evaluating the Diagnostic Process. Annals of Clinical and Laboratory Science. 323-331, Vol 19/No 5, 1989

[7] Pitkeathly, DA, Evans, AL, James, WB. The Use of Information Theory in Evaluating the Contribution of Radiological and Laboratory Investigations to Diagnosis and Management. Clinical Radiology, 643 - 647, Vol 30, 1979

[8] Rifkin, RD, Maximum Shannon Information Content of Diagnostic Medical Testing, Including Application to Multiple Non-Independent Tests. Med. Decis. Making, 179 - 189, 5(2), 1985

[9] Shannon, CE, Weaver, W. The Mathematical Theory of Communication Urbana IL: Univ. of Illinois Press, Chicago (1949).

[10] Sorenson, DK, Cundick, RM, Fan, C, Warner, HR. Passing Partial Information among Bayesean and Boolean Frames. Proceedings of the Symposium on Computer Applications in Medical Care, 13, Washington, DC: IEEE Computer Society Press, 50-53, 1989.

[11] Thornbury, JR, Fryback, DG, Edwards, W. Likelihood Ratios As a Measure of the Diagnostic Usefulness of Excretory Urogram Information. Diagnostic Radiology, 561 - 566, Vol 114, 1975.

[12] Usher, MJ. Information Theory for Technologists London, UK: Macmillan Publishers, 1984

[13] Warner, HR, Haug, P, Bouhaddou, O, Lincoln, M. Warner, HR. Jr, Sorenson, D, Williamson, JW, Fan, CL. Iliad As An Expert Consultant to Teach Differential Diagnosis. Proceedings of the Symposium on Computer Applications in Medical Care, 12, Washington, DC: IEEE Computer Society Press, 371 - 376, 1988.

Appendix A

The calculation of x and y values for deterministic clusters (nested Boolean frames) and probabilistic clusters (nested Bayesean frames) is discussed below.

Calculation of x and y for a deterministic cluster. A Boolean frame is designed as a decision module built around a Boolean relationship among its findings. Any one or some combination of findings in the frame may be sufficient for the frame to come true or false. When there is not enough information to make the frame true, we use two terms, x and y to express the true state and false state respectively [1, 9]. For example, if the logic is "true if 3 of (A,B,C,D)", and assume A and B are true, C is false, D is unknown and each item in the logic has the same frequency. If only one item is true, and other items are unknown in the logic, there is 1/3 of what is needed to be true. Since A and B are true, x = 0.67. The negative logic (derived from the "true" statement) for this cluster would be "false if 3 of (A, B, C, D) are false". The calculation of y is similar to the above calculation. Here C is false, y = 0.33. This is a simple case. When items in the logic have different frequencies, normalization is done to calculate x and y.

Calculation of x and y for a probabilistic cluster. If AP is the prior probability of the cluster before a finding is known, P is the posterior probability of the cluster after a finding is known, the rule is:

if $P > AP$ $\quad x = (P - AP) / (1 - AP), \quad y = 0$

if $P < AP$ $\quad x = 0, \quad y = (AP - P) / AP$