A Qualitative Expert System for Clinical Trial Assignment

Sanjukta Bhanja, Lynn M. Fletcher-Heath, Lawrence O. Hall, Dmitry B. Goldgof and Jeffrey P. Krischer†
Computer Science and Engineering
†Moffitt Cancer Center and Research Institute
4202 E. Fowler Avenue
University of South Florida
Tampa, FL 33620
E-mail: hall@csee.usf.edu

Abstract

Assigning patients into clinical trials is a knowledge and data intensive task. Determining the eligibility of a patient for admission into a clinical trial is based upon specific criteria. These criteria may be shared among several protocols or may be unique to one protocol. A major difficulty in assigning patients to clinical trial protocols is the absence of complete information regarding the patient. Much of the needed data can be time-consuming or expensive to obtain, or needed tests can cause pain or discomfort to the patient. Another difficulty is that there are many open trials at an institution at any one time and it is very difficult to keep track of criteria for each trial. This paper investigates the use of a fuzzy expert system joined with a dependency analysis to handle uncertainty and sort needed data for several protocols in order of influence. The system's output is an evaluation of the patient's eligibility for one or more clinical trials. Preliminary tests show that the system presents important data as high priority data while finding an appropriate order to obtain all needed data. We have implemented four breast cancer protocols and successfully tested 15 cases which were clinically eligible for one of the four protocols.

Introduction

The purpose of a clinical trial is to evaluate new treatments for disease. Each trial is designed to increase scientific understanding of therapy and to find better ways to help patients recover from disease. Patients may participate in a clinical trial if certain criteria are met involving demographics, laboratory results, symptoms, physical findings, current and prior medications and treatments, and drug allergies. However, obtaining enough patients for a given trial to enhance statistical significance may be hindered by a large number of open trials. It becomes difficult for a clinician to recall the exact eligibility requirements for every clinical trial for which a patient may be eligible. For example, there may be information that rules a patient in or out of one trial, but the person doing the assignment may be considering a different trial entirely. Therefore, the need for a fast and efficient method of assigning patients to clinical trials is the main goal for this work; a problem well-suited for an expert system (Giarratano & Riley 1994).

Assigning patients into clinical trials is a task requiring certain data about the patient and knowledge about clinical trials (Tu *et al.* 1993). Eligibility determination for inclusion in a clinical trial is based upon specific inclusion and exclusion criteria. This paper describes an attempt to solve this problem using a qualitative expert system. The expert system uses a fuzzy rule-base since there may be missing data or uncertain data due to the imprecision of modeling medical test results, signs and symptoms. (Zimmerman 1991; Jang & Gulley 1995). The fuzzy rule-base will be covered in the first subsection of Interface and Algorithms. In addition, data needed by the system to determine if a patient is eligible for a protocol, is listed in order of importance. This prioritization is done using dependency analysis via graphs of the fuzzy rules and the facts available, which will be discussed in the Dependency Analysis subsection of Interface and Algorithms.

Background

Inference from rule-bases has traditionally been handled in various ways. The rule-base can be probablistic or possiblistic (Kruse & Borgelt 1997). Analysis may be used to predict the relative importance or sensitivity of each rule in the rule-base (Laskey 1995). Some methods use Bayesian belief networks (Laskey 1995; Theocharous 1996) or Dempster-Shafer's belief networks (Wang & Valtorta 1992) rather than rules. Our goal was to use a simple tool which would enable us to compute the relative influence of each fact on the patient's eligibility score.

There has been little research on building systems to determine patient eligibility for clinical trials. Mostly, there are two general approaches taken, a probabilistic one using Bayesian belief networks (Pearl 1988; Wang & Valtorta 1992; Ohno-Machado *et al.* 1993; Laskey 1995; Theocharous 1996) and a qualitative approach (Zimmerman 1991; Tu *et al.* 1993). The dif-

ficulty in using a Bayesian network is that it requires large numbers of exact conditional probabilities to be specified as the number of network nodes increases. The number of probabilities may be so large that it may be necessary to apply learning techniques to get acceptable values for the prior probabilities (Heckerman 1995; Theocharous 1996). Moreover, while implementing multiple protocols, a very complex network may be created. Hence the issues are scalability, time and design complexities when using probabilistic approaches such as Bayesian belief networks.

Another of our goals was to ensure that the developed system for clinical trial eligibility scales easily to groups of protocols for the same area, such as all open protocols for breast cancer. Therefore it should be easy to incorporate new protocols and remove closed protocols as is possible while pursuing a qualitative approach. Some exploration of the qualitative approach to clinical trial eligibility assessment has helped determine that scaling up may best be accommodated by a qualitative system (Tu *et al.* 1993). Also, in work with fuzzy qualitative models, it has been found that small changes in the models do not perturb the results, attesting to their stability (Berenji & Khedkar 1992; Wang 1994). The fuzzy qualitative models provide the possibility of flexible and robust decision making which forms the basis for our research.

The Rete algorithm implements a rule-base using acyclically directed dependency graphs (Nayak, Gupta, & Rosenbloom 1988). The nodes in the graph represent variables or patterns and the links represent the antecedent of the rules. The algorithm we propose is conceptually similar since relevant rule-base information is also stored in the links and nodes of the graph. This paper describes an implementation where the nodes represent fuzzy facts and intermediate steps within each rule, which will be discused in more detail in the next section. Also, when a fuzzy fact is introduced into the system that information is propagated to effect all relevant nodes through the links.

The combination of dependency analysis with a fuzzy qualitative model allows the system to determine which facts are needed to improve the eligibility score (Nayak, Gupta, & Rosenbloom 1988; Kusiak & Wang 1995; Chandwani & Chaudhury 1996). Those facts are then ordered by potential impact on the decision to include or exclude the patient in a protocol adding to the efficiency of the system.

Interface and Algorithms

A web-based interface is used to collect patient data from a user and offer patient eligibility information based on that data. The fuzzy qualitative model was implemented using a fuzzy rule-base where each rule is either a conjuction or a disjunction of fuzzy facts. The dependency analysis examines the relationships of chained facts, represented as nodes of a graph, built with the antecedents and consequences of the fuzzy rules. The details of both the fuzzy qualitative model and the dependency analysis of this system are discussed in the following subsections.

Fuzzy Qualitative Model

Fuzzy rules provide a means for modeling the imprecision of medical signs, symptoms and test results) (Zimmerman 1991; Jang & Gulley 1995). The fuzzy rules used to determine clinical eligibility were encoded in the fuzzy expert system tool, fuzzy CLIPS (Orchard 1995). The advantages of using fuzzy CLIPS are that it is easily expandable with user-defined C functions, portable, and it executes quickly.

The fuzzy rules each provide some partial membership for a patient as either eligible or ineligible for a clinical trial protocol. At each point in the process of determining eligibility, a membership score within the range of [0,1] is available. A membership value of 0 reflects no membership in the eligibility class while 1 is full eligibility within the eligibility class for the trial. Although the person entering information into the system can decide when an eligibility score is high enough, a threshold membership value of 0.75 in the eligibility class was used in our tests to indicate that a patient is eligible for a protocol.

An example of two fuzzy rules with the same goal node are shown in Figure 1. Since both rules lead to the same conclusion, either or both rules could influence the eligibility of a patient to Protocol 1. Let us examine this example in detail. The rule on the left side of the graph, a AND b AND c → Protocol 1, uses information from three leaf nodes, a,b and c. The fuzzy values for a and b have already been determined, denoted by the same values for the lower and upper bounds. Although node a has a value of 1, node b only has a value of 0.2. If we propagate the known information to the intermediate node abc_{AND}, note that the minimum values of the lower and upper bounds are chosen. This means that any new information obtained from node c will not increase the value of abc_{AND}. Since low values (based on some predefined threshold) will not add to the eligibility of a patient to Protocol 1, there is little need to determine the value of node c. The rule on the right side of the graph, e OR d → Protocol 1, uses information from two leaf nodes, d and e. Node d has been predetermined with a value of 0.5. Since this rule uses an OR node, the maximum bound values will be propagated to de_{OR}. Presently, the range for de_{OR} is above all possible values of abc_{AND}, so the final range passed to the goal node Protocol 1 (also an OR node) is received from Rule 2. Therefore, any further information from Rule 1 would indeed be meaningless.

Our implementation strategy allows for ease in knowledge maintenance. Since each protocol is implemented as a separate set of rules, the ability to modify the rules of one protocol exists as that protocol may evolve over time. Also, this modular approach will make for simpler addition of a new protocol, as well as removal of a protocol that is closed to patient accrual.

With each protocol implemented as a separate set

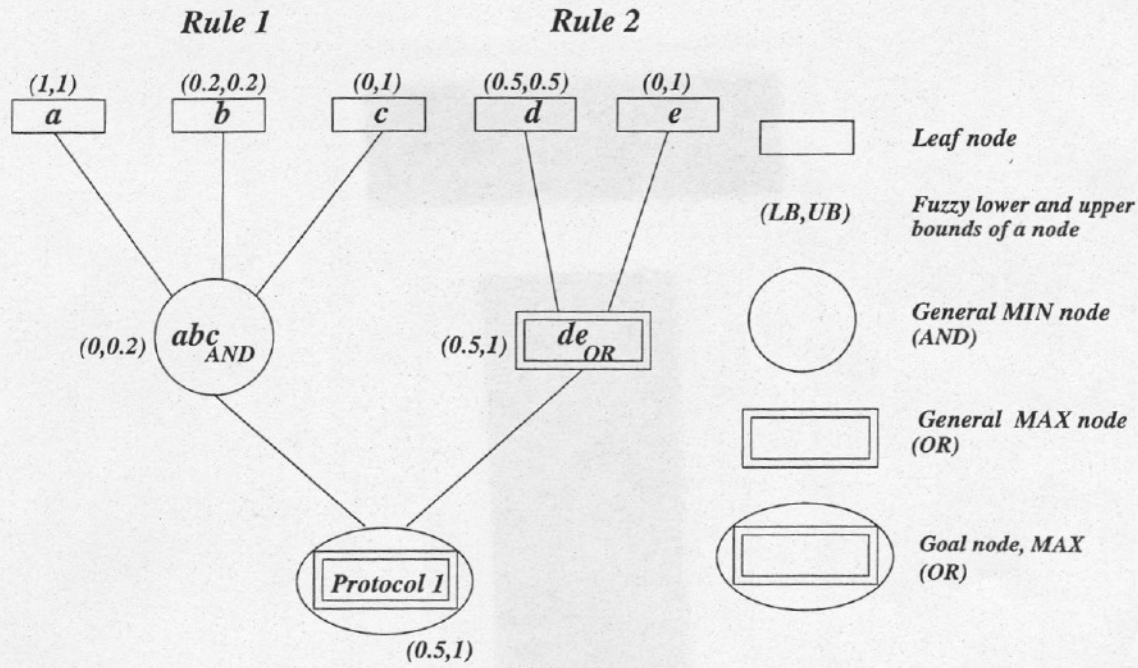

Figure 1: Fuzzy rule-base with two rules leading to the same conclusion. Rule 1: a AND b AND c → Protocol 1. Rule 2: e OR d → Protocol 1. The fuzzy AND is implemented as a minimum function while the fuzzy OR is implemented as a maximum function. Therefore the fuzzy bounds reported to the intermediate nodes (abc_{AND} and de_{OR}) are the minimum lower and upper bound values for the AND node and the maximum lower and upper bounds for the OR node. Similarly, the fuzzy values reported to the goal node (Protocol 1) are the maximum bound values of the intermediate nodes since the goal node is also an OR node.

of rules, a separate instantiation of CLIPS is provided. The needed facts for each instantiation are merged into one set of grouped facts which will then be sorted in order of priority during the dependency analysis.

Dependency Analysis

Again, Figure 1 demonstrates that some facts may not need to be collected for conclusions about the goal node to be made. For example, since the abc_{AND} node from Rule 1 takes the minimum value of all three upper bounds and the minimum value of all lower bounds, and node b has previously been determined as having a fuzzy value of 0.2, the upper bound of abc_{AND} can be no larger than 0.2. Given this information there is no need to acquire the value of node c since it will never match or exceed the minimum value for the de_{OR} node from Rule 2. A general description of the impact of OR nodes and AND nodes is: unknown facts at an OR node can raise the eligibility membership value bound to 1, but unknown facts at an AND node can lower the membership value for that node bound to 0. Similarly, other considerations may be made which allow for efficient and meaningful data collection.

The analysis described above is applied to all unknown facts in the graph allowing them to be ranked within the bound membership values. For example, Figure 2 demonstrates the technique used to rank the needed facts in order of relative impact on the goal node. Nodes between the leaf nodes and the goal node are referred to as intermediate nodes. The intermediate node for Rule 1 is abc_{AND}, implemented with the minimum function, shown with a value bound by (0,1). Although node a has a fuzzy membership value of 1, nodes b and c must be determined before abc_{AND} can be fully determined. In contrast, since the intermediate node for Rule 2 is an OR node, implemented with the maximum function, a high fuzzy value from either d or e could positively impact the goal node. However, node d has been pre-determined with a value of 0. Therefore, information regarding node e will cause Rule 2 to fire, activating the goal node with a fuzzy value of the maximum of d or e. Hence, there is only one fact needed to impact the goal node via Rule 2 so node e is labeled as the most important fact to obtain.

The dependency graph is implemented as a weighted directed graph using an adjacency list. This allows the record keeping necessary to associate the relationships between nodes. For example, we must keep track of each node's parent so that we may easily conduct a depth first search from the goal node, looking for the number of unknown leaf nodes in the goal node's subtree. From this search the impact of each leaf node on the goal node may be determined.

There are two major rules when sequencing the unknown leaf nodes in order of greatest impact. First, a subtree with several unknown leaf nodes is more expen-

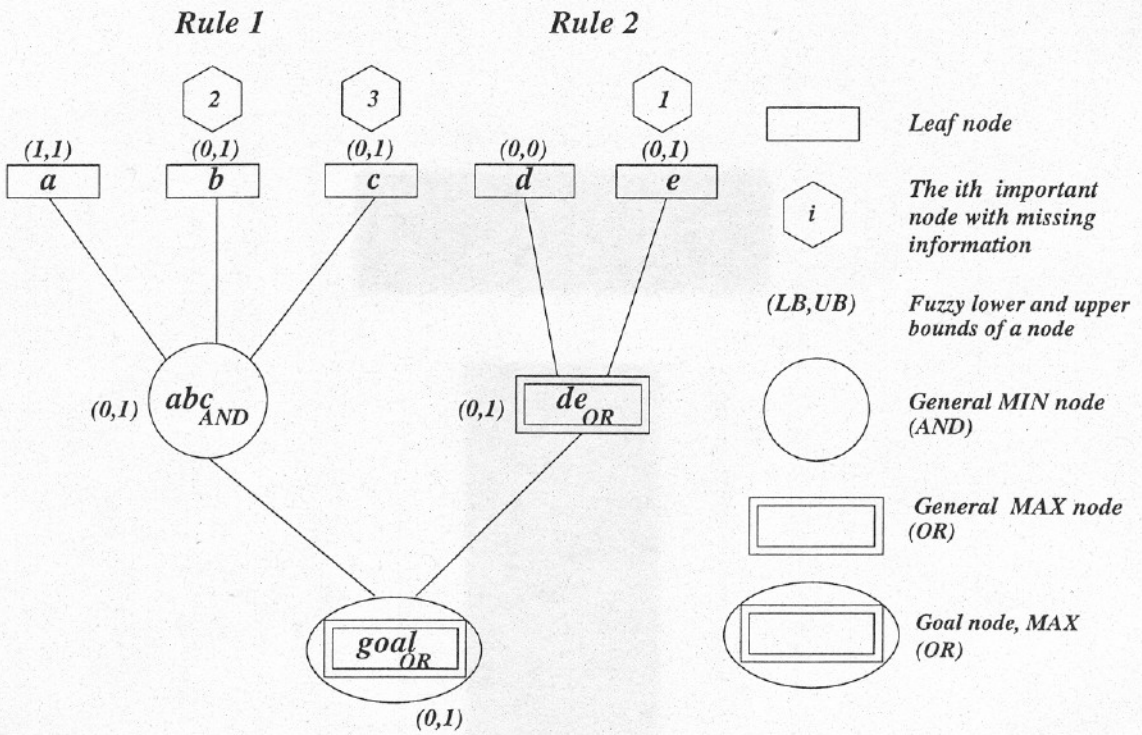

Figure 2: Rule-base with two rules each leading to the same goal node. Rule 1. a AND b AND c → goal. Rule 2. e OR d → goal. This dependency graph demonstrates the technique used to rank the needed facts in order of relative impact on the goal node. Since the intermediate node for Rule 2 is an OR node, a high fuzzy value from either d or e could positively impact the goal node, however, node d has been pre-determined with a value of 0. Therefore, information regarding node e will cause Rule 2 to fire, activating the goal node with a fuzzy value of the maximum of d or e. Hence, there is only one fact needed to impact the goal node via Rule 2 so node e is labeled as the most important fact to obtain. In contrast, since the intermediate node for Rule 1 is an AND node, the minimum fuzzy value will be propagated to the goal node. Therefore, all three leaf nodes of de_{OR} must be considered before propagating the minimum fuzzy value to the goal node, demanding nodes b and c be determined second and third, not necessarily b before c.

sive to explore than a subtree with fewer unknown leaf nodes. Second, the antecedent leaf nodes of an AND node do not need to be fully determined if one of the known leaf nodes has a value of 0, or if the known value of that leaf is less than the minimum value of a competing intermediate node. Similarly, the antecedent leaf nodes of an OR node do not need to be fully determined if one of the known leaf nodes has a value of 1, or if the known value of that leaf is more than the maximum value of a competing intermediate node. Once the unknown leaf nodes (needed patient data) have been properly sequenced and displayed through the web interface, the user may provide values to those questions if the information is known. Then at any time the eligibility membership may be recalculated and displayed.

Results and Discussion

One parameter on which our system may be tested is how effective our approach prioritizes the needed facts, judged by how effectively it reduces the number of questions a user must answer. We tested the system on two breast cancer protocols. Table 1 shows results from two rule-bases for which the dependency analysis was tested with random organization of facts and cases. There were 50 test cases generated for each protocol in two ways. The first group represents when the facts are randomly generated and the second represents the usage of the dependency analysis to suggest the most important facts. In both cases, the fuzzy value for each fact is selected randomly. The average number of facts needed to determine eligibility and standard deviation is shown in Table 1.

It is clear that answering the ordered facts results in significantly less necessary facts before a decision is made. The standard deviation using the ordered facts is also significantly less for both protocols. This data indicates that the dependency analysis is effective in ordering the facts needed, thereby reducing the number of questions asked before the goal node is reached. However, when we have a large set of alternative parents with equal impact on the child node, it may choose a deciding fact (one which rules in or out a goal) with an appropriate value. In this case, a random choice of facts may work better than the sorted one.

	Protocol 1		Protocol 2	
Order of Facts	Random	Sorted	Random	Sorted
Avg. Needed	7.72	4.26	5.24	2.86
Std. Dev.	6.45	2.02	3.09	1.01

Table 1: Performance of Dependency analysis algorithm over 50 cases on two rule-bases.

For one of the protocols implemented, there were 15 available clinical examples from patients who were entered into the protocol. We have tested our system on these 15 cases and found that each patient was found to be eligible for this protocol and not the other three implemented protocols. This eligibility was based on a minimum membership threshold of 0.75. The membership values for all other protocols were 0.5 or less. Therefore, our system successfully classified the available 15 clinical test cases.

Conclusion

The dependency analysis aids in the assignment of patients into clinical trials in an effective manner. The dependency analysis algorithm helps to sort missing information from patients with respect to their impact on the protocol eligibility. Moreover, this algorithm recognizes when acquiring further facts would not aid in the analysis once the maximum possible fuzzy value of the goal node reaches below the acceptance threshold. The algorithm has the potential to sort the important nodes with respect to other factors such as cost and discomfort along with the causal impact on the goal. We are in the process of testing this algorithm on multiple clinical protocols while comparing results with those from medical experts.

Acknowledgements

This work was supported in part by the Moffitt Cancer Center. Many thanks to Dr. Susan Minton for her help in system analysis and to Margaret Gross-King for her help in data collection.

References

Berenji, H., and Khedkar, P. 1992. Learning and tuning fuzzy controllers through reinforcements. *IEEE Transactions on Neural Networks* 3(5):724–740.

Chandwani, M., and Chaudhury, N. S. 1996. Knowledge representation using fuzzy deduction graph. *IEEE Transactions on Systems, Man and Cybernetics* 26(6):848–854.

Giarratano, J., and Riley, G. 1994. *Expert Systems: Principles and Programming.* Boston, MA: PWS Publishing Company.

Heckerman, D. 1995. A tutorial on learning baysian networks. Technical report, Microsoft.

Jang, J. S. R., and Gulley, N. 1995. *Fuzzy logic toolbox.* Natick, MA: The Math Works, Inc.

Kruse, R., and Borgelt, C. 1997. *Learning probabilistic and possibilistic networks : Theory and applications.* Parague: Seventh IFSA World Congress.

Kusiak, A., and Wang, J. 1995. Dependency analysis in constraint negotiation. *IEEE Transactions on Systems, Man, and Cybernetics* 25(9):1301–1313.

Laskey, K. B. 1995. Sensitivity analysis for probability assessments in bayesian networks. *IEEE Transactions on Systems, Man and Cybernetics* 25(6):901–909.

Nayak, P.; Gupta, A.; and Rosenbloom, P. 1988. Comparison of the rete and treat production matchers for soar (a summary). In *Proceedings of AAAI*, 693–698.

Ohno-Machado, L.; Parra, E.; Tu, S. W.; and Musen, M. A. 1993. Aids2: A decision-support tool for decreasing physicians uncertainty regarding patient eligibilty for hiv treatment protocols. In *Annual Symposium Applied Medical Care*, 429–433.

Orchard, R. A. 1995. *FuzzyCLIPS Version 6.04.* Canada: National Research Council.

Pearl, J. 1988. *Probalistic Reasoning in Intelligent Systems.* San Mateo, CA: Morgan Kaufman.

Theocharous, G. 1996. An expert system for assigning patients into clinical trials based on bayesian networks. Master's thesis, University of South Florida.

Tu, S.; Kemper, C. A.; Lane, N. M.; Carlson, R. W.; and Musen, M. A. 1993. A methodology for determining patient's eligibility for clinical trials. *Methods of Information in Medicine* 32:317–325.

Wang, S., and Valtorta, M. 1992. The conversion of rule bases into belief networks. *ACM Computing Surveys* 363–368.

Wang, L. X. 1994. *Adaptive fuzzy systems and control: Design and stability analysis.* NY, NY: Prentice-Hall.

Zimmerman, H. 1991. *Fuzzy set theory and its applications.* Boston, MA: Kluwer Academic, second edition.