Stepwise Knowledge Acquisition in a Fuzzy Knowledge Representation Framework

Thomas E. Rothenfluh[1], Karl Bögl[2], and Klaus-Peter Adlassnig[2]

[1]Department of Psychology
University of Zurich, Zürichbergstraße 43, Zurich, Switzerland
e-mail: thomas.rothenfluh@access.unizh.ch
[2]Department of Medical Computer Sciences, Section on Medical Expert and Knowledge-Based Systems
University of Vienna Medical School, Spitalgasse 23, A-1090 Vienna, Austria
e-mail: karl.boegl@akh-wien.ac.at

Abstract. *We describe the knowledge acquisition process that is used in MedFrame/CADIAG-IV, a medical computer consultation system. Fuzzy medical knowledge is used to model the vagueness and the uncertainty of medical concepts and fuzzy logic reasoning mechanisms provide the basic inference engines. Knowledge acquisition procedures and computer tools have been implemented in order to make the tasks of (a) defining medical concepts, (b) providing appropriate interpretations for patient data, and (c) constructing inferential knowledge easier and more accessible. This paper explains how the knowledge acquisition tasks are supported both by special representations and by a stepwise knowledge acquisition process.*

1. Introduction

MedFrame/CADIAG-IV is a fuzzy medical consultation system that provides diagnostic hypotheses and therapeutic suggestions based on symptoms, signs, test results, and clinical findings of a patient. It is a successor of earlier implementations of CADIAG expert systems that have used Boolean logic, first-order predicate calculus formulas, and fuzzy sets [1]. MedFrame/CADIAG-IV extends its predecessors in that fuzzy representations are used for almost all representations of knowledge: it accepts (or fuzzifies) its inputs, operates on fuzzy sets with fuzzy rules, and produces fuzzy sets or defuzzified values as output.

The major application domain of MedFrame/CADIAG-IV is internal medicine. Within an elaborate representation and inferencing framework [2], medical concepts such as symptoms, findings, diagnoses, and therapies have to be entered. These basic entities are then connected together with inference relations that allow the system to infer new interpretations based on real patient data.

While the definition of knowledge structures and the basic relations between entities is provided by knowledge engineers and system designers, the expert medical knowledge has to be acquired from domain experts. The knowledge acquisition process has thus to be supported by various means that help physicians to “translate” their expert knowledge into computational representations. Although

medical scientific literature is full with fuzzy statements and qualifications, there is usually no simple translation from experts' linguistic statements of uncertainty into computer representations. Thus, special support is needed to allow physicians to express and refine their expert knowledge. This paper discusses a stepwise knowledge acquisition process that leads physicians from entering basic associative knowledge to manipulating fuzzy membership functions (for more detailed information, examples, and implementation issues see [3] or [4]).

2. Fuzzy Representation

2.1. Basic Fuzzy Representations

MedFrame/CADIAG-IV uses *fuzzy sets* that define the degree of membership for an element in a set (for definitions of fuzzy terminology see [5]). *Fuzzy numbers* are used to specify fuzzy membership functions $\prod(x;\alpha,\beta,\gamma,\delta)$ and are constrained to values in [0,1] (Figure 1). Additionally, a special value representing 'unknown' is provided to distinguish an element that is known to have some, but (yet) unknown, membership relation to a set from other elements whose membership degree has not yet been assessed at all.

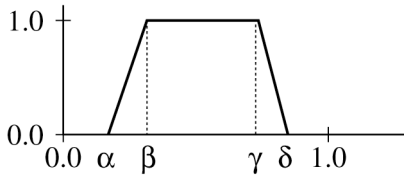


Figure 1: Graphical representation of fuzzy membership function $\prod(x;\alpha,\beta,\gamma,\delta)$.

Membership functions can be represented (and acquired, as will be explained in a following section) through numerical or graphical representations. Transition functions between endpoints (between α and β or between γ and δ) can accept linear or other shapes.

Type 2 fuzzy sets are fuzzy sets whose degrees of memberships are themselves fuzzy sets. They are applied whenever combinations of fuzzy elements are used.

Fuzzy relations between two (ordinary) sets are defined as the fuzzy set of the Cartesian product between the elements of those sets. Every element in this set is characterized by a membership function. Since fuzzy numbers are used to define the basic sets themselves (see next section), fuzzy relations in MedFrame/CADIAG-IV's knowledge base are essentially a combination of type 2 fuzzy sets.

Whenever appropriate or desired, the system can defuzzify or approximate its statements to crisp values or defined sets with the help of user-definable thresholds and specific, domain-dependent algorithms. For example, rank-ordered lists of confirmed, possible, and excluded diagnoses can be produced in order to help physicians to direct their next examination steps.

2.2. Knowledge Types

In MedFrame/CADIAG-IV, we distinguish between two basic knowledge types that are defined in the knowledge representation framework: (a) *medical entities* represent findings, diseases, and therapies as the basic building blocks for all possible statements about medical concepts and (b) *medical data* that describe quantitative medical concepts such as measurements, results from physical examinations, and laboratory data (e.g., height, duration of morning stiffness, serum glucose levels).

Since MedFrame/CADIAG-IV's reasoning mechanisms operate at the level of symbolic concepts (i.e., medical entities), a data-to-entity conversion has to be employed to transform quantitative medical data into medical entities. The transformation of medical data into meaningful interpretation categories (medical entities) can be compared to the definition of a linguistic variable in other fuzzy systems. This definition has to be established in the knowledge acquisition phase for all meaningful data values of a medical parameter. At run-time, when actual patient data is used, these definitions will translate data values (e.g., white blood cell count) and assessments (e.g., morning stiffness lasting more than 15 minutes but less than 30 minutes) into symbolic, but fuzzy medical entities.

In the first step of a data-to-entity conversion, the defined range of possible values needs to be partitioned into an appropriate number of categories. These categories usually define some "normal" category and any number of "abnormal" or "pathologic" categories. In a second step, the selected categories can be defined as *exclusive* or *inclusive* categories. In a last step, the compatibility functions for the interpretation categories need to be defined. In MedFrame/CADIAG-IV, a data-to-entity conversion rule builder supports this process with several assistants. The data-to-entity conversion process is completed when the whole, defined range of possible data values is covered. It is, however, possible to just define parameter ranges that are 'pathologic' with respect to a certain class of diseases.

In many situations, the interpretation of actual patient data is only reasonable in special circumstances. MedFrame/CADIAG-IV allows the specification of *fuzzy contexts* that are used to qualify specific interpretations. In terms of the knowledge acquisition process, a default context, which is used whenever no specialized context is applicable, is always defined in a first step. Subsequently, an unlimited number of appropriate fuzzy contexts can be defined (or reused); thereafter, contexts can be adapted for all related parameters individually.

2.3. Inference Knowledge

MedFrame/CADIAG-IV's inference processes are reasoning mechanisms that deal with symbolic medical entities. These entities are connected by means of fuzzy relations. The basic inference process follows these relations and recursively calculates (fuzzy) values for connected entities. A

controlled medical vocabulary defines a hierarchy of the medical entities which is used for logical inferences like abstractions and generalizations.

Two special knowledge representations, *disease profiles* and *explicit rules*, combine several medical entities in more complex ways. *Disease profiles* are intermediate representations that combine, in a table-like manner, all the defined medical entities (e.g., symptoms, findings, examinations, syndromes, diseases, therapies) and their relations to other entities (usually diseases or diagnostic hypotheses). A *rule builder* has been implemented in MedFrame/CADIAG-IV that facilitates the definition of *explicit rules*, which are composed of medical entities and a set of operands (e.g., arithmetic, Boolean, fuzzy, and magnitude operators). Medical entities are unrelated to each other unless an expert adds some knowledge about a specific relation by defining *disease profiles* or *explicit rules*. The relation between two entities is a *fuzzy membership relation* and is defined with the help of a stepwise refinement process, which is outlined in detail below. The sum of all fuzzy relationships constitutes a network of linked concepts that defines the knowledge base, which is used by the inference processes.

3. Stepwise Refinement of Fuzzy Relations

A *guided, stepwise knowledge-acquisition process* has been established in MedFrame/CADIAG-IV to support this complex task. An iterative, stepwise definition of (1) associations, (2) relations, (3) fuzzy linguistic categories, and (4) fuzzy membership functions refines the system's knowledge.

(1) *Associations* are appropriate whenever causal relations or at least empirical correlations are accepted as scientific facts. For example, a *positive association* between a symptom and a disease implies that medical knowledge is available to always infer the presence of a disease whenever the symptom is present at least to some extent. Confirmation is thus the maximum value of a positive association. To exclude the disease whenever the symptom is present, a *negative association* with a maximum value would have been used.

(2) The basic *relations* are "*frequency of occurrence"* and "*strength of confirmation"*. They are used to differentiate positive and negative associations into nine basic relations. Any relation is directed from an antecedent A to a consequent D (usually from a finding to a disease) and is characterized by the frequency of occurrence F_p and the strength of confirmation S_p. Additionally, the same relations have to be defined for the *absence of the consequent* (negation of D, ¬D), because a low or zero value of strength of confirmation is semantically different from an exclusion (overall, the user has nine possible relations by specifying F_p and S_p, or F_n and S_n, respectively).

(3) MedFrame/CADIAG-IV supports the definition of vagueness and uncertainty—that is, to qualify the degree of a relationship with the possibility to use *linguistic terms*. For example, if F_p or F_n were established in a previous step not to be 1 (obligatory occurring) they can be further refined

with linguistic terms. For frequency of occurrence they denote the concepts “almost never”, “very seldom”, “seldom”, “medium”, “often”, “very often”, and “almost always”. Similarly, for S_p and S_n, the linguistic terms “almost no”, “very weak”, “weak”, “medium”, “strong”, “very strong”, and “almost definitely” can be used to modify the strength of confirmation. Associated with these terms are predefined fuzzy membership functions.

(4) Alternatively, or as a further refinement to the use of linguistic terms, the fuzzy membership functions can be manipulated directly. In the user interface, textual definitions (i.e., function type, values, bounds, and ranges) as well as the corresponding graphical representations (i.e., the membership graph) can be manipulated to define or adapt fuzzy intervals or fuzzy values.

The summarizing example in Figure 2 does stop with this refinement. However, if enough knowledge becomes available to provide exact values for the frequency of occurrence or for the strength of confirmation, the fuzzy functions can be converted to values. Formally, even these values remain defined as fuzzy values.

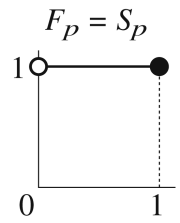
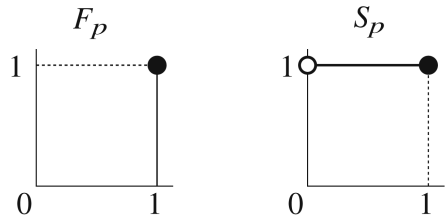
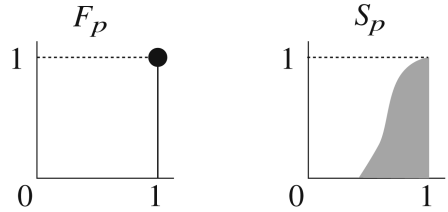
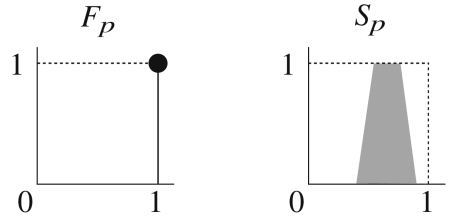
Step 1 **Positive association** between INCREASED SERUM GLUCOSE LEVEL and DIABETES	
Step 2 INCREASED SERUM GLUCOSE LEVEL is **obligatory occurring** with DIABETES but **not confirming**	
Step 3 INCREASED SERUM GLUCOSE LEVEL is **obligatory occurring** with DIABETES the strength of confirmation is **'strong'** (linguistic term)	
Step 4 INCREASED SERUM GLUCOSE LEVEL is **obligatory occurring** with DIABETES the strength of confirmation is defined as a **fuzzy interval between 0.5 and 0.8**	

Figure 2: Illustration of the stepwise refinement process of the relationship between two medical entities, INCREASED SERUM GLUCOSE LEVEL and DIABETES. F_p, F_n, S_p, and S_n are defined in the text. Full/empty circle denotes full/no membership.

4. Discussion

In this final section, we address possibilities and problems with further improvements to the knowledge acquisition process.

Additional knowledge of the local patient population or well-researched patient samples or hypothetical cases can be used during knowledge acquisition to serve as "gold standards" or at least as robustness indicators to evaluate changes in the knowledge base. With additional computations, fuzzy membership functions for F_P, F_N, S_P, and S_N can be calculated based on reference patient databases. However, different assumptions about the influence of prevalences (i.e., the frequency with which symptoms or diseases are present in the patient population) and about the nature of the patient samples (e.g., what is the interpretation of patients not having the disease) influence the validity of the required calculations. Thus, these statistically derived values will always need critical review by domain experts.

MedFrame/CADIAG-IV has been designed to be backwards compatible with its predecessors to allow the integration of their data, knowledge bases, and inference rules. Given the fact that reasoning with negative evidence (introduced in MedFrame/CADIAG-IV) is not fully understood, it should come as no surprise that difficulties in using the full spectrum for negative evidence representations (e.g., F_N, S_N) are reported.

The use of *linguistic variables* is not without problems if multiple domains are aggregated in a single system—as is the case for MedFrame/CADIAG-IV's ambitious goal to integrate many subfields of internal medicine. Any modification in the set of a linguistic variable itself (e.g., adding some new fuzzy quantifier) may require adaptations of previously entered knowledge. However, by providing even finer-grained, guided access to the fuzzy membership functions, MedFrame/CADIAG-IV is not restricted to the use of linguistic terms—but at the price of imposing additional decision tasks on the domain expert.

5. References

[1] Adlassnig, K.-P., Kolarz, G., Scheithauer, W., and Grabner, H. (1986) Approach to a Hospital-Based Application of a Medical Expert System. *Medical Informatics* 11, 205–223.

[2] Kolousek, G. (1997) The Systems Architecture of an Integrated Medical Consultation System and its Implementation Based on Fuzzy Technology. *Ph.D. Thesis*, Technical University of Vienna, Austria.

[3] Bögl, K. (1991) Design and Implementation of a Web-Based Knowledge Acquisition Toolkit for Medical Expert Consultation Systems. *Ph.D. Thesis*, Technical University of Vienna, Austria.

[4] Rothenfluh, T.E., Bögl, K., and Adlassnig, K.-P. (2000) Representation and Acquisition of Knowledge for a Fuzzy Medical Consultation System. In Szczepaniak, P.S., Lisboa, P.J.G., and Kacprzyk, J. (eds.) *Fuzzy Systems in Medicine*. Springer-Verlag, Berlin, 636–651.

[5] Dubois, D. and Prade, H. (1980) *Fuzzy Sets and Systems*. Academic Press, New York.