Aggregating Features and Matching Cases on Vague Linguistic Expressions

Alfons Schuster*, Werner Dubitzky*[1], Philippe Lopes**, Kenneth Adamson*, David A. Bell*,
John G. Hughes*, John A. White***

University of Ulster/Faculty of
Informatics, [§]Northern Ireland
Bio-Engineering Centre
Co. Antrim, BT37 OQB
Northern Ireland

**University of Wales
Dep of Physical Education, Sport
and Exercise Science Unit
Aberystwyth, Dyfed, SY23 3DE
Wales

***Department of Public Health
Medicine and Epidemiology
Queen's Medical Centre
Nottingham NG7 2UH
England

Abstract

Decision making based on the comparison of multiple criteria of two or more alternatives, is the subject of intensive research. In many decision making situations, a single criterion consists of more than one piece of information, and therefore might be regarded as a lump of aggregated information. This paper proposes a general method for aggregating information. To accomplish information aggregation we have developed a fuzzy expert system. Results from an application of our approach in the domain of Coronary Heart Disease Risk Assessment (CHDRA) indicate the value of the information aggregation process of the system. We also show in this paper, how a case-based reasoning (CBR) system can greatly benefit—in its time performance and ability to manage uncertainty—from the information aggregation method.

1 Introduction

Decision making situations very frequently require an ability to compare multiple criteria of two or more alternatives [Chen and Hwang, 1992]. In many cases a single criterion has a complex structure, but even if the meaning of such a complex criterion can be represented in terms of simpler ones, the need for a higher level entity persists [Wilensky, 1986].

As an example, consider a job application scenario, where a manager is confronted with the decision to choose between two candidates, A and B. Let us suppose the manager decides to employ candidate A. Then, when asked to explain his decision, he might say that comparing the two candidates, A has *better prepared* and also showed a *better personality profile.* A closer look at the applicants' documents will probably show, that candidate A indeed has *better references, more experience* and *better working skills* than B. Furthermore, asked in more detail about the *better personality profile,* the manager might answer that candidate A had *better communication skills* and was *more confident* during the interview—therefore his decision was right.

So:

- All information about the candidates is expressed by rather vague or imprecise linguistic terms.
- To explain his decision, the manager uses the linguistic terms *better prepared and better personality profile,* rather than details *{better references, more experience, better working skills, better communication skills, more confidence).*
- The manager would possibly have arrived at the same choice (candidate A) if his only information was that candidate A is *better prepared* and has a *better personality profile* than candidate B.

Our observations are:

(1) Many decision making situations require the capability to manage and process vague or imprecise information, perhaps via linguistic terms.

(2) In many decision making situations, information can be crudely but usefully classified as *higher level information,* and *lower level information* respectively.

In CBR for example, complex case features represent higher level information, and primitive case features represent lower level information respectively. Higher level information can be composed of: (a) lower level information, (b) other higher level information, or (c) a mixture of both information types. In our job application scenario, the higher level information *better preparation* aggregates the lower level information *better references, more experience* and *better working skills.*

And so: It is possible to arrive at useful decisions using information at various levels.

A study of this analysis might lead to a hierarchical structure of information as it is shown in Figure 1.

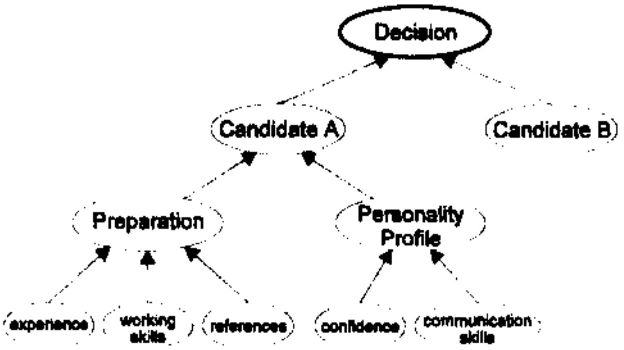

Figure 1: Hierarchical structure of information.

The purpose of this paper is to propose a method that aggregates available lower level information to units of higher level information. In many cases, the lower level information will be imprecise or vague, and so the proposed method should be able to manage uncertainty [Bonissone, 1985]. We therefore have used an expert system shell to develop a fuzzy expert system that accomplishes such aggregation. The applicability and usefulness of this approach was tested in the domain of CHDRA. We show that our fuzzy expert system is able to manage the uncertainty contained in the aggregation process, and that its reasoning process (information aggregation) leads to meaningful, consistent and valuable outcomes. Furthermore, we show how a CBR system can benefit from an implementation of this information aggregation method. For example, we model primitive case features and complex case features via fuzzy sets. This, significantly increases the CBR system's ability to manage uncertainty.

The remainder of this paper is organized as follows: Section 2 briefly describes the advantages that fuzzy primitive and fuzzy complex case features provide for CBR. In Section 3, we describe our fuzzy expert system and the information aggregation process in more detail. The results of applying the approach are outlined in Section 4. Finally, in Section 5, we finish with a discussion, conclusions and future work.

2 CBR, Fuzzy Primitive Case Features and Fuzzy Complex Case Features

CBR is a problem-solving model that allows reasoning to be performed by using past experience [Brown, 1992; Kolodner, 1993, Riesbeck and Schank, 1989]. Past experience—i.e., knowledge about situations that have been solved in the past—is represented by entities, called *cases.* These cases are stored and organized in a memory-like construct called a *case knowledge base* or simply *case base.* Reasoning in CBR systems is accomplished by retrieving the base case(s) most relevant to a new situation or problem at hand, called the *query case,* and then adapting the solution(s) to the actual problem. Because the knowledge contained in a case base is basically determined by its constituents (the stored base cases) the representation of base cases is an important issue in CBR. We describe cases in a compact, characteristic fashion by *abstract* or *salient features,* here simply referred to as *features.* There exist two types of features, primitive features and complex features. Complex features are composites of several (primitive or complex) features.

For example, in the domain of CHDRA the features Smoking and Cholesterol have been identified (among other factors) to be main risk factors for myocardial infarction and subsequent sudden death. In the assessment process, Smoking is regarded as a primitive feature, used to indicate the number of cigarettes a person smokes per day, whereas the complex feature Cholesterol is a composite of the three primitive feature cholesterol types: TOTAL cholesterol, LDL cholesterol and HDL cholesterol.

Cholesterol travels in the blood in distinct particles called lipoprotein. The two major types of lipoproteins are low-density lipoproteins (LDL) and high-density lipoproteins (HDL). LDL, often called 'bad' cholesterol, delivers the cholesterol to the arterial walls with the ultimate consequence of narrowing the arteries [Slyper, 1994]. HDL, often called 'good' cholesterol, protects against heart disease by removing excess cholesterol from the blood [Gordon et al., 1989]. In a fasting blood test, a clinician first finds out what a person's TOTAL, cholesterol level is. If the TOTAL cholesterol level is too high then further measurements of LDL and HDL are required (note: a *high* HDL value 'compensates' a *high* TOTAL cholesterol value, and therefore, a person's cholesterol can be still described as *normal).* In this paper we use 'cholesterol' when we are discussing generally, and Cholesterol when we talk about a complex case feature; but they mean the same thing—an aggregate or composite of three cholesterol type values.

Possible instances of Smoking and Cholesterol might be given by [Smoking/<40cigarettes per day>], and [Cholesterol/<TOTAL 4.6 mmol/1 >, <LDL 3.0 mmol/l >, <HDL 1.0 mmol/1 >].

Frequently, it is not possible to obtain or assess a value of a feature precisely [Dubitzky et al., 1995]. In situations like this, use is often made of linguistic terms. For example, it is not possible to 'measure' a person's cholesterol value, because it is a composite of three cholesterol type values. But, asked about it, the doctor might describe the person's cholesterol to be simply as *normal,* rather than state: TOTAL 4.6, LDL 3.0 mmol/1, and HDL 1.0 mmol/1. Even in situations where precise values are obtainable, humans often fall back upon to use vague or imprecise linguistic terms. For example, the doctor might describe a LDL value of 3.0 mmoi/1 simply as *normal,* and one of

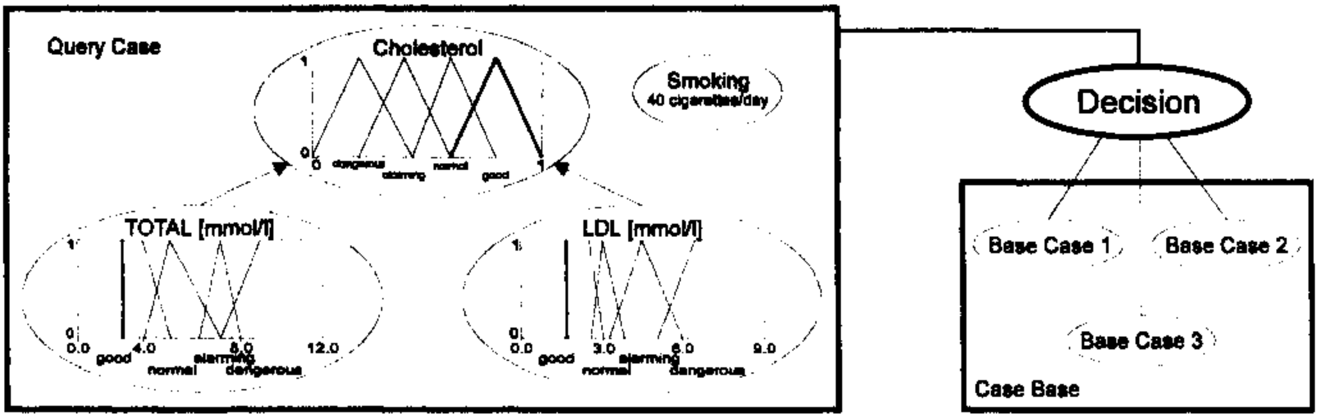

Figure 2: Fuzzy primitive and fuzzy complex case features.

4.5 mmol/l simply as *alarming.* Furthermore, there is no exact boundary between a *normal* and an *alarming* LDL value; that is, the transition between *normal* and *alarming* is gradual or fuzzy, rather than abrupt or crisp [Zadeh, 1973; Klir and Folger, 1988]. In this work we use fuzzy set theory to model primitive features and complex features, and therefore those features will be called fuzzy primitive features, and fuzzy complex features respectively [Na and Park, 1996; Dubitzky et al., 1995]. Figure 2 illustrates the aforementioned CHDRA example in a CBR context. It shows, a query case consisting of the primitive feature Smoking and the fuzzy complex feature Cholesterol, where the fuzzy complex feature Cholesterol is a composite of the fuzzy primitive features Total and LDL (for the sake of simplicity the fuzzy primitive feature HDL is omitted in Figure 2). The cases: Base Case 1, Base Case 2 and Base Case 3 in Figure 2 constitute a simplified case base.

The reasoning process performed should retrieve the base case(s) most relevant to the query case. Therefore, corresponding feature-value pairs for Smoking, Total and LDL of the query case and the base case(s) have to be compared and to be aggregated to an overall similarity score.

3 Information Aggregation via a Fuzzy Expert System

Instead of performing the reasoning task by comparing each single feature value (Smoking, TOTAL, LDL and HDL) of the query case and the base case(s), our approach allows us to compare features on a fuzzy complex feature level. This means that only the feature values Smoking and Cholesterol of the query case and the base case(s) are used. This reasoning process has two main advantages: firstly, for a very large case base the promise of better performance; secondly, sometimes data at primitive feature level is not available, but a description is available on complex feature level. For example, patients may not know the value of each cholesterol type, but possibly remember that during their last health test the cholesterol was *normal.*

To make information on fuzzy complex feature level available, we have developed an inference process based on fuzzy set theory that maps (aggregates) fuzzy primitive feature values to fuzzy complex feature level. For example, the two fuzzy primitive feature values <Total/3.0> and <LDL/2.0> in Figure 2, might map on fuzzy complex feature level to <Cholesterol/(good)>. Such a mapping (aggregation) should satisfy the following requirements:

(1) The aggregated values on complex feature level should be intuitively appealing to an expert's understanding of the problem in question.

(2) Using aggregated values on complex feature level in a decision making process should lead to meaningful, justifiable and consistent results.

To manage the proposed information aggregation process, the normal steps of knowledge acquisition, knowledge representation, and design of an inference engine were realized.

Within the knowledge acquisition process for our application the knowledge engineer and the domain expert were involved to extract the domain knowledge for its use in the fuzzy expert system (e.g., establishing the various fuzzy sets for the different cholesterol types). The basis for the knowledge acquisition was a data set, consisting of 133 records. One record for each person initially held values for TOTAL cholesterol, LDL cholesterol and HDL cholesterol, as well as the two ratios TOTAL/HDL and LDL/HDL. These two ratios are also important because they provide more meaningful indicators of coronary heart disease risk than TOTAL cholesterol per se [Kinosian et el., 1994]. The expert was asked to provide expertise for determining each person's cholesterol value, and so was asked to indicate one of the fields *(dangerous, alarming, normal* and *good)* for each data record as illustrated in Table 1.

Nr.	Cholesterol Data					Expert's Decision			
	Total	LDL	HDL	Total/HDL	LDL/HDL	dangerous	alarming	normal	good
1	6.5	4.9	1.2	5.4	4.1		X		
2	6.3	4.7	1.0	6.3	4.7		X		
...	...	...	...	...	...				
133	6.4	4.4	0.6	10.7	7.3	X			

Table 1: Cholesterol values, taken from 133 persons. Associated with each data record is an expert's decision, representing the expert's interpretation of the person's cholesterol value.

Typically the category that a cholesterol type value or ratio value belongs to is expressed in intervals [Pyorala et al., 1994]. For example, a TOTAL/HDL ratio between 4 and 4.5 is considered as *good,* and one below 4 is regarded to be even *better.* There is no doubt that such a representation is not intuitive to a human's understanding of the problem. In our understanding, the transition from *good* to *better* should be gradual, rather than abrupt. To represent such categories, the three different cholesterol types, and the two ratios are modeled via fuzzy sets. As an example, Figure 3 shows the fuzzy sets for the cholesterol type TOTAL, and for the ratio TOTAL/HDL.

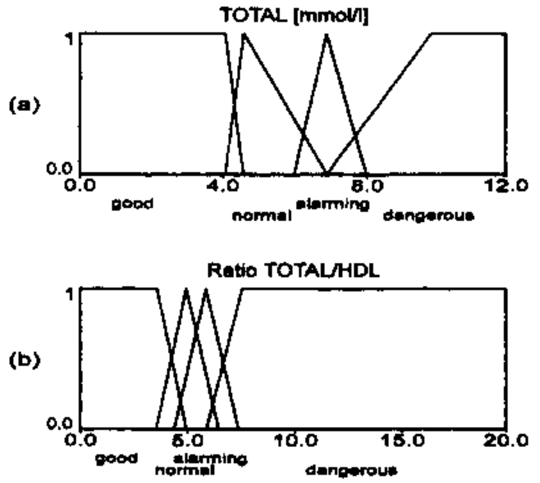

Figure 3: Fuzzy sets for (a) the cholesterol type TOTAL and (b) the ratio TOTAL/HDL.

The knowledge representation scheme used for the proposed information aggregation process was that of production rules, formulated as if-then statements, where the if-part of a rule (the antecedent) is the input, and the then-part of the rule (the consequent) is the output of the fuzzy expert system. Here, the rule base consisted of four rules only. As an example, Figure 4 shows a typical rule used in the fuzzy expert system.

A crucial concept of the proposed fuzzy expert system is, that all rules apply at all times, but some may have more influence than others. This means that if more than one rule is active, the separate responses have to be combined to a composite output. This idea is central to fuzzy logic systems.

If (Total is dangerous) or (LDL is dangerous) or (HDL is dangerous) or (Ratio_Total_HDL is dangerous) or (RatioJLDL_HDL is dangerous)
Then Cholesterol is dangerous

Figure 4: Example rule.

Therefore, the inference process performed by the fuzzy expert system consists of three sub-processes: (a) scaling of the fuzzy input, (b) combination of the output, and (c) defuzzification of the output. There exist different methods for all three sub-processes, and it is part of the knowledge engineer's work to find the methods appropriate to the actual problem. Scaling was done via the *correlation-product* encoding, the combination step via *sum combination* and finally, for the defuzzification of the output, the *center of gravity* method is applied. As an example, Figure 5a shows the fuzzy sets for the system's output (cholesterol), and Figure 5b shows a possible output activation.

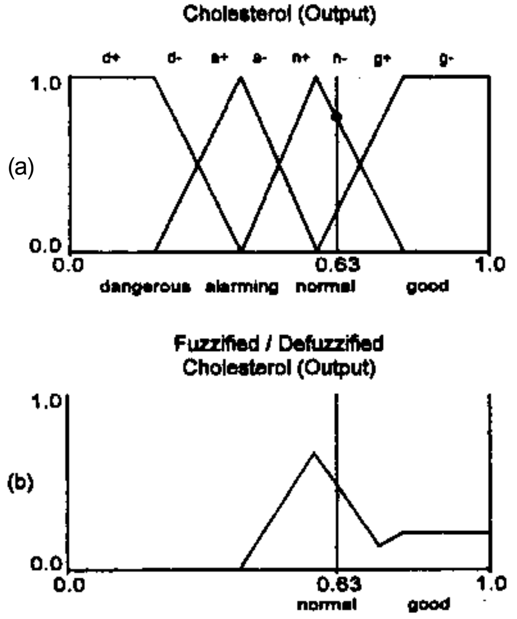

Figure 5: (a) Fuzzy sets for the cholesterol output, and (b) activated cholesterol output and defuzzification via the center of gravity method.

	Cholesterol Data					Expert's Decision				System Output	
Nr.	Total	LDL	HDL	Total/HDL	LDL/HDL	dangerous	alarming	normal	good	COG	Cholesterol
1	6.5	4.9	1.2	5.4	4.1		X			0.63	a+
2	6.3	4.7	1.0	6.3	4.7		X			0.49	d–
...	...	...	...	...	...						...
133	6.4	4.4	0.6	10.7	7.3	X				0.28	d+

Table2: Expert's decision and system output for each cholesterol data record.

Figure 5b also illustrates that the two fuzzy sets *normal* and *good* have been activated by the rules, and that the defuzzification of the output via the center of gravity method results in an output value of 0.63. The location of this value in Figure 5b also shows, that the aggregated cholesterol should be interpreted as *normal* 'with a tendency' to *good.* The tendency of an output is indicated here by a plus (+) or a minus (-) sign, attached to the corresponding fuzzy set and derived as illustrated in Figure 5a. The output value (0.63) intersects with the fuzzy sets *normal* and *good.* When the value intersects with two or more fuzzy sets like this, we take and qualify by 'tendency' the fuzzy set where the output value intersects with the highest score. In this example, the system's output would look like <Cholesterol/n->, and should be interpreted as: 'The aggregated cholesterol value of the person is *normal* with a tendency to *good".* Such a result is intuitively appealing and close to a human expert's explanation in such a situation. In the next section, we investigate the usefulness, the validity and the consistency of the system's output.

4 Results

After the inference process was accomplished for all 133 data records, the fuzzy expert system output of each record was compared with the expert's judgment of the record in question. Table 2 is similar to Table 1, but additionally contains two columns for the system's output. The first column displays the center of the gravity (COG) of the system output, and the second column shows, the system's decision on the cholesterol for the corresponding record.

The results have been evaluated in two steps. In the first step the number of direct matches was computed, and in the second step the number of 'tendency' matches. A direct match was considered to be the case when the expert and the system evaluated the data record belonging to the same category. For example, this is the case for the first and the last record in Table 2. Both the expert and the system evaluated the first record to be *alarming* and the last record to be *dangerous.* Record two in Table 2 represents a tendency match. The expert considers the cholesterol of record 2 to be *alarming,* whereas the system's response is d-, which means *dangerous* with a tendency to *alarming* (see Figure 5a). This is a meaningful result because, as pointed out before, the transition from *alarming* to *dangerous* is gradual.

Our approach led to the following results. A direct match happened 83 times, and a tendency match 34 times. Therefore, the system derived 118 meaningful results, i.e. in 88.7% of the sample. The inference process was not to be expected to establish a direct match of 100% for a number of reasons. Firstly, asked about the same situation or problem twice (e.g. repeated after some weeks), even a single expert's decision-making diverges very often. Secondly, when several experts are available, it is very likely, that they will disagree in some cases. Thirdly, during knowledge acquisition, the expert was enforced to chose one of the four categories *(dangerous, alarming, normal, good)* for a record invoking one of the weaknesses of a discrete choice; very often it is not possible to express intermediate values.

5 Discussion, Conclusions and Future Work

A general method to aggregate information has been presented. Based on fuzzy set theory and fuzzy logic the aggregation process was implemented in a fuzzy expert system. The aggregated information, derived by the fuzzy expert system is meaningful, valuable and consistent. According to [Hall and Kandel, 1992], the proposed fuzzy expert system displays most of the characteristics of 'class one' expert systems; e.g.: (a) the domain of the problem is limited and very well defined, (b) an expert was available during the development, (c) the complexity of the problem is not extreme in the eyes of the knowledge engineer, and (d) the uncertainty prevailing in the domain was manageable.

We applied this process to a specified problem, but its applicability to similar problems is manifest (e.g. at the moment the approach is tested on a data set of cancer patients), and therefore, its potential is obvious. The integration of the proposed information aggregation method into a CBR environment is very promising because the reasoning process in CBR can benefit in two ways: (1) in cases with a lack of data (e.g. unavailable data at the primitive feature level) the higher level information, available at the fuzzy complex feature level, can be used in the reasoning process, and so the CBR system's capability to handle uncertainty increases significantly; and (2) CBR systems with a large case base will improve in their performance in the time domain.

record carries data about a patient suffering a cancerous disease. The support the proposed information aggregation method can provide to the CBR system will be investigated from the point of view of management of uncertainty and performance. There is also work underway to use the aggregation method in a multiple expert scenario and to relate this work with the theory of evidence.

References

[Bonissone, 1985] Piero P. Bonissone. Editorial: Reasoning with uncertainty in expert systems. *International Journal Man-Machine Studies,* 22:241-250, 1985.

[Brown, 1992] Mike Brown. Case-Based Reasoning: principles and potential. *AI Intelligence,* January 1992.

[Chen and Hwang, 1992] Shu J. Chen and Ching L Hwang. *Fuzzy multiple attribute decision making, methods and applications.* Springer Verlag, Berlin, Heidelberg, 1992.

[Dubitzky et al., 1995] Werner Dubitzky, Alfons Schuster, John G. Hughes, and David A. Bell. Conceptual distance of numerically specified case features. In *Proceedings of the Second New Zealand International Two-Stream Conference on Artificial Neural Networks and Expert Systems,* pages 210-213, Dunedin, New Zealand, 1995.

[Gordon et al., 1989] D.J. Gordon, J.L. Probstfield, R.J. Garrison. High density lipoprotein cholesterol and cardiovascular disease. *Circulation,* 79(8):8-15, 1989.

[Hall and Kandel, 1992] Lawrence O. Hall and Abraham Kandel. The evolution from expert systems to fuzzy expert systems. Abraham Kandel, editor. *Fuzzy Expert Systems,* pages 3-21, CRC Press, Boca Raton, Florida, 1992.

[Kinosian et al., 1994] B. Kinosian, H. Glick, G. Garland. Cholesterol and coronary heart disease - predicting risks by levels and ratios. *Annals of Internal Medicine,* 121(9):641-647, 1994.

[Klir and Folger, 1988] George J. Klir and Tina A. Folger. *Fuzzy Sets, Uncertainty and Information.* Prentice Hall, Englewood Cliffs, New Jersey, 1988.

[Kolodner, 1993] Janet Kolodner. *Case-Based Reasoning.* Morgan Kaufmann, San Mateo, California, 1993

[Na and Park, 1996] Selee Na and Seog Park. Management of fuzzy objects with fuzzy attribute values in a fuzzy object-oriented data model. In *Proceedings of Flexible Query-Answering Systems,* pages 19-40, Rosklide, Denmark, 1996.

[Pyorala et al., 1994] Kaveli Pyorala, Guy De Backer, Ian Graham, Philip Poole-Wilson, and Wood David. Prevention of coronary heart disease in clinical practice: Recommendations of the Task Force of the European Society of Cardiology, European Atherosclerosis Society and European Society of Hypertension. *Atherosclerosis,* 110(2):21-61, 1994

[Riesbeck and Schank, 1989] Christopher K. Riesbeck and Roger C. Schank. Inside Case-Based Reasoning. Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1989.

[Schneider and Kandel, 1992] Mordechay Schneider and and Abraham Kandel. General purpose fuzzy expert systems. Abraham Kandel, editor. *Fuzzy Expert Systems,* pages 23-41, CRC Press, Boca Raton, Florida, 1992.

[Slyper, 1994] A.H. Slyper. Low-density-lipoprotein density and atherosclerosis - unravelling the connection. *JAMA,* 272(4):305-308, 1994.

[Wilensky, 1986] Robert Wilensky. Knowledge representation—A critique and a proposal. In J. Kolodner and K. Riesbeck, editors, *Experience, Memory, And Reasoning,* pages 15-28, Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1986.

[Zadeh, 1973]. Lotfi A. Zadeh. Outline of a new approach to the analysis of complex systems and decision processes. *IEEE Transactions on Systems, Man, and Cybernetics,* SMC-3(I):28-45, January 1973.