Reasoning in Expert Systems

1 Diagnostic and Causal Rules

In general, we might write our rules in two ways.

Diagnostic rules These rules can be used to infer the presence of a 'hidden' property from observable properties. For example:

if *animalGivesMilk* **then** *animalIsMammal*

Causal rules These rules reflect the way we think causality works in the world: 'hidden' properties cause observable properties. E.g.:

if *animalIsMammal* **then** *animalGivesMilk*

To some extent, it doesn't matter which we use in a rule-based system because we can change the direction of reasoning, e.g. forwards-chaining on diagnostic rules gives much the same effect as backwards-chaining on causal rules.

In fact, most rule-based systems are probably written using diagnostic rules. For example, the link between measles and spots will probably be represented using a diagnostic rule:

if *spots* **then** *measles*

rather than a causal rule:

if *measles* **then** *spots*

This makes sense because, using diagnostic rules, forwards-chaining takes us from observables (symptoms that the patient might report) towards hidden properties (medical conditions), and backwards-chaining takes us from a hypothesis (a medical condition that a doctor is considering) towards observables (things to look for/ask the patient about).

But our measles example reveals something else about the rules used in many rule-based expert systems.

The rules in a rule-based expert system are sometimes said to encode *heuristic associations*. This is to say that they link (associate) two things (e.g. symptoms and diseases, medical conditions and related medical conditions, etc.) but the link is not certain. Taken literally, the rules are not true. (The word 'heuristic' is used in AI for 'rule of thumb' knowledge. We saw heuristic functions in the context of search, and now we have heuristic associations.)

Consider our diagnostic rule (written using logic notation):

$$spots \Rightarrow measles$$

This simply isn't true. Not all patients who have spots have measles. Spots may be suggestive of measles, but spots are suggestive of other conditions too (acne, insect bites, chicken pox, ...). To make it true, we would have to put more literals into the antecedent (although it's not clear exactly how many literals we would need before we made it true).

Equally our causal rule (written using logical notation):

$$measles \Rightarrow spots$$

is not true either. It is possible (but perhaps rare) to have measles but to have no spots (e.g. German measles can manifest itself as more of a rash). Again, to make it true we would have to put more literals into the antecedent to say exactly when measles will give rise to spots. (Again, it's not clear what literals are needed).

In many domains (medical diagnosis being one of the best examples), it is not possible to repair these rules (by adding literals into their antecedents) to make them true. The reasons for this include:

- Human frailty: We, as knowledge engineers, find it too much work to write exceptionless rules.
- Theoretical ignorance: The domain experts have only a partial theory of the domain.

The alternative to trying to make the rules true is to explicitly represent their uncertainty. This is usually done by assigning numeric measures to the rules to designate the degree of belief we can have in conclusions inferred. For example, we might rewrite our diagnostic rule as follows:

if *spots* **then** 0.6 *measles*

This says that if we are certain a patient has spots, then we can, with 0.6 degree of belief, conclude that they have measles.

But, on top of this, we may not even be certain that the patient has spots. Suppose we are only 0.5 certain that the patient has spots, then our belief that the patient has measles ought to be accordingly lower than 0.6. Somehow we need to combine the numbers 0.5 and 0.6.

How we combine these numbers will depend on exactly what the numbers mean. The approach with the firmest mathematical grounding is if the numbers are probabilities. Then they can be combined using the laws of probability theory. But AI researchers have invented numerous alternative schemes, including some which have no mathematical rationale at all.

(Probabilities might seem the best approach, but using probability is not without problems. In particular, the computations will be more complicated and less efficient unless we assume that multiple symptoms are independent. If new information is not independent of existing information, it should not contribute as much to any change in probabilities we are computing. Recent research has been tackling these issues. See work on belief networks, also called Bayesian networks. These provide an alternative to the whole rule-based approach.)

2 Models of Diagnosis

Let's assume we want to build an expert system for a diagnosis task. Recently, AI researchers have spent time developing models of how diagnosis proceeds. These models state, in general terms, how an expert might reach a diagnosis on the basis of some symptoms. The models are useful in two ways. First, they provide a way of understanding the wide range of existing diagnostic expert systems. We can see how well they fit the model, what technology they use to implement different parts of the models, and how and why they deviate from the model. Second, they might guide the development of new diagnostic expert systems.

We're going to look at a model of diagnosis and then we'll look at an expert system that exhibits aspects of this model.

There are three main search spaces:

The data space: the observables (or measurables, findings, labdata, signs or symptoms). For each, we might know the possible values, the normal values, abnormal values, and, for a particular patient, observed values and predicted values. There are relationships to capture, e.g. knowing the value of one attribute may allow derivation of the value of another, and there may need to be, e.g., procedures for obtaining the observations.

The hypothesis space: the diagnostic hypotheses — what could be wrong with the system being diagnosed (faults, diseases or causes). Relationships in this space show which hypotheses are refinements of other hypotheses or which hypotheses co-occur or cannot co-occur with each other, etc.

The therapy space: the possible actions we can take when treating or repairing a system. There may again be relationships, e.g. there may be a taxonomy of repairs (e.g. of drugs), and procedures may explain how to administer the repair.

As well as relationships within each space, there will be relationships between spaces, e.g. a network of causal relationships.

Diagnosis is then about the interplay of observation and prediction. The main subtasks are:

Recognise abnormalities: Some data is presented, and the system must determine whether the data is abnormal.

Generate hypotheses: From the newest data, you form hypotheses about the likely cause of the abnormal data. In later iterations, new hypotheses might be determined both by the new data and the set of hypotheses already under consideration. For example, related hypotheses (ones that would exhibit similar symptoms) might be entertained; old hypotheses might be ruled out; old hypotheses might be refined; and so on.

Discriminate among hypotheses: You might rank hypotheses using information available at the moment. And you might gather new data that could help to discriminate among the hypotheses (ruling them in or ruling them out). At this point, we go round the loop again.

One aspect of diagnosis that we should make more explicit is that much of the skill of human diagnostic experts is the way they generate hypotheses and reject hypotheses. It would seem that experts are good at these steps because their knowledge is very highly organised and their reasoning methods are tailored to this highly organised knowledge. In other words, you become an expert by restructuring your knowledge so it reflects the uses to which you will put it. As briefly mentioned above when we introduced the hypothesis space, it seems likely that an expert's set of hypotheses are highly inter-related. A hypothesis might be related to its complementary hypotheses (those which tend to co-occur with the hypothesis) and competing hypotheses (those which could account for the data instead). With your knowledge so highly structured, you might be able to capture human reasoning strategies and hence dialogue structure better. For example, you will ask questions to obtain highly discriminating pieces of information. These help you rule out red herrings and trivial causes, choose the most likely hypothesis, and if this hypothesis proves inadequate then your questions should help you to make a lateral shift to a competing or a complementary hypothesis.

3 CASNET

Weiss,S. *et al.*: 'A Model-Based Method for Computer-Aided Medical Decision-Making', *Artificial Intelligence*, vol. 11, 1978, pp.145-172

Weiss,S. and Kulikowski,C.: *A Practical Guide to Designing Expert Systems*, Chapman and Hall, 1983

CASNET is an expert system for the long-term management of diseases whose mechanisms are well-known. Diseases are modeled in terms of Causal ASsociative NETworks. CASNET is, in principle, a general tool for building expert systems for the diagnosis and treatment of such diseases. It has been best demonstrated in the diagnosis of glaucoma (a condition involving increased pressure within the eyeball and gradual loss of sight) and is said to have exhibited experts' performance levels. (In fact, CASNET was only partially implemented but it has a design that is simple yet can be related to the model of diagnosis that I presented in the previous section.)

3.1 Knowledge Base

CASNET's model of diseases (the network) is split into three 'planes' (or, in the terminology used earlier, 'spaces'):

1. Observations
2. Pathophysiological States
3. Disease Categories

There is, in fact, a fourth plane, that of Therapy Plans, which is orthogonal to the three above, and which we will not consider in detail. A plane is basically a network of interrelated knowledge on the same subject in this system. Then there are connections between the planes. See the diagram attached to these notes.

When used for diagnosing glaucoma, CASNET has 400 observation states, 100 pathophysiological states, 75 disease categories/subcategories and 200 treatments.

3.1.1 Plane of Observations

The *observations plane* has nodes for signs, symptoms and laboratory tests. To make questioning more intelligent, there are some arcs relating observation nodes, which establish the order in which questions might be asked, consistent with medical practice, aiming for a natural and sensible dialogue, e.g. that O_1 should be asked about before O_2. They can also say that if O_1 is established then O_2 is (or isn't) also established and so doesn't need to be asked about (e.g. don't ask whether the patient is an alcoholic if we know that the patient's age is less than 13, don't ask what the cup-to-disc ratio at the optic nerve head is unless an opthalmoscopic examination has been performed).

Observations have numbers on them that reflect the 'cost' of obtaining a result. An observation with a high cost might be a ghastly test or very expensive to carry out. CASNET can then prefer cheap painless tests to expensive retch-inducing tests.

Observations are related by *implication links* to pathophysiological states that they can confirm or deny. The links have numbers on them giving the degree of support, i.e. how much that observation confirms the state it is linked to.

3.1.2 Plane of Pathophysiological States

The *plane of pathophysiological states* is a causal network of pathophysiological states, i.e. it contains states of a patient linked together by *causal links*. States are not diseases but detailed dysfunctions. There are final states with no outgoing arcs, which are not taken to cause any other states (in the model). And there are starting states with no incoming arcs, which are not caused by any other states (in the model). A link from one state to another represents the progress of a disease. Roughly, if you are in one state you will progress to the next if you have the disease (unless you receive successful treatment!). Links have numbers that give causation strength.

A complete disease process is a complete pathway from a starting state to a final state. A partial pathway from a starting state to a non-final state represents the degree of evolution within the disease process. Progression along a pathway is associated with development and increasing seriousness of the disease.

The purpose of this network is to help with diagnosis. It is not necessarily a deep causal model of disease development. It's just saying what happens next; it's not saying *why* what happens next should happen next. All it's saying is if you're in this stage of the disease then you're likely to progress to the state of being in the next stage of the disease.

3.1.3 Plane of Disease Categories

In the *plane of disease categories* we represent disease categories (amazing!). These are linked by *classification links* to paths in the plane of pathophysiological states. A disease category node will be linked to some final state. Each path that leads to that state is classified as that disease category. There are also disease subcategories. These are linked to non-final states. Each path that leads to the non-final state is classified as that disease subcategory.

An example is shown in another diagram attached to these notes. The following two paragraphs relate to this diagram.

Disease category D_1 is the three ordered sequence of states $\langle n_1, n_3, n_5, n_6\rangle$, $\langle n_2, n_3, n_5, n_6\rangle$ and $\langle n_4, n_5, n_6\rangle$. Similarly, category D_2 is the two ordered sequences $\langle n_1, n_8, n_9, n_{10}\rangle$ and $\langle n_7, n_9, n_{10}\rangle$.

Suppose that n_3 and n_5 are important intermediate states of the sequences in disease category D_1, then the subcategory D_{12}, corresponding to intermediate state n_5, is the three sequences $\langle n_1, n_3, n_5\rangle$, $\langle n_2, n_3, n_5\rangle$ and $\langle n_4, n_5\rangle$.

The subcategory D_{13}, corresponding to intermediate state n_3, is the two sequences $\langle n_1, n_3\rangle$ and $\langle n_2, n_3\rangle$. Similar subcategories can be defined for D_2 where, say, the important intermediate states are n_8 and n_9.

3.1.4 Plane of Therapy Plans

For completeness I will just mention that this plane has nodes representing therapies. Therapies are linked to nodes in the plane of disease categories. They can also be related to nodes in the plane of observations to show that an observation favours of disfavours a particular treatment, e.g. from observation of a certain age you might want to rule out a certain treatment.

3.2 Inference Engine

The basic idea is to infer the disease processes that are taking place within the patient, i.e. to determine which part of the causal network is actually operative in the patient. Observations are obviously used as direct evidence of certain pathophysiological states. Then, given some states, we can hypothesise that a certain path has been and is being followed.

1. At the beginning of a consultation, the user either enters a set of initial observations or responds to a preliminary sequence of questions.

2. Since these initial observations are linked to pathophysiological states, these related states can be given a status, e.g. confirmed or denied or something in between. This clearly depends on the confidence the user has in each observation and the strengths on the links. Thus we hypothesise certain states.

3. Given the hypothesis of certain states, we can hypothesise certain paths. We want to work out which pathways might explain the observations. An admissible pathway is one that contains no denied states. We are interested in admissible pathways that start from the undenied starting states, and have at least one confirmed state, no undenied states and end on a confirmed state. (They can be of any length including just one state). We then want to choose the most likely of these. This depends on the most likely starting states.

   The most likely starting states are the minimal subset of the starting states of those admissible pathways which collectively contain all the confirmed states.

   We can look at the possible extensions to these paths, i.e. where we can get to from their last state (this being a confirmed state). An extension which covers undetermined states but does not include any denied states gives the possible current aspects of the disease process, i.e. things the patient might have but we've got no evidence for yet. Extensions which then go on to include denied states give possible future developments of the disease, i.e. they predict states which might arise (a prognosis).

4. Each undetermined state on the most likely admissible pathways is chosen for further exploration.

   In particular, we will choose to explore the state that is most likely to be confirmed or denied. Since these are undetermined states, we have no direct evidence from observations to help us to make this choice. Instead, we need to compute the circumstantial evidence for each undetermined state, as follows.

   A state is likely to occur if things prior to it in the chain have occurred, and the more of these that have occurred, the better. Also if later states in a pathway have occurred the undetermined states earlier on the pathway are also more likely to have occurred. Also if a later state is denied, then the likelihood of earlier undetermined states is lower.

   Note that this is all done with 'magic numbers'. To compute the circumstantial evidence for a node in the plane of pathophysiological states, given the values of the surrounding nodes, evidence from confirmed earlier states is cumulated forwards, and evidence from confirmed later states is selected backwards. The stronger of the forward and backward numbers is the weight of the state. So this allows the system to choose to investigate the undetermined state with the highest circumstantial evidence.

   The system will ensure that subsequent questions to the user will help it to determine the status of these states.

5. Thus more questions are asked, specifically about observations that might help to determine the status of these states. Then we go round again. We do this until no more questioning is judged to be fruitful (whatever that means). At the end, the most likely disease process is chosen and therapy can be looked at.

CASNET can give very basic explanations of its reasoning. It can show parts of the networks and the evidence it has for the nodes in the networks. An interesting idea is that nodes might have bits of text associated with them such as references to publications or the results of clinical experiments. These can be displayed to help increase user confidence.

In trials, CASNET had a very good success rate. 77% of its results were rated as of an expert level of performance by domain experts. However, it is limited to use in domains where you can describe things in terms of causes. Some domains don't seem to fit this, or if they do the model might be too big or not well understood enough to be reasonably captured. CASNET was used for glaucoma, aneamias, thyroid dysfunction, diabetes and hypertension. All of these are rather narrow medical domains where the pathophysiology is well understood. It was also used as the basis for a expert system shell, EXPERT, which was used for domains such as endocrinology, opthalmology and rheumatology.