KNOWLEDGE ACQUISITION BY ENCODING EXPERT RULES VERSUS COMPUTER INDUCTION FROM EXAMPLES: A CASE STUDY INVOLVING SOYBEAN PATHOLOGY

by

Ryszard S. Michalski
R.L. Chilausky

International Journal for Man-Machine Studies, No. 12. 1980.

Int. J. Man-Machine Studies (1980) **12**, 63–87

Knowledge acquisition by encoding expert rules versus computer induction from examples: a case study involving soybean pathology

R. S. MICHALSKI AND R. L. CHILAUSKY
University of Illinois, Urbana, Illinois 61801, U.S.A.

(*Received 15 June 1979*)

In view of growing interest in the development of knowledge-based computer consulting systems for various problem domains, the problems of knowledge acquisition have special significance. Current methods of knowledge acquisition rely entirely on the direct representation of knowledge of experts, which usually is a very time and effort consuming task. The paper presents results from an experiment to compare the above method of knowledge acquisition with a method based on inductive learning from examples. The compaiison was done in the context of developing rules for soybean disease diagnosis and has demonstrated an advantage of the inductively derived rules in performing a testing task (which involved diagnosing a few hundred cases of soybean diseases).

1. Introduction

The amount of diagnostic and therapeutic knowledge existing today in the area of human medicine, animal medicine, pathology of plants, etc. surpasses by far what a single expert can encompass. Also, due to the rapid growth of the above disciplines, it is increasingly difficult for an expert to continually update once acquired knowledge. A prospective solution to this problem is the development of expert computer consulting systems which can interactively provide information, advice, and support in decision-making. Such systems could shorten or improve decision-making by suggesting most likely problems or areas of investigation, by calling attention to information which might be overlooked, by suggesting non-typical cases which are possible within the accumulated evidence, etc. In the area of medicine, several experimental consulting systems have been developed, e.g.:

(a) INTERNIST for general medical diagnosis (Myers & Pople, 1977);
(b) MYCIN for antimicrobial therapy advice (Shortliffe, 1976);
(c) CASNET for disease modelling (Kulikowski, 1977, 1978);
(d) CONSULT I and CONSULT II (Patrick, 1979).

Recently, there has been also developed a consulting system in the area of geology, called "PROSPECTOR", for the purpose of providing consultation about mineral exploration (Duda *et al.*, 1978).

A consulting system consists of a knowledge base and an inference mechanism, which matches the queries of users with rules in the knowledge base in order to compute advice. A knowledge base is a symbolic representation of factual and, as well, judgmental knowledge in the subject domain. In each of the above-mentioned consultation systems, the knowledge base was established by handcrafted encoding of the knowledge of human experts. Such encoding can be a very time consuming task, requiring close collaboration between experts of the subject domain and computer scientists trained as "knowledge engineers". This task can be simplified somewhat by special computer programs which facilitate the debugging, modification and maintenance of the knowledge base (Davis, 1976).

An attractive alternative would be to construct a knowledge base by presenting examples of expert decisions to the system and have the system determine the general rules. This means that a consulting system would have to include a module capable of performing inductive inference. The research on computer inductive inference is still at an early stage of development; however, it is already possible to obtain practical results, if the problem is sufficiently well defined and specialized. The papers (Buchanan & Feigenbaum, 1978; Mitchell, 1977; Hayes-Roth &McDermott, 1978; Dietterich & Michalski, 1979) describe some more recent work in this area.

In this paper, we present the results of applying an inductive computer program to the problem of learning from examples the decision rules for the diagnosis of soybean diseases. Then we contrast these decision rules with the decision rules obtained by direct interrogation of experts in soybean pathology. The results may be somewhat surprising to the reader: in the conclusion we have attempted to explain them.

2. The formalism used for knowledge representation

A good formalism for knowledge representation should have not only adequate operators for representing many different aspects of knowledge of human experts, but also be well suited for implementing inference processes on this knowledge. The latter issue seems to be sometimes neglected by workers in the area of knowledge representation.

One of the basic ways for representing expert knowledge is in the form of decision (or production) rules (Davis, Buchanan & Shortliffe, 1975):†

$$\text{CONDITION} \overset{\alpha}{::>} \text{DECISION} \tag{1}$$

The interpretation of such a rule is that if a *situation* satisfies CONDITION then infer DECISION. The parameter α denotes the "strength of implication". Typically, the CONDITION is a conjunction of binary statements and the DECISION is some action, decision, or assignment of values to a variables (e.g., in Shortliffe, 1976). In general, the CONDITION can be any description expressed in some formal language.

A *situation* is a description of some object or processes under consideration. For example, in medical diagnosis, a situation may be some observed manifestations or results of tests performed on a patient. In plant pathology, a situation may be a description of symptoms of a diseased plant.

Another way of representing expert knowledge is in the form of a semantic net (Brachman, 1978) whose general form is a labeled graph with nodes representing various conceptual entities and links representing relationships among these entities.

This way of representing knowledge is quite natural for certain problems. The network representation has, however, several drawbacks. First, since everything is interconnected, it is difficult to modify and incrementally update or extend the knowledge base. Also, it is difficult to represent non-binary relationships. For example,

† We use symbol ::> instead of → which is often used here to indicate a difference between the decision assignment operator and the logical implication.

it is difficult to represent a statement indicating that a certain logical product of concepts (associated with various nodes) implies some other concept, and that the "strength of the implication" is so and so. Such statements are, however, very common in human decision processes, and, therefore, a decision rule representation is often preferable. In the study by Duda *et al.* (1978), the initial representation of knowledge is in terms of rules but in the final stage, these rules are incorporated into a so-called partitioned semantic net. Moreover, individual rules can be made to represent individual "chunks" or "modules" of human knowledge, and therefore, it is relatively easy to modify or incrementally build-up the knowledge base. Also, it seems that it is easier to explain to a user the inference process done by a system by listing the involved decision rules, than by showing a part of a network. Knowledge aquisition by learning from examples also seems to be easier to implement using a rule representation.

The accurate encapsulating of knowledge in the form of rules, however, encounters a number of problems. Typically, an expert's knowledge is expressed in terms of imprecise concepts and involves operators that are not well defined. Also, much of this knowledge is accompanied by statements indicating varying degrees of credibility and varying levels of importance assigned to expressed conditions.

In this paper, we use the rule representation of knowledge. The knowledge here involves descriptions of plant conditions indicating one of 15 soybean diseases. The format of the rules is based on the variable-valued logic calculus VL_1 (Michalski, 1974). This calculus was developed for formally representing in a simple, compact and self-explanatory way decision and inference processes involving many-valued variables. Commonly, the variables in such processes have semantically determined value sets, which can differ both in the scope and in the structure relating its elements. For example, "sex" is a 2-valued variable with no structure relating its possible values, "height" or "temperature" of a human being varies in certain range of possible values, and the values constitute a linearly ordered set.

A simple way of characterizing, e.g., a person is by a list of attribute-value pairs, which in VL_1 is written in the form

$$[\text{sex} = \text{male}][\text{height} = \text{medium}][\text{blood-type} = \text{O}+]$$

A form in brackets [ ] is called a *selector*, and generally is a relational statement relating a variable to one or more values from its domain. A concatenation of selectors denotes the logical product. VL_1 does not include functions or predicates; in many applications, however, descriptions using only variables are sufficient. (A richer language developed in the same spirit which includes functions, predicates and some other forms is VL_{21} (Michalski, 1978).)

In discussions with experts who are trying to describe their decision processes, in particular diagnostic processes, we observed that they often state a condition for a specific diagnosis as a sequence of observations or symptoms (which can be represented by a conjunction of appropriate selectors). However, these experts often also indicate that certain observations are more important than others. In our experiment, observations have ranged from very important to merely supportive or confirmatory. Therefore, we extended here the concept of a selector as defined in (Michalski, 1974) by adding to it a *weight*. A *weighted selector* S^w is a form:

$$[x_i \# \text{R} : \text{W}] \tag{2}$$

where x_i is a variable, R, called the *reference*, is a list of one or more values from the value set of this variable, # stands for one of the relational operators $= \neq \geq \leq > <$, and w is the *weight* of the selector, $w \in [0, 1]$. Is assumed to be 1, if not specified. Before explaining further the weighted selector, we will define some preliminary concepts.

An *event e* is defined as a list of values of an assumed set of variables. For example, assuming the variables: sex, height and blood-type, an event can be

$$e: (\text{male}, 5\text{ ft } 11\text{ in}, \text{A}+)$$

An event e is said to *satisfy* a selector S: $[x_i \# \text{R}]$ if the value of x_i in e is related by # to at least one element of R. For example, selector

$$[\text{albumin} = \text{low}, \text{medium}]$$

is satisfied by e, if the value of albumin in e is low or medium.

It is easy to see that if the *reference* of a selector has more than one element, the selector is equivalent to a disjunction of selectors with one element references:

$$[x_i \# a, b, \ldots] = [x_i \# a] \vee [x_i \# b] \vee \ldots. \tag{3}$$

A selector with a reference consisting of more than one element denotes the so-called *internal disjunction* (disjunction on values of the same variables).

In medical or other applications, the knowledge of values of variables (of tests, observations, etc.) may not be certain. It is usually possible to estimate this uncertainty. Let $D(S, e) \in [0, 1]$ denote the *degree* to which event e satisfies the condition S: $[x_i \# \text{R}]$.

Given an event and a weighted selector S^w, the *degree of confirmation of selector* S^w *by event e* is defined:

$$v(\text{S}^w, e) = v(\text{S}, e) + (1 - w)(1 - v(\text{S}, e)) \tag{4}$$

To explain the idea behind the rule (4), let us assume that in a decision rule, C::>D, the condition, C, is a logical product of selectors, each of which can either be satisfied $v(\text{S}, e) = 1$) or not satisfied $v(\text{S}, e) = 0$). If the weight of each selector is 1, then, when a single selector is not satisfied, the condition, C, is not satisfied. If, however, the weight ("importance") of this selector is small ($\ll 1$), then one would like to see the effect of not satisfying this selector weakened. Formula (4) provides a means for capturing this property. A product of selectors is called a *term*, and a logical union of terms is called a *disjunctive* VL_1 expression (or *weighted* DVL_1 expression).

A simple way of expressing decision rules is in the form

$$\text{C} \overset{\alpha}{::>} \text{D} \tag{5}$$

where C is a DVL_1 expression D (DECISION) is a single selector, or a product of selectors, and α measures the "strength" of the implication ($\alpha \in [0, 1]$).

An example of such a rule is the following description of post-necrotic cirrhosis of the liver:

[albumin = low][regeneration: bile ducts & fibrosis: diff or focal = present]
[fat: diff or zonal ≠ strongly present][fibrosis: portal or central = absent]
[liver nodules = no]

$$\vee$$

[nausea = no][albumin ≠ above normal][regeneration: retic. endo. = absent]

[cells: central or portal, fibrosis: diff or focal = present]
[cells: monos. or epithel. ≠ strongly present]
::> [Diagnosis = Postnecrotic Cirrhosis]

(When α is not specified then $\alpha = 1$.)

The above example illustrates the form of inductively-derived decision rules used in this study (section 5 and Appendix 2). Expert-derived rules had somewhat more complex form (section 4 and Appendix 1).

3. Description space

In the case study, 15 soybean diseases were selected as being representative of the nature and scope of the problems which are faced in the diagnosis of plant diseases. The task was to develop a knowledge base which contained sufficient information to diagnose the following subset of soybean diseases:

D1: *Diaporthe stem canker*
D2: *Charcoal rot*
D3: *Rhizoctonia root rot*
D4: *Phytophthora root rot*
D5: *Brown stem rot*
D6: *Powdery mildew*
D7: *Downy mildew*
D8: *Brown spot*
D9: *Bacterial blight*
D10: *Bacterial pustule*
D11: *Purple seed stain*
D12: *Anthracnose*
D13: *Phyllosticta leaf spot*
D14: *Alternaria leaf spot*
D15: *Frog eye leaf spot*

A description space for diagnosing the selected soybean diseases was developed in conference with an expert in soybean pathology. The variables used were 35 plant and environmental descriptors and one decision variable (specifying diagnosis). The intent in selecting the particular descriptors and their associated values was to provide a description space which was sufficient to describe the diseases of soybeans in terms of macro-symptoms, i.e. those symptoms which could be clearly observed with no sophisticated mechanical assistance. The reason is that an Extension Service Field Agent, a farmer, or even a layman should be able to make reliable observations. A descriptor is a function which assigns to the plant or its environment a specific value from the set called the domain of the descriptor. For example, descriptor Time of Occurrence (TOC) specifies for the diseased plant the time of occurrence of the disease in the field. The descriptor Condition of Roots (COR) assigns a value describing the state of the roots of the plant. The domains of these descriptors for this knowledge base were:

D(TOC) = (April, May, June, July, August, September, October)
D(COR) = (Normal, Rotted, Galls or Cysts Present)

TABLE 1
Plant descriptors used in the experiment

	Number of values	Variable
1. *Environmental descriptors*		
1.1 Time of occurrence	(7)	(x_1)
1.2 Plant stand	(2)	(x_2)
1.3 Precipitation	(3)	(x_3)
1.4 Temperature	(3)	(x_4)
1.5 Occurrence of hail	(2)	(x_5)
1.6 Number years crop repeated	(10)	(x_6)
1.7 Damaged area	(4)	(x_7)
2. *Plant global descriptors*		
2.1 Severity	(3)	(x_8)
2.2 Seed treatment	(3)	(x_9)
2.3 Seed germination	(3)	(x_{10})
2.4 Plant height	(2)	(x_{11})
3. *Plant local descriptors*		
3.1 Condition of leaves	(2)	(x_{12})
3.1.1 Leafspots—halos	(3)	(x_{13})
3.1.2 Leafspots—margin	(3)	(x_{14})
3.1.3 Leafspot size	(3)	(x_{15})
3.1.4 Leaf shredding or shot holing	(2)	(x_{16})
3.1.5 Leaf malformation	(2)	(x_{17})
3.1.6 Leaf mildew growth	(3)	(x_{18})
3.2 Condition of stem	(2)	(x_{19})
3.2.1 Presence of lodging	(2)	(x_{20})
3.2.2 Stem cankers	(4)	(x_{21})
3.2.3 Canker lesion color	(4)	(x_{22})
3.2.4 Fruiting pod on stem	(2)	(x_{23})
3.2.5 External decay	(3)	(x_{24})
3.2.6 Mycelium on stem	(2)	(x_{25})
3.2.7 Internal discoloration	(3)	(x_{26})
3.2.8 Sclerotia—internal or external	(2)	(x_{27})
3.3 Condition of fruits—pods	(4)	(x_{28})
3.3.1 Fruit spots	(5)	(x_{29})
3.4 Condition of seed	(2)	(x_{30})
3.4.1 Mold growth	(2)	(x_{31})
3.4.2 Seed discoloration	(2)	(x_{32})
3.4.3 Seed size	(2)	(x_{33})
3.4.4 Seed shrivelling	(2)	(x_{34})
3.5 Condition of roots	(3)	(x_{35})

Table 1 lists the selected 35 descriptors. The number in parentheses following each descriptor indicates the number of possible values the descriptor can take. In addition, there is a decision variable which specifies the diagnosis of a disease from the assumed set of soybean diseases.

Individual diseased plants were described in terms of the above 35 descriptors. Thus, the total description space, (i.e. the set of all possible sequences of values of descriptors) has the size $7\times2\times3\times\cdots\times2\times2\times3 =$ approx. 3×10^{15} events.

4. Expert-derived decision rules

Diagnostic decision rules for the above-mentioned 15 soybean diseases were obtained from discussions with plant pathologists during several conferences. Approximately 20 hours were required to developed the descriptions for the above 15 diseases. The descriptions of diseases were expressed in the form of modified DVL_1 rules. This modification provided a way to express the statements by experts which indicated different levels of significance for applicable conditions. Significant conditions which must be present in a plant when afflicted by a particular disease are grouped in a term preceded by Q_s; conditions which, although generally present, merely confirm the information which is given by significant conditions are grouped in a term preceded by Q_c. When this representation is used, a sum of these terms constitutes a description of disease.

Additionally, we distinguish a new form of selector, called a *functional* selector, which is defined:

$$[x_i : @fn]$$

where fn is a function which assigns a weight to the selector dependent upon the value of the variable x_i, and @ indicates the nature of fn. It can be $\uparrow, \downarrow, \cap, \cup$, where $\uparrow(\downarrow)$ indicates that fn is monotonically increasing (decreasing) over the domain of x_i and $\cap(\cup)$ indicates that fn has the greatest (smallest) weight around some mean and decreases (increases) with the distance from this mean.

For example, in [# years crop repeated: $\uparrow$ER1] the $\uparrow$ indicates that the weight assigned by the function ER1 grows as the number of years the soybean crop is repeated in the same field. The function ER1 can be defined, e.g.:

$$\text{ER1: w} = \begin{cases} 1{\cdot}0, \text{ if the crop is repeated 3 or more years} \\ 0{\cdot}8, \text{ if the crop is repeated 2 years} \\ 0{\cdot}7, \text{ if the crop is repeated 1 year} \\ 0{\cdot}2, \text{ if the crop has not been repeated.} \end{cases}$$

which is graphically shown in Fig. 1.

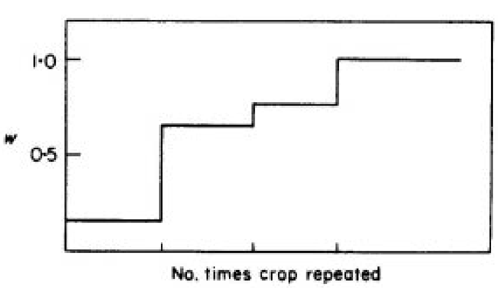

FIG. 1.

TABLE 2

An example of a learning event
(completed questionnaire describing a diseased plant)

Environmental descriptors
- Time of occurrence = July
- Plant stand = normal
- Precipitation = above normal
- Temperature = normal
- Occurrence of hail = no
- Number years crop repeated = 4
- Damaged area = whole fields

Plant global descriptors
- Severity = potentially severe
- Seed treatment = none
- Seed germination = less than 80%
- Plant height = normal

Plant local descriptors
- Condition of leaves = abnormal
  - Leafspots—halos = withour yellow halos
  - Leafspots—margin = without watersoaked margin
  - Leafspot size = greater than 1/8 inch
  - Leaf shredding or shot holding = present
  - Leaf malformation = absent
  - Leaf mildew growth = absent
- Condition of stem = abnormal
  - Presence of lodging = no
  - Stem cankers = above the second node
  - Canker lesion color = brown
  - Fruiting bodies on stem = present
  - External decay = absent
  - Mycelium on stem = absent
  - Internal discoloration of stem = none
  - Sclerotia—internal or external = absent
- Condition of fruits—pods = normal
  - Fruit spots = absent
- Condition of seed = normal
  - Mold growth = absent
  - Seed discoloration = absent
  - Seed size = normal
  - Seed shriveling = absent
- Condition of roots = normal

Diagnosis
Diaporthe stem canker( ) *Charcoal rot*( ) *Rhizoctonia root rot*( )
Phytophthora root rot( ) *Brown stem root rot*( ) *Powdery mildew*( )
Downy mildew( ) <u>*Brown spot*(X)</u> *Bacterial blight*( )
Bacterial pustule( ) *Purple seed stain*( ) *Anthracnose*( )
Phyllosticta leaf spot( ) *Alternaria leaf spot*( ) *Frog eye leaf spot*( )

The following is an example of an expert decision rule (describing diaporthe stem canker):

Q_s([time = Aug . . . Sepresent][precipitation:↑EP][fruiting bodies = present]
[stem cankers = above second node][fruit pods = absent])
+
Q_c([temperature ≥ n][canker lesion color = brown]
[# years crop repeated: ↑ER1]
: : >[Diagnosis = *diaporthe stem canker*]

The complete set of the expert-derived decision rules and the weight assigning functions are given in Appendix 1.

5. Inductively-derived decision rules

5.1. BACKGROUND INFORMATION

The inductively-derived decision rules were generated by applying the computer program AQ11 (Michalski & Larson, 1978) to a set of events (descriptions of individual diseased plants) with known diagnosis. The events were specified in the form of questionnaires completed by plant pathologists. Table 2 is an example of a completed questionnaire which describes a case of brown spot. All available events (630) were partitioned into a learning and testing set (Table 3).

TABLE 3
Events available for learning and testing

Disease	Learning events	Testing events	Available events
Diaporthe stem canker	10	10	20
Charcoal rot	10	10	20
Rhizoctonia root rot	10	10	20
Phytophthora root rot	40	48	88
Brown stem rot	20	24	44
Powdery mildew	10	10	20
Downy mildew	10	10	20
Brown spot	40	52	92
Bacterial pustule	10	10	20
Bacterial blight	10	10	20
Purple seed stain	10	10	20
Anthracnose	20	24	44
Phyllosticta leaf spot	10	10	20
Alternaria leaf spot	40	51	91
Frog eye leaf spot	40	51	91
Total	290	340	630

Also, rules describing some *a priori* knowledge of the problem were specified. These rules included the following:

1. A description of known relationships among variables, specifically relations stating that if some part of a plant is healthy then all the descriptors which specify the particular

conditions of that part do not apply. For example,

[leaves = normal] ⇒ [leafspots halos = *][leafspots margin = *]
[leafspot size = *][leaf shredding = *]
[leaf malformation = *][leaf mildew growth = *]

where * denotes "does not apply" and ⇒ is the logical implication. Table 4 gives the rules used.

TABLE 4
Rules describing a priori *knowledge*

1. [leaves = normal] ⇒ [leafspots halos = *][leafspots margin = *]
[leafspot size = *][leaf shredding = *]
[leaf malformation = *][leaf mildew growth = *]
2. [leafspots halos = absent] ⇒ [leafspots margin = *][leafspot size = *]
3. [stem = normal] ⇒ [presence of lodging = *][stem cankers = *]
[canker lesion color = *][fruiting bodies on stem = *]
[external decay of stem = *][mycelium on stem = *]
[internal discoloration = *]
[sclerotia internal or external = *]
4. [fruit pods = normal] ⇒ [fruit spots = *]
5. [seed = normal] ⇒ [seed mold growth = *][seed discoloration = *]
[seed size = *][seed shriveling = *]

2. Definitions of generalization trees which relate to each other the values of structured variables (Michalski & Larson, 1978) from the viewpoint of their generality. Two structured descriptors were used:

DAMAGED AREA

0 Scattered plants
1 Groups of plants in low areas → 4 Not whole fields
2 Groups of plants in upland areas
3 Whole fields

LEAF SPOTS HALOS

0 Absent
1 With yellow halos → 3 Present
2 Without yellow halos

The learning events and the above rules were the input to the inductive program AQ11. Before presenting the rules and discussing them, we will briefly describe the basic algorithm underlying the program. (Michalski & Larson, 1978).

5.2. DESCRIPTION OF THE TOP-LEVEL ALGORITHM

Suppose there is given a set of hypothesis, $V = \{V_i\}$, $i = 1, \ldots, m$, and a family of event sets ("facts"), $F = \{F_i\}$, which these hypotheses are supposed to describe. Suppose that for any i, V_i describes correctly only a part of the events from F_i.

The problem is to produce a new set of hypotheses, $V^1 = \{V_i^1\}$, where each V_i^1 describes all events from set F_i, and does not describe events from other event sets F_j, $j \neq i$.

The following solution to this problem is based on an algorithm for determing a *cover*, $C(E_1/E_0)$, of an *event set* E_1 *against* the *event set* E_0. Such a cover can be interpreted as a DVL_1 expression which is satisfied by every event in E_1 and not satisfied by any event in E_0 (or in $E_0 \setminus E_1$, if $E_0 \cap E_1 \neq \phi$). The covering algorithm is based on the effective use "negative events" (i.e. those in E_0), and is especially efficient when the negative examples are expressed as a cover. For the lack of space we have to omit here a review of the covering algorithm, and describe only the process of hypothesis generation which uses the algorithm as the basic book. The algorithm is described in Michalski (1971, 1975). The solution consists of 3 major steps.

Step 1

The first step isolates those facts which are not consistent with the given hypotheses. For each hypothesis, two sets are created:

F^+—a set of events which should be covered by the hypothesis, but are not;
F^-—a set of events which are covered by the hypothesis, but should not be covered.

(An event is said to *be covered* by a hypothesis if the event satisfies the VL_1 formula which represents the hypothesis.) Specifically, this step determines, for each i, $i = 1, 2, \ldots, m$, the sets:†

$$F_i^+ = F_i \setminus \tilde{V}_i,$$
$$F_{ij}^- = \tilde{V}_i \cap F_j, \qquad j = 1, 2, \ldots, m; \quad j \neq i.$$

Thus, F_i^+ denotes events which should be covered by V_i but are not, and F_{ij}^- denotes "exception" events, i.e. events in F_j, $j \neq i$, which are covered by V_j, but should not be covered.

Step 2

This step determines, for each i, a generalized formula V_i^- describing all exception events (the union of sets F_{ij}^-, $j = 1, 2, \ldots, m$, $j \neq i$). This is done by generating, for given i and each j, a cover of F_{ij}^- against the events in the sets $\tilde{V}_i \cup F_i^+$, $i = 1, 2, \ldots, m$:

$$V_{ij}^- = C\left(F_{ij}^- \Big/ \bigvee_{i=1}^{m} (\tilde{V}_i \cup F_i^+)\right)$$

and then taking the logical union of V_{ij}^-:

$$V_i^- = \bigvee_{\substack{j=1 \\ j \neq i}}^{m} V_{ij}^-.$$

The reason for this step is that it is computationally more efficient to use formulas V_i^- than the union of E_{ij}^-, $j = 1, 2, \ldots, m$; $j \neq i$.

Step 3

New "correct" hypotheses could be obtained now by "subtracting" from each V_i the formula V_i^- and "adding" to it the set F_i^+. To do this however, is difficult. Again, an advantage is taken of the available covering techniques. Namely, the new hypotheses,

† $\tilde{V}_i$ denotes the set of events covered for formula V_i.

V_i^1, $i = 1, 2, \ldots, m$, are determined as covers:

$$V_i^1 = C\left(F_i \Big/ \bigvee_{\substack{k=1\\k\neq i}}^{m} [(\tilde{V}_k \backslash \tilde{V}_{\bar{k}}) \quad F_k]\right).$$

(The point is that directly simplifying a union of terms is difficult; but subtracting a term from a term or generating a cover of an event set against a formula is easier.)

Step 4

This step determines the final representation of hypotheses V_i^1. The V_i^1 are expressions which are unions of terms. Some terms in a V_i^1 may represent (cover) only a few events in F_i. Such "low weight" terms can be replaced by the events (facts) themselves (since an event takes less memory than a term). (They may also indicate errors in data.)

The rules for the generalization of structured descriptors were applied after the decision rules had been generated.

5.3. THE INDUCTIVELY-DERIVED RULES

AQ11 produced decision rules in which the CONDITION part is a DVL_1 expression involving selectors with $w = 1$. The following is an example of an inductively-derived decision rule (describing *Phytophthora root rot*):

[plant stand < n][precipitation ≥ n][temperature ≤ n][stem = abn]
[plant height = abn][leaves = abn][leaf malformation = abs] (24, 6, 24)

$\vee$

[time = Ar. .Aug][plant stand = abn][damaged area = low areas]
[plant height = abn][leaves = abn][stem = abn] (16, 16, 34)
[external decay ≠ firm & dry]
⇒[Diagnosis = *Phytophthora root rot*]

The complete set of inductively derived decision rules is given in Appendix 2. (AQ11, written in PL/I, took approximately 4 minutes and 30 seconds on an IBM 360/75 to generate the rules.) The triplet of numbers given with each term (a product of selectors) of the rule indicates the performance of that term in covering the learning set of events. The first element of the triplet indicates the number of new events covered by this term (those which were not covered by previously generated terms); the second, the number of events which only this term covered; the third, the number of events which this term covered totally. This triplet provides information about the relative importance of each term to a given decision rule.

The program ESEL (Michalski & Larson, 1978) was used to select the learning events from the set of available events. This program attempts to select the most representative events from each disease set using a "distance" measuring technique. This method of selecting the learning events biases the testing set in some sense since the testing events are those which were not selected by the program. To eliminate this effect one could acquire a distinct set of testing events or select learning events totally randomly. The point of this study was, however, not to test the learning method using a teacher which randomly selects examples, but a "good" teacher which selects representative learning examples. The program ESEL was such a teacher. The selected events were analysed by AQ11 to produce the decision rules.

6. Comparison of the performance of the rules

Both the inductively derived rules and the expert-derived rules were tested using the same testing events (340 cases in total of soybean diseases—Table 3). The experiment involved the application of several inference techniques (Michalski & Chilausky, 1980). Here we present the results which were obtained with the best performing technique for each set of rules.

A. EVALUATION TECHNIQUES USED FOR EXPERT-DERIVED RULES

(Scheme ⟨P, A, M⟩ as described in Michalski & Chilausky, 1980.)

(a) Evaluation of a selector:

$$D(S^w) = \begin{cases} 1, & \text{if the value of the variable in the event satisfies the selector,} \\ 1 - w, & \text{otherwise.} \end{cases}$$

(b) Evaluation of a functional selector (i.e. $[x_i : @\,fn]$):

$$v(S^w) = \text{value of } fn \text{ for the value of the variable in the event}$$

(c) Evaluation of a term:

$$v(T) = \sum_i (v(S_i^w) / \# \text{ of selectors in the term})$$

where i indexes each selector in the term.

(d) Evaluation of an expression. Each rule was a sum of two terms, T_s (conditions preceded by Q_s) and T_1 (conditions preceded by Q_1). (In two rules T_1 was empty.) T_s contributed 90% and T_1 contributed 10% to the degree of confirmation of the rule:

$$v(F) = 0.9 \cdot v(T_s) + 0.1 \cdot v(T_c)$$

The coefficients 0.9 and 0.1 were determined experimentally. (When T_1 was empty, the coefficient for T_s was 1.)

B. EVALUATION TECHNIQUES USED FOR THE INDUCTIVELY-DERIVED RULES

(Scheme ⟨N, A, S⟩ as described in Michalski & Chilausky, 1980.)

(a) Evaluation of a selector:

$$D(S) = \begin{cases} w, & \text{if the value of the variable in the event satisfies the selector,} \\ -w, & \text{otherwise.} \end{cases}$$

(The rules consisted of only selectors with $w = 1$.)

(b) Evaluation of a term:

$$v(T) = \sum_i v(S_i) \Big/ \# \text{ of selectors in the term}.$$

(c) Evaluation of an expression:

For $F = T_1 \vee T_2$

$$v(F) = v(T_1) + v(T_2) - v(T_1) \cdot v(T_2)$$

(For the rules which consisted of more than two terms the evaluation was appropriately extended.)

TABLE 5

Confusion matrix summarizing the diagnosis of 340 testing events using expert-derived VL rules

Correct diagnosis	Indecision ratio	Ties	Maximum # of altern	Test cases	Assigned decision														
					D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11	D12	D13	D14	D15
Diaporthe stem canker (D1)	1·8	7	3	10	100				40							40			
Charcoal rot (D2)	1·0	0	1	10		100													
Rhizoctonia root rot (D3)	0·9	0	1	10			90												
Phytophthora root rot (D4)	1·4	18	2	48		27	8	100	2										
Brown stem rot (D5)	0·96	2	3	24					87			4						4	
Powdery mildew (D6)	1·0	0	1	10						100									
Downy mildew (D7)	3·4	10	5	10							80	100	30				30	70	30
Septoria brown spot (D8)	4·9	52	8	52	37						40	100		38		37	90	44	100
Bacterial blight (D9)	2·7	9	4	10							50		100	90			30		
Bacterial pustule (D10)	3·2	9	5	10							10	70	50	100	30		30	20	10
Purple seed stain (D11)	2·1	8	5	10					20		10		10		80			60	30
Anthracnose (D12)	2·1	21	4	24	50			4	4						54	96			
Phyllosticta leaf spot (D13)	4·1	10	6	10							20	100		50			90	80	70
Alternaria leaf spot (D14)	3·1	51	5	51							39	100	20				8	94	69
Frog eye leaf spot (D15)	4·2	51	6	51				4	39		63	100				4	6	100	100

TABLE 6

Confusion matrix summarizing the diagnosis of 340 testing events using inductively-derived VL rules

Correct diagnosis	Indecision ratio	ties	Maximum # of altern	Test cases	Assigned decision D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11	D12	D13	D14	D15
Diaporthe stem canker (D1)	2·7	10	3	10	100											100			70
Charcoal rot (D2)	1·0	0	1	10		100													
Rhizoctonia root rot (D3)	2·0	10	2	10			100	100											
Phytophthora root rot (D4)	1·0	0	1	48				100											
Brown stem rot (D5)	1·3	3	5	24				8	100			4						8	8
Powdery mildew (D6)	1·0	0	1	10						100									
Downy mildew (D7)	4·1	10	5	10							100	90					30	90	100
Septoria brown spot (D8)	4·0	52	5	52				10				100					88	100	100
Bacterial blight (D9)	3·2	10	4	10									100	50			10	80	80
Bacterial pustule (D10)	1·6	4	4	10								20		100	10		30		
Purple seed stain (D11)	2·8	7	4	10								40	10		100	10		60	60
Anthracnose (D12)	1·1	2	3	24				8								100			4
Phyllosticta leaf spot (D13)	3·9	10	4	10								100					100	100	90
Alternaria leaf spot (D14)	3·2	51	4	51								100					22	100	100
Frog eye leaf spot (D15)	3·9	51	5	51				2				100				82	4	100	100

Tables 5 and 6 show the results of testing both sets of rules (expert-derived and inductively derived) to determine the accuracy with which they classified testing cases of plant diseases. The correct diagnoses for testing events were determined by plant pathologists. If two or more rules were satisfied by a testing event (i.e. a description of a sick plant), the event was multiply classified (i.e. assigned a set of alternatives). The labels for the confusion matrices are defined as follows.

Correct diagnosis
The correct diagnosis for the given testing event.

Indecision ratio
The ratio of the number of alternative diagnoses for the events of the given disease over the number of testing events in the set. An increase in the indecision ratio indicates an increase in the average number of alternative diagnoses for the cases of the given disease. A small indecision ratio does not imply correct diagnoses.

Ties
The number of testing events of the disease which were not uniquely diagnosed.

Maximum # of altern
The maximum number of alternatives in diagnosing a case of the given disease.

Test cases
The number of testing events of the given disease.

Assigned decision
Each column under this label gives the percentage of decisions indicating the corresponding disease for the testing events (for which the correct diagnosis is indicated by the label in the row).

Thus, the percent of correctly assigned diagnoses are on the diagonal of each confusion matrix.

TABLE 7
Performance of the rules

Type	% correct diagnosis	% preferred diagnosis	% not diagnosed	Indecision ratio	Threshold
Inductively-derived	100·0	97·6	—	2·64	0·80
Expert-derived	96·2	71·8	2·1	2·90	0·65

Table 7 gives a comparison of the overall performance of the two sets of rules. The rules which satisfied a *criterion of acceptability* were selected as alternative diagnoses. The *criterion of acceptability* was that the degree of confirmation of a rule must be greater than the *THRESHOLD*, and be either maximum or smaller than maximum by no more than *MARGIN OF UNCERTAINTY*. The *THRESHOLD* was 0·65 for the expert-derived rules and 0·8 for the inductively derived rules. The label "% Correct diagnosis" indicates the percentage of cases when the correct disease (according to experts) was one of alternative diagnoses. The *MARGIN OF UNCERTAINTY* was specified as 0·2 for both sets of rules. The label "% Preferred diagnosis" indicates the percentage of cases when the disease which had the highest degree of confirmation was the correct one. Both inductive and expert rules performed well in selecting the correct disease as one of the diagnostic alternatives. However, the inductively derived rules performed better in selecting the correct disease as the preferred diagnosis. The indecision ratio (total decisions over total events) for the two sets of rules were comparable and the number of alternative diagnoses were distributed quite similarly (Tables 5 and 6). Seven cases could not be diagnosed by the expert rules using the given *THRESHOLD*. The *THRESHOLDS* (determined experimentally) were significantly different. This appears to indicate that the inductive rules are "cleaner", i.e. there is less information in them which is non-essential to diagnosis.

7. Conclusion

The comparison of 2 knowledge acquisition techniques indicates that decision rules derived inductively performed somewhat better than the rules derived by representing the knowledge of experts (in the specific context of soybean disease diagnosis). Since this result was contrary to the initial expectations of the authors, the experiment was repeated several times introducing various corrections to the expert-derived rules and the input events and using different inference techniques. The results always had basically the same pattern. There can be several explanations for this outcome.

(1) The information obtained during the conference with the experts was not sufficiently adequate.
(2) Our knowledge representation scheme was not adequate. (It may be interesting to notice here that expert-derived rules were basically single conjunctions of selectors having varying weight, while inductively derived rules were either a single conjunction of unweighted selectors or a logical union of such conjunctions.)
(3) The inference techniques used to evaluate the decision rules were not adequate.
(4) Experts in making diagnoses are not necessarily experts in explaining the process of diagnosis. These functions are different. If this is the case, it means that the reliability of the data describing diagnoses made by experts (i.e. reliability of the learning events) will tend to be better than the diagnostic decision rules which they formulate. This would provide an additional argument for knowledge acquisition by induction from examples.

The major conclusion of this experiment is that the current computer induction techniques can already offer a viable knowledge acquisition method if the problem domain is sufficiently simple and well defined.

The research presented here was supported in part by the National Foundation Grants NSF MCS 76-22940 and NSF MCS 79-06614. The authors would like to thank Professor James Sinclair and Professor Barry Jackobsen, from the Plant Pathology Department of the University of Illinois, for providing the expertise and the data for the experiments reported here, and for their strong interest in this work.

References

BRACHMAN, R. J. (1978). On the epistemological status of semantic networks. *Report No. 3807.* Bolt Beranek and Newman, Inc., Cambridge, Massachusetts, April.

BUCHANAN, B. G. & FEIGENBAUM, E. A. (1978). Dendral and meta-dendral, their applications dimension. *Artificial Intelligence*, **11**, 5–24.

DAVIS, R., BUCHANAN B. & SHORTLIFFE, E. (1975). Production rules as a representation for a knowledge-based consultation program. *Memo AIM-266.* Stanford Artificial Intelligence Laboratory, Stanford, California, October.

DAVIS, R. (1976). Application of meta level knowledge to the construction, maintenance and use of large knowledge bases. *STAN-CS-76-552.* Department of Computer Science, Stanford University, Standford, California, July.

DIETTERICH, T. G. & MICHALSKI, R. S. (1979). Learning and generalization of characteristic descriptions: evaluation critera and comparative review of selected methods. In *Proceedings of Sixth International Joint Conference on Artificial Intelligence*, Tokyo, Japan, August.

DUDA, R. O. *et al.* (1978). Development of the PROSPECTOR consultation system for mineral exploration. *Final Report.* SRI International, Menlo Park, California, October.

HAYES-ROTH, F. & MCDERMOTT, J. (1978). An inference matching technique for inducing abstractions. *Communications of the ACM*, **21**(5), 401–410.

KULIKOWSKI, C. A. (1977). Problems in the design of knowledge bases for medical consultation. In *Proceedings of 1st Annual Symposium on Computer Applications in Medical Care*, IEEE, New York.

KULIKOWSKI, C. A. (1978). Artificial Intelligence approaches to medical consultation. In *Proceedings of the Fourth Illinois Conference on Medical Information Systems*, May.

MICHALSKI, R. S. (1971). A geometrical model for the synthesis of internal covers. *Report No. 461.* Department of Computer Science, University of Illinois, Urbana, Illinois, June.

MICHALSKI, R. S. (1974). Variable-valued logic: system VL_1. In *Proceedings of Fourth International Symposium on Multiple-Valued Logic*, Morgantown, West Virginia, May.

MICHALSKI, R. S. (1975). Synthesis of optimal and quasi-optimal variable-valued logic formulas. *Proceedings of Fifth Internal Symposium on Multiple-valued Logic*, Indiana University, Bloomington, Indiana.

MICHALSKI, R. S. (1978). Pattern recognition as knowledge-guided computer induction. *Report No. 927.* Department of Computer Science, University of Illinois, Urbana, Illinois.

MICHALSKI, R. S. & LARSON, J. B. (1978). Selection of most representative training examples and incremental generation of VL_1 hypothesis: the underlying methodology and the descriptions of programs ESEL and AQ11. *Report No. 877.* Department of Computer Science, University of Illinois, Urbana, Illinois, May.

MICHALSKI, R. S. & CHILAUSKY, R. L. (1980). An experimental comparison of several many-valued logic inference techniques in the context of computer diagnosis of soybean diseases. *International Journal of Man–Machine Studies*, to appear.

MITCHELL, T. M. (1977). Version spaces: a candidate elimination approach to rule learning. In *5th International Joint Conference on Artificial Intelligence*, Vol. 1, Cambridge, Mass.

MYERS, J. D. & POPLE, H. E. (1977). INTERNIST: A consultative diagnostic program in internal medicine. *Proceedings of the 1st Annual Symposium on Computer Applications in Medical Care*, IEEE, New York.

PATRICK, E. A. (1979). *Decision Analysis in Medicine.* West Palm Beach, Florida: CRC Press.

SHORTLIFFE, E. H. (1976). *Computer-based Medical Consultations: MYCIN.* New York: American Elsevier.

Appendix 1

EXPERT-DERIVED RULES FOR 15 SOYBEAN DISEASES

Q_s indicates significant conditions.

Q_c indicates corroborative conditions.

Abbreviations used: n—normal; abn—abnormal; p—present; abs—absent.

D1: Q_s([time = Aug . . Sep][precipitation: ↑EP]
[stem cankers = above second node][fruiting bodies = p]
[fruit pods = n])
+
Q_c([temperature ≥ n][canker lesion color = brown]
[# years crop repeated: ↑ER1])
: :> [Diagnosis = *Diaporthe stem canker*]

D2: Q_s([time = Jul . . . Aug][precipitation ≤ n][temperature ≥ n]
[plant growth = abn][leaves = abn][stem = abn][sclerotia = p]
[roots = rotted][internal discoloration = black])
+
Q_c([damaged area = upland areas][severity = severe][seed size < n]
[# years crop repeated: ↑ER2])
: :> [Diagnosis = *Charcoal rot*]

D3: Q_s([time = May . . . Jun][plant stand < n][temperature < n]
[precipitation < n][leaves = abn][stem = abn]
[canker lesion color = brown][roots = rotted]
([occurrence of hail = no]⇒[stem cankers = below soil line, at or slightly above soil line])
([occurrence of hail = yes]⇒[stem cankers = above second node]))
+
Q_c([fruiting bodies = abs][external decay = firm & dry][mycelium = abs])
: :> [Diagnosis = *Rhizoctonia root rot*]

D4: Q_s([time: ∩ET][plant stand < n]
([time = Apr . . . Jun]⇒[precipitation = n])
([time = Jul . . . Aug]⇒[precipitation = above n])
([time = Apr]⇒[temperature = above n])
([time = May . . . Aug]⇒[temperature = n])[damaged areas = low areas]
[plant growth = abn][leaves = abn][stem = abn]
[stem cankers = at or slightly above soil line]
([time = May , . . Aug]⇒[canker lesion color = dark brown or black])
[roots = rotted])
+
Q_c([# years crop repeated ≥ 2])
: :>[Diagnosis = *Phytophthora root rot*]

D5: Q_s([time = Jul . . . Sep][precipitation > n][temperature ≤ n][leaves = abn]
[stem = abn][internal discoloration = brown][lodging = p])
+
Q_c([seed size < n][# years crop repeated: ↑ER3])
: :> [Diagnosis = *Brown stem rot*]

D6: Q_s([leaves = abn][leaf mildew growth = upper leaf surface])
+
Q_c[time = Aug . . . Sep]
: :>[Diagnosis = *Powdery mildew*]

D7: Q_s([time = Jun . . . Aug][precipitation ≥ n][damaged areas = whole fields]
[leaves = abn][leafspots halos = no yellow halos]
[leaf mildew growth = lower leaf surface]
([time = Sep . . . Oct]⇒[see = abn])[mold growth on seed = p])
: :>[Diagnosis = *Downy mildew*]

D8: Q_s([leaves = abn][leafspots halos = p]
[leafspots watersoaked margin = abs][leafspot size > 1/8 inch])
+
Q_c([time = May, Aug . . . Sep][precipitation ≥ n])
: :>[Diagnosis = *Brown spot*]

D9: Q_s([time = Apr . . . Jun, Aug . . . Sep]
([time = Apr . . . Jun]⇒[precipitation = n, above n])
([time = Aug . . . Sep]⇒[precipitation⇒above n])
([time ≠ Aug]⇒[temperature = n])
([time = Aug]⇒[temperature = below n])[leaves = abn]
[leafspots halos = with yellow halos][leafspots watersoaked margin = p]
[leafspot size < 1/8 inch][leaf shredding = p])
: :>[Diagnosis = *Bacterial blight*]

D10: Q_s([time = Jun . . . Aug][precipitation ≥ n][leaves = abn]
[leafspots halos = no yellow halos][leafspots watersoaked margin = abs]
[leafspot size < 1/8″][leaf shredding = p])
+
Q_c[# years crop repeated ≥ 1]
: :>[Diagnosis = *Bacterial pustule*]

D11: Q_s([time = Sep . . . Oct][seed = abn][seed discoloration = p]
[seed size = smaller than n])
+
Q_c([time = Aug . . . Sep][precipitation ≥ n][leaves = abn])
: :>[Diagnosis = *Purple seed stain*]

D12: Q_s([time = Aug . . .Oct][precipitation ≥ n][stem = abn]
[canker lesion color = brown][fruiting bodies = p]
([time = Sep . . . Oct]⇒[seed = abn])
[fruit spots = abs, brown spots with black specks])
+
Q_c [damaged area = whole fields]
: :>[Diagnosis = *Anthracnose*]

D13: Q_s([time = Apr . . . Jul][precipitation ≥ n][leaves = abn]
[leafspots halos = no yellow halos][leafspots watersoaked margin = abs]
[leafspot size > 1/8 inch][leaf shredding = p])
+
Q_c([damaged area = whole fields][time ≠ Jun]⇒[temperature = n])
([time = Jun]⇒[temperature = below n])
: :>[Diagnosis = *Phyllosticta leaf spot*]

D14: Q_s([time = Jul . . Oct][leaves = abn][leafspots halos = no yellow halos]
[leafspots watersoaked margin = abs][leafspot size > 1/8 inch]
[leaf shredding = abs])
+
Q_c(([time = Sep . . . Oct] ⇒ [fruit pods = diseased])
([fruit pods = diseased] ⇒ [fruit spots = colored spots])
([seed = abn] ⇒ [seed discoloration = p]))
: :>[Diagnosis = *Alternaria leaf spot*]

D15: Q_s([time = Jul . . . Sep][precipitation ≥ n][leaves = abn]
[leafspots halos = no yellow halos][leafspots watersoaked margin = abs]
[leafspot size > 1/8 inch])
+
Q_c(([time = Sep] ⇒ [fruit spots = colored spots])
[stem canker = above second node][canker lesion color = tan]
[fruiting bodies = abs])
: :>[Diagnosis = *Frog eye leaf spot*]

DEFINITION OF WEIGHT ASSIGNING FUNCTIONS

$$\text{EP}: \begin{cases} 1{\cdot}0, & \text{if precipitation = above normal} \\ 0{\cdot}7, & \text{if precipitation = normal} \\ *, & \text{otherwise} \end{cases}$$

$$\text{ER1}: \begin{cases} 1{\cdot}0, & \text{if \# years crop repeated} \geq 3 \\ 0{\cdot}8, & \text{if \# years crop repeated} = 2 \\ 0{\cdot}7, & \text{if \# years crop repeated} = 1 \\ 0{\cdot}2, & \text{if crop not repeated} \end{cases}$$

$$\text{ER2}: \begin{cases} 1{\cdot}0, & \text{if \# years crop repeated} \geq 2 \\ 0{\cdot}6, & \text{if \# years crop repeated} = 1 \\ 0{\cdot}2, & \text{if crop not repeated} \end{cases}$$

$$\text{ET}: \begin{cases} 1{\cdot}0, & \text{if time of occurrence = May . . . Jul} \\ 0{\cdot}7, & \text{if time of occurrence = Apr, Aug} \\ *, & \text{otherwise} \end{cases}$$

$$\text{ER3}: \begin{cases} 1{\cdot}0, & \text{if \# years crop repeated} \geq 2 \\ 0{\cdot}5, & \text{if \# years crop repeated} = 1 \\ 0{\cdot}1, & \text{if crop not repeated} \end{cases}$$

Appendix 2

INDUCTIVELY-DERIVED RULES FOR 15 SOYBEAN DISEASES

Abbreviations used: n—normal; abn—abnormal; p—present; abs—absent.

D1: [time = Jul . . . Oct][precipitation > n][leaf malformation = abs]
[stem = abn][stem cankers = above second node] (10, 10, 10)
[external decay = firm & dry][fruit pods = n]
: :>[Diagnosis = *Diaporthe stem canker*]

D2: [leaf malformation = abs][stem = abn]
[internal discoloration = black] (10, 10, 10)
: :>[Diagnosis = *Charcoal rot*]

D3: [leaves = n][stem = abn][stem cankers = below soil line]
[canker lesion color = brown] (9, 9, 9)
$\vee$
[leaf malformation = abs][stem = abn]
[stem cankers = below soil line][canker lesion color = brown] (1, 1, 1)
: :>[Diagnosis = *Rhizoctonia root rot*]

D4: [plant stand > n][precipitation ≥ n][temperature ≤ n]
[plant height = abn][leaves = abn][leaf malformation = abs] (24, 6, 24)
[stem = abn]
$\vee$
[time = Apr . . . Aug][plant stand = abn][damaged area = low]
[plant height = abn][leaves = abn][stem = abn] (16, 16, 34)
[external decay = abs, soft and watery]
: :>[Diagnosis = *Phytophthora root rot*]

D5: [leaf malformation = abs][stem = abn]
[internal discoloration = brown] (13, 13, 13)
$\vee$
[leaves = n][stem = abn][internal discoloration = brown] (7, 7, 7)
: :>[Diagnosis = *Brown stem rot*]

D6: [leaves = abn][leaf malformation = abs]
[leaf mildew growth = on upper leaf surface][roots = n] (10, 10, 10)
: :>[Diagnosis = *Powdery mildew*]

D7: [leafspots halos = p][leaf mildew growth = on lower leaf surface]
[stem = n][seed mold growth = p] (10, 10, 10)
: :>[Diagnosis = *Downy mildew*]

D8: [precipitation ≥ n][# years crop repeated > 1]
[damaged area ≠ whole fields][leaves = abn]
[leafspots halos = no yellow halos]
[leafspots watersoaked margin = abs][leafspot size > 1/8 inch]
[leaf malformation = abs][roots = n] (19, 2, 19)
$\vee$
[precipitation > n][leaves = abn]
[leafspots halos = no yellow halos]
[leafspots watersoaked margin = abs][leafspot size > 1/8 inch]
[root = n] (15, 11, 30)
$\vee$
[time = Apr . . . Jun][damaged area ≠ whole fields][leaves = abn]
[leafspots halos = no yellow halos]
[leafspots watersoaked margin = abs][leafspot size > 1/8 inch]

[leaf shredding = abs][leaf malformation = abs][roots = n] (6, 6, 12)
: :>[Diagnosis = *Brown spot*]

D9: [time = Jun . . . Sep][temperature ≥ n][leaves = abn]
[leafspots halos = p][leafspots watersoaked margin = p]
[leafspot size < 1/8 inch][fruit pods = n][roots = n] (10, 10, 10)
: :>[Diagnosis = *Bacterial blight*]

D10: [leaves = abn][leafspots halos = with yellow halos]
[leafspots watersoaked margin = abs][leafspot size < 1/8 inch]
[stem = n][fruit pods = n] (7, 6, 7)
$\vee$
[leafspots halos = p][leafspot size < 1/8 inch][stem = n]
[roots = rotted] (2, 2, 2)
$\vee$
[time = May][precipitation = n][leaves = abn]
[leafspots halos = with yellow halos] (1, 1, 2)
: :>[Diagnosis = *Bacterial pustule*]

D11: [plant stand = n][precipitation > n][severity = minor]
[plant height = n][leafspots halos = no yellow halos][seed = abn]
[seed discoloration = p][seed size = n] (5, 5, 5)
$\vee$
[leaves = n][seed = abn][seed size = n] (5, 5, 5)
: :>[Diagnosis = *Purple seed stain*]

D12: [precipitation > n][leaf malformation = abs][stem = abn]
[stem cankers = at or slightly above soil line, above second node]
[seed = abn][roots = n] (10, 8, 10)
$\vee$
[time = Aug . . . Oct][precipitation > n][leaves = n]
[stem cankers = above second node][fruit pods = diseased]
[fruit spots = brown spots with black specks] (5, 5, 5)
$\vee$
[temperature > n][leafspots halos = abs][leaf malformation = abs]
[stem = abn][external decay = firm and dry] (5, 5, 7)
: :>[Diagnosis = *Anthracnose*]

D13: [time = Jun . . . Jul][precipitation ≤ n][severity = minor]
[leafspots halos = no yellow halos]
[leafspots watersoaked margin = abs][stem = n][roots = n] (6, 5, 6)
$\vee$
[precipitation < n][leaves = abn]
[leafspots halos = no yellow halos]
[leafspots watersoaked margin = abs][roots = n] (3, 3, 4)
$\vee$
[plant stand < n][precipitation = n][occurrence of hail = no]
[leafspots halos = no yellow halos]
[leafspots watersoaked margin = abs][stem = n][roots = n] (1, 1, 1)
: :>[Diagnosis = *Phyllosticta leaf spot*]

D14: [time = Aug][precipitation > n][seed treatment = none]
[leaves = abn][leafspots halos = no yellow halos]
[leafspots watersoaked margin = p][leafspot size > 1/8 inch]
[leaf mildew growth = abs][stem = n][fruit pods = n] (8, 5, 8)

$\vee$

[time = Sep . . . Oct][precipitation > n]
[damaged area = scattered plants, low areas, whole fields]
[seed germination ≥ 80%][leaves = abn]
[leafspots halos = no yellow halos]
[leafspots watersoaked margin = p][leafspot size > 1/8 inch]
[stem = n] (13, 4, 13)

$\vee$

[time = Aug . . . Oct][damaged area = scattered plants, low areas]
[seed germination < 80%][plant height = n][leaves = abn]
[leafspots halos = no yellow halos]
[leafspots watersoaked margin = p][leafspot size > 1/8 inch]
[leaf mildew growth = abs][stem = n] (7, 3, 10)

$\vee$

[time = Oct][seed germination < 90%][leaves = abn]
[leafspots halos = no yellow halos]
[leafspots watersoaked margin = p][leafspot size > 1/8 inch]
[leaf mildew growth = abs][stem = n] (4, 2, 7)

$\vee$

[time = Aug . . . Oct][damaged area = upland areas, whole fields]
[seed treatment = none, other][seed germination ≥ 80%]
[leaves = abn][leafspots halos = no yellow halos]
[leafspots watersoaked margin = p][leafspot size > 1/8 inch]
[leaf mildew growth = abs][stem = n][fruit pods = n] (3, 3, 3)

$\vee$

[occurrence of hail = no][damaged area = scattered plants]
[severity = potentially severe][seed germination ≥ 80%]
[leaves = abn][leafspots halos = no yellow halos]
[leafspots watersoaked margin = p][leafspot size > 1/8 inch]
[leaf mildew growth = abs][stem = n] (3, 3, 11)

$\vee$

[time = Aug . . . Oct][temperature = n][seed treatment = fungicide]
[seed germination = 80–89%][leaves = abn]
[leafspots halos = no yellow halos]
[leafspots watersoaked margin = p][leafspot size > 1/8 inch]
[leaf mildew growth = abs][stem = n][fruit pods = n] (1, 1, 6)

$\vee$

[time = Sep . . . Oct][leaves = abn]
[leafspots halos = no yellow halos]
[leafspots watersoaked margin = p][leafspot size > 1/8 inch]
[leaf shredding = p] (1, 1, 1)

: :>[Diagnosis = *Alternaria leaf spot*]

D15: [precipitation ≥ n]
[damaged area = low areas, upland areas, whole fields]
[leaves = abn][leafspots halos = no yellow halos]
[leafspots watersoaked margin = p][leafspot size > 1/8 inch]
[leaf shredding = abs][leaf mildew growth = abs][stem = abn]
[roots = n] (13, 0, 13)

∨

[time = Jul . . . Sep][precipitation ≥ n][temperature = n]
[occurrence of hail = no][damaged area = low areas, whole fields]
[seed treatment = fungicide][leaves = abn]
[leafspots halos = no yellow halos]
[leafspots watersoaked margin = p][leafspot size > 1/8 inch]
[leaf shredding = abs][leaf malformation = abs][roots = n] (7, 5, 8)

∨

[time = Aug . . . Sep][precipitation ≥ n]
[damaged area = low areas, upland areas][severity = minor]
[leaves = abn][leafspots halos = no yellow halos]
[leafspots watersoaked margin = p][leafspot size > 1/8 inch]
[leaf shredding = abs][leaf mildew growth = abs][seed = n]
[roots = n] (8, 4, 20)

∨

[time = Jul . . . Aug][precipitation > n][# years crop repeated ≥ 1]
[damaged area = scattered plants][seed treatment = none, other]
[leaves = abn][leafspots halos = no yellow halos]
[leafspots watersoaked margin = p][leafspot size > 1/8 inch]
[leaf shredding = abs][leaf mildew growth = abs][roots = n] (4, 3, 8)

∨

[precipitation > n][# years crop repeated ≤ 2]
[damaged area = scattered plants, upland areas]
[severity = potentially severe][seed germination < 80%]
[leaves = abn][leafspots halos = no yellow halos]
[leafspots watersoaked margin = p][leafspot size > 1/8 inch]
[leaf mildew growth = abs][roots = n] (4, 3, 9)

∨

[time = Jul][occurrence of hail = yes][leaves = abn]
[leafspots halos = no yellow halos]
[leafspots watersoaked margin = p][leafspot size > 1/8 inch]
[leaf mildew growth = abs][stem = n] (2, 2, 4)

∨

[plant stand = n][precipitation ≥ n][# years crop repeated = 2]
[leaves = abn][leafspots halos = no yellow halos]
[leafspots watersoaked margin = p][leafspot size > 1/8 inch]
[leaf shredding = abs][leaf mildew growth = abs][seed = n]
[roots = n] (2, 2, 5)

: :>[Diagnosis = *Frog eye leaf spot*]