Precision Agriculture, 6, 213–228, 2005

Modelling Wild-Oat Density in Terms of Soil Factors: A Machine Learning Approach

BEATRIZ DIAZ bdiaz@iai.csic.es
ANGELA RIBEIRO angela@iai.csic.es
RICARDO BUENO AND rbueno@iai.csic.es
DOMINGO GUINEA domingo@iai.csic.es
Industrial Automation Institute-CSIC, Ctra.Campo Real. Km 0.2, 28500 Madrid, Spain

JUDIT BARROSO judith@ccma.csic.es
DAVID RUIZ AND david.ruiz@ccma.csic.es
CESAR FERNADEZ-QUINTANILLA cesar@ccma.csic.es
CCMA-CSIC, Serrano 115 B, 28006 Madrid, Spain

Abstract. In crop fields, weed density varies spatially in non-random patterns. Initial knowledge of weed distribution would greatly improve weed management for Precision Agriculture operations. Site properties could be correlated to weed distribution, since the former vary among crop fields and also certain factors such as soil texture or nitrogen may condition the weed growth. This paper presents a method, based on artificial intelligence techniques, for inducing a model that appropriately predicts the heterogeneous distribution of wild-oat (*Avena sterilis* L.) in terms of some environmental variables. From several experiments, distinct rule sets have been found by applying a genetic algorithm to carry out the automatic learning process. The best rule set extracted was able to explain about 88% of weed variability.

Keywords: artificial intelligence, data mining, genetic algorithms, machine learning, rules, weed density

Introduction

Weed infestations in crops are still a challenge that has to be met in agriculture. Usually, weeds are heterogeneously distributed in agricultural fields (Cardina *et al.*, 1997). Thus, different sampling procedures have been used to detect and describe the spatial distribution of weeds within a field (Rew and Cousens, 2001). However, weed discrimination is often a difficult task, particularly when weeds and crops have similar morphological and/or spectral characteristics. In spite of this, the spatial variability of weed abundance constitutes the basis for site specific weed management systems. Using these systems, farmers could spray selectively to reduce the amount of herbicide usage thereby diminishing environmental impact as well as economic cost (Earl *et al.*, 1996).

The persistence of high-density weed areas in fields over time suggests a non-random distribution that probably depends on environmental variability in the field. Moreover, soil characteristics as well as the properties of plant species, have a strong influence on the growth and reproduction of both crop and weed. Some studies

conducted with several weed species indicate that the weed abundance could be associated with soil properties. In fact, local relationships among weed abundance and certain factors have already been derived using different types of approaches. For example, Dieleman *et al.* (2000) used canonical correlation analysis to interpret associations among several weed species and certain site properties by covariance-based coefficients. Different canonical correlations were able to describe between 25% and 80% of weed variance and about 27–55% of variance in site properties. From this, the authors concluded that associations among the presence of certain weeds and variables such as herbicide activity, topography or soil type vary from year to year. In recent years, geostatistics have been widely used to analyse the spatial properties of the data. Cardina *et al.* (1995) used a combination of semivariograms and an interpolation method (kriging) to model weed variability based on spatial dependence. Spatial autocorrelation, another spatial statistical technique, has been utilised to obtain spatial patterns for herbicide resistance of wild-oat (Maxwell *et al.*, 1995). Here, autocorrelation for wild-oat resistance has been found at different spatial scales due to biological effects and cultural practices.

Although some of these authors have obtained such spatial dependencies for certain weeds, there is no unique or generic method to study associations among the site properties for all weed types. In addition, these statistical approaches require extreme control of both data acquisition and data analysis processes. For instance, in a canonical correlation analysis, the data should be analysed and modified until the variables are completely independent spatially. Statistical procedures present some other restrictions as they are based on stationary data, which is assumed to have a normal distribution (Fortin *et al.*, 2002). Sometimes the interpretation of results, for example numerical indices of correlation and significance, do not allow a clear determination of whether or not the analysed variables are related. Furthermore, the statistical methods are based on measures of central tendency and spread (i.e. average and variance) and, as a consequence, the results are affected notably by noisy data. The important consequence of all this is that obtaining a precise analytical model for weed dynamics becomes quite a complex task. An alternative and complementary solution could focus on finding an approximate model, such as a set of rules that explain the data, in the same way as an expert farmer.

Currently, artificial intelligence methods such as neural networks, decision trees and genetic algorithms are important in agricultural data modelling (Farkas, 2003; Murase, 2000). These methods belong to an artificial intelligence area named machine learning. A machine learning system aims to determine a description of a given concept from a set of input training examples (Mitchell, 1997). The main machine learning systems are listed below.

1. The artificial neural networks (ANN) is a data-driven model, which can be constructed by a learning procedure such as a backpropagation algorithm employing input–output data; therefore, it can be applied to the cases where sufficiently detailed information about the phenomena to be controlled or modelled is not available. For data analysis, the major disadvantage of ANN lies in the knowledge representation. Acquired knowledge in the form of a network of units connected by weighted links is difficult for humans to interpret (Berthold and Hand, 1999).

2. In the decision tree learning process (one of the most widely used and practical methods for inductive inference), the learnt model is represented by means of a decision tree. In contrast to the ANN, a decision tree is a structure that can be represented as sets of IF-THEN rules to improve human understanding. One of the important drawbacks of decision-tree-building algorithms is the fragmentation problem (Friedman *et al.*, 1996), where the set of examples belonging to each tree node gets smaller and smaller as the depth of the tree increases, making it difficult to induce reliable rules from deep levels of the tree. To overcome this problem Carvalho and Freitas (2002) proposed a decision-tree/genetic-algorithm approach.
3. Genetic algorithms (Goldberg, 1989) offer a powerful search methodology that is independent of the problem. Genetic algorithms have most commonly been applied to optimisation problems outside machine learning, such as for design optimisation problems. When applied in machine learning processes, they are especially suited for cases in which the hypotheses are complex (e.g. sets of rules) and the objective to be optimised may be an indirect function of the hypothesis. Rather than search from general-to-specific hypotheses, or from simple-to-complex as do the decision tree methods, genetic algorithms generate successor hypotheses by repeatedly changing and recombining parts of the best currently known hypotheses, as we will show later on.

The present study aims to identify complex associations among eight soil properties and the abundance of wild-oat within agricultural fields. The final goal is to find a descriptive model (a set of IF-THEN rules) that adequately express available knowledge covering the set of training examples. The use of genetic algorithms has been considered as the most appropriate (machine learning system) hypothesis search method due to their robustness and flexibility. The proposed approach has been tested for discovering a rule set that explains the abundance of winter wild-oat (*Avena sterilis* L.) based on eight soil properties.

Material and methods

Data

Data description. Data was obtained from a quadrangular grid sampling carried out in five barley fields in two different locations in South-East Madrid (Spain). Field size ranged from 0.5 to 1.6 ha. At each grid point, soil samples and wild-oat abundance data was obtained according to the sampling features shown in Table 1. Wild-oat density was estimated in a 0.33×0.33 m quadrat, either as seedlings emerged early in the life-cycle or as seeds produced at the end of the life-cycle (seed rain). Inside the quadrats, soil samples of approximately 2 kg were extracted from the top 150 mm of soil. Soil attributes analysed for each grid point were: pH value and content of extractable nitrogen (N), phosphorus (P), potassium (K), organic matter (OM), sand, silt and clay as detailed in Table 2.

Data pre-processing. Descriptive statistical analysis of the data (see Table 2) showed that, although soil properties in all fields did not have the same range of values, all fields showed spatial variation in weed density. To reduce the effect of

Table 1. Sampling features for the five tested fields

Field	Size (ha)	Topography	Grid spacing (m)	Number of points	Wild-oat count	Date
1	0.5	Flat	10 × 10	50	Seedlings	Feb-2001
2	1.6	Flat	12 × 6	38	Seed rain	Jul-2000
3	0.9	Hilly	10 × 10	96	Seed rain	Jul-2000
4	1.2	Hilly	10 × 10	124	Seedlings	Feb-2001
5	1.6	Flat	12 × 6	228	Seed rain	Jul-2001

Table 2. Descriptive statistical analysis of the soil factors (i.e. Organic matter (OM), nitrogen (N), pH, phosphorous (P), potassium (K), sand, silt, and clay) in the five fields

	OM (%)	N (%)	pH	P (mg 100 g^{-1})	K (mg 100g^{-1})	Sand (%)	Silt (%)	Clay (%)
Field (1) : 50 samples								
Max	2.63	0.154	8.14	374	700	37	53	22
Min	1.09	0.116	7.69	158	280	26	41	18
Avg	2.00	0.136	7.87	243	520	31.5	47.7	20.8
SD	0.29	0.011	0.1	56.1	93	2.8	3	1.3
Field (2) : 38 samples								
Max	2.6	0.178	7.95	347	490	60	58	29
Min	1.33	0.077	7.43	214	240	16	27	13
Avg	1.85	0.121	7.7	281.2	370.5	33.6	46.0	20.4
SD	0.34	0.028	0.13	36	62.5	11.2	7.4	4.6
Field (3) : 96 samples								
Max	2.07	0.149	8.31	127	360	45	43	30
Min	1.12	0.064	7.1	22	140	34.0	31.0	20.0
Avg	1.41	0.092	7.83	65	224.1	38.4	37.1	24.5
SD	0.21	0.015	0.23	22.3	52.3	2.1	2.7	2.4
Field (4) : 124 samples								
Max	2.2	0.144	7.95	186.6	620	51	38	53
Min	0.76	0.054	6.09	17.8	215	16.0	23.0	25.0
Avg	1.36	0.088	7.14	67.9	350.6	36.0	30.0	34.0
SD	0.32	0.019	0.61	41.8	75.1	9.8	3.6	7.2
Field (5) : 228 samples								
Max	2.69	0.181	7.99	940	590	60	60	32
Min	0.82	0.063	7.25	450	260	13	27	12
Avg	1.77	0.117	7.69	648.6	399	34.2	45.8	20.0
SD	0.3	0.025	0.14	90.1	63.6	10.8	6.5	4.9
MaxT.	2.69	0.181	8.31	940	700	60	60	53
MinT.	0.76	0.054	6.09	17.8	140	13	23	12

Some basic statistical values such as the Maximum, the minimum, the average (Avg) and the standard deviation (SD) for the samples taken in each field are shown. The last two rows give the Maximum and minimum values among all fields.

factors such as field history and landscape characteristics on the weed evolution for each field, the data were normalised by a linear scaling technique (Pyle, 1999) making them comparable. Thus, all input data have continuous values in the range from 0 to 1. Summarising, a value V_i was represented by a normalised value $V_{(\text{Normal})i}$ computed using the expression in Eq. (1).

$$V_{(\text{Normal})\text{i}} = \frac{V_i - \min(V_1, \ldots, V_n)}{\max(V_1, \ldots, V_n) - \min(V_1, \ldots, V_n)} \tag{1}$$

where $V_1, \ldots, V_n$ are all input values for a variable V. Linear scaling is a simple, straightforward technique to normalise a range of numerical values. Its greatest advantage is that it introduces no distortion to the variable distribution and only requires knowledge of the maximum and minimum values.

After the scaling process, the variables were categorised into high, middle or low classes based on specific intervals as shown in Table 3. Reasoning with symbolic data has an immediate advantage because the patterns will be expressed by linguistic terms and, consequently they will be directly interpretable by and comprehensible to a human operator. In addition, the use of this categorisation technique allows better handling of data uncertainty. Categorisation thresholds, except for weed density, were established for building regular intervals of values. For the soil variables, three regular intervals were defined, except for pH values where two intervals were enough due to their limited range. In the case of the wild-oat density variable, the value 0.2 was selected as an appropriate threshold as it gave rise to a similar partition for the input data set. Using the 0.2 threshold, two classes, of the same size, for use in the training step were defined: a high-density class and a low-density class, where each sampled point belongs to one or the other class according to its wild-oat density.

As a consequence of previous categorisation, several instances with equal label values belonged to both classes at the same time. This data conflict was resolved by a cleaning step that eliminated the overlap. After the cleaning step, the training set was formed from a total of 414 examples. The high-density class contained 271 (50.6%) points having a weed density greater than 0.2. The other 265 examples (49.4%) formed the low-density class. Although relative values of wild-oat abundance cannot

Table 3. Intervals and linguistic labels for the variables

Normalized variable	Range	Linguistic labels
pH	[0, 0.5]	Low
	(0, 0.5–1]	High
OM, P, K, sand, silt, clay	[0, 0.333]	Low
	(0.333, 0.666)	Medium
	[0.666, 1]	High
Wild-oat density	[0, 0.200]	Low
	(0.200, 1]	High

We use the following notation for defining intervals. Square bracket means that the border value is included as achievable variable value and round bracket represents the opposite situation.

be used directly for preventative treatment purposes, they may allow identification of zones of the field that would be susceptible to infestation by this weed species. This information could be very useful for defining management zones and for adjusting more precisely the treatment dosage.

Finally, the sampled data was integrated into a database where each record stored the nine sampled variables (i.e. eight soil properties and weed density value) at each grid point. Formally, a record m was defined by a features vector $V_{\mathrm{m}} = (l_1, l_2, \ldots, l_k, \ldots, l_n)$, where l_k is the linguistic label for attribute k at the sampled point m. An important characteristic of this data structure is that it allows additional information from later experiments to be easily added to the database.

Machine learning algorithm

The rule-based model. The extraction of an approximate model that explains the input data would greatly improve management decisions. However, the knowledge discovered should be as accurate and as comprehensible to humans as possible. Knowledge comprehensibility is essential for at least two related reasons. First, the user wants to use new knowledge in decision-making. In other words, it is important to get an accurate prediction and also to know why a particular situation (the set of conditions) has been classified in a specific group, because this information can improve the field management. This feature of comprehensibility and/or justifiability is lacking in ANN models, which can make good predictions but without explanation or justification (Henery, 1994). Second, comprehensible knowledge allows a user to validate it. The goal of discovering understandable knowledge can be facilitated by using a high-level knowledge representation, such as the IF-THEN rules. A linguistic rule can be expressed as follows:

$$IF\ cond_1\ \wedge \ldots \wedge\ cond_i\ \wedge \ldots \wedge\ cond_n\ THEN\ pred$$

The IF part, i.e. rule antecedent, contains conjunctions ($\wedge$, i.e. "and") or disjunctions ($\vee$, i.e. 'or') of n conditions (*cond*) about attribute values (i.e. $Attr_i = V_i$), whereas the THEN part, i.e. rule consequent (*pred*), contains a prediction/classification for the value of a goal attribute. The semantic underlying this kind of prediction/classification rule is as follows: If all the conditions specified in the rule antecedent are satisfied by the attributes of a given data instance (a record in the database) then the goal attribute of that instance will take on the value specified in the consequent part of the rule. In the proposed approach, each of the conditions in the rule antecedent has the form $Attr_i\ ,\ O_pC\ ,\ v_{ij}$, where $Attr_i$ denotes the ith attribute in the set of attributes, v_{ij} the jth value of the domain of $Attr_i$, and O_pC is a comparison operator; in the present context, $O_pC \in \{=, \neq\}$. Definitively, the rules induced by the proposed learning process are similar to the following rule:

$$IF\ pH = Low\ \wedge\ Sand = Low\ \wedge\ Clay = high$$
$$THEN\ Wild - Oat - Density = High$$

Genetic search. The induction algorithm to extract the model can be considered as a hypothesis-space search, where a hypothesis/model is essentially a (or a part of a) candidate solution and operators transform one hypothesis into another until a specified model is reached. In the proposed approach, a hypothesis corresponds to a candidate solution defined through a set of rules. On the other hand, operators are usually implemented by generalisation/specialisation operations that transform one candidate set of rules into another (Michalski, 1983). These deterministic operators perform a kind of local search in the hypothesis space, in the sense that a single application of an operator modifies a small part of a candidate set of rules. Moreover, typical rule-induction algorithms construct and evaluate a candidate set of rules in an incremental way; that is, many of the hypotheses evaluated are only part of the whole solution. In contrast, genetic algorithms working as search procedures avoid some of the weaknesses of more classical search approaches. In fact, genetic algorithms typically use stochastic operators and some of these operators, such us crossover, usually perform a more global search in the hypothesis space, in the sense that a single application of an operator can modify a relatively large part of a candidate hypothesis. As a result, intuitively, genetic algorithms tend to cope better with attribute (or condition) interaction problems (Freitas, 2002). Briefly, the basic ideas of the genetic algorithm search are as follows (Holland, 1975): A genetic algorithm maintains a population of individuals, each of them being a candidate solution to a given problem; here, a hypothesis or set of rules. Each individual is evaluated by a fitness function, which measures the quality of its corresponding candidate solution. A fitness function defines the criterion for ranking potential hypotheses and allows their selection according to their probability for inclusion in the next generation population. Individuals evolve towards better and better individuals through a procedure based on natural selection, i.e. survival and reproduction of the fittest, and operators based on genetics, e.g. crossover (recombination) and mutation. In essence, crossover swaps some genetic material between two or more individuals, whereas mutation changes the value of a small part of the genetic material of an individual to a new random value, simulating the erroneous self-replication of individuals (Goldberg, 1989). The search process finishes when the best hypothesis is reached, that is, when the candidate solution that produces the maximum value for the fitness function has been identified. In our case, this means the model or set of rules that can explain all input examples.

The development of a genetic algorithm for rule discovery involves a set of non-trivial decisions. The remainder of this section describes in detail the hypothesis representation (called individual representation in genetic approaches terminology) and the fitness function. The most used and traditional representation for an individual is a binary fixed-length string. In the present approach, individuals are encoded by a set of conditions. Each condition is associated with a specific $Attr_i$ and encoded as a pair: $\{OpC_i, Label_{ij}\}$ where O_pC denotes a comparison operator ($=$ or $\neq$) and $Label_{ij}$ denotes the jth value of the domain of $Attr_i$.

Note that this approach implicitly assumes a positional encoding of attributes in the string. In other words, in each individual, the first condition refers to the first soil property (pH), the second condition (OM) refers to the second attribute, and so on. This positional convention simplifies the action of the genetic operators. The

attribute ordering in our case is: pH, OM, N, P, K, Sand, Silt, and Clay. Keeping in mind Table 3, two bits are needed to code each $Label_{ij}$, so the binary codification of labels is as follows: (low, 01), (medium, 10), (high, 11) and, for pH attribute (low, 10) and (high, 11). The configuration 00 for the three-label attributes, and 01 and 00 for the two-label ones, are used to represent the absence of a condition in the rule antecedent. In the same way, only a single bit is needed to codify the comparison operator $(C)_i$, i.e. 1 for $=$ and 0 for $\neq$. On the other hand, in the approach proposed here, conjunction or disjunction of conditions generates individuals. Then, the logical operators (OpL), that is, $\wedge$ and $\vee$ are coded by 1 and 0, respectively. Disjunction operator between conditions allows the splitting of the individual in more than one rule. Finally, the rule consequent or the value of the wild-oat attribute is coded by one bit, so that 0 means low and 1 high. An example of an individual, described with the guidelines above is represented as follows:

(1) The part rule antecedent

Cond1	OpL	cond2	OpL	cond3	OpL	cond4	OpL	cond5	OpL	cond6	OpL	cond7	OpL	cond8
00 1	1	00 1	1	00 1	1	10 1	0	11 1	1	01 0	1	00 1	1	00 1
pH = –	AND	OM = –	AND	N = –	AND	P = medium	OR	K = high	AND	Sand ≠ low	AND	Silt = -	AND	Clay = -

(2) The part rule consequent

OpL	Class
1	1
AND	wild-oat = high

This binary-string is decoded as:

$$P = medium \ \vee \ K = high \ \wedge \ Sand \neq low \rightarrow Wild - Oat = high$$

Notice that the antecedent of IF-THEN rules is a set of conjunctions and/or disjunctions. Therefore, the previous statement represents the following set of rules according to universal assumptions of logic:

$$\textbf{IF } P = medium \textbf{ THEN } Wild - Oat = high$$

$$\textbf{IF } K = high \ \wedge \ Sand \neq low \textbf{ THEN } Wild - Oat = high$$

Finally, using the basic ideas described above, different types of approaches have been tested for representing a set of rules as an individual codified by a binary string. All of them give rise to different results that are analysed in the following section. The representations that have been tested are as follows:

Case A. Logical operators ($\wedge$ and) are allowed to combine premises in the rule antecedent indistinctly (i.e. first order relationship). The individual length is restricted to contain only eight attributes. More concretely, 32 bits to codify the

rule antecedent and 1 bit to codify the consequent are needed. In the limit and as a consequence of the use of the logical operator $\vee$, the set of rules could be formed by eight conditions for each rule.

Case B. Operators among rule conditions are always $\wedge$ operators. To code a set of rules, the antecedents are encoded in an individual with a number of bits which is a multiple of 32. For example, an individual that is defined by means of 3 rules, would have 97 bits (32×3 bits for antecedents and 1 bit to codify rule consequent). In this case, the objective is to build more generic models that could provide more accurate rules.

Case C. In this case, the construction of the antecedents of the rule set is similar to case B, but only the comparison operator $=$ is used. This case highlights the preference for specific rules versus more general ones.

The other essential key in the design of a genetic algorithm, besides the representation, is the selection of the most appropriate fitness function that guarantees a good trade-off between exploitation of the best candidate solutions and exploration of the search space. In essence, selection based on fitness is the source of exploitation of the best current candidate solutions, whereas genetic operators such as crossover and mutation are the source of exploration of the search space, creating new candidate solutions.

To simplify the explanation without loss of generality, let us assume that there are two classes. Let positive (+) class be the class predicted by a given set of rules, and let negative (–) class be any class other than the class predicted by the rule. In the proposed approach, each hypothesis or set of rules intends to explain positive examples without covering negatives ones. To evaluate the quality of an individual, the genetic algorithm uses the fitness function given by Eq. (2).

$$\text{fitness} = \left(\frac{T_P}{T_P + F_N}\right) \times \left(\frac{T_N}{F_P + T_N}\right) \tag{2}$$

where T_P (true positives) is the number of + examples that were correctly classified as + examples; F_P (false positives) is the number of – examples that were wrongly classified as + examples; F_N (false negative) is the number of + examples that were wrongly classified as – examples; T_N (true negatives) is the number of – examples that were correctly classified as – examples. In the above formula, the term (T_P / (T_P+ F_N)) is often called *sensitivity*, whereas the term (T_N /(F_P + T_N)) is named *specificity*. These two terms are multiplied to force the genetic algorithm to discover rules that have both high sensitivity and high specificity, since it would be relatively easy to maximise one of these terms by reducing the other.

Results and discussion

Out of the experimental data, about 90% (373 instances) of the inputs were chosen randomly to form the training set of examples where the genetic algorithm process searched for the best set of rules that modelled or explained this input data. The rest of the set (10%) was used in the validation process, to test the predictive ability of the

Table 4. Experimental results

Case	Number of rules	Fitness	T_P True$_{high}$	F_N False$_{low}$	T_N True$_{low}$	F_P False$_{high}$	Accuracy (%)
A1	2	0.613	160	44	132	37	78.28
B1	1	0.546	153	51	123	46	73.99
B2	2	0.644	151	53	147	22	79.89
B3	3	0.685	170	34	139	30	82.84
B4	4	0.742	162	42	158	11	85.79
B5	5	0.757	161	43	162	7	86.60
B6	6	0.778	172	32	156	13	87.94
B7	7	0.782	175	29	154	15	88.20
B8	8	0.830	181	23	158	11	90.88
C1	1	0.390	133	71	101	68	62.73
C2	2	0.575	156	48	127	42	75.87
C3	3	0.618	142	62	150	19	78.82
C4	4	0.670	164	40	141	28	81.77
C5	5	0.683	167	37	141	28	83.11
C6	6	0.711	175	29	140	29	84.72
C7	7	0.722	166	38	150	19	84.72
C8	8	0.743	173	31	148	21	86.10

Fitness was computed from expression in Eq (2).

set of rules that were obtained by the genetic search. In the experiments, we considered as positive examples those where wild-oat density was high and the opposite case as negative examples. Thus, our objective is to model the high density wild-oat class. Current work used *AGLearner*, a genetic algorithm software environment developed by us for experimentation with genetic algorithms and related techniques. Results of the different experiments are shown in Table 4.

The different cases displayed in the table correspond to the representation cases described in the previous section. As expected, the use of all comparison and logical operators in the first test (A) resulted in a set of more general rules than in the other tests (B and C). In fact, in case (A) two rules (A1:R1 and A1:R2) were found which explained 78.28% of input data. These preliminary results were presented by Díaz *et al*. (2003).

$$\textbf{IF } OM = low \ \wedge \ P = high \ \wedge \ Clay = low \ \textbf{THEN } Wild - Oat = high \tag{A1:R1}$$

$$\textbf{IF } Silt = high \ \wedge \ Sand = low \ \textbf{THEN } Wild - Oat = high \tag{A1:R2}$$

The second series of experiments (B) was able to discover a rule set that covered a larger number of examples but, as was logically expected, was more complex. In each series of experiments (for example, C3, C4, C5), the possible number of rules for each set was increased by one, since one extra rule in the set can explain a new group of data and, as a consequence can improve the classification accuracy. Experiments were done for individuals that codified among 1–8 rules. Experiments showed that a larger number of rules did not improve the classification accuracy greatly, whereas the model became unnecessarily more complex. For case B, the set of rules with the highest value of fitness was the following:

IF $N = low\ AND\ P \neq high\ AND\ K \neq medium\ AND\ clay \neq medium\ AND\ clay \neq medium\ AND\ sand = high$

THEN $WildOat = high \quad (\mathbf{15}T_P, \mathbf{166}T_N)$ (B8:R1)

IF $P = medium\ AND\ K = medium\ AND\ silt = low$

THEN $Wild - Oat = high \quad (\mathbf{37}T_P, \mathbf{167}T_N)$ (B8:R2)

IF $clay = high\ AND\ silt \neq high\ AND\ sand = high$

THEN $Wild - Oat = high \quad (\mathbf{33}T_P, \mathbf{167}T_N)$ (B8:R3)

IF $K \neq high\ AND\ clay \neq high\ AND\ silt = low\ AND\ sand \neq high$

THEN $Wild - Oat = high \quad (\mathbf{46}T_P, \mathbf{168}T_N)$ (B8:R4)

IF $pH = low\ AND\ P \neq high\ AND\ K \neq high\ AND\ silt \neq high\ AND\ sand \neq low$

THEN $Wild - Oat = high \quad (\mathbf{52}T_P, \mathbf{163}T_N)$ (B8:R5)

IF $OM \neq medium\ AND\ P = medium\ AND\ K \neq medium\ AND\ clay \neq low\ AND\ silt \neq low\ AND\ sand = low$

THEN $Wild - Oat = high \quad (\mathbf{7}T_P, \mathbf{169}T_N)$ (B8:R6)

IF $OM = medium\ AND\ P \neq medium\ AND\ K \neq medium\ AND\ clay \neq high\ AND\ silt = medium$

THEN $Wild - Oat = high \quad (\mathbf{37}T_P, \mathbf{166}T_N)$ (B8:R7)

IF $pH = low\ AND\ N = high\ AND\ P \neq high\ AND\ clay \neq low$

THEN $Wild - Oat = high \quad (\mathbf{25}T_P, \mathbf{169}T_N)$ (B8:R8)

In these rules, the number of true positives (T_P) and true negatives (T_N) (defined in Eq. (2)) that each rule covers is shown in parenthesis.

Finally, the third set of tests (C) showed a model with more specific rules, which were discovered by a genetic search. In this case the rule set that showed the best classification accuracy was as follows:

IF $OM = medium\ AND\ clay = low\ AND\ silt = medium$

THEN $Wild - Oat = high \quad (\mathbf{23}T_P, \mathbf{165}T_N)$ (C8:R1)

IF $P = low\ AND\ silt = low$

THEN $Wild - Oat = high(58TP, \mathbf{160}T_N)$ (C8:R2)

IF $OM = medium\ AND\ P = medium\ AND\ K = medium\ AND\ clay = medium$

THEN $Wild - Oat = high \quad (\mathbf{21}T_P, \mathbf{167}T_N)$ (C8:R3)

IF $OM = medium\ AND\ P = low\ AND\ K = low\ AND\ silt = medium$

THEN $Wild - Oat = high \quad (\mathbf{31}T_P, \mathbf{167}T_N)$ (C8:R4)

IF $P = medium\ AND\ clay = high$

THEN $Wild - Oat = high \quad (\mathbf{30}T_P, \mathbf{165}T_N)$ (C8:R5)

IF $OM = high\ AND\ clay = medium$

THEN $Wild - Oat = high \quad (\mathbf{17}T_P, \mathbf{168}T_N)$ (C8:R6)

IF $N = medium\ AND\ P = high\ AND\ K = high\ AND\ sand = medium$

THEN $Wild - Oat = high \quad (\mathbf{3}T_P, \mathbf{168}T_N)$ (C8:R7)

IF $pH = low\ AND\ N = high\ AND\ P = low\ AND\ clay = high$

THEN $Wild - Oat = high \quad (\mathbf{8}T_P, \mathbf{169}T_N)$ (C8:R8)

As before, the true positives and true negatives are shown in parenthesis.

The best results were obtained in the experiments of series B, in particular when the model contained eight rules since this case gave the best fitness value (0.830) and produced the highest classification accuracy (91%). However, the set of rules C8 in case C, with only a small difference in the classification accuracy compared to rule set B8, was both more specific and more comprehensible. The rules in case C8 contain fewer conditions than rules in case B8 and the conditions embed an equality test on an attribute value, so they are more specific. Although it is extremely important that farmers spray every infested weed area for weed control purposes, in this application domain it is more important to find a model to explain high density examples (T_P) rather than to procure high accuracy. Of course, a trade-off between true positives and accuracy would be meaningless when the accuracy approaches 100%. This premise is satisfied by some rule sets, like for instance, C6 compared to B6 that belong to the same complexity level. This is because while B6 explains 172 examples, C6 describes perfectly 175 high density examples which is the more important requirement for a farmer. Notice that the classification accuracy of B6 is higher than accuracy of C6 because of the larger number of negative examples covered by B6 versus those covered by C6 ($T_{\mathrm{NB6}} < T_{\mathrm{NC6}}$).

In the last step of this process, the rules sets were validated with about 10% of the input data (the validation data set). This validation analysis, shown in Table 5, confirmed previous results of the training set (Table 4), resulting in similar accuracy

Table 5. Percentages of well-estimated examples in the validation data set

Case	T_P	F_N	T_N	F_P	Accuracy (%)
A1	19	7	13	2	78.05
B8	21	5	13	2	82.93
C8	19	7	14	1	80.5

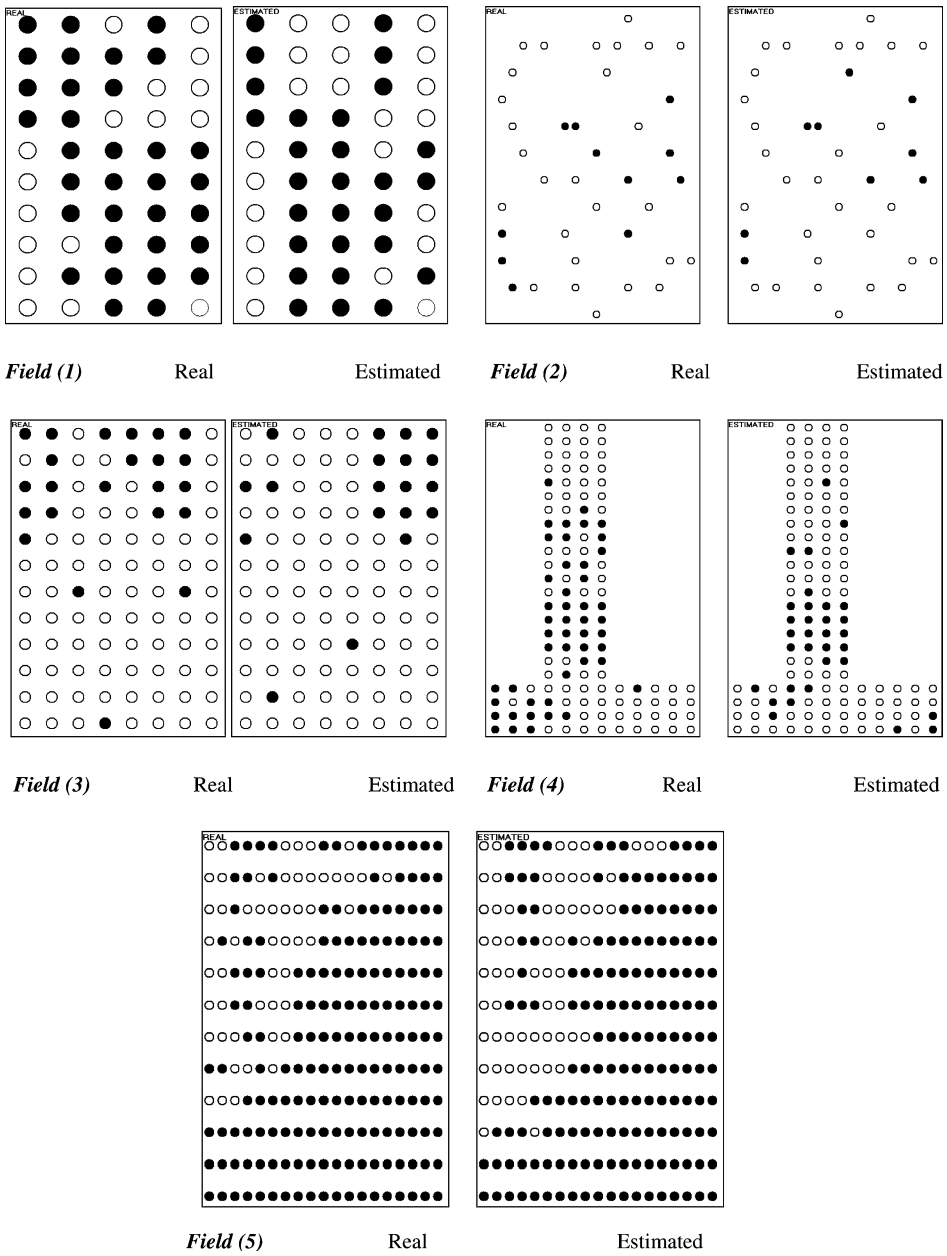

Figure 1. Actual and estimated infestation maps of five studied fields at Madrid site (Spain). Dark circles represent high density weed, while light ones are samples with low density weed.

percentages. Rule set A1, composed of A:R1 and A:R2, classified correctly 19 of the existing 26 high weed density instances and 13 of the 15 low weed density instances, giving a classification accuracy of about 78.1%. B8 and C8 estimated correctly 21 records from the 26. However, B8 set, which consists of rules B:R1 to B:R8, produced a better result as it correctly classified one more example of low class of the weed density. Therefore, B8 accuracy (90%) was higher than accuracy (86%) of C8, which consists of rules C:R1 to C:R8.

Spatial distribution of results is shown in Figure 1 which depicts the actual and the predicted infestation estimated by the best set of rules, B8. It can be seen that the model estimates the weed distribution especially well in field number 5.

As a further verification, the performance of the method proposed here has been compared with the C5.0 algorithm, included in *SPSS Clementine* ©, using the same data set.

The C5.0 algorithm can induce two kinds of model: decision trees and rule sets working in two ways, namely finding general models or specific ones. In the specific category, C5.0 induced a model composed of 15 rules that classified positive examples as well as negatives examples; that is, the model contains rules for describing both classes at the same time. Eight rules from this set described the high weed density class and seven rules the low weed density class. The model obtained by C5.0 produced the following values for accuracy: $T_P = 186$, $F_N = 18$, $T_F = 147$, $F_P = 22$, and 89.43% of examples properly classified. With the general approach, the rule set induced by the C5.0 algorithm was composed of 7 rules (3 for positives and 4 for negatives) and values for accuracy were: $T_P = 174$, $F_N = 30$, $T_F = 132$, $F_P = 37$, in consequence 82.03% of the input data were well classified. Obviously, the specific approach gives more complex models. In fact, it needed 15 rules in order to reach a similar, or even lower, accuracy than that achieved by our genetic algorithm model that has only eight rules. In contrast with the specific model, the general model was less complex, but it was almost 10% less accurate than the rule set B8.

In spite of the fact that categorisation is a fundamental key to handle uncertainty, the threshold selection for defining the attribute boundaries still remains a non-trivial matter. More experimental work for classification of high weed infestation is required using expert criteria in the attribute categorisation. Then, for the categorisation task, other machine learning techniques such as fuzzy logic (Zadeh, 1988) can be used to manage the inherent uncertainty. Some experiments and preliminary results in this area are presented in Ribeiro *et al.* (2003).

Conclusions

This paper presents a general technique that has been developed to induce a set of rules that describe wild-oat density in terms of soil properties from a set of input data. The proposed approach is based on a machine learning technique that performs a genetic search to discover the best rule set according to the classification instances of an experimental database. The use of linguistic terms in the rules antecedent facilitates straightforward interpretation and analysis of the rules discovered. The proposed approach has produced a set of eight rules that can explain

around 91% of the particular input data set for wild-oats. Comparisons of the performance of our proposed genetic approach tool versus that of a commercial one were carried out and show the improvements obtained with the proposed approach.

Acknowledgments

The authors wish to express their gratitude to Dr. Maria C. García-Alegre and Begoña Rebollo of the Industrial Automation Institute (CSIC) for their valuable comments. The Spanish Ministry of Science and Technology is providing full and continuing support for this research work through projects: CICYT-AGF 1999–1125-C03–03 and MCYT- AGL2002-04468-C03-01.

References

Berthold, M. and Hand, D. J. 1999. *Intelligent Data Analysis. An Introduction* (Springer, Germany), ISBN: 3540430601

Bojarczuk, C., Lopes, H. and Freitas, A. A. 2001. Data mining with constrained-syntax genetic programming: Applications in medical data sets. In: *Proceedings Intelligent Data Analysis in Medicine and Pharmacology (IDAMAP 2001)* (MedInfo-2001 workshop, London).

Cardina, J., Johnson, G. A. and McCoy, E. 1995. Analysis of spatial distribution of common lambs quarters in no till soybean. Weed Science **43**, 258–269.

Cardina, J., Johnson, G. A. and Sparrow, D. 1997. The nature and consequence of weed spatial distribution. Symposium Importance of weed biology to weed management. Weed Science **45**, 364–373.

Carvalho, D. and Freitas, A. 2002. A genetic-algorithm for discovering small-disjunct rules in data mining. Applied Computing **2**, 75–88.

Diaz, B., Ribeiro, A., Ruiz, D., Barroso, J. and Fernandez-Quintanilla, C. 2003. A genetic algorithm approach to discover complex associations between wild-oat density and soil properties. In: *Proceedings of the Fourth European Conference on Precision Agriculture*, edited by J. Stafford and A. Werner (Wageningen Academic publishers, The Netherlands) pp. 149–157.

Dieleman, J., Mortensen, D., Buhler, D., Cambardella, C. and Moorman, T. 2000. Identifying associations among site properties and weed species abundance I. Multivariate analysis. Weed Science **48**, 567–575.

Earl, R., Wheeler, P., Blackmore, B. and Godwin, R. 1996. Precision farming: The management of variability. Landwards **51**, 18–23.

Farkas, I. 2003. Special issue: Artificial intelligence in agriculture. Computers and Electronics in Agriculture **40**, 1–3.

Fortin, M. J., Dale, M. and Hoef, J. 2002. Spatial analysis in ecology. In: *Encyclopedia of Envirometrics,* edited by A. H. El-Shaarawi and W. W. Piegorsch (John Wiley and Sons, Ltd, Chichester, UK) pp. 2051–2058.

Freitas, A. A. 2002. *Data Mining and Knowledge Discovery with Evolutionary Algorithms. Series: Natural Computing Series* (Springer Verlag, USA). ISBN: 3-540-43331-7

Friedman, J., Kohavi, R. and Yun, Y. 1996. Lazy Decision Trees. In: *Proceedings of the Thirteenth National Conference on Artificial Intelligence and the Eighth Innovative Applications of Artificial Intelligence Conference, American Association of Artificial Intelligence (AAAI)*, (AAAI Press/MIT Press, USA), pp. 717–724.

Goldberg, D. 1989. *Genetic Algorithms in Search, Optimization, and Machine Learning* (Addison Wesley Publishers Professional, USA).

Henery, R. 1994. Classification. In: *Machine Learning, Neural and Statistical Classification*, edited by D. S. D. Michie and C. Taylor (Ellis Horwood, England), pp. 6–16.

Holland, J. 1975. *Adaptation in Natural and Artificial Systems* (University of Michigan Press, USA).
Maxwell, B. D., Davidson, R. M., Malchow, W. E. and Popay, A. J. 1995. Spatial patterns and spread of herbicide resistance. In: *Proceedings of the forty-eighth New Zealand plant protection conference* (New Zealand plant protection society, New Zealand), pp. 1–6. ISSN: 1172–0719
Michalski, R. 1983. A theory and methodology of inductive learning. Artificial Intelligence **20**, 111–161.
Mitchell, T. 1997. *Machine Learning* (McGraw-Hill Companies, Inc, USA), ISBN: 0070428077
Murase, H. 2000. Special issue: Artificial intelligence in agriculture. Computers and Electronics in Agriculture **29**, 1–2.
Pyle, D. 1999. *Data Preparation for Data Mining* (Morgan Kaufmann, USA), ISBN: 1-558-60529-0
Rew, L. and Cousens, R. 2001. Spatial distribution of weeds in arable crops: Are current sampling methods appropriate? Weed Research **41**, 1–18.
Ribeiro, A., Diaz, B. and Alegre, M. G. 2003. Extracting fuzzy rules to describe weed infestation in terms of soils factors. In: *Proceedings of IEEE International Conference on Fuzzy Systems*, edited by O. Nasraoui and H. Frigui (IEEE, USA), pp. 1032–1037.
Zadeh, L. 1988. Fuzzy logic. IEEE-CS Computer **21**(4), 83–93.