Development of an Expert System to control Pest and Disease Of Groundnut with CN2 algorithm

May Hnin Wai Oo, Nyein Nyein Myo
University of Computer Studies, Yangon
mayhninwaioo@gmail.com

Abstract

Disease and pest are important factors in agriculture ecosystem. In this paper, an expert system is applied in agriculture. It uses a rule induction method to generate the rules for the pests or diseases that occurred in groundnuts and gives the accuracy for the system. It assists the users to give the knowledge of diseases and pests that found in groundnut. The system accepts the behavior of infection and defines pest or disease. Then it diagnoses what pests and disease based on the symptoms that have been selected by the user. CN2, the rule induction method is integrated into a tool for agricultural decision support. The usage of this data mining method allows discovering new agricultural knowledge in the field of pest and disease data that can help to make the optimal solution of the preliminary diagnosis.

Keywords: expert system, data mining, rule induction, pest and disease data, optimal solution, decision making.

1. Introduction

Agriculture is a primary factor of economy in this country. Nowadays, agriculture requires information and application knowledge from different interacting fields to do appropriate decision-making. The production of crops depend on so many factors like fertility of soil, type of seed, climate condition, water logging, application of fertilizers, pests and diseases etc.. It requires proper planning and management that in turn needs correct decision making for growers based on information and knowledge obtained from different related areas.

Detecting those pests and diseases at early stages to enable growers to overcome and treat them appropriately pesticides. Identifying the plant pests and diseases are not easy tasks; it needs experience and knowledge of plants and their pest and disease. Moreover, it requires accuracy in describing the symptoms of plant pest and disease. A pheasant can depend on a system that posses experience and knowledge (expert systems) to enable him or her in identifying any type of pests and disease, making the decision and choosing the right treatment.

The methods that expert system used differ from one system to another because that depends on the user's basic knowledge of the problem. Decision making depends mainly on the way of receiving that knowledge.

In the task of developing an expert system, methods for inducing concept descriptions form examples have proved useful in easing the bottleneck of knowledge acquisition. Two families of systems, based on the IDE and AQ algorithms, have been especially successful. These basic algorithms assume no noise in the domain, searching for a concept description that classifies training data perfectly. However, application to real-world domains requires methods for handling noisy data.

In this paper, CN2 is used to overcome these noisy data. CN2 is to modify the AQ algorithm itself in ways that removed this dependence on specific examples and increased the space of rules searched. This lets one apply statistical techniques, analogous to those used for tree pruning, in the generation of if-then rules, leading to a simpler induction algorithm.

CN2, a new induction algorithm, combines the efficiency and ability to cope with noisy data of ID3 with the if-then rule form and flexible search strategy of the AQ family. The representation for rules output by CN2 is an ordered set of if-then rules, also known as a decision list. CN2 uses a heuristic function to the noise present in the data. This results in rules that may not classify all the training examples correctly, but that perform well on new data [1].

2. Related Work

Rule Induction methods are applied in data mining, knowledge acquisition for knowledge base, rule-based expert systems, etc. Rule learning is typically used in solving classification and prediction tasks. Branko Kavˇsek and Nada Lavraˇc summarizes the modifications needed for the adaptation of the CN2 rule learner to subgroup discovery and presents some results of its application to a real-life data set of UK traffic accidents, together with an initial evaluation of results by the traffic expert. The comparative results show that CN2-SD induced on average smaller rule sets that included rules that had on average a higher coverage that those induced by the standard CN2 algorithm. The latter fact makes CN2-SD more suitable for the subgroup discovery task as each rule with high coverage represents potentially an interesting subgroup in the data. On the other hand the average accuracy of the CN2-SD rule sets was more or less the same as the accuracy of standard CN2 rules, which is very good given that the CN2-SD algorithm does not optimize rule accuracy. The above findings are not new and reflect the findings in [2]. Jolita Bernatavicien and his friends applied the rule induction methods for the investigation of the parameter system of the optic nerve disc. The rule induction methods, CN2 and Classification tree Algorithm can be integrated into a tool for medical decision support. This tool would be useful especially for the medics that are not experts in medical field. The usage of these data mining methods would allow discovering new medical knowledge in the field of ophthalmologic data that can help to make the optimal solution of the preliminary diagnosis. Such a tool is important in e-health context [3].

3. CN2 Algorithm

The CN2 algorithm inductively learns a set of prepositional if/then rules from a set of training examples. To do so, it performs a general to specific beam search through rule –space for the “best” rule: then it removes the training example covered by that rule and repeats the previous two steps until no more good rules can be found.

```
Repeat
        -start with the general rule:
        “everything  → <class>”
        -specialize the rule;
        -retain the more significant disjunctive term;
Until no more rules to find [4].
```

The CN2 algorithm consists of two main procedures: a search algorithm performing a beam search for a good rule and a control algorithm for repeatedly executing the search [3]. The characteristic of CN2 are as follows:

- The representation language for the include knowledge
- The performance engine for executing the rules
- The learning algorithm and its associated search heuristic

3.1 Rule Induction with CN2

The simplest setting of classification rule induction is usually called propositional rule induction or attribute-value rule learning. The classification form of rule, where the training examples can be represented in a single table and the output are if-then rules.

IF Conditions THEN Class [Class Distribution]

Number in the Class Distribution list denote, for each individual class, how many training examples of this class are covered by the rule. This form of if-then rules is induced by the CN2 learner [3].

3.2 The CN2 Unordered Rules Algorithm

```
    Procedure CN2 unordered (all examples, classes);
    Let ruleset = {}
        For each class in classes:
            Generate rules by CN2 ForOneClass (all examples, class)
            Add rules to ruleset
    Return ruleset

    Procedure CN2 ForOneClass (examples, classes);
    Let rule = {}
    Repeat
        Call FindBestCondition (examples, classes) to find bestcond
        If bestcond is not null
        Then add the rule ‘if bestcond then predict class’ to rules
 & remove from examples all exs in class covered by bestcond
    Until bestcond is null
Return rules
```

3.3 The CN2 Rules Search Algorithm

```
    Procedure FindBest Condition(examples, [class]);
    let mgc=the most general condition(‘true’)
    let star initially contain only the mgc(i.e = {mgc})
    let bestcond = null
    while star is not empty
        let newstar={ }
        for each condition cond in star:
        for each possible attribute test not already tested on in cond
let cond’ = a specialization of cond, formed by adding test as an extra conjunct to cond(i.e cond’ = ‘cond & test’)
if cond’ is better than bestcond & cond’ is statistically significant
then let bestcond = cond’
add cond’ to newstar
if size of newstar > maxstar(a user-defined constant)
then remove the worst condition in newstar.
            let star = newstar
Return bestcond
```

4. System Implementation

This system implements CN2 rule induction algorithm. Among rule induction methods, CN2 is used as CN2 approximately equals to particle results. This paper focuses on diagnosis of pest or disease in groundnuts. CN2 algorithm is applied to generate the specific rules.

4.1 Overview of the system

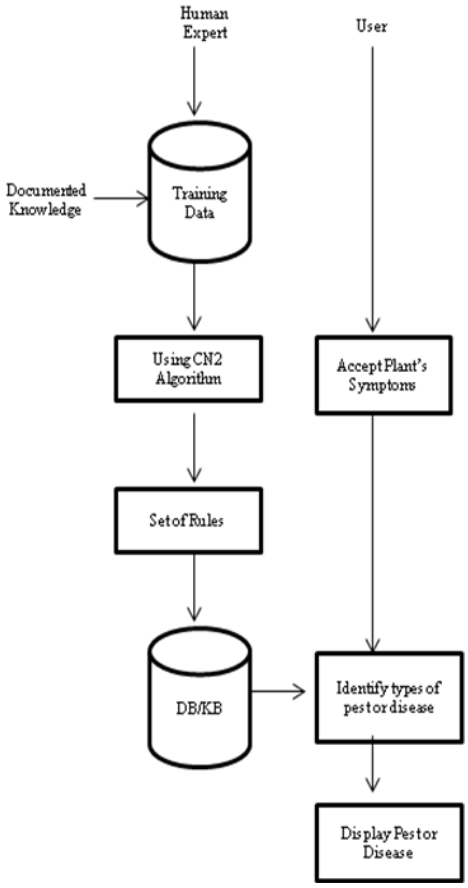

Figure 1. Overview of the System

The proposed system mainly focuses on prediction of pests and diseases according to the user-defined symptoms. First, the facts are acquired from expert and documented knowledge. The training data are applied to generate rules by using CN2. The plant's symptoms are accepted from the user and are matched with the generated rules, and then the types of pest or disease are defined according to the plant's symptoms that are given by user. The system displays type of pest or disease together with image of pest or disease, name and pesticides for treatment.

In this system, there are five menus such as View Training Dataset, Add Training Data, Find Refine Rule, Operation and System Accuracy. Each of menus has a choice to select pest or disease of groundnut. The system operates according to the user's choice.

In Operation menu and Add Training menu, the user has to fill the plant's symptoms that are seen on his or her plant. View Training Dataset is used to view the training data in the database, Add Training Data is to add training data into the database, Find Refine Rule is to generate the rules according to the training data set, Operation is to define the types of pest or disease of groundnuts according to the plant's symptoms that are given by the user and display the result pest or disease and System accuracy is to view the system accuracy for the system. The overview of the system is shown in figure 1.

4.2 Attributes and Classes of Pest and Disease of Groundnuts

In this system, there are ten attributes and five classes for pest; eleven attributes and six classes for disease. The attributes are Time, Soil, Blotch, Seed, Pod, Top of leaf, Lower surface of leaf, Leaf, Surface of leaf, Lower leaves, Stem, Plant, Stem touch with soil, Underground stem and Roots. The classes are Chafer Grub, Army Worm, Aphids, Hairy Caterpillar, Leaf Miner & Binder, Damping-off, Leaf Spot, Rust, Wedge Blotch, Web Blotch, and Crown Rot.

4.3 Sample of Rules

When the user clicks the Find Refine Rule, the rules are shown as a set of propositional if...then... classification rules by using CN2 algorithm. The sample rules of the system are shown in figure 2:

No.	Description
1	IF Seed=feed & destroy AND Underground Stem= bite & feed the bark AND Plant= be bluntly AND Stem= Bite & feed AND Lower Surface of Leaf=suck & feed fresh juice AND Top of leaf=droop & dry AND Leaf= be pierced & torn THEN Pest=Aphids
2	IF Blotch=spread through the leaf and dried the whole plant AND Stem= light brown spots or stripes were web blotch AND Plant= Be bluntly AND Lower leaf=light brown spots and stripes were web blotch AND Stem Touch With Soil=be in a mess AND Lower Surface of Leaf= Became yellow spots and turn to maroon AND Surface of leaf=dark brown, black irregular blotch AND Leaf= be dried THEN Disease=Crown rot
3	IF Soil= Sandy soil/ Dung Soil AND Roots= Bite & Feed AND Pod= Bite AND Seed= Eat & Destroy AND Underground Stem= Bite & Feed the bark THEN Pest=Chafer Grub

4	IF Leaf= Scrape green tissue of leaf and remain petiole AND Leaf=Bite & Feed AND Leaf= Be pierced & torn THEN Pest=Hairy Caterpillar
5	IF Leaf=make cobweb with thread and feed & destroy inside the cobweb AND Leaf= Be dried THEN Pest=Leaf Minder & Binder

Figure 2. Sample Rules of the System

4.4 Accuracy for the System

The evaluation of accuracy is motivated by considering the CN2 as knowledge acquisition tools for expert system. A useful system should induce rules that are accurate, so that they perform well, and comprehensible, so that they can be validated by an expert and used for explanation.

The accuracy is measured by splitting the data into a training set and a test set, presenting the algorithm with the training set to induce a concept description on the test set [2].

4.4.1 Experiments on domain

In this system, the accuracy for pest and disease of groundnuts are calculated. In each test, 70% of the training examples were selected and the remaining 30% of the data were used for testing. Figure 4.2 shows the characteristics of the pest and disease used in the system.

The testing data set is matched with the rules that are generated by CN2. If all the values of attributes are equal to rules, these data are correct. If it is not equal to the specific rules, these data are missing. Then the data is matched with rules one after another and increase the total correct when examples correctly predicted to be positive. In information retrieval, Recall is defined as the fraction of positive that are covered by the rule.

Recall (H←B) =P(B\H)

Where,

H = the set of instances for which the head is true

B = the set of instance for which the body of the rule is true

The relative frequency estimate of its accuracy is computed as p\p+n. The Laplace estimate is obtained by adapting the relative frequency to p+1\p+n+k, where p is the number of positive or correct, n is the number of negative or miss and k is the number of classes.

Domain Property	Pest	Disease
Number of Attributes	10	11
Values Per Attributes		
Minimum	1	1
Maximum	6	5
Number of Classes	5	6

Figure 3. Description of Pest and Disease

For example, when the user clicks Operation menu and choose pest of groundnut to view the system accuracy for pest, the testing data in the database is used to match the rules that are generated by using CN2 algorithm. The classes of pest are Aphids, Army Worm, Chafer Grub, Hairy Caterpillar and Leaf Minder & Binder. The attributes of pest are Soil, Roots, Pod, Seed, Underground stem, Plant, Stem, Lower Surface of Leaf, Top of Leaf and Leaf.

The system takes correct for one class when the values of all attributes are true and miss for incorrect. By this way, finally, the system gets the correct for class (p) and missing for class (n). Then accuracy calculation is stated as the following:

Class name		=Aphids
Number of classes	(k)	=5
Total number of data set for all classes		=128
Correct for Aphids	(p)	=13
Missing for Aphids	(n)	=115
So,		
Recall for one class		=p+1\p+n+k
Recall for Aphids		=13+1\13+115+5 =0.105263157

Then Recall for all classes of pest are calculated by this way. Finally, Recall for all classes are added and total is multiplied by100, and then the system accuracy for pest is shown with percentage in figure4.

No	Pest	Total	Correct	Missing	Recall
1	Aphids	128	13	115	0.105263
2	Army Worm	128	17	111	0.135338
3	Chafer Grub	128	21	107	0.165413
4	Hairy Caterpillar	128	23	105	0.180451
5	Leaf Miner & Binder	128	28	100	0.218045
				System Accuracy	80.451%

Figure 4. System Accuracy for Pest

The classes of diseases are Web Blotch, Leaf Spot, Crown Rot, Damping-off, Wedge blotch and Rust. The attributes of diseases are Surface of Leaf, Soil, Time, Lower Leaf, Blotch, Plant, Stem, Lower Surface of leaf, Stem Touch with soil, Leaf and Top of Leaf. The system accuracy of diseases is calculated above the way as in the pest. The system accuracy for disease is shown in figure 5.

No	Pest	Total	Correct	Missing	Recall
1	Web Blotch	150	20	130	0.120877
2	Leaf Spot	150	25	125	0.110283
3	Crown Rot	150	30	120	0.190179
4	Damping_ off	150	25	125	0.087761
5	Wedge blotch	150	25	125	0.149304
6	Rust	150	25	125	0.226417
System Accuracy					88.4821%

Figure 5. System Accuracy for Disease

5. Conclusion

This paper has been discussed the diagnosis of pest and disease in groundnut by using rules induction with CN2 algorithm. Rules can be generated by using rule induction method with data (or) experience without expert's entire help. CN2 is an algorithm designed to induce `if...then...' rules in domains where there might be noise. The proposed system is mainly focus on prediction of pests and diseases according to the user-defined symptoms. It saves time to consult an agriculture specialist for infection of pests or diseases. For the agriculture specialists, if they find new pest or disease, they can add new pest or disease easily. By developing an expert system in agriculture, it helps growers to know the facts and trusts in increasing the production.

References

[1] P. Clark and T. Niblett., " The CN2 induction algorithm in Machine Learning", 3(4):261:283, TheTuring Institute, 36 N.HanoverSt., Glasgow, G1 2AD,U.K, 1989.

[2] B. Kavˇsek and N. Lavraˇc1, "Using Subgroup Discovery to Analyze The UK Traffic Data", Metodoloski zvezki, Vol. 1, No.1, 249-264, 2004.

[3] J. Bernatavicien,, G. Dzemyda, O. Kurasova1,V.Barzdžiukas, D. Buteikien, A.Paunksnis, "Rule Induction For Ophthamological Data Classification", International Conference 20th EURO Mini Conference, Neringa, Lithuania, May20-23, 2008.

[4] H. Lounis, M. Boukadoum and V. Siveton, "Assessing Hydro Power System Relevant Variables: a Comparison Between a Neural Network and Different Machine Learning approaches", ALCAN, Ltd, group and CRIM, 2002.