*J. Ind. Soc. Agril. Statist.*
*63(1), 2009 : 97-107*

# Machine Learning for Forewarning Crop Diseases

Rajni Jain, Sonajharia Minz[1] and Ramasubramanian V.[2]
*National Centre for Agricultural Economics and Policy Research, New Delhi*
(Received: June 2006, Revised: November 2008, Accepted: December 2008)

**SUMMARY**

With the advent of computers, the development of accurate forewarning systems for incidence of crop diseases has been increasingly emphasized. Timely forewarning of crop diseases will not only reduce yield losses but also alert the stakeholders to take effective preventive measures. Traditionally, Logistic Regression (LR) and discriminant analysis methods have been used in forewarning systems. Recently, several machine learning techniques such as decision tree (DT) induction, Rough Sets (RS), soft computing techniques, neural networks, genetic algorithms etc. are gaining popularity for predictive modelling. This paper presents the potential of three machine learning techniques viz. DT induction using C4.5, RS and hybridized rough set based decision tree induction (RDT) in comparison to standard LR method. RS offers mathematical tools to discover hidden patterns in data and therefore its application in forewarning models needs to be investigated. A DT is a classification scheme which generates a tree and a set of rules representing the model of different classes from a given dataset. A java implementation of C4.5 (CJP) is used for DT induction. A variant of RDT called RJP, combines merits of both RS and DT induction algorithms. *Powdery mildew of Mango* (PWM) is a devastating disease and has assumed a serious threat to mango production in India resulting in yield losses of 22.3% to 90.4%. As a case study, prediction models for forewarning PWM disease using variables viz. temperature and humidity have been developed. The results obtained from machine learning techniques viz. RS, CJP and RJP are compared with the prediction model developed using LR technique. The techniques RJP and CJP have shown better performance over LR approach.

*Key words:* Forewarning crop diseases, Machine learning, Rough sets, RDT, Decision tree, Logistic regression, Powdery mildew of mango.

## 1. INTRODUCTION

With the advent of computers, the development of accurate forewarning systems for incidence of crop diseases has been increasingly emphasized. Crop diseases are one of the major causes of reduction in crop yield and hence timely application of remedial measures may combat the yield loss to a great extent. Forewarning systems can help in providing prior knowledge of the time and severity of the outbreak of such diseases. Crop diseases are influenced by interaction of various factors with the most significant of them being weather. Normally data on crop disease status and information on related variables (including weather) over years are utilized for developing models/rules for forewarning of diseases. Developing forewarning systems for crop diseases is now made relatively easier by increasingly research efforts in the application of advanced and complicated statistical computing concepts which include inter alia soft computing techniques such as neural networks, fuzzy theory, rough sets etc. Timely forewarning of crop diseases will not only reduce yield losses but also alert the stakeholders to take effective preventive measures. Forewarning consists of examining the features of a newly presented case and assigning it to a predefined class. In general it can be treated as task

[1] *School of Computer Science, Jawaharlal Nehru University, New Delhi 110067*

[2] *Indian Agricultural Statistics Research Institute, New Delhi 110012.*

of classification which is characterized by the well-defined classes, and a training set consisting of pre-classified examples. The task is to build a model called classifier that can be applied to unclassified data in order to classify it. Machine Learning offers many techniques like decision tree induction algorithms, neural networks, genetic algorithms, rough sets, fuzzy sets as well as many hybridized strategies for the classification (Han and Kamber, 2001; Pujari, 2000; Komorowski *et al.*, 1999; Witten and Frank, 1999). On the other hand, traditional statistical techniques such as Logistic Regression (LR) and discriminant analysis may be employed for the task of classification. The potential of three machine learning techniques viz. DT induction using C4.5, RS and hybridized rough set based decision tree induction (RDT) has been compared with the standard LR method. As a case study, prediction models for forewarning Powdery mildeW of Mango (PWM) disease using causal variables viz. temperature and humidity have been developed. While developing the models, the study also identifies best set of variables and the suitable algorithms for forewarning of PWM disease.

The purpose of this study has arisen out the need for developing crop disease forewarning systems which are evolved upon reliable, robust and improved soft computing methods. Various approaches are in vogue to build such early warning systems. Every approach has its own advantages and limitations. Soft computing techniques can be advantageously used in certain situations to convert abstract knowledge and heuristics into easily comprehensible rules. The expected gain in accuracy by using soft computing concepts such as rough sets or its hybridized model may justify the effort involved in using them in preference to the conventional models.

The rest of the paper is organised as follows. Section 2 deals with the preliminaries. Section 3 describes a case study. Section 4 presents the methodology used followed by results and discussion in Section 5. Finally, the conclusions are presented in the Section 6.

## 2. PRELIMINARIES

### 2.1 Logistic Regression

Let class variables are of 0-1 type. To handle the task of classification (Hastie *et al*. 2001) in Logistic Regression (LR) approach, the probability of membership in the first group, $p_1(x)$, is modelled directly as in equation (1) for the two categories problem where $\alpha$ and $\beta$ are the parameters.

$$p_1(x) = \frac{e^{\alpha+\beta' x}}{1+e^{\alpha+\beta' x}} \quad (1)$$

### 2.2 Decision Tree

Decision tree induction represents a simple and powerful method of classification which generates a tree and a set of rules, representing the model of different classes, from a given dataset (Winston 1992). Decision Tree (DT) is a flow chart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test and each leaf node represents the class (Hans and Kamber 2001). The top most node in a tree is the root node. For DT induction, ID3 algorithm and its successor C4.5 algorithm by Quinlan (1993) are widely used. Algorithm CJP (java implementation of C4.5) is used in this paper for DT induction. One of the strengths of decision trees compared to other methods of induction is the ease with which they can be used for numeric as well as non-numeric domains. Another advantage of decision tree is that it can be easily mapped to rules. The classical DT induction algorithm i.e. C4.5 by Quinlan (1993) is presented below for better understanding to the readers.

#### 2.2.1 C4.5 algorithm

Let the training data is a set $S = s_1, s_2, ...$ of already classified samples. Each sample $s_i = x_1, x_2, ...$ is a vector where $x_1, x_2, ...$ represent attributes or features of the sample. The training data is augmented with a vector $C = c_1, c_2,...$ where $c_1, c_2,...$ represent the class that each sample belongs to. C4.5 uses the fact that each attribute of the data can be used to make a decision that splits the data into smaller subsets. C4.5 examines the normalized Information Gain (difference in entropy) that results from choosing an attribute for splitting the data. Entropy($S$) can be thought of as a measure of how random the class distribution is in $S$. Information gain is a measure given to an attribute a. Attribute a can separate $S$ into subsets $S_{a1}, S_{a2}, S_{a3}, ..., S_{an}$. The information gain of a is then Entropy($S$) – Entropy($S_{a1}$) – Entropy($S_{a2}$) – ... – Entropy($S_{an}$). Information gain is then normalized by multiplying the entropy of each attribute choice by the proportion of attribute values that have that choice. The attribute with the highest normalized information

gain is the one used to make the decision. The algorithm then recurs on the smaller sub lists. The pseudo code of the algorithm is as follows:

1. Check for base cases
2. For each attribute *a*

   2.1 Find the normalized information gain from splitting on *a*
3. Let a-best be the attribute with the highest normalized information gain
4. Create a decision node *dnode* that splits on a-best
5. Recur on the sub lists obtained by splitting on a-best and add those nodes as children of *dnode*

### 2.3 Rough Set Theory

RS theory was introduced in early 1980s by *Z*. Pawlak, a Polish mathematician and has been widely explored for pattern discovery since then. RS emerged as an important mathematical tool for managing uncertainty that arises in the indiscernibility between objects in a set, and has proved to be useful in a variety of knowledge discovery processes (Pawlak 1991, Komorowski 1977). Some of the basic terms and concepts pertaining to RS are discussed below.

#### 2.3.1 Information system and decision table

In RS, knowledge is a collection of facts expressed in terms of the values of attributes that describe the objects. These facts are represented in the form of a data table. Entries in a row represent an object.

A data table described as above is called an information system. Formally, an information system S is a 4-tuple, $S = (U, Q, V, f)$ where, $U$ a nonempty, finite set of objects is called the universe; $Q$ a finite set of attributes; $V = \cup Vq, \ \forall\, q \in Q$ and $Vq$ being the domain of attribute $q$; and $f : U \times Q \rightarrow V$, $f$ be the information function assigning values from the universe $U$ to each of the attributes $q$ for every object in the set of examples.

#### 2.3.2 Indiscernibility relation

For a subset $P \subseteq Q$ of attributes of an information system $S$, a relation called indiscernibility relation denoted by IND is defined in equation (2).

$$\mathrm{IND}_s(P) = \{ (x, y) \in U \times U : f(x, a) = f(y, a) \ \forall\, a \in P\} \quad (2)$$

The function $f(x, a)$ assigns the value of the attribute a for an object $x$. If $(x, y) \in \mathrm{IND}_s(P)$ then objects $x$ and $y$ are called indiscernible with respect to $P$. The subscript s may be omitted if information system is implied from the context. IND($P$) is an equivalence relation that partitions universe $U$ into equivalence classes, the sets of objects indiscernible with respect to $P$. Set of such partitions are denoted by $U$/IND($P$).

#### 2.3.3 Approximation of sets

Let $X \subseteq U$ be a subset of the universe. A description of $X$ is desired that can determine the membership status of each object in $U$ with respect to $X$. Indiscernibility relation is used for this purpose. If a partition defined by IND(P) (denoted by $Y$ in Equation 3) partially overlaps with the set $X$, the objects in such an equivalence class can not be determined without ambiguity. The description of such a set $X$ is defined in terms of P-lower approximation (denoted as $\underline{P}$) and P-upper approximation (denoted as $\overline{P}$) where for $P \subseteq Q$:

$$\underline{P}X = \cup\{Y \in U / \mathrm{IND}(P) : Y \subseteq X\}$$
$$\overline{P}X = \cup\{\mathrm{Y} \in U / \mathrm{IND}(P) : Y \cap X \neq \varnothing\} \quad (3)$$

A set $X$ for which $\underline{P}X = \overline{P}X$ is called an exact set otherwise it is called rough set with respect to $P$.

#### 2.3.4 Dependency of attributes

RS introduces a measure of dependency of two subsets of attributes $P$, $R \subseteq Q$. The measure is called a degree of dependency of $P$ on $R$, denoted by $\gamma_R(P)$. It is defined as

$$\gamma_R(P) = \frac{\mathrm{Card}(\mathrm{POS}_R(P))}{\mathrm{Card}(U)} \ \text{ where } \ \mathrm{POS}_R(P)$$
$$= \bigsqcup_{X \in U / \mathrm{IND}(P)} \underline{R}X \quad (4)$$

The set $\mathrm{POS}_R$(P), positive region, is the set of all the elements of $U$ that can be uniquely classified into partitions U/IND($P$) by $R$. Here, Card refers to the cardinality of the set included in the parenthesis. Thus, numerator and denominator are the number of objects in the positive region denoted by $\mathrm{POS}_R(P)$ and the universe $U$ respectively. Coefficient $\gamma_R(P)$ represents the fraction of the number of objects in the universe

which can be properly classified. If $P$ totally depends on $R$ then $\gamma_R(P) = 1$, else $\gamma_R(P) < 1$.

### 2.3.5 Reduct

The minimum set of attributes that preserves the indiscernibility relation is called a reduct. The relative reduct of the attribute set $P$, $P \subseteq Q$, with respect to the dependency $\gamma_P(Q)$ is defined as a subset $\text{RED}(P, Q) \subseteq P$ such that:

1. $\gamma_{\text{RED}(P, Q)}(Q) = \gamma_P(Q)$, i.e. relative reduct preserves the degree of inter attribute dependency
2. For any attribute $a \in \text{RED}(P, Q)$, $\gamma_{\text{RED}(P,Q)-\{a\}}(Q) < \gamma_P(Q)$ i.e. the relative reduct is a minimal subset with respect to property 1.

Computation of a minimal optimum reduct is a NP hard problem. However a single relative reduct can be computed using efficient heuristics. Johnson's algorithm is one such method which is available in Rosetta software (http://www.idi.ntnu.no/~aleks/rosetta/).

### 2.3.6 Rule discovery

Rules can be perceived as data patterns that represent the relationships between attribute values. RS theory provides mechanism to generate rules directly from the dataset by reading the values of the attributes present in reduct from the given decision table.

## 2.4 Proposed Model – Rough Set based Decision Tree (RDT)

RDT model as proposed by Minz and Jain (2003a) combines merits of both RS and DT induction algorithm. It aims to improve efficiency, simplicity and generalization capability of both the base algorithms as shown by Minz and Jain (2003b). In the present study, a variant of RDT called RJP (Table 3) is used as a representative of RDT approach. Algorithm RJP for the induction of rough decision tree is presented below (Jain and Minz 2003).

**Algorithm RJP**

1. Input the training dataset say T1.
2. Discretize the continuous attributes if any, and label the modified dataset as T2.
3. Obtain the minimal decision relative reduct of T2, say R.
4. Reduce T2 based on reduct R and label the reduced dataset as T3.
5. Apply C4.5 algorithm on T3 to induce decision tree.
6. Convert the decision tree to rules (if needed) by traversing all the possible paths from root to each leaf node.

The training data-T1 is a collection of examples used for supervised learning to develop the classification model. In step 2, continuous attributes of the dataset (if any) are discretized. The next step involves computation of a reduct R. The reduct helps in reducing the training data, which is finally used for decision tree induction. Algorithms like Boolean reasoning algorithm, Johnson's algorithm or Genetic Algorithms can be used for the computation of the optimal reduct. In this paper, Johnson's algorithm based on efficient heuristics (implemented in Rosetta software), is used for the computation of a single reduct. More details pertaining to RDT model are available in Minz and Jain (2005).

## 3. CASE STUDY

Powdery Mildew of Mango (PWM) caused by Oidium mangiferae Berthet is responsible for foliar as well as inflorescence infection in mango. Generally, PWM epidemic occurs in the third and fourth week of March when the inflorescences are of the age of 6-7 weeks. The spread of the disease is greatly manifested by factors such as temperature, humidity, wind velocity, dews, wind direction etc. because it is an airborne disease.

The PWM dataset for the study has been taken from the project "Epidemiology and forecasting of PWM" undertaken at Central Institute for Subtropical Horticulture, Uttar Pradesh. From the original data, the attributes relative humidity and maximum temperature are selected because of the prior information available about contribution of these factors to the occurrence of PWM (Misra *et al*. 2004). As repeated life cycles of PWM are around 4-7 days, periods from 8th of March up to 14th of March i.e. one day a prior to the start of possible occurrence of epidemic (3rd week of March) were taken for developing forewarning models. Moving averages of maximum temperature and relative humidity are computed for March $8^{th}$ –$11^{th}$ , $8^{th}$ –$12^{th}$ , $8^{th}$ –$13^{th}$ and $8^{th}$ –$14^{th}$ and are referred by the set of corresponding pair of variables as{(T811, H811), (T812, H812), (T813,

**Table 1.** Pre-processed dataset for Powdery Mildew of mango

| Year | T811 | H811 | T812 | H812 | T813 | H813 | T814 | H814 | STATUS |
|---|---|---|---|---|---|---|---|---|---|
| 1987 | 28.20 | 91.25 | 28.48 | 88.60 | 29.50 | 85.50 | 30.14 | 82.86 | 1 |
| 1988 | 32.05 | 75.75 | 31.64 | 80.60 | 31.27 | 77.83 | 30.66 | 79.57 | 0 |
| 1989 | 26.60 | 86.25 | 26.28 | 87.20 | 26.47 | 87.33 | 26.31 | 89.14 | 0 |
| 1990 | 27.50 | 91.25 | 28.12 | 91.20 | 28.17 | 92.00 | 28.43 | 91.00 | 1 |
| 1991 | 28.43 | 87.00 | 28.70 | 86.20 | 29.00 | 83.50 | 29.57 | 80.57 | 0 |
| 1992 | 30.12 | 69.23 | 30.45 | 68.58 | 30.80 | 68.31 | 31.25 | 67.82 | 1 |
| 1993 | 30.50 | 61.75 | 30.48 | 61.13 | 30.37 | 60.56 | 30.33 | 61.76 | 0 |
| 1994 | 30.45 | 89.25 | 30.56 | 85.80 | 30.63 | 83.17 | 30.71 | 81.14 | 1 |
| 1995 | 28.63 | 61.38 | 29.10 | 61.20 | 29.58 | 61.17 | 30.71 | 61.57 | 0 |
| 1996 | 31.63 | 60.33 | 31.90 | 60.87 | 32.67 | 60.89 | 33.07 | 59.76 | 1 |
| 1997 | 32.13 | 71.00 | 32.20 | 69.40 | 31.67 | 69.00 | 31.50 | 68.29 | 0 |
| 2000 | 29.00 | 78.33 | 29.23 | 78.60 | 29.36 | 78.83 | 29.52 | 79.14 | 0 |

H813), (T814, H814)} in Table 1. Data is partitioned into train and test pairs as shown in Table 2. For example, the entry 1987-94 under the first column called MODEL means the data for the years 1987-94 is used for learning the model while the data for the years 1995-97 and 2000 is used for the model validation.

## 4. METHODOLOGY

All the eight independent variables as shown in Table 1 along with STATUS as dependent variable were used as input for the machine learning algorithms. However in case of LR, only two variables can be taken at a time as the number of observations in the dataset (Table 1) is less. Machine learning algorithms as well as traditional logistic regression method are employed using the training and test data pairs as identified in Table 2. The algorithms and the corresponding software that are used in this paper for forewarning of PWM disease are presented in Table 3. The Logistic Regression (LR) model has already been applied to the dataset by Misra *et al*. (2004). The redundant variables (if any) are filtered using concepts of information theory in CJP algorithm and using concept of reducts in RS and RJP algorithm. In this section, we demonstrate the characteristics of the output from each of the algorithms (Table 3) with the help of an example using the train data for the years1987-97 and the test data for the year 2000 (Table 2). The overall mean accuracy of the models for each algorithm is presented and discussed in Section 5.

**Table 2**. Train and test data pairs for different models

| Model | Train Data | Test Data |
|---|---|---|
| 1987-94 | 1987-94 | 1995, 1996, 1997, 2000 |
| 1987-95 | 1987-95 | 1996, 1997, 2000 |
| 1987-96 | 1987-96 | 1997, 2000 |
| 1987-97 | 1987-97 | 2000 |

**Table 3.** Learning algorithms used for PWM dataset

| Id | Algo | Description | Software | Model |
|---|---|---|---|---|
| 1 | LR | Logistic Regression | SAS | Coeffi-cients |
| 2 | RS | Rough Set reducts (decision relative full discernibility global) | Rosetta | Rules |
| 3 | CJP | Java Implementation of C4.5 Pruned | Weka | DT |
| 4 | RJP | Rough set based DT induction embedding J4.8 for DT induction with Pruned tree | Rosetta, Weka, C++ programs | DT |

### 4.1 LR

Consider the LR model given in equation (1) noting that for the two variables case, the expression $\alpha + \beta' x$ would be $a + bT(.) + cH(.)$. Table 4 shows estimates of parameters *a*, *b* and *c* for the model using the attribute pairs i.e. (*T*811, *H*811), (*T*812, *H*812), (*T*813, *H*813) or (*T*814, *H*814) separately for the train data 1987-97 (Table 2). Using these parameter estimates, the outcomes for the training set and the test set can be predicted by plugging in the corresponding parameter estimates from Table 4 in equation (1).

To decide whether the status of the disease is of epidemic nature, it is necessary to have a cut off value beyond which the probability value lies. It is fairly realistic to keep as a thumb rule the cut off value of probability as 0.5. Then if probability is less than this value then the event that epidemic will occur will be minimal, otherwise there is more chance of occurrence of disease in epidemic proportions. It is emphasized here that there is no objective procedure to be considered as a general rule. If one wants to be more stringent, then the cut off value can be increased as per requirement. Statistically speaking, depending upon the problem under consideration there is always a possibility of error because we deal with sample data for model development. Thus, the consideration of 0.5 as a cut off value in the present study is to a greater extent appropriate.

For the test data 2000, using (*T*811, *H*811), (*T*812, *H*812), (*T*813, *H*813) or (*T*814, *H*814), the corresponding $p(x)$ values as defined in equation (1) are 0.44, 0.44, 0.39 and 0.29. All these probabilities being less than 0.5 imply that the predicted STATUS is 0 which is same as the observed STATUS (Misra *et al.* 2004).

**Table 4.** Parameters of the LR model developed for the PWM prediction

| Model | Years | 1987-97 |
|---|---|---|
| $8^{th}$ to $11^{th}$ day | *a* | –10.79 |
| | *b* | 0.19 |
| | *c* | 0.06 |
| $8^{th}$ to $12^{th}$ day | *a* | –13.97 |
| | *b* | 0.3 |
| | *c* | 0.06 |
| $8^{th}$ to $13^{th}$ day | *a* | –36.95 |
| | *b* | 0.88 |
| | *c* | 0.13 |
| $8^{th}$ to$14^{th}$ day | *a* | –70.43 |
| | *b* | 1.73 |
| | *c* | 0.24 |

### 4.2 RS

Employing RS approach for the different train data (Table 2), the set {*H*811, *T*814} is a computed reduct (Section 2.3.5). By using the discretized train data of the years 1987-97, the following three rules are generated. The rules are simple to comprehend for applying to the unseen dataset.

1. If (*H*811> = 88.2) AND (*T*814 <31.0) => then STATUS =1
2. If (*H*811 < 88.2)) AND (*T*814<31.0) => then STATUS = 0
3. If (*H*811<88.2)) AND (*T*814> = 31.0) => then STATUS = 1

The rules when applied to the test data i.e. the year 2000, correct prediction is obtained as illustrated in the following example.

**Example 1:** For the year 2000, *H*811 = 78.33 and *T* 814 = 30.71. Observing the three rules, we can identify that the Rule 2 is applicable to this dataset. Therefore, predicted value of the STATUS is 0. This is verified by the observed value of the STATUS (Table 1).

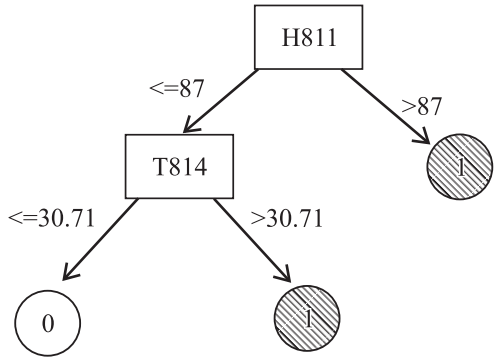
### 4.3 CJP

The prediction model which is obtained by employing the CJP algorithm using data of 1987-97 as the train data, is represented as decision tree in Figure 1. The corresponding rules are obtained by following the path from the root of the decision tree towards its leaf (Fig. 1).

**Example 2:** For the year 2000, we observe from Table 1 that *H*811 = 78.33 and *T*814 = 30.71. Consider the decision tree (output of CJP) in Figure 1. Starting from the root node and following the tree as per matching of the conditions in each branch, we reach the final node, also called leaf, having value 0. Therefore the predicted STATUS = 0 which is same as the observed value of the STATUS (Table 1). The prediction method using the rules from the decision tree is similar to Example 1.

### 4.4 RJP

Like CJP, the model obtained from RJP algorithm is a decision tree which can be mapped to rules (Fig. 2). However, it is observed that the branches representing the conditions are different from the branches of the decision tree from CJP algorithm.

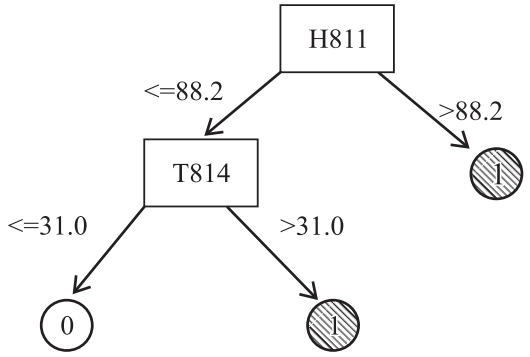



**Rules:**

1. If (H811>87) then Status = 1
2. If (H811<=87) and (T814<=30.71) then Status = 0
3. If (H811<=87) and (T814>30.71) then Status = 1

Note:
The Status = 1 means that epidemic will occur while Status=0 means no epidemic.

**Fig. 1**. The Prediction model for PWM Epidemic as obtained using CJP Algorithm on 1987-97 data as the training dataset




**Rules:**

1. If (H811>88.2) then Status = 1
2. If (H811<=88.2) and (T814<=31.0) then Status = 0
3. If (H811<=88.2) and (T814>31) then Status = 1

Note:
The Status = 1 means that epidemic will occur while Status=0 means no epidemic.

**Fig. 2**: The Prediction model for PWM Epidemic as obtained using RJP Algorithm on 1987-97 data as the training dataset

### 4.5 Performance Evaluation

Considering the costs associated with the wrong prediction of the PWM disease, accuracy is considered as the most important evaluation measure. In order to estimate the average accuracy initially the models are developed using the algorithms listed in Table 3. The average accuracy of the corresponding models was computed for each algorithm using training data and test data pairs (Table 2) and by considering each of the attribute sets - {*T*811, *H*811}, {*T*812, *H*812}, {*T*813, *H*813}, {*T*814, *H*814}. To determine average accuracy using all variables attribute set {*T*811, *H*811, *T*812, *H*812, *T*813, *H*813, *T*814, *H*814} is used for each algorithm. STATUS is used as the decision variable for each run of the algorithm. The superset of all the attributes, for example {*T*811, *H*811, *T*812, *H*812, *T*813, *H*813, *T*814, *H*814}, has not been used for LR because of its limitation in handling datasets when the order of the number of attributes is same as the number of observations. However, all the variables along with the pairs of variables are used for the RS, CJP and RJP algorithms to investigate whether the accuracy estimates improve by including all the variables for the training. Inclusion of all variables helps to identify the best set of input attributes for learning the model.

## 5. RESULTS AND DISCUSSION

The comparative discussion of the algorithms is done using the Fig. 3 and Fig. 4. Fig. 3(i) shows the overall mean accuracy of the models using two variables at a time. For example, mean accuracy of RJP algorithm in Fig. 3 (i) is computed using the formula:

$$\text{Mean Accuracy} = \frac{\sum_{v}\sum_{m} \text{Accuracy}_{vm}}{n} \qquad (5)$$

where each *v* i.e. variable pair in {{*T*811, *H*811}, {*T*812, *H*812},{*T*813, *H*813},{*T*814, *H*814}}; each *m* in {1987-94, 1987-95, 1987-96, 1987-97}; and *n* = number of observations.

Further, part (ii) of Fig. 3 presents the estimated mean accuracy using the variable set (*T*811, *H*811, *T*812, *H*812, *T*813, *H*813, *T*814, *H*814). As LR does not permit use of all variable pairs together, the results are not shown for LR in this figure.

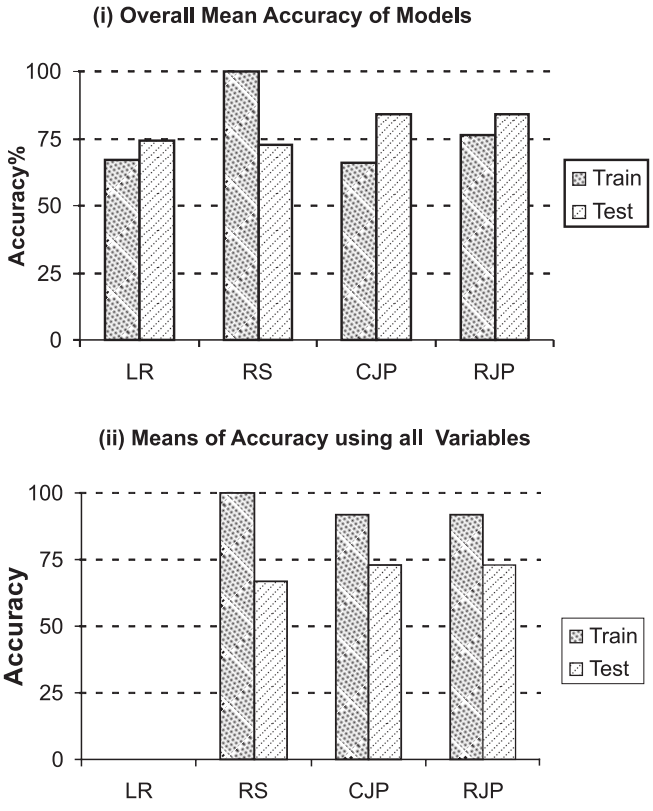



**Fig. 3**. (i)The Mean of accuracy with two variables at a time
(ii) The Mean of accuracy from with all variables at a time. As LR does not allow using all variables, the bar is not shown for LR in this graph

Fig. 4 helps in comparing and selecting the appropriate pair of variables for disease prediction. Using the results for the test data, the variables *T*811, *H*811 can be selected for better predicting capabilities

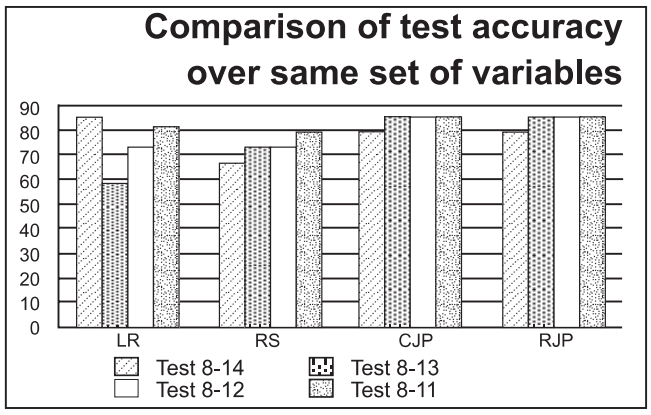



**Fig. 4.** Comparison of accuracy estimated by using data over the same set of variables

**Note:** LR gives its best performance on the test data using moving averages of max. temperature and humidity covering 8-14 days while other algorithms can give best accuracy estimates using 8-11 days of temperature and humidity information.

in all machine learning algorithms. Although the results for the training data exhibited different scenario, yet the average accuracy estimates for the test data are considered more realistic. The behaviour of each of the algorithm as observed from the Fig. 3 and Fig. 4 is discussed separately.

### 5.1 LR

Trends pertaining to the mean of computed accuracies are shown in Fig. 3(i). Training accuracy is observed to be lesser as compared to test accuracy (75%) for this method. For the purpose of identifying the best pair of variables, mean accuracies are presented for each set of the variables in Fig.4. When the attributes corresponding to the moving averages of more number of days are used the test accuracy estimate worsens except for the *T*814 and *H*814 (Fig. 4). It is in contrast to the general opinion that more data, either in terms of the size of the training data or in terms of more days in computing the moving averages would result in improved performance of the model (Fig. 4). As per the test data estimates, it is suggested to use *T*811, *H*811 variables for forewarning of PWM disease using LR algorithm because this results in forewarning well ahead in time without much loss of accuracy as compared to *T*814, *H*814 variables.

### 5.2 RS

Attribute pairs used for LR were also used for RS for the sake of comparison between the two. The mean accuracy on the training set is observed as 100 per cent for each case irrespective of the size of the training data or the set of attributes (Fig. 3(i) and Fig. 3(ii)). Mean test accuracy for RS is observed to be much less as compared to training accuracy for paired variables as well as while using all the 8 variables together. When performance of RS is compared with LR, it is observed that RS performs well on training data while LR is better for test data (Fig. 3). To identify the best pair of variables, it is observed from Figure 4 that with addition of one day in computation of moving averages, test accuracy deteriorates from 79.2 per cent for the variables (*T*811, *H*811) to 66.7 per cent for the variables (*T*814, *H*814). Thus, the pair (*T*811, *H*811) is recommended for forewarning of PWM disease using RS algorithm. However, use of all the 8 attributes as input to the RS algorithm has resulted in identification of *H*811 and *T*814 as relevant attributes with mean accuracy shown in Fig. 3(ii).

### 5.3 CJP

Comparison of overall mean accuracies of CJP with LR and RS shows that CJP performs well on the test data unlike training data (Fig. 3). As the aim of any forewarning model is to have better prediction for the test cases which are unseen as well, CJP is preferable in comparison to LR and RS approaches for disease prediction. Figure 4 shows that increasing one day at a time for computation of moving averages of attributes does not affect the test accuracy except for the case of *T*814, *H*814 showing little decrease in test accuracy. Thus for the reasons mentioned as above for LR algorithm, *T*811, *H*811 is recommended for predicting PWM disease. Parallel use of all attributes results in the selection of the attribute set {*H*811} or {*H*811, *T*814} as the most relevant attributes. However, it has not resulted in improvement of the test accuracy over the pair of attributes.

### 5.4 RJP

Test accuracy is improved for RJP algorithm as per expectations even though it does not show 100 per cent accuracy on the training data unlike RS (Fig. 3). Like CJP, increasing the number of days in computation of attributes does not show any impact on test accuracy but decrease the test accuracy for the case of *T*814, *H*814 (Fig. 4). Hence variables *T*811, *H*811 are recommended as best set of variables for predicting PWM. Here, we would also like to mention that slight decrease in test accuracy on adding an extra day for computing moving average is not a strange behaviour because biological cycle of a pathogen depends more on the weather conditions as compared to the exact date of a calendar month. Use of all variables as input to RJP identifies *H*811 and *T*814 as the most significant attributes. However, it has not resulted in improvement of the test accuracy over the pair of attributes.

In Fig. 3, we observe that for three algorithms namely LR, CJP and RJP, test accuracy is more than the training set accuracy. Although this behaviour is not commonly observed, yet it is not unusual. There have been a number of published reported results (Table 5) in the literature on different datasets using different models and algorithms where the training set accuracy is observed to be lesser as compared to the test set accuracy [Clark et al ( 1989), Mitra *et al.* (1997), Duch (2001)].

**Table 5.** Some reported cases where training set accuracy is less than test set accuracy

| S.No. | Model | Training | Test | Dataset | References |
|---|---|---|---|---|---|
| 1 | NN | 76.9 | 80.4 | Hepatobiary disorder | Duch *et al.* 2001 |
| 2 | AN | 87.5 | 88.69 | Vowel data | Mitra *et al.* 1997 |
| 3 | SN | 98.11 | 100 | Hepato | Mitra *et al*. 1997 |
| 4 | C4.5 | 89.0 | 89.8 | Quadrant-200 | Klaus *et al*. 1995 |
| 5 | Default Rule | 54 | 56 | Lymphography | Clark *et al*. 1989 |
| 6 | Default Rule | 70 | 71 | Breast Cancer | Clark *et al.* 1989 |
| 7 | Default Rule | 23 | 26 | Primary Tumor | Clark *et al.* 1989 |

A special mention is also needed regarding 100 per cent accuracy of prediction in some cases (Fig. 3). It is emphasised that100 per cent accuracy for the training data may occur due to over fitting. But, whether the high accuracy over the training data holds good for future prediction will be substantiated if similar performance is observed for the test data as well. For example, in the present analysis, RS exhibits 100 per cent accuracy over the training data. But this can not be substantiated because RS performs badly over the test data. Thus, performance of RS on the training data is attributed to over fitting as is evident from its relatively worse performance on the test data.

As training set accuracy is not considerably important for the purpose of final comparison of algorithms, mean accuracies of the test data as obtained from all the algorithms is compared in Table 6. It is evident that test performance of CJP and RJP are comparable. Hence, among the algorithms CJP and RJP are recommended for prediction of PWM.

**Table 6.** Comparison of average of test accuracy (in per cent) of various algorithms

| Variables used | LR | RS | CJP | RJP |
|---|---|---|---|---|
| Pairwise | 75 | 74 | 83 | 84 |
| All | * | 62 | 74 | 74 |

**Note: '*'** indicates all variables were not used together in LR because of its limitation in handling all 8 variables together.

### 5.5 Differential Behaviour of Various Techniques and Contribution of the Study

The behaviour of the various techniques on training data and test data and their pair wise comparison can be explained by putting them into four categories (Table 7). For example, If we compare any two algorithms say A1 and A2 using the training and the test data, then the behaviour of the algorithm A1 over the algorithm A2 would belong to one of the four categories say 1, 2, 3 or 4 (see column Catg in Table 7). In this table, the entry 'worse' under the column 'Training Data' means that algorithm A1 (the first in the pair (A1,A2)) is shown to perform worse than the algorithm A2 for the training data. Similarly the entry 'better' under the column 'Test Data' means that algorithm A1 is shown to perform better than or equal to the algorithm A2 for the test data. The overall preference of the algorithm for the corresponding category is known by the comment on the overall performance of algorithm A1 over algorithm A2 (Table 7).

**Table 7.** The categories of different behaviour on pair wise comparison of prediction accuracy of an algorithm $A_1$ with algorithm $A_2$

| (Algo A1, Algo A2) | Accuracy of algo A1 over algo A2 on | | CatG | Comment on Performance of A1 with respect to A2 |
|---|---|---|---|---|
| | Training data | Test data | | |
| (LR, RJP) | worse | worse | 1 | worse |
| (LR, CJP), (RS, LR), (RS, CJP), (RS, RJP) | better | worse | 2 | worse |
| (LR, RS), (CJP, RS), (CJP, RJP), (RJP, RS) | worse | better | 3 | better |
| (CJP, LR), (RJP, LR), (RJP, CJP) | better | better | 4 | better |

Category 1 includes the situation where algorithm A1 performs worse than the algorithm A2 over training as well as test data. Under this situation A2 is recommended over A1 for prediction. Comparison of LR with RJP denoted by (LR, RJP) belongs to this category (Table 7, Fig. 3). Thus, RJP emerges as the better performer than LR in this comparison.

Category 2 includes the behaviour that perform exclusively better for training data but contrastingly worse for the test data e.g. RS approach gives better accuracy over training but worse for test when compared with LR, CJP or RJP algorithms (Table 7, Fig.3,). The good performance over the training data is not important but test data performance is certainly important while comparing the algorithms. Hence in this category, algorithm 2 is considered better over algorithm 1. Comparisons of (RS, LR), (LR, CJP), (RS, CJP), (RS, RJP) fall under category 2. Here, it is observed that LR performs better than RS and CJP performs better than LR. This implies CJP is better than RS as well as LR. Further, (RS, RJP) implies that RJP is better than RS.

Category 3 includes the situation where algorithm A1 is worse than the algorithm A2 on training data but better than the algorithm A2 on the test data. In such cases algorithm A1 is to be selected for forewarning because they have shown better performance on the test data due to their least tendency towards over fitting during model learning. The cases of (LR, RS), (CJP, RS) and (CJP, RJP) and (RJP, RS) were observed to belong into this category (Table 7, Fig. 4). Here, the algorithms LR, CJP and RJP emerge better in pair wise comparison, but LR is being rejected in its comparison with other algorithms falling under category 4.

Category 4 includes the behaviour where the algorithm A1 performs better over algorithm A2 on training data as well as test data. Whenever any algorithm is able to achieve this, it means the model is perfect, model has truly captured the causing agents of the disease. Naturally, algorithm A1 is considered better over A2 in this category. The cases of (CJP, LR), (RJP, LR) and (RJP, CJP) were observed to belong into this category (Table 7, Fig.3). Consequently, CJP and RJP are recommended for prediction of the PWM disease.

Based on the discussion in this section, contributions of this study in predicting PWM disease are

1. CJP and RJP model are recommended for forewarning PWM because of better predicting accuracy over conventional method namely LR.
2. Temperature and humidity values pertaining to 8-11 days is found more appropriate for predicting PWM disease.
3. The underlying assumption regarding normal distribution of the values of the variables is not necessary to have better prediction.

4. The resulting model i.e. rules and DT are easy to interpret as well as easy to apply in comparison to classical method of LR.

## 6. CONCLUSIONS

Powdery Mildew of Mango (PWM) is a devastating disease and a prediction model to forewarn the epidemic outbreak of PWM using data from historical years is required. Predictive models are developed using the algorithms LR, RS, CJP and RJP by using different training-test pairs and attributes representing weather parameters. The results support the recommendation of CJP and RJP for prediction in crop diseases as it performs better than LR and RS in terms of performance parameters. The resulting models are easy to understand and implement without much technical expertise. The temperature and humidity variables relating to 8th-11th days of month of March are recommended for predicting PWM disease.

## ACKNOWLEDGEMENTS

The authors are thankful to the referee for the very useful comments that helped in improving the quality of the manuscript.

## REFERENCES

Clark, P., Niblett, T. (1989). The CN2 induction algorithm. *Machine Learning,* **3(4),** 261-283.

Duch,W., Adamczak, R. and Grabczewski K. ( 2001). A new methodology of extraction, optimization and application of crisp and fuzzy logical rules. *IEEE Transactions on Neural Networks,* **12(2),** 277-305.

Han, J. and Kamber, M. (2001). *Data Mining Concepts and Techniques.* Morgan Kaufmann Publisher.

Hastie, T., Tibshirani, R. and Friedman (2001). *The Elements of Statistical Learning: Data Mining, Inference and Prediction.* Springer.

Jain, R. and Minz, S. (2003). Should decision trees be learned using rough sets? In *Proc. 1st Indian International Conference on Artifical Intelligence* (IICAI-03), 1466-1479, Hyderabad.

Komorowski, J., Pawlak, Z., Polkowki, L. and Skowron, A. (1999). Rough Sets: A Tutorial. In : *Rough Fuzzy Hybridization, Pal S.K. and Skowron, A.(eds.),* Springer, 3-99.

Krzanowski, W. J. (1977). The performance of Fisher's linear discriminant function under non-optimal conditions. *Technomettrics,* **19(2)**, 191-200.

Minz, S. and Jain, R. (2005). Refining decision tree classifiers using rough set tools. *Int. J. Hybrid Intell. Sys.,* **2(2),** 133-148.

Minz, S. and Jain, R. (2003a). Rough set based decision tree model for classification. *Proc. 5th International Conference on Data Warehousing and Knowledge Discovery, (DaWaK 2003) Prague,* Czech Republic, September 3-5, 2003, LNCS 2737, 172-181.

Minz, S. and Jain, R. (2003b). Hybridizing rough set framework for classification: An experimental view. In : *Design and Application of Hybrid Intelligent Systems, A. Abraham et al*. (eds.), IOS Press, 631-640.

Misra, A.K., Prakash, O. and Ramasubramanian V. (2004). Forewarning powdery mildew caused by oidium mangiferae in mango *(Mangifera Indica)* using logistic regression models. *Ind. J. Agric. Sci.,* **74(2)**, 84-87.

Mitra, S., De, R.K. and Pal, S.K. ( 1997). Knowledge-based fuzzy MLP for classification and rule generation. *IEEE Trans. Neural Networks,* **8(6)**, 1338-1350.

Pawlak, Z. (1991). *Rough Sets-Theoretical Aspects of Reasoning about Data*. Kluwer Academic Publishers, Dordecht.

Pujari, A.K. (2000). *Data Mining Techniques.* Universities Press.

Quinlan, J.R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kauffman.

Rosetta, Rough set toolkit for analysis of data available at http://www.idi.ntnu.no/~aleks/rosetta/

Winston, P.H. (1992). *Artificial Intelligence.* Addison-Wesley.

Witten, I.H. and Frank E. (1999). *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations.* Morgan Kaufmann Publishers.