FEATURE SELECTION BY ATTRIBUTE CLUSTERING OF INFECTED RICE PLANT IMAGES

SANTANU PHADIKAR1, JAYA SIL2, ASIT KUMAR DAS3
1Department of Computer Science and Engineering, West Bengal University of Technology, Salt Lake, Kolkata, India
2Department of Computer Science and Technology, Bengal Engineering and Science University, Shibpur, Howrah, India
3Department of Computer Science and Technology, Bengal Engineering and Science University, Shibpur, Howrah, India
*Corresponding Author: Email- sphadikar@yahoo.com

Received: August 10, 2011; Accepted: September 05, 2011

Abstract- Automatic diagnosis of rice plant diseases at an early stage and taking corrective measures in time saves damages of rice crop across the world. The paper aims at developing an appropriate methodology to classify diseases with the help of feature sets obtained by analyzing images of infected rice plants acquired from the field. Since all features are not important in classifying diseases; selection of optimum features is a challenging task to address the problem. The work is performed in three steps. Firstly thirty six features of different category are extracted from the diseased plant images using image processing techniques. Secondly information gain (IG) of each attribute with respect to other attributes is calculated following the concept of information entropy theory. Thirdly using IG, functional dependency of the attributes are evaluated based on which fourteen significant attributes out of thirty six are selected, sufficient to classify the diseases. The proposed method has been applied on four hundred fifty infected rice plant images having three different classes. With the reduced feature set, classification accuracy is calculated using different classifiers demonstrating effectiveness of the proposed model.

Key words - Information Entropy Theory, Rice Diseases, Feature selection, Attribute Clustering, Reduct Generation.

INTRODUCTION

Rice is one of the most widely cultivated food crops throughout the world. Damages due to various reasons affect productivity of rice, which can be arrested to some extent by automatic diagnosis of the diseases at an early stage. Rice ‘blast’ disease caused by the fungus Pricularia Grisea [1-3], occurs in most of the rice fields across the globe. The damages caused by ‘blast’, depend on the degree of severity of the disease. Another critical rice disease, ‘leaf brown spot’ caused by the fungus Bypolaris Orzya [4-6] is viable throughout the rice growing season. ‘Sheath rot’ disease, caused by the fungus Sarocladium Orzya [2, 4, 7] usually occurs on the flag leaf sheath (boot) that encloses the panicle. The lesions first appear as oblong or irregular spots about 3/16 to 5/8 of an inch long with a gray center and a reddish-brown margin. Abundant white powdery growth of the fungus is later observed inside the affected leaf sheaths and on the surface of rotted panicles. Panicles of sheaths affected before emergence rot, turn brown or reddish brown and fail to produce any grain.

With the advancement of information technology, remote sensing techniques have been used in the field of crop management, described in [8-10]. A relation among the ground disease index and remote sensing data is established in [11] to classify the diseases. Very recently, Data mining techniques [12-14] are used to discover classification rules of rice diseases and image processing and soft computing techniques are applied to automatic diagnose the field problems, as reported in [15-16]. Studies in the field reveal that accurate diagnosis depends on the visual properties of the plants such as change of colour, shape, orientation (textures) of the infected portion of the images. However, handling large number of features increase complexity of the system and unimportant features may lead to improper classification of diseases. One of the most important problems of the automatic diagnosis process is to identify the significant information from large volume of data using appropriate data mining techniques [17-19]. Therefore, feature selection [20-23] has become an important pre-processing step to reduce complexity in building an efficient classifier [24, 25] for diagnosing the diseases.

The goal of feature selection is to avoid selecting too many or too few features than is necessary. If too few features are selected, there is a high chance that the information content in this set of features is low. On the other hand, if too many (irrelevant) features are selected, the effect of noise may overshadow the information present. Hence, a trade-off is essential that must be addressed by feature selection method. Rough set based reduct [25-27] generation methods, statistical methods [28-31] and correlation-based methods [32] contributed towards developing improved dimensionality reduction techniques. Statistical methods are generally lower in computational complexity compared to the reducts and the correlation-based methods. However, reduct generation methods are significant in reducing attributes
without lose of important information, therefore, producing better classification accuracy compare to other methods. In the proposed method, different diseased rice plant images acquired from the paddy field are used as training dataset to design the classifier. Various types of image features are extracted [33-36] using image processing techniques, which are categorized based on colours, shapes and texture. Change of colour, deviation from the actual shape and non-uniformity of the infected leaf provide important information to diagnose the diseases. However, information contained in the features is not all important. In the work following steps are executed to select features to design the classifier more accurately.

(i) Information gain (IG) [12] of each attribute with respect to others is calculated based on the concept of information entropy, (ii) using IG, the IG table is formed that expresses dependency relationship between the attributes, (iii) the size of the IG table is reduced by removing the elements which don't have significant influence to classify the objects (images of infected rice plants), (iv) functional dependencies [37] of the attributes are evaluated using IG values, (v) based on functional dependency of attributes, a dependency graph is constructed [38] whose vertices and edges represent attributes and degree (in-degree / out-degree) of dependency among the attributes, respectively, (vi) the attributes are clustered [39-40] depending on the in-degree / out-degree values and elements (attributes) of each cluster are sorted in ascending / descending order, (vii) score representing significance of an attribute is calculated giving equal importance of its presence both in-degree and out-degree cluster and finally (viii) the attributes are partitioned based on their score and a single attribute from each partition is taken to generate a single attribute set consisting of optimal number of attributes of the system.

In the experiment, dataset with thirty six features are prepared from the collection of four hundred fifty infected rice plant images. The proposed method reduces number of features to fourteen and used for building the classifiers. Ten-fold cross validations are carried out to compute accuracy of various classifiers. Result shows that important information about the infected leaf is retained that generates accurate and complete classifier able to diagnose the diseases.

The paper is organized as follows: Feature extraction procedures are discussed in first section. Next section describes the single reduct generation method using information entropy and functional dependency for feature selection. Experimental procedure and result obtained from the rice plant dataset are discussed in next section and finally, conclusions are summarized in last section.

**FEATURE EXTRACTION**

From the studies of the diseased rice plant images, it has been noticed that colour change of the infected region, shape of the spot created by the diseases and orientation (texture) of the shapes are the most important visual properties to identify the diseases. The attributes or properties are first categorized into three sub domains namely colour, shape and texture. For each sub domain, different attribute (features) values are extracted using spatial information of the image. Leave of the rice plants are infected by ‘blast’ and ‘brown spot’ diseases while stem by ‘sheath rot’ disease, described in Fig. (1).

**Colour Feature Extraction**

In order to extract features following important observations of the field experts are presented here that helps the feature extraction procedure: (i) ‘brown spots’ are dull yellow margin and dark brown center, (ii) lesions of blast create spots with a gray or white center surrounded by a reddish brown border and spots with gray center and (iii) a reddish-brown margin are created by ‘sheath rot’.

The images are first segmented using Otsu’s threshold based method [41] and then complemented to identify the background (BC) of infected region as shown in Fig. (2). Now to separate the core (CR) and border (BR) regions of the infected images, second level segmentation is performed, respective results are shown in Fig. (3) and Fig. (4).

Colour features are obtained by calculating mean (M) and standard deviation (SD) of the intensity of pixels creating spots in three classical planes; red (R), green (G) and blue (B) of the segmented images. All 36 extracted features are listed in Table 1 using their abbreviated name. For example, BC_M_R, and BC_SD_R represent mean and standard deviation of the spot in the background region by considering red colour plane. Similarly for border and core regions, the feature values are extracted by considering green and blue colour planes.

**Shape Based Feature Extraction**

Area (AR), Sharpness (SH), Area-discrepancy (AD), Perimeter (PR), Eigen values (EV) and Aspect-ratio (ASR) are computed as shape based features to detecting the diseases. Here, area is determined by counting the number of pixels in the infected region while perimeter is obtained by counting the number of pixels in the boundary of the spot. These features are able to detect the variation of the shape of the spots from standard elliptical shape. Sharpness of the shape of the spot is determined by calculating the average distance between two boundary points along the major axis, labeled as T in Fig. (5). Aspect ratio is the ratio of the major and minor axis of the ellipse that provides information regarding stretch of the spot either horizontally or vertically. The eigen value [34-35] represents valuable information, about the image of the infected region. The eigen values of the Dirichlet Laplacian [35] are preserved if the underlying domain is translated or rotated. These properties make eigen values as very useful features in recognizing shapes of different sizes and orientations.

Various central moments [42] φ₁ to φ₇, invariant to scaling, translation and rotation are extracted from the infected images as shape features, as described in Table 1.
1. An image moment is usually chosen to depict the global properties of the image and computed using weighted average (moment) of the intensity of pixels. Moments of all orders i.e. a complete moment set can be computed and used uniquely to describe the information contained in the image. A simple moment of a region of degree $p + q$ is defined by equation (1), where, $p$ and $q$ are integers vary from 0 to 3.

$$m_{pq} = \sum x^p y^q f(x, y)$$

In equation(1) summation is taken over all points in the region, assuming uniform gray value in the infected region, and $f(x, y)$ represents the brightness at a particular point $(x, y)$. In two dimensional space, the coordinates of the centre of mass are $m_x$ and $m_y$ calculated by equation (2), define a unique location of the image $f(x, y)$. It can be used as a reference point to describe the position of the image and also known as center of gravity.

$$m_x = \frac{m_{x0}}{m_{00}} ; m_y = \frac{m_{y0}}{m_{00}}$$

Invariance to translation is achieved by referencing all points to the center of gravity, producing the “central moments,” as described in equation (3).

$$\mu_{pq} = \sum (x - m_x)^p (y - m_y)^q f(x, y)$$

The normalized central moments is defined as $\eta_{pq} = \frac{\mu_{pq}}{\mu_{00}}$, where, $\gamma = \frac{p+q}{2} + 1$. Finally, the invariant moments ($\phi_1$ to $\phi_7$) are computed using formulae given in Table 2.

**Texture Feature Extraction**

Orientation of the shape is represented by the texture. Different texture features extracted from the diseased images are energy (EG), entropy (ET), contrast (CT), homogeneity (HG) and co-relation (CR), as mentioned in Table 1. The basic assumption of selecting EG as a feature is based on the concept that the energy distribution in frequency domain able to identify a texture. Besides providing acceptable retrieval performance from large texture database, EG based approaches are partly supported by physiological studies of the visual cortex. Another feature ET is a statistical measure of randomness and invariant to scaling, translation and rotation, used to characterize the texture of the image. It does not depend on the actual value of the gray level but only on the probabilities of gray level distribution. Local variations present in an image are measured by texture feature CT that helps to distinguish objects by their colour and brightness within the same field of view. In general, HG is defined as the quality or state of being homogeneous, used to evaluate the intensity uniformity of a local region based on high-pass operators as texture. CR measures the pixel linear dependencies of neighboring pixels, based on which uniformity in neighboring regions of image is determined.

To obtain these features, colour spot images are converted to gray scale images and the co-occurrence matrix $C$ is calculated by equation (4). Using the co-occurrence matrix the texture features [36, 43] are calculated, as described in Table 3.

$$c_{\Delta x \Delta y}(i, j) = \sum_{p=1}^{n} \sum_{q=1}^{m} t$$

where,

$$t = \begin{cases} 1, & \text{if } f(p, q) = i \text{ and } f(p + \Delta x, q + \Delta y) = j \\ 0, & \text{otherwise} \end{cases}$$

Where $f$ is intensity of the image of size $m \times n$ and $(\Delta x, \Delta y)$ is the offset, considered each as one.

**SINGLE REDUCT GENERATION**

Once the features are extracted, the decision table is constructed with 36 features, 450 infected rice plant images and three diseases as described in “feature extraction” section. The proposed method redefines the decision table by selecting only the relevant features, required for disease classification of infected rice plants without compromising its accuracy. The set of relevant features, called reduct [26-27, 44] used to determine the optimal set of features based on the conditional entropy and functional dependency of the attributes.

**Conditional Entropy and Information Gain**

Information gain is the concept applied for reducing/removing uncertainty or randomness in classifying objects with respect to some given features. Uncertainty is measured using information entropy that quantifies the expected value of information. Uncertainty relative to the given feature value is called conditional entropy contained in the features of objects. The entropy $H(A)$ of an attribute A is defined in equation (5) and conditional entropy $H(C_B^A)$ is referred as the entropy of A with observation of attribute B, given in equation (6).

$$H(A) = \sum_i p(A_i) \log p(A_i)$$

$$H(C_B^A) = \sum_j p(B_j) \sum_i p(A_i | B_j) \log p(A_i | B_j)$$

Where $p(A_i)$ is the prior probability of $i^{th}$ value of $A$, $P(C_B^A)$ is the post prior probability of $A_i$ for given $B_j, j = 1, 2, .., N$. The information gain $IG(C_B^A)$ of an attribute $A$ with respect to another attribute $B$ measures the reduction in uncertainty about the value of $B$ when the value of $A$ is known, defined as the difference between the entropy and conditional entropy values, given in equation (7).

$$IG(C_B^A) = H(A) - H(C_B^A)$$

The information gain of an attribute $A$ with respect to attribute $B$ (i.e., $IG(C_B^A)$) is nothing but mutual information of $A$ with respect to $B$. According to this measure, an attribute $B$ is regarded to be more correlated to attribute $A$.
The algorithm to computing information gain consists of two procedures namely, "Individual_Entropy_Computation()" and "Conditional_Entropy_Information_Gain()".

Algorithm: Individual_Entropy_computation (DS)
Begin
Input: DS, the decision system.
Output: Individual entropy; \( H(A) \) of the attribute \( A \)
For \( I = 1 \) to \( N \) do
{ /*computation of the distinct values and their frequencies for each attribute */
Let \( M = \) No. of distinct \( A_I \) values
Let \( DA_I = \{ DA_{I1}, DA_{I2}, \ldots \ldots \ldots \ldots \ldots DA_{IM} \} \)
INA = \( M; /* store the index value */
For \( J = 1 \) to \( M \) do
FA\(_{IJ}\) = Frequency of \( DA_{IJ} \).
/*Compute individual entropy of each attribute using eq. (5) */
\( H[A_I] = 0; /* initialize the entropy */
For \( J = 1 \) to \( M \) do
\( H[A_I] = H[A_I] - \frac{FA_{IJ}}{|U|} \times \log_2 \left( \frac{FA_{IJ}}{|U|} \right) \)
}
End.

Algorithm: Conditional_Entropy_Information_Gain (DS, H)
Begin
Input: Decision system DS and individual entropy H.
Output: Information gain of the system.
For \( I = 1 \) to \( N \) do /*as there are \( N \) attributes in dataset*/
{ /*compute conditional entropy */
For \( J = 1 \) to \( N \) do
{ /*initialize the conditional entropy */
\( H[\frac{DA_{IJ}}{DA_{IK}}] = 0; /* compute conditional entropy by eq. (6) */
If \( (I != J) \) then
{ /*sum(l) = sum(l) + \( H[\frac{DA_{IJ}}{DA_{IK}}] \) */
sum(l) = sum(l) + \( H[\frac{DA_{IJ}}{DA_{IK}}] \)
}
}
For \( L = 1 \) to \( |DA_I| \) do /*initialize the total conditional entropy against each distinct value of attribute */
\( T = 0; /* clear the count */
For \( L = 1 \) to \( |DA_I| \) do /*end of L loop */
\( T = T - \frac{1}{\sum(l)} \times \log_2 \left( \frac{DA_{IL}}{|U|} \right) \times \sum(l) \)
\( H[\frac{DA_{IJ}}{DA_{IK}}] = T \times \log_2 \left( \frac{DA_{IL}}{|U|} \right) \)
}
/* end of K loop */
}
/* end of J loop */
/* end of I loop */
End.

Functional Dependency of Attributes
To know the mutual information gain among attributes, equation (7) is applied for each pair of distinct attributes in the system. Thus an \( N \times N \) information gain table \( IG[\times N] \) is obtained, where the first row indicates the information gain of first attribute with respect to all attributes and so on. Each entry \( IG[\times]\) in the table represents the information gain value \( IG \left( \frac{A_J \rightarrow A_i}{A_I} \right) \) obtained using equation(7).

Then, for a given attribute, say in column \( j \) of IG table, average information gain is calculated and the attribute, say in row \( i \), having greater information gain than average depends on attribute \( j \). A functional dependency (FD) of the attributes \( j \rightarrow i \) is established and thus all \( N \) rows are checked for the dependency and set of attributes depends on attribute \( j \) are obtained. Repeating the process for \( j = 1, 2, \ldots, N \), all possible mutual dependencies of the attributes are determined. Then from the functional dependencies, a dependency graph \( DG = (V, E) \) is obtained, where a directed edge \( V_j \rightarrow V_i \) corresponds to functional dependency \( j \rightarrow i \).

For each vertex of the graph, in-degree and out-degree parameters are evaluated where, in-degree is the number of edges incident to the vertex and out-degree is the number of edges leaves the vertex. So higher the out-degree of a vertex implies more attributes are dependent on the attribute, mapped at that vertex and so it is considered as a valuable attribute of the system. Similarly, lower the in-degree of a vertex implies it is dependent on less number of attributes, mapped at the vertices and so considered as a valuable attribute of the system. Therefore, a higher out-degree and lower in-degree is expected for an attribute in a decision system. The functional dependencies and in-degree / out-degree of the attributes are evaluated, as described by the following algorithms.

Algorithm: Functional_Dependency_of_Attributes (DS, IG)
Begin
\( u = 1 /*Compute functional dependencies of attributes */
For \( J = 1 \) to \( N \) do
{ /* counting frequency of distinct values in condition */
\( \text{sum}(l) = \text{sum}(l) + H[\frac{DA_{IJ}}{DA_{IK}}] \)
\( T = 0; /* initialize the total conditional entropy against each distinct value of attribute */
For \( L = 1 \) to \( |DA_I| \) do /*end of L loop */
\( T = T - \frac{1}{\sum(l)} \times \log_2 \left( \frac{DA_{IL}}{|U|} \right) \times \sum(l) \)
\( H[\frac{DA_{IJ}}{DA_{IK}}] = T \times \log_2 \left( \frac{DA_{IL}}{|U|} \right) \)
}
/* end of K loop */
/* end of J loop */
/* end of I loop */
End.
Algorithm: Degree_of_Dependency (DS, FD)
Begin
  /*Compute in-degree, out-degree of attributes in DS from FD and store in first and second columns array deg[ ][ ] respectively*/
  u = u - 1; /* u is the number of Functional dependencies*/
  For I = 1 to N do
    /* compute in and out degree for each attribute*/
    deg[I][1] = deg[I][2] = 0;
    For J = 1 to u - 1 do
      /* Jth loop compute the in and out degree of ith attribute*/
      If (FD[J][1] = = I) then /* out-degree*/
        deg[I][1]++;
      If (FD[J][2] = = I) then /* in-degree*/
        deg[I][2]++;
    /*end of J loop*/
  /*end of I loop*/
End

Algorithm: Partition_based_on_Out_Degree(deg)
Begin
  /* partition into groups w.r.t. out-degree */
  CLUS_out_degree = Ø /* it is a 2-D array, each row contains one group, initially all empty*/
  row = 1;
  While(1)
    /* select maximum out-degree*/
    max = deg[1][1];
    For I = 2 to N do
      If (max < deg[I][1]) then
        max = deg[I][1];
      If (max = = -1) then
        break; /* partitioning done, so go out of while loop */
    /* end of while loop */
  /* this loop compute one group of the partition*/
    For I = 1 to N do /* this loop compute one group of the partition*/
      If (deg[I][1] = = max) then
        deg[I][1] = -1;
        CLUS_out_degree[row] = CLUS_out_degree[row] ϵ {Ai}
        row = row + 1;
    /* end of while loop */
  /* number of clusters in CLUS_out_degree */
  No_out_grp = row - 1;
End.

Algorithm: Partition_based_on_In_Degree(deg)
Begin
  /* partition into groups w.r.t. in-degree */
  CLUS_in_degree = Ø /* it is a 2-D array, each row contains one group, initially all empty*/
  row = 1;
  While(1)
    min = First non-negative value in deg[ ][2] /* select minimum in-degree*/
    For I = 1 to N do
      If ((min > deg[I][2]) && (deg[I][2] >0))
        min = deg[I][2];
    break; /* partitioning done, so go out of while loop */
  /* compute one group of the partition*/
    For I = 1 to N do /* compute one group of the partition*/
      If (deg[I][2] = = min) then
        deg[I][2] = -1;
        CLUS_in_degree[row] = CLUS_in_degree[row] ϵ {Ai}
        row = row + 1;
    /* end of while loop */
  /* no of clusters in CLUS_in_degree */
  No_in_grp = row - 1;
End.

Finally, a single partition of attributes is obtained from the clusters of attributes having similar in-degree and out-
Algorithm: Partition_based_on_Attribute_Score(Score)
Begin
/* Computation of score for each attribute*/
No_out_grp = |CLUS_{out-degree}|
No_in_grp = |CLUS_{in-degree}|
For I = 1 to N do
{
  If (A \in CLUS_{out-degree}[I]) then
    rank_out = I;
    break;
  
  If (A \in CLUS_{in-degree}[I]) then
    rank_in = I;
    break;
}

Score[I] = \frac{\text{rank}_in + \text{rank}_out}{2} \quad (10)
End

Algorithm: Score_of_Attributes(CLUS_{out-degree}, CLUS_{in-degree})
Begin
/* Generation of reduced set of attributes*/
RED = \emptyset
For I = 1 to N do
  If (\{A \in CLUS_{out-degree}[I]\}) then
    RED = RED \cup \{A\};
  
  If (\{A \in CLUS_{in-degree}[I]\}) then
    RED = RED \cup \{A\};
End

RESULT AND DISCUSSIONS
The proposed method is applied on a dataset generated from 450 infected rice plant images of three diseased classes (brown spot, blast and sheath rot). The dataset
contains 36 features and a decision attribute with 3 different class labels. Sample datasets for different kinds of features, calculated using the methodologies described in “feature extraction” section, and given in Table 4, Table 5 and Table 6 that contain extracted colour features, shape-based features and texture features respectively. All the numeric attributes are discretized by ChiMerge [45] discretization algorithm and is described graphically in Fig. (6).

For each pair of 36 features, information gain is calculated as shown in Fig. (7). Functional dependencies of the attributes are shown as dependency graph, depicted in Fig. (8).

Now, from the dependency relationship, in-degree and out-degree of each attribute (i.e., vertex of the graph) are calculated. Then the attributes are partitioned based on their in-degree/out-degree values and are stored in CLUS-in-degree and CLUS-out-degree as shown in Table 7 and Table 8 respectively.

The score of each attribute is calculated using Table 7 and Table 8 while new cluster CLUS is formed based on their score, as listed in Table 9, such that score of any two attributes are same in a cluster and different in different clusters. For example, score of attribute BR_M_G is (6+1)/2 = 3.5 as it is in 6th cluster of Table 7 and in 1st cluster of Table 8. Similarly, score of all attributes are computed, where BR_M_G is of lowest score. Repeat the overall process on each element in a cluster and consider the feature from each cluster with minimum score, in case of multiple minimum score, arbitrarily one attribute is selected. For example, in case of cluster 3 in Table 9, repeat the process with feature set {BC_M_R, CR_M_G, EV, φ} and ultimately attribute EV is obtained with minimum score, shown in third column of cluster 3. Finally, combining all 14 features with minimum score, a single reduct RED = {BR_M_G, φ, EV, CT, BR_SD_G, CR_SD_R, ϕ, ET, SH, AD, AR, BC_SD_G, BC_SD_B, BR_SD_B} is obtained.

The well known dimensionality reduction method, of attributes from 36 to 18, “Cfs Subset Eval” method Principal Component Analysis (PCA) [46] reduces number with Genetic Search [47] (CFS) selects 19 attributes and “Consistency Subset Eval” with Rank Search(CON) [48] method finds 20 attributes out of thirty-six extracted features of the disease images. So the rate of dimensionality reduction is higher for the proposed method compare to the existing methods, shown in Fig. (9). The method does not reduce dimension of data by losing its decision making capability, rather it provides compatible classification accuracy obtained by various classifiers when run using “weka” tool [49] where 10-fold cross-validations are carried out, as listed in Table 10. In Table 10, other dimension reduction methods like “Chi-Squared Attribute Eval”(CHI), “Classifier Subset Eval”(CLS)[50] and “Support Vector Machine Attribute Eval”(SVM)[51] are used where first fourteen ranked attributes are considered for classification, as the proposed method selects only fourteen attributes. The accuracy of classifiers show that the proposed method is at least comparable with other dimensionality reduction methods like, PCA, CFS, CON, SVM, CLS and so on.

CONCLUSION

In the paper, functional dependencies of the attributes represent the dependency graph for the attribute set. From the dependency set in-degree and out-degree of the vertices (i.e., attributes) are measured which finally helps to compute the score of the attributes. Then, attributes are partitioned according to their scores and reduct is generated. The results show significant efficiency of the proposed method. Moreover, the proposed method is envisaged on the concept of information gain, which is an established theory of measuring uncertainty and quantified the information contained in the system.

References


Bioinfo Publications http://www.bioinfo.in
Feature selection by attribute clustering of infected rice plant images

Table 1 - Description of Features in the Dataset

<table>
<thead>
<tr>
<th>Feature Category</th>
<th>Feature Names</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shape Feature</td>
<td>AR, SH, AD, PR, EV, ASR, φ1, φ2, φ3, φ4, φ5, φ6, φ7</td>
</tr>
<tr>
<td>Texture Features</td>
<td>EG, ET, CT, HG, CR</td>
</tr>
</tbody>
</table>

Table 2 - Formulae for Normalized Central Moments

<table>
<thead>
<tr>
<th>Central Moments</th>
<th>Formulae</th>
</tr>
</thead>
<tbody>
<tr>
<td>φ1</td>
<td>η20 + η02</td>
</tr>
<tr>
<td>φ2</td>
<td>(η20 + η02)2 + 4η02</td>
</tr>
<tr>
<td>φ3</td>
<td>(η30 + 3η12)2 + (η30 - η03)2</td>
</tr>
<tr>
<td>φ4</td>
<td>(η30 + η12)2 + (η21 + η03)2</td>
</tr>
<tr>
<td>φ5</td>
<td>(η30 + η12)2 (η30 + η12)2 - 3(η21 + η03)2 + 3(η30 + η12)2</td>
</tr>
<tr>
<td>φ6</td>
<td>(η20 + η02)2 (η30 + η12)2 - η21 + η03)2 + 4η11(η30 + η12)2</td>
</tr>
<tr>
<td>φ7</td>
<td>(η30 - η03)2 (η30 + η12)2 - 3(η21 + η03)2 - η30 + η12)2</td>
</tr>
</tbody>
</table>

Table 3 - Generation of Texture Features:

<table>
<thead>
<tr>
<th>Texture Feature</th>
<th>Formulae</th>
</tr>
</thead>
<tbody>
<tr>
<td>Energy</td>
<td>( \sum \sum C^2(i,j) )</td>
</tr>
<tr>
<td>Entropy</td>
<td>( \sum \sum C(i,j) \log C(i,j) )</td>
</tr>
<tr>
<td>Contrast</td>
<td>( \sum \sum (i - j)^2 C(i,j) )</td>
</tr>
<tr>
<td>Homogeneity</td>
<td>( \sum \sum C(i,j)/(1 +</td>
</tr>
<tr>
<td>Correlation</td>
<td>( \sum \sum ((i - \mu_x)(j - \mu_y))/\sigma_x \sigma_y )</td>
</tr>
</tbody>
</table>
### Table-4 - Sample Colour Features of the Images:

<table>
<thead>
<tr>
<th>BC_M_R</th>
<th>BC_SD_R</th>
<th>BC_SD_B</th>
<th>BR_M_R</th>
<th>...</th>
<th>CR_M_G</th>
<th>CR_SD_G</th>
<th>CR_M_B</th>
<th>CR_SD_B</th>
</tr>
</thead>
<tbody>
<tr>
<td>69.04</td>
<td>8.3035</td>
<td>5.2566</td>
<td>71.181</td>
<td>...</td>
<td>31.541</td>
<td>9.0277</td>
<td>17.609</td>
<td>7.4044</td>
</tr>
<tr>
<td>125.91</td>
<td>10.909</td>
<td>9.8703</td>
<td>161.25</td>
<td>...</td>
<td>175.15</td>
<td>12.535</td>
<td>156.44</td>
<td>11.673</td>
</tr>
<tr>
<td>143.69</td>
<td>10.215</td>
<td>10.244</td>
<td>179.97</td>
<td>...</td>
<td>192.82</td>
<td>13.986</td>
<td>178.53</td>
<td>12.822</td>
</tr>
<tr>
<td>134.12</td>
<td>8.7062</td>
<td>9.1771</td>
<td>137.92</td>
<td>...</td>
<td>142.34</td>
<td>12.479</td>
<td>121.37</td>
<td>12.208</td>
</tr>
<tr>
<td>58.43</td>
<td>8.3836</td>
<td>5.9226</td>
<td>71.966</td>
<td>...</td>
<td>50.534</td>
<td>10.374</td>
<td>23.477</td>
<td>8.6921</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>127.49</td>
<td>30.278</td>
<td>14.054</td>
<td>71.881</td>
<td>...</td>
<td>17.268</td>
<td>37.245</td>
<td>32.074</td>
<td>20.081</td>
</tr>
</tbody>
</table>

### Table-5 - Sample Shape Features of the Images:

<table>
<thead>
<tr>
<th>AR</th>
<th>SH</th>
<th>AD</th>
<th>...</th>
<th>ASR</th>
<th>...</th>
<th>$\phi_1$</th>
<th>...</th>
<th>$\phi_4$</th>
<th>$\phi_5$</th>
<th>$\phi_7$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.61</td>
<td>0.059</td>
<td>0.758</td>
<td>...</td>
<td>0.897</td>
<td>0.003179</td>
<td>1.10E-10</td>
<td>2.82E-21</td>
<td>2.01E-13</td>
<td>6.95E-21</td>
<td></td>
</tr>
<tr>
<td>3.99</td>
<td>0.043</td>
<td>0.701</td>
<td>...</td>
<td>0.232</td>
<td>0.002022</td>
<td>7.16E-10</td>
<td>5.67E-19</td>
<td>1.26E-12</td>
<td>2.11E-20</td>
<td></td>
</tr>
<tr>
<td>1.84</td>
<td>0.083</td>
<td>0.648</td>
<td>...</td>
<td>0.154</td>
<td>0.000992</td>
<td>3.68E-12</td>
<td>7.05E-24</td>
<td>4.03E-16</td>
<td>6.21E-23</td>
<td></td>
</tr>
<tr>
<td>1.37</td>
<td>0.076</td>
<td>0.782</td>
<td>...</td>
<td>0.108</td>
<td>0.001008</td>
<td>1.93E-13</td>
<td>6.98E-26</td>
<td>6.53E-17</td>
<td>1.49E-23</td>
<td></td>
</tr>
<tr>
<td>3.10</td>
<td>0.039</td>
<td>0.710</td>
<td>...</td>
<td>0.183</td>
<td>0.004395</td>
<td>4.74E-10</td>
<td>7.43E-19</td>
<td>1.45E-12</td>
<td>6.39E-19</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td></td>
</tr>
<tr>
<td>4.04</td>
<td>0.201</td>
<td>0.725</td>
<td>...</td>
<td>0.111</td>
<td>0.003629</td>
<td>3.58E-10</td>
<td>1.23E-19</td>
<td>1.11E-12</td>
<td>1.13E-20</td>
<td></td>
</tr>
<tr>
<td>1.49</td>
<td>0.160</td>
<td>0.872</td>
<td>...</td>
<td>0.885</td>
<td>0.001899</td>
<td>3.81E-11</td>
<td>1.55E-21</td>
<td>1.16E-14</td>
<td>1.15E-22</td>
<td></td>
</tr>
</tbody>
</table>

### Table-6 - Sample Texture Features of the Images

<table>
<thead>
<tr>
<th>Energy (EG)</th>
<th>Entropy (ET)</th>
<th>Contrast (CT)</th>
<th>Homogeneity (HG)</th>
<th>Correlation (CR)</th>
<th>$\phi_1$</th>
<th>$\phi_4$</th>
<th>$\phi_5$</th>
<th>$\phi_7$</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.088483</td>
<td>23.927</td>
<td>28.936</td>
<td>1.8458</td>
<td>0.000228</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.056449</td>
<td>20.47</td>
<td>159.04</td>
<td>1.7422</td>
<td>0.011142</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.182287</td>
<td>20.956</td>
<td>216.75</td>
<td>1.8249</td>
<td>0.017118</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.605002</td>
<td>20.634</td>
<td>149.28</td>
<td>1.7773</td>
<td>0.017302</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.447494</td>
<td>22.079</td>
<td>21.57</td>
<td>1.8021</td>
<td>0.000822</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.598053</td>
<td>17.595</td>
<td>79.094</td>
<td>1.8049</td>
<td>0.17761</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.397743</td>
<td>17.144</td>
<td>120.6</td>
<td>1.7677</td>
<td>0.2006</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.598453</td>
<td>16.742</td>
<td>59.271</td>
<td>1.7986</td>
<td>0.23098</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Table-7 - Clustering Based on In-degree:

<table>
<thead>
<tr>
<th>Cluster # in CLUS\text{in-degree}</th>
<th>Features in Cluster #</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(ET, HG)</td>
</tr>
<tr>
<td>2</td>
<td>(SH, AR)</td>
</tr>
<tr>
<td>3</td>
<td>(EG)</td>
</tr>
<tr>
<td>4</td>
<td>(AD, $\phi_1$)</td>
</tr>
<tr>
<td>5</td>
<td>(CR_M_B, PR, $\phi_1$, $\phi_3$, $\phi_5$)</td>
</tr>
<tr>
<td>6</td>
<td>(BC_M_R, BC_M_B, BR_M_G, BR_M_B, CR_M_G, CR_SD_B, ASR, EV, $\phi_2$, $\phi_4$)</td>
</tr>
</tbody>
</table>
Feature selection by attribute clustering of infected rice plant images

Table 8 - Clustering Based on Out-degree:

<table>
<thead>
<tr>
<th>Cluster # in CLUS</th>
<th>Features in Cluster #</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>{BC_M_B, φ₂}</td>
</tr>
<tr>
<td>3</td>
<td>{BR_SD_R, BR_SD_G, CR_M_B}</td>
</tr>
<tr>
<td>4</td>
<td>{CR_SD_R}</td>
</tr>
<tr>
<td>5</td>
<td>{φ₃}</td>
</tr>
<tr>
<td>6</td>
<td>{CR_SD_G}</td>
</tr>
<tr>
<td>7</td>
<td>{BR_M_B}</td>
</tr>
<tr>
<td>8</td>
<td>{φ₇}</td>
</tr>
<tr>
<td>9</td>
<td>{BC_SD_R}</td>
</tr>
<tr>
<td>10</td>
<td>{BC_SD_G}</td>
</tr>
<tr>
<td>11</td>
<td>{BC_SD_B, CR_SD_B, AR, PR, EG}</td>
</tr>
<tr>
<td>12</td>
<td>{BR_SD_B, SH, AD, ASR, ET, HG}</td>
</tr>
</tbody>
</table>

Table 9 - Clustering of Features Based on Scores and Final Reduct:

<table>
<thead>
<tr>
<th>Cluster # in CLUS</th>
<th>Features in Cluster # based on score</th>
<th>Selected Attribute</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>{BR_M_G}</td>
<td>BR_M_G</td>
</tr>
<tr>
<td>2</td>
<td>{φ₁, φ₂}</td>
<td>φ₁</td>
</tr>
<tr>
<td>3</td>
<td>{BC_M_R, CR_M_G, EV, φ₆}</td>
<td>EV</td>
</tr>
<tr>
<td>4</td>
<td>{BC_M_G, BC_M_B, BR_M_R, CR_M_R, CR_M_B, φ₂, CT, CR}</td>
<td>CT</td>
</tr>
<tr>
<td>5</td>
<td>{BR_SD_R, BR_SD_G, φ₃}</td>
<td>BR_SD_G</td>
</tr>
<tr>
<td>6</td>
<td>{CR_SD_R}</td>
<td>CR_SD_R</td>
</tr>
<tr>
<td>7</td>
<td>{φ₇}</td>
<td>φ₇</td>
</tr>
<tr>
<td>8</td>
<td>{BR_M_B, ET, HG}</td>
<td>ET</td>
</tr>
<tr>
<td>9</td>
<td>{CR_SD_G, SH, ASR, EG}</td>
<td>SH</td>
</tr>
<tr>
<td>10</td>
<td>{BC_SD_R, AD, PR, φ₆}</td>
<td>AD</td>
</tr>
<tr>
<td>11</td>
<td>{CR_SD_B, AR}</td>
<td>AR</td>
</tr>
<tr>
<td>12</td>
<td>{BC_SD_G}</td>
<td>BC_SD_G</td>
</tr>
<tr>
<td>13</td>
<td>{BC_SD_B}</td>
<td>BC_SD_B</td>
</tr>
<tr>
<td>14</td>
<td>{BR_SD_B}</td>
<td>BR_SD_B</td>
</tr>
<tr>
<td>Final Reduct (RED)</td>
<td>{BR_M_G, φ₁, EV, CT, BR_SD_G, CR_SD_R, φ₇, ET, SH, AD, AR, BC_SD_G, BC_D_B, BR_SD_B}</td>
<td></td>
</tr>
</tbody>
</table>

Table 10 - Accuracy of Different Classifier for Reduced Dataset

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Proposed Method</th>
<th>PCA</th>
<th>CHI</th>
<th>CLS</th>
<th>SVM</th>
<th>CFS</th>
<th>CON</th>
</tr>
</thead>
<tbody>
<tr>
<td>C4.5</td>
<td>86.21</td>
<td>83.79</td>
<td>84.6</td>
<td>84.85</td>
<td>84.34</td>
<td>84.85</td>
<td>84.85</td>
</tr>
<tr>
<td>PART</td>
<td>90.3</td>
<td>87.77</td>
<td>84.85</td>
<td>84.45</td>
<td>84.34</td>
<td>86.36</td>
<td>87.12</td>
</tr>
<tr>
<td>K-STAR</td>
<td>88.19</td>
<td>88.05</td>
<td>91.16</td>
<td>89.65</td>
<td>89.9</td>
<td>86.87</td>
<td>88.89</td>
</tr>
<tr>
<td>NaïveBaye's</td>
<td>85.68</td>
<td>83.74</td>
<td>79.55</td>
<td>80.3</td>
<td>85.61</td>
<td>84.09</td>
<td>80.05</td>
</tr>
<tr>
<td>SMO</td>
<td>90.28</td>
<td>88.33</td>
<td>88.89</td>
<td>88.13</td>
<td>90.44</td>
<td>89.39</td>
<td>89.14</td>
</tr>
<tr>
<td>Boosting</td>
<td>79.9</td>
<td>74.44</td>
<td>75.25</td>
<td>75.25</td>
<td>75.76</td>
<td>75.5</td>
<td>75.25</td>
</tr>
<tr>
<td>Bagging</td>
<td>88.52</td>
<td>86.56</td>
<td>86.11</td>
<td>85.61</td>
<td>85.61</td>
<td>86.62</td>
<td>86.11</td>
</tr>
<tr>
<td>MCS</td>
<td>90.77</td>
<td>88.63</td>
<td>90.4</td>
<td>89.5</td>
<td>92.42</td>
<td>90.4</td>
<td>90.66</td>
</tr>
<tr>
<td>Average</td>
<td>87.48</td>
<td>85.16</td>
<td>85.10</td>
<td>84.72</td>
<td>86.05</td>
<td>85.51</td>
<td>85.26</td>
</tr>
</tbody>
</table>
Feature selection by attribute clustering of infected rice plant images

Fig. 1- Rice plant Images (a) stem infected by 'Sheath Rot' (b) leaf infected by 'Blast' (c) leaf infected by 'Brown Spot'.

Fig. 2- Segmented Spots in Background Region

Fig. 3- Segmented Spots in Core Region

Fig. 4- Segmented Spots in Border Regions
Fig. 5 - Boundary of the spots corresponding to Fig. (2(b))

Fig. 6 - Decision Table of Infected Rice Images

Fig. 7 - Conditional Information Gain of the Decision Table
Feature selection by attribute clustering of infected rice plant images

Fig. 8 - 3D Plotting of Attribute Dependency Matrix

Fig. 9 - Percentage of Reduction using different methods