# A SURVEYAL STUDY ON FEATURE SELECTION BASED CLASSIFICATION TECHNIQUES

[1]M.INDIRA, [2]Dr. S.JAYASANKARI
[1]Research Scholar, [2]Assistant Professor
[1]Department of Computer Science,
[1]P.K.R. Arts College for Women(Autonomous),
Gobichettipalayam, Erode(Dt), India.

## ABSTRACT

Data mining is the computing process of analyzing and classifying data from large dataset to detect the patterns and determine the relationships through data analysis. In data mining, the feature selection and classification are major challenging tasks while classifying the data as normal or defected one. The feature selection is process of detecting the subset of input variables by discarding irrelevant features. However, developed techniques failed to provide the better improvement on enhancing the classification accuracy. This research work is focused on improving the classification accuracy by introducing new techniques in order to solve the limitations from existing methods.

**Keywords: Data mining, Classification, Feature selection, Image analysis.**

## 1. INTRODUCTION

Data mining discovers valid information from database to determine significant model and rules by examining the huge data sets. In data mining, classification and self assured classifier rule generation are the widely used mining techniques. In order to perform classification, feature selection process is to be carried out. The feature selection is, the process of identifying relevant features from huge database. Based on relevant features, the data are classified using different classification techniques.

This paper is organized as follows: Section II discusses reviews on data classification based on feature selection techniques, Section III describes existing classification methods, Section IV identifies possible comparison between them, Section V explains limitations and Section VI concludes the paper.

## 2. RESEARCH METHODOLOGY

In [1], PLS-DA and SVM-DA were developed with the objective of providing better results in sensitivity and specificity. But, PLS-DA and SVM-DA failed to improve performance of sensitivity and specificity. In [2], new approaches were developed to address challenges. But, designed approach failed to improve classification accuracy.

The novel technique was implemented in [3] to analyze physical features of castor seed varieties. However, designed technique consumes more time to extract feature from gathered castor seeds. In local neighborhood of each point, the geometric and cloud features of radiometric point in Terrestrial Laser Scanning (TLS) data was presented by [4] to execute the comparative classification analysis for post-harvest growth detection. But, time for classification remained unaddressed.

Multi-class agricultural datasets was classified with Hybrid Kernel based Support Vector Machine (H-SVM) in [5]. However, designed method was not able to perform feature selection in maximal error rate. C-band, dual polarimetric and temporal satellite of RISAT-1 was determined in [6]. But, the divergence method was not efficient to reduce the error rate.

The multi-class disease classification issues were addressed by developing improved Random Forest Classifier (RFC) approach in [7]. However, improved RFC approach failed to improve classification accuracy. In [8], crops were organized from satellite images by implementing agricultural expert knowledge and machine learning algorithm through the hybrid intelligent system. But, computational complexity remained unaddressed. According to the intra- and inter-provenance levels, the seed germination and seedling growth differences were discussed in [10]. Hybrid ensemble approach was developed in [9] by combining machine learning algorithms, a feature ranking method and supervised instance filter. However, hybrid ensemble approach consumes more time to perform classification.

## 3. DATA CLASSIFICATION

Data mining is the major process to predict the data outcomes by identifying the anomalies and correlations from data sets. The data classification is one of the significant ways to predict future data.

### 3.1 Classifying castor seeds based on morphological and color features

The new precise image analysis and data mining technique is implemented for organizing seeds of castor according to measurement of morphological and color features. The designed technique is involved with processes namely data collection, image acquisition and correction, digital image processing stage and classification algorithms to group castor seeds samples based on features such as size and color.

At first, data collection process is carried out to collect samples of seeds. The data set is constructed while considering five hundred seeds. The seeds are collected points from roadways, on soil slopes, crops, and close to water channels. Then, these seeds bunches are stored in paper bags for future analysis. The image acquisition and correction process is initialized by placing seed samples on rotating electromechanical system.

In order to extract the features of seeds, image segmentation, and feature extraction methods are performed. In the image segmentation process, subtraction operation is carried out between input image holding the seed and background without it. The image is divided into two regions as pixels belonging to the seed, and those from background by applying binarization operation.

The classification process is carried out to cluster castor seeds with similar features. The seed classification is done by characterizing the castor seeds with the help of machine learning algorithms. Through classification algorithm, the seeds are classified which is utilized to design the oil extraction equipment.

### 3.2 Crop classification by using Artificial Neural Network (ANN)

The C-band RISAT-1 Satellite Dataset is used with the aim of performing classification by using artificial neural network.

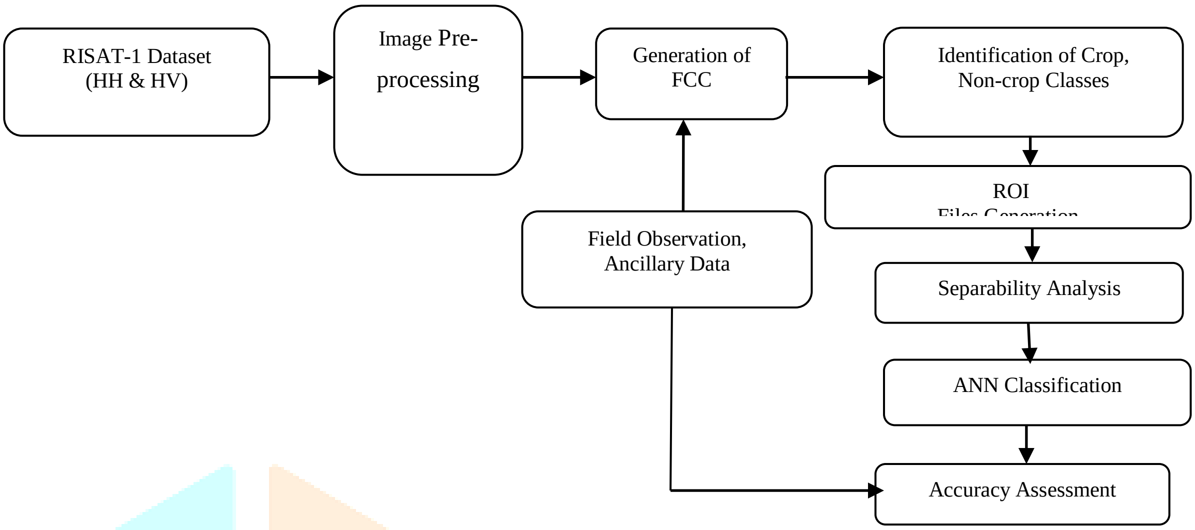

**Figure 1 Methodology of ANN**

Figure 1 shows methodology of the artificial neural network. HH and HV polarization data are taken to perform the task. At image preprocessing phase, images are filtered by implementing frost filter to the remove speckle noise. The calibration constant in metadata is ensured to create sigma naught images. The non-linear and complex patterns are generated by using ANN classification algorithm with appropriate topological structures. The structure of ANN is enclosed with three layers namely input layer, one hidden layer and output layer. In the input layer, neuron is indicated as one of the input features as one satellite image band. The input layer comprises of 3 bands as a number of neuron and single hidden layer comprise of 8 neurons. The output layer is enclosed with 6 neurons as crop classes to perform classification. ANN algorithm is designed with standard back propagation for supervised learning. By using ANN algorithm, root mean square error (RMSE) is reduced between the actual outputs of multilayer feed forward ANN.

### 3.3 Multiclass classification by Random Forest Classifier (RFC)

The problems in multi class disease classification are addressed by enhancing accuracy of random forest algorithm through implementation of improved Random Forest Classifier (RFC). The improved-RFC approach is formed by combining random forest machine learning algorithm, an attribute evaluator method and instance filter method.

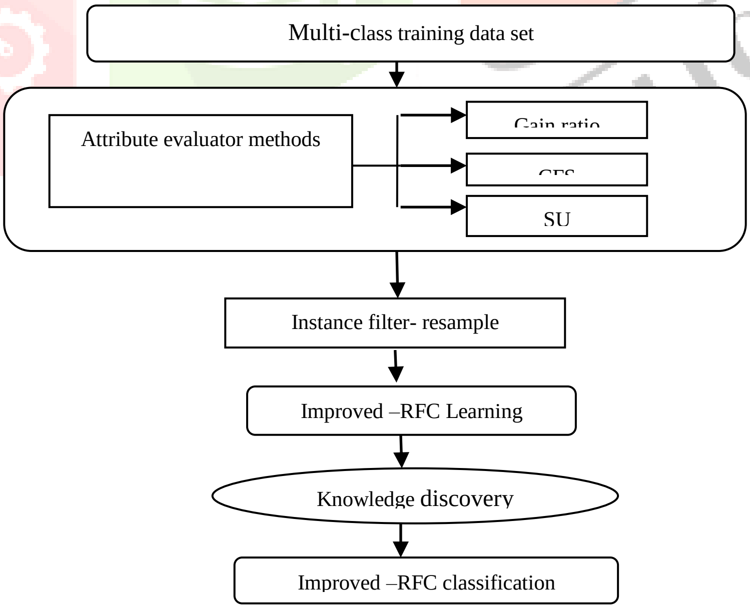
Multi-class training data set
Attribute evaluator methods
Gain ratio
CFS
SU
Instance filter- resample
Improved –RFC Learning
Knowledge discovery
Improved –RFC classification

**Figure 2 Process for Improved –RFC classification**

The attribute evaluator method is type of feature selection method utilized to identify the related features and to avoid the irrelevant feature. The attribute evaluator method is utilized for data analysis and capability of data reduction by estimating the Correlation-based Feature Selection (CFS), Symmetrical Uncertainty (SU) and Gain Ratio.

CFS is the process of grading feature subsets consistent with correlation based heuristic estimation function. The correlation among features is identified through symmetrical measurement. The bias of mutual information is balanced by determination of SU. SU is measured as the fraction between the two features of Information Gain (IG) and Entropy (H).

The instance filter is fundamental type of random sampling technique. The real-world dataset comprises of non-uniform class distribution. In training phase, performance of classification algorithm is controlled by non-uniformity of class distribution.

The figure 2 illustrates the process involved in achieving the improved-RFC classification. At first, an improved RFC-approach selects the multi-class training dataset to perform the classification. The attribute evaluator method determines the CFS, SU and Gain Ratio. Then, the attribute evaluator method is applied to the training dataset to identify the relevant attributes which is used for classification.

## 4 RESULTS AND DISCUSSION

In order to compare the feature selection and classification techniques, number of features is taken as input to carry out the experiments. Various parameters are used for analyzing the performance of seed classification with improved classification accuracy.

### 4.1 Classification accuracy (CA)

The classification accuracy is measured as the ratio of number of data correctly classified to the total number of data.

$$CA = \frac{\text{Number of Data Classified correctly}}{\text{Total number of data}} * 100$$

..Eqn (4.1)

The classification accuracy is measured in terms of percentage (%). Higher classification accuracy ensures the better performance of the method.

**Table 4.1 Tabulation of Classification accuracy**

| Number of Data | Classification accuracy (%) | | |
|---|---|---|---|
| | Precise image analysis and data mining technique | Artificial Neural Network (ANN) classification | Improved Random Forest Classifier (RFC) approach |
| 5 | 50.45 | 47.67 | 45.56 |
| 10 | 52.67 | 49.75 | 46.78 |
| 15 | 57.87 | 51.23 | 48.23 |
| 20 | 59.23 | 53.34 | 50.23 |
| 25 | 63.21 | 56.12 | 51.78 |
| 30 | 67.35 | 59.34 | 53.42 |
| 35 | 71.57 | 62.43 | 55.34 |
| 40 | 72.56 | 65.78 | 57.23 |
| 45 | 73.78 | 67.23 | 59.32 |
| 50 | 75.24 | 69.12 | 61.23 |

Table 4.1 shows the experiment results of classification accuracy with respect to number of data. The number of data is considered from 5 to 50 which are considered as an input. For the simulation purposes, the three methods such as feature based classification algorithm, Artificial Neural Network (ANN) classification and Improved Random Forest Classifier (RFC) approach are compared. The precise image analysis and data mining technique effectively improves the classification accuracy when compared to other methods. In precise image analysis and data mining technique by implementing random tree method, the data are classified as normal or affected one based on image segmentation, and feature extraction methods. As a result, the classification accuracy in precise image analysis and data mining technique is improved by 11% when compared to ANN classification and improves 21% when compared to Improved RFC approach respectively.

### 4.2 Feature Selection Time

The feature selection time (FST) is evaluated as the difference between the ending time and starting for selecting the features of data and it is given by,

$$\text{FST} = \text{Ending time} - \text{Starting time of feature selection}$$

..Eqn (4.2).

From Eqn (4.2), feature selection time is measured. The feature selection time is measured in terms of milliseconds (ms). When the feature selection time is less, method is said to more efficient.

Figure 4.2 illustrates the measurement of feature selection time by using three methods namely precise image analysis and data mining technique, ANN classification and Improved RFC approach. It is evident that, the artificial neural network effectively reduces feature selection time when compared to other methods.

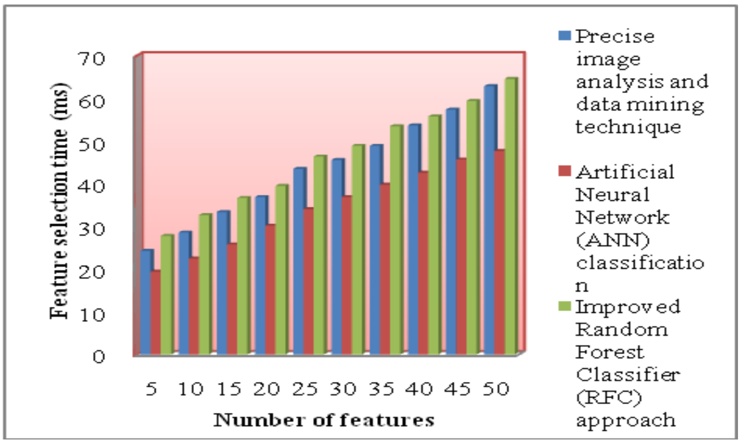

**Figure 4.2 Measurement of Feature selection time**

The separability analysis is carried out with transformed divergence (TD) and Jefferies Matusita (JM) method to classify the crops by implementing ANN classification algorithm. The feature selection time in ANN classification is reduced by 21% when compared to precise image analysis and data mining technique and improves 26% when compared to Improved RFC approach respectively.

### 4.3 Classification Time (CT)

The classification time is defined as difference between ending time and starting time for classifying the number of data based on the features by conducting the experiments.

CT = Ending time – Starting time of data classification

...Eqn (4.3)

From Eqn (4.3), classification time is measured. The classification time is measured in terms of milliseconds (ms). If the classification time is low, then the method is said to be efficient.

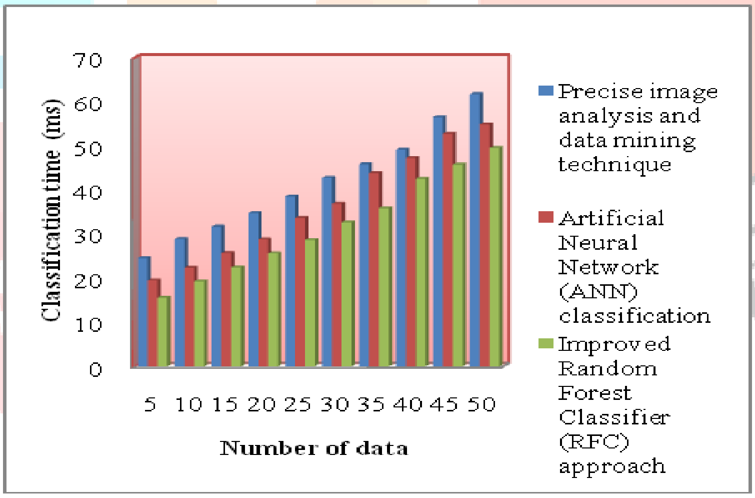

**Figure 4.3 Measurement of Classification time**

Figure 4.3 illustrates measurement of classification time by using three methods namely precise image analysis and data mining technique, ANN classification and Improved RFC approach. The artificial neural network effectively reduces classification time when compared to other methods. In Improved RFC approach, instance filter method is used to balance the class distributions. This in turns, the classification time to classify the data gets reduced. As a result, the classification time in Improved RFC approach is reduced by 25% when compared to precise image analysis and data mining technique and improves 13% when compared to ANN classification respectively.

## 5 DISCUSSION ON LIMITATIONS TO PERFORM SEED CLASSIFICATION BY FEATURE EXTRACTION

The classification of seeds for oil extraction equipment is carried out by developing the precise image analysis and data mining technique. However, the time was too high to extract the feature from the gathered caster seeds.

ANN classification of crop is performed by utilizing the C-band RISAT-1 Satellite Datasets. By the combination of transformed divergence and Jefferies Matusita distance techniques, the separability analysis is performed for artificial neural network. Then, the transformed divergence method ensures better results in performing the separation among classes. But, error rate remained unaddressed while using the divergence method.

The improved-RFC approach is implemented with the objective of resolving the issues in the multi-class disease classification. However, Improved-RFC approach failed to enhance the classification accuracy.

### 5.1 FUTURE DIRECTION

The future direction of the proposed scheme is to enhance the data classifications using improved data mining techniques. Besides, some of classification techniques are developed in future with the objective of improving the classification accuracy by reducing the time for classifying the data. Another future direction is carried out to detect the relevant features for performing the effective classification with minimized feature selection time.

## 6 CONCLUSION

The survival study is carried out to ensure the improved seed growth in the agricultural area. The feature selection based classification techniques are tested with the metrics such as classification accuracy, feature selection time and classification time. Finally, further research work is performed with effective seed classification to increase the seed growth by achieving improved classification accuracy with minimal feature selection time.

## REFERENCES

[1] Bruna Tassi Borille, Marcelo Caetano Alexandre Marcelo, Rafael Scorsatto Ortiz, Kristiane de Cássia Mariotti, Marco Flores Ferrao, Renata Pereira Limberger, “Near infrared spectroscopy combined with chemometrics for growth stage classification of cannabis cultivated in a greenhouse from seized seeds”, Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, Elsevier, Volume 173, 2017, Pages 318–323

[2] Christian Bauckhage and Kristian Kersting, “Data Mining and Pattern Recognition in Agriculture”, KI - Künstliche Intelligenz, Springer, Volume 27, Issue 4, November 2013, Pages 313–324

[3] Cesar Isaza, Karina Anaya, Jonny Zavala de Paz, Jose F. Vasco-Leal, Ismael Hernandez-Rios and Jose D. Mosquera-Artamonov, “Image analysis and data mining techniques for classification of morphological and color features for seeds of the wild castor oil plant”, Multimedia Tools and Applications, Springer, Pages 1–18

[4] Kristina Koenig, Bernhard Höfle, Martin Hämmerle, Thomas Jarmer, Bastian Siegmann and Holger Lilienthal, “Comparative classification analysis of post-harvest growth detection from terrestrial LiDAR point clouds in precision agriculture”, ISPRS Journal of Photogrammetry and Remote Sensing, Elsevier, Volume 104, 2015, Pages 112–125

[5] K. Aditya Shastry, H.A. Sanjay and G. Deexith, “Quadratic-Radial-Basis-Function-Kernel for classifying multi-class agricultural datasets with continuous attributes”, Applied Soft Computing, Elsevier, Volume 58, September 2017, Pages 65–74

[6] P. Kumar, R. Prasad, V. N. Mishra, D. K. Gupta and S. K. Singh, “Artificial Neural Network for Crop Classification Using C-band RISAT-1 Satellite Datasets”, Russian Agricultural Sciences, Volume 42, Issue 3–4, 2016, Pages 281–284

[7] Archana Chaudhary, Savita Kolhe and Raj Kamal, “An improved random forest classifier for multi-class classification”, Information Processing in Agriculture, Elsevier, Volume 3, 2016, Pages 215–222

[8] Stefan Contiu and Adrian Groza, “Improving remote sensing crop classification by argumentation-based conflict resolution in ensemble learning”, Expert Systems with Applications, Elsevier, Volume 64, 1 December 2016, Pages 269-286

[9] Archana Chaudhary, Savita Kolhe and Raj Kamal, “A hybrid ensemble for classification in multiclass datasets: An application to oilseed disease dataset”, Computers and Electronics in Agriculture, Elsevier, Volume 124, 2016, Pages 65–72

[10] Romulo Santelices Moya, Sergio Espinoza Meza, Carlos Magni Díaz, Antonio Cabrera Ariza, Sergio Donoso Calderon and Karen Peña-Rojas, “Variability in seed germination and seedling growth at the intra- and interprovenance levels of Nothofagus glauca (Lophozonia glauca), an endemic species of Central Chile”, New Zealand Journal of Forestry Science, Volume 47, Issue 10, 2017, Pages 1-9