Identification of Plant Species using Supervised Machine Learning

Ankita Tripathi
Amity institute of biotechnology
Amity University, Gurgaon, IN

Ravi Datta Sharma
Amity institute of biotechnology
Amity University, Gurgaon, IN

Shrawan Kumar Trivedi
Indian Institute of Management
Sirmaur, HP, IN

ABSTRACT

This research emphasizes on the plant species recognition which is considered as an important area of research in plant biotechnology. Artificial intelligence and machine learning have a prominent place in such research. In this study, a boosted evolutionary plant species classifier has been developed that works on ensemble of classifier methods. This classifier identifies different species of plants with the help of different texture and shape features of leaf image. A publicly available plant image dataset has been incorporated where features are extracted with the help of image processing tools. The proposed classifier is trained and tested with the help of these features. Further, proposed classifier is compared with other popular machine learning classifier viz. Bayesian, Naïve Bayes, SVM, J48, Random forest, Genetic Programming. Proposed evolutionary classifier was found to be good in terms of F-Value, FP rate and TP rate whereas SVM was found to be underperforming predictor in this study. However, the training time of the proposed classifier was high.

General Terms

Plant Species Identification, Machine Learning Classifiers, Pattern Recognition

Keywords

Plant Species, Leaf image, Genetic programming, Machine learning, F-Value, FP rate, Training time.

1. INTRODUCTION

We ask that authors follow some simple guidelines. In essence, we ask you to make your paper look exactly like this document. The easiest way to do this is simply to download the template, and replace the content with your own material.

Plants play a great role in medicine, foodstuff, industry and also essential for environmental protection [1]. Approximately 3, 00,000 plant species have already been recognized throughout the world but still a number of species are unknown [2]. Plant identification is promising research in the plant research domain.

Plant classification has generally been done by various taxonomists who use many different features/attributes of plants like shape, colour and fragrance of the leaves, flowers, bark, seedlings, size & shape of fruits etc. These features/attributes play a vital role in the identification of various plant species. Leaf shape is useful attribute to identify the different characteristics of plants [3].

Nowadays, many automated tools are playing promising role in such research domain. Plant recognition is the complex task that needs a huge financial expenditure. Recently, many automatic digitalized systems have been developed for identification of plant species. In today's digitalised world, smart phones are considered more personalized medium of technology [4] and it may play an important role in plant recognition. Different plant images are captured from smart phone for identifying different shape features. Some benefits are seen in plant recognition by capturing the shape features of images. These features help in avoiding the use of large amount of chemicals, also huge time consumption and complexity from laboratory practical work. Plant recognition from image features may provide a costless process which saves time and money with efficient image processing tools.

Plethora of research has been conducted in the plant identification domain. Such research is divided in three parts viz. generic plant identification systems, agriculture systems and also systems for the intra-specific variation, ecological effects as well as geographical distribution.

A number of systems of automatic plant recognition have been proposed in last few years. The system proposed by [5] was one of these proposed automatic systems where the local shape properties were taken instead of global shape based approach to tackle with the damaged or overlapped leaves. The approach uses the shape analysis by involving Fourier descriptors and dynamic programming in combination with polygon fitting. The authors analysed a good accuracy rate by using existing database.

The new version of mobile phone devices has been altering the purpose of plant recognition system. The users both specialities and non-specialities show interest by capturing plant leaves images in field to identify plants. One of the current on-going projects related with the development of field guide of plants in United States of America. This system permits the user to capture a picture of leaf in simple background and gives result of twenty similar leaves which are closest to the input query. There is another project [6] named clover systems with same philosophy but this project uses some other shape analysing techniques. These both projects have given results by using at least a few specimens of leaves. The image processing technique was used in one oldest paper of [7] for recognition of weeds in crops.

Hemming et al. in 2001 [8] differentiated between two crop species and weed plants. Many other papers also focused on improvement of recognition rates.

In this work, leaf image features data [9] has been incorporated where the shape and texture features were extracted from the images of the different leaves of different plant species [3]. While such methods provide benefits in plant recognition but in reality Various factors as Inappropriate digitalization, typical geometry of leaf and contamination in leaves by diseases etc. can affect this process. Many advance technology and image processing tools are being used in such research to tackle above problems [10, 11, 12].

In this research, an improved genetic programming classification model is being proposed for plant species recognition. This classifier works on ensemble of classifiers technique where a number of genetic programming classifier is ensemble to get accurate classification research. Further, the proposed classifier is compared with popular machine learning.

2. METHODOLOGY

2.1 Data Description:

This data was created by [9] where earlier 40 plant species were taken for experiment. Dataset was built by capturing approximately 10 leaf specimen images from each unique plant and thus approximately 443 images of leaves were collected. Canon EOS 40D reex camera and an Apple iPAD2 tablet were used for capturing the images of each leaf specimen.

Leaf database overview - 40 class types (Pedro et al. 2013)

2.2 Image Feature Extraction Process:

The image pixel for each image specimen was kept 720x920 pixel with 24 bit RGB and the background of most of the images were kept in contrast (e.g. Green leaves image were captured in reddish background). However, for some special cases like acer palmatum leaves grey colour was used as image specimen background. The choice of colour was based on the condition of the leaves of the plant. After capturing all the required images, a survey was done on the image complexity where two classes of leaves were created named simple leaves and complex leaves

The Simple leaves classes have been assigned to the species numbered from 1 to 15 and from 22 to 36 whereas species numbered 16 to 21 and 37-40 were assigned complex leaves class. Only simple leaves have been considered for this research because complex leaves can be deficient with the proposed systems. EFD can be used for a little description but due to shape variability in the leaves of these classes can yield senseless results. One other reason is that there will be difficulty in distinguishing results of EFA & other techniques. The complex leaves specimens were collected for near future work & for accumulation of database of complex leaves.

3. MACHINE LEARNING MODELS

3.1 Genetic Programming (GP):

Genetic programming (GP) [13] is a machine learning (ML) classifier that works on the biological evolution concept to find an optimum solution for a particular problem. In this technique, an individual is considered as the solution of a problem. Initially, few individuals are randomly selected from a population. A fitness function is used to assign a fitness value for each individual in the population for evaluating the performance of every individual. Thereafter, Genetic Operators (i.e. selection, crossover and mutation) are engaged for generating "Offspring". This evolution process will continue until the optimum solution is not found.

Individual Representation:

Genetic programming generally constitutes a binary tree based structure [14] to represent an individual. The non-leaf-nodes of the tree perform operations on the terminals of the leaf node. It has been identified that simple arithmetic operators are sufficient to achieve higher classification accuracy and lesser computation cost [15]. Functions like $+, -, \times, \div$ are used for individual representation. Two kinds of terminal are observed in the tree i.e. feature terminals and constant terminals. Feature terminals are evaluated by the training data corresponding to the features selected from the leaves image of different plants. Usually, transformed link based features (such as Log of in degree, Log of out degree, and so on) are preferred. Constant terminals are 14 different floating numbers which are 0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0, 3.14159, 2.71828, and 1.5708. These numbers will remain unchanged during evolutionary progression.

Fitness function:

Fitness function $F(\mathrm{I}^x)$ for any individual I^x is represented as $A_{ccu}(\mathrm{I}^x, P^{Leaves})$ and can be calculated by the following equation:

$$\mathrm{A_{ccu}}(\mathrm{I}^{\mathrm{x}}, \mathrm{P}^{\mathrm{Leaves}}) = \frac{\mathrm{P}^{\mathrm{Leaves}}_{\mathrm{I}^{\mathrm{x}}} \to \mathrm{correct}}{\left|\mathrm{P}^{\mathrm{Leaves}}\right|} \qquad (1)$$

Where, P^{Leaves} is termed as Leaves Image dataset and $P^{Leaves}_{I^x} \to correct$ is the total number of correctly classified leaves species by the individual I^x.

Algorithm for GP:

Input: Training Set T^x, Number of individuals N^x, Maximum depth of binary tree D^x, Number of Generations R^x.
Output: A best individual with a unique discriminant function

Initialize population P^x from arbitrary generated individuals with respect to N^x and D^x.

Assign values for operators i.e. mutation (m^x=0.07), Crossover (c^x=0.9), New program (N_p=0.03) and Re-production (r_p=0.0).

Calculate the fitness value $F(\mathrm{I}^x)$ for individuals I^x, where $\mathrm{I}^x \in P^x$ with training set T^x.

Perform Genetic operators:

Re-production operator: select two most fit individuals from the population P^x and put them in to new population P_n^x.

Mutation operator: Mutation operator is applied on a randomly selected individual to compute its fitness for mutant and compare it with best fit individual and put in the P_n^x

Crossover operator: Crossover operator is performed in to two best fit individual selected above. The properties of parent individuals exchanged to form two offspring. Compute the fitness value of these new offspring and compare then put then in new population P_n^x .

Repeat until $\left|P^x\right| < N^x$.

Let $P^x = P_n^x$, $P_n^x = \phi$, and $r_p = r_p$ +1.

Repeat until $r_p < R^x$

Evaluate and compare the fitness function of every individual in population P^x with training set T^x , for observed the best output.

3.2 Boosted Genetic Programming (IGP)

The proposed Boosted Genetic Programming (IGP) works on the ensemble of classifiers methods [16] to generate optimum classification output. Ensemble based methods obtain a strong classifier by ensemble of many weak classifiers by combining their individual decisions. IGP technique uses the same concept and combines many weak Genetic Programming (GP) classifiers. Adaptive Boosting mechanism (AdaBoost) is incorporated to alter the sample dataset with re-weight method. The proposed method uses Genetic Programming (GP) classifier as a weak classifier which has to be modify. In this research, initially different set of number for ensemble member M_x are tested in the ascending order. After combining more than 40 weak classifiers ($M_x \geq 40$), an Improved Genetic Programming classifier with excellent classification accuracy have been identified.

Considering a set of output M_x of weak Genetic Programming classifiers $GP_t^{m_x}$ for learning process with Adaptive Boosting which will visualize the decision for the final classifier IGP_t^x .

Algorithm for IGP

Input: Training set $T_r = t_{1,}t_2, t_3 ... t_n$ with $t_i = \left(x^i, y^i\right)$.

Number of sample version of training set B .

Output: An Enhanced Genetic Programming classifier EGP_t^x

.

Initialise the weights $w_i^t = \frac{1}{N}, \quad i \in \{1, 2, 3, ...N\}$

From $m = 1, 2, 3, ...M_x$

Train the weak classifier $GP_t^{m_x}$ with the training outset using weights w_i^t .

Calculate the error term $E_{rror}^m = \frac{\sum_{i=1}^N w_i^t I\,(\mathrm{y}_i \neq GP_t^{m_{xi}})}{\sum_{i=1}^N w_i^t}$.

Calculate weight contribution $\theta_m = 0.5 \log\left(\frac{1-E_{rror}^m}{E_{rror}^m}\right)$

Substitute $w_i^t \leftarrow w_i^t Exp\left(-\theta_{(m)} I\left(y_i \neq GP_t^{m_{xi}}\right)\right)$

Then renormalize $\sum_i w_i^t = 1$.

The final Enhanced GP classifier is :–

$$BGP_t^x = \theta_m sign\left(\sum_{m=1}^{M_x} GP_t^{m_x}\right) \qquad (2)$$

3.3 Probabilistic Classifiers

Bayesian Model

Bayesian model is a probabilistic classifier [17, 18] with a nature of white box. It is used to predict particular class of membership samples. This model works on Bayesian theory and explained by the following model. Let us consider a training sample set is $D = \{u_1......u_n\}$, where mission of the classifier is to evaluate the training sample and determine its function $f:(x_1....x_n) \to C$ for deciding the label of the sample x= $(x_1....x_n)$ with respect to the highest probability of the class as per the label $P\left(c_j / x_1,....x_n\right)$. According to minimum error probability criterion:

If $p\left(c_i\middle/x\right) = \max_{j=1,...i} P\left(c_j\middle/x\right)$ then we can determine that $x \in c_i$

The two commonly used models are Naïve Bayes and Bayesian belief. Naïve Bayesian classifier assumes that independent samples are used. Even if the calculation is simplified in this model the variables are correlated really. It is a graphical model where conditional independencies are characterized between subsets of variables. Bayesian network consists of two sections namely: cyclic graph and Conditional probability tables.

3.4 Support Vector Machine (SVM):

In recent years SVM [19] is the best classifier developed for Pattern Classification. It does not limit the distribution of data and mostly used for small samples. This model also achieves good robustness which is based structural risk. Let us consider S is a dataset and M with observations set defined as $\left\{\left(x_{i,}y_i\right)\middle/x_i \in R^n; y_i \in \{-1,+1\}, i = 1, 2....M\right\}$ where $\{x_i, z, y_i \in \{-1,+1\}$ denotes equivalent binary class label, suggesting whether the client or customer is default. The main purpose of this categorization is to find a maximal hyper plane by which the examples of opposite labels are separated. This constraint is written as:

$$y_i\left((w, x_i) + b\right) - 1 \geq 0, i = 1, 2...M \qquad (3)$$

Where w is defined as the plane’s normal and b is defined as intercept. (w,b) denote linear set. $2\big/\|w\|$ is margin of separation. The optimal hyper plane is the point where margin $2\big/\|w\|$ is maximum. Subject to constraints of $y_i\left((w, x_i) + b\right) - 1 \geq 0, i = 1, 2...M$ .Then solving the quadratic equation $\min_{w,b} 1\big/2\|w\|2$ is the classification problem.

$$y_i\left((w, x_i) + b\right) - 1 \geq 0, i = 1, 2, ...M \qquad (4)$$

By bringing in langrage multipliers $\alpha = \left(\alpha_{1,}\alpha_2, ...\alpha_m\right)$ the problem is changed to solve the dual program as follows:

$$\max_{\alpha} Q(\alpha) = \sum_{i=1}^{M}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\alpha_i\alpha_j y_i y_j \left(x_i, x_j\right) \qquad (5)$$

$$\text{s.t} \quad \sum_{i=1}^{M} y_i\alpha_i = 0, \alpha_i \ge 0, i = 1, 2...M \qquad (6)$$

If α >0 then xi is called support vector. From the above problem the decision function obtained is formulated as

$$f(x) = \text{sgn}\left(\left(w, x\right) + b\right) \qquad (7)$$

$$\varepsilon^i = \text{sgn}\left\{\sum_{i=1}^{M}\alpha_i y_i \left(x_{i,} x\right) + b\right\} \qquad (8)$$

The decision function obtained above defines that examples are classified as class +1 when $(w,x)+b>0$ and class -1 when $(w,x)+b<0$. The mapping of input vectors in the form of high-dimensional feature space via an inferred chosen mapping function φ is to be done if the mapping is non-separable. By means of a Kernel function $k\left(x_i, x_j\right) = \varphi\left(x_i\right)\varphi\left(x_j\right)$ mapping can be done implicitly. There are four kinds of kernels like Linear, Polynomial with degree d, sigmoid and RBF kernels. In linear non separable case the training errors are allowed. So called slack variables ε^i are thus introduced in order to be tolerant of classification error.

3.5 Decision Tree (J48):

J48 [20] is open source implementation in JAVA platform which is a kind of decision tree classifier. This algorithm was suggested and implemented by Ross Quinlan. It is based on C4.5 algorithms. This classifier divides the dataset based on their attributes taken from training data. The idea of entropy is to train the data to develop the decision tree.J48 neglects the values which are missing. Leaf node is formed in the decision tree to select the class.

Algorithm for C4.5:

Checking the above mentioned base cases

- For every feature x^i , find normalise information gain by splitting capability on x^i .
- If the x_b^i is an informative feature with extreme normalise gain, establish a decision node that split on x_b^i .
- Repeat the above on the sub lists formed by splitting on x_b^i

3.6 Random Forest (RF):

Random Forest [21, 22] is a classifying technique used in data mining which works on ensemble of classifier method. It combines decision of many classifiers to generate a suitable result. Bagging technique approach is the first technique which alters the samples of data and randomly selected feature input is the second technique.

Algorithm of Random Forest:

Given: n^T - number of training examples, x^i -number of all features, x^e -number of features selected for Ensembles, m^i - number of all Ensemble members

Create Random Forest for m^i trees

For each m^i iterations:

do,

- Bagging: Sample n^T with substitution from training data
- Random feature selection: Grow the decision tree without trimming. For each step, choose informative features by considering only x^e arbitrarily selected features and achieving the Gini index.
- Classification:

  Apply text set to each of the m^i decision trees starting from the root node. Assign it to a particular from respective leaf node. Combine the decisions of each member by mass majority voting and resulting ideal classification.

4. EXPERIMENTAL DESIGN

4.1 Instruments for Evaluation

Three measures are used for evaluating the performance: Accuracy, F-value and False Positive rate.

Accuracy [23]

It is the ratio of total correctly classified Leaf Image to the total text Leaf s and represented as:

$$\text{A}_{\text{cc}} = \frac{\text{Corrctly\_Classified\_Leaf\_Image}}{\text{Total\_Image}} \qquad (8)$$

F-Value [23]

It is defined as the harmonic sum of Precision (i.e. fraction of retrieved classified leaf Images that are relevant) and Recall (fraction of accurate classified Leaf Images that are retrieved) and represented as:

$$\text{F}_{\text{value}} = \frac{\text{2*Precision*Recall}}{\text{Precision+Recall}} \qquad (9)$$

False Positive Rate [24]

This instrument performs to measure the sensitivity of accurate classification and tells how many positive instances are misclassified. It is represented as:

$$\text{FP}_{\text{rate}} = \frac{\text{Misclassified\_Leaves\_Images}}{\text{Misclassified\_Leaves\_Images+Correctly\_Classified\_Leaves\_Images}} \qquad (10)$$

4.2 Software:

MATLAB 2008 and JAVA based implementation on the WINDOW 7 operating system with 4GB RAM has been preferred for this study.

4.3 Design Parameters

After Pre-processing, 36 different plant species have been incorporated for building robust classification model. Whole dataset is split to obtain 66% files for training and 34% for testing.

TABLE I. Values For Parameters Of Proposed Igp Classifier

Parameters	Value
Target Fitness ($F(\text{I}^x)$)	90%
Maximum Generation (R^x)	20
Maximum tree depth (D^x)	5

Mutation rate (m^x)	7%
Crossover rate (c^x)	90%
New Program generation (N_p)	3%
No. of classifiers for ensemble generation (M_x)	40

5. RESULTS AND DISCUSSIONS

The analysis section is divided in three segments. The first segment demonstrates the performance with F-Value (Table II & Fig 2) of proposed IGP classifier and various other machine learning classifiers such as “Bayesian, NB, SVM, Logistic Regression, J48, RF, and GP”. Second part presents False Positive rate and True Positive rate (Table II & Fig. 2) of proposed IGP classifier and other classifiers. The last third segment discuss about the training time.

TABLE II. F-Value For Machine Learning Classifiers

Machine Learning Classifiers							
	Bayesian	NB	SVM	LR	J48	GP	IGP
FV (%)	65.3	76.4	32.8	71.2	66.4	60.5	78
FP (%)	1.10	0.70	1.40	0.90	1.10	1.20	0.70
TP (%)	66.4	76.7	37.1	71.6	67.2	70	77.67
TT (s)	0.05	0.01	10.6	7.8	0.05	19	678.2

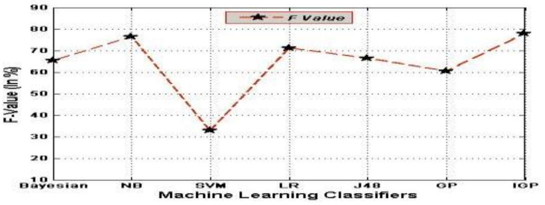

FIG 1.: Value (Leaf Corpus)

5.1 Analysis of F-Value:

F-Value is an important metric to evaluate performance accuracy of the classifiers. It takes the harmonic sum of precision and recall to compute the accuracy value. After testing proposed IGP classifier and other classifiers on the features of plant species, different observations have been identified.

Observation 1: Table II and Fig 1 show the F-Value of proposed IGP and other ML classifiers. In the first observation, proposed IGP classifier (with F-Value 78%) is found to be good in comparison with other classifiers. However, Naïve Bayes (with F-Value 76.4%) and Logistic regression (with F-Value 71.2%) classifiers are found to be second and third best classifier respectively.

Observation 2: This observation identifies SVM (with F-Value 32.8%) as the worst classifier. SVM works on the statistical learning theory that maximize the hyper plane which is created between two classes and found good for two class problems. For multi-classification filter, SVM performs worst.

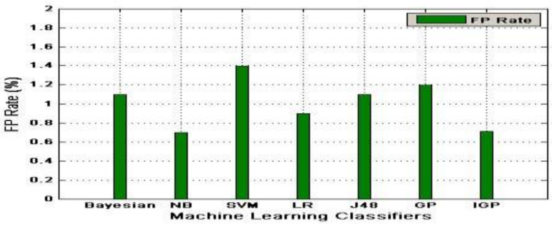

Fig 2. False Positive Rate for Machine Learning Classifiers

5.2 Analysis with False Positive (FP) Rate:

False positive rate is the important metric that evaluate how many instances are misclassified. After analysis of FP rate of the concerned classifiers, following observations have been identified.

Observation 1:
Table II and Fig 2, observe the FP rate of the proposed and other classifiers. It has been observed that FP rate of proposed IGP classifier and Naïve Bayes are comparable and found low (0.7%). This indicates IGP classifier is good in terms of FP rate also. However, Logistic Regression is found to be second best with FP rate 0.9%.

Observation 2:
This observation shows SVM as underperforming classifier with high FP rate i.e. 1.4%. However, the performance of GP and J48 are also not satisfactory i.e. 1.2% and 1.1% FP rate.

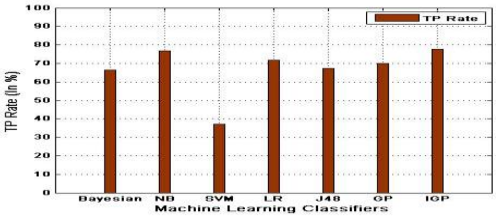

Fig 3. True Positive rate for Machine Learning Classifiers

5.3 Analysis with True Positive (TP) Rate:

In this section, proposed IGP classifier and other classifiers are evaluated with the True Positive (TP) rate. This metric observes that how many instances are correctly classified. The observations of this section are given below.

Observation 1:
Table 2 and Fig 3 show TP rate of proposed and other classifiers of this research. After comparing all the classifiers, proposed IGP classifier is again found to be a promising classifier with 77.67% TP rate. Naïve bayes classifier is the second best classifier with 76.7% TP rate.

Observation 2:
In this observation, SVM is again found underperforming model with 37.1% TP rate for multi classification.

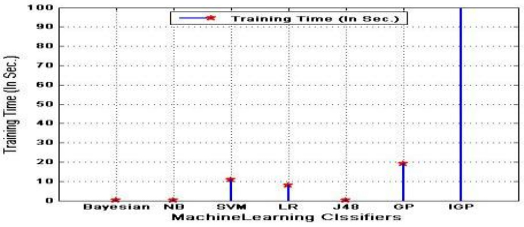

Fig 4. Training Time for Machine Learning Classifiers

5.4 Analysis with Training Time:

Training time is considered the important metric for machine learning classifiers. The observations of this metrics for different classification model have been mentioned below.

Observation 1:
Table 2 and Fig 4, show the results of training time in the unit seconds. The proposed IGP classifier takes huge training time and only on this metric this classifier is underperforming with 678.2 sec training time. This result is not surprising because the proposed classifier works on ensemble of classifiers method. In such method multiple training is performed for multiple classifiers and hence it takes much training time but once the model is trained, training time is not that much important

Observation 2:
In this research Naïve bayes predicted to be good in training time with 0.01 sec training time whereas Bayesian is second best classifier in this research.

5.5 Analysis of F-Value and False Positive rate with TenFold cross validation:

Cross-validation is a method used to evaluate classification models by splitting the corpus into a training set to train the model, and a test set to evaluate it. In TenFold cross-validation [25], the corpus is randomly split into 10 equal-sized subsamples. Of these, a single subsample is taken as the validation data for testing the model, and the remaining 9 are used as training data. The cross-validation process is then repeated 10 times ("folds"), with each of the 10 subsamples used exactly once as validation data. The 10 results from the folds can then be averaged to obtain a single estimation. The advantage of this method is that all observations are used for both training and validation, and each observation is used for validation exactly once.

Twenty-fold cross-validation of the F-value and FP rate for the IGP and other classifiers are shown in Table II tested on plant species dataset. The results clearly validate the performance accuracy of the IGP and other classifiers found using 66/34% split method. A comparison of Table I and Table II shows that the MGP classifier more accurately identifies the plant species category compared to the other classifiers tested in this study.

Table 3: Tenfold Cross Validation of F-Value And FP Rate

Machine Learning Classifiers (TenFold Cross Validation)							
	Bayesian	**NB**	**SVM**	**LR**	**J48**	**GP**	**IGP**
FV (%)	66.1	76.9	33	72.1	67.8	61	78.4
FP (%)	1	0.6	1.3	0.7	1	1.1	0.6

5.6 Wilcoxon Signed Ranked test of accuracy

This study also uses Wilcoxon signed-rank test [26] for further verifying the predictive accuracy of the Improved Genetic Programming (IGP) and other classifiers used in this study. Wilcoxon signed-rank test is used to check whether the results of the plant species dataset are significantly different or not [27]. The null hypothesis of this setup may be formulated as "the predictive capability of two classifiers are same" hence the mean difference of the accuracy of the classifiers would be zero. In this study, all the machine learning classifiers are tested on plant species dataset and Wilcoxon signed ranked test was performed on them. Table IV is showing the p-values of pair wise machine learning classifiers based on the values of F-measures for given datasets. In the Table III, the p-values that are greater than significant level (i.e., $p > 0.05$) have been mentioned bold. Three observations are found from the table IV.

Observation 1: in terms of F-measure, most of the pair wise machine learning classifiers are significantly found different (p-value < 0.05) from the others with only a small number of exceptions (p-value>0.05).

Observation 2: Bayesian and Decision Tree (J48) classifiers are found not significant (p-value>0.05) and hence the performance capability of both classifiers is comparable.

Observation 3: IGP turns out to be good classifier in comparison to other classifiers of this study as the p-value of the IGP classifier is high with the other paired classifiers and hence in the view of F-Value the performance capability of IGP classifier is strong in comparison to other classifiers.

Table 4: P-Values of the paired classifiers

P-Value*	**NB**	**SVM**	**LR**	**J48**	**GP**	**IGP**
Bayesian	2E-06	2E-06	2E-06	0.06	2E-06	2E-06
NB		2E-06	3E-03	2E-06	2E-06	2E-06
SVM			3E-06	2.5E-06	2.5 E-06	2E-06
LR				2E-06	2E-06	3E-06
J48					2E-06	2E-06
GP						3E-06

95% confidence interval

6. CONCLUSION

Plant species recognition is gaining interest of researchers due to advancement of the technology and artificial tools. At the starting of this research, the aim was to construct a robust and accurate filter with the help of machine learning classifiers and artificial intelligence. The motive of this research has been achieved successfully by developing an Improved Genetic Programming (IGP) filter that works on ensemble of classifiers technique. The proposed classifier was trained with the publicly available dataset and found good when it has been compared with other machine learning classifiers.

In addition, to validate the performance of the proposed plant species classifier and other machine learning classifiers, 10 fold cross validation of accuracy and FP rate have been done. In the validation process, the results show the strong support to the results obtained from the 66% - 34% training and testing split setup.

Further, to check the significant difference on the performance accuracy of the proposed plant species classifier and other machine learning classifiers, Wilcoxon signed-rank test of matched data was performed for both datasets. The results of this test suggest that most of the machine learning classifiers are significantly different on 95% confidence interval (p-value < 0.05) from others with only a small number of exceptions (p-value > 0.05). Finally, this research concludes that Genetic Programming (GP) with Adaboost algorithms are good to classify plant species.

In the future, same classifiers can be tested on some other datasets. Different studies can also be done to check the credibility of GP by comparative analysis with other machine learning classifiers.

7. REFERENCES

[1] Rao, P. V., & Gan, S. H. (2014). Cinnamon: a multifaceted medicinal plant. Evidence-Based Complementary and Alternative Medicine, 2014.

[2] Estimated Number of Animal and Plant Species on Earth" Fact Monster. 2000–2013 Sandbox Networks, Inc., publishing as Fact Monster. 18 Apr. 2017,https://www.factmonster.com/science/animals/estimated-number-animal-and-plant-species-earth/

[3] Kadir, A., Nugroho, L. E., Susanto, A., & Santosa, P. I. (2013). Leaf classification using shape, color, and texture features. arXiv preprint arXiv:1401.4447.

[4] Brantes Ferreira, J., Zanela Klein, A., Freitas, A., & Schlemmer, E. (2013). Mobile learning: Definition, uses and challenges. In Increasing student engagement and retention using mobile applications: Smartphones, skype and texting technologies (pp. 47-82). Emerald Group Publishing Limited.

[5] Du J.-X., Wang X.-F., Zhang G.-J. 2007. Leaf shape based plant species recognition. Applied Mathematics and Computation 185: 883–893

[6] Nam, Yun young, Eenjun Hwang, and Dongyoon Kim. "CLOVER: a mobile content-based leaf image retrieval system." In International Conference on Asian Digital Libraries, pp. 139-148. Springer Berlin Heidelberg, 2005.

[7] Jensen, John R., and Kalmesh Lulla. "Introductory digital image processing: a remote sensing perspective." (1987): 65-65.

[8] Hemming, J., & Rath, T. (2001). Computer-vision-based weed identification under field conditions using controlled lighting. Journal of agricultural engineering research, 78(3), 233-244.

[9] Pedro F.B. Silva, Andre R.S. Marcal, Rubim M. Almeida da Silva (2013)Evaluation of Features for Leaf Discriminatio, n',. Springer Lecture Notes in Computer Science, Vol. 7950, 197-204

[10] P Anderson, R., Dudík, M., Ferrier, S., Guisan, A., J Hijmans, R., Huettmann, F., ... & A Loiselle, B. (2006). Novel methods improve prediction of species' distributions from occurrence data. Ecography, 29(2), 129-151.

[11] Guisan, A., Edwards, T. C., & Hastie, T. (2002). Generalized linear and generalized additive models in studies of species distributions: setting the scene. Ecological modelling, 157(2), 89-100.

[12] Skowronek, S., Asner, G. P., & Feilhauer, H. (2017). Performance of one-class classifiers for invasive species mapping using airborne imaging spectroscopy. Ecological Informatics, 37, 66-76.

[13] Trivedi, S. K., & Dey, S. (2013, December). An Enhanced Genetic Programming Approach for Detecting Unsolicited Emails. In Computational Science and Engineering (CSE), 2013 IEEE 16th International Conference on (pp. 1153-1160). IEEE.

[14] J.R. Koza, Genetic Programming: on the Programming of Computers by Means of Natural Selection, MIT Press, Cambridge, MA, 1992.

[15] J.K. Kishore, L.M. Patnaik, V. Mani, V.K. Agrawal, Application of genetic programming for multi-category pattern classification, IEEE Trans. Evol. Comput. 4 (3) (2000) 242–258.

[16] Trivedi, S. K., Dey, S., & Dey, S. (2016). A novel committee selection mechanism for combining classifiers to detect unsolicited emails. VINE Journal of Information and Knowledge Management Systems, 46(4), 524-548.

[17] Lewis, D. D. (1998). Naive (Bayes) at forty: The independence assumption in information retrieval. In Machine learning: ECML-98(pp. 4-15). Springer Berlin Heidelberg.

[18] Tripathi, A., & Trivedi, S. K. (2016, October). Sentiment analyis of Indian movie review with various feature selection techniques. In Advances in Computer Applications (ICACA), IEEE International Conference on (pp. 181-185). IEEE.

[19] V.N Vapnik, "An Overview of Statistical Learning Theory", IEEE Trans.on Neural Network, Vol. 10, No. 5, pp.988-998 , 1999. 6

[20] Bhargava, N., Sharma, G., Bhargava, R., & Mathuria, M. (2013). Decision tree analysis on j48 algorithm for data mining. Proceedings of International Journal of Advanced Research in Computer Science and Software Engineering, 3(6).

[21] Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. R news, 2(3), 18-22.

[22] Trivedi, S. K., & Dey, S. (2014). Interaction between feature subset selection techniques and machine learning classifiers for detecting unsolicited emails.

[23] Provost, F., & Fawcett, T. (2001). Robust classification for imprecise environments. Machine learning, 42(3), 203-231.

[24] Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern recognition, 30(7), 1145-1159.

[25] Moreno-Torres, J. G., Raeder, T., Alaiz-RodríGuez, R., Chawla, N. V., & Herrera, F. (2012). A unifying view on dataset shift in classification. Pattern Recognition, 45(1), 521-530.

[26] Woolson, R. F. (2007). Wilcoxon signed- rank test. Wiley encyclopedia of clinical trials, 1-3.

[27] Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. Journal of Machine learning research, 7(Jan), 1-30.

[28] ACM SIGAPP Applied Computing Review, 14(1), 53-61.