8th IEEE International Conference Humanoid, Nanotechnology, Information Technology
Communication and Control, Environment and Management (HNICEM)
The Institute of Electrical and Electronics Engineers Inc. (IEEE) – Philippine Section
9-12 December 2015 Water Front Hotel, Cebu, Philippines

Using Naïve Bayesian Method for Plant Leaf Classification Based on Shape and Texture Features

Francis Rey F. Padao* and Elmer A. Maravillas†
Department of Computer Science, Cebu Institute of Technology - University
Cebu City, 6000, Philippines
*francisreypadao@gmail.com, †elmer.maravillas@gmail.com

Abstract - As an initial step in protecting different plant species from extinction, establishment of database for plant becomes necessary in order to catalogue various plant diversities. Therefore, automatic and accurate recognition and classification system of plants is important. Thus, the research aims to study plant classification using naïve Bayes (NB) method. Leaf shape and texture serves as input features to the model classifier. The test result shows that the classification accuracy of the model is high. The ROC curve area is 0.981. It indicates that the true positive rating is excellent and the weighted average of the false positive rating is 0.09%, which is considered very minimal and acceptable.

Index Terms—Data Analytics, Plant leaf Classification, Naïve Bayesian, Probabilistic Classification, Supervised Learning.

INTRODUCTION

Plant is a vital component of the ecosystem that provides source of oxygen and food to support humans and other living organism that exist on earth. It is also the raw material for the production of medicines, fibers, building material, ornamental and other plant based products. The United Nations Environment Program (UNEP) discussed that there should be a sustainable approach in the preservation and utilization of natural resources because it is a critical component for economic development. It is also considered as a critical resource to support and sustain human life and human well-being [18].

However, human society has systematically undermined these natural allies, treating ecosystems, such as forests, arable land and rivers as though they are inexhaustible based on the 2014 annual report published by UNEP on Ecosystem Management [18]. If this situation will continuously progress, it will endanger many plant species and would probably lead to its extinction if there will be no early prevention. One sustainable approach is to document plant information across different diversities in a form of plant database to support plant protection and conservation.

978-1-5090-0360-0/15/$31.00 ©2015 IEEE

The United States Department of Agriculture (USDA) has established a database that contains large collection of plant information in order to promote land conservation and encourage information utilization for academic and educational purposes. This approach opens new challenges and opportunities in the field of plant research. Development of effective and efficient automatic plant recognition as well as classification system becomes a vital area of application [3]. Accurate identification of plants is essential in knowing how it grows as well as how to care and protect it from pests and diseases.

Plant classification is a method used to categorize a plant species according to its correct group or taxon based from its most frequent characteristics. The common feature used in classification is the morphological characteristics of leaf extracted from leaf images because it is low-cost and convenient.

In this study, naïve Bayesian (NB) method was used for leaf classification. The identifying features utilized are the texture and shape variables.

As explained by [7], the fundamental basis of NB is Bayes' theorem. It is a statistical approach in classification that predicts classes based on probabilities. NB is also considered to be more efficient compared to decision tree and some applications of neural network in terms of accuracy performance and computational speed [7].

RELATED WORKS

Most studies in plant research have considered physical characteristics of either leaves or petals as basis for the construction of plant classifiers. [12] cited that the most remarkable discriminating plant leaf character is the shape attribute.

[12] further discussed that Shape Features (SF) is one of the most interesting computational approach among different methods in the context of leaf analysis due to its

intuitiveness. Its geometric properties are explicit since it is identical on how the human perception is visually acquired. In addition, the computational cost of this method is relatively low.

Eccentricity, aspect ratio, elongation, solidity, stochastic convexity, isoperimetric factor, maximal indentation depth and lobedness are the attributes used to describe the leaf shape as product of SF and becomes the major consideration in the construction of the model classifier based on the study of [12], [13].

Leaf texture analysis methods were also considered in the study and have generated commonly used texture variables [12]. [12] presented these 7 variables as part of the image processing, including average intensity, average contrast, smoothness, third moment, uniformity and entropy.

In consideration to the classifiers accuracy performance, both shape and texture attributes have been used as input feature for training and testing [12].

PLANT LEAF CLASSIFICATION

The procedure for NB classification is presented below:

1) Let ***D*** be the dataset that will be used for training the classifier. This set is composed of *n* number of attributes, $A_1, A_2,..., A_n$. Every record in the dataset is associated with a class identifier or label. These records are represented by a vector, $\boldsymbol{X} = (x_1, x_2,...,x_n)$.

2) Given an instance ***X***, compute for the probability of every attribute value in ***X*** from all classes in the training set. Class prediction for ***X*** is performed in reference to classes with leading posterior probability.

   Consider that there are only 2 classes in the dataset, 1 and 2. Eq. 1 must be expanded, where $i = 1, 2$ based from the frequency of classes.

$$P(X|C_i)P(C_i) \qquad [7]\ (1)$$

   As explained by [7], the probability of instance ***X*** across classes is unchangeable, only the probability of ***X*** for a particular class must multiplied to the probability of the same class across the training set should expanded. This process should be repeated to each class within the training set.

3) The probabilities of the attribute values belonging to every class in the training set can be estimated. Equation 2.1 presents how these attribute probability values can be use to determine the class label of an instance, ***X***. Equation 2.2 elaborates the probability estimation in equation 2.1

$$P(X|C_i) = \prod_{k-1}^{n} P(x_k|C_i) \qquad [7]\ (2.1)$$

$$P(X|C_i) = P(x_1|C_i)\times P(x_2|C_i)\times\ldots\times P(x_n|C_i) \qquad [7]\ (2.2)$$

4) When computing for the probability of an attribute value, evaluate first if it is continous or categorical. To get the probability of a continous valued attribute, first compute the mean μ and standard deviation σ of the values under a particular attribute for every class in the training set. Afterwards, equation 3, Gaussian Distribution will assume these value, mean, standard deviation and the given attribute value of an instance so that the probability of the attribute value is generated, as presented in equation 4.

$$g(x,\mu,\sigma)=\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad [7]\ (3)$$

$$P(x_k|C_i)=g(x_k,\mu c_i,\sigma c_i) \qquad [7]\ (4)$$

5) To determine the class of instance ***X***, equation 1 is applied for all classes. The class with the highest probability value is considered for class label prediction.

RESULTS AND DISCUSSIONS

I. Plant Leaf Dataset

The leaf dataset from [13] was used to train and test NB classifier. The dataset is composed of 340 instances and 30 classes. Table 1 shows that there is an average of 11.33 instances for each class. It is notable that the frequency distribution of class instances is balance. The standard deviation is 1.6470 among the frequency distribution of class instances. It means that frequency distribution of class instances is closely concentrated around the mean due to small value of the standard deviation.

978-1-5090-0360-0/15/$31.00 ©2015 IEEE

TABLE I
PLANT DATABASE: PLANT SPECIES (CLASS) AND NUMBER OF SPECIMENS

Class	Scientific Name	No. of Specimen
1	Acca-sellowiana	11
2	Acer-palmatum	16
3	Alnus-sp	8
4	Arisarum-vulgare	9
5	Betula-pubescens	14
6	Bougainvillea-sp	13
7	Buxus-sempervirens	12
8	Castanea-sativa	12
9	Celtis-sp	12
10	Corylus-avellana	13
11	Crataegus-monogyna	8
12	Erodium-sp	11
13	Euonymus-japonicus	12
14	Geranium-sp	10
15	Hydrangea-sp	11
16	Ilex-aquifolium	10
17	Ilex-perado-ssp-azorica	11
18	Magnolia-grandiflora	11
19	Magnolia-soulangeana	12
20	Nerium-oleander	11
21	Podocarpus-sp	11
22	Populus-alba	10
23	Populus-nigra	10
24	Primula-vulgaris	12
25	Pseudosasa-japonica	11
26	Quercus-robur	12
27	Quercus-suber	12
28	Salix-atrocinera	10
29	Tilia-tomentosa	13
30	Urtica-dioica	12
	Average	11.33

TABLE II
ACCURACY EVALUATION RESULT

Class	TP Rate	FP Rate	ROC Area
Acca-sellowiana	0.545	0.012	0.98
Acer-palmatum	0.938	0.003	1.0
Alnus-sp	0.25	0.021	0.949
Arisarum-vulgare	0.889	0	1.0
Betula-pubescens	0.643	0.009	0.985
Bougainvillea-sp	0.615	0.024	0.965
Buxus-sempervirens	0.917	0.003	0.999
Castanea-sativa	0.5	0.006	0.92
Celtis-sp	0.75	0.024	0.981
Corylus-avellana	0.769	0.012	0.987
Crataegus-monogyna	0.875	0.003	1.0
Erodium-sp	1.0	0	1.0
Euonymus-japonicus	0.583	0.006	0.955
Geranium-sp	1.0	0.003	0.999
Hydrangea-sp	0.636	0.009	0.983
Ilex-aquifolium	0.6	0.018	0.987
Ilex-perado-ssp-azorica	0.636	0.012	0.963
Magnolia-grandiflora	0.636	0.009	0.966
Magnolia-soulangeana	0.583	0.024	0.973
Nerium-oleander	1.0	0	1.0
Podocarpus-sp	0.636	0.003	0.997
Populus-alba	0.9	0	0.999
Populus-nigra	0.8	0.006	0.992
Primula-vulgaris	0.333	0.015	0.978
Pseudosasa-japonica	0.909	0.012	0.996
Quercus-robur	0.917	0	1.0
Quercus-suber	0.833	0.006	0.97
Salix-atrocinera	0.6	0.018	0.917
Tilia-tomentosa	0.923	0.006	0.999
Urtica-dioica	0.917	0	0.999
Weighted Average	0.741	0.009	0.981

II. Accuracy Performance Evaluation Result

10 folds cross validation method were used to measure the accuracy performance of naïve Bayesian classifier in plant leaf classification. As explained by [7], the dataset is divided into 10 smaller subsets with the same number of instances. The training and testing was repeatedly executed 10 times. This method works as follows:

1) During the initial execution, the first subset was used as test set and the remaining subset was used as training set for naïve Bayes classification model.
2) In the second execution, the second subset was used as testing set and the first, third and succeeding subsets was used as training set.
3) This incrementing process was continuously repeated until 10th subset and on the 10th times [8].
4) After the cross validation, the classification accuracy is estimated based on the overall number of correctly classified instances and divided by the total number of instances from the initial dataset. The 10 folds cross validation method was used in this study because it is relatively low bias and variance [8].

III. ROC Curve Analysis Result

(*Receiver Operating Characteristic*) curve analysis was performed. It is a graphing technique for visualizing, organizing and selecting classifiers based on their performance. ROC graphs have long been used in signal detection theory to depict the trade-off between the true positive rate and the false positive rate of classifiers [5], [14].

To determine the rate at which naïve Bayesian classifier can accurately recognize the class of the test data, ROC

In ROC curve analysis, the accuracy performance of the classifier can be categorized into five classifications based on the measurement of the area under the ROC curve. These classifications are patterned from the traditional academic system which are presented as follows, (i) 0.90 to 1.0 is excellent, (ii) 0.80 to 0.90 is good, (iii) 0.70 to 0.80 is fair,

978-1-5090-0360-0/15/$31.00 ©2015 IEEE

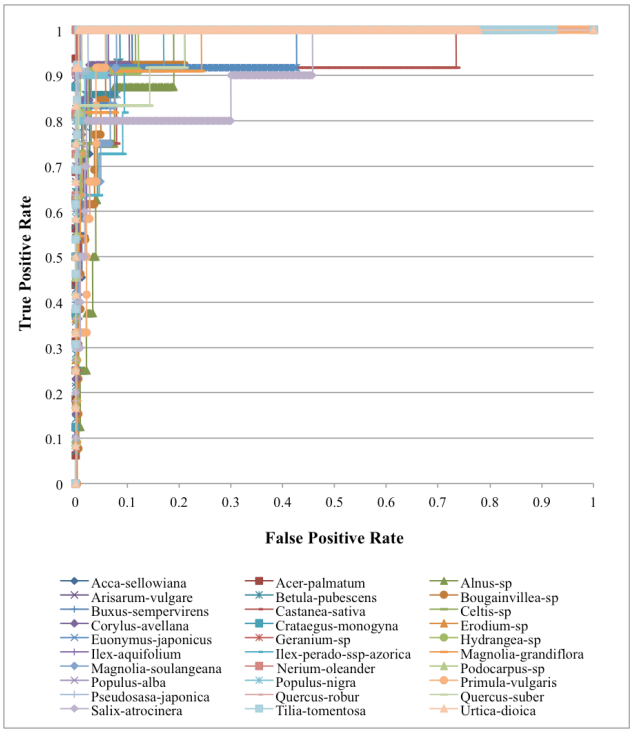

FIGURE 1. ROC Curve Analysis Result

(iv) 0.60 to 0.70 is poor, and (v) 0.50 to 0.60 is fail. This measurement indicates that when the area under the ROC curve is close to 0.50, the classifier becomes less accurate while the classifier with perfect accuracy will have an area of 1.0 [7].

Figure 2 shows the graph that represents the ROC curve for all classes of the plant leaf classifier that is based from naïve Bayesian classification method. In general, the classifier's accuracy performance is interesting because the ROC graph is inclined more towards the y-axis that represents the true positive rate.

It can also be observed in Table 2 that the weighted average of the ROC area for naïve Bayes classifier is 0.981. It is nearly closed to perfect accuracy performance. In addition to this, the standard deviation of the ROC area is 0.023. It indicates that there is a high concentration of ROC area among all classes around the weighted average presented above.

Furthermore, 6 classes have an ROC area of 1.0 namely, Acer-palmatum, Arisarum-vulgare, Crataegus-monogyna, Erodium-sp, Nerium-oleander, and Quercus-robur. This indicates that the classifier did not commit any classification error during the testing process.

Classes	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	*Total*	*Recognition Rate (%)*
1	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	3	0	0	11 (3.24%)	54.50
2	0	15	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	16 (4.71%)	93.80
3	0	0	2	0	0	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8 (2.35%)	25.00
4	0	0	0	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	9 (2.65%)	88.90
5	0	0	1	0	9	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	2	0	0	0	0	0	1	0	14 (4.12%)	64.30
6	0	0	3	0	1	8	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	13 (3.82%)	61.50
7	0	0	0	0	0	1	11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	12 (3.53%)	91.70
8	1	0	0	0	0	0	0	6	0	0	0	0	0	0	0	0	0	1	3	0	0	0	0	1	0	0	0	0	0	0	12 (3.53%)	50.00
9	0	0	0	0	0	0	0	0	9	0	0	0	0	0	0	0	0	0	1	0	0	0	0	2	0	0	0	0	0	0	12 (3.53%)	75.00
10	0	0	0	0	0	0	0	0	0	10	0	0	0	0	2	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	13 (3.82%)	76.90
11	0	1	0	0	0	0	0	0	0	0	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8 (2.35%)	87.50
12	0	0	0	0	0	0	0	0	0	0	0	11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	11 (3.24%)	100.00
13	0	0	0	0	0	0	0	0	0	2	0	0	7	0	0	2	0	0	0	0	0	0	0	0	0	0	0	1	0	0	12 (3.53%)	58.30
14	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10 (2.94%)	100.00
15	0	0	0	0	0	0	0	0	1	0	0	0	0	0	7	0	2	0	1	0	0	0	0	0	0	0	0	0	0	0	11 (3.24%)	63.60
16	0	0	1	0	0	0	0	0	2	0	0	0	0	0	0	6	0	0	0	0	0	0	0	0	0	0	0	1	0	0	10 (2.94%)	60.00
17	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	7	0	0	0	0	0	0	0	0	0	1	1	0	0	11 (3.24%)	63.60
18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7	3	0	0	0	0	1	0	0	0	0	0	0	11 (3.24%)	63.60
19	0	0	0	0	0	0	0	1	2	0	0	0	0	0	0	0	0	2	7	0	0	0	0	0	0	0	0	0	0	0	12 (3.53%)	58.30
20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	11	0	0	0	0	0	0	0	0	0	0	11 (3.24%)	100.00
21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7	0	0	0	4	0	0	0	0	0	11 (3.24%)	63.60
22	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	9	0	0	0	0	0	0	0	0	10 (2.94%)	90.00
23	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8	0	0	0	0	0	0	0	10 (2.94%)	80.00
24	0	0	0	0	0	0	0	1	4	0	0	0	0	0	0	3	0	0	0	0	0	0	0	4	0	0	0	0	0	0	12 (3.53%)	33.30
25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	10	0	0	0	0	0	11 (3.24%)	90.90
26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	11	0	0	0	0	12 (3.53%)	91.70
27	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	12 (3.53%)	83.30
28	3	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	6	0	0	10 (2.94%)	60.00
29	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	12	0	13 (3.82%)	92.30
30	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	11	12 (3.53%)	91.70
Total	10	16	9	8	12	16	12	8	17	14	8	11	9	11	10	12	11	10	15	11	8	9	10	9	14	11	12	12	14	11	340	74.10

FIGURE 2. Confusion Matrix of Naive Bayesian for Plant Leaf Classification

978-1-5090-0360-0/15/$31.00 ©2015 IEEE

Table 2 also shows that the true positive rate or recognition rate is higher in many instances and false positive rate is generally minimal having a weighted average of 0.9%.

CONCLUSION

In this study, the researchers applied naïve Bayesian method in plant classification based on leaf shape and textures as input features in order to produce model classifier that is accurate and efficient. The dataset with 30 different plant species was used to construct and train the classifier. The accuracy of the classifier was also tested using the testing dataset. The test results showed that the naïve Bayesian classifier was able to produce an accurate classification result compared to other classification methods conducted in the previous studies.

REFERENCES

[1] A. Bhardwaj *et al*., "Recognition of plants by Leaf Image using Moment Invariant and Texture Analysis," *Int. J. of Innovation and Applied Studies*, vol. 3, no. 1, pp. 237-248, May 2013.

[2] D. Bhattacharyya *et al*., "Leaf Image Analysis towards Plant Identification," in *Signal Process., Image Process. and Pattern Recognition*, vol. 260, pp. 113-125, 2011 © Springer-Verlag GmbH Berlin Heidelberg. doi: 10.1007/978-3-642-27183-0

[3] J. Chaki and R. Parekh, "Plant Leaf Recognition using Shape based Features and Neural Network classifiers," *Int. J. of Advanced Comput. Sci. and Applicat.*, vol. 2, no. 10, 2011. doi: 10.14569/IJACSA.2011.021007

[4] J. X. Du *et al.*, "Recognition of Leaf Image Based on Outline and Vein Fractal Dimension Feature," in *Advanced Intelligent Computing*, vol. 6838, pp. 364-369, 2011 © Springer-Verlag GmbH Berlin Heidelberg. doi: 10.1007/978-3-642-24728-6_49

[5] T. Fawcett, "ROC Graphs: Notes and Practical Considerations for Data Mining Researchers," HP Labs., Palo Alto, CA, Tech. Rep. HPL-2003-4, Jan. 7, 2003.

[6] M. Hall *et al.,* "The WEKA Data Mining Software: An Update," *ACM SIGKDD Explorations Newslett.*, vol. 11, no. 1, pp. 10-18, June 2009. doi: 10.1145/1656274.1656278

[7] J. Han and M. Kamber, "Classification and Prediction" in *Data Mining: Concepts and Techniques,* 2nd ed. San Francisco, CA: Morgan Kaufmann, 2006, ch. 6, sec. 6.4, pp. 310-315.

[8] N. Kumar *et al.*, "Leafsnap: A Computer Vision System for Automatic Plant Species Identification," in *Computer Vision–ECCV 2012*, pp. 502-516, 2012 © Springer-Verlag Berlin Heidelberg. doi: 10.1007/978-3-642-33709-3_36

[9] A. M. Morrison, "Receiver Operating Characteristic (ROC) Curve Preparation - A Tutorial," MWRA, Boston, MA, Tech. Rep. 2005-20, 2005.

[10] Y. Shen *et al.*, "Leaf Image Retrieval Using a Shape Based Method," in *Artificial Intelligence Applications And Innovations*, vol. 187, pp. 711-719, 2005 © Springer US. doi: 10.1007/0-387-29295-0_77

[11] P. F. Silva, "Development of a System for Automatic Plant Species Recognition," M.S. thesis, Dept. Math., Univ. Porto, Portugal, 2013.

[12] P. F. Silva *et* al., "Evaluation of Features for Leaf Discrimination," in *Image Analysis and Recognition*, vol. 7950, pp. 197-204, 2013 © Springer-Verlag Berlin Heidelberg. doi: 10.1007/978-3-642-39094-4_23

[13] P. F. Silva *et al.* (2014, February 24). *Leaf Data Set* [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Leaf

[14] J. Swets. (1988). *Measuring the accuracy of diagnostic systems* [Online]. Available FTP: norbif.uio.no Directory: pub/outgoing/runeho/KR File: Swet88Science240-1285.pdf

[15] B. Vijayalakshmi, "A New Shape Feature Extraction Method for Leaf Image Retrieval," in *Proc. 4th Int. Conf. on Signal and Image Process*., vol. 221, pp. 235-245, 2013 © Springer India. doi: 10.1007/978-81-322-0997-3_22

[16] I. Witten and E. Frank,"The Weka machine learning workbench" in *Data Mining: Practical Machine Learning Tools and Techniques, 2nd ed*. San Francisco, CA: Morgan Kaufmann, 2005, ch. 10, sec. 10.1-10.4, pp. 369-414.

[17] S. G. Wu *et al.*, "A Leaf Recognition Algorithm for Plant Classification Using Probabilistic Neural Network," *in 2007 IEEE Int. Symp. Signal Process. and Inform. Technology,* Giza, Egypt, 2007 © IEEE, pp. 11-16. doi: 10.1109/ISSPIT.2007.4458016

[18] United Nations Environment Program. (2014). *UNEP Annual Report 2014* [Online]. Available: http://www.unep.org/annual report/2014/en/ecosystem-management.html

978-1-5090-0360-0/15/$31.00 ©2015 IEEE