Contents lists available at ScienceDirect

Artificial Intelligence in Agriculture

journal homepage: http://www.keaipublishing.com/en/journals/artificial-intelligence-in-agriculture/

Identification of maize (*Zea mays* L.) progeny genotypes based on two probabilistic approaches: Logistic regression and naïve Bayes

D. Seka [a,*], B.S. Bonny [a], A.N. Yoboué [b], S.R. Sié [a], B.A. Adopo-Gourène [a]

[a] *School of Natural Sciences, UFR-SN, University Nangui Abrogoua, 02 B.P. 801, Abidjan 02, Cote d'Ivoire*
[b] *School of Environmental Sciences, Jean Lorougnon Guédé University, B.P. 150, Daloa, Cote d'Ivoire*

ARTICLE INFO

Article history:
Received 11 November 2018
Received in revised form 12 March 2019
Accepted 12 March 2019
Available online 21 March 2019

Keywords:
Gaussian naïve Bayes
Logistic regression
Maize genotype
Prediction
Selection

ABSTRACT

We used two probabilistic methods, Gaussian Naïve Bayes and Logistic Regression to predict the genotypes of the offspring of two maize strains, the BLC and the JNE genotypes, based on the phenotypic traits of the parents. We determined the prediction performance of the two models with the overall accuracy and the area under the receiver operating curve (AUC). The overall accuracy for both models ranged between 82% and 87%. The values of the area under the receiver operating curve were 0.90 or higher for Logistic Regression models, and 0.85 or higher for Gaussian Naïve Bayes models. These statistics indicated that the two models were very effective in predicting the genotypes of the offspring. Furthermore, both models predicted the BLC genotype with higher accuracy than they did the JNE genotype. The BLC genotype appeared more homogeneous and more predictable. A Chi-square test for the homogeneity of the confusion matrices showed that in all cases the two models produced similar prediction results. That finding was in line with the assertion by Mitchell (2010) who theoretically showed that the two models are essentially the same. With logistic regression, each subset of the original data or its corresponding principal components produced exactly the same prediction results. The AUC value may be viewed as a criterion for parent-offspring resemblance for each set of phenotypic traits considered in the analysis.

1. Introduction

The development of the learning algorithms and their implementations have brought new ways to process data, to extract information from data, and to formulate a basis for decision. The application of these learning algorithms spans to all fields of human activities. The literature on the development and the use of these algorithms abounds. Hastie et al. (2001), and Webb (2002) provide a good overview of these algorithms. Sambo et al. (2012) developed an algorithm for genetic biomarker selection and subjects' classification from the analysis of genome-wide SNP data. They found a significantly higher classification accuracy with their algorithm in the identification of biomarkers for Type 1 diabetes. Douglass et al. (2016) used Naïve Bayes classifier with small RNA deep-sequencing data and genomic features to identify plant miRNAs in soybean, peach, rice and Arabidopsis. They found the classifier for all four plants to be highly accurate in the identification of miRNA. They based the performance of the classifier on the values of the area under the curve of the receiver operating characteristic (ROC) curve. They reported AUC values between 0.9771 and 0.9980, with the identification of Arabidopsis miRNAs having the highest value. These learning algorithms have extensively been used in the classification of plant organisms as well. In most cases, the identification of plants was based on leaf attributes because the shape, the color and the texture of a leaf are specific to the plant species (El-Hariri et al., 2014) and unlike flowers and fruits, the plant always carries functional leaves throughout its existence (Siravenha and Carvalho, 2015). Wu et al. (2007) developed an algorithm that uses probabilistic neural network with image and data processing techniques for an automated leaf recognition and plant classification. Its application resulted in the classification of 32 different plants based on leaf features, with an accuracy greater than 90%. They found their algorithm to be fast in execution, easy to implement, and efficient in plant recognition and classification. Hemming and Rath (2001) constructed an identification system that distinguished weeds from two major crops, cabbage and carrots. In their experiment, they initially used eight morphological features along with three color features, then conducted feature selection to determine the most relevant features for discriminating between weed species and crops. They found that color features helped to increase the classification accuracy. They reported that between 51% and 95% of the plants were correctly classified. The mean classification accuracy was 88.15% for cabbage and 72.46% for carrots. Giselsson et al. (2013) constructed a data set representing shape features by using distance data from images of seedling of *Centaurea cyanus* and *Solanum nigrum*. They fitted the distance data to a high degree Legendre polynomial and used the coefficients of the polynomial which they called

[*] Corresponding author.
E-mail address: seka.d@iugb.edu.ci (D. Seka).

https://doi.org/10.1016/j.aiia.2019.03.001

Legendre polynomial feature set. They collected another set of data and named it standard feature set. With the two set of data, they ran four classification algorithms. They reported that the Legendre polynomial feature set was very robust and its use with the classification algorithms resulted in an accuracy of 98.75% compared with the standard feature set which yielded an accuracy of 87.1%. However, they suggested more testing of this attribute data collection method in order to evaluate its true value. Siravenha and Carvalho (2015) explored a method of data collection on leaf features that involved the contour-centroid distance and data transformation with the Fast Fourier Transform. They applied feature-selection techniques for dimensionality reduction and improved classification accuracy. They then used various classification algorithms for plant identification. They reported classification accuracies ranging from 65.14% obtained with the C4.5 algorithm to 97.45% resulting from the Pattern Net algorithm applied to principal components. The learning algorithms combined with other statistical methods have significantly contributed in the advancement of phenotypic predictions of quantitative traits of sequenced whole genome. Guzzetta et al. (2010) explained a learning pipeline that use the L1L2 regularization method based on the naive elastic net for the prediction of phenotypes. They found it to be very effective regarding the accuracy of phenotype prediction. Zhao et al. (2016) used machine learning methods to identify root traits that led to the differentiation of cultivars in a binomial setting. And Lippert et al. (2017) developed models for the prediction of phenotypic traits. They combined those models in a single machine learning model for genome re-identification from the phenotypic prediction. In all the studies, different learning methods have been used and in each case the performance of the method has proven to be very satisfactory to excellent in prediction or in classification. Kell et al. (2001) mentioned that all machine learning methods have their strength and weaknesses and no single method works best for all situations.

Given the acclaimed high performance of these algorithms in the identification and classification of the subjects under studies, we conducted this study to predict plant genotypes based on parents' phenotypic traits. This study also provides a numerical proof of the elegant demonstration by Mitchell (2010) who showed that naïve Bayes with continuous features and logistic regression are essentially the same, even though one is generative and the other is discriminative (Ng and Jordan, 2001).

2. Materials and methods

We used two maize (*Zea mays* L.) strains, BLC and JNE, in this study. The two strains were different by the color of their kernels. Kernels from BLC were translucent giving the appearance of white kernels. Kernels from JNE were yellow. In this study, the BLC strain underwent three cycles of selection for longer and wider leaves, a larger number of leaves and taller plants while the JNE strain was subjected to random selection after each cycle. Seed from selected plants of each strain were obtained by pollination with a bulk of pollen from several male inflorescences of the same strain. The controlled pollination was insured by the protection of ear shoot with paper bag soon after it emerged.

In spring 2017, seed from the third generation (parents in this study) and the fourth generation (offspring) of the two genotypes were planted in a completely randomized experiment with two factors (generation and strain) at the experimental station of University Nangui Abrogoua, Abidjan, Cote d'Ivoire. We added fertilizer to the soil. Each experimental unit consisted of a row 4 m long of either one of the two strains. Rows were separated by 0.75 m. The plots were overseeded, and three weeks after planting, were thinned to 42,000 plants ha^{-1}. At silking, the day when about 50% of the plants in a plot developed silks, we determined the number of leaves (ln) by counting nodes, we measured ear leaf length (*ll*) and maximum ear leaf width (*lw*) as suggested by Cross (1991). We measured plant height (*ht*) as the length of the plant from the base at the soil level to the tip of the male inflorescence. We determined the rate of grain filling (*gf*) as the linear coefficient from the orthogonal contrast (Carlone and Russell, 1987) on the sequential sampling of kernel dry weight during the linear phase of the grain filling period.

In predicting the genotypes of the offspring, we used two classification methods: the Gaussian naïve Bayes and the logistic regression. We used the naïve Bayes algorithm to determine the posterior probability of an individual progeny to be a descendant of either the BLC or the JNE strain given its attribute values of plant height, leaf number, leaf width, number of leaves and rate of grain filling. The probabilistic model is simplified as follows:

$$\Pr(Class|A_1, A_2, \ldots, A_5) = \Pr(Class) * \prod_{i=1}^{5} \Pr(A_i|Class)$$

where *Class* is the class label with outcomes BLC or JNE, and A_i is the attribute. The model computes the posterior probability of a progeny, with respect to its lineage, as the product of the prior probability of the class label and the conditional probabilities of the attributes with respect to that class label. The application of this model in the prediction or the classification of subjects in a population is said to be discriminative as opposed to logistic regression that is generative (Ng and Jordan, 2001).

The measures on the parents of BLC and JNE genotypes were also used to fit a logistic regression model (McCullagh and Nelder, 1989) and to determine the parameters of the model in order to compute the posterior probabilities of the progeny to belong to the JNE or the BLC genotypes.

$$\Pr(Class|A_1, A_2, \ldots, A_5)$$

The parametric model assumed by logistic regression in the case where the class variable was binomial is given as follows (Mitchell, 2010):

$$\Pr\left(Class = {''}\mathrm{BLC}{''} \middle| A_1, A_2, A_3, A_5\right) = \frac{1}{1 + \exp\left(b_0 + \sum_{i=1}^{5} b_i A_i\right)}$$

and

$$\Pr\left(Class = {''}\mathrm{JNE}{''} |A_1, A_2, \ldots, A_5\right) = \frac{\exp\left(b_0 + \sum_{i=1}^{5} b_i A_i\right)}{1 + \exp\left(b_0 + \sum_{i=1}^{5} b_i A_i\right)}.$$

Mitchell (2010) showed that the parametric form of Pr(*Class*|A_1,A_2,…A_5) used by logistic regression is precisely the form implied by the assumptions of a Gaussian naïve Bayes classifier, and in most cases both methods produce similar results. However, it should be noted that logistic regression directly estimates the parameters of Pr(*Class*|A_1,A_2,…A_5), whereas Gaussian naïve Bayes directly estimates parameters of Pr(*Class*) and Pr(A_1,A_2,…A_5|*Class*) (Ng and Jordan, 2001).

To test the assertion by Mitchell (2010) that Gaussian naïve Bayes and logistic regression produce the same results, we created several subsets and their principal components from the original data by iteratively removing attributes. We then predicted the progeny with each of the two models based on the retained attributes or the corresponding principal components of each subset. The principal components are an orthogonal transformation of the original observations that inherit the total variability. Their use should remove any bias associated with the assumption of independence between attributes in comparison to the use of the original variables and should serve to assess any such bias in the prediction results. We performed an importance ranking (Kuhn et al., 2017) of the traits with respect to their ability to discriminate between the two classes. We computed statistics such as accuracy, sensitivity, specificity (Tze-Wey, 2003) and the maximum value of the area under the receiver operating characteristic curve (Bradley, 1997) for each model. We assessed the performance of each model with the accuracy and the AUC value. In running the two models, the BLC genotype was considered the positive class. We used the statistical software R (R Core Team, 2017) for all the analyses.

3. Results and discussion

The variable importance ranking technique (Kuhn et al., 2017) identified *lw*, *gf* and *ht* as the most important traits among the attributes of the data set in their ability to discriminate between the two genotypes. Their importance values were, respectively, 0.85, 0.82 and 0.72 on a 1-point scale. That finding was corroborated by the analysis of the logistic regression model that showed *lw* ($p\text{-}value = 0.03888$), *gf* ($p\text{-}value = 0.04667$), and *ht* ($p\text{-}value = 0.04942$) as the most significant predictor variables in the data set. Table 1 shows the means and standard deviations of the 5 attributes. The trait ln showed little variability across generations and genotypes. It was also ranked last in its ability to differentiate the two genotypes with an importance value of 0.56.

We predicted the progeny of the two parents using different subsets of the original data and their principal components. We reported the performances of the two models on those subsets in Table 2. Comparison of the confusion matrices generated from the models on each subset was done with a Chi-square test (Agresti, 2007; Hope, 1968). The test of homogeneity of the confusion matrices produced *p*-values that ranged from 0.7692 to 1.000, proving the close similarity of the prediction performances of the two models. Gaussian naïve Bayes updates prior knowledge from the parents with current evidence of the offspring to compute the conditional probability that an offspring belongs to a genotype. And, logistic regression estimates the parameters of the model with the parents' phenotypic values to compute the same conditional probabilities. However, Mitchell (2010) gave an elaborate and elegant demonstration that under some conditions of binary response variable (Y) with parameter $\pi = \Pr\ (Y = \text{"positive class"})$, Gaussian distributed attributes (A_i) that are conditionally independent with respect to Y, Pr ($A_i|Y = y_k$) ~ N(μ_{ik},σ_i), the conditional probabilities under the Gaussian naïve Bayes model can be written in parametric forms as

$$\Pr\left(Class = \text{"BLC"}\,\middle|A_1, A_2, \ldots A_5\right) = \frac{1}{1+\ \exp\left(w_0 + \sum_{i=1}^{5} w_i A_i\right)}$$

and

$$\Pr\left(Class = \text{"JNE"}\,|A_1, A_2, \ldots A_5\right) = \frac{\exp\left(w_0 + \sum_{i=1}^{5} w_i A_i\right)}{1+\ \exp\left(w_0 + \sum_{i=1}^{5} w_i A_i\right)}$$

where

$$w_i = \frac{\mu_{i0} - \mu_{i1}}{\sigma_i^2}$$

and

$$w_0 = Ln\frac{1-\pi}{\pi} + \sum_i \frac{\mu_{i1}^2 - \mu_{i0}^2}{2\sigma_i^2}.$$

Table 1
Mean and standard deviation of *lw*, *ll*, ln, *ht* and *gf* of the parents and the offspring.

Trait	Parent		Offspring	
	BLC	JNE	BLC	JNE
lw (cm)	9.66 (0.61)	8.34 (1.18)	9.47 (0.67)	8.93 (0.82)
ll (cm)	89.30 (6.47)	84.34 (7.36)	89.90 (5.23)	85.16 (5.86)
ln (cm)	12.43 (1.61)	12.13 (1.11)	12.73 (1.05)	12.20 (1.19)
ht (cm)	235.87 (23.59)	216.57 (24.32)	231.50 (21.12)	218.40 (21.12)
gf ($\text{mg}\cdot\text{d}^{-1}$)	5.41 (0.32)	4.98 (0.34)	5.49 (0.30)	5.10 (0.40)

These probabilities are exactly the same given by logistic regression when the response variable has binary outcomes and the predictor variables are Gaussian distributed. The results of this experiment provide a numerical evidence of that demonstration produced by Mitchell (2010), and simulation studies by Bhowmik (2015). We should mention that logistic regression directly estimates the parameters for Pr($Class|A_i$), and Gaussian naïve Bayes directly estimates the parameters for Pr($Class$) and Pr($A_i|Class$) (Ng and Jordan, 2001).

Table 2 shows that the logistic regression gave exactly the same prediction results whether we used subsets of the original data or their corresponding principal components. This observation finds its explanation in the fact that the right hand-side of the logistic equation with the original data is exactly the same as the right hand-side of the logistic equation with the principal components,

$$\frac{1}{1+\ \exp\left(\hat{\beta}_{obs}^{t}\mathrm{A}\right)} = \frac{1}{1+\ \exp\left(\hat{\beta}_{pc}^{t}\mathrm{Z}\right)}$$

which leads to

$$\hat{\beta}_{obs}^{t}\mathrm{A} = \hat{\beta}_{pc}^{t}\mathbf{Z}$$

where $\hat{\beta}_{obs}^{t}$ is the vector of estimated regression coefficients from the original data, A is the matrix formed with the subset of the original variables, $\hat{\beta}_{pc}^{t}$ sis the vector of estimated regression coefficients from the principal components, and Z is the matrix of principal components derived from the subset of the original data.

Therefore,

$$\Pr\left(Class = \text{"BLC"}\,\middle|\mathrm{A}, \beta_{obs}\right) =\ \Pr\left(Class = \text{"BLC"}\,\middle|\mathrm{Z}, \beta_{pc}\right)$$

and

$$\Pr\left(Class = \text{"JNE"}\,|\mathrm{A}, \beta_{obs}\right) =\ \Pr\left(Class = \text{"JNE"}\,|\mathrm{Z}, \beta_{pc}\right).$$

With Gaussian naïve Bayes, the confusion matrices obtained from the subsets of the original data and their corresponding principal components were not always exactly the same. But the differences were not significant (*p*-values ≥ 0.8227).

The two models yielded very good prediction performances. The overall prediction accuracy ranged from 82% to 87% for all data tested (Table 2). The values of the area under the curve were between 0.85 and 0.93. In general, both model predicted the BLC genotypes with higher accuracy than they predicted the JNE genotypes. The relatively low specificity and high sensitivity observed with the two models may be linked to the structures of the BLC and JNE populations and how they were developed. The BLC population was developed with a selection procedure directed to wider and longer leaves, a larger number of leaves and taller plants. Three cycles of selection may have contributed to the partial accomplishment of that goal. Except for the trait ln, more individuals of the BLC genotypes have developed the distinctive traits defined by the selection criteria and have progressively formed a group that was more homogenous and that was better predicted by the models. Kudaravalli et al. (2009), Bersaglieri et al. (2004), Enattah et al. (2005), and Tishkoff et al. (2007) reported corroborating evidence of the effect of selection on a trait. They proved that natural or artificial selection affected the expression of the gene(s) that translated into a modified phenotype. And Wang et al. (2018) found that the maize leaf width is controlled by dominant genes not linked with gene

Table 2
Prediction characteristics of the logistic regression (LR) and the Gaussian naïve Bayes (GNB) models with subsets of the original data and their principal components.

Data	Attribute	Model	Confusion Matrix				AUC	Accuracy	Chi-square (p-value)
			TP	FN	TN	FP			
Original data	(a) $lw + ll + \text{ln} + ht + gf$	LR	26	4	23	7	0.90	0.82	0.9674
		GNB	27	3	24	6	0.90	0.85	
	(b) $lw + gf + ht + ll$	LR	27	3	22	8	0.93	0.82	1.000
		GNB	27	3	22	8	0.91	0.82	
	(c) $lw + gf + ht$	LR	27	3	22	8	0.93	0.82	0.9931
		GNB	27	3	23	7	0.90	0.83	
	(d) $gf + ht$	LR	26	4	26	4	0.93	0.87	0.8868
		GNB	24	6	27	3	0.92	0.85	
	(e) $lw + gf$	LR	27	3	23	7	0.91	0.83	0.9835
		GNB	26	4	23	7	0.89	0.82	
Principal component	$\text{pc}_1 + \text{pc}_2 + \text{pc}_3 + \text{pc}_4 + \text{pc}_5$ from (a)	LR	26	4	23	7	0.90	0.82	0.9931
		GNB	26	4	22	8	0.85	0.80	
	$\text{pc}_1 + \text{pc}_2 + \text{pc}_3 + \text{pc}_4$ from (b)	LR	27	3	22	8	0.93	0.82	0.8227
		GNB	28	2	19	11	0.86	0.78	
	$\text{pc}_1 + \text{pc}_2 + \text{pc}_3$ from (c)	LR	27	3	22	8	0.93	0.82	1.000
		GNB	27	3	22	8	0.90	0.82	
	$\text{pc}_1 + \text{pc}_2$ from (d)	LR	26	4	26	4	0.93	0.87	0.7692
		GNB	23	7	25	5	0.90	0.80	
	$\text{pc}_1 + \text{pc}_2$ from (e)	LR	27	3	23	7	0.91	0.83	0.9690
		GNB	26	4	22	8	0.90	0.80	

TP = true positive; FN = false negative; TN = true negative; and FP = false negative.
Note: pc_i ($i = 1, 2, \ldots, 5$) are the principal components.

(s) controlling maize leaf length. That assertion implies that the maize leaf width could be modified through selection without any direct effect on other traits of the canopy. The results of this study support that assertion. In this study, the comparative increase in maize leaf width and plant height of the BLC genotype could be attributed to the direct effect of selection for wider leaves and taller plants. As a result, the two traits became the two most important discriminating traits between the two populations. The higher grain filling rate observed with the BLC genotypes compared with the JNE genotypes may be attributed to an indirect selection for size of canopy. In the other hand, the JNE genotypes were randomly selected. The measured traits consequently had larger variability, and the individuals of the JNE genotypes were relatively less homogeneous. The relatively larger dispersion that characterized the JNE strains resulted in the reduced specificity of the models. Many individuals of the JNE strains had traits similar to the traits that characterized most of the BLC genotypes and the models identified them as belonging to the class of the BLC genotype. However, the majority of the individuals of the JNE genotype conserved their traits and the models correctly predicted those individuals.

4. Summary and conclusion

We used field data to predict progeny genotype of two maize strains using logistic regression and Gaussian naïve Bayes. The overall prediction accuracy ranged from 78% to 87% and the values of the area under the receiver operating characteristic curve were between 0.85 and 0.93. Sensitivity, defined as the correct identification of the BLC genotype was high with a modal value of 0.90 for both Gaussian naïve Bayes and logistic regression. On the other hand, specificity which is the correct identification of the JNE genotype had a modal value of 0.73 for both Gaussian naïve Bayes and logistic regression. The observed higher sensitivity and lower specificity of the two tests were due to the structures of the two populations more than the quality of the tests. The BLC population was more homogeneous and much easily identifiable compared to the JNE population which was more diverse and a high proportion of its progeny was misclassified as BLC genotypes.

The tests of prediction of offspring's genotypes was more a test of parent-offspring resemblance. The models trained with parents' phenotypic values and identified progenies having similar traits. With a single predictor variable or attribute, the area under the receiver operating characteristic curve from such tests may be a good predictor of heritability. Wray et al. (2010) formulated an equation that related maximum area under the receiver operating characteristic curve to heritability and the prevalence of the trait. And Dreyfuss et al. (2012) mentioned that the accuracy of a test was a direct reflect of the heritability of the trait being considered in that test. They added that a phenotypic trait with high heritability would be predicted with high accuracy whereas a trait with low and non-significant heritability would not result in a high prediction accuracy. Such trait was less influenced by genetic factors and more by environmental factors. In this study, several phenotypic traits were considered in the determination of offspring's genotype. We could not infer that the AUC value was a good estimator of heritability. Instead, we may view the AUC value in this study as a good indicator, or criterion, of parent-offspring resemblance. And the higher the AUC value, the greater the resemblance between parents and offspring.

We used several subsets of the data and their corresponding principal components to conduct the genotype-prediction tests. For each data set, the Chi-square test comparing the prediction results of the logistic regression and Gaussian naïve Bayes showed that the two models produced similar prediction performances. The results of our field study were consistent with the finding by Mitchell (2010) and the work by Bhowmik (2015). With logistic regression, the use of the subset of the data or its corresponding principal components gave exactly the same prediction results. The principal components retained the total variability of the data and the product of the data matrix with the vector of estimated regression coefficients from a subset was the same as the product of the matrix of principal components with the vector of estimated coefficients from the principal components. With Gaussian naïve Bayes, the prediction results with a subset of the data or its corresponding principal components were not always exactly the same. But where there is a difference, the difference was not significant.

Conflict of interest

On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

Agresti, A., 2007. An Introduction to Categorical Data Analysis. 2nd ed. John Wiley & Sons, New York, p. 38.

Bersaglieri, T., Sabeti, P.C., Patterson, N., Vanderploeg, T., Schaffner, S.F., Drake, J.A., Rhodes, M., Reich, D.E., Hirschhorn, J.N., 2004. Genetic signatures of strong recent positive selection at the lactase gene. Am. J. Hum. Genet. 74, 1111–1120.

Bhowmik, T.K., 2015. Naive Bayes vs logistic regression: theory, implementation and experimental validation. Intell. Artif. 18 (56), 14–30.

Bradley, A.P., 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognit. 30, 1145–1159.

Carlone, M.R., Russell, W.A., 1987. Response to plant densities and nitrogen levels for four maize cultivars from different eras of breeding. Crop Sci. 27, 465–470.

Cross, H.Z., 1991. Leaf expansion rate effects on yield and yield components in early-maturing maize. Crop Sci. 31, 579–583.

Douglass, S., Hsu, S.W., Cokus, S., Goldberg, R.B., Harada, J.J., Pellegrini, M., 2016. A naïve Bayesian classifier for identifying plant microRNAs. Plant J. 86 (6), 481–492.

Dreyfuss, J.M., Levner, D., Galagan, J.E., George, M., Church, G.M., Ramoni, M.F., 2012. How accurate can genetic predictions be? BMC Genom. 13, 340.

El-hariri, E., El-Bendary, N., Hassanien, A.E., 2014. Plant classification system based on leaf features. Proceedings of 2014 9th IEEE International Conference on Computer Engineering and Systems. ICCES 2014 https://doi.org/10.1109/ICCES.2014.7030971.

Enattah, N., Pekkarinen, T., Välimäki, M.J., Löyttyniemi, E., Järvelä, I., 2005. Genetically defined adult-type hypolactasia and self-reported lactose intolerance as risk factors of osteoporosis in Finnish postmenopausal women. Eur. J. Clin. Nutr. 59, 1105–1111.

Giselsson, T.M., Midtiby, H.S., Jørgensen, R.N., 2013. Seedling discrimination with shape features derived from a distance transform. Sensors 13, 5585–5602.

Guzzetta, G., Jurman, G., Furlanello, C., 2010. A machine learning pipeline for quantitative phenotype prediction from genotype data. BMC Bioinforma. 11 (Suppl. 8), S3.

Hastie, T., Tibshirani, R., Friedman, J., 2001. The Elements of Statistical Learning. Springer, New York.

Hemming, J., Rath, T., 2001. Computer-vision-based weed identification under field conditions using controlled lighting. J. Agric. Eng. Res. 78 (3), 233–243.

Hope, A.C.A., 1968. A simplified Monte Carlo significance test procedure. J. Roy. Stat. Soc. B. 30, 582–598.

Kell, D.B., Darby, R.M., Draper, J., 2001. Genomic computing. Explanatory analysis of plant expression profiling data using machine learning. Plant Physiol. 126, 943–951.

Kudaravalli, S., Veyrieras, J.B., Stranger, B.E., Dermitzakis, E.T., Pritchard, J.K., 2009. Gene expression levels are a target of recent natural selection in the human genome. Mol. Biol. Evol. 26 (3), 649–658.

Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., Kenkel, B., the R Core Team, Benesty, M., Lescarbeau, R., Ziem, A., Scrucca, L., Tang, Y., Candan, C., Hunt, T., 2017. caret: Classification and Regression Training (R package version 6.0-78).

Lippert, C., Sabatini, R., Maher, M.C., Kang, E.Y., Lee, S., Arikan, O., Harley, A., Bernal, A., Garst, P., Lavrenko, V., Yocum, K., Wong, T., Zhu, M., Yang, W.-Y., Chang, C., Lu, T., Lee, C.W.H., Hicks, B., Ramakrishnan, S., Tang, H., Xie, C., Piper, J., Brewerton, S., Turpaz, Y., Telenti, A., Roby, R.K., Och, F.J., Venter, J.C., 2017. Identification of individuals by trait prediction using whole-genome sequencing data. PNAS 114 (38), 10166–10171.

McCullagh, P., Nelder, J., 1989. Generalized Linear Models. Chapman & Hall.

Mitchell, T.M., 2010. Generative and discriminative classifiers: naive Bayes and logistic regression. https://www.cs.cmu.edu/~tom/mlbook/NBayesLogReg.pdf.

Ng, A.Y., Jordan, M.I., 2001. On discriminative vs. generative classifiers: a comparison of logistic regression and naive Bayes. Adv. Neural Inform. Process. Syst. 14, 605–610.

R Core Team, 2017. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (URL).

Sambo, F., Trifoglio, E., Di Camillo, B., Toffolo, G.M., Cobelli, C., 2012. Bag of naive Bayes: biomarker selection and classification from genome-wide SNP data. BMC Bioinforma. 13 (Suppl. 14), S2.

Siravenha, A.C.Q., Carvalho, S.R., 2015. Exploring the use of leaf shape frequencies for plant classification. 2015 28th SIBGRAPI Conference on Graphics, Patterns and Images, Salvador, pp. 297–304.

Tishkoff, S.A., Floyd, A.R., Alessia, R., Voight, B.F., Babbitt, C.C., Silverman, J.S., Powell, K., Mortensen, H.M., Hirbo, J.B., Osman, M., Ibrahim, M., Omar, S.A., Lema, G., Nyambo, T.B., Ghori, J., Bumpstead, S., Pritchard, J.K., Wray, G.A., Deloukas, P., 2007. Convergent adaptation of human lactase persistence in Africa and Europe. Nat. Genet. 39 (1), 31–40.

Tze-Wey, L., 2003. Understanding sensitivity and specificity with the right side of the brain. BMJ 327 (7417), 716–719.

Wang, B., Zhu, Y., Zhu, J., Liu, Z., Liu, H., Dong, X., Guo, J., Li, W., Chen, J., Gao, C., Zheng, X., E, L., Lai, J., Zhao, H., Song, W., 2018. Identification and fine-mapping of a major maize leaf width QTL in a re-sequenced large recombinant inbred lines population. Front. Plant Sci. 9, 101.

Webb, A., 2002. Statistical Pattern Recognition. 2nd ed. Wiley, New York.

Wray, N.R., Yang, J., Goddard, M.E., Visscher, P.M., 2010. The genetic interpretation of area under the ROC curve in genomic profiling. PLoS Genet. 6 (2), e1000864.

Wu, S.G., Sheng, B.F., You, X.E., Wang, Y.X., Chang, Y.F., Xiang, Q.L., 2007. A leaf recognition algorithm for plant classification using probabilistic neural network. IEEE International Symposium on Signal Processing and Information Technology, pp. 11–16.

Zhao, J., Bodner, G., Rewald, B., 2016. Phenotyping: using machine learning for improved pairwise genotype classification based on root traits. Front. Plant Sci. 7, 1864.