Computers and Electronics in Agriculture 74 (2010) 250–257

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

journal homepage: www.elsevier.com/locate/compag

Original paper

Supervised machine learning and heterotic classification of maize (*Zea mays* L.) using molecular marker data

Leonardo Ornella[a,*], Elizabeth Tapia[a,b]

[a] CIFASIS-CONICET, Av. 27 de Febrero 210bis, Rosario, Argentina
[b] FCEIA-UNR, Department of Electronic, Riobamba 210 bis, Rosario, Argentina

ARTICLE INFO

Article history:
Received 1 October 2009
Received in revised form 11 March 2010
Accepted 27 August 2010

Keywords:
Maize
Supervised learning
Heterotic groups

ABSTRACT

The development of molecular techniques for genetic analysis has enabled great advances in cereal breeding. However, their usefulness in hybrid breeding, particularly in assigning new lines to heterotic groups previously established, still remains unsolved. In this work we evaluate the performance of several state-of-art multiclass classifiers onto three molecular marker datasets representing a broad spectrum of maize heterotic patterns. Even though results are variable, they suggest supervised learning algorithms as a valuable complement to traditional breeding programs.

1. Introduction

Since the first maize hybrid was bred and produced in USA, hybrid breeding has become one of the primary goals in any maize breeding programs (Hallauer and Miranda, 1988); however, varietal development has become more competitive and costly. For example, in USA, development of one variety of maize or soybean requires 0.5–7.0 million dollar. The lifetime of a variety is usually 3–6 years before it succumbs to the challenges of the production environment (biotic and abiotic stress) and demands of consumers (Lee, 1998). Consequently, grouping parent lines into heterotic groups is fundamental in both private and public breeding programs in order to reduce the number of crosses, and therefore field tests, necessary to evaluate potential high-yielding hybrids (Hallauer and Miranda, 1988). By heterotic groups we mean a population of genotypes that, when crossed with individuals from another heterotic group or population, consistently outperform intra-population crosses (Hallauer and Miranda, 1988). Molecular markers, such as RAPD (random amplified polymorphic DNA), AFLP (amplified fragment length polymorphism) and microsatellites, among others, have facilitated the development of new varieties by reducing the time required for the detection of specific traits in progeny plants and the identification of disease resistance genes (Korzun, 2003). Even though they have been proposed to assign new inbred to heterotic groups previously established (dos Santos Dias et al., 2004; Xia et al., 2004), their usefulness in this task still remains uncertain (dos Santos Dias et al., 2004). Machine-learning techniques, such as decision trees and artificial neural networks, are increasingly used in agriculture to deal with classification, prediction, and modeling problems (Mitchell et al., 1996; Kirchner et al., 2004); however, we found no reports about machine learning algorithms (Witten and Frank, 2005; Kotsiantis, 2007) and heterotic group assignment using molecular marker data. We conjecture that traditional distance-based methods (Reif et al., 2005) currently available for assigning new inbreds to heterotic groups in corn do not capture the possible non-linear relation between parental data and progeny performance (dos Santos Dias et al., 2004; Springer and Stupar, 2007) and that such type of non-linearity may be easily captured by supervised machine learning models.

In this paper, we evaluate the performance of several state-of-art supervised learning algorithms on molecular marker data for heterotic assignation, and delineate perspectives for further research.

2. Multiclass classifiers

The goal of supervised learning is to build a concise model of the distribution of class labels in terms of predictor features, the resulting classifier is then used to assign class labels to the testing instances where the values of the predictor features are known, but the value of the class label is unknown (Kotsiantis, 2007). There are numerous learning algorithms reported in the bibliography (Kotsiantis, 2007; Witten and Frank, 2005), for this introductory work we considered four well-known supervised learning algorithms implemented in Weka workbench (Hall et al., 2009): (i) Naive Bayes (John and Langley, 1995), (ii) Bayes Net (Friedman et

[*] Corresponding author. Tel.: +54 341 4821771x104; fax: +54 341 4821772.
E-mail address: ornella@cifasis-conicet.gov.ar (L. Ornella).

0168-1699/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.compag.2010.08.013

al., 1997), (iii) Simple Logistic (Landwehr et al., 2005) and (iv) Support Vector Machines (SVMs) with linear and radial basis function kernels (Burges, 1998).

2.1. Simple logistic

Landwehr et al. (2005) proposed Logistic Model Trees or LMT: trees that contain linear logistic regression functions at the leaves. In that work they report that at low number of training instances ($n \leq 100$) Simple Logistic (logistic model tree of size one) performs as well as more complex LMT and better than decision tree C4.5 (Quinlan, 1993), with less computational requirements (Landwehr et al., 2005).

Linear logistic regression models the posterior class probabilities $Pr(C = c\,|\,\mathbf{X} = \mathbf{X})$ for the J classes via functions linear in $\mathbf{x}$ and ensures that they sum to one and remain in [0, 1] (Sumner et al., 2005). The model is:

$$P(C = c|X = x) = \frac{e^{F_c(x)}}{\sum_{k=1}^{C} e^{F_k(x)}} \tag{1}$$

where $F_j(\mathbf{x}) = \sum_{m=1}^{M} f_{mj}(x) = \beta_j^T \cdot \mathbf{x}$. Estimates of β_j^T are obtained by numeric optimization algorithms that approach the maximum likelihood solution iteratively (Sumner et al., 2005). In Simple Logistic such iterative method is the LogitBoost algorithm (Landwehr et al., 2005). In each iteration, it fits a least-squares regressor to a weighted version of the input data with a transformed target variable. y_{ij}^* are the binary pseudo-response variables which indicate group membership of an observation like this:

$$y_{ij}^* = \begin{cases} 1 & \text{if} \quad y_i = j \\ 0 & \text{if} \quad y_i \neq j \end{cases} \tag{2}$$

By constraining f_{mj} to be a linear function of only the attribute that results in the lowest squared error arrives at an algorithm that performs automatic attribute selection (Sumner et al., 2005); also, by using cross-validation (5-folds) to determine the best number of LogitBoost iterations, only those attributes that improve the performance on unseen instances are included (Landwehr et al., 2005; Sumner et al., 2005).

2.2. Naive Bayes

NB learns from training data the conditional probability of each attribute A_i given the class label C. Classification is then done by applying Bayes rule to compute the probability of C given the particular instance of $A_1, \ldots, A_n$; and then predicting the class with the highest posterior probability. This computation is rendered feasible by making a strong independence assumption: all the attributes A_i are conditionally independent given the value of the class C. Independence means probabilistic independence, i.e, A is independent of B given C whenever $P(A\,|\,B, C) = P(A\,|\,C)$ for all possible values of A, B and C, whenever$P(C) > 0$ (Friedman et al., 1997). Even though the above assumption is clearly unrealistic, its predictive performance is competitive with state-of-the-art classifiers (Friedman et al., 1997; Kohonen et al., 2008).

2.3. Bayes Net

A Bayesian network is an annotated directed acyclic graph that encodes a joint probability distribution over a set of random variables U (Friedman et al., 1997). The graph G encodes independence assumptions: each variable X_i is independent of its nondescendants given its parents in $G(\Pi_{x_i})$:

$$p(x_1, x_2, \cdots, x_n) = \prod_{i=1}^{n} p(x_i|\Pi_{x_i}) \tag{3}$$

To use a BN as classifier, a search algorithm find a network B, $P_B(A_1, A_{.} \ldots, A_n, C)$, that best matches a training set D according to some scoring function (Friedman et al., 1997; Cooper and Herskovits, 1992). Once a network is learned, B returns the label c that maximizes the posterior probability $P_B(c/a_1, \ldots, a_n)$ (Friedman et al., 1997; Cooper and Herskovits, 1992). Naive Bayes can be considered a Bayes Net in where the structure of the graph is constrained (Friedman et al., 1997).

2.4. Support vector machines

The support vector machine (SVM) algorithm is based on the statistical learning theory and the Vapnik–Chervonenkis (VC) dimension introduced by Vladimir Vapnik and Alexey Chervonenkis (Cortes and Vapnik, 1995); the underlying idea is to calculate a maximal margin hyperplane (the decision function) separating two classes of the data (Cortes and Vapnik, 1995), such decision function is fully specified by a usually small subset of the data (the support vectors) which defines the position of the separator. New samples are classified according to the side of the hyperplane they belong to (Cortes and Vapnik, 1995; Devos et al., 2009).

In the case of non-separable data, the “ideal boundary” must be adapted to tolerate errors for some objects i:

$$\text{minimize} \quad \frac{1}{2}|\mathbf{w}|^2 + C\sum_{i=1}^{n} \zeta_i \tag{4}$$

under the constraints $\zeta_i \geq 0$, $\zeta_i + y_i(\mathbf{w} \cdot \mathbf{x}_i + b) - 1 \geq 0$, $\mathbf{w}$ and b are respectively the normal vector and the bias of the hyperplane, and each ζ_i corresponds to the distance between the object i and the corresponding margin hyperplane (Devos et al., 2009).

The parameter C is a regularization meta-parameter, when C is small, margin maximization is emphasized whereas when C is large, the error minimization is predominant (Cortes and Vapnik, 1995; Devos et al., 2009).

To learn non-linearly separable functions, data are implicitly mapped to a higher dimensional space by means of mercer kernels which can be decomposed into a dot product, $K(\mathbf{x}_i, \mathbf{x}_j) = \phi(x_i) \cdot \phi(x_j)$ (Burges, 1998). Examples of kernels are the linear kernel $K = (\mathbf{x}_i \cdot \mathbf{x}_j - 1)^{p=1}$ and the radial basis function kernel $K = e^{-\gamma(\mathbf{x}_i - \mathbf{x}_j)^2}$.

2.5. ECOC codes

SVMs have particular high generalization abilities and have become very popular in the recent years; nevertheless, they are inherently binary classifiers and a combination scheme is necessary to extend SVMs for problems with more than two classes (Rifkin and Klautau, 2004). In this work, the One Against All (OAA) (Rifkin and Klautau, 2004) and the Error Correcting Output Coding (ECOC) (Dietterich and Bakiri, 1995) combination schemes are used.

Briefly, OAA classifiers rely on the discrimination of individual classes against the others while ECOC codes are defined by a more general decomposition or “‘coding matrix”’ $M \in \left\{0, 1\right\}^{L \times N}$, which converts a L-multiclass problem into N binary tasks (Dietterich and Bakiri, 1995). There are several coding matrices reported in the bibliography (Dietterich and Bakiri, 1995; Allwein et al., 2000; Rifkin and Klautau, 2004). In particular we worked with random codes, where each entry of the matrix is chosen to be 0 or 1 with equal probability, N is limited by the maximum number of different and non-complementary binary vectors that can be generated for dichotomization (Dietterich and Bakiri, 1995).

The original approach to ECOCs predicts the class whose corresponding row vector has minimum Hamming distance to the vector of 0/1 predictions obtained from the N classifiers (Dietterich and Bakiri, 1995). Allwein et al. (2000) presented an alternative, loss-based decoding, which notices the magnitude of the predictions, sometimes interpreted as a measure of "confidence" of a prediction. Several authors verified that Loss-decoding indeed produces more accurate classifiers than the Hamming distance (Allwein et al., 2000; Rifkin and Klautau, 2004; Frank and Kramer, 2004).

3. Materials and methods

3.1. Datasets

We compiled three molecular marker datasets representing a broad spectrum of temperate and tropical germplasm. The Liu data (Liu et al., 2003) comprises 197 inbreeds (instances) of both temperate and tropical germplasm characterized by 188 attributes derived from 94 microsatellites. The number of distinct values per attribute ranges from 4 to 48 with a mean of 18.18. Missing data represents a 4.75 % of the total, ranging from 0% to 25.38%, depending on the attribute. Instances are distributed into 10 heterotic groups (classes) and the number of instances per group is $\{61, 13, 11, 8, 9, 13, 28, 17, 29, 8\}$. The Morales data (Morales Yokobori et al., 2005) comprises 26 temperate inbreeds of germoplasm characterized by 42 attributes derived from 21 microsatellites. The number of distinct values per attribute ranges from 2 to 13 with a mean of 4.72. Missing data represents a 8.60% of a total, ranging from 0% to 42% of missing data per attribute. Instances are distributed into 4 heterotic groups and the number of instances per group is $\{4, 8, 6, 8\}$. The Xia data (Xia et al., 2004) comprises 73 inbreeds of tropical germplasm characterized by 166 attributes derived from 83 microsatellites. The number of distinct values per attribute ranges from 2 to 14 with a mean of 5.93. Missing data represents the 8.02% from the total, ranging from 0% to 43.84% of missing data per attribute. Instances are grouped into 8 heterotic groups and the number of instances per group is $\{22, 17, 7, 5, 5, 5, 5, 7\}$.

3.1.1. Classifiers

Simple Logistic, Naive Bayes and Bayes Nets were all implemented with defaults parameters of Weka (Witten and Frank, 2005). SVMs were evaluated using lineal kernel and radial basis function (RBF) kernel, both also with default parameters ($C = 1$ for linear kernel and $C = 1$, $\gamma = 0.01$ for radial basis function kernel). In both SVM alternatives we choose the option "to fit Logistic regression models" of Weka's SMO (Sequential Minimal Optimization) algorithm for SVMs, which allows to emit an estimate of the confidence for the binary prediction instead of (0,1) hard outputs.

Concerning the implementation of ECOC classifiers, in a preliminary research we evaluated the data with variable length codes and we did observed a positive correlation between ECOC accuracy and code length. As a trade off between classifier's performance and computational complexity we choose random codes of length $N = 6$ for Morales data, $N = 55$ for Xia data and $N = 75$ for Liu data. Therefore, 75 SVMs were used for the ECOC classification of Liu data, 55 for Xia data, and 6 for Morales data. The multiclass schemes were implemented as a new WEKA classifier and integrated into the original package (Witten and Frank, 2005).

3.1.2. Evaluation of classifier's performance

The predictive power of supervised learning algorithms on molecular marker data was evaluated by means of the error rate (Borra and Ciaccio, 2005) and the Cohen's Kappa coefficient (Cohen, 1960) exhibited across 30 Montecarlo runs of stratified 10-fold Cross Validation (CV) experiments (Kohavi, 1995; Kirchner et al., 2004). At each Montecarlo run, the data was split into 10 different segments of almost the same size and containing approximately the same proportion of categories as the original dataset. For each segment, classifiers were respectively trained and evaluated on the samples derived by omitting the selected segment and on selected segment. At the end of this procedure, the average classification error and the average Kappa coefficient were reported. The choice of the Kappa coefficient was motivated by its ability to better measure the agreement between binary inter-annotators than the traditional classification error. In particular, the Kappa coefficient takes into account chance agreements (Cohen, 1960; Kirchner et al., 2004) and it is well suited for unequal class distribution datasets.

Two main classification scenarios were considered: (i) NB, BN, SL, OAA-rbf (SVM with radial basis function), ECOC-rbf, OAA-lineal (SVM with lineal kernel) and ECOC lineal classifiers on full molecular marker data, and (ii) the same classifiers evaluated on CFS reduced data.

3.1.3. Missing data

Regarding missing data, all associated to nominal attributes, imputation depends on the classifier evaluated (Su et al., 2008). In Weka, Naive Bayes ignores the missing values whereas SMO globally replaces all missing values by a default value, e.g., "unknown" (Su et al., 2008). Finally, in Bayes Net and Simple Logistic classification, missing values of training and test set are filled in using the mode of the corresponding attribute valuated on the training data (Bouckaert, 2008; Landwehr et al., 2005).

3.1.4. Statistical comparison among classifiers

It is important to assess whether the observed difference in classification performance is statistically significant or simply due to chance (Luengo et al., 2009). Comparisons of arithmetic means and visual inspection of Kappa boxplots was supplemented with Kolmogorov–Smirnov (KS-test) provided by the R[1] environment (stats package). KS is a nonparametric test and it has the advantage of making no assumption about the distribution of data (Luengo et al., 2009). For each dataset and condition evaluated (Full and CFS reduced data), all possible pairs of (A,B) Kappa coefficients distributions were assessed under the alternative hypothesis "distribution B is greater than distribution A" (The R Development Core Team, 2009).

3.2. Feature Selection

Reducing the feature space to non-redundant features results in improved classification accuracy and helps avoid overfitting of the classifiers. In this study, we experimented with Correlation-based Feature Subset selection (CFS) (Hall, 2000). The CFS strategy uses a correlation-based heuristic to evaluate the merit of feature subsets with respect to classification categories and the correlation between features. CFS selection implemented in WEKA is fully automatic and does not require a priori specification of the number of features to be included in the final subset (Hall, 2000). We apply a second feature selection method, Relief (Kononenko, 1994), on Morales data. This method ranks the worth of an attribute by repeatedly sampling an instance and considering the value of the given attribute for the nearest instance of the same and different class (Kononenko, 1994). In other words, Relief assigns more weight to those attributes that have the same value for instances from the same class and differentiate between instances from dif-

[1] http://www.rproject.org/.

ferent classes (Witten and Frank, 2005). Filtering algorithm was calibrated in order to retain 25, 50 and 75% of the original number of attributes.

3.2.1. SVM parameters optimization

Optimization of the meta-parameters, C (regularization parameter) of linear kernel and C and γ (RBF kernel), is the key step in SVM performance (Devos et al., 2009). Globally, when C is small the margin maximization is emphasized leading to large margin and smooth boundary. The number of support vectors included in the solution depends on this parameter and, usually, if the number of support vectors is high the solution is unstable and leads to lower classifications rates (Forman and Cohen, 2004; Devos et al., 2009). Also, when the value of γ is large, the separating boundary has a large number of support vectors and can become tortuous. Again, this risks overfitting the training set data to yield an SVM model that is not robust. In contrast, a small value of γ can lead to separating boundaries described with a small number of support vectors but that may be too smooth to classify the training set examples with sufficient accuracy (Jorissen and Gilson, 2005; Devos et al., 2009). In RBF kernels it has been reported that different combinations of C and γ lead to similar classification rates (Devos et al., 2009). To perform the optimization we implemented an exhaustive grid search: 30 points (C = 0.25, 0.5, 1, 2, 4 and G = 0.0001, 0.001, 0.01, 0.1, 1, 10) for radial basis function kernel and 5 points (C = 0.25, 0.5, 1, 2, 4) for the linear kernel. This approach enables to visualize directly the effect of both parameters and provides useful information of base classifiers performance. In order to minimize the risk of overfitting all parameters were estimated by external leaving out one Cross Validation (Morales) or 10-fold Cross Validation (Liu and Xia datasets) over the training data (Ambroise and McLachlan, 2002).

4. Results and discussion

Three native multiclass classifiers plus Support Vector Machines classifiers under the OAA (Rifkin and Klautau, 2004) and ECOC frameshifts (Dietterich and Bakiri, 1995) were evaluated on three molecular marker datasets representing a broad spectrum of maize heterotic patterns. Generalization error of classifiers in this domain was estimated by means of the error-rate and the Kappa Cohen's Coefficient. Error-rate, defined as the ratio between the number of misclassified cases and the total number of cases examined, is the common measure used in nonparametric classification models (Borra and Ciaccio, 2005). However, it does not compensate for classifications that might have been due to chance. Hence, we also used the Cohen's Kappa as a statistically robust alternative, especially in datasets with an unequal distribution of classes. Both statistics were determined by 30 runs of Montecarlo 10-fold CV experiments. Arithmetic means of these statistics, with and without feature selection, are shown in Table 2. It can be observed that results according to mean error-rate and Kappa values do not always agree. For example, in Liu Full data, SL and NB display identical error rates and different Kappa values; in Liu CFS reduced data the four SVM ensembles rank different either we consider Kappa or error rate values; also in Xia CFS data OAA schemes rank different whatever we choose error rate or Kappa (Table 2). Overall, classification results seem to be problem-dependent, indefinite and not always normal. Therefore arithmetics means may be not always provide representative measures of classification performance. Consequently, comparison of means and visual inspection of Kappa boxplots was supplemented with Kolmogorov–Smirnov (KS) tests (Luengo et al., 2009). We recall that KS is a nonparametric test which does not rely on an assumption of normality (Luengo et al., 2009).

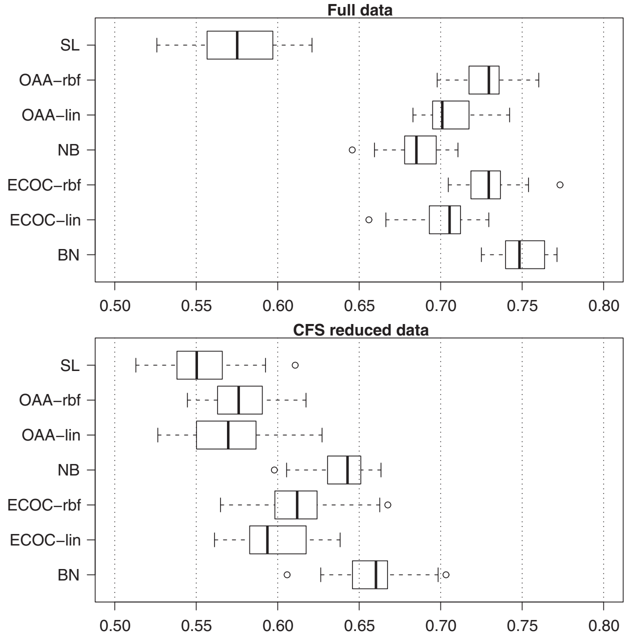

Fig. 1. *Liu* data. Boxplots of the Cohen's Kappa coefficient in 30 Montecarlo runs of 10-fold CV experiments. Native multiclass classifiers: Bayes Network (BN), Naive Bayes (NB), and Simple Logistic (SL). Multiclass extensions of Support Vector Machines (SVM): One Against All (OAA) and Error Correcting Output Coding (ECOC). Base classifiers: lin - SVM with linear kernel, rbf - SVM with radial basis function kernel. Results on full (top) and Correlation-based Feature Selection (CFS) reduced data (bottom) are shown.

4.1. Results on full data

Bayes Net exhibited the best *mean* performance on full Liu data (Table 2). Visual inspection of Kappa boxplots and KS test agreed with this result. All KS tests were significant when comparing the rest of classifiers to BN, for example; p-value = 6.55$e-$05 when comparing ECOC-rbf and OAA-rbf, the closest classifiers according to Kappa coefficient, to BN.

In Xia data, ECOC-rbf significantly exceeds the rest of classifiers (Table 2 and Fig. 2). In all KS test (any classifier vs. ECOC-rbf) the null hypothesis was rejected, as an example; p-value = 0.0015 when comparing ECOC-linear (the second ranked classifier) against this ensemble.

Finally, Simple Logistic exhibited the best mean performance on full Morales data (Table 2), a fact that was confirmed by corresponding Kappa boxplots (Fig. 3). Moreover, when comparing the rest of the classifiers with Simple Logistic using KS, the highest p-value obtained was 0.0006, i.e., all null hypothesis was rejected.

Concerning SL, our results are in agreement with Landwehr et al. (2005). When evaluating Liu and Xia data, which are more complex respect to a number of classes, number of attributes and number of instances; the classifier displays the worst performance (Figs. 1 and 2). Even though, we included this classifier in the analysis because of its good performance on Morales data, and this dataset is similar, with regard to number of instances and/or attributes, to most works reported in the literature, specially those from development countries (dos Santos Dias et al., 2004).

4.2. Impact of feature selection

The genetic basis of heterosis has been debated for nearly a century without a clear resolution. The two main hypotheses that advanced to explain this phenomenon are dominance and overdominance (Hallauer and Miranda, 1988; Springer and Stupar, 2007). It is also well documented that not all markers will be linkage

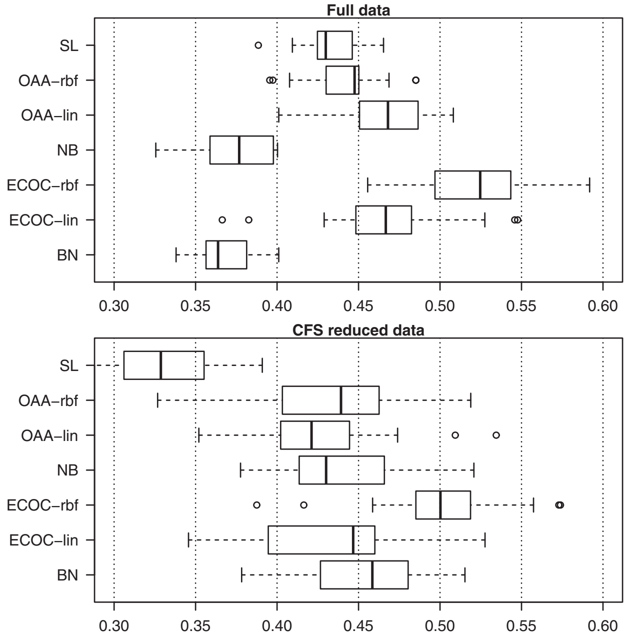

Fig. 2. *Xia* data. Boxplots of the Cohen's Kappa coefficient in 30 Montecarlo runs of 10-fold CV experiments. Native multiclass classifiers: Bayes Network (BN), Naive Bayes (NB), and Simple Logistic (SL). Multiclass extensions of Support Vector Machines (SVM): One Against All (OAA) and Error Correcting Output Coding (ECOC). Base classifiers: lin - SVM with linear kernel, rbf - SVM with radial basis function kernel. Results on full (top) and Correlation-based Feature Selection (CFS) reduced data (bottom) are shown.

Table 1
Number of features preserved by Correlation-based Feature Selection (CFS). *Liu*, *Xia*, and *Morales* are the original molecular marker datasets. Full data denotes the initial number of features of each dataset. Min and Max are respectively the arithmetic means of the maximum and minimum number of features selected during the 30 Montecarlo runs of 10-fold CV experiments.

	Dataset		
	Liu	Xia	Morales
Full data	188	166	42
Min	26	29	8
Max	50	42	20

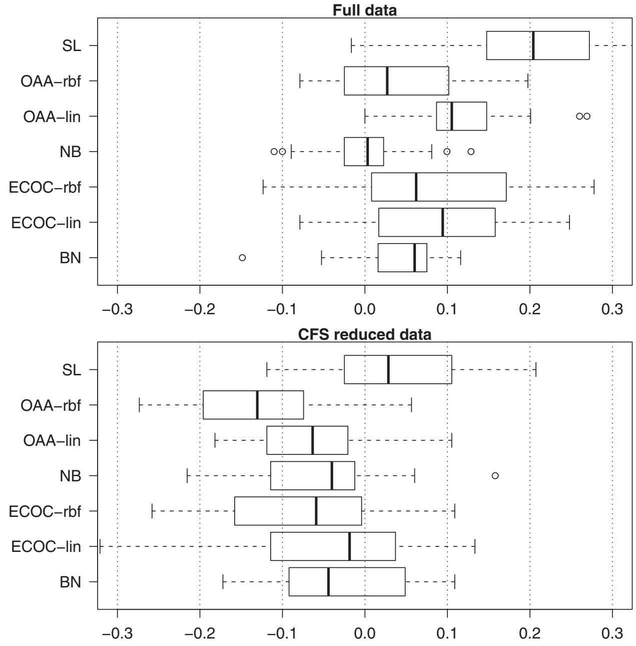

Fig. 3. *Morales* data. Boxplots of the Kappa coefficient in 30 Montecarlo runs of 10-fold CV experiments. Native multiclass classifiers: Bayes Network (BN), Naive Bayes (NB), and Simple Logistic (SL). Multiclass extensions of Support Vector Machines (SVM): One Against All (OAA) and Error Correcting Output Coding (ECOC). Base classifiers: lin - SVM with linear kernel, rbf - SVM with radial basis function kernel. Results on full (top) and Correlation-based Feature Selection (CFS) reduced data (bottom) are shown.

to genes or QTL (quantitative trait locus) associated with heterosis (Austin et al., 2000). Moreover, the diploid nature of data and the characteristics of the instances (homozygous lines) allow us to infer the existence of some redundancy in attributes. Therefore, we implemented CFS (Correlation-based Feature Selection) in order to remove attributes not related to the class. The number of CFS selected attributes was variable, depending on the dataset; extreme values ranged from 13.83 to 47.62 % of the initial number of features (Table 1).

Almost none of the classifiers improve their performance with filtered data (Table 2 and boxplots). The only exception was Naive Bayes and Bayes net evaluated on Xia data (Fig. 2). Even though, ECOC-rbf was still the best classifier. All KS tests were statistically significant when comparing the rest of classifiers to this ensemble.

In Morales reduced data and according to arithmetic means (Table 2 and boxplot of Fig. 3) SL was still the best classifier but; when ECOC lineal was compared to SL, the p-value was 0.0672. The rest of classifiers did show significant p-values in KS test. Finally, in Liu data, thought Naive Bayes degraded its performance with CFS filtering, like the rest of the classifiers; it ranked second after Bayes Net (p-value < 0.05).

Theory suggests that interactions between genes associated with molecular markers could play an important role in the generation of the observed heterosis (Dudley and Johnson, 2009), so it would be possible that using filters that contemplate interactions between attributes would contribute the classification improvements.

Table 2
Means of the error rate and Kappa values in 30 Montecarlo runs of 10-fold CV experiments. Native multiclass classifiers: Bayes Net (BN), Naive Bayes (NB), and Simple Logistic (SL). Multiclass extensions of Support Vector Machines: One Against All (OAA) and Error Correcting Output Coding (ECOC). Three molecular marker datasets, namely *Liu*, *Xia*, and *Morales*, are considered. Results on full and Correlation-based Feature Selection (CFS) reduced data are reported. Best results are shown in boldface.

Classifier	Full data						CFS reduced data					
	Liu		Xia		Morales		Liu		Xia		Morales	
	Error	Kappa	Error	Kappa	Error	Kappa	Error	Kappa	Error	Kappa	Error	Kappa
BN	**0.205**	**0.749**	0.475	0.368	0.715	0.039	**0.280**	**0.658**	0.428	0.455	0.755	−0.032
NB	0.345	0.685	0.472	0.372	0.751	0.000	0.294	0.638	0.432	0.439	0.772	−0.057
ECOC lineal	0.252	0.701	0.435	0.469	0.660	0.087	0.341	0.598	0.459	0.436	0.753	−0.039
ECOC rbf	0.223	0.730	**0.385**	**0.523**	0.681	0.078	0.320	0.613	**0.402**	**0.500**	0.786	−0.078
OAA lineal	0.245	0.706	0.415	0.465	0.645	0.116	0.348	0.571	0.460	0.424	0.768	−0.059
OAA rbf	0.223	0.730	0.429	0.442	0.690	0.043	0.357	0.579	0.462	0.433	0.819	−0.127
SL	0.345	0.576	0.436	0.433	**0.572**	**0.210**	0.367	0.552	0.537	0.326	**0.703**	**0.033**

4.3. Data complexity

Molecular marker data showed to be complex enough to require the careful exploration of non-trivial multiclass classifiers: the attribute-class relationship is possibly non-linear (dos Santos Dias et al., 2004; Springer and Stupar, 2007) and datasets present noisy and/or missing features (Jones et al., 1997). Also, the dimensionality of molecular marker data is between that of the classic Machine Learning setting ($n/p > 10$) (Kohavi, 1995; Asuncion and Newman, 2007) and that posed by recent challenging microarray data classification problems ($n/p << 1$) (Mukherjee et al., 2003), where n is the number of instances and p the number of attributes. Actually, the number of classes ranges from 4 to 10 and the number of instances per class is generally less than 30, which is a very low number of training instances (Liu et al., 2003; Xia et al., 2004; Morales Yokobori et al., 2005; dos Santos Dias et al., 2004).

When comparing classifiers performance on full data scenarios we did observe significative differences between Liu, Xia and Morales data results (Table 2). Kappa values ranging between 0.61 and 0.80 indicate a substantial agreement between observed and predicted data whereas values below 0.20 indicate only a slight agreement (Landis and Koch, 1977).

From a genetic point of view, differences of methods used to established the heterotic groups could be reflecting differences between mechanisms relating attributes (molecular markers) with classes (heterotic groups): heterotic groups of Xia and Morales data were established on the basis of field essays (topcross or diallel) and, according to Xia et al. (2004), the mixed genetic constitution of the populations and pools of Cymmit germplasm (Xia data) made the task of assigning them to genetically diverse and complementary heterotic groups difficult. A similar situation was reported for Morales data (Eyhérabide et al., 2006). Liu data clusters, on the other side, were established on the basis of genetic origin (Liu et al., 2003) so it was easy to assign new lines to groups solely on molecular data.

From a Machine Learning point of view, these differences could be due to a challenging ratio between the number of instances (n) and the number of attributes (p) of training data (Mukherjee et al., 2003; Kohavi, 1995). For example, for microarray data (extremely low n/p ratios) achieving error rates around 0.1–0.2% requires in the order of 75–100 training samples (Mukherjee et al., 2003), whereas Kohavi (1995) reported error rates from 5.8 to 53.2% when working with datasets comprising a number of instances and a number of attributes similar to those used in this work. However, if the bad classification performance for Morales and Xia data bases is only due to the n/p ratios (specially for Morales data set), a good feature selection method should improve the results. It can be seen from Figs. 2 and 3 that attribute CFS selection did not improve the accuracy of the classifiers. We performed an additional experiment on Morales dataset using another filter method implemented in Weka, Relief (Kononenko, 1994), and selecting 25, 50 and 75% of the original number of attributes. Filtered data was evaluated with Simple Logistic and the four SVM ensembles and as stated in Materials and Methods. It can be see from Fig. 4 that, except a few and non-significant exceptions, all classifiers degrades their performance with higher n/p ratios.

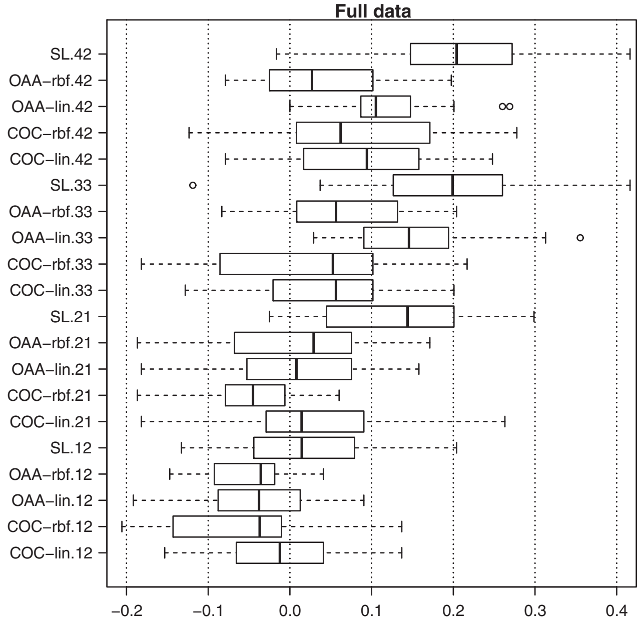

Fig. 4. *Morales* data. Boxplots of the Kappa coefficient in 30 Montecarlo runs of 10-fold CV experiments. Full and Relief Filtered data: Simple Logistic (SL). Multiclass extensions of Support Vector Machines (SVM): One Against All (OAA) and Error Correcting Output Coding (ECOC). Base classifiers: lin - SVM with linear kernel, rbf - SVM with radial basis function kernel. 42, 33, 21 and 12 indicate the number of attributes retained after filtering.

Table 3
Means of the error rate and Kappa values in 30 Montecarlo runs of 10-fold CV experiments of optimized SVM with lineal kernel and under two decomposition schemes (OAA and ECOC).

Classifier	One Against All			Random code		
	Error	Kappa	KS test (Kappa) [a]	Kappa	Error	KS test (Kappa) [a]
Morales data	0.6308	0.1338	p-Value = 0.1184	0.6500	0.1021	p-Value = 0.3012
Xia data	0.4160	0.4631	p-Value = 0.5866	0.4438	0.4576	p-Value = 0.9672
Liu data	0.2302	0.7160	p-Value = 0.9560	0.2330	0.721	p-Value = 0.9354

[a] Kolmogorov–Smirnov test was performed between outputs of classifier with default parameter ($C = 1$) and outputs of classifier with optimized parameter and as stated in Section 3.

Table 4
Means of the error rate and Kappa values in 30 Montecarlo runs of 10-fold CV experiments of optimized SVM with radial basis function kernel and under two decomposition schemes (OAA and ECOC).

Classifier	One Against All			Random code		
	Error	Kappa	KS test (Kappa) [a]	Kappa	Error	KS test (Kappa) [a]
Morales data	0.6795	0.0509	p-Value = 0.0761	0.7556	−0.0410	p-Value = 1.0000
Xia data	0.4201	0.4550	p-Value = 0.0357	0.3583	0.5540	p-Value = 0.0327
Liu data	0.2200	0.7350	p-Value = 0.9030	0.2430	0.7500	p-Value = 0.9350

[a] Kolmogorov–Smirnov test was performed between outputs of classifier with default parameter ($C = 1$, $\gamma = 0.01$) and outputs of classifier with optimized parameters and as stated in Section 3.

It has been reported SVM classification is quite sensitive to meta-parameters (Rifkin and Klautau, 2004; Devos et al., 2009). However, we could not observe a significative enhancement of ensembles performance with the optimization of the meta-parameters (*C* in linear kernel and *C* and γ in radial basis function kernel). None of the optimized lineal SVM-ensembles significant outperformed their standard counterparts (Table 3). In Xia data both, OAA and ECOC, optimized RBF ensembles outperform classifiers with default values provided by Weka (Table 4). In Morales data, only OAA-RBF shows a significant improvement with optimized parameters (Table 4) with respect to Morales data, this is reasonable because with small training sets optimization of parameters, even by cross-validation, may only lead to over fitting the training set (Forman and Cohen, 2004). Surprisingly, in Liu data none of the optimized SVM ensembles (significantly) outperformed their counterparts with default parameters. This could be attributed to the number of missing data and the imputation technique of SMO (Su et al., 2008), or to the robustness of ensembles to base classifier error (Dietterich and Bakiri, 1995). Last but not least, classification accuracy not only depend on the number of instances and or attributes but also on the relationships between the attributes and classes so; if we apply the incorrect model, the expected performance will be poor. Classifiers were selected upon their reported performance on similar data and because they used different approaches to classify the data. However, remains to explore new algorithms from the bibliography.

5. Summary and conclusions

The information on germplasm diversity and relationships among elite materials is a fundamental importance in crop improvement (Hallauer and Miranda, 1988). Assigning lines to different heterotic groups would avoid the development and evaluation of many of the crosses that would eventually be discarded (Terron et al., 1997). Our proposal was to complement traditional breeding using molecular markers information and supervised learning algorithms. Three well-known multiclassifiers and support vector machine (a binary classifier) with linear and radial basis function kernels and under two decomposition schemes were evaluated using three molecular datasets representing a broad spectrum of maize heterotic patterns. Morales dataset includes 26 lines, mostly derived from orange flint (temperate) germplasm, clustered in four heterotic groups by topcross field essays (Eyhérabide et al., 2006), Liu data includes 248 inbred lines of importance to temperate breeding and many important tropical and subtropical lines (Liu et al., 2003) and Xia data 73 inbreds of tropical germplasm grouped mainly by diallel (Xia et al., 2004). We also used CFS filtering to improve classifiers performance, but we only obtained a slight improvement in Xia data. We also evaluated Relief filtering on Morales data, with negative results. However, CFS removes noisy attributes non-correlated between them and theory suggests that interactions between genes associated with molecular markers could play an important role in the generation of the observed heterosis (Pea et al., 2008) so filters that contemplates this situation remain to be explored. Finally, although results obtained with heterotic groups established by field essays (top cross or diallel) are relatively poor, there is a strong evidence that using data with more training instances could generate successful classifiers. Also it is necessary to evaluate other algorithms; the potential impact, in time and money, on crop sustainability makes our research worth to try: while traditional genetic breeding requires expensive fields test and a time scale in the order of years for obtaining an heterotic assignment, in our proposed framework costs are significantly lower and the time scale is in the order of weeks, two weeks for growing an small plant plus a week to obtain molecular data and a couple of days for computational analysis.

Acknowledgments

Authors would like to acknowledge Dr. Marilyn Warburton from Cimmyt, Mexico, for permitting use of Xia data. We also thank J. Coronel and L. Angelone for technical support. Authors also would like to thank the reviewers for their comments that help improve the manuscript. E. Tapia's work is supported by project PICT 11-15132, Agencia Nacional de Promoción Científica y Tecnológica from Argentina. Leonardo Ornella is a doctoral fellow of CONICET, Argentina.

References

Allwein, E.L., Schapire, R.E., Singer, Y., 2000. Reducing multiclass to binary: a unifying approach for margin classifiers. Journal of Machine Learning Research 1, 113–141.

Ambroise, C., McLachlan, G.J., 2002. Selection bias in gene extraction on the basis 405 of microarray gene-expression data. Proceedings of the National Academy of Sciences 99, 6562–6566.

Asuncion, A., Newman, D., 2007. UCI Machine Learning Repository. School of Information and Computer Sciences, University of California, Irvine.

Austin, D.F., Lee, M., Veldboom, L.R., Hallauer, A.R., 2000. Genetic mapping in maize with hybrid progeny across testers and generations: grain yield and grain moisture. Crop Science 40 (1), 30–39.

Borra, S., Ciaccio, A., 2005. Methods to compare nonparametric classifiers and to select the predictors. In: Vichi, M., Monari, P., Mignani, S., Montanari, A. (Eds.), New Developments in Classification and Data Analysis. Springer, pp. 11–19.

Bouckaert, R.R., 2008. Bayesian Network Classifiers in Weka for Version 3-5-7.

Burges, C.J.C., 1998. A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery 2, 121–167.

Cohen, J., 1960. A coefficient of agreement for nominal scales. Educational and Psychological Measurements 20, 37–46.

Cooper, G.F., Herskovits, E., 1992. A bayesian method for the induction of probabilistic networks from data. Machine Learning 9 (4), 309–347.

Cortes, C., Vapnik, V., 1995. Support-vector networks. Machine Learning 20, 273–297.

Devos, O., Ruckebusch, C., Durand, A., Duponchel, L., Huvenne, J.P., 2009. Support vector machines (SVM) in near infrared (NIR) spectroscopy: focus on parameters optimization and model interpretation. Chemometrics and Intelligent Laboratory Systems 96 (1), 27–33.

Dietterich, T.G., Bakiri, G., 1995. Solving multiclass learning problems via error-correcting output codes. Journal of Artificial Intelligence Research 2, 263–286.

dos Santos Dias, L., de Toledo Picoli, E., Rocha, R.B., Alfenas, A.C., 2004. A priori choice of hybrid parents in plants. Genetics and Molecular Research 3, 356–368.

Dudley, J.W., Johnson, G.R., 2009. Epistatic models improve prediction of performance in corn. Crop Science 49 (3), 763–770.

Eyhérabide, G., Nestares, G., Hourquescos, M., 2006. Development of a heterotic pattern in orange flint maize. In: Lamkey, K., Lee, M. (Eds.), Plant Breeding: The Arnel R. Hallauer International Symposium. Blackwell Publishing, pp. 352–379.

Forman, G., Cohen, I., 2004. Learning from little: comparison of classifiers given little training. In: 8th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD), pp. 161–172.

Frank, E., Kramer, S., 2004. Ensembles of nested dichotomies for multi-class problems. In: Proceedings of the 21st International Conference of Machine Learning (ICML-2004). ACM Press, pp. 305–312.

Friedman, N., Geiger, D., Goldszmidt, M., 1997. Bayesian network classifiers. Machine Learning 29, 131–163.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H., 2009. The WEKA data mining software: an update. SIGKDD Exploration Newsletter 11, 10–18.

Hall, M.A., 2000. Correlation-based feature selection for discrete and numeric class machine learning. In: Proc. 17th International Conf. on Machine Learning. Morgan Kaufmann, San Francisco, CA, pp. 359–366.

Hallauer, A.R., Miranda, J.B., 1988. Quantitative Genetics in Maize Breeding, 2nd edition. Iowa State University Press, Ames.

John, G.H., Langley, P., 1995. Estimating continuous distributions in Bayesian classifiers. In: Eleventh Conference on Uncertainty in Artificial Intelligence. Morgan Kaufmann, pp. 338–345.

Jones, C., Edwards, K., Castaglione, S., Winfield, M., Sala, F., 1997. Reproducibility testing of RAPD, AFLP and SSR markers in plants by a network of European laboratories. Molecular Breeding 3, 381–390.

Jorissen, R.N., Gilson, M.K., 2005. Virtual screening of molecular databases using a support vector machine. Journal of Chemical Information and Modeling 45, 549–561.

Kirchner, K., Tölle, K., Krieter, J., 2004. The analysis of simulated sow herd datasets using decision tree technique. Computers and Electronics in Agriculture 42, 111–127.

Kohavi, R., 1995. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. In: IJCAI, pp. 1137–1145.

Kohonen, J., Talikota, S., Corander, J., Auvinen, P., Arjas, E., 2008. A naive Bayes classifier for protein function prediction. In Silico Biology 9, 0003.

Kononenko, I., 1994. Estimating attributes: analysis and extensions of relief. In: Bergadano, F., Raedt, L.D. (Eds.), European Conference on Machine Learning. Springer, pp. 171–182.

Korzun, V., 2003. Molecular markers and their application in cereals breeding. In: Marker Assisted Selection: A fast Track to Increase Genetic Gain in Plant and Animal Breeding Session I: MAS in plant. Tech. rep., FAO.

Kotsiantis, S.B., 2007. Supervised machine learning: a review of classification techniques. Informatica 31, 249–268.

Landis, J.R., Koch, G.G., 1977. The measurement of observer agreement for categorical data. Biometrics 33 (1), 159–174.

Landwehr, N., Hall, M., Frank, E., 2005. Logistic model trees. Machine Learning 95 (1–2), 161–205.

Lee, M., 1998. Genome projects and gene pools: new germplasm for plant breeding? Proceedings of the National Academy of Sciences USA 95, 2001–2004.

Liu, K., Goodman, M., Muse, S., Smith, J., Bucklerd, E., Doebley, J., 2003. Genetic structure and diversity among maize inbred lines as inferred from DNA microsatellites. Genetics 165, 2117–2128.

Luengo, J., García, S., Herrera, F., 2009. A study on the use of statistical tests for experimentation with neural networks: analysis of parametric test conditions and non-parametric tests. Expert Systems with Applications 36 (4), 7798–7808.

Mitchell, R.S., Sherlock, R.A., Smith, L.A., 1996. An investigation into the use of machine learning for determining oestrus in cows. Computers and Electronics in Agriculture 15, 195–213.

Morales Yokobori, M., Decker, V., Ornella, L., 2005. Analysis of heterotic maize (Zea mays L.) populations using molecular markers. Maize Genetics Cooperation Newsletters 79, 36.

Mukherjee, S., Tamayo, P., Rogers, S., Rifkin, R., Engle, A., Campbell, C., Tr, G., Jp, M., 2003. Estimating dataset size requirements for classifying DNA microarray data. Computational Biology 10, 119–142.

Pea, G., Ferron, S., Gianfranceschi, L., Krajewski, P., Pè, M.E., 2008. Gene expression non-additivity in immature ears of a heterotic F1 maize hybrid. Plant Science 174 (1), 17–24.

Quinlan, R., 1993. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, San Mateo, CA.

Reif, J., Melchinger, A., Frisch, M., 2005. Genetical and mathematical properties of similarity and dissimilarity coefficients applied in plant breeding and seed bank management. Crop Science 45, 1–7.

Rifkin, R., Klautau, A., 2004. In defense of one-vs-all classification. Journal of Machine Learning Research 5, 101–141.

Springer, N.M., Stupar, R.M., 2007. Allelic variation and heterosis in maize: how do two halves make more than a whole? Genome Research 17, 264–275.

Sumner, M., Frank, E., Hall, M.A., 2005. Proc 9th European Conference on Principles and Practice of Knowledge Discovery in Databases, Eeding up Logistic Model Tree Induction. In: PKDD, pp. 675–683.

Terron, A., Preciado, E., Cordova, H., Mickelson, H., Lopez, R., 1997. Determinación del patrón heterótico de 30 líneas de maíz derivadas de la población 43 SR del CIMMYT. Agron. Mesoamericana 8, 26–34.

The R Development Core Team, dic 2009. R: A Language and Environment for Statistical Computing. Reference index http://www.r-project.org/.

Witten, I.H., Frank, E., 2005. Data Mining: Practical machine learning tools and techniques, 2nd Edition. Morgan Kaufmann, San Francisco.

Xia, X.C., Reif, J.C., Hoisington, D.A., Melchinger, A.E., Frisch, M., Warburton, M.L., 2004. Genetic diversity among CIMMYT maize inbred lines investigated with SSR markers. I. Lowland tropical maize. Crop Science 44, 2230–2237.

Su, X., Khoshgoftaar, T.M., Greiner, R., 2008. Using imputation techniques to help learn accurate classifiers. Tools with artificial intelligence. In: IEEE International Conference on 1, pp. 437–444.