Reducing the number of DNA primers for classifying pejibaye palm races using SVM

Jose L Vásquez
University of Costa Rica
15 PUBLICATIONS **27** CITATIONS
SEE PROFILE

Javier Vásquez
University of Costa Rica
6 PUBLICATIONS **5** CITATIONS
SEE PROFILE

Juan Briceno
University of Costa Rica
26 PUBLICATIONS **132** CITATIONS
SEE PROFILE

Carlos M. Travieso
Universidad de Las Palmas de Gran Canaria
380 PUBLICATIONS **2,962** CITATIONS
SEE PROFILE

Reducing the number of DNA primers for classifying Pejibaye Palm races using SVM

JOSÉ LUIS VÁSQUEZ[1], JAVIER VÁSQUEZ[2], JUAN CARLOS BRICEÑO[2], ELENA CASTILLO[3], CARLOS M. TRAVIESO[4]

[1]Sede del Atlántico, Universidad de Costa Rica, COSTA RICA
jose.vasquez@ucr.ac.cr
[2]Computer Science Department. University of Costa Rica.
Sede "Rodrigo Facio Brenes", Montes de Oca, Post-Code 2060, San José. COSTA RICA
{javier.vasquez, juancarlos.briceno}@ecci.ucr.ac.cr
[3]Centro de Investigación en Biología Molecular, Ciudad Universitaria Rodrigo Facio, 2500,
Universidad de Costa Rica, COSTA RICA
elena.castillo@ucr.ac.cr
[4]Signals and Communications Department. Technological Center in Communication Innovation.
University of Las Palmas de Gran Canaria
Campus de Tafira, Edificio de Telecomunicación, Pabellón B, E-35017 Las Palmas de Gran Canaria,
SPAIN. ctravieso@dsc.ulpgc.es

Abstract**.** This paper presents a feature reduction method, applying to Deoxyribonucleic Acid (DNA) primer, obtaining 100% classes identification based on Support Vector Machines (SVM). In particular, the biochemical parameterization has 89 Random Amplified polymorphic DNA (RADP) primers of Pejibaye Palm races, and it has been reduced to 10 RADP primers. The interest of this application is economic and computational, because it is so much cheaper to calculate less primers (only 11.24% from the previous dataset); and therefore, this supervised classification system is faster in order to do a method of origin denomination plant certification.

Keywords: Dimensionality Reduction, feature selection, DNA analysis, supervised classification

1 Introduction

Nowadays, the reduction of characteristic is a main aim on pattern recognition due to increasing volume of data and the expected cost / benefits rate, because each irrelevant feature excluded from the needed set to certify Pejibaye seeds, will spare chemicals products and time inverted in an inefficient procedure. It also will avoid to record spurious data.

In this paper, authors present a method for reducing the needed number of Pejibaye palm DNA primers, produced by the use of RAPD technique (Random Amplified polymorphic DNA). In order to certify this method, Support Vector Machines (SVM) has been used.

Pejibaye palm presents a large variety of morphology genus and large distribution over Central and South America. Since last century, due to the crop origin controversy [2][13] till now unsolved, mayor concern has been to identify biologically, domestic races and the research has been aimed to obtain genetic improvement and preservation instead of varieties identification. Economically, because different “landraces” (varieties), promote more or less one or other product and, in order to obtain origin denominations, there is an evident interest to correctly certify each one of different seed varieties.

The proposed system is based on feature reduction comparing three different methods and feedback with exhaustive search authenticated by Support Vector Machines (SVM).

After feature reduction, the goal was to do an optimization of the selected primer set; therefore, an exhaustive method is proposed to remove the worst significant primers, but keeping the discrimination between classes.

On this study we have obtained three important results. In the first place a corroboration of RAPD traces analysis technique, obtaining an inexpensive straight forward method to validate Pejibaye Palm

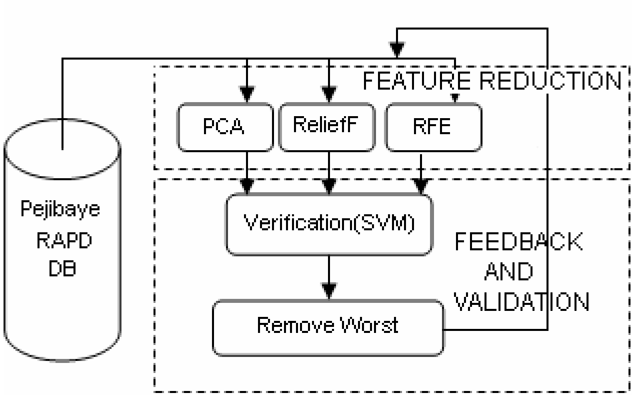

Fig.1. Dimensionality reduction method proposes in this present work.

parameterization of DNA chains, and obtaining similar grouping on selected landraces than morphological analysis. Second a substantial reduction of RAPD parameters to account for, and therefore concluding in a real time system response. And finally a 100% correct identification of each palm variety.

2 Database and its parameterization

The germ plasma bank of the University of Costa Rica has been stabilized about 30 years ago and account for more than 1200 different introductions of Pejibaye palms from Central and South America, becoming one of the most World wide completed. In this present work, we have used a database with 6 classes of Pejibaye (Utilitis - Costa Rica, Tuira - Panama, Putumayo - Colombia, Yurimagua - Peru, Tembé - Bolivia and Pará - Brazil), and each one has 13 samples with 89 RAPD primers per sample. Travieso et al. showed is possible to recognize with 100% assurance different pejibaye palm races using this 89 RAPD primers [17]. Raw DNA analysis is a very expensive and time consuming technique but, the interest of such analysis is based on the fact that it is used on decision making, management and preservation of genetic resources, taxonomy and systematic molecular studies.

Technical and financial aspects avoid us to use another type of data but the produced by the RAPD technique.

Several techniques have been developed in order to diminish this description extension. RAPD trace analysis is one of those finger printing technique based on PCR (Polymerase Chain Reaction) [13], [3], [5], [14], [15], and [19] (see Fig. 2). This study was realized over each individual's genetic material, with 89 OPC primers (from the Operon Company) producing information variables with clear and well defined fragments, after multiples reactions amplifications for each individual it is obtained an 89 long parameter binary description vector associated with a nominal classifier. That is to say, primers and individuals produced a binary matrix, indicating enough presence of a particular RAPD primer, from the six different Pejibaye races considered. From the beginning was unclear if the considered Boolean features were interdependent or not.

On the Fig. 2, we see some examples of Utilitis-Tucurrique pejibayes amplified DNA description, through the application of the PCR OPC-20 primer, with the RAPD technique. From left to right Columns 1 to 11: samples of the amplified DNA. Column 12 shows the molecular weight primer (1Kb).

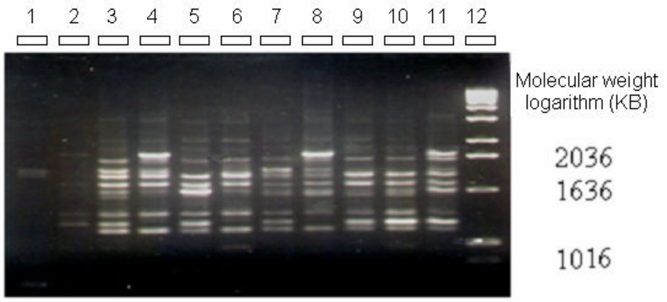

Fig.2. Some examples of Utilitis-Tucurrique pejibaye amplified DNA description, through the application of the PCR OPC-20 primer, with the RAPD technique.

3 Dimensionality reduction method

To avoid the appearance of spurious patterns, two different types of feature selection techniques were used; in first place we applied 2 different types of filters that evaluate attribute's relevance using general characteristics of the data. In second place we use a wrapper, which evaluates the attribute's merit by using the estimates generated from a learning algorithm. [1][7]

The first selected filter was an unsupervised method: the principal components analysis (PCA) [6]. The second selected filter was an instance base attribute ranking scheme nominated ReliefF [10] that handle noise and works fine with multiple-class data sets.

The selected wrapper technique was the Recursive Feature Elimination [5], which evaluates the worth of an attribute by using an SVM classifier [18].

It was used a ranking method, which measures the merits from each attribute. SVM was the approach for classification, which is supervised learning.

3.1 Principal Component Analysis (PCA)

Principal Components Analysis (PCA) is a technique proposed by Pearson [14] for identifying patterns in data, and expressing the data in such a way as to highlight their similarities and differences [8]. Since patterns in data can be hard to find in data of high dimension, where the feasibility of graphical representation is not available, PCA is a powerful tool for analyzing data. The other main advantage of PCA is that once you have found these patterns in the data, you can compress the data by reducing the number of dimensions, without much loss of information.

PCA is an orthogonal linear conversion that transforms a number of possible correlated data into a smaller number of uncorrelated data, bringing it in a new coordinate system such that the greatest variance by any projection of the data comes to lie on the first coordinate (called the first principal component), the second greatest variance on the second coordinate, and so on. PCA is theoretically the optimum transform for a given data in least square terms.

In PCA, the basis vectors are obtained by solving the algebraic eigenvalue problem $R^T XX^T R = \Lambda$ where X^T is a data matrix where each row represents a different repetition of the experiment, XX^T is the matrix of observed covariances, R is a matrix of eigenvectors, and Λ. is the corresponding diagonal matrix of eigenvalues. The projection of data, $C_n = R^T{}_nX$ from the original p dimensional space to a subspace spanned by n principal eigenvectors is optimal in the mean squared error sense.

This method is very well known and extensively used in many different applications for feature selection and/or dimensionality reduction.

3.2 ReliefF

In 1992 Kira and Rendell proposed the Relief algorithm that works with two classes. ReliefF was proposed by Kokonenko [10][16] and it works with multiple data sets and handles noise. ReliefF assigns a merit (grade of relevance) to each attribute and every time a feature has a value over this threshold, is the feature selected. ReliefF is a filter feature selection technique that begin generating a random example and searching its two closer neighbors, one from its same class and the other from another class. This technique updates the attribute's weight depending on the similarity between its values and the values from its neighbors and estimates the relevance of a feature, based on the ability to distinguish which class this instance belongs to. This process continuous until a threshold is reached.

3.3 Recursive Feature Elimination

Recursive Feature Elimination is a wrapper technique proposed by Guyon et al. [5]

This algorithm reduces the possibility of overfitting the learning scheme, what is usual when the amount of features is not supported with at least 10 samples per attribute as was stated by Jain et al. [9]. In this data bank we have 89 attributes and only 87 morphologically different samples.

A reduction of attributes was realized before classifying with SVM. The reduction of attributes combined two strategies, we calculated a ranking to be used with SVM and then we compute the SVM model. Originally were considered 89 primers, and the algorithm to make the ranking joins the individual assessment of each attribute based on the Recursive Feature Elimination, thereafter was used the SVM. We also tested the "Backward elimination" algorithm to perform a heuristic search [1]. The results of this ranking process and the elimination of the worst attribute were also tested in association of Support Vector Machines.

4 Classification system

In this work, we have used Support Vector Machine (SVM) as classifier, in order to evaluate and analyze performance and behaviour of the dimensionality reduction on Pejibaye Palm DNA Primers. For the based on the SVM classification system, we have calculated error, success and rejected rates to establish the efficiency of the model.

Particularly, we have used an implementation of Vapnik`s Support Vector Machine known as SVM light [18] which is a fast optimization algorithm for pattern recognition, regression problem, and learning retrieval functions from unobtrusive feedback to propose a ranking function. The optimization algorithms used in SVM light are described in [18]. The algorithm has scalable memory requirements and can handle problems with many thousands of support vectors efficiently.

As SVM is a supervised learning approach, and to be sure we don't have overfitted the learning scheme, we have used a cluster algorithm as unsupervised learning technique to validate our results.

4.1 Proposed algorithm to reduce dimensionality

Be K the set of the selected attributes, *tech* one of the following techniques (PCA, ReliefF and Recursive Feature Elimination), *merit*(K) an ordered set of the individual merit each attribute according *tech, num* (K) the number of attributes in K, *worst*(K) the attribute with the worst merit, *best*(n, K) the best *n* attributes according with *merit*(K), and *SVM*(K) a learning scheme over K using SVM as classifier, which uses as learning approach a cross validation from 4 partitions, that is it manages 4 iterations lo learn a model, using in each one 25% of the cases to learn and the other 75% to test the model.

In each iteration of the cross validation, the model tries to correct the mistakes made in the previous iterations. Steps to reduce dimensionality are:

```
1. For each of the following techniques (PCA,
   ReliefF and Recursive Feature Elimination) do:
a. Load the file with the K original attributes
b. Set n = num (K) +1
c. While (((SVM(K) = = 100%) && (n != num (K)
   ))
       // 100%assurance && there are
       // irrelevant features
     i. Compute merit(K)
    ii. Set K' = first(num(K)/2, K)   //take the
        main attributes
   iii. If (SVM(K') == 100%))
           1. Set K = K'
    iv. Else
           1. Set K' = K
           2. While (SVM(K') == 100%)
                a. Compute merit(K')
                b. Set feature = worst(K')
                c. Set K' = K' – feature
           3. Set K' = K' + feature
     v. Set n = num (K')
    vi. K = K'
d. Save K
```

5 Experiments and results

In order to develop experiments of the dimensional reduction, we have done 3 different tests and after each one, we built a classifier suitable for Pejibaye palm certification based on the SVM technique; conducted test were PCA, attribute subset selection with ReliefF feature elimination and recursive feature elimination.

5.1 Attribute subset selection with PCA and feedback to SVM

Table 1 shows a set of independent iterations (one per row), using PCA as algorithm for eliminating irrelevant data. For each iteration was generated a file including the transformation to the original dimensions of the main components generated with PCA and the data were classified using SVM with a polynomial kernel. The successful results were reduced quickly. It is shown that by using only 12 attributes, the success rate has been reduced to 27.59%.

Table 1. Classifying with PCA and SVM.

ID	Number of Primers	Time (sg.)	Folds	Epsilon	kernel	Success
a	89	2.28	10	1×10^{-12}	Pol.	100%
b	44	2.08	10	1×10^{-12}	Pol.	98.8 %
c	22	1.45	10	1×10^{-12}	Pol.	63.22%
d	12	1.48	10	1×10^{-12}	Pol.	27.59%

5.2 Attribute selection with ReliefF and feedback to SVM

It was tested a technique for ranking attributes, that is independent of the classifier. The worst attribute was removed before using the SVM. The results are shown in Table 2.

Table 2. Metrics based in the isolated primer elimination

ID	Number of Primers	Epsilon	Time (sg.)	Folds	Success
1	89	1×10^{-12}	2.25	4	100%
2	44	1×10^{-12}	2.2	4	100%
3	22	1×10^{-12}	2.39	4	100%
4	14	1×10^{-12}	1.56	4	100%

5.3 Attribute subset selection with Recursive Feature Elimination and feedback to SVM

The use of a wrapper was iteratively applied with SVM and the elimination of the worst attribute. Results are shown in Table 3. The final dimensionality was tested by a support vector machine using a polynomial kernel, which maintained 100% accuracy with only 30% of instances. Subsequently elimination of the worst attribute was maintained while the resulting model had 100% accuracy.

Table 3. Metrics based on the use of a wrapper with SVM.

ID	Number of Primers	Epsilon	Time (sg.)	Folds	Success
1	89	1×10^{-12}	1.47	4	100%
2	44	1×10^{-12}	1.41	4	100%
3	22	1×10^{-12}	1.44	4	100%
4	10	1×10^{-12}	1.58	4	100%

6 Conclusion

A method has been implemented for automatic landraces identification, using the RADP method, and being classified by an SVM. The success rate achieves 100% reducing to 11.24% of the original dataset, checked with our database of Pejibaye DNA. These good results suggest the use of this technique in the field of the Biochemist, as it should provide Biochemists with assistance in carrying out their tasks and reduce their costs. At summary, this present work gives a tool with high feature reduction, keeping the discrimination between classes.

7. Acknowledgment

This work was supported by the Spanish Government under Grants from “Programa José Castillejo” in 2008/09.

References:

[1] Blum, A.; Langley P., Selection of relevant features and examples in machine learning, Artificial Intelligence, 1997, pp. 245-271

[2] Clement, C.R.; Aguiar, J.; Arkcoll, D.B.; Firmino, J.; Leandro, R., Pupunha brava (Bactris dahlgreniana Glassman): progenitora da pupunha (Bactris gasipaes H.B.K.), Boletim do Museu Paraense Emilio Goeldi, Botánica Vol.5, No.1, 1989, pp. 39-55

[3] Dellaporta, S.L.; Wood, J.; Hick, J.B., Plant DNA minipreparation. Version II: Plant, Mol. Biol. Rep. 1, 1983, pp. 19-21

[4] Ferrer, M.; Eguiarte, L.E.; Montana, C., Genetic structure and outcrossing rates in Flourensia cernua (Asteraceae) growing at different densities in the South-western Chihuahuan Desert, Annals of Botany, Vol. 94, 2004, pp. 419-426

[5] Guyon, I.;Weston, J.; Barnhill, S.; Vapnik, V., Gene Selection for Cancer Classification using Support Vector Machines, Machine Learning, Vol.46, No. 1-3, 2002

[6] Hall, Mark A., Correlation-based Feature Selection for Machine Learning. PhD Thesis. University of Waikato, Department of Computer Science, Hamilton, New Zealand, 1998

[7] Hall, Mark A. & Geoffrey Holmes., Benchmarking Attribute Selection Techniques for Discrete Class Data Mining, IEEE Transactions On Knowledge And Data Engineering, Vol.15, No.3, 2003

[8] Jolliffe I.T., Principal Component Analysis, Series: Springer Series in Statistics, 2nd ed., Springer, NY, 2002

[9] Jain, A.K.; Duin, R.P.W.; Mao,J., Statistical Pattern Recognition: A Review, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.22, 2000, pp. 4-37

[10] Kokonenko, Igor., Estimating Attributes: Analysis and Extensions of RELIEF. Proceeding of the 7th European Conference on Machine Learning: ECML-94, 1994

[11] Mora-Urpí, J.; Clement C.; Patiño. V., Diversidad Genética en Pejibaye: I. Razas e Híbridos. IV Congreso Internacional sobre Biología, Agronomía e Industrialización del Pijuayo. Universidad de Costa Rica, 1993, pp. 11-20

[12] Mora-Urpí, J.; Arroyo, C., Sobre origen y diversidad en pejibaye. Serie Técnica Pejibaye (Guilielma). Boletín Informativo. Editorial de la Universidad de Costa Rica. Vol.5, No.1, 1996, pp. 18-25

[13] Porebski, S.; Grant, L.; Baun, B., Modification of a CTAB DNA extraction protocol for plants containing high polysaccharide and polyphenol components. Plant Molecular Biology Reporter Vol. 15, 1997, pp: 8-15

[14] Pearson, K., On Lines and Planes of Closest Fit to Systems of Points in Space. Philosophical Magazine, Vol.2, No.6, 1901, pp. 559-572.

[15] Ravishankar, K.V.; Anand L.; Dinesh M.R., Assessment of genetic relatedness among mango cultivars of India using RAPD primers, Journal of Horticultural Sci. & Biotechnology, Vol.75, 2000, pp. 198-201

[16] Robnik-Šikonja, M. & Kononenko, I. "Theoretical and Empirical Analysis of ReliefF and RReliefF" Machine Learning 53(1-2): 23-69, 2003

[17] Travieso, C M.; Briceño, J.C.; Vásquez J.L.; Vásquez, J.; Castillo, E., Automatic recognition system for pejibaye palm DNA using SVM. Recent Advances In Computer Engineering. Proceedings of the 2nd conference on European computing conference, 2008, pp. 262-266

[18] Vapnik, V.; Golowich, S.; Smola, A., Support vector method for function approximation, regression estimation, and signal processing. In M. Mozer, M. Jordan, and T. Petsche, editors, Advances in Neural Information Processing Systems 9, 1997 pp. 281—287.

[19] Williams J.G.K., DNA polymorphisms amplified by arbitrary oligonucleotide primers are useful as genetics primers” Nucleic Acids Research Vol.18, 1990, pp. 6531-6535