Feature selection of RAPD haplotypes for identifying Peach Palm (Bactris gasipaes) landraces 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 present work presents a robust system for the feature reduction, using Deoxyribonucleic Acid (DNA) primer. This system reaches up to 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 development of this application provides economic and computational advantages. When it is reduced the number of primers, this application reduces the economic cost, being a process so much cheaper, up to 11.24% from the initial process. On the other hand, the use of our supervised classification system is faster in order to do a method of origin denomination plant certification, due to reduce the dataset up to 11.24%.

Keywords: Dimensionality Reduction, feature selection, DNA analysis, supervised classification, SVM, Artificial Neuronal Network, Cluster analysis

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. This certification allows value to the seed grant, to be able to identify their origin and to predict their future genetic characteristics. In turn, this conducts to an increase in market value and an incentive for producers to conserve this genetic resource. It also will avoid to record spurious data.

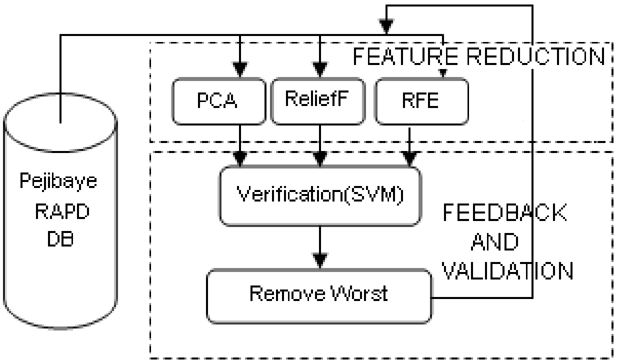

Fig.1. Proposed method for dimensionality reduction

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, see fig.1.

The pejibaye palm presents a large variety of morphology features as well as a large distribution over Central and South America, see figura 2. This palm (Bactris gasipaes) has been developed since the second half of last century in an atmosphere overwhelmed of controversies, some of which are the following:

1. The origin is unclear (still undefined), but the hypothesis of regionalization proposed by Mora Urpí [16],Clement et al. [5] is strongly supported.
2. Since this palm is considered a phytogenetic resource, it always was identified as a cultigens (the gene pool is the set of all genes of all individuals and during 30 years were considered only domesticated species in this gene pool). Today, knowing the existence of wild species their classification and phylogenetic relationships are modified.
3. The identification of this genetic material analyzed since the beginning as a landraces based on the weight and the form of the fruit (morphological features) and the geographic distribution, this criteria persists in the formation of taxonomic groups, both Henderson [11] and Mora Urpí [12] utilized this same criteria for the identification of the domesticated and the wild populations.

Due to the crop origin controversy [4] 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 of the different “landraces” (varieties) with a high concentration of oil, the races can enter new markets as oil for cosmetology product or as biofuel, in other words their use promotes 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 parameterization of DNA fingerprinting 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 germoplasm 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 (landraces) of Pejibaye (Utilitis - Costa Rica, Tuira - Panama, Putumayo - Colombia, Yurimagua - Peru, Tembé - Bolivia and Pará - Brazil), and each one has 13-16 samples with 89 RAPD haplotypes per sample see fig 2.

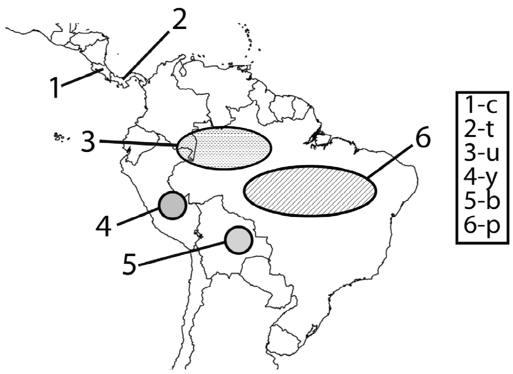

Fig.2. Origin and geographic distribution of The six Landraces of Peach Palm: 1. Utilitis(c), 2. Tuira(t) 3. Putumayo(3), 4. Yurimagua(y), 5. Tembé(b), 6. Pará(p).

Travieso et al. showed is possible to recognize with 100% assurance different pejibaye palm races using this 89 RAPD haplotypes [24]. 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 molecular systematic 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 amplified band analysis is one of those fingerprinting techniques based on PCR (Polymerase Chain Reaction) that are of easy manipulation, and of a detected high polymorphic level, and its feasibility to analyze big samples [17], [6], [8], [18], [19], and [26] (see Fig. 3). Furthermore the RAPD are widely applied in this cultivar in order to identify the phylogenetic relations and the molecular characterization of the domesticated or landrace populations [3], [19], [21], [22], [23]. In Peru the genetic structure of the oriental races has been evaluated (var. without spines) in agroforestry systems [1].

This study was realized over each individual's genetic material (87 samples), 30 primers from Operon Company series A, C, and 12 of these primers were classified as informative due to they generated polymorphic fragments with a high reproducibility rate, 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 haplotypes, from the six different Pejibaye races considered. From the beginning was unclear if the considered Boolean features were interdependent or not.

On the Fig. 3, 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).

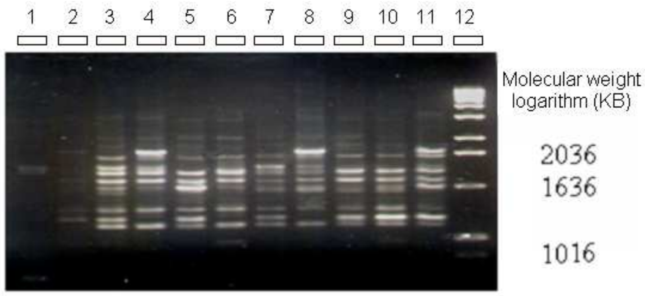

Fig.3. Some examples of Utilitis-Tucurrique pejibaye amplified DNA description, through the application of the PCR OPC-20 primer, with the RAPD technique.

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. [2][9]

The first selected filter was an unsupervised method: the principal components analysis (PCA) [12]. The second selected filter was an instance base attribute ranking scheme nominated ReliefF [14] that handle noise and works fine with multiple-class data sets.

The selected wrapper technique was the Recursive Feature Elimination [8], which evaluates the worth of an attribute by using an SVM classifier [25].

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 [18] for identifying patterns in data, and expressing the data in such a way as to highlight their similarities and differences [12]. Since patterns 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 [18][20] 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. [8]

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. [13]. 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 [2]. 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 [25] 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 [25]. The algorithm has scalable memory requirements and can handle problems with many thousands of support vectors efficiently.

In the next figure, we can see the detection of support vectors and the creation of a boundary, one per each class, because it is a bi-class classifier. In our implementation, we have built a multi-classes classification module, from this SVM light.

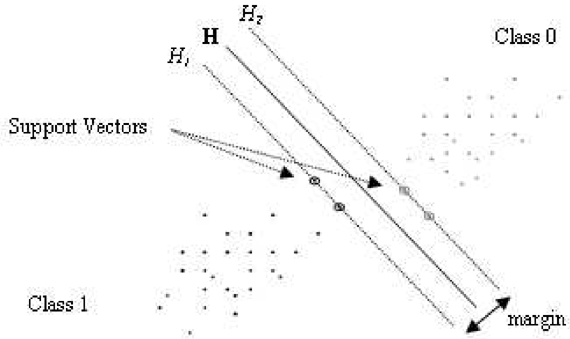

Fig. 4. Separate lineal Hyperplane in SVM

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 showed in figure 5.

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.

```
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')
```

Fig. 5. Reduce algorithm dimensionality

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%

5.4 Cluster Analysis

With the purpose of validating the results and verifying the relevance of the RAPD haplotypes obtained by the characteristics selection algorithm, we proceeded to review the data base throughout the analysis of conglomerates. The analysis of conglomerates is a method of hierarchy classification belonging to the unsupervised learning type.

The cluster analysis was employed to determine the discriminatory power of the ten attributes selected in the sample classification of pejibaye. The idea of analyzing the data using a technique of a different nature –unsupervised learning- to the one utilized in the characteristic selection, has as its purpose to discard that the obtained result in the selection process corresponds to a overtraining (overfitted) of the learning algorithm. On the other hand, it is worth mentioning that the cluster analysis technique is frequently utilized in biological data classification studies.

In the cluster analysis two fundamental parameters are defined, the similarity index and the grouping index. Since in our case the pejibaye samples are characterized with binary data, it is important to consider the similarity indexes that take this fact into notice, being the Jaccard index one of the most utilized. The Jaccard index expression is:

$$d = \frac{a_{ij}}{a_{ij} + b_{ij} + c_{ij}}$$

where:

$a_{ij} = \sum_{k=1}^{p} w_{ik} w_{jk}$ Number of attributes on objects w_i y w_j simultaneously present.

$b_{ij} = \sum_{k=1}^{p} w_{ik}(1 - w_{jk})$ Number of attributes present in the object w_i and absent in the object $w_{j.}$

$c_{ij} = \sum_{k=1}^{p} (1 - w_{ik}) w_{jk}$ Number of attributes absent in the object w_i and present in the object w_j.

As far as the grouping index is concerned, the average distance has been utilized on each group.

To corroborate with the hypothesis of the six landraces of pejibaye throughout the cluster analysis, a grouping algorithm was applied and the sample distribution for the last six groups was analyzed. From the sample distribution on those groups and their previous identification, a confusion table was constructed in order to determine the resulting classification percentage. It is important to clarify that the sample labels are not part of the grouping algorithm (unsupervised algorithm), they are only used to analyze the classification.

As a first step in the use of cluster analysis, the pertinence of utilizing this type of analysis was determine with the pejibaye data, for which de groping algorithm was applied considering the 89 attributes. The result is shown in the dendrogram of figure 6, where the tree cut with six groups highlights itself with the horizontal line. The table 4 corresponds to the sample distribution of each group, being the result a rating of 100%.

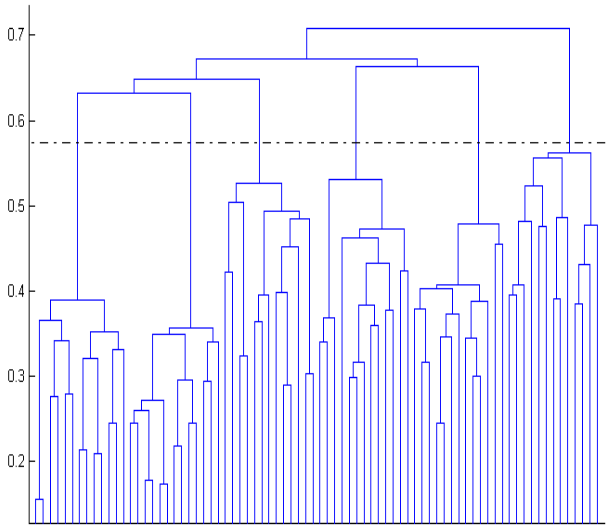

Fig. 6. Cluster tree for pejibaye samples characterized with 89 binary attributes.

Table 4. Data classification with cluster analysis

c	t	p	y	u	b
1c 0	1t 1	1p 2	1y 3	1u 4	1b 5
2c 0	2t 1	2p 2	2y 3	2u 4	2b 5
3c 0	3t 1	3p 2	3y 3	3u 4	3b 5
4c 0	4t 1	4p 2	4y 3	4u 4	4b 5
5c 0	5t 1	5p 2	5y 3	5u 4	5b 5
6c 0	6t 1	6p 2	6y 3	6u 4	6b 5
7c 0	7t 1	7p 2	7y 3	7u 4	7b 5
8c 0	8t 1	8p 2	8y 3	8u 4	8b 5
9c 0	9t 1	9p 2	9y 3	9u 4	9b 5
10c 0	10t 1	10p 2	10y 3	10u 4	10b 5
11c 0	11t 1	11p 2	11y 3	11u 4	11b 5
12c 0	12t 1	12p 2	12y 3	12u 4	12b 5
13c 0	13t 1	13p 2	13y 3	13u 4	13b 5
14c 0	14t 1		14y 3	14u 4	14b 5
15c 0	15t 1			15u 4	15b 5

Once the cluster analysis applicability was verified with the pejibaye data, the cluster algorithm was employed considering only the 10 selected attributes. The result of the algorithm shown as a dendrogram in the figure 7, obtains again a success rate of 100% for the six generated groups.

Finally, a grouping algorithm was utilized in order to analyze the pejibaye samples classification based on the transformed data; such transformation is generated by the SVM model.

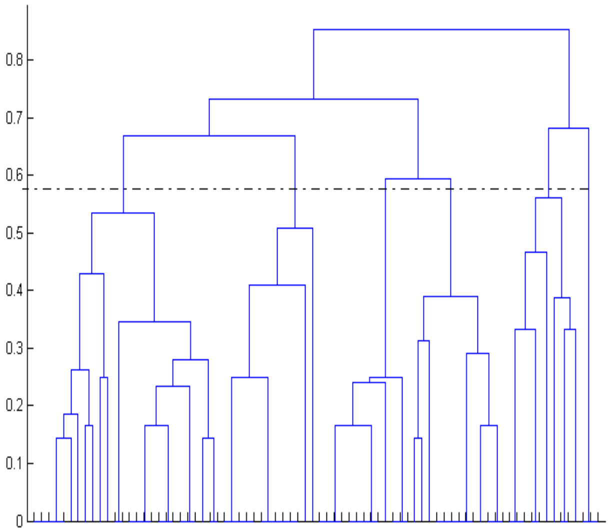

Fig. 7. Cluster tree for pejibaye samples characterized with 10 binary attributes.

From the characteristics reduction process the transformation equations for the 10 attributes were obtained, which were then applied to the pejibaye data in order to get a new data base with real values.

Since there is a new domain in the value of attributes, the Euclid distance was utilized in the grouping algorithm as a similarity index, instead of the Jaccard index. The result of the algorithm shown as dendrograms in the figure 8, obtains again a success rate of 100% for the six generated groups.

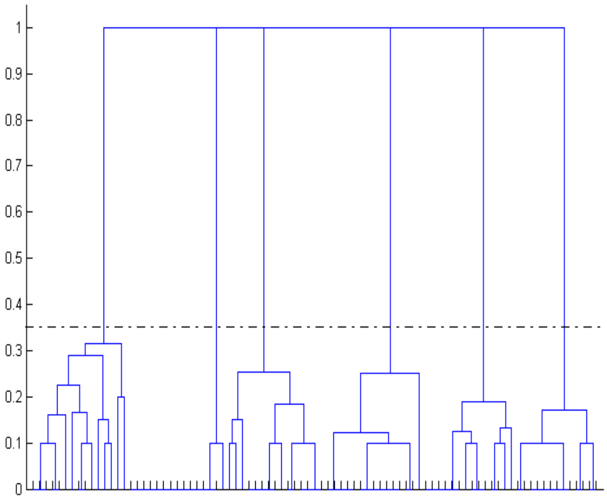

Fig. 8. Cluster tree for pejibaye samples characterized with 10 attributes obtained from SVM.

Furthermore it is evident that the data transformation increases the separation of the classes simplifying, its classification.

5.6 Verification system based on NN

The widely used Neural Networks techniques are much known on applications of pattern recognition. The perceptron of a simple layer establishes its correspondence with a rule of discrimination between classes, based on the lineal discriminant. However, it is possible to define discriminations for not lineally separable classes using multilayer perceptrons that are networks without refreshing (feed-forward) with one or more layers of nodes between the input layer and exit layer. These additional layers contain hidden neurons or nodes, are directly connected to the input and output layer.

An objective of this research is to generate an automatic classification system that allows identifying, based on its genome, the pejibaye race that allows an unclassified sample. Between options for the creation of this classification system there is the use of a normal neuronal network. Travieso et al. in [24] employed a neuronal network to analyze the utilized data based in the present study, utilizing every attributes in the data base. However, the results of such classification were not satisfying (65% success), probably because of having few samples of the data base in relation to the quantity of attributes utilized to characterize them.

Once the reduction of characteristics process has been realized and proven that the selected attributes set allow a correct classification of the pejibaye samples, we proceeded to analyze the data with a neuronal network for which a neural network multilayer perceptron (NN-MLP) Feed-Forward with Back-Propagation training algorithm was used, and with only one hidden layer (eight neurons). The number of input fits in with the reduced number of DNA elements, and the number of outputs with the Pejibaye palms landraces, see figure 9.

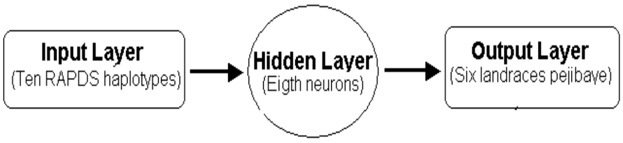

Fig. 9. Multilayer Perceptron Neural network structure used

In table 7, it is shown the success rates for NN classifier, for ten binary attributes. The obtained results applying the 10 selected attributes by the SVM are better in comparison with the ones obtained utilizing the complete data set. From the 10 executions made, with a 30% training sample, the minimum success percentage was of 94% and the maximum 100%, with a 97% average.

Table 7. Average result with the NN classifier

	c	t	p	y	u	b
C	15	1	0	0	0	0
T	0	14	0	1	0	0
P	0	0	13	0	0	0
Y	0	0	0	13	0	0
U	0	0	0	0	15	0
B	0	0	0	0	0	15

The values of the 10 selected attributes were transformed in order to be used with the neuronal network, generating a 100% success rate, with a training sample of just 20%. In a population of 15 instances, to require only a 20% of them, evidences the effectiveness and quality of the realized process.

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.

This useful tool can be used for: the identification of domesticated populations according with the products wanted in the improvement programs; that also reduces the repeated materials percentage that are found in the germplasm banks ex situ, with a high maintenance cost and to increase the genetic diversity with the regional introduction of wild species.

These good results suggest the use of this technique in the field of the molecular biology, as it should provide biotechnologists 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.

At the same time it allows to consolidate the trust in the RAPD markers, especially on these studies where much ignorance still exists as far as the taxonomy and the identification of domesticated and wild populations are concerned and the possible natural hybrids of this genetic resource, that could be of great use in the Southern cone and in Central America, and with the potential to expand into other continents like Africa, as a alternative cultivar.

The obtained transformation simplifies the characterizations of the pejibaye races, generating simpler dendrograms to analyze.

Converting the binary values into new ones, leaning on the equations generated by the SVM provides the opportunity to create very efficient algorithms in time and space for the classification of instances.

Even when the binary attributes suggested by the SVM allow to be used in combination of neuronal networks (NN), the transformation of which has generated better results.

7. Acknowledgment

This work was supported by the Laboratory in Automate and Intelligent Systems in Biodiversity of the University of Costa Rica and the AECID from the Spanish Government in 2010/01. Besides, this work was supported in part by the Spanish Government under Grants from "Programa José Castillejo" in 2008/09.

References:

[1] Adin, A.; Weber, J., Sotelo Montes, C., Vidaurre, C.H., Vosman, B. Smulders, M. J. M. 2004. Genetic differentiation and trade among populations of peach palm (*Bactris gasipaes Kunth*) in the Peruvian Amazon-implications for genetic resource management. Theor Appl Genet. 108:1564–1573

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

[3] Castillo, E., 2005. Evaluación de la Diversidad genética de cultivares regionales y locales de pejibaye (Bactris gasipaes) utilizando marcadores moleculares (RAPDs y AFLPs). Magister Scientia. Universidad de Costa Rica.

[4] 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

[5] Clement, C.R. ; Santos RP, Desmouliere SJM, Ferreira EJL, Neto JTF. Ecological Adaptation of Wild Peach Palm, Its *In Situ* Conservation and Deforestation-Mediated Extinction in Southern Brazilian Amazonia. PLoS ONE 4(2): e4564. doi:10.1371/journal.pone.0004564. 2009

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

[7] 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

[8] 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

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

[10] 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

[11] Henderson **A.** 2000.Bactris (Palmae). Flora Neotrópica Neotrópica 79-1181

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

[13] 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

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

[15] 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

[16] 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

[17] 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

[18] 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.

[19] 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

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

[21] Rodríguez D.P. Astolfi- Filho S., Clement C.R. 2004 .Molecular marker- mediated validation of morphologically defined landraces of Pejibaye (Bactris gasipaes) and their phylogenetic relationships. Genetic Resources and Crop Evolution 51: 871–882, 2004.

[22] Silva, C.C. 2004. Análise molecular e validação de raças primitivas de pupunha (*Bactris gasipaes*) por meio demarcadores RAPD. Dissertação de Mestrado, Univ. Fed. São Carlos/Univ. Fed. Amazonas, Manaus

[23] Sousa N. R., Rodríguez D. P., Clement C. R., Nagao E. O., Astolfi-Filho S**.** 2001. Discriminação de raças primitivas de pupunha (*Bactris gasipaes*) na Amazônia Brasileira por meio de marcadores moleculares (RAPDS). Amazónica 31: 539-545.

[24] 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

[25] 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.

[26] 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.