Soft Computing
https://doi.org/10.1007/s00500-020-05551-x

METHODOLOGIES AND APPLICATION

A database for automatic classification of gender in *Araucaria angustifolia* plants

Jefferson G. Martins[1] · Luiz E. S. Oliveira[2] · Daniel Weingaertner[2] · Andersson Barison[2] · Gerlon A. R. Oliveira[3] · Luciano M. Lião[3]

Abstract
Forests have been disorderly exploited, and many species are considered endangered. Some initiatives have been taken in order to prevent forests from being destroyed. A good alternative would be to plan a spatial distribution of plants, with higher number of females than males. Determining the gender of seedlings would provide important information for a possible strategy. Another common problem that researchers in this field very often face, in order to perform their experiments, is the lack of a representative database. To overcome this difficulty, we introduce a new database in this work, which is composed of nuclear magnetic resonance of adult *Araucaria angustifolia* plants. In order to gain better insight into this database, we have tested different strategies and classifiers. A first set of experiments took three classifiers trained to discriminate males from females considering the original database. A second round of experiments applied the genetic algorithm technique to select subsets of attributes based on single-objective and two-objective functions. After analyzing the achieved results, we have also proposed a new strategy based on statistical measures for selecting subsets from the attributes. A comprehensive set of experiments has shown that the proposed selecting strategy has achieved better performances, with an accuracy of 80.3% (AUC = 79.4). We believe that researchers will find this database a useful tool in their work on determining the *Araucaria angustifolia* gender. On the other hand, the proposed selecting strategy would be useful for reducing the complexity of databases and accelerating the process of building classification models.

Keywords Pattern recognition · Gender plant classification · Feature selection

1 Introduction

Natural resources have been misused. Minerals are widely employed in civil construction and several industries. Forests are being widely destroyed around the world. Rivers, seas and oceans are being polluted. Animals have been expelled from their original environments and left under improper living conditions. Many vegetable and animal species are at risk of extinction because they are not able to produce new individuals.

Under the described scenario, there is the *Araucaria angustifolia* specie. Popularly known as araucaria or pine of Paraná, it is an endemic conifer in the Brazilian southern and southeastern region. Besides the extraction of its wood, two facts reinforce this species as being endangered. Firstly, Araucaria is a dioecious plant, and only the female plants produce seeds. Consequently, there must be individuals from both genders, male and female, close enough to each other in order to allow a possible fecundation. Such phenomenon happens by pollination, when pollen is transferred from males to females, mainly through the wind. Secondly, seeds are used as food and have an important economic influence on people who collect and sell them in order to make a living (Freitas et al. 2009; Guerra et al. 2002).

In natural populations of *Araucaria angustifolia*, where there is no human interference, the relationship between

This research has been supported by The National Council for Scientific and Technological Development (CNPq) [grants 301653/2011-9, 471050/2013-0].

✉ Jefferson G. Martins
martins@utfpr.edu.br

[1] Federal University of Technology - Paraná (UTFPR), Rua Cristo Rei, 19, Toledo, PR 85902-490, Brasil

[2] Federal University of Paraná (UFPR) , Rua Cel. Francisco H. dos Santos, 100, Curitiba, PR 81531-990, Brazil

[3] Federal University of Goiás (UFG), Av. Esperança, s/n, Goiânia, GO 74690-900, Brazil

Published online: 22 January 2021

females and males is 1:1 (Bandel and Gurgel 1967; Carvalho 2003; Zanon et al. 2009). However, it is possible to find a slight predominance of males (Atanazio et al. 2018; Bandel and Gurgel 1967; Carvalho 2003; Zanon et al. 2009). In this context, some studies with a focus in the maximization of seed production define the ideal conditions to maximize the earnings of planted areas with proportion of females and males in 80%-20% (Zanette 2014), and 70–30% (Zanette et al. 2017).

Determining the sex of seedlings would provide important information for a possible strategy to plan their spatial distribution in fields and forests. A good option could be planting a higher number of females than males. It could solve both previously presented problems, by simply increasing the seed production. The existence of more seeds would provide a higher number of new individuals, making this species much more attractive for reforestation purposes and, at the same time, preserving the existent areas. Simultaneously, seed collectors would have higher earnings.

In spite of the perspectives mentioned before, there is no technique for identifying the sex of the seedlings before they express their sexuality. This takes place after twenty years when the trees are natives. The same degree of maturity can be developed after twelve or fifteen years, when the individuals are planted. However, if plants are planted close to each other, it could take up to 26 years (Freitas et al. 2009; Guerra et al. 2002).

Some studies have unsuccessfully tried to find out phenotype features related to the *Araucaria angustifolia* individuals' gender. Bandel and Gurgel (1967) were not able to define a relationship between morphological features of the individuals and their genders. Amaral et al. (1971) studied 36 individuals, 18 males and 18 females. They analyzed a possible influence of growing rates and wood density, but they did not find any relation between them and the individual genders. Males and females presented the same behavior and patterns for both analyzed features.

Zanon et al. (2009) studied the diameter sizes in a population of 4888 individuals, being 2754 males and 2134 females. The authors found a higher number of females in the classes with bigger diameters and a higher number of males in the classes with smaller diameters. However, there was no significant difference between the medium diameter from one class to another, having the diametrical distribution followed a normal distribution.

Stefenon et al. (2008) assessed the influence of genetic diversity levels in natural and planted populations of *Araucaria angustifolia*. The study was conducted on 512 seeds, where 192 were collected from five plantations and 320 were collected from five natural populations. They used Amplified Fragment Length Polymorphism (AFLP) and nuclear microsatellites, in order to assess the usefulness of planted forests in programs of species' genetic resource conservation. In general, the genetic structures found were not capable of differentiating planted from natural populations.

Murakami (2003) performed DNA (deoxyribonucleic acid) tests in young leaves of 20 adult individuals, 10 males and 10 females. The author applied Random Amplified Polymorphic DNA in order to find genes that could be used to identify the gender. No element was found that could be useful to distinguish males from females. The author concluded that a possible region of the genome involved in determining the gender of individuals is relatively restricted or the gene that controls the sexual expression could be found in a genomic region, not very frequent, while recombination events take place. Murakami reassured that *Araucaria angustifolia* genders cannot be identified by morphological or physiological features, and there are no sex chromosomes or differences in karyotype that could help to differentiate males from females (Murakami 2002, 2003).

Carvalho (2012) suggested that possible plants' metabolic differences could support a gender definition, furthermore, those supposed differences would be a consequence of genetic and environmental variations. The author used a total of 61 samples: 26 males, 26 females and 9 undefined genders. The author was not able to discriminate males from females using such a theory as supervisioned chemometric analysis. Oliveira (2016) also analyzed genetic and environmental variations and their influence on metabolic differences. The author used 18 from the 61 original samples from Carvalho (2012) and 80 new samples with Infrared and Nuclear Magnetic Resonance (NMR). The author concluded that none of them were able to differentiate males from females. Oliveira (2016) reassured that genetic and environmental differences are intrinsically related to each other. Usually, individuals close to each other have a higher probability of generating new ones, which creates a microenvironment in which there is a lower genetic variation.

So far, grafting has been the base for techniques that try to identify the sex of seedlings before a definitive transplant. Herein, we have two different types of stems: orthotropic and plagiotropic. Orthotropic stems grow vertically and bifurcations are uncommon. Although they have a normal crown, it is difficult to use them in large scale because they are rare. Plagiotropic stems have branches and produce seeds precociously, but they also grow horizontally and have a short life cycle. Grafting demands a great level of knowledge from botanists and, even though it grants increased productivity and more homogeneous forests, it reduces the plants' genetic variability by an unnatural process (Constantino and Zanette 2016; Wendling 2011; Zanette et al. 2011).

A major challenge to pursue any pattern recognition research is the lack of consistent and reliable databases. To overcome this difficulty, we have introduced a database in this work composed of NMR of adult *Araucaria angustifolia* plants. We have also presented a new strategy, based

on statistical analysis, in order to select subsets of attributes. We have evaluated the proposed selecting strategy on the new database through a comprehensive set of experiments. Our best result was an accuracy of 80.3% ($\sigma = 0.0$, AUC = 79.6), using selected subsets of only 13.7 ($\sigma = 6.0$) attributes. The database introduced in this work will make future benchmark and evaluation possible. The proposed strategy to select attributes can also be useful to reduce the complexity of any database and accelerate the process of building classification models.

This paper is structured as follows: Section 2 presents the used materials and methods. Section 3 reports our experiments and discusses the results. Finally, Sect. 4 concludes the work.

2 Materials and methods

Here, we briefly present the NMR technique (Sect. 2.1) and the proposed database (Sec. 2.2). Section 2.3–2.6 describes strategies applied to identify the gender of *Araucaria angustifolia* plants. In order to show that choices, in terms of experiment settings, do not have a significant impact on the accuracy, each experiment was performed three times. The low values for standard deviation demonstrate the stability of our methods and results.

2.1 Nuclear magnetic resonance

Plants are affected by their gender and such influence carries on distinct metabolic behaviors. Thus, further studies could provide a good alternative to distinguish between males and females. In spite of there being other options, such as Mass Spectrometry, Raman and Infrared, we have chosen the well-known Nuclear Magnetic Resonance (NMR) technique, because of its superiority in terms of repeatability and accuracy. It simultaneously provides information about a set of metabolites in a single run, without specifying a particular element. In addition, NMR can also be used to identify structures of metabolites and quantify their absolute and relative concentration measures (Nicholson and Lindon 2008).

NMR spectroscopy was originated with the development of pulsed Fourier transform (Ernst and Anderson 1966) and the concept of multidimensional NMR spectroscopy (Jeener and Broekaert 1967). Nuclear magnetism and NMR spectroscopy are the manifestations of a nuclear spin angular momentum (Cavanagh et al. 1995). Good references can be found in Butterworth et al. (2001), Cavanagh et al. (1995), and Rith and Schafer (1999).

Ernst and Anderson (1966) reassure that the frequency response function and the unit impulse response of a linear system form a Fourier transform pair. Both of them contain the exact same information. In magnetic resonance, the first one is usually called spectrum, while the second is represented by the free induction decay. The authors have shown, theoretical and experimentally, that Fourier transform spectroscopy is able to improve the sensitivity of the magnetic resonance in up to a factor of 10, in sensitivity, in a restricted time, and in up to a factor of 100, in time, for a given sensitivity.

NMR spectroscopy and x-ray crystallography are the only techniques capable of determining three-dimensional structures of macromolecules in an atomic resolution (Cavanagh et al. 1995). NMR spectroscopy is an analytical tool that can provide information on the chemical nature of metabolites, and, if performed in vivo, it can also provide non-invasive information on cellular environments and metabolism (Suarez et al. 1999). Figure 1 presents a representative ^{1}H-NMR spectrum of *Araucaria Angustifolia* needles. The intensity of the peaks in the NMR spectra is proportional to the compounds that originated them. The challenging problem of identifying the gender of such plants, by using NMR spectroscopy, is to find the peaks having different intensities in the two groups of spectra.

Analyzing spectra from different perspectives is important, when using NMR. A possible way to do that is by using buckets, where each bucket summarizes a set of points. Thus, the usage of buckets decreases the resolution, due to a set of points, which are reduced to only one point. In addition, buckets reduce the size of spectra, making chemometric calculations faster, as well as decreasing the redundancy of variables. However, the most important advantage of using buckets is the alignment of spectra (Euceda et al. 2015; Keun et al. 2003; Rinnan et al. 2009; Sousa et al. 2013; Vu and Laukens 2013).

Different alternatives are available to determine how buckets will be defined. We can determine whether all buckets will have the same size or each one will have a specific size. Another important question is to define how a set of points will be summarized. A simple way to produce buckets is to sum the point intensities. It is not a simple task to define the ideal bucket size. Good references can be found in Sousa et al. (2013), and Vu and Laukens (2013).

2.2 Database

A major challenge to pursue research involving pattern recognition is the lack of consistent and reliable databases. Usually researchers perform their experiments on particular databases that are not accessible to the others. Therefore, it is difficult to develop new studies and reproduce the previous ones. Additionally, as we have shown in Sect. 1, most databases relating to NMR or discriminating the gender of *Araucaria angustifolia* plants, usually contain few instances.

It is worth noting that there is no prohibitive aspect concerning the gathering of more samples to create a database.

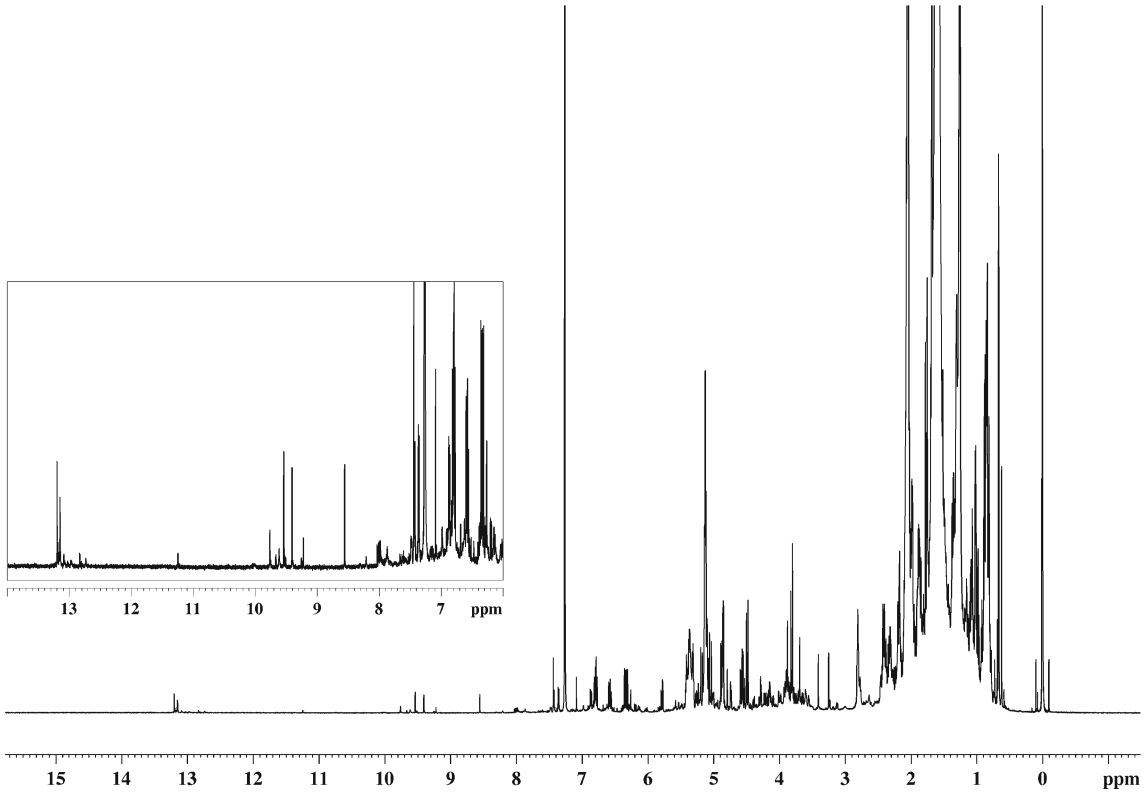

Fig. 1 ^{1}H-NMR spectrum of needles of *Araucaria angustifolia* needles ($CDCl_3$, 600 MHz). *Y*-axis has arbitrary units

However, there are some restrictions in terms of logistic and technical limitations that could, in the least case scenario, raise the final cost. Such restrictions include the flowering period, a botanist to identify male and female plants, as well as a set of procedures to be followed and materials to be used in the process.

To overcome such a difficulty, we have introduced a new database in this work, composed of only two classes, but useful for any research carried on to identify the gender of *Araucaria angustifolia* plants. This database was built in collaboration with the Laboratory of Nuclear Magnetic Resonance at the Federal University of Paraná (UFPR) and the Laboratory of Nuclear Magnetic Resonance at the Federal University of Goiás (UFG), both of them in Brazil. It is available upon request for research purposes.[1]

In order to create the database, during the flowering season (July 2015), we collected needles from adult *Araucaria angustifolia* plants, in the metropolitan region of Curitiba, state of Paraná, Brazil. A botanist visually determined the gender, based on the presence of flowers in females and strobilus in males (Fig. 2). From the 76 plants used, 35 were males, while 41 were females. Thus, the introduced database is slightly imbalanced, with six (7.9%) more female than male individuals.

At this stage, several pre-treatments were evaluated. In the best one, we macerated the needles with liquid nitrogen until a fine powder was obtained. In our analysis, 30.0 mg of needle powder was put in a 2-mL Eppendorf recipient. After that, 600 μL of deuterated chloroform solvent was added. Extracts were taken for 10 minutes under ultrasound assistance. Finally, after a centrifugation process, the liquid fraction was put in a 5-mm diameter tube.

Analysis concerning ^{1}H NMR spectra was carried on in a BRUKER Avance III 600 MHz. In order to proceed with the spectra acquisition, we used the Bruker's Quadruple (QXI) Resonance Probes equipment and the following set of parameters: relaxation delay of a second, receiver gain of 145.6, spectral window of 30 ppm (parts per million), 256 scans and 90^{O}-pulse sequence zg (zero-go).

Spectra processing applied the Bruker's TopSpin 3.1 and Amix® v. 3.9.14 software systems. By using the first one, the baseline was automatically adjusted with an 'absn' function. A manual phase correction was performed. Taking tetramethylsilane (TMS) as reference, spectra were calibrated to zero. While performing the acquisition, spectra were smoothed by a function in which the factor 'lb' was set to 0.3. The Amix® v. 3.9.14 software was used to define

[1] https://web.inf.ufpr.br/vri/databases/araucaria-nuclear-magnetic-resonance/.

Fig. 2 *Araucaria angustifolia*: **a**, **b** male strobilus, and **c**, **d** female flowers

the buckets. Finally, noise and solvent spectral regions were eliminated (Sousa et al. 2013).

Taking the original spectra, we have analyzed different strategies to define the size of the buckets, while looking for a perspective that could improve accuracy. We have tested rectangular buckets, as well as the optimized bucketing method for NMR spectra presented by Sousa et al. (2013). For both strategies, a new value for each bucket was taken as the sum of intensities of those points belonging to that bucket. The best results were produced by individuals composed of 260 rectangular buckets taken from the 3189 original attributes. After a reduction of more than 12 times, each bucket was taken as an attribute and each individual was represented by a set of only 260 attributes.

2.3 Classification models

After identifying the 260 buckets, we used them to build our classifiers. While carrying out the experiments, we analyzed different machine learning algorithms: kNN (k-Nearest Neighbors), LDA (Linear Discriminant Analysis) and Support Vector Machine (SVM). After that, while performing the following experiments, we took the previous results as a parameter and assessed the real improvements brought by each tested strategy.

For all classifiers, the data were normalized by the Z-Score method in order to rescale each column to the [-1...+1] interval. Different values were tested for the kNN neighborhood size. For the SVM, different kernels were tried, but the Gaussian kernel produced the best results. The kernel parameters γ and C were defined empirically through a grid search on the validation set.

LDA and kNN, different from SVM, do not provide probabilities for each possible class. However, such information is necessary in order to calculate the Area under the Curve (AUC). These probabilities were provided by normalizing the amount of votes for each class by the total of neighbors taken in each variation in k. For the LDA, the value generated

by each discriminant function was divided by the sum of all of them.

The introduced database represents an important improvement concerning the problem of identifying the gender of *Araucaria angustifolia* plants and any experimental protocol can be used with it. In our experiments, different strategies were tried out, but *leave-one-out* produced the best results, due to the low number of instances. Thus, each instance from the original database was used in the testing set, while the training set was composed of the remaining 75.

This strategy is identified as 'A'. As we have a 2-class problem, different measures can be calculated from a 2×2 confusion matrix (Fig. 3), according to Eqs. 1–4, where FP, FN, TP and TN stand for False Positive, False Negative, True Positive and True Negative, respectively. This representation is useful while analyzing the frequency of mistakes made at each class or pair of classes in terms of Type I (FP) and Type II (FN) errors. In spite of the mentioned possibility of planning a spatial distribution of males and females in fields and forests, and the fact that there is a preference for a higher number of females than males, while performing the experiments, female was always considered as the positive class and Type I errors should be minimized.

As stated previously, the introduced database is slightly imbalanced. So, it is interesting to evaluate the achieved results, through different measures other than accuracy. Accuracy (Eq. 1) calculates how many real positive instances were labeled as positive (TP) and how many real negative instances were labeled as negative (TN). In other words, accuracy is evaluated by counting the instances that were correctly classified, i.e., females and males truly classified as females and males, respectively. Precision (Eq. 2) assesses how many individuals predicted as positive were actually positive, i.e., how many instances classified as females belonged to the female class. This is a good measure to be used in cases where the cost of FP is high, which for this case would be to classify males as females. Recall (Eq. 3) calculates how many real positive instances were labeled as positive (TP) by a model, i.e., how many instances classified as females were actually females. By applying the same understanding, Recall would be a metric model to be used for selecting the best model, when there is a high cost associated with FN, which for this case would be to classify females as males. On the other hand, F1 Score (Eq. 4) is useful when we are seeking a balance between precision and recall, and where there is an uneven class distribution.

ROC (Receiver Operating Characteristic) curves are also attractive, because they are insensitive to changes in the class distribution. If the proportion of positive to negative instances changes in a test set, the ROC curves will not change. The AUC (Area Under the ROC Curve) of a classifier is a numeric value that represents the ability of a classifier to rank positive instances relative to negative instances, i.e., it is equivalent to

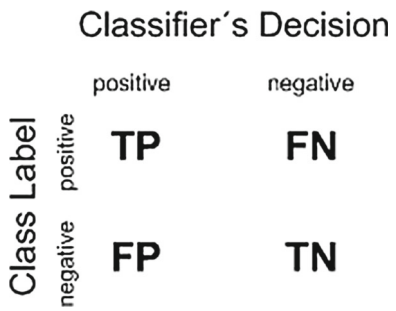

Fig. 3 2×2 confusion matrix

the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance (Fawcett 2006).

The evaluation of those metrics with values in the interval [0...100] is presented for all the experiments.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} * 100 \quad (1)$$

$$Precision = \frac{TP}{TP + FP} * 100 \quad (2)$$

$$Recall = \frac{TP}{TP + FN} * 100 \quad (3)$$

$$F1\ Score = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (4)$$

2.4 Genetic algorithms

Optimization is in almost every aspect of life, inspiring the development of numerical procedures that satisfy prevailing constraints by systematically choosing the values of variables from within an allowed set (Arqub and Abo-Hammour 2014). It is also a branch of mathematics concerned with obtaining the conditions that give the best solution represented by an extreme (maximum or minimum) value of a real function, or many functions, under given circumstances to find feasible solutions to real-life problems (Sahab et al. 2013).

Modern optimization techniques have been emerged as powerful and popular methods for solving complex real-life optimization problems (Sahab et al. 2013). Among them, we can enumerate genetic algorithms (Holland 1975; Goldberg 1989a; Goldberg et al. 1989), simulated annealing (Kirkpatrick et al. 1983), particle swarm optimization (Kennedy and Eberhart 1995), ant colony optimization (Dorigo 1992; Dorigo et al. 2006; Blum 2005), bee colony optimization (Teodorović and Dell'Orco 2005), harmony search algorithm (Geem et al. 2001), firefly algorithm (Yang 2010), cuckoo search (Yang and Deb 2009) and bat algorithm (Yang and He 2013).

The well-known genetic algorithm (GA) is an optimization technique based on the principles of genetics and natural selection. GA has been widely used to build good solutions

for different types of problems in different disciplines. Good references for GA can be found in Arqub and Abo-Hammour (2014), Holland (1975), Goldberg (1989a), Goldberg et al. (1989) and Sastry et al. (2005). Looking to increase the accuracy achieved for the recognition of gender of the *Araucaria angustifolia* plants, we have used GA to select subsets from the 260 original attributes. GA was a useful tool in the process of finding subsets of attributes with higher relevance for this problem.

In the GA technique, once the problem is encoded in a chromosomal manner and a fitness function is defined to measure and discriminate good solutions from bad ones, we start evolving solutions to the search problem. In this process, there is an initialization step and a circular repeating subprocess with the following steps: evaluation, selection, crossover, mutation and replacement. Such a subprocess goes on until a solution that satisfies the stop condition is met. It repeatedly modifies a population of individual solutions and randomly selects individuals on the basis of fitness value from the current population to be parents and uses them to produce the offspring for the next population.

Evaluation is based on a fitness function. The design of fitness function is very essential in genetic algorithm as the desired output depends heavily on it. The fitness value of each individual is computed by applying the fitness function to it. A fitness function is an application-specific objective function used to evaluate relative effectiveness of the potential solutions. For the standard optimization algorithm, it is known as objective function.

Crossover and mutation perform two different roles. Crossover (like selection) is a convergence operation which is intended to pull the population toward a local minimum/maximum. Mutation is a divergence operation. It is intended to occasionally break one or more individuals of a population out of a local minimum/maximum space and potentially discover a better minimum/maximum space. The parameters that guide these two operations are defined empirically.

The heart of GA is its fitness function and a set of parameters that allows to generate and select promising solutions. However, such GA parameters are problem specific, and hence, success depends on the choice of them (Arqub and Abo-Hammour 2014; Holland 1975; Goldberg 1989a; Goldberg et al. 1989; Sastry et al. 2005). Taking it into account, we evaluated different combinations of parameters. For each combination, we performed three runs from which we took the average and standard deviations. While presenting our results, only the best ones are discussed. It is worth noting that the standard deviation was calculated for the results achieved in each GA generation, and the general low resultant values demonstrated the stability of our methods and results.

It is known that dissimilarity among the individuals is necessary in order to generate a diverse new population in each GA generation. Thus, a stop condition was empirically defined based on the similarity degree of the individuals in the population. We identified that, at certain similarity degree, GA was not able to generate new individuals capable of producing an increase in accuracy and a decrease in the selected attributes.

All GA runs started and kept populations of 1000 individuals in all their generations. Each individual was represented by a numeric array with 260 columns filled with ones and zeros. Columns with the value one indicated that such features were present in an individual. On the other hand, columns with zeros indicated that such features were not present in the respective individual. In this case, each GA individual represented an evaluation of the original database filtered only by those attributes set with ones among the original ones. In other words, the *leave-one-out* strategy was applied for each new configuration of each individual in the GA population.

In the first generation, each individual had its initial configuration randomly defined, i.e., its 260-column arrays were filled with zeros and ones. Hereafter, in each new GA generation, mutation and crossover operations were performed on individuals in order to create new ones. Such operations were performed only if a randomly given number in the [0...1] interval was higher than specific threshold values, i.e., 0.5 and 0.2 for crossover and mutation, respectively. As previously mentioned, such values were defined empirically through tests with a wide range of configurations in our experiments. In fact, the mutation probability is higher than the reported in the literature, but in our experiments this value produced the best results. After creating a new population, the individual with the worst evaluation was always replaced by the best-evaluated individual from the previous generation. Such an option granted a stable increase in accuracy.

While using the GA technique, we assessed two different objective functions. In the first one, named strategy 'B,' the objective function was given by Eq. 1 and took into account only accuracy maximization. While evaluating such a strategy, we identified a stagnation concerning an increase in accuracy and a decrease in selected attributes (nf). Thus, experiments based on a new strategy, named 'C,' were performed. Here, a new criterion concerning nf was added to the objective function used in 'B'. Equation 5 illustrates the new function in which we have weights for the real accuracy (Eq. 1) and the percentage of nf attributes (Eq. 6) selected from the N ($N = 260$) original attributes. Different values for the weights pa and pf were tried out to produce the weighted accuracy tw, but they were always complementary to each

other (Eq. 7).

$$tw = \text{Accuracy} * pa + tf * pf \tag{5}$$

$$tf = \left(1 - \frac{nf}{N}\right) * 100 \tag{6}$$

$$pf = (1 - pa) \tag{7}$$

2.5 Pre-selecting the initial configuration for the GA population

When analyzing the previous results, we identified a delay of 150 to 200 GA generations until the accuracy could be stabilized. Taking this into account, we looked for alternatives to perform a pre-selection of attributes to be used in order to initialize individuals of the first GA population. Such a strategy has been shown useful, mainly for an early convergence of accuracy in terms of GA generations, as well as a decrease in computational complexity and cost to build classification models.

The achieved results corroborated with the well-known 'curse of dimensionality,' for which the performance of similarity measures and classification models tends to degrade as fast as the dimensionality of the data increases (Houle et al. 2010). In other words, when working with problems for which the number of parameters far exceeds the number of samples to learn from, built models are not able to distinguish elements from distinct classes (Kuo and Sloan 2005; Worzel et al. 2007).

In addition, when dealing with complex problems, a high number of attributes can cause troubles in terms of scalability (processing, memory, time etc.). At the same time, noisy, irrelevant or redundant attributes can confuse the learning algorithm, as well as hide a distribution of small sets of relevant attributes and harm building models and their accuracy.

The pre-selection of attributes was based on two hypotheses. For the first one, instances belonging to a same class would be close to each other. At the same time, instances belonging to a same class would be far from those belonging to other classes. Consequently, there would be lower variances internally to each class and higher variances when the whole database was considered. For the second hypothesis, class' centroids given by vectors of medium values calculated for each column would be far from each other.

Based on the previously stated hypotheses, a diverse set of statistical analysis was carried out in order to select a subset of attributes considering only average and standard deviation calculated in columns, classes and the whole database. For Eq. 8 to 13, we have N (N = 260) columns (attributes), M (M = 76) instances, M_c instances for each class c, and $v_{i,j,c}$ represents the original value for each attribute. Equations 8 and 9 provide attribute vectors with averages $\mu_{j,c}$ and standard deviation $\sigma_{j,c}$ calculated for each column j of a class c, respectively. In Eq. 10, σ_c^* represents the average of standard deviation calculated for each column of a class c. Equations 11 and 12 provide attribute vectors with averages μ_j and standard deviation σ_j calculated for each column j when the whole database is considered. Finally, in Eq. 13, σ^* represents the average of standard deviation calculated for each column when the whole database is considered.

As stated previously, in Sect. 2.4 each individual of the first GA generation had its attribute configuration randomly defined, i.e., its 260-column arrays were filled with zeros and ones. In the present strategy, all of the 1000 individuals were also represented by 260-column arrays filled with zeros and ones. However, in the first GA generation, only the selected attributes were set with ones and the remaining ones were set with zeros. As a result, all individuals in the population were equal to each other, but a diversified population was produced by GA operations randomly applied during each run throughout the GA generations, producing distinct partial and final states. This way, even those previously unselected attributes in the first generation could be included throughout the following GA generations. The low values for standard deviation reassured the GA technique robustness and accuracy stability.

A diverse set of arrangements of the conditions presented in Eqs. 8 to 13 were tried out. Best results were achieved by taking all of the 1000 individuals in the first GA generation, initialized with only a 22-attribute subset, which simultaneously satisfied both conditions (Eqs. 14 and 15). In other words, only the 22 selected attributes were set with ones, while all the remaining ones were set with zeros.

$$\mu_{j,c} = \frac{\sum_{i=1}^{M_c} v_{i,j,c}}{M_c} \qquad j = [1, N], c = [1, 2] \tag{8}$$

$$\sigma_{j,c} = \sqrt{\frac{\sum_{i=1}^{M_c} (v_{i,j,c} - \mu_{j,c})^2}{M_c}} \qquad j = [1, N], c = [1, 2] \tag{9}$$

$$\sigma_c^* = \frac{\sum_{j=1}^{N} \sigma_{j,c}}{N} \qquad c = [1, 2] \tag{10}$$

$$\mu_j = \frac{\sum_{i=1}^{M} v_{i,j,c}}{M} \qquad j = [1, N], c = [1, 2] \tag{11}$$

$$\sigma_j = \sqrt{\frac{\sum_{i=1}^{M} (v_{i,j,c} - \mu_j)^2}{M}} \qquad j = [1, N], c = [1, 2] \tag{12}$$

$$\sigma^* = \frac{\sum_{j=1}^{N} \sigma_j}{N} \tag{13}$$

$$|\mu_{j,1} - \mu_{j,2}| > \frac{\sum_{l=1}^{N} |\mu_{l,1} - \mu_{l,2}|}{N} \tag{14}$$

$$\sigma_{j,c} > \sigma_c^* \qquad j = [1, N], c = [1, 2] \tag{15}$$

From Eq. 14, we have a condition in which the distance between medium values $\mu_{j,c}$ of a column j for classes c ($c = [1, 2]$) must be higher than the average distance calculated for all columns l ($l = [1, N]$). In other words, such a condition defines that columns j that have medium values (class centroid) $\mu_{j,c}$ farther from each other are selected.

Equation 15 defines that a column j is selected if its standard deviation $\sigma_{j,c}$ calculated for instances belonging to a class c is higher than the average standard deviation σ_c^* ($c = [1, 2]$). σ_c^* is calculated for all columns l ($l = [1, N]$), but internally to each class c. This condition determines that columns that have values very close to each other are useless in the process of identifying or distinguishing individuals from different classes, and only columns with higher dispersion in their values are selected.

Hereafter, all the previous experiments performed for the original 260-attribute database were repeated. In strategy 'A,' only the 22 selected attributes were used. On the other hand, in strategies 'B' and 'C,' all of the 1000 individuals were represented by 260-column arrays filled with zeros and ones, but in the first GA generation only the 22 selected attributes were set with ones and the remaining ones were set with zeros.

2.6 Multiple classifier systems

When looking for alternatives to improve our previous results, we have also evaluated two alternatives for Multiple Classifier Systems (MCS) while selecting attribute subsets. The first one was the Random Forest Classifier (RFC). RFC is a meta-estimator that fits a number of decision tree classifiers on various sub-samples of the database and uses averaging to improve the predictive accuracy and control over-fitting. The sub-sample size is always the same as the original input sample size, but samples are drawn with replacements (Pedregosa et al. 2011).

The Gradient Boosting Classifier (GBC) was also used for classification. GBC builds an additive model in a forward stage-wise fashion; it allows for the optimization of arbitrary differentiable loss functions. As we are dealing with a binary classification, a single regression tree is fit on the negative gradient of the binomial or multinomial deviance loss function (Pedregosa et al. 2011).

3 Results

Table 1 presents the best results for the original database with 260 columns. In this case (strategy 'A'), only one run was performed and the standard deviation (σ) was not calculated. For kNN, the best k is presented. SVM presented slightly higher results and was elected as the machine learning algorithm to be used in the following experiments.

Taking only accuracy maximization (strategy 'B') and the SVM classifier into account, the best results were: 61.2% ($\sigma = 7.8$) of accuracy; 62.1% ($\sigma = 8.4$) of precision; 63.5% ($\sigma = 8.9$) of recall; 62.7% ($\sigma = 8.4$) of f1 score; and 65.2 ($\sigma = 5.5$) for AUC. These results were reached in the 53^{a} GA generation by using individuals composed of 124.3 ($\sigma = 4.7$) selected attributes. The achieved minimum (min), maximum (max), average (avg) and standard deviation (std) values for these measures are shown in Table 2.

Table 1 Best single classifier results (in %) on the original database using strategy 'A'

Classifier	Accuracy	Precision	Recall	F1 Score	AUC	k
SVM	60.5	62.2	68.3	65.1	67.3	–
k-NN	60.5	62.8	65.9	64.3	60.8	1
LDA	57.9	60.0	65.9	62.8	57.4	–

Table 2 Best SVM results (in %) on the original database using strategy 'B': minimum, maximum, average and standard deviation

Classifier	Accuracy	Precision	Recall	F1 Score	AUC
Mim	50.0	53.3	51.4	52.9	58.5
Max	69.7	72.5	73.2	71.6	71.8
Avg	61.2	62.1	63.5	62.7	65.2
Std	7.8	8.4	8.9	8.4	5.5

3.1 Multi-objective genetic algorithm

Figure 4a presents the evolution of weighted accuracy (Eq. 5) for some tested arrangements for the weights *pa* and *pf*, while using strategy 'C'. Legend values represent the *pf* weights used in each set of experiments. The other ones are not presented in the whole report, because they reached intermediate results.

Figure 4a, b shows that higher values for *pf* imposed a reduction in selected attributes. Such values also boosted the weighted accuracy used as objective function by the GA technique to select subsets of attributes throughout its generations. That fact can be explained by both Eqs. 5 and 6. From Eq. 6, *tf* tends to 1 as *nf* tends to zero. Thus, lower values for *nf* and higher values for *pf* produce higher values for the weighted accuracy, which almost has no influence from the real accuracy (Eq. 1).

While Fig. 4a shows the expected increase in weighted accuracy, Fig. 4b, c depicts a higher instability. Sometimes, the weighted accuracy improved by a new smaller subset of attributes, but at other times it happened with a new bigger subset of attributes that provided an increase in real accuracy (Eq. 1).

By using individuals composed of 14.3 ($\sigma = 2.5$) attributes in the 284^{a} GA generation with *pa* and *pf* values of 0.5, the best results were: 79.4% ($\sigma = 0.6$) of accuracy; 77.5% ($\sigma = 0.7$) of precision; 87.0% ($\sigma = 1.1$) of recall; 82.0% ($\sigma = 0.6$) of f1 score; and 79.2 ($\sigma = 1.4$) for AUC. The best values for weighted accuracy were achieved by using balanced values for *pa* and *pf*. Such values neither privi-

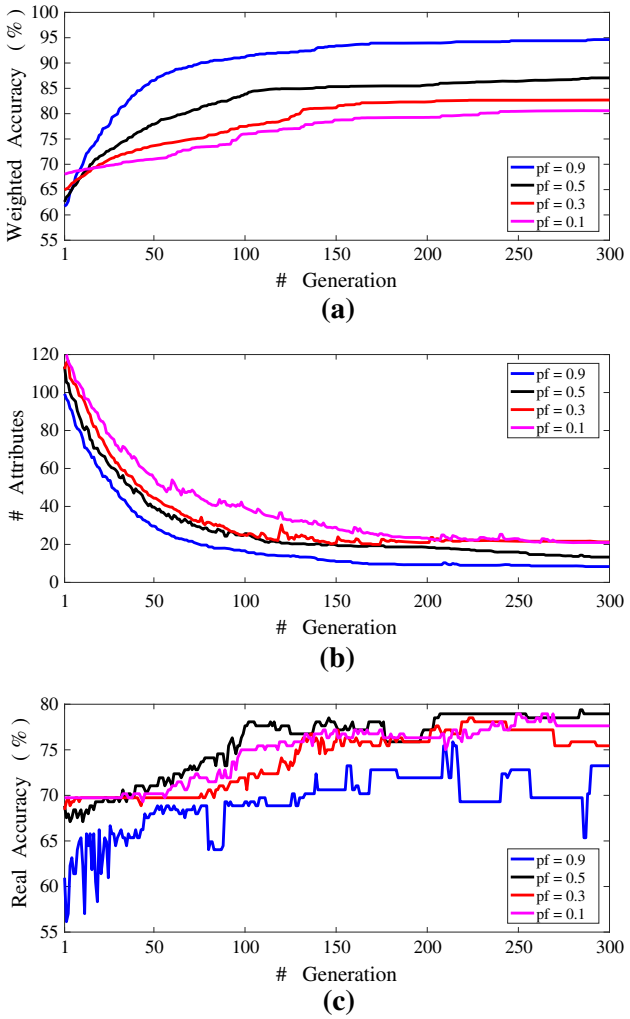

Fig. 4 Results achieved by the best-evaluated individual in each GA generation: **a** weighted accuracy; **b** number of attributes; and **c** real accuracy

Table 3 Best SVM results (in %) on the original database using strategy 'C': minimum, maximum, average and standard deviation

Classifier	Accuracy	Precision	Recall	F1 Score	AUC
Mim	78.9	76.6	85.4	81.4	77.2
Max	80.3	78.3	87.8	82.8	80.3
Avg	79.4	77.5	87.0	82.0	79.2
Std	0.6	0.7	1.1	0.6	1.4

leged only an accuracy increase nor a reduction in selected attributes, but both of them. The achieved minimum (min), maximum (max), average (avg) and standard deviation (std) values for these measures are shown in Table 3.

From the three runs, 37 different attributes were selected 43 times, in order to form the best general individuals. There was a high dispersion concerning the selected attributes. Only four out of 37 attributes were selected to form the best general individual in more than one run. Additionally, considering the original database, only three out of the 260 attributes were never selected to form the best individual in at least one GA generation.

Table 4 Best single classifier results (in %) on the database with only the pre-selected 22-attribute subsets and the strategy 'A'

Classifier	Accuracy	Precision	Recall	F1 Score	AUC	k
SVM	68.4	67.3	80.5	73.3	69.4	–
k-NN	69.7	67.3	85.4	75.3	60.8	7
LDA	61.8	64.3	65.9	65.1	66.8	–

Table 5 Best SVM results (in %) on the database with only the pre-selected 22-attribute subsets and the strategy 'B': minimum, maximum, average and standard deviation

Classifier	Accuracy	Precision	Recall	F1 Score	AUC
Mim	76.3	75.6	82.9	79.1	77.8
Max	78.9	77.8	85.4	81.4	79.8
Avg	77.6	76.9	83.7	80.2	78.7
Std	1.1	1.0	1.1	1.0	0.8

3.2 Pre-selecting the initial configuration for the GA population

Table 4 presents the results achieved by using only the 22 pre-selected attributes and the strategy 'A'. It is worth mentioning that in this case we have ignored all remaining 238 attributes. When analyzing these results, we noticed an improvement in about eight percentual points in accuracy and a decrease in about 91.5% in the number of attributes.

As stated previously in Sect. 2.5, in our next experiments (strategies 'B' and 'C'), all of the 1000 individuals were also represented by 260-column arrays filled with zeros and ones. However, in the first GA generation only the 22 preselected attributes were set with ones and the remaining ones were set with zeros. Consequently, all individuals in the population were equal to each other, but a diversity population was produced by GA operations throughout the GA generations. Thus, even those previously unselected attributes in the first generation could be included later in the following GA generations.

We have achieved the following results, by using strategy 'B': 77.6% ($\sigma = 1.1$) of accuracy; 76.9% ($\sigma = 1.0$) of precision; 83.7% ($\sigma = 1.1$) of recall; 80.2% ($\sigma = 1.0$) of f1 score; and 78.7 ($\sigma = 0.8$) for AUC. They were achieved by individuals with 21.3 ($\sigma = 2.6$) selected attributes in the 288[a] GA generation. The achieved minimum (min), maximum (max), average (avg) and standard deviation (std) values for these measures are shown in Table 5.

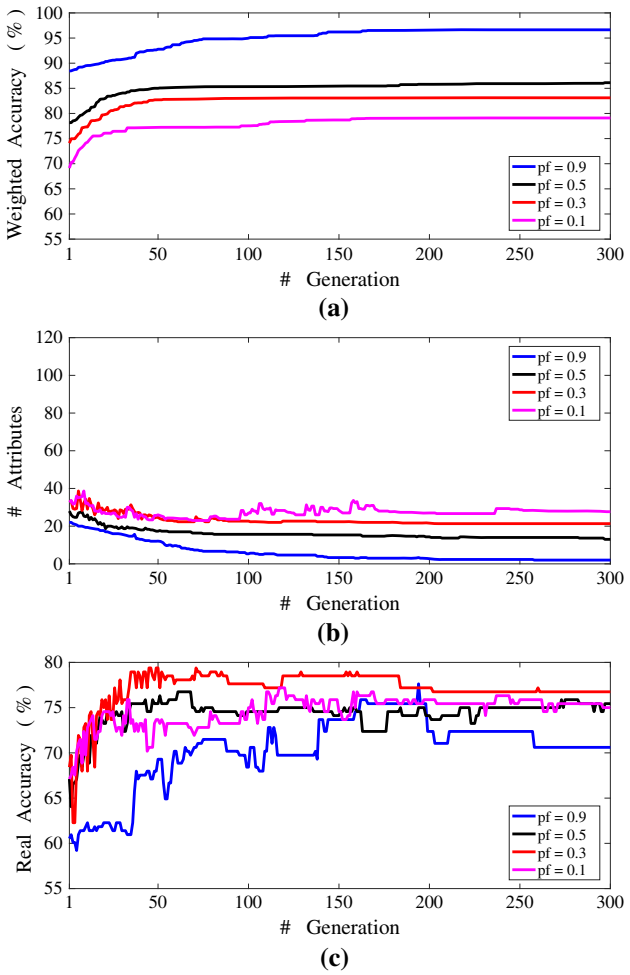

Fig. 5 Accuracy achieved by the best-evaluated individual in each GA generation: **a** weighted accuracy; **b** number of attributes; and **c** real accuracy

Figure 5a presents the evolution of weighted accuracy (Eq. 5) for some tested arrangements for the weights *pa* and *pf*, while using strategy 'C'. Again, legend values represent the *pf* weights used in each set of experiments. The other ones are not presented in the whole report, due to the fact that they reached intermediate results.

Figure 5a, b shows, once more, that higher values for *pf* imposed a reduction in selected attributes. They also boosted the weighted accuracy (Eq. 5) used as an objective function by GA technique to select subsets of attributes through its generations. Again, such a fact can be explained by both Eqs. 5 and 6 .

Similar to Fig. 4, while Fig. 5a shows the expected increase in weighted accuracy, Fig. 5b, c depicts a higher instability. Once more, sometimes weighted accuracy was improved by a new smaller subset of attributes, but at other times it happened with a new bigger subset of attributes that provided an increase in real accuracy (Eq. 1).

Table 6 Best SVM results (in %) on the database with only the pre-selected 22-attribute subsets and the strategy 'C': minimum, maximum, average and standard deviation

Classifier	Accuracy	Precision	Recall	F1 Score	AUC
Mim	78.9	76.6	85.4	81.4	77.1
Max	80.3	78.3	87.8	82.8	80.1
Avg	79.4	77.5	87.0	82.0	78.4
Std	0.6	0.7	1.1	0.6	1.3

The best values for weighted accuracy were achieved by using balanced values for $pa = 0.7$ and $pf = 0.3$. Such weights represent a lower preference for an accuracy increase than a reduction in selected attributes. The best results were achieved by using individuals composed of 24.7 ($\sigma = 6.7$) attributes in the 49^{a} GA generation: 79.4% ($\sigma = 0.6$) of accuracy; 77.5% ($\sigma = 0.7$) of precision; 87.0% ($\sigma = 1.1$) of recall; 82.0% ($\sigma = 0.6$) of f1 score; and 78.4 ($\sigma = 1.3$) for AUC. The achieved minimum (min), maximum (max), average (avg) and standard deviation (std) values for these measures are shown in Table 6.

From the three runs, 45 different attributes were selected 74 times to form the best general individuals. Considering the 22 pre-selected attributes, only 18 were selected 41 (55.4%) times to form the best final individual in at least one run. The remaining four attributes were never selected to form the best general individual. Other 27 new attributes were selected 33 (44.6%) times and formed the best general individual in at least one run.

When analyzing the 45 different selected attributes, a high frequency of those relating to the highest buckets was identified. In addition, although the general selecting rates (55.4% and 44.6%) for both subsets are similar, the frequencies in which the attributes from each one were selected are very different. Table 7 demonstrates that the 18 initially pre-selected attributes were chosen more than once, while the new ones were usually chosen only once, possibly as complementary information.

Based on Table 7 and the 22 pre-selected attributes, we realized that two of them were selected only once, nine were selected twice, and seven were selected in all of the three runs. Considering the 27 complementary attributes, 22 of them were selected only once, four were selected twice, and only one was selected in all of the three runs.

3.3 Multiple classifier systems

When applying the GA technique on the original 260-attribute and 22-pre-selected-attribute databases, we evaluated two alternatives for MCS. Tables 8 and 9 present the best results achieved for the GA technique, and the RFC and GBC algorithms. Both of them were evaluated only for

Table 7 Frequency of attributes selected to form the best general individuals

Attribute subset/ # Selection	22		27	
	#	%	#	%
1	2	4.9	22	66.7
2	18	43.9	8	24.2
3	21	51.2	3	9.1
Totals	41	100.0	33	100.0

Table 8 Best RFC and GBC accuracy while applying the GA selection (strategy 'C') on the original 260-attribute and 22-pre-selected-attribute databases

Attribute Subset	Strategy	Accuracy			Attributes		
		%	σ	AUC	#	σ	# Gen.
22	RFC	71.5	7.6	69.1	25.7	8.7	87
	GBC	80.3	0.0	79.4	13.7	6.0	284
260	RFC	67.1	7.7	68.1	105.0	1.6	7
	GBC	79.4	1.2	79.3	22.0	3.6	293

Table 9 Precision, recall and f1 score concerning the accuracy presented in Table 8

Attribute Subset	Strategy	Precision		Recall		F1 Score	
		%	σ	%	σ	%	σ
22	RFC	72.8	7.5	75.6	6.0	74.2	6.7
	GBC	80.2	2.3	84.6	4.1	82.2	0.7
260	RFC	68.2	7.2	74.0	5.7	70.9	6.4
	GBC	77.5	1.0	87.0	1.1	82.0	1.1

Table 10 Best RFC results (in %) on the original database using strategy 'C': minimum, maximum, average and standard deviation

Classifier	Accuracy	Precision	Recall	F1 Score	AUC
Mim	56.6	58.7	65.9	62.1	59.3
Max	75.0	76.2	78.0	77.1	76.6
Avg	67.1	68.2	74.0	70.9	68.1
Std	7.7	7.2	5.7	6.4	7.1

Table 11 Best RFC results (in %) on the database with only the pre-selected 22-attribute subsets using strategy 'C': minimum, maximum, average and standard deviation

Classifier	Accuracy	Precision	Recall	F1 Score	AUC
Mim	63.2	65.1	68.3	66.7	64.1
Max	81.6	82.9	82.9	82.9	71.9
Avg	71.5	72.8	75.6	74.2	69.1
Std	7.6	7.5	6.0	6.7	3.6

Table 12 Best GBC results (in %) on the original database using strategy 'C': minimum, maximum, average and standard deviation

Classifier	Accuracy	Precision	Recall	F1 Score	AUC
Mim	77.6	76.1	85.4	80.5	78.6
Max	80.3	78.3	87.8	82.8	79.9
Avg	79.4	77.5	87.0	82.0	79.3
Std	1.2	1.0	1.1	1.1	0.5

Table 13 Best GBC results (in %) on the database with only the pre-selected 22-attribute subsets using strategy 'C': minimum, maximum, average and standard deviation

Classifier	Accuracy	Precision	Recall	F1 Score	AUC
Mim	80.3	77.1	80.5	81.5	77.7
Max	80.3	82.5	90.2	83.1	82.4
Avg	80.3	80.2	84.6	82.2	79.4
Std	0.0	2.3	4.1	0.7	2.2

the best values found in strategy 'C'. In general, the results achieved by using the pre-selecting strategy surpassed the original database for GBC and RFC. In addition, the GBC results surpassed those achieved by RFC. The same superiority was shown for the number of selected attributes. The achieved minimum (min), maximum (max), average (avg) and standard deviation (std) values for these measures are shown in Tables 10, 11, 12, 13.

Considering the original database, from the three runs of GBC, 54 different attributes were selected 66 times to form the best general individuals. Only 4 out of the 22 pre-selected attributes were selected 6 (9.1%) times to form the best final individual in at least one run. Other 50 attributes were selected 60 (90.9%) times and formed the best general individual in at least one run. For the same database, from the three RFC runs, 204 different attributes were selected 315 times to form the best general individuals. Sixteen out of the 22 pre-selected attributes were selected 28 (8.9%) times to form the best final individual in at least one run. Other 188 attributes were selected 287 (91.1%) times and formed the best general individual in at least one run.

Considering the 22 pre-selected attribute database, from the three GBC runs, 31 different attributes were selected 41 times to form the best general individuals. Thirteen out of the 22 pre-selected attributes were selected 22 (53.7%) times to form the best final individual in at least one run. Eighteen other attributes were selected 19 (46.3%) times and formed the best general individual in at least one run. For the same database, from the three RFC runs, 43 different attributes were selected 77 times to form the best general individuals. All of the 22 pre-selected attributes were selected 54 (70.1%) times to form the best final individual in at least one run; 21 other attributes were selected 22 (28.6%) times and formed the best general individual in at least one run.

Table 14 Synthesis of the achieved results

Attribute Subset	Strategy	Accuracy %	Accuracy σ	Accuracy AUC	Attributes #	Attributes σ	# Gen.
22	A	68.4	–	69.4	22.0	–	–
	B	77.6	1.1	78.7	21.3	2.6	288
	C	79.4	0.6	78.4	24.7	6.7	49
	RFC	71.5	7.6	69.1	25.7	8.7	87
	GBC	80.3	0.0	79.4	13.7	6.0	284
260	A	60.5	–	67.3	260.0	–	–
	B	61.2	7.8	65.2	124.2	4.7	53
	C	79.4	0.6	79.2	14.3	2.5	284
	RFC	67.1	7.7	68.1	105.0	1.6	7
	GBC	79.4	1.2	79.3	22.0	3.6	293

Table 15 Precision, recall and f1 score concerning the accuracy presented in Table 14

Attribute Subset	Strategy	Precision %	Precision σ	Recall %	Recall σ	F1 Score %	F1 Score σ
22	A	67.3	–	80.5	–	73.3	–
	B	76.9	1.0	83.7	1.1	80.2	1.0
	C	77.5	0.7	87.0	1.1	82.0	0.6
	RFC	72.8	7.5	75.6	6.0	74.2	6.7
	GBC	80.2	2.3	84.6	4.1	82.2	0.7
260	A	62.2	–	68.3	–	65.1	–
	B	62.1	8.4	63.5	8.9	62.7	8.4
	C	77.5	0.7	87.0	1.1	82.0	0.6
	RFC	68.2	7.2	74.0	5.7	70.9	6.4
	GBC	77.5	1.0	87.0	1.1	82.0	1.1

3.4 Discussion

Tables 14 and 15 present a general synthesis of our results, and Fig. 6 compares the ROC curves for the best ones. Our best results were: 80.3% ($\sigma = 0.0$) of accuracy; 80.2% ($\sigma = 2.3$) of precision; 84.6% ($\sigma = 4.1$) of recall; 82.2% ($\sigma = 0.7$) of f1 score; and 79.4 ($\sigma = 2.2$) for AUC. Except for the recall measure, these results were reached by using the 22-pre-selected attributes to set the initial GA population and the GBC algorithm in the 284^{a} GA generation. The final individuals were composed of only 13.7 ($\sigma = 6.0$) selected attributes.

In order to achieve such results, 31 different attributes were used, which were selected 41 times, from which 13 belonged to the 22 pre-selected attributes and were selected 22 (53.7%) times. In general, there is always some advantage to the database, based on the proposed pre-selecting strategy when its results are compared to those achieved by using the original database. Our experiments and their results assure the effectiveness of our proposed strategy of pre-selecting attributes. However, there are certainly other attribute subsets in the original database that could produce similar or better results.

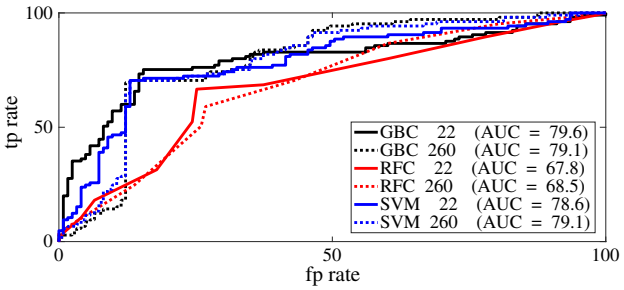

Fig. 6 ROC curves for the best results in Table 14

For strategies 'A' and 'B,' accuracy was about eight and sixteen percentage points higher, respectively. For RFC and GBC, accuracy surpassed by about four and one percentage points, respectively. Only for strategy 'C,' did the accuracy achieve similar results, but we selected fewer attributes when using the original database. On the other hand, even in this case, when using the pre-selected attributes, there was a faster convergence in terms of increasing accuracy and reducing the selected attributes (Fig. 7).

As we have shown, the pre-selected attributes presented a higher frequency in the final GA individual, even when the original 260-attribute database was used. The other attributes usually appeared once and could be considered as complementary information. From Sec. 3.1 and 3.2 , we have 11 attributes selected by using the original database and also the pre-selected 22-attribute database. Four attributes were always selected to form the best general GA individual. Only one attribute was not selected by GA using the SVM classifier on the original database. Also, only one attribute was not selected by GA using RFC on the pre-selected 22-attribute database.

4 Conclusion

In this paper, we introduced a new database composed of 76 NMR spectra of *Araucaria angustifolia* plants, which have shown to be worthy for the challenging problem of identifying their gender. A new strategy based on statistical analysis to select subsets of attributes was also presented. Such a strategy helped to reduce the size of the original set of attributes and improve the performance of the built models in relation to recognition rates and computational costs. The results reported in Tables 1 and 4 indicate that, by using only the 22-pre-selected attributes from the 260 original set, there was an accuracy improvement of 9.2 percentage points and a reduction of 91.5% in the amount of attributes.

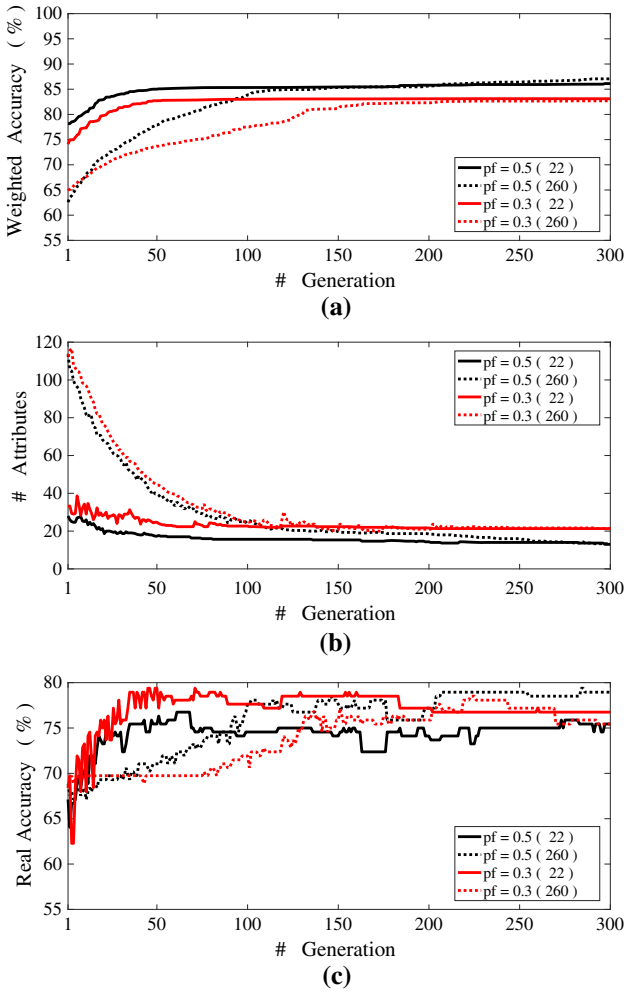

Fig. 7 Accuracy achieved by the best-evaluated individual in each GA generation: **a** weighted accuracy; **b** number of attributes; and **c** real accuracy

While looking for improving our results, we used the GA technique to select new subsets of attributes. Here, we carried experiments in which the initialization step took into account all the original 260 attributes and also only the 22-pre-selected attributes. Although similar results were achieved, using a smaller initial set of attributes produced a faster convergence of the GA technique.

Our best results were: 80.3% ($\sigma = 0.0$) of accuracy; 80.2% ($\sigma = 2.3$) of precision; 84.6% ($\sigma = 4.1$) of recall; 82.2% ($\sigma = 0.7$) of f1 score; and 79.4 ($\sigma = 2.2$) for AUC. Except for the recall measure, these results were reached by using the 22-pre-selected attributes to set the initial GA population and the GBC algorithm in the 284^{a} GA generation. The final individuals were composed of only 13.7 ($\sigma = 6.0$) selected attributes, which represent the use of only 5.3% from the 260 original set but an accuracy improvement by about 20.0 percentage points.

The results for the 2-class problem presented in this work are very promising, since they can be used as an alternative to plan a spatial distribution of plants in fields and forests. In future work, we plan to test other machine learning algorithms to solve any remaining confusion, as well as apply techniques to select and combine their results. Our expectation is that this database will contribute to the field of forest management and motivate more researchers to work in this field. Additionally, the proposed selecting strategy would be useful in order to reduce the complexity of any databases and accelerate the process of building classification models. In this context, one more future work arises, i.e., evaluating the GA behavior by considering only pre-selected attributes and not allowing the remaining ones to be added throughout the GA generations.

References

Amaral ACB, Ferreira M, Bandel G (1971) Variação da densidade básica da madeira produzida pela Araucaria angustifolia (bert.) o. KTZE no sentido medula-casca em árvores do sexo masculino e feminino. Inst Pesqui Estud Florestais 2–3:119–127

Arqub OA, Abo-Hammour Z (2014) Numerical solution of systems of second-order boundary value problems using continuous genetic algorithm. Inf Sci 279:396–415

Atanazio KA, Hess AF, Krefta SM, Schorr LPB, da Rosa GT, Filho MDHV, da Silva GO, Abatti R, Simas M, Galvani LV (2018) Proporção da dioicia de Araucaria angustifolia em um povoamento localizado em lages, sc. In: Buzatto CR, Prestes NP, Martinez J, Nienow AA (eds) Procs of Seminário Sul-Brasileiro sobre a Sustentabilidade da Araucária, vol 3. Lew, Tapera, RS, pp 243–246

Bandel G, Gurgel JTA (1967) A proporção do sexo em pinheiro-brasileiro *Araucaria angustifolia* (bert) o. KTZE. Silvicultura: Revista Técnica do Serviço Florestal do Estado de São Paulo 6:209–220

Blum C (2005) Ant colony optimization: introduction and recent trends. Phys Life Rev 2(4):353–373

Butterworth I, Ellis J, Gabathuler E, Sloan T (2001) The spin structure of the nucleon. Philos Trans R Soc Lond Ser A Math Phys Eng Sci 359(1779):379–389

Carvalho BG (2012) Diferenciação sexual de Araucaria Angustifolia por RMN HR-MAS e análise multivariada. PhD thesis, Universidade Federal de Goiás, Goiânia, GO

Carvalho PER (2003) Espécies florestais brasileiras. Embrapa Informação Tecnológica, Brasília

Cavanagh J, Fairbrother WJ, Palmer AG, Rance M, Skelton NJ (1995) Protein NMR spectroscopy: principles and practice. Elsevier Science, Amsterdam

Constantino V, Zanette F (2016) Produção de borbulhas ortotrópicas para enxertia de *Araucaria angustifolia*. Acta Biol Parana 44(1–4)

Dorigo M (1992) Optimization, learning and natural algorithms. PhD thesis, Politecnico di Milano, Milano, Italy

Dorigo M, Birattari M, Stutzle T (2006) Ant colony optimization. IEEE Comput Intell Mag 1(4):28–39

Ernst RR, Anderson WA (1966) Application of Fourier transform spectroscopy to magnetic resonance. Rev Sci Instrum 37(1):93–102

Euceda LR, Giskeødegård GF, Bathen TF (2015) Preprocessing of NMR metabolomics data. Scand J Clin Lab Investig 75(3):193–203

Fawcett T (2006) An introduction to roc analysis. Pattern Recognit Lett 27:861–874

Freitas AM, Almeida MTR, Andrighetti-Fröhner CR, Cardozo FTGS, Barardi CRM, Farias MR, Simões CMO (2009) Antiviral activity-guided fractionation from *Araucaria angustifolia* leaves extract. J Ethnopharmacol 126(3):512–517

Geem ZW, Kim JH, Loganathan GV (2001) A new heuristic optimization algorithm: harmony search. Simulation 76(2):60–68

Goldberg D, David Edward G, Goldberg D, Goldberg V (1989) Genetic algorithms in search, optimization, and machine learning. Artificial intelligence. Addison-Wesley Publishing Company, Boston

Goldberg DE (1989a) Genetic algorithms in search. Optimization and machine learning. Addison-Wesley, New York, NY

Guerra MP, Silveira V, dos Reis MS, Schneider L (2002) Exploração, manejo e conservação da Araucária (Araucaria angustifolia). Editora SENAC São Paulo, São Paulo, SP, pp 85–101

Holland JH (1975) Adaptation in natural and artificial systems, 2nd edn. University of Michigan Press, Ann Arbor, MI, p 1992

Houle ME, Kriegel HP, Kröger P, Schubert E, Zimek A (2010) Can shared-neighbor distances defeat the curse of dimensionality? In: Gertz M, Ludäscher B (eds) Scientific and statistical database management. Springer, Heidelberg, pp 482–500

Jeener J, Broekaert P (1967) Nuclear magnetic resonance in solids: thermodynamic effects of a pair of RF pulses. Phys Rev 157:232–240

Kennedy J, Eberhart R (1995) Particle swarm optimization. In: Proceedings of ICNN'95—international conference on neural networks, vol 4, pp 1942–1948

Keun HC, Ebbels TMD, Antti H, Bollard ME, Beckonert O, Holmes E, Lindon JC, Nicholson JK (2003) Improved analysis of multivariate data by variable stability scaling: application to NMR-based metabolic profiling. Anal Chim Acta 490(1):265–276

Kirkpatrick S, Gelatt CD, Vecchi MP (1983) Optimization by simulated annealing. Science 220(4598):671–680

Kuo FY, Sloan IH (2005) Lifting the curse of dimensionality. Not AMS 52:1320–1329

Murakami MH (2002) Identificação de marcador molecular associado à expressão sexual em Araucaria Angustifolia (BERT) o. KTZE. PhD thesis, Universidade Federal do Paraná, Curitiba, PR

Murakami MH (2003) Effects of forest management on the genetic diversity in a population of Araucaria angustifolia (BERT.) o. KTZE. Silvae Genet 52:5–6

Nicholson JK, Lindon JC (2008) Systems biology: metabonomics. Nature 455(7216):1054–1056

Oliveira GAR (2016) Avaliação da composição química de indivíduos adultos de *Araucaria Angustifolia* (BERT.) Kutz. através da RMN aliada a quimiometria visando a distinção sexual. PhD thesis, Universidade Federal de Goiás, Goiânia, GO

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E (2011) Scikit-learn: machine learning in Python. J Mach Learn Res 12:2825–2830

Rinnan A, van den Berg F, Engelsen SB (2009) Review of the most common pre-processing techniques for near-infrared spectra. TrAC Trends Anal Chem 28(10):1201–1222

Rith K, Schafer A (1999) The mystery of nucleon spin. Sci Am 281(1):58–63

Sahab MG, Toropov VV, Gandomi AH (2013) A review on traditional and modern structural optimization: problems and techniques. In: Gandomi AH, Yang XS, Talatahari S, Alavi AH (eds) Metaheuristic Appl Struct Infrastruct. Elsevier, Oxford, UK, pp 25–47

Sastry K, Goldberg D, Kendall G (2005) Genetic algorithms. Springer, US, Boston, MA, pp 97–125

Sousa SAA, Magalhães A, Ferreira MMC (2013) Optimized bucketing for NMR spectra: three case studies. Chemom Intell Lab Syst 122:93–102

Stefenon VM, Gailing O, Finkeldey R (2008) Genetic structure of plantations and the conservation of genetic resources of Brazilian pine (*Araucaria angustifolia*). For Ecol Manag 255(7):2718–2725

Suarez C, Kohler SJ, Allen MM, Kolodny NH (1999) NMR study of the metabolic 15n isotopic enrichment of cyanophycin synthesized by the cyanobacterium synechocystis sp strain PCC 6308. Biochim Biophys Acta (BBA) Gen Subj 1426(3):429–438

Teodorović D, Dell'Orco M (2005) Bee colony optimization: a cooperative learning approach to complex transportation problems. In: Advanced OR and AI methods in transportation. Proceedings of the 10th meeting of the EURO working group on transportation, Poznan, Poland, pp 51–60

Vu TN, Laukens K (2013) Getting your peaks in line: A review of alignment methods for NMR spectral data. Metabolites 3:259–276

Wendling I (2011) Enxertia e florescimento precoce em *Araucaria angustifolia*. Comunicado Técnico - Embrapa 1(272):1–7

Worzel WP, Almal A, MacLean CD (2007) Lifting the curse of dimensionality. Springer, US, Boston, MA

Yang X, Deb S (2009) Cuckoo search via lévy flights. In: 2009 World congress on nature biologically inspired computing (NaBIC). Coimbatore, India, pp 210–214

Yang XS (2010) Nature-inspired metaheuristic algorithms, 2nd edn. Luniver Press, Cambridge

Yang XS, He S (2013) Bat algorithm: literature review and applications. Int J Bio-Inspired Comput 5(3):141–149

Zanette F (2014) Enxertia de araucária para produção de pinhão

Zanette F, Oliveira LS, Biasi LA (2011) Grafting of *Araucaria angustifolia* (BERTOL.) KUNTZE through the four seasons of the year. Rev Bras Frutic 33(4):1364–1370

Zanette F, Danner MA, Constantino V, Wendling I (2017) Particularidades e biologia reprodutiva de *Araucaria angustifolia*. In: Wendling I, Zanette F (eds) Araucária: particularidades, propagação e manejo de plantios. Embrapa, Brasília, DF, pp 15–39

Zanon MLB, Finger CAG, Schneider PR (2009) Proporção da dióicia e distribuição diamétrica de árvores masculinas e femininas de Araucaria angustifolia (BERT.) KUNTZE. em povoamentos implantados. Ciência Florest 9(4):425–431