FORTHCOMING PAPER ·#68T05-21-02-01

DIVERSITY-BASED SELECTION OF LEARNING ALGORITHMS: A BAGGING APROACH

Leidys Cabrera-Hernández[1] *, Alejandro Morales Hernández*, Maricel Meneses Gómez **, Alfredo Meneses Marcel **,Gladys M. Casas Cardoso*, María M. García Lorenzo *
*Departamento de Computación, Facultad Matemática, Física y Computación, Universidad Central "Marta Abreu" de Las Villas, Cuba.
** Centro de Bioactivos Químicos, Universidad Central "Marta Abreu" de Las Villas, Cuba.

ABSTRACT
Nowadays, classification problems are becoming increasingly important in many real-world applications. As the problems become more complex and the consequences of a bad decision are more serious, more advanced techniques, as the combination of classifiers, need to be applied. When combining classifiers, it is important to ensure diversity between them as it does not make sense to combine classifiers whose classification is the same. There are several techniques to ensure diversity in systems like these and generally it consider modify the data set, use different learning algorithms or make a process of improvement or learning on the individual classification. Although the relationship between diversity and system accuracy has not been fully established, it is clear that diversity remains a factor to be taken into account in the construction of multiclassifiers. In this paper we present a modification to the bagging algorithm to consider different learning algorithms during the training process and optimize the classifiers built to obtain diverse systems and as accurate as possible. Executed simulations suggest the use of the Double Failure pairwise measure to quantify the diversity of the system. With respect to the number of classifiers used, it was observed that the systems built had approximately half of the total classifiers they should have. After, the superiority of the proposed method with respect to five state-of-the-art multiclassifiers was verified and it is suggested the incorporation of a learning process like the one executed in Stacking. Finally, are shown results in biochemical real applications and the general conclusions are exposed.

KEYWORDS: Diversity measures, classifiers combination, Bagging, supervised learning.

MSC: 68T05

RESUMEN
En la actualidad, los problemas de clasificación cada día cobran mayor importancia en muchas aplicaciones reales. A medida que los problemas se hacen más complejos y las consecuencias de una mala decisión son más graves se necesita aplicar técnicas más avanzadas como la combinación de clasificadores. Cuando se combinan clasificadores es importante garantizar la diversidad entre ellos ya que no tiene sentido combinar clasificadores cuya clasificación sea la misma. Existen varias técnicas para garantizar la diversidad en sistemas de este tipo y de forma general consideran modificar el conjunto de datos, utilizar diferentes algoritmos de aprendizaje o efectuar un proceso de mejora o aprendizaje sobre la clasificación individual. Aunque la relación entre la diversidad y la exactitud del sistema no ha sido establecida del todo, sí queda claro que la diversidad sigue siendo un factor a tener en cuenta en la construcción de los multiclasificadores. En este trabajo se presenta una modificación al algoritmo de Bagging para considerar diferentes algoritmos de aprendizaje durante el proceso de entrenamiento y optimizar los clasificadores construidos para obtener sistemas diversos y lo más exacto posibles. Las simulaciones ejecutadas sugieren la utilización de la medida pairwise de Doble Fallo para cuantificar la diversidad del sistema. Con respecto a la cantidad de clasificadores usados, se observó que los sistemas construidos tenían aproximadamente la mitad del total de clasificadores que debían tener. Después, se comprobó la superioridad del método propuesto con respecto a cinco multiclasificadores reportados en la literatura y se sugiere la incorporación de un proceso de aprendizaje como el ejecutado en Stacking. Finalmente, se muestran los resultados en aplicaciones reales de bioquímicas y las conclusiones generales son expuestas.

PALABRAS CLAVES: Medidas de diversidad, combinación de clasificadores, Bagging, aprendizaje supervisado.

1. INTRODUCTION

A classifier has as main objective to assign the category or class to an example that represents an instance of a certain problem. Hence, one of the distinctive stages in them is the training of a learning algorithm to build a model, mathematical or not, that allows classification. Although the use of classifiers has proven to be an effective tool in different classification processes, many researchers suggest combining them instead of using

[1] leidysc@uclv.edu.cu, mmgarcia@uclv.edu.cu

just one [35]. There are several works that use the combination of some classifiers with success to solve many real-life problems [17, 32-34, 43, 48].
In these combinations, a key factor is the use of classifiers whose performance is different. In fact, combining similar classifiers would not improve individual classification. There are a large number of learning algorithms and the configuration possibilities of their parameters make it difficult to select a set of classifiers that guarantee a better classification than that which can be obtained using a single classifier. The classic methods to build this type of system are Bagging [41], Boosting [44] and Stacking [49]. In general, the individual classifiers used to build the system should complement each other since, if one classifier fails, the others can prevent the system from failing as well [18].
In relation to diversity in multiclassifier systems, not always greater diversity is related to better accuracy in the system [8]. In fact, in a multiclassifier system, diversity is not always associated with a higher classification than the individual classification [37, 45]. However, diversity remains an important element to take into account due to its own definition: two classifiers are diverse if they make mistakes different and as was mentioned before not have sense combine classifiers with the same errors, because then the results of the combination not will be better.
Although there are many measures of diversity reported in the literature to quantify the diversity [25, 35, 44], the presence of this in the construction of the system is normally assumed and not used directly the measures. These are the cases of the classical approaches in which diversity is generated by modifying the training set, using different learning algorithms or introducing meta-learners to learn from the outputs given by the individual classifiers. However, some methods have been developed that explicitly use the measures to quantify the diversity between the classifiers and build better systems [10, 12]. Although the results obtained in these investigations show the possibility of finding better combinations of classifiers that ensure an accuracy superior to the best individual accuracy, they have not been extended to the traditional construction approaches mentioned above. Also, there are other works [1] that use diversity and accuracy as important aspects but they neither have been extended to the traditional construction approaches mentioned above.
In this work, a new method of construction of multi-classifiers based on the resampling performed by Bagging is presented, incorporating the possibility of using different learning algorithms and selecting the best combination of classifiers according to the diversity between the combined classifiers and the accuracy of the system formed. More explicitly, this work has two main contributions. First, the scheme for combining classifiers generated from different training samples and whose combination is optimized by means of a Genetic Algorithm is proposed in order to obtain the most exact and diverse combination possible. Finally, an experimental study is carried out using 20 data sets recognized in the literature to validate the proposed method.
The rest of the article is organized as follows: Section 2 and 3 present the main elements associated with multiclassifier systems and the diversity measures reported in the literature. In section 4, reference is made to the method of construction of multi-classifiers proposed in this work. The empirical study carried out on the proposed method applied to 20 data sets is presented in Section 5. In Section 6 is presented results in biochemical real applications. Finally, the main results found are summarized in the conclusions.

2. MULTICLASSIFIER SYSTEMS

A multiclassifier system is built in several ways but generally it requires that the classification models be exact and that there be diversity among them [50]. Diversity can be guaranteed in different ways and is generally considered implicitly in the construction of the system. Methods that modify the training set do so by taking different subsets of training examples or by selecting different subsets of features. Bagging [41] is the classic example of classifiers of this type. Its operation is simple: the construction of the system is done from classifiers with the same learning algorithm, but trained in different samples taken from the training set. The classification given by the system will be the one with the most votes within the total set of classifiers formed. Boosting [44] can be considered in this group too, if it is taken into account that in each iteration the set of data is “modified” by weighing the incorrectly classified examples in the previous iteration to try to classify them correctly in the next one.
On the other hand, in the group of methods that use different learning algorithms to build the system, the training set is not modified. The simplest is Vote [35], that starts from the individual training of different learning algorithms and uses different ways to combine the outputs; either by majority vote, average of the class assignment probabilities, among others. The variants in this type of method are generally given by the modification of the vote to carry out a heavy vote [35, 39]. Another method is Stacking [49] but instead of using a vote to obtain the result of the system, it uses another learning algorithm to learn from the individual

outputs.
Hybrid approaches are more common because they try to take advantage of the previous methods. In [4] a modification is made to the majority vote implemented in Vote so that each learning algorithm is trained in a subset of traits of the training set. In turn, in [40] three levels of diversity generation are used to analyze their influence on the construction of the multiclassifier; doing feature selection, resampling the training set and using different learning algorithms.
Another very common technique is the use of metaheuristics for the optimization of the multiclassifier. For example, in [40] a Genetic Algorithm is used to carry out the selection of traits, while in [31, 36] other variants are explored to also select classifiers and applied to digit recognition.
On the other hand, the works presented in [10-12] make use of the diversity measures to guide the search in metaheuristics that select a combination of classifiers whose performance in the system is superior to that obtained individually.

3. DIVERSITY MEASURES

As mentioned before, it does not make sense that multiclassifiers combine identical classifiers between them because a good performance would not be attained, so it is important to know how diverse a classifier ensemble is.
According to Kuncheva and Whitaker [37], there is no measure of diversity explicitly involved in classic methods of multiclassifiers construction, although diversity is the key point in any of the previously discussed methods. A number of diversity measures have been proposed in the literature. They are divided into two categories: pairwise and non-pairwise.
The first set of measures is calculated for pairs of classifiers. Its outputs are binary (0, 1) indicating whether the instance was correctly classified or not. Table 1 shows the results of two classifiers (*Ci, Cj*) for one given instance, depending on whether or not it was correctly classified. If we consider all N instances between the pair of classifiers (*Ci, Cj*), the results summarized in the Table 2 are obtained. In this case, A is equal to the amount of instances where both classifiers have correct classification. D is equal to the amount of instances where both classifiers have incorrect classification. B and C are equals to the amount of instances where one classifier is correct and the other is incorrect. It should be observed that a set of L classifiers have associated L (L-1)/2 pairs of values, so to obtain a single result these values must be averaged. N is the total number of case. Some of more common pairwise measures are shown below.

	C_j correct (1)	C_j incorrect (0)
C_i correct (1)	a	b
C_i incorrect (0)	c	d
a + b + c + d = 1		

Table 1: Binary matrix for one instance.

	C_j correct (1)	C_j incorrect (0)
C_i correct (1)	A	B
C_i incorrect (0)	C	D
A +B + C + D = N		

Table 2: Binary matrix for N instances.

Correlation coefficient ρ
The coefficient of correlation [35] is one of the measures for pairs of classifiers. A better diversity is obtained for smaller values of ρ. The value of ρ will be in the interval [-1,1]. The Correlation coefficient is Pearson`s for the particular case of (0,1) variables (known as ϕ-coefficient). The coefficient is calculated as:

$$\rho_{c_i,c_j} = \frac{A \times D - B \times C}{\sqrt{(A+B) \times (C+D) \times (A+C) \times (B+D)}} \quad (1)$$

The Measure of Differences
The measure of differences was introduced by Skalak [46], it is the most intuitive measure between a pair of classifiers, and it is equal to the probability that the two classifiers disagree in their predictions. The value of D will be in the interval [0,1]. The diversity increases when the value of D increases.

$$D_{c_i,c_j} = \frac{\mathrm{B+C}}{\mathrm{N}} \quad (2)$$

The Double Fault Measure

Another measure to be analyzed is known as double fault measure, which was introduced by Giacinto and Roli [25]and considers the failure of two classifiers simultaneously. This measure is based on the concept that it is more important to know when simultaneous errors are committed, that when both have a correct classification. The value of DF will be in the interval [0,1]. The diversity increases when the value of DF decreases.

$$DF_{c_i,c_j} = \frac{\mathrm{D}}{\mathrm{N}} \quad (3)$$

On the other hand, the non-pairwise measures take into account the outputs of all classifiers at the same time and calculate a unique value of diversity for the whole ensemble, some of them are shown below.Entropy

This measure was introduced by Cunningham and Carney [15], where Y_{ji} will be 1 if the classifier i was correct in the case j and 0 otherwise. The value of E will be in the interval [0,1]. If E is equal to zero, then there is not a difference between the classifiers and if E is equal to 1 then there is the most diversity.

$$E = \frac{1}{N} \times \frac{2}{L-1} \times \sum_{j=1}^{N} min\left\{\left(\sum_{i=1}^{L} Yji\right), \left(L - \sum_{i=1}^{L} Yji\right)\right\}, \quad Yji \in \{0,1\} \quad (4)$$

Measurement of Interrater Agreement

The Measurement of Interrater Agreement was presented in [22]. This measure is known too as kappa and their origin is in [21]. The k is calculated by equation 5, this equation is formed by the subtraction between the unit and the measure of Kendall concordance. In this last term p is the mean of the accuracy in the individual classification, this term is calculated by Equation 6. The value of K will be in the interval [-1,1]. In this measure the diversity is lower when the k value is higher.

$$K = 1 - \frac{\frac{1}{L} \times \sum_{j=1}^{N} Y(Zj) \times \left(L - Y(Zj)\right)}{N \times (L-1) \times p \times (1-p)} \quad (5)$$

$$p = \frac{1}{N \times L} \times \sum_{j=1}^{N} \sum_{i=1}^{L} Yji \quad (6)$$

Difficult Measure

The difficulty Measure, comes from the study carried out for Hansen and Salamon [29]. It is calculated through the variance of a discrete random variable X that takes values in the set (0/L, 1/L, 2/L, …, 1) and it denotes the probability that exactly i classifiers has classified well all the instances.

The intuition of this measure can be explained in the following way: a diverse classifier set has a small value of difficulty measure, since each training sample at least can be classified correctly by a proportion of all the classifiers base, that which is more probable with a low variance of X. In equation 7 the symbol μ represent the mean of the discrete random variable X and x represent each possible value of X, $f(x)$ represent the own probability of each value. For convenience, this measure is usually climbed lineally in the interval [0,1].

$$DIF = \Theta = Var\,(X) = \left[\sum\left(x^2 \times f(x)\right)\right] - \mu^2 \quad (7)$$

As a general rule, we may notice that non-pairwise measures are more computationally complex than pairwise measures; the latter are simpler and the results lend themselves to an easier interpretation given their mathematic formulation. After the implementation of each one of all these measures, was implemented a standardization method over them used in [9, 10], to facilitate the analysis of the results of the simulations carried out in section 5, because as we observe they have different intervals in their values.

4. BAGGING MODIFICATION

The method proposed in this paper (*mulGA*) starts from the idea of Bagging by obtaining different subsets to train. The main difference with the classic method of Bagging is that a single learning algorithm is not used and that not all the trained classifiers are used to build the multi-classifier. The combination of the individual outputs of the classifiers is done using a majority vote.

The proposed method can be divided as follows:

1. Obtain the set of samples to train.
2. Carry out the training of a learning algorithm, taken from a previously defined set **P**, with one of the samples previously formed and form the set **Ω**.
3. Select the best set **ω**, **ω⊂Ω** to form the multiclassifier, using a Genetic Algorithm.

The general operation of the proposed method can be seen in Figure 1.

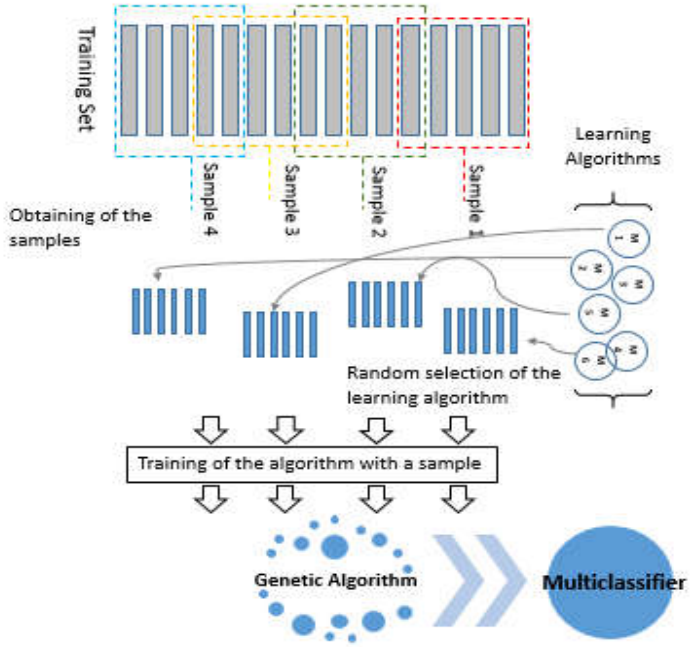

Figure 1: Construction of the multiclassifier from the training data set.

The following sections describe the main parts of the method.

4.1 . Resampling method and obtaining the trained classifiers

The way in which Bagging guarantees diversity in results is by dividing the training set into several replicas and using a single learning algorithm to train on them. Taking advantage of the variations present in the training sets requires that the classifier be unstable; that is, small modifications in the training set should lead to large changes in the classifier output. Otherwise, the built system would be a collection of almost identical classifiers and it would be almost unlikely to improve the performance that only one of them can have [35]. In *mulGA*, the resampling with replacement carried out by Bagging is maintained and unlike this it incorporates more than one learning algorithm to construct the individual classifiers. The Algorithm 1 describes this process.

Algorithm 1: Obtaining the set of classifiers **Ω**

Input: number of samples to form T, set P of learning algorithms to train, set S of examples
Output: set Ω of already trained classifiers

```
1  Begin
2  | Ω ← { }
3  | i ← 1
4  | Repeat
5  |   | s ← resampling with replacement(S)
6  |   | p ← select randomly(P)
7  |   | to train(p, s)
8  |   | Ω ← Ω ∪ p
9  |   | i ← i + 1
10 | until i == T
11 End
```

Figure 2: Algorithm 2: obtaining the set of classifiers.

4.2 . Learning algorithms

Among the most widely used learning algorithms are case-based algorithms, decision trees, Bayesian networks, discriminant analysis, logistic regression, and artificial neural networks.
The closest neighbor rule is to assign a given example the associated class of the closest prototype. An extension to this rule is the k closest neighbors rule [14]. In this case, an example x is classified according to the most frequent class in the k closest samples.
In a classification tree [5], the root node is located above the structure and is connected to the descendant nodes by links or branches. Leaf nodes can be considered as the class to assign. When assigning a class to a new example, the analysis begins starting from the root to one of the leaves of the tree.
A Bayesian network is a probabilistic model that graphically represents a set of variables and the relationships between them [7]. Naive Bayes [38] is the simplest Bayesian classifier that can be formed. Although this cannot always be assured in real problems, its performance can remain acceptable in many situations.
Discriminant analysis is a mathematical technique that helps to identify the characteristics that discriminate two or more groups and to create a function capable of distinguishing with the greatest possible precision the members of one or the other group. For a two-class problem, linear discriminant analysis [20] constructs a hyperplane that separates the data set according to each class.
The logistic regression is defined as equation 8, where *x* represents an example to classify and $\beta_0, \beta_1, \cdots, \beta_n$ are the parameters of the model. These parameters must be estimated from the data to obtain a model fitted to them. The purpose of discriminant analysis is to predict the probability of a certain event occurring, based on the characteristics presented by the learning examples analyzed during training.

$$p(C = 1|x) = 1/\left[1 + e^{-(\beta_0 + \sum_{i=1}^{n} \beta_i x_i)}\right] \quad (8)$$

Artificial neural networks (ANNs) arise inspired by the modeling of the human brain through the mathematical expression of human intellectual abilities [27]. Of all the existing RNA models, the most popular has been multilayer perceptron (MLP). Among the parameters that are specified in an MLP are the learning ratio and the influence of the old weights on the new ones (momentum). Although these are not the only ones, they can be used to generate a set of RNAs with different performances, varying their values and establishing different combinations, even if the topology of the network remains identical.

4.3 . Genetic algorithms

Once the set of trained classifiers has been obtained, the Genetic Algorithm (GA) metaheuristics [26] is used to obtain a combination of diverse classifiers that guarantee the highest possible accuracy. GA modeling is done using the binary representation of chromosomes. In this way, each chromosome represents a possible combination of classifiers in the multiclassifier, according to equation 9.

$$C_x = (g_1, g_2, \cdots, g_i, \cdots, g_T), \quad g_i = \begin{cases} 0 & ,if\ the\ classifier\ i\ is\ not\ present \\ 1 & ,if\ the\ classifier\ i\ is\ present \end{cases} \quad (9)$$

The quality function of each chromosome contains two variables; the diversity determined between classifiers of the combination and the accuracy of the multiclassifier formed. In general, it can be seen as:

$$f(C_x) = E(C_x) * D(C_x) \quad (10)$$

Where, $E(C_x)$ is the accuracy of the multiclassifier and $D(C_x)$ is the value resulting from measuring the diversity in the built system. Taking into account this quality function, it is defined as an objective function $\max_{0 \le x \le P} f(C_x))$, where P is the size of the population.
The crossover operator used is based on a crossover point [30] and the implementation of the mutation operator is done by exchanging the information of the gene to be mutated. In relation to the size of the populations, in the present work a size equal to 50 is used, because, in the combination of classifiers, the process of obtaining the quality function in each chromosome can be computationally expensive. The initial population is generated randomly so that the genes of each chromosome take on a value of 0 or 1 depending on a random value $r \in [0, 1]$, if *r* is greater than 0,5 the classifier represented by the i^{th} gene is included; otherwise, it is not included. Furthermore, the best individual classifiers are included in the combination according to the results obtained in [10, 11].
Each generation of GA begins with a population with the number of chromosomes specified above. With the roulette selection method [30], half of the chromosomes are taken to form an intermediate population. On these, the genetic recombination operators are applied and their result is added to the intermediate population

until it reaches the established size. The AG runs until the generations for which it is sent to evolve are reached.

5. SIMULATIONS

In this section, the proposed method is validated in a real scenario where the objective is to build multiclassifiers with the highest possible accuracy and using the diversity between the combined classifiers.

5.1 . Data sets, learning algorithms and evaluation measure

In order to validate the proposed method is used 20 data sets available in the automatic learning repository of the University of California Irvine UCIML [2] and in the PMB (Penn Machine Learning Benchmarks) [42] developed by the Laboratory of Computational Genetics of the University of Pennsylvania. Table 3 summarizes the main characteristics of these sets.

ID	Data Sets	Examples	Features
1	acute-inflammation	120	6
2	appendicitis	120	6
3	australian	690	14
4	blood-transfusion	748	4
5	bupa	345	6
6	Clean1*	476	24
7	echocardiogram	131	11
8	fertility diagnosis	100	9
9	German Credit	1000	20
10	HCC	165	49
11	Heart-statlog	270	13
12	ionosphere	351	34
13	liver-disorder	345	6
14	new-tyroid	215	5
15	parkinsons	195	22
16	Pima diabetes	768	8
17	promoters	100	57
18	saheart	462	9
19	sonar	208	60
20	vote	435	16

Table 3: Data sets used in simulations.

* The original data set underwent a principal component analysis to reduce the number of features.

In relation to learning algorithms, 11 algorithms recognized in the literature were used, which are implemented in the WEKA tool. They are:

1. *MultilayerPerceptron (MLP)*
2. *Logistic (L)*
3. *IBk*
4. *J48*
5. *DecisionStump (DS)*
6. *REPTree(RepT)*
7. *NaiveBayes (NB)*
8. *ZeroR (ZR)*
9. *RandomTree (RanT)*
10. *SimpleLogistic (SL)*
11. *SMO*

Were used five RNAs; one with the WEKA default parameters and four more with random values for momentum and learning rate. In the case of IBk, the values of k = 1; k = 3; k = 5 and k = 7. The rest of the algorithms kept the default parameters.

The analysis of the results is made from the average of the results obtained in 10 executions of the method and the accuracy of the system is used in comparison with the performance of other methods reported in the literature. The execution of all the algorithms used for the validation of the method and that are executed

externally to the method, is done using a 10-fold cross-validation process. Furthermore, all random numbers were generated with the same seed to ensure subsequent replication of the experiments.
The parameters used by default in the proposed method consider 30 samples to be formed from the training set with 85% of the original size. Regarding the genetic algorithm to optimize the combination of classifiers, the configuration used in [10] was maintained with 50 generations to evolve, a mutation probability equal to 0,25 and a crossing probability equal to 0,75.

5.2 . Diversity

The Table 4 shows the diversity obtained according to the measure considered in the quality function of the chromosomes, in each of the data sets studied.

ID	ρ	D	DF	E	k	DIF
1	0,499	0,194	**1,000**	0,180	0,517	0,999
2	0,315	0,171	0,944	0,202	0,325	**0,948**
3	0,361	0,249	0,948	0,376	0,411	**0,963**
4	0,245	0,167	0,852	0,228	0,235	**0,902**
5	0,450	0,386	0,902	0,593	0,458	**0,966**
6	0,463	0,319	0,968	0,466	0,486	**0,983**
7	0,389	0,299	0,903	0,440	0,407	**0,958**
8	0,309	0,125	**0,950**	0,175	0,281	**0,950**
9	0,388	0,232	0,924	0,345	0,385	**0,954**
10	0,453	0,334	0,947	0,497	0,465	**0,977**
11	0,396	0,258	0,949	0,341	0,401	**0,964**
12	0,390	0,237	0,965	0,305	0,439	**0,977**
13	0,446	0,391	0,900	0,606	0,454	**0,965**
14	0,422	0,076	**0,996**	0,116	0,433	0,995
15	0,425	0,156	**0,987**	0,256	0,434	**0,987**
16	0,370	0,256	0,901	0,390	0,354	**0,945**
17	0,475	0,322	**0,980**	0,478	0,484	0,990
18	0,398	0,290	0,891	0,469	0,400	**0,943**
19	0,470	0,355	0,956	0,519	0,483	**0,985**
20	0,374	0,125	**0,985**	0,158	0,424	0,988
AVG	**0,402**	**0,247**	**0,942**	**0,357**	**0,414**	**0,967**

Table 4: Diversity obtained according to the measure used in the quality function.

The results observed in the calculated measures of diversity indicate that the greatest diversity is obtained with the DIF measure, followed by the DF measure. Taking into account that DIF considers the number of classifiers that correctly classify an instance, it can be deduced that for the data sets and the constructed classifiers there is a proportion of these that can correctly classify each training sample.

Figure 3: Correlation between the diversity and the accuracy of the systems formed. * Significance at 95% ** Significance at 99%.

Despite the foregoing, it is not advisable to consider only the magnitude of the measure as a criterion to select the one to be used in the proposed method. Taking into account that the quality function includes both the diversity and the accuracy of the system formed, the relationship between these must be considered. In some works [8, 37] a poor relationship between diversity and system accuracy has been reported and in others [1, 28] it is concluded that it is better to form systems with a medium diversity than to form them too diverse. By applying a Pearson correlation analysis between the diversity values and the accuracy of the multiclassifier, it is possible to determine which measure could be selected. In Figure 3 it can be seen that DF measure is the one that has a greater correlation with the accuracy of the system, even when the Pearson coefficient is not high.
Note that the DIF measure, which was the one that had measured the greatest diversity, has practically a zero relationship with the accuracy of the system. Therefore, analyzing diversity according to the successes made by individual classifiers does not seem to be a very good option and it is better to take into account the existence of diversity when the errors in the individual classification coincide. This is precisely the idea that the DF measure considers.

5.3 . System size and accuracy

The Table 5 shows the accuracy of the systems formed depending on the diversity measure used in the quality function of chromosomes.

ID	ρ	D	DF	E	k	DIF
1	**100,0**	**100,0**	**100,0**	**100,0**	**100,0**	**100,0**
2	**90,0**	**90,0**	**90,0**	**90,0**	**90,0**	**90,0**
3	82,8	82,2	83,0	81,9	**83,6**	84,1
4	**93,2**	**93,2**	**93,2**	**93,2**	**93,2**	**93,2**
5	59,7	60,3	**61,2**	60,0	57,4	60,6
6	90,4	91,5	89,2	93,4	**94,3**	90,2
7	56,2	61,5	59,2	57,7	57,7	**62,3**
8	88,0	**90,0**	**90,0**	88,0	89,0	**90,0**
9	73,4	74,1	**75,9**	73,9	74,8	74,7
10	56,9	**64,4**	59,4	59,4	61,9	56,9
11	81,5	81,1	**81,9**	80,7	80,7	**81,9**
12	**99,1**	98,0	97,7	98,3	98,9	98,9
13	**69,7**	68,8	67,1	65,9	**69,7**	67,9
14	99,5	98,6	**100,0**	97,6	**100,0**	**100,0**
15	37,4	37,4	36,8	36,3	**37,9**	37,4
16	79,2	78,3	**80,3**	77,2	79,5	79,2
17	93,0	91,0	**97,0**	88,0	95,0	93,0
18	78,3	**80,0**	79,8	79,1	78,7	78,9
19	76,0	77,5	79,0	76,5	**83,0**	76,0
20	95,3	95,1	**95,4**	94,4	95,3	95,3
AVG	**80,0**	**80,6**	**80,8**	**79,6**	**81,0**	**80,5**

Table 5: Accuracy obtained in the multiclassifiers built according to the diversity measure used in the quality function.

The numerical results do not highlight the use of one measure of diversity over the other to obtain more accurate systems. In order to determine if there were significant differences in the measured accuracy values, the Friedman aligned rank test is applied [23]. This test rejects the null hypothesis (p-value = 1,081E-4 <0,05) for a confidence interval of 95%, so it can be concluded that there are significant differences in at least two pairs of values.
As a second step, a pairwise significance analysis is performed to determine where significant differences exist and with what measure of diversity the best accuracy was obtained. Taking this into account, the Wilcoxon test is applied with the correction of the p-values given by the Finner method [13]. In Figure 4 it can be seen that using the DF diversity measure in the quality function the best results are obtained, considering the range of values determined by the test.

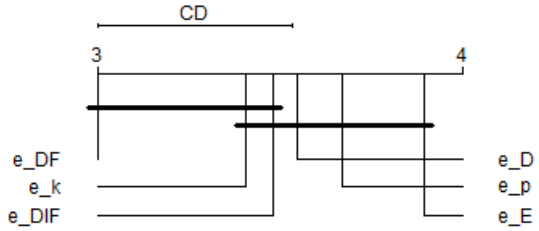

Figure 4: Significant differences between accuracy of the systems formed according to diversity measure used.

Regarding the number of classifiers included in the combinations of the multiclassifiers formed, solutions were obtained with approximately half of the classifiers of the maximum that could be obtained (30 according to the experimental design carried out), to see Table 6.

ρ	D	DF	E	k	DIF
16	**15**	**16**	**14**	**15**	**17**

Table 6: Average of classifiers included in the systems formed according to diversity measure used.

5.4 COMPARISON WITH OTHER ALGORITHMS

In this subsection, the performance of the multiclassifiers formed with the proposed method is compared with other multiclassifiers widely used in the literature. Since *mulGA* constitutes a modification to Bagging (BAG), the existing implementation in WEKA software is used. Similarly, the AdaboostM1 algorithm is used as an implementation of Boosting (BST), with a DecisionStump as classifier.

Both Bagging and *mulGA* use majority vote to combine the outputs of the combined classifiers, hence the existing multiclassifier Vote (VOT) in WEKA is used with the majority vote combination rule and as individual algorithms are used those mentioned in the section 5.1, which are also part of the Ω set used in *mulGA*. These algorithms are also used by Stacking (STK), in addition to using a vector support machine (SMO) as a meta-classifier.

Although many times the Random Forest algorithm (RNF) [6] is considered one of the decision tree algorithms, it can also be considered as an implementation of Bagging but using a decision tree. The algorithm creates its own structure depending on the data set used for training and as a result a set of trees (forest) already trained is obtained. All algorithms kept the default configuration of the parameters, except those that were mentioned previously.

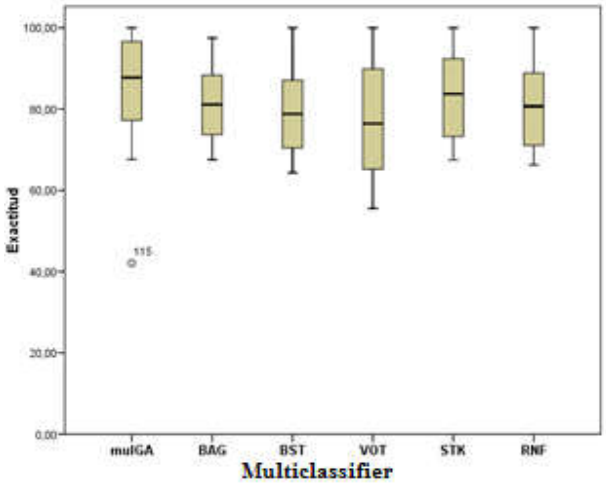

Figure 5: Accuracy obtained in each of the analyzed multiclassifiers.

The Figure 5 shows the accuracy values obtained for each analyzed multiclassifier, including the proposed method. From the results observed in *mulGA*, the system formed using the DF diversity measure in the quality function of chromosomes is used and the best result achieved by this is reported. It can be noted that the proposed method is superior to all the multiclassifiers compared followed by Stacking, which suggests as future work to execute a learning process similar to the one performed by the meta-learner in this algorithm, after the combination of classifiers is optimized with the Genetic Algorithm. We use Friedman's aligned rank

test to determine if there are significant differences between the results of the multiclassifiers. This test is a non-parametric statistical test that compares three or more matched or paired groups where no normality assumption in data is required. Also, this test demonstrating its effectiveness in the comparison of machine learning algorithms [3, 16, 24]. The test reports a p-value = 0,005 <0,05, so differences can be established in at least one pair of multiclassifiers.
On the other hand, the p-values corrected for the Wilcoxon test with Finner's method in Table 7 suggest rejecting, in all cases, the null hypothesis, for a confidence level of 95%.
The Table 7 shows the negative (R-) and positive (R +) ranges obtained in the test, where the latter represent the number of times in which the accuracy of the proposed method is greater than the other multi-classifiers. Taking this into account, the statistical test confirms the superiority of the proposed method over the rest of the multi-classifiers analyzed.

Mult.	*p-value*	R^+	R^-	Finner
BAG	0,0117	14	6	0,0310
BST	0,0034	15	4	0,0282
VOT	0,0052	14	5	0,0282
STK	0,0065	15	4	0,0282
RNF	0,0047	14	5	0,0282

Table 7: Paired analysis of multi-classifier results, using *mulGA* as control.

6. BIOCHEMICAL APPLICATIONS

In the pharmacology environment, the in-silico methods make reference to computers technical applied to the summary, analysis and integration of biological data coming from diverse sources, dedicated to the virtual development of models and simulations able to predict or to outline hypothesis with the purpose of to provide discoveries and relative advances to the medicine and the therapy [19].
The in-silico methods that relate the chemical structure of the molecules with their biological activity are divided in two big groups: the SAR (Structure-Activity Relationship) methods and the QSAR (Quantitative Structure-Activity Relationship) methods. With regard to the last ones, these consist on the construction of mathematical models that relate the molecules structure with their chemical properties and biological effects, assuming that similar molecules have similar properties [47], i.e., the QSAR methods establishes that the biological activity of the medicine is a dependent function of the structural characteristics of the molecule and the activity prediction is carried out by means of the calculation of Molecular Describers (MD) and the application of statistical or quimio-metric techniques [19]. The MD are the final result of a logical and mathematical procedure that transforms coded chemical information, inside a symbolic representation of a molecule, in a number of utility or in the result of some standardized experiment.
In this case, two data sets of QSAR models are used to identify compounds with activity in presence of parasite Trypanosoma cruzi (illness of Chagas) and in presence of inflammatory processes of the human body (Inflammatory processes) respectively, this helps in the process of medicines elaboration for the prevention or cure of these illnesses or sufferings. These data sets are provided by the Center of Chemicals Bioactive (CCB) of Central University “Marta Abreu” of Las Villas. For the construction of the data sets several MD were used and also the statistical technique Lineal Discriminant Analysis (LDA), implemented in the Statistical package.
The classification of a compound (an example or an instance) is active or inactive and it can be expressed in two ways: the first one is the probability granted by the classifier to classify the compound as active/inactive, or the second based on the difference of the probabilities (ΔP), in the last case when the difference is bigger than zero the compound will be classified as active, otherwise it will be classified as inactive.
These data sets are characterized to present big features number, each one of feature represents coded chemical information, in Table 8 the characteristics of them are described.

Data sets	**Nominal features**	**Numeric features**	**Classes**	**Cases (examples)**	**Distribution by class**
Trypanosoma cruzi	1	22	2	650	325-325
Inflammatory	1	45	2	592	260-332

Table 8: Principal characteristics of data sets.

As part of the pre-processing of these sets for their specific characteristics was carried out an analysis of consistency in the data to avoid the existence of equal compounds (examples) assigned to different classes. As

a result of this analysis is proven the consistency in the same ones. Then it was thought of using some technique of features selection, but for the importance that is assigned to the features in this applications type is not carried out.
The objective of the researchers in the CCB is to select the models that combined increase the precision of the analysis, if the certainty in the prediction of the activity of a compound in presence of certain illness is bigger, then bigger quantity of money and time can be saved in the conformation process of the medicine to combat this illness, for this reason the importance of this process.
In summary, it is wanted to look for the best combinations of classifiers to increase the precision of the analysis, for this is taking into account the same 11 algorithms in previous section with their default parameters. Regarding the genetic algorithm is maintained the same configuration and is used the cross-validation process too. Also, according with the results in previous section is used the DF measure to quantify the diversity between the classifiers.
At the request of the researchers interested in these applications the evaluation measures used in the classification were several: Mathews coefficient, Accuracy, Sensibility, Specificity, Reason of False Positive (RFP), Precision and Reason of False Discoveries (RFD). Also, as the request of the researchers in these cases, the combinations of classifiers are selected using several combination methods for the classifiers outputs as: Majority vote (MAJ), Average (AVG), Product (PROD), Maximum (MAX) and Minimum (MIN). The results of *mulGA* are taking into account to look for the best combinations in both data sets.
As selection approach for the best combinations the researchers interested in these applications have mainly in consideration the Accuracy and the Reason of False Positive (RFP). The accuracy should be the biggest possible value, near to 100%, while the RFP should be the smallest possible value, closer to 0%. In many occasions for these applications type the selection approach for the best combinations is subjective and it depends on the researcher reasoning, independently that the proposed method will offer the best combinations as for accuracy and diversity.
In the case of *Trypanosoma cruzi*, the best individual classifier has an accuracy of 79%, while the best classifier according to the RFP has a value of 16%. These will be the values to overcome in the search of the best combinations. The results are shown in Table 9, where is presented the five better combinations of classifiers that are found. The accuracy of the best classifier is overcome, also, is pointed out with more importance those combinations where the RFP diminished with regard to the smallest value obtained by the individual classifiers. Only in one combination the RFP doesn't diminish regarding this value.
In the case of *Inflammatory* the best individual classifier has an accuracy of 90%, while the best classifier according to the RFP has a value of 5,45%. These will be the values to overcome in the search of the best combinations. The results are shown in Table 10, where is presented the five better combinations of classifiers. The accuracy of the best classifier is overcome, also, is pointed out with more importance those combinations where the RFP diminished with regard to the smallest value obtained by the individual classifiers. In three combinations the RFP doesn't diminish regarding this value.

Combinations	C´ Matthews	Accuracy	Sensibility	Specificity	RFP	Precision	RFD
mulGA							
J48_MLP_IBk (MAJ)	66,74	80,33	84,67	80	16	80,8	15,2
DS_L (AVG)	68,01	**83**	84	82	**14**	82,31	14,69
MLP_NB_ZR (AVG)	68,69	**83,33**	84,67	82	**14**	82,42	14,58
SMO_NB_RepT (MAX)	60,01	**81,33**	68	88,67	**7,33**	86,24	9,76
21_4_20 (MAJ)	67,34	**82,67**	83,33	82	**14**	82,21	13,79

Table 9: Statistical results of the combinations for *Trypanosoma cruzi*.

Combinations	C´ Matthews	Accuracy	Sensibility	Specificity	RFP	Precision	RFD
mulGA							
SMO_NB_L_J48_SL (MAJ)	89,08	**91,59**	92,37	92,76	**5,45**	91,88	2,12
MLP_SL_SMO_IBk (AVG)	88,62	**92,36**	91,85	92,76	**5,45**	92,85	2,15
MLP_L_NB (AVG)	90,46	**93,27**	92,89	93,57	**4,43**	92,89	3,11
RepT_J48 (PROD)	80,76	91,96	87,19	89,52	8,48	89,67	3,33
J48_MLP_SL (MAJ)	90,49	**93,27**	93,93	92,76	**4,24**	91,94	2,06

Table 10: Statistical results of the combinations for *Inflammatory*.

7. CONCLUSIONS

This paper has presented a modification of the method followed by Bagging to build a multiclassifier. The modification consists of allowing the use of more than one learning algorithm in the training stage and the use of optimization with metaheuristic to obtain a combination of diverse and precise classifiers.
The proposed method considers the diversity measured in the training of each learning algorithm and at the same time the accuracy of the multiclassifier formed. Six diversity measures reported in the literature were analyzed and it was determined that, although with the DIF measure the highest diversity values were obtained, the DF measure is the one that best relates to the accuracy of the systems formed.
The analysis of significant differences between the accuracy obtained with the proposed method and some multiclassifiers reported in the literature shows the superiority of the proposed method. However, according to the results observed for the Stacking multiclassifier, the use of a meta-learner can be suggested in the proposed method to learn from the behavior of the combined classifiers after the optimization of the metaheuristic. Finally, real applications of biochemical are presented, where satisfactory results are reached with the proposed method.

RECEIVED: OCTOBER , 2020.
REVISED: JANUARY, 2021

REFERENCES

[1] ALBUQUERQUE, R. A. S. (2018): **Seleção dinâmica de comitês de classificadores baseada em diversidade e acurácia para detecção de mudança de conceitos**. PhD Thesis, Universidade Federal do Amazonas, Brasil.

[2] ASUNCION, A. and D. J. NEWMAN. (2017): **UCI Machine Learning Repository.** Available: http://www.ics.uci.edu/ml

[3] BENAVOLI, A., CORANI, G., and MANGILI, F. (2016): Should we really use post-hoc tests based on mean-ranks?. **The Journal of Machine Learning Research**, 17, 152-161.

[4] BONET, I., FRANCO, P.E., RIVERO, V., TEIJEIRA, M., BORGES, F., URIARTE, E. and MORALES, A. (2013): Classifier ensemble based on feature selection and diversity measures for predicting the affinity of A2B adenosine receptor antagonists, **Journal of chemical information and modeling**, 53, 3140-3155.

[5] BREIMAN, L., FRIEDMAN, J., STONE, C.J. and OLSHEN, R. A. (1984): **Classification and regression trees**. Taylor & Francis. Wadsworth statistics/probability series. Monterey, CA. CRC Press.

[6] BREIMAN, L. (2001): Random forests, **Machine Learning**, 45, n 5-32.

[7] BUCZAK, A. L. and GUVEN, E. (2015) : A survey of data mining and machine learning methods for cyber security intrusion detection,**IEEE Communications Surveys & Tutorials**, 18, 1153-1176.

[8] BUTLER IV, H. K., FRIEND, M. A., BAUER JR, K. W. and BIHL, T. J. (2018): The effectiveness of using diversity to select multiple classifier systems with varying classification thresholds, **Journal of Algorithms & Computational Technology**, 12, 187-199.

[9] CABRERA, L. (2019): **Método para la selección de combinaciones de clasificadores**. PhD Thesis, Universidad Central "Marta Abreu" de Las Villas, Cuba.

[10] CABRERA, L., MORALES, A. and CASAS, G. M. (2016): Medidas de diversidad para la construcción de sistemas multiclasificadores usando algoritmos genéticos, **Computación y Sistemas**, 20, 729-747.

[11] CABRERA, L., MORALES, A., CASAS , G. M. and MARTÍNEZ, Y. (2015): Genetic Algorithms with diversity measures to build classifiers systems, **Investigación Operacional**, 36, 206-224.

[12] CABRERA, L., SANTOS, L. R., NÁPOLES, G., MORALES, A., CASAS, G. M., GARCÍA, M. M. and MARTÍNEZ, Y. (2017): Building multi-classifier systems with ant colony optimization, **Investigación Operacional**, 38, 407-423.

[13] CALVO, B. and SANTAFÉ, G. (2016): scmamp: Statistical comparison of multiple algorithms in multiple problems, **The R Journal**, 8, 248-256.

[14] COOMANS, D. and MASSART, D. L. (1982): Alternative k-nearest neighbour rules in supervised pattern recognition: Part 1. k-Nearest neighbour classification by using alternative voting rules, **Analytica Chimica Acta**, 136, 15-27.

[15] CUNNINGHAM, P. and CARNEY, J. (2000): Diversity versus Quality in Classification Ensembles Based on Feature Selection, in the **Machine Learning: ECML 2000**, editors: R. López de Mántaras, E. Plaza. Lecture Notes in Computer Science, vol. 1810, Springer, Berlin, Heidelberg.

[16] DEMŠAR, J. (2006): Statistical comparisons of classifiers over multiple data sets, **Journal of Machine Learning Research**, 7, 1-30.

[17] DU, S., LIU, C. and XI, L. (2015): A selective multiclass support vector machine ensemble classifier for engineering surface classification using high definition metrology, **Journal of Manufacturing Science and Engineering**, 137, 011003.

[18] DUVAL, M. A., SHULCLOPER, J. R. and VEGA, S. (2012): Combinación de clasificadores supervisados: estado del arte, **Reporte Técnico Reconocimiento de patrones**. Serie Azul RNPS No 2142. CENATAV, RT-048. ISSN: 2072-6287. La Habana-Cuba.

[19] EKINS, S., MESTRES, J. and TESTA, B. (2007): In silico pharmacology for drug discovery: methods for virtual ligand screening and profiling, **British Journal of Pharmacology**, 152, 9-20.

[20] FISHER, R. A. (1936): The use of multiple measurements in taxonomic problems, **Annals of eugenics**, 7, 179-188.

[21] FLEISS, J. L. (1971): Measuring nominal scale agreement among many raters, **Psychological Bulletin**, 76, 378-382.

[22] FLEISS, J. L. (1981): **Statistical Methods for Rates and Proportions**, John Wiley & Sons.

[23] FRIEDMAN, M. (1937): The use of ranks to avoid the assumption of normality implicit in the analysis of variance, **Journal of the american statistical association**, 32, 675-701.

[24] GARCÍA, S. and HERRERA, F. (2008): An extension on "statistical comparisons of classifiers over multiple data sets" for all pairwise comparisons, **Journal of Machine Learning Research**, 9, 2677-2694.

[25] GIACINTO, G. and ROLI, F. (2001): Design of effective neural network ensembles for image classification purposes, **Image vision and Computing Journal**, 19, 699-707.

[26] GOLDBERG, D. E. and HOLLAND, J. H. (1988): Genetic algorithms and machine learning, **Machine Learning**, 3, 95-99.

[27] GRAU, I., NÁPOLES, G., BONET, I. and GARCÍA, M. M. (2013): Backpropagation through time algorithm for training recurrent neural networks using variable length instances, **Computación y Sistemas**, 17,. 15-24.

[28] HADJITODOROV, S. T., KUNCHEVA, L. I. and TODOROVA, L. P. (2006): Moderate diversity for better cluster ensembles, **Information Fusion**, 7, 264-275.

[29] HANSEN, L. K. and SALAMON, P. (1990): Neural Network Ensembles, **IEEE Transactions on Pattern Analysis and Machine Intelligence**, 12,. 993-1001.

[30] HERRERA, F., LOZANO, M. and VERDEGAY, J. L. (1998): Tackling real-coded genetic algorithms: Operators and tools for behavioural analysis, **Artificial intelligence review**, 12, 265-319.

[31] IMPEDOVO, D., PIRLO, G. and BARBUZZI, D. (2012): Multi-classifier System Configuration Using Genetic Algorithms, in the **International Conference on Frontiers in Handwriting Recognition**. IEEE, 560-564.

[32] KORYTKOWSKI, M., RUTKOWSKI, L. and SCHERER, R. (2016): Fast image classification by boosting fuzzy classifiers, **Information Sciences**, 327, 175-182.
[33] KRAWCZYK, B., MINKU, L. L., GAMA, J., STEFANOWSKI, J. and WOŹNIAK, M. (2017): Ensemble learning for data stream analysis: A survey, **Information Fusion**, 37, 132-156.
[34] KUMAR, A., KIM, J., LYNDON, D., FULHAM, M. and FENG, D. (2016): An ensemble of fine-tuned convolutional neural networks for medical image classification, **IEEE Journal of Biomedical and Health Informatics**, 21, 31-40.
[35] KUNCHEVA, L. I. (2004): **Combining pattern classifiers: Methods and Algorithms.** John Wiley & Sons, Inc., New Jersey.
[36] KUNCHEVA, L. I. and JAIN, L. C. (2000): Designing classifier fusion systems by genetic algorithms, **IEEE Transactions on Evolutionary Computation**, 4, 327-336.
[37] KUNCHEVA, L. I., and WHITAKER, C. J. (2003): Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy, **Machine Learning**, 51, 181-207.
[38] MINSKY, M. (1961): Steps toward artificial intelligence, **Proceedings of the IRE**, 49, 8-70.
[39] MORALEZ, A., CABRERA, L. and FERNÁDEZ, C. (2018): Majority Vote Modification to consider the classifier experience on multiclassifiers systems, in the **IV International Conference on Informatics and Computer Sciences**, "CICCI 2018", Habana, Cuba.
[40] NASCIMENTO, D., COELHO, A. and CANUTO, A. (2014): Integrating complementary techniques for promoting diversity in classifier ensembles: A systematic study, **Neurocomputing**, 138, 347-357.
[41] NOWLAN, S. J. and HINTON, G. E. (1991): Evaluation of adaptive mixtures of competing experts, **Advances in Neural Information Processing Systems (NIPS)**, 3, 774-780.
[42] OLSON, R. S., LA CAVA, W., ORZECHOWSKI, P., URBANOWICZ, R. J. and MOORE, J. H. (2017): PMLB: a large benchmark suite for machine learning evaluation and comparison, **BioData Mining**, 10, 1-37.
[43] QUINTANA, J. C., QUINTANA, N., GIRÁLDEZ, R., MOLINA, R. and SANTIESTEBAN, C. E., (2017): Predictor de interacciones entre estructuras secundarias de proteínas, **Revista Cubana de Ciencias Informáticas**, 11, 105-113.
[44] SCHAPIRE, R. E. (1990): The strength of weak learnability, **Machine Learning**, 5, 197-227.
[45] SHIPP, C. A. and KUNCHEVA, L. I. (2002): Relationships between combination methods and measures of diversity in combining classifiers, **Information Fusion**, 3, 135-148.
[46] SKALAK, D. B. (1996): The sources of increased accuracy for two proposed Boosting algorithms, in the **Proc. American Association for Artificial Intelligence, AAAI-96**, Integrating Multiple Learned Models Workshop, 120-125.
[47] TROPSHA, A. (2010): Best practices for QSAR model development, validation, and exploitation, **Molecular Informatics**, vol. 29, nro 6-7, pp. 476-488.
[48] VERDECIA-CABRERA, A., BLANCO, I., DOMÍNGUEZ, L. and SARABIA, Y. (2018): Learning with ensembles from non-stationary data streams, **Inteligencia Artificial**, 21, 145-158.
[49] WOLPERT, D. H. (1992): Stacked generalization. **Neural networks**, 5, 241-259.
[50] ZHANG, Y., ZHANG, H., CAI, J. and YANG, B. (2014): A weighted voting classifier based on differential evolution, **Abstract and Applied Analysis**, ID 376950, 376950, ISSN: 1085-3375.