Journal of Information Technology Management

Print ISSN: 2008-5893

Online ISSN: 2423-5059

Feature Selection Using a Genetic Algorithms and Fuzzy logic in Anti-Human Immunodeficiency Virus Prediction for Drug Discovery

Houda Labjar

Researcher, Laboratory Processes and Environment, Faculty of Sciences and Technology, University Hassan II Casablanca, Mohammedia, Morocco. E-mail: h.labjar@gmail.com

Mohammad Al-Sarem

Associate Professor, Information System Departement, Taibah University, Al-Madinah Al-Monawarah, Saudi Arabia. E-mail: mohsarem@gmail.com

Mohamed Kissi*

*Corresponding author, Full Professor, LIM Laboratory, Computer Science Department, Faculty of Sciences and Technology, University Hassan II Casablanca, Mohammedia, Morocco. E-mail: mohamed.kissi@fstm.ac.ma

Abstract

This paper presents an approach that uses both genetic algorithm (GA) and fuzzy inference system (FIS), for feature selection for descriptor in a quantitative structure activity relationships (QSAR) classification and prediction problem. Unlike the traditional techniques that employed GA, the FIS is used to evaluate an individual population in the GA process. So, the fitness function is introduced and defined by the error rate of the GA and FIS combination. The proposed approach has been implemented and tested using a data set with experimental value anti-human immunodeficiency virus (HIV) molecules. The statistical parameters q2 (leave many out) is equal 0.59 and r (coefficient of correlation) is equal 0.98. These results reveal the capacity for achieving subset of descriptors, with high predictive capacity as well as the effectiveness and robustness of the proposed approach.

Keywords: Feature Selection, Machine Learning, Computational Chemistry, QSAR, Fuzzy Logic, Genetic Algorithms.

DOI: https://orcid.org/10.22059/jitm.2022.84883 Manuscript Type: Research Paper University of Tehran, Faculty of Management

Received January 12, 2020 Accepted March 25, 2020

Introduction

Molecular computing is an emerging discipline that addresses mathematical and computer problems, related to the search for the relationship and information that exist between the structure and activity of molecules. The aim is to discover new knowledge in several fields such as pharmacology and environmental science (Danishuddin & Khan, 2016; Fenner & Tratnyekc, 2017). In particular, the modeling of quantitative structure-activity relationships (QSAR) is an active area of research in molecular modeling. QSAR models have been proposed to estimate biological property, such as activity (Swathik et al., 2019), with the aim of providing adequate and relevant information for drug discovery (Hdoufane et al., 2019).

QSAR models require the representation of the chemical, physical and electronic structure of compounds by their molecular parameters (Todeschini & Consonni, 2009), such as the constitutional, geometric and quantum descriptors. The development of a QSAR model consists of calculating all molecular descriptors, but only a subset of these descriptors provides the in-formation needed to obtain the powerful predictive model (Eklund et al., 2014). Subsequently, obtain the powerful predictive models is based on the good selection of the molecular descriptors used during the generation of the QSAR model (Grisoni et al., 2018).

Several algorithms of machine learning have been used for processing the selection of molecular descriptors, which are known in informatics as feature selection methods. Generally, the used approaches are divided into three types: filter, wrapper, and embedded methods. A dis-advantage of the filter approaches is that they examine each feature independently and ignore the individual performance of the feature in relation to the all-group features (Liu et al., 2018). This problem can affect the machine learning results used below. For wrapper and embedded meth-ods, the machine learning algorithms requires an implementation heavy in time complexity, which constitutes an important problem.

The genetic algorithm (GA) is a method that can be used for feature selection (Wutzl et al., 2019). Often, feature selection methods used GA requires a classification method as support vectors machines (SVM) (Nagasubramanian et al., 2018), artificial neural networks (ANN) (Lab-jar et al., 2016), k-nearest neighbor (KNN) (Salari et al., 2014) and decisions trees (DT) (Sri-vastava et al., 2019), for evaluate each individual in the population. All features found by these methods are common and it is not possible to explain the influence nature of these features on the treated class. In addition, these approaches require considerable training time for each individual set.

In this work, a new approach feature selection based on GAs is proposed, which incorporates with a classification module based on a fuzzy inference system (FIS) for each iteration in this GA. The individual evaluation is performed in each iteration using the same classification method. The fitness function is represented through the classification error found by the FIS as-sociated with features selected via the GA. More specifically, the FIS is formed for each iteration before GA starts to search new features set by improving the FIS error. This approach will be tested on the anti-HIV QSAR problem, then it will be compared with the new approaches in the literature dealing with the QSAR modeling problem.

The structuration of the rest paper is as follows: in section 2, the approach of feature selection is briefly motivated by presenting feature selection problem in general. GAs and FIS are reviewed in Section 3 and the feature selection approach, based on GA and FIS, is presented. Section 4 is devoted to experimental results with analysis and discussion, and this work is con-cluded in Section 5.

Literature Review

The feature selection is a technique for choosing characteristics, variables, and attributes that more interesting, relevant and informative for supervised classification problem well as unsupervised. It consists of choosing from a set of large variables, a subset of interesting features for the studied problem. The application fields of feature selection techniques are varied, for example: modeling areas, classification, machine learning and data mining, but the aim of this work is to propose an approach for the feature selection in the supervised classification case. In this case, the purpose feature selection is to find an optimal sub-set of features that are relevant and non-redundant. In addition, this sub-set should satisfy accuracy and as well as rapid learning, or yet the applicability of the proposed model classification, that is to say that the sub-set must be confirmed by the human expert.

Feature selection algorithms are used to extract a non-redundant information and relevant to the efficient exploitation of growth data. They are divided into three categories: filters, wrappers (Guyon et al., 2002) and embedded methods (Lal et al., 2006). Filter methods operate directly on data set and providing a subset of output variables. These methods are rapid and independent of the classification model. Wrapper methods searching in the space of variable subsets, guided by model results such as cross-validation performance of training data. They often have better results than filtering methods, but they are costly in terms of computing time. Finally, embedded methods witch uses internal information model classification (for example, the weight vector in the case of SVM), these methods are therefore close envelopes methods, because they combine the process exploration with a learning algorithm without validation step for maximizing the quality of fit and minimize the number of attributes.

QSAR models establish the relationship between structure and activity of chemical or biological compounds (Swathik et al., 2019). The process, for achieving QSAR models, is based on supervised machine learning algorithms. These models have the ability to classify correctly the compounds forming the training database, and to predict the activity of newly synthesized compounds (Liu et al., 2019). The realization of these models begins with a selection step of the most relevant molecular descriptors for the modeling of the activity target (Racz et al., 2019), which constitutes one of the characteristic selection problems. A training database of compounds is used to carry out this descriptor selection process. This training database includes compound examples with their molecular descriptors and the measured activity of each compound. The machine learning strategy consists in selecting and evaluating the different descriptors in order to identify the smallest and best subset of these descriptors.

The proposed approach

In this part, the concepts and the proposed approach for feature selection methodology is described, it includes the concepts of genetic algorithms and fuzzy inference systems.

a) Genetic Algorithms

The GAs is one of the evolutionary-based techniques that are used nowadays intensively in the feature selection field. The application of GAs, to solve an optimization problem, requires encoding all potential solutions to this problem in chromosome forms. This is to find a good selection function for discrimination between the set of chromosomes using genetic operators. Recently, many works in pattern recognition systems used evolutionary algorithms for feature selection. In this sense, genetic algorithms method has proven effective as a feature selection technique QSAR studies (Pourbasheer et al., 2014).

The GAs is a stochastic optimization algorithms that are based on the mechanisms of natural selection and genetic (Holland, 1992). The algorithm started with an initial chromosome population arbitrarily chosen and the performance or fitness of each chromosome is evaluated. A genetic algorithm is an iterative algorithm for optimum search; it manipulates a population of constant size. Constant population size causes a phenomenon of competition between chromosomes. Each chromosome represents the coding of a potential solution to the asked problem; it consists of elements called genes. For each iteration, called generation, a new population is created with the same number of chromosomes. This generation involves better chromosomes adapted to their environment using selection function. As the generations, the chromosomes will tend towards the optimum selective function. The creation of a new population from the previous one by applying the genetic operators: selection, crossover and mutation (See Figure 1).

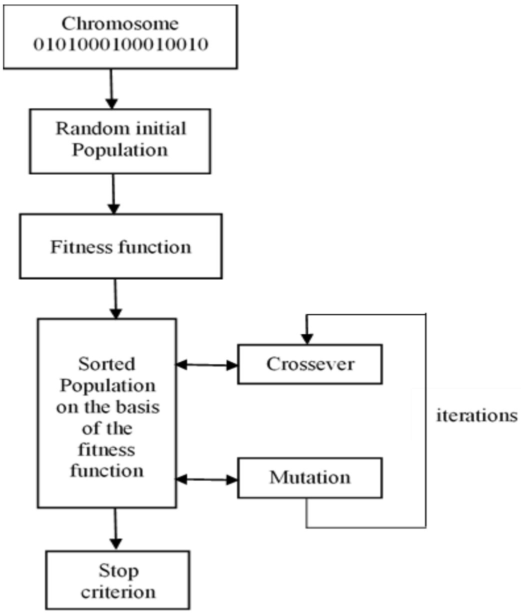

Figure 1. Principle of the genetic algorithm

The are several operators that guide the GA process, some of them are listed below:

- **Selection**: is a process in which a chromosome is copied into new next population based on the function values, to be optimized, for this chromosome. That is to give the chromosomes, whose fitness function is great, higher probability to contribute in the next generation.
- **Crossover**: the simple crossover or at a cut point consists in step to choose a pair of chromosomes with probability p, generally defined by user, and then in a second step, the representative chains are cut in the same random position in both parents. This then produces two head segments and two tail segments, finally the two tail segments of parents are permuted to obtain two children who inherit some characteristics from their parents.
- **Mutation**: is defined as a bit inversion in a chromosome. That is to randomly change the parameter value. Mutations play the noise role and prevent evolution from freezing. They make it possible to ensure a global as well as local search, depending on the weight and the number of mutated bits. In addition, they ensure that the mathematically optimum overall can be achieved.

b) Fuzzy inference systems

Fuzzy inference systems (FIS) are techniques for approximating complicated nonlinear functions. They take the inputs and process them according to the predefined rules to produce the outputs. Outputs and inputs have crisp value, while internal processing is based on fuzzy arithmetic and fuzzy rules. The five functional components for FIS are as follows:

- A data base defining the membership functions of fuzzy sets.
- A rule base containing a number of fuzzy if–then rules.
- A fuzzification interface which transforms crisp inputs to linguistic variables.
- A decision-making unit as the inference engine.
- A defuzzification interface converting fuzzy outputs to crisp outputs.

Adaptive neural network fuzzy inference system (ANFIS) is an architecture, which is functionally equivalent to a Takagi-Sugeno fuzzy (TSF) rule base, whose parameters are tuned by a learning algorithm using input–output data available (Takagi & Sugeno, 1985). Assume a simple Takagi-Sugeno FIS with two inputs x and y and one output f and a rule base with two fuzzy if–then rules as follows:

- Rule1. If x is A_1 and y is B_1 then $f_1 = p_1x + q_1y + r_1$
- Rule2. If x is A_2 and y is B_2 then $f_2 = p_2x + q_2y + r_2$

where A_1, $A2$ and B_1, B_2 are, respectively fuzzy sets of input premise variables x and y; and p_1, q_1, r_1 and p_2, q_2, r_2 are parameters of the consequent or output variable. The ANFIS structure is presented in Figure 2 wherein square nodes are fixed nodes and circle nodes are adaptive which change their values during the training process. In ANFIS, a hybrid learning algorithm is used to optimize the parameters of membership functions of input variables in antecedent part of fuzzy rules, using a steepest descent algorithm. While in the consequent part, the linear parameters of the output variable are optimized using a least square method. The final output of the given network with two inputs and one output in according to the above parameters is calculated as follows (Jang, 1993):

$$f=\frac{w_1}{w_1+w_2}f_1+\frac{w_2}{w_1+w_2}f_2=\overline{w}_1 f_1+\overline{w}_2 f_2$$

$$=\overline{w}_1(p_1\,x+q_1\,y+r_1)+\overline{w}_2(p_2\,x+q_2\,y+r_2)$$

$$=(\overline{w}_1 x)\,p_1+(\overline{w}_1 y)q_1+\overline{w}_1 r_1)+(\overline{w}_2 x)p_2+(\overline{w}_2 y)q_2+\overline{w}_2 r_2)$$

$$w_i=\mu_{A_i}(x)\mu_{B_i}(y)\quad i=1,2$$

$$\mu_{A_i}(x)=\exp\left[-\left(\frac{x-c_i}{a_i}\right)^2\right]$$

$$\mu_{B_i}(y)=\exp\left[-\left(\frac{y-d_i}{e_i}\right)^2\right]\quad i=1,2$$

where w_i is called the firing strength of rule i, $\mu_{A_i}(x)$ and $\mu_{B_i}(y)$ are, respectively the membership degrees of x and y in Ai and Bi, ci and di are, respectively the mean values of the gaussian membership functions defined for fuzzy sets Ai and Bi and ai and ei are, respectively the standard deviations of the membership functions.

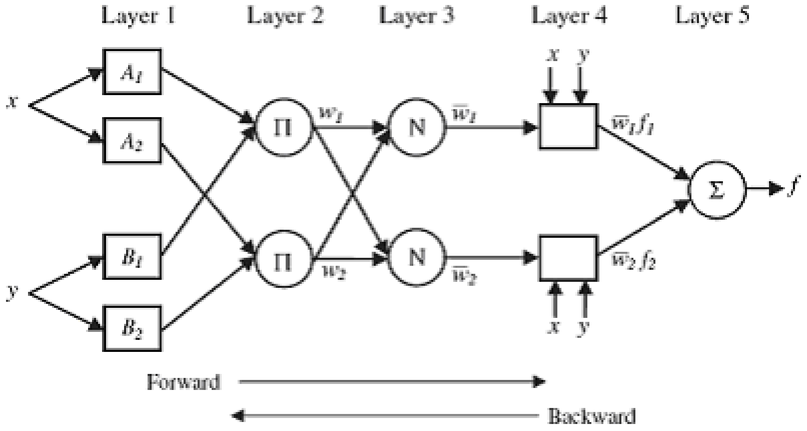

Figure 2. ANFIS model structure

c) Hybrid feature selection techniques

In this part, the choice of the fitness function which constitutes the originality of the proposed approach is explained. Then, the different steps of this approach are as follows:

- For the fuzzy inference classification, the proposed approach is presented for the implementation of a GA, the feature selection and the confirmation of this selection by constructing a fuzzy inference system (ANFIS) based on these features.
- For the identification of relevant features, the GA is used to extract them. These features will be as basic parameters for the construction of an optimized fuzzy inference system. During the learning, the elements of each database constituted by the features found by the GA, will each be presented to the FIS.

An architecture of the general approach is presented in the Figure 3:

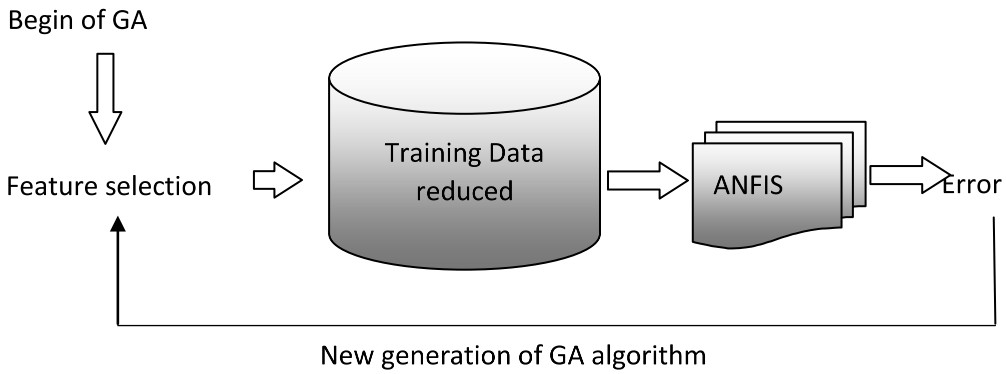

Figure 3. Model structure of the proposed approach

The algorithm of the approach is as follows:

Algorithm

```
begin
    Step 1: Initial population of nb descriptors
    Step 2: Selection of the best features according to their correlation coefficients
    Step 3: Injection of the best features to generate a FIS and quadratic error calculation
    of learning examples
    Step 4: If stop test is checked
        then give the features found
        else
        o Crossover: addition of a descriptors: new population with nb + 1 descriptors
        o Mutation: alteration of nb+1 descriptor
        o Step 2
        End if
End
```

The proposed approach determines the best features (the number nb is fixed at the beginning) according to their correlation coefficients by applying the multiple linear regression and their influence on the class. This step is the first population of the GA. The built learning base will be injected into the FIS system of the TSK type, a classification error will be retained as a fitness function for the genetic algorithm. If the stopping criterion is not satisfied (Error), a second generation of GA is constructed by applying the crossover principle. The Crossover for the proposed approach based on adding a feature to the selected feature set before. The goal is to have the minimum of features that gives a better FIS with a minimal classification error. The mutation operator consists to alter different nb+1 descriptors of all possible descriptors in the initial population.

Experiment's setups

In the experiment's setups, the dataset of recognition of anti-HIV activity is used. This dataset consists of 79 anti-HIV molecules with their inhibitory activity. It is taken from articles published by Tanaka et al., (1992) and Garg et al., (1999). The anti-HIV activity of the compounds has been expressed by the ability of the compound to protect cells against the cytopathic effects of the virus. This activity, the concentration needed for 50% effect, was measured and expressed in pIC50. The objective of QSAR studies is to search the relationship that exists between the structure of molecules and their activity, in particular the study activity anti-HIV.

A major step in all QSAR studies is to select and compute descriptors as coded numeric variables representing chemical, physical, geometric and electronic structures. As all the compounds studied have the same basic skeleton, each molecule is described by means of properties of the substituents, linked to the basic skeleton. Determining the relevant properties for a given substituent is useful for assessing local interactions between the molecule and the receptor site. In this work, the properties of the molecule that looked are its size, its height, its molecular weight and its lipophilicity etc. All these properties have been measured to know the possibility of being transported, accessed and interacted with the given receiving site. These descriptors show the hydrophobic, steric, geometric and electronic aspects.

The descriptors of the molecules used are:

- *LogP*: logarithm of the partition coefficient between the water and the octanol of the molecule.
- X_1: 4th order connectivity index
- X_2: 6th order connectivity index
- *S*: surface of the molecule
- *V*: volume of the molecule
- B_1, B_2: the parameters of the substituents of the molecule
- *MR*: molecular reactivity
- *MW*: Molecular weight
- *Ov*: Ovality estimation

- *L*: Length is the distance along the screen x-axis between the left- and rightmost atoms.
- *W*: Width is the distance along the screen y-axis between the top and bottommost atoms.
- *H*: Height is the distance along the screen z-axis between the nearest and farthest atoms.
- The ratios *V/L*, *V/W*, *W/H* were also calculated.

The number of all descriptors or variables for each molecule is 16, and calculated using Molecular Modelling Pro software (MMP, 2020).

Results and discussion

In this section, the application of the proposed approach is described. The results of the evaluation are given, which concerns the number of selected features, the specificity, sensibility and accuracy rates.

Each example contains values for these variables, is associated with a class value (anti-HIV). Consider an instance (example) of the dataset studied, its description is associated with the value "anti-HIV = 1", to obtain an example of learning with the all descriptors (See Table 1).

Table 1. Description of an instance example

X_1	B_1	V	X_2	Ov	B_2	L	MW	W	H	V/L	V/W	W/H	$LogP$	MR	S	Class
0.26	1	1.55	0.042	0.58	1.52	2.25	250	4.5	3.25	0.69	0.34	1.38	1.87	0.57	0.65	1

Once all the examples (molecules) described by all different descriptors, the anti-HIV activity (the class) is coded by a variable indicating the absence, the weak or the strong presence of the activity.

The GA will start with two descriptors chosen at random from the all-descriptors givens at the beginning. The two best features will be fed into the FIS and the classification error is calculated for all learning data examples.

The proposed genetic algorithm will increase the number of descriptors for each iteration and look for the best features and inject them back into the fuzzy inference system. After a certain number of generations fixed at the beginning, the results found are recapitulated in Table 2.

For each selected features, sensitivity, specificity and accuracy are showed of the obtained classification result. Accuracy is calculated as the ratio between the number of correctly classified molecules and the total number of molecules. Sensitivity and Specificity are defined and calculated as follows:

$$Sensibilty = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

Where TP: true positive is correctly anti-HIV=1, FP: false positive incorrectly anti-HIV=1, TN: true negative is correctly anti-HIV=0, FN: false negative incorrectly anti-HIV=0.

Table 2. Evaluation of classification performance for each selected descriptor

Number of descriptors	Descriptors	Error	Sensibility (%)	Specificity (%)		Accuracy (%)
2	X_1, S	0.134	92	81		87
2	L, H	0.23	82	71		77
3	$X_1, logP, S$	0.033	99	91		97
3	X_1, V, W	0.25	81	73		76
4	$X_1, B_2, X_2, logP$	0.201	85	74		80
4	$L, V/W, X_1, logP$	0.27	78	72		73
5	$S, B_2, X_2, LogP, X_1$	0.128	95	82		88
5	$S, X_1, X_2, LogP, L$	0.14	90	80		86
6	$S, X_1, X_2, LogP, L, B_1$	0.33	77	64		67
6	V, H, X_2, L, MR, B_2	0.36	76	62		64

Following the results of this table, it is noted that the features number for most descriptors is reduced by improving the accuracy, sensibility and specificity rates. On the other hand, it is remarkable the efficiency the propose approach to select the most relevant features after selecting the following three features: X_1, *LogP* and *S*.

The statistical quality of the proposed approach, in particular the best features selected, was examined by different parameters like coefficient of correlation *r*, regression standard error *s*. The generated QSAR models were validated by leave-one out (LOO), and cross-validation q^2 (Buglak et al., 2019).

Table 3. The best selected features results

Number of descriptors	Descriptors	r	s	q^2
2	X_1, S	0.87	0.35	0.49
3	X_1, logP, S	0.98	0.24	0.59
4	X_1, B_2, X_1, logP	0.81	0.45	0.38
5	S, B_2, X_2, LogP, X_1	0.89	0.36	0.50

The number of selected descriptors which is equal to 3 were accepted for the following reasons (see Table 3):

- $r \geq 0.95$.
- *s* is not much larger than the standard deviation of the biological experiment.
- $q^2 \geq 0.55$.

The predict activity (predicted pIC50) by the proposed approach, genetic algorithm with ANFIS, and real activity (pIC50) for each example are compared in Figure 4. In this Figure, the real activity for compounds data and the predict activity training data are same degrees for 66 examples and with minor error equal 0.00901 for the 15 remaining compounds. It proved that the GA-ANFIS approach used in this work is feasible and could be used to predict the pIC50 activity.

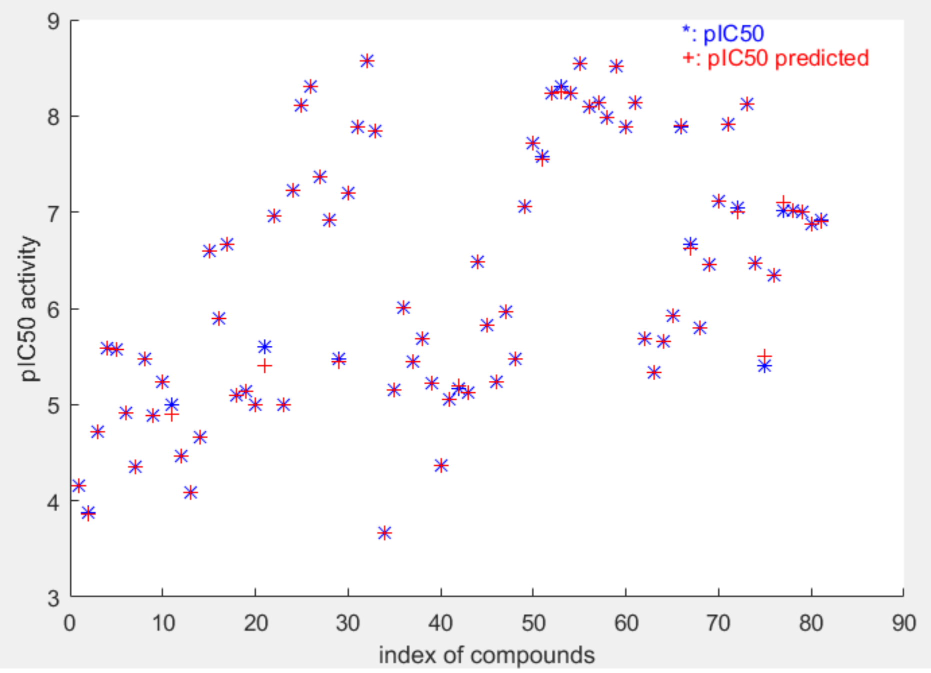

Figure 4. Comparison of real and predicted activity

For more investigation, artificial neural networks (ANN) (Anacleto et al., 2019), support vector machines (SVM) (Wen et al., 2017), multi linear regression (MLR) with SVM (Marunnan et al., 2018) and deep learning with long short-term memory neural networks (LSTM-NN) (Chakravarti & Alla, 2019) techniques are used to research the selected descriptors and predicting the anti-HIV activity.

Table 4. Comparison the proposed approach with other approach

Approach	*r*: coefficient of correlation	Number of selected descriptors
ANN	0.92	5
SVM	0.88	8
MLR-SVM	0.89	7
LSTM-NN	0.95	6
The approach: GA-ANFIS	0.98	3

The results are that are presented in Table 4 show that, the results of the proposed approach GA-ANFIS are superior compared with those of stepwise ANN, SVM, MLR-SVM and deep learning with LSTM-NN. The coefficient correlation is high and GA-ANFIS model uses fewer descriptors for predicting the pIC50 activity.

Conclusion

In this article, a new approach for the selection of molecular descriptors in QSAR modeling is presented. This new feature selection approach, called GA-ANFIS, was designed to address the problems of regression and classification.

The GA-ANFIS approach is organized in two phases. The first uses a genetic algorithm technique which identifies promising subsets of molecular descriptors following a linear regression technique. The second phase completes the first and makes it possible to confirm the subset of descriptors identified in a machine learning method: adaptive neural fuzzy inference.

The GA-ANFIS approach was illustrated and tested in a case study which constitutes an example of QSAR classification modeling, where the estimated property corresponds to the anti-HIV activity of the chemical compounds. Comparisons with the results obtained in the literature by other QSAR models were discussed, showing the potential and the usefulness of the proposed approach. The GA-ANFIS hybrid approach is an important idea for QSAR modeling, helping to minimize the time and money costs in the field of drug discovery.

Conflict of interest

The authors declare no potential conflict of interest regarding the publication of this work. In addition, the ethical issues including plagiarism, informed consent, misconduct, data fabrication and, or falsification, double publication and, or submission, and redundancy have been completely witnessed by the authors.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

Danishuddin, M., & Khan, A. U. (2016). Descriptors and their selection methods in QSAR analysis: paradigm for drug design. *Drug Discovery Therapy*, 21(8), 1291-1302.

Fenner, K., & Tratnyekc, P. G. (2017). QSARs and computational chemistry methods in environmental chemical sciences. *Environmental Science: Processes & Impacts,*19, 185-187.

Swathik, C. P., Jaspreet, K. D., Vidhi, M., Navaneethan, R., Mannu, J., & Durai S. (2019). Quantitative Structure-Activity Relationship (QSAR): Modeling Approaches to Biological Applications. *Encyclopedia of Bioinformatics and Computational Biology*, 2, 661-676.

Hdoufane, I., Stoycheva, J., Tadjer, A., Villemin, D., Najdoska-Bogdanov, M., Bogdanov, J., & Cherqaoui, D. (2019). QSAR and molecular docking studies of indole-based analogs as HIV-1 attachment inhibitors. *Journal of Molecular Structure*, 1193, 429-443.

Todeschini, R., & Consonni, V. (2009). *Molecular Descriptors for Chemoinformatics*, Wiley-VCH.

Eklund, M., Norinder, U., Boyer, S., & Carlsson, L. (2014). Choosing feature selection and learning algorithms in QSAR. *Journal of Chemical Information and Modeling*, 54(3), 837-843.

Grisoni, F., Consonni, V., & Todeschini, R. (2018). Impact of molecular descriptors on computational models. *In Computational Chemogenomics. Methods in Molecular Biology*. J. Brown, Ed., vol. 1825, Humana Press, New York, USA.

Liu, X. Y., Liang, Y., Wang, S., Yang, Z. Y., & Ye, H., S. (2018). Hybrid Genetic Algorithm With Wrapper-Embedded Approaches for Feature Selection. *IEEE Access*, 6, 22863-22874.

Wutzl, B., Leibnitz, K., Rattay, F., Kronbichler, M., Murata, M., & Golaszewski, S. M. (2019). Genetic algorithms for feature selection when classifying severe chronic disorders of consciousness. *PLoS ONE* 14(7), 1-16.

Nagasubramanian, K., Jones, S., Sarkar, S., Singh, A. K., Singh, A., & Ganapathysubramanian, B. (2018). Hyperspectral band selection using genetic algorithm and support vector machines for early identification of charcoal rot disease in soybean stems. *Plant Methods*, 14(86), 1-13.

Labjar, H., Kissi, M., Mouhibi, R., Khadir, O., Chaair, H., & Zahouily, M. (2016). QSAR study of 1-(3, 3-diphenylpropyl)-piperidinyl amides and ureas using genetic algorithms and artificial neural networks. *International Journal of Bioinformatics Research and Applications*,12(2), 116-128.

Salari, N., Shohaimi, S., Najafi, F., Nallappan, M., & Karishnarajah, I. (2014). A novel hybrid classification model of genetic algorithms, modified k-Nearest Neighbor and developed backpropagation neural network. *PLoS One*. 9(11), 1-50.

Srivastava, A. K., Singh, D., Pandey, A. S., & Maini, T. (2019). A Novel Feature Selection and Short-Term Price Forecasting Based on a Decision Tree (J48) Model. *Energies*, 12, 1-17.

Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning*, 46, 389-422.

Lal, T.N., Chapelle, O., Weston, J., & Elisseeff, A. (2006). Embedded Methods. *In: Guyon I., Nikravesh M., Gunn S., Zadeh L.A. (eds) Feature Extraction*. Studies in Fuzziness and Soft Computing, vol 207. Springer, Berlin, Heidelberg.

Swathik, C. P., Jaspreet, K. D., Vidhi, M., Navaneethan, R., Mannu, J., & Durai, S. (2019). Quantitative Structure-Activity Relationship (QSAR): Modeling Approaches to Biological Applications. *Encyclopedia of Bioinformatics and Computational Biology*, 2, 661-676

Liu, B., He, H., Luo, H., Zhang, T., & Jiang, J. (2019). Artificial intelligence and big data facilitated targeted drug discovery. *Stroke & Vascular Neurology*, 4, 206-213.

Racz, A., Bajusz, D., & Héberger, K. (2019). Intercorrelation Limits in Molecular Descriptor Preselection for QSAR/QSPR. *Molecular informatics*, 38, 1-6.

Pourbasheer, E., Aalizadeh, R., Ganjali, M. R., Norouzi, P., Shadmanesh, J. (2014). QSAR study of ACK1 inhibitors by genetic algorithm–multiple linear regression (GA–MLR). *Journal of Saudi Chemical Society* ,18, 681-688.

Holland, J. H. (1992). *Adaptation in Natural and Artificial Systems*. Ann Arbor, MI, university of Michigan Press.

Takagi, T., & Sugeno, M. (1985). Fuzzy identification of systems and its application to modelling and control. *IEEE Trans on Systems, Man and Cybernetics*, 15, 116-132.

Jang, J.S.R. (1993). ANFIS: Adaptive-Network-Based Fuzzy Inference systems. *IEEE Trans. Syst. Man Cybernet.*, 23 (3), 665–685.

Tanaka, H., Takashima, H., Ubasawa, M., Sekiya, K., Nitta, I., Baba, M., Shigata, S., Walker, R. T., De Clercq, E., Miyasaka, T. (1992). Structure-activity relationships of 1-[(2-hydroxyethoxy) methyl]-6-(phenylthio) thymine (HEPT) analogues: Effect of substitutions at the C-6 phenyl ring and the C-5 position on anti-HIV-1 activity. *J. Med. Chem*. 35, 337-345.

Garg, R., Gupta, S. P., Gao, H., Babu, M. S., & Debnath, A. K. (1999). Comparative Quantitative Structure-Activity Relationships Studies on Anti-HIV Drugs. *Chem. Rev*. 99, 3525-3601.

MMP, molecular modelling pro-Demo (TM) Revision 301 demo. ChemSW Software (TM). http://www.chemistry-software.com/modelling/molecular_modeling_pro_plus.htm

Buglak, A. A., Zherdev, A. V., Lei H. T., & Dzantiev, B. B. (2019). QSAR analysis of immune recognition for triazine herbicides based on immunoassay data for polyclonal and monoclonal antibodies. *PLoS ONE*, 14(4),1-19.

Anacleto de Souza, S., Leonardo Ferreira, L. G., Aldo de Oliveira, S., & Adriano Andricopulo, D. (2019). Quantitative Structure–Activity Relationships for Structurally Diverse Chemotypes Having Anti-Trypanosoma cruzi Activity. *Int. J. Mol. Sci*., 20, 1-21.

Wen, L., Li, Q., Li, W., Cai, Q., & Cai., Y. M. (2017). A QSAR Study Based on SVM for the Compound of Hydroxyl Benzoic Esters. *Bioinorganic Chemistry and Applications*, 1-10.

Marunnan, S. M., Pulikkal, B. P., Jabamalairaj, A., Bandaru, S., Yadav, M., Nayarisseri, A., & Doss, V. A. (2017). Development of MLR and SVM Aided QSAR Models to Identify Common SAR of GABA Uptake Herbal Inhibitors used in the Treatment of Schizophrenia. *Current Neuropharmacology*, 15(8), 1085-1092.

Chakravarti, S. K., & Alla S. R. M. (2019). Descriptor Free QSAR Modeling Using Deep Learning With Long Short-Term Memory Neural Networks. *Front. Artif. Intell*, 2(17), 1-18.

Bibliographic information of this paper for citing:

Labjar, Houda; Al-Sarem, Mohammad & Kissi, Mohamed (2022). Feature Selection Using a Genetic Algorithms and Fuzzy logic in Anti-Human Immunodeficiency Virus Prediction for Drug Discovery. *Journal of Information Technology Management*, Special Issue, 23-36.