Electronic Journal of Applied Statistical Analysis
EJASA, Electron. J. App. Stat. Anal.
http://siba-ese.unisalento.it/index.php/ejasa/index
e-ISSN: 2070-5948
DOI: 10.1285/i20705948v13n1p86

A QSAR classification model of skin sensitization potential based on improving binary crow search algorithm

By Algamal

Published: 02 May 2020

Electronic Journal of Applied Statistical Analysis
Vol. 13, Issue 01, May 2020, 86-95
DOI: 10.1285/i20705948v13n1p86

A QSAR classification model of skin sensitization potential based on improving binary crow search algorithm

Ghada Yousif Ismail Abdallh[a] and Zakariya Yahya Algamal*[b]

[a] *Departmet of Mechanical Technology, Technical Institute in Mosul, Northern Technical University*
[b] *Department of Statistics and Informatics, University of Mosul*

Published: 02 May 2020

Classifying of skin sensitization using the quantitative structure-activity relationship (QSAR) model is important. Applying descriptor selection is essential to improve the performance of the classification task. Recently, a binary crow search algorithm (BCSA) was proposed, which has been successfully applied to solve variable selection. In this work, a new time-varying transfer function is proposed to improve the exploration and exploitation capability of the BCSA in selecting the most relevant descriptors in QSAR classification model with high classification accuracy and short computing time. The results demonstrate that the proposed method is reliable and can reasonably separate the compounds according to sensitizers or non-sensitizers with high classification accuracy.

keywords: QSAR; crow search algorithm; skin sensitization; transfer function; descriptors selection.

1 Introduction

Skin-related illness such as occupational contact dermatitis, eczema and skin cancer made up fourteen percentages of total occupational illness as surveyed by the bureau of labor statistics of which 90-95% cases are due to occupational contact dermatitis (Burnett et al., 1998; Gunturi et al., 2010). The occupational contact dermatitis is an

*Corresponding author: zakariya.algamal@uomosul.edu.iq

ISSN: 2070-5948
http://siba-ese.unisalento.it/index.php/ejasa/index

important environmental and occupational health concern both in the United States and in Europe (Diepgen, 2003).

In chemometrics, "the quantitative structure-activity (property) relationship (QSAR/ QSPR) is a powerful and a promising model used to better understand the structural relationship between the chemical activity (property) and the chemical compounds by explicitly considering the mathematical, statistical, and informatical methods (Qasim et al., 2018). A common task in these models is the selection of relevant descriptors (variables), where researchers try to determine the smallest possible set of descriptors that can still achieve good predictive performance (Eklund et al., 2012; Algamal, 2019a,b; Algamal et al., 2017). A typical data in QSAR/ QSPR modeling consist of a small sample size of compounds (molecules) and a very large number of descriptors. Consequently, QSAR/ QSPR modeling is challenged by the high dimensionality of the descriptors.

In chemometrics, today, it easily comes out with thousands of molecular descriptors, such as Dragon 7, which is a commercial software. It can calculate 5270 molecular descriptors. In high dimensional QSAR/ QSPR modeling, where the number of descriptors, exceeds the number of compounds, , the traditional statistical classification methods are not feasible. In addition, the large number of descriptors can degrade the generalizable performance of the used classifier or the prediction performance. Therefore, selecting descriptors that truly affect the biological activity is an attractive way in QSAR/ QSPR modeling .

Variable (Descriptor) selections can be reported as a non-polynomial (NP) hard problem. The objective of variable selection is to provide faster and more effective models, and also to avoid overfitting and the curse of dimensionality. Variable selection is a typical combinatorial optimization problem. A considerable effort has been devoted to developing variable selection procedures. With the development of computational intelligence, evolutionary algorithms, such as particle swarm optimization (PSO), bat algorithm (BA), and grey wolf optimization (GWO), are the most effective and core technology to address high-dimensional data.

The crow search algorithm (CSA), which was proposed by Askarzadeh (2016), has certain outstanding merits, such as a simple computational process, simple implementation, and easy understanding with only a few parameters for tuning. Due to its good properties, CSA has become a useful tool for many real-world problems (Abdelaziz and Fathy, 2017; Allaoui et al., 2018; Anter et al., 2019; Gupta et al., 2018a,b; Hassanien et al., 2018; Horng and Lin, 2017; Liu et al., 2017; Mohammadi and Abdi, 2018; Rizk-Allah et al., 2018,?). The CSA is inspired by the social behavior of the crow. In the case of variable selection, the search space is modeled as an n-dimensional Boolean lattice, in which the selected variable is coded as 1 and the not selected variable is coded as 0. Therefore, a binary version of the CSA was proposed. The efficiency of the binary crow search algorithm (BCSA) is depending on the transfer function which is responsible to map a continuous search space to a discrete search space.

In this study, a new time-varying transfer function is proposed to improve the exploration and exploitation capability of the BCSA in selecting the most relevant descriptors in QSAR classification model with high classification accuracy and short computing time of identifying whether a compound is a sensitizers or non-sensitizers.

The rest of the paper is organized as follows: The explanation of the crow search algorithm and the proposed time-varying transfer function are given in Section 2. In section 3, the experimental setting is covered. Section 4, the results are summarized with their discussion. Finally, Section 5 contains a conclusion of this work.

2 Methodology

2.1 Crow search algorithm

The crow search algorithm is one of the most recent evolutionary algorithms inspired from the social behavior of the crow. This algorithm was introduced in 2016 by Askarzadeh (2016). In CSA, the idea is motivated from the storing process of the excess food in hiding places then restoring it in the necessary time. It is known that the crow is very intelligent bird that observes the others hide their food and steal it once they leave. After committing the theft, it hides to avoid being a victim in the future. It is assumed that a flock of n_c crows, the crow number i has position at iteration t is x_i^t. The hiding place of the food followed by crow i is memorized. Crow moves in the search plane and tries to find the best food source which is defined as M_i^t. The searching approach in CSA has two probable scenarios; the first one is that the owner crow $j$of food source M_j^t does not know the thief crow follows it therefore the thief crow reaches to the hide place of owner crow. The updating process of the crow thief position is done by

$$x_i^{t+1} = x_i^t + \tau \times fl \times (M_j^t - x_i^t), \quad i = 1, 2, ...n_c, \tag{1}$$

where fl is the flight length and $\tau.$ is a random number in the interval $[0, 1]$.

The second scenario is that the owner crow j knows that the thief crow $i$follows it therefore, the owner crow will deceive crow $i$by going to any another position of search space. The position of crow $i.$ is updated by a random position. In CSA, the scenario is determined by the following expression:

$$x_i^{t+1} = \begin{cases} x_i^t + \tau \times fl \times (M_j^t - x_i^t), & \textbf{if } \theta \geq \textbf{AP} \\ \textbf{random position}, & \textbf{otherwise} \end{cases} \tag{2}$$

where θ a random number in the interval $[0, 1]$ and AP is the probability of awareness.

To perform the variable selection, a binary crow search algorithm was proposed Sayed et al. (2017). Unlike the standard CSA, in which the solutions are updated in the search space towards continuous-valued positions, in the BCSA, the search space is modeled as an n-dimensional Boolean lattice and the solutions are updated across the corners of a hypercube. In addition, as the problem is to select or not a given variable, a solution binary vector is employed, where 1 corresponds whether a variable will be selected to compose the new dataset, and 0 otherwise. In any binary algorithm, where one uses the step vector to calculate the probability of changing positions, the transfer functions significantly impact the balance between exploration and exploitation (Islam et al., 2017; Mafarja et al., 2018).

2.2 The proposed time-varying transfer function

In BCSA, the transfer function is used to map a continuous search space to a binary one, and the updating process is designed to switch positions of stars between 0 and 1 in binary search spaces. In order to build this binary vector, a transfer function in Eq. (4) can be used, in which the new solution is constrained to only binary values

$$x_i^t = \begin{cases} 1 & \textbf{if} \quad T(x) > \alpha \\ 0 & \textbf{otherwise} \end{cases} \tag{3}$$

where $\alpha \in [0,1]$ is a random number, $T(x)$ is the transfer function. This transfer function is defined as:

$$T_{BCSA}\left(x_i^t\right) = \frac{1}{1+e^{10(x_i^t-0.5)}}, \tag{4}$$

In optimization algorithm, it is expected that the focus of the early stages of the implementation the algorithm will be on exploration to avoid falling into the local point, but in later stages of implementation, the algorithm focuses more on exploitation to improve the quality of the solution (Islam et al., 2017; Mafarja et al., 2018). As in Mafarja et al. (2018) and Islam et al. (2017), in this paper, a dynamic transfer function is proposed to improve the BCSA. In our proposed time-varying transfer function, (TV), a new control parameter β is added in the original transfer function. This β is a time-varying variable which starts with a large value and gradually decreases over time. The proposed β is defined as

$$\beta = \beta_{\min} + (\beta_{\max} - \beta_{\min})e^{-t}, \tag{5}$$

where $\beta_{\max}$ and $\beta_{\min}$ are, respectively, the minimum and maximum values of the control parameter β. Accordingly, the proposed transfer function is defined as

$$T_{\mathbf{TV}}\left(x_i^t\right) = \frac{1}{1+e^{10(x_i^t-0.5)/\beta}}, \tag{6}$$

3 Experimental setting

3.1 Dataset

A data set of 255 compounds classified as sensitizers or non-sensitizers was collected from Gunturi et al. (2010). As in Gunturi et al. (2010), these compounds were divided into three parts: a training set with 184 compounds, containing 90 sensitizers and 94 non-sensitizers, a test set with 45 compounds, containing 21 sensitizers and 24 non-sensitizers, and an external data set used with 26 compounds containing 12 sensitizers and 14 non-sensitizers.

Dragon software (version 6.0) was used to generate the molecular descriptors. To include consistent and useful descriptors, preprocessing steps were carried out as follows: First, those that had zero values for all molecules were discarded. Second, those that had a constant value for all molecules were excluded from the study. Then, descriptors in which 95% of their values were zeros were removed. And, finally, descriptors with a relative standard deviation of less than 0.001 were removed.

3.2 BCSA parameters initialization

There are six control parameters in our proposed time-varying transfer function: The number of crows (n_c) (Population size), the maximum number of iterations ($\mathbf{t}_{\max}$), the flight length fl, the probability of awareness (AP), and the minimum and maximum values of the control parameter β of Eq. (6). The specific parameter values are outlined in Table 1. The position for each crow is a vector of 0 and 1 values with size equals the number of the descriptors. Initially, the positions were randomly generated from a uniform distribution between 0 and 1. Further, the best fitness function that can combine the maximum classification performance and the minimum number of selected descriptors is preferable. The fitness function used in BCSA to evaluate each crow position is defined as

$$\textbf{fitness} = 0.9 \times \textbf{CA} + 0.1 \times \left(\frac{d - \tilde{d}}{d} \right), \tag{7}$$

where CA is the classification accuracy obtained, d represents the number of descriptors in the dataset, and $\tilde{d}$ represents the number of selected descriptors.

Table 1: Parameter setting for BCSA

Parameter	Value
n_c	25
$\mathbf{t}_{\max}$	500
fl	1.8
AP	0.1
$\beta_{\max}$ in Eq. (6)	2.5
$\beta_{\min}$ in Eq. (6)	0.1

4 Results and discussion

With the aim of correctly assessing the performance of our proposed time-varying transfer function, $T_{\mathbf{TV}}$, comparative experiments with the original $T_{\mathbf{BCSA}}$, and with the SVM were carried out. In this study, the experiments were carried out using a support vector machine (SVM) with a Gaussian radial basis function (RBF) kernel function. The classification accuracy ($\mathbf{CA} = (\mathbf{TP} + \mathbf{TN})/(\mathbf{TP} + \mathbf{FP} + \mathbf{FN} + \mathbf{TN}) \times 100\%$) and the number of selected descriptors are reported in Table 2. It can be seen in Table 2 that, from among the three methods, the $T_{\mathbf{TV}}$ function performs the best with results of 98.16%, 95.00%, and 97.14%, in terms of classification accuracy, for training, testing, and external datasets, respectively. That is mean that $T_{\mathbf{TV}}$ is reliable and can reasonably separate the compounds according to sensitizers or non-sensitizers. Further, as it can be observed from Table 2, $T_{\mathbf{TV}}$ overtakes the standard $T_{\mathbf{BCSA}}$. The most remarkable result for $T_{\mathbf{TV}}$ is that it obtained 97.14% accuracy with 4 descriptors for external dataset.

Moreover, Table 2 demonstrates that the $T_{\mathbf{TV}}$ is significantly better than T_{BCSA} in terms of the number of selected descriptors. $T_{\mathbf{TV}}$ selects descriptors approximately 2 times fewer than $T_{\mathbf{BCSA}}$. The names of the selected descriptors and their descriptions for each used transfer function are presented in Table 3. The selected descriptors are came in agreement with Gunturi et al. (2010) indicating that the selected descriptors have related to the skin sensitization potential profile.

Table 2: Classification performance of the proposed time-varying transfer function

Method	Training dataset		Testing dataset	External dataset
	# selected descriptors	CA	CA	CA
All	88.04%	84.45%	80.76%	80.57%
T_{BCSA}	9	91.84%	88.89%	88.46%
T_{TV}	4	96.73%	93.33%	92.30%

Table 3: The selected descriptor names and their descriptions by the $T_{\mathbf{TV}}$

Descriptor name	Group type	Description
SsCH3	Atom-type E-state indices	Sum of sCH3 E-states
SdO	Atom-type E-state indices	Sum of dO E-states
P-VSA-LogP-9	P-VSA-like descriptors	P-VSA-like on LogP, bin 9
SdssC	Atom-type E-state indices	Sum of dssC E-states

To further highlight the efficiency of the proposed time-varying transfer function, Figure 1 displays the execution time in seconds. The computational efficiency of $T_{\mathbf{TV}}$ (102.54 sec.) is comparable to $T_{\mathbf{BCSA}}$ (139.66 sec.). It is noteworthy that $T_{\mathbf{TV}}$ has the faster convergence speed beating the $T_{\mathbf{BCSA}}$, where it requires the least amount of time to complete the optimized target.

TV	101.23
BCSA	137.64

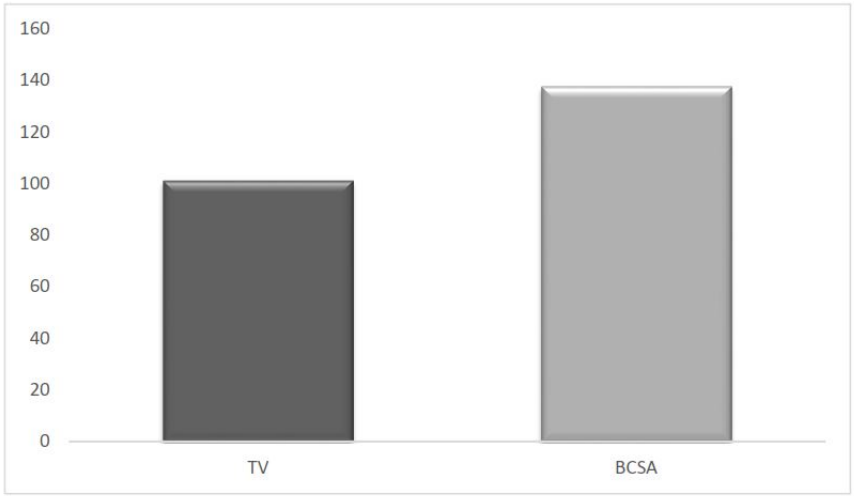

Figure 1: CPU time, in seconds, of the proposed time-varying transfer function.

To farther show the advantage of our proposed method, the 229 compounds are randomly divided in 70% of samples as a training dataset and the remaining 30% of the samples are used as a testing dataset. Then the same 26 compounds are used as an external validation set. This partition repeated 50 times independently and the average area under the curve (AUC) values are reported in Table 4. Based on the results in Table 4, $T_{\mathbf{TV}}$ achieves the best and most stable classification results. Compared with the $T_{\mathbf{BCSA}}$, the $T_{\mathbf{TV}}$ achieved an 5.49% performance improvement for the external validation dataset.

Table 4: Averaged AUC values of the proposed time-varying transfer function

Method	Training dataset	Testing dataset	External dataset
	AUC	AUC	AUC
SVM	88.73 ± 0.118	83.12 ± 0.121	80.57 ± 0.121
$T_{\mathbf{BCSA}}$	91.12 ± 0.113	87.21 ± 0.119	87.51 ± 0.118
$T_{\mathbf{TV}}$	96.37 ± 0.111	93.11 ± 0.117	92.60 ± 0.116

5 Conclusion

In this work, a classification of skin sensitization using the QSAR model was proposed, in which a new time-varying transfer function was proposed to improve the exploration and exploitation capability of the binary crow search algorithm. The results gained by the classification accuracy for the training dataset, the testing dataset, and the external validation dataset proved the better predictive power of the QSAR model compared with other two methods in separating the compounds according to sensitizers or non-sensitizers. Further, the proposed method selected fewer descriptors than the others".

6 Acknowledgment

The author is very grateful to the University of Mosul/ College of Computer Sciences and Mathematics for their provided facilities, which helped to improve the quality of this work.

References

Abdelaziz, A. Y. and Fathy, A. (2017). A novel approach based on crow search algorithm for optimal selection of conductor size in radial distribution networks. *Engineering Science and Technology, an International Journal*, 20(2):391–402.

Algamal, Z. (2019a). A particle swarm optimization method for variable selection in beta regression model. *Electronic Journal of Applied Statistical Analysis*, 12:508–519.

Algamal, Z. (2019b). Variable selection in count data regression model based on firefly algorithm. *Statistics, optimization and information computing*, 7:520–529.

Algamal, Z. Y., Qasim, M. K., and Ali, H. T. (2017). A qsar classification model for neuraminidase inhibitors of influenza a viruses (h1n1) based on weighted penalized support vector machine. *SAR and QSAR in Environmental Research*, 28:415–426.

Allaoui, M., Ahiod, B., and El Yafrani, M. (2018). A hybrid crow search algorithm for solving the dna fragment assembly problem. *Expert Systems with Applications*, 102:44–56.

Anter, A. M., Hassenian, A. E., and Oliva, D. (2019). An improved fast fuzzy c-means using crow search optimization algorithm for crop identification in agricultural. *Expert Systems with Applications*, 118:340–354.

Askarzadeh, A. (2016). A novel metaheuristic method for solving constrained engineering optimization problems: Crow search algorithm. *Computers & Structures*, 169:1–12.

Burnett, C. A., Lushniak, B. D., McCarthy, W., and Kaufman, J. (1998). Occupational dermatitis causing days away from work in us private industry, 1993. *American journal of industrial medicine*, 34(6):568–573.

Diepgen, T. L. (2003). Occupational skin-disease data in europe. *International archives of occupational and environmental health*, 76(5):331–338.

Eklund, M., Norinder, U., Boyer, S., and Carlsson, L. (2012). Benchmarking variable selection in qsar. *Mol Inform*, 31(2):173–9.

Gunturi, S. B., Theerthala, S. S., Patel, N. K., Bahl, J., and Narayanan, R. (2010). Prediction of skin sensitization potential using d-optimal design and ga-knn classification methods. *SAR QSAR Environ Res*, 21(3-4):305–35.

Gupta, D., Rodrigues, J. J. P. C., Sundaram, S., Khanna, A., Korotaev, V., and de Albuquerque, V. H. C. (2018a). Usability feature extraction using modified crow search algorithm: a novel approach. *Neural Computing and Applications*.

Gupta, D., Sundaram, S., Khanna, A., Ella Hassanien, A., and de Albuquerque, V. H. C. (2018b). Improved diagnosis of parkinson's disease using optimized crow search algorithm. *Computers & Electrical Engineering*, 68:412–424.

Hassanien, A. E., Rizk-Allah, R. M., and Elhoseny, M. (2018). A hybrid crow search algorithm based on rough searching scheme for solving engineering optimization problems. *Journal of Ambient Intelligence and Humanized Computing*.

Horng, S.-C. and Lin, S.-S. (2017). Merging crow search into ordinal optimization for solving equality constrained simulation optimization problems. *Journal of Computa-*

tional Science, 23:44–57.

Islam, M. J., Li, X., and Mei, Y. (2017). A time-varying transfer function for balancing the exploration and exploitation ability of a binary pso. *Applied Soft Computing*, 59:182–196.

Liu, D., Liu, C., Fu, Q., Li, T., Imran, K. M., Cui, S., and Abrar, F. M. (2017). Elm evaluation model of regional groundwater quality based on the crow search algorithm. *Ecological Indicators*, 81:302–314.

Mafarja, M., Aljarah, I., Heidari, A. A., Faris, H., Fournier-Viger, P., Li, X., and Mirjalili, S. (2018). Binary dragonfly optimization for feature selection using time-varying transfer functions. *Knowledge-Based Systems*, 161:185–204.

Mohammadi, F. and Abdi, H. (2018). A modified crow search algorithm (mcsa) for solving economic load dispatch problem. *Applied Soft Computing*, 71:51–65.

Qasim, M. K., Algamal, Z. Y., and Ali, H. M. (2018). A binary qsar model for classifying neuraminidase inhibitors of influenza a viruses (h1n1) using the combined minimum redundancy maximum relevancy criterion with the sparse support vector machine. *SAR and QSAR in Environmental Research*, 29:517–527.

Rizk-Allah, R. M., Hassanien, A. E., and Bhattacharyya, S. (2018). Chaotic crow search algorithm for fractional optimization problems. *Applied Soft Computing*, 71:1161–1175.

Sayed, G. I., Hassanien, A. E., and Azar, A. T. (2017). Feature selection via a novel chaotic crow search algorithm. *Neural Computing and Applications*.