# A Novel firefly algorithm of solving feature selection of infrared spectrum

# Zhonghong Xi<sup>1,a</sup>

# <sup>1</sup>Gansu Normal University for Nationalities, Hezuo, gansu, 747000, China <sup>a</sup>zhonghongxi@163.com

Keywords: Firefly Algorithm; feature selection; infrared spectrum; Global convergence

**Abstract:** Feature selection is an important part during the process of qualitative and quantitative analysis of infrared spectrum. In order to solve the disadvantage of traditional methods, such as multi-parameters, slow convergence, poor accuracy, prone to premature, etc. a novel feature selection firefly algorithm was proposed, which combined the basic firely algorithm and Lévy flights search strategy. The algorithm was experimented and the experimental results show that the novel algorithm is better suited to handle feature selection of infrared spectrum. The feasibility and effectiveness of the method are demonstrated.

# **1. Introduction**

Infrared spectral analysis is a new analytical technique, Its characteristics is analysis speed, high efficiency, low cost and easy to implement on-line analysis, it has widely application in the petroleum, chemical industry, agriculture, medicine, tobacco, food, environmental monitoring, aerospace and other industries <sup>[1-3]</sup>. Because of the wavelength, the wavelength between serious collinearity and other characteristics, in the use of full spectrum calibration models were established for the former, usually need to operate wavelength selection with spectral information, that is, through the proper selection of spectral, or transform, excluding no variable information and noise, so as to reduce the prediction error and correction model to enhance the reliability.

At present, the wavelength selection methods are mainly based on the physical meaning of the spectral selection  $(\text{KBS})^{[4]}$ , continuous projection algorithm  $(\text{SPA})^{[5]}$ , no information variable elimination  $(\text{UVE})^{[6]}$ , simulated annealing (SA) algorithm, genetic algorithm  $(\text{GA})^{[7]}$ , artificial neural network (ANN), interval partial least squares (iPLS)<sup>[8]</sup>.

# 2. The basic principle of the firefly algorithm

# 2.1The bionics principle of algorithm

The firefly algorithm is the development of simulating biological characteristics of 1 the adults firefly, but the algorithm abandons some biological significance of luminous firefly, the firefly searches partners according to the search area only using the emission properties, and the firefly moves to position better firefly in the neighborhood structure, so as to realize the position evolution. In this algorithm, the firefly attracts one another reason depends on two factors, namely, its brightness and attraction. Among them, the firefly fluorescence brightness depends on the location of the target value, the brightness higher the location of the target value the better. The more light of fireflies has more attractive, it can attract brightness than its weak firefly in sight range to the direction of movement. If the fireflies have the same brightness, then they randomly move. The brightness and attraction are inversely proportional to the distance of the fireflies, they have increased as the distance decreases, which is equivalent to the simulation of the fluorescence characteristics of propagation in the space by the media absorption and fading. The firefly algorithm is adopted to simulate the firefly group behavior to construct a class of stochastic optimization algorithm. The bionic principle is: The points in the search space simulate the nature of firefly individuals, the search and optimization process simulation to attract and movement of the firefly individuals, the objective function of solving the problem is measured into the pros and cons of the location of the individual, the fittest of individual process for search and optimization process of a good analogy of feasible solutions to replace less feasible solutions.

### 2.2 Mathematical description and analysis of the algorithm

As mentioned above, the firefly algorithm includes two factors: the brightness and the degree of attraction. The brightness of the firefly reflects the pros and cons of location and determines its direction of movement, attraction degree determines the distance of the firefly mobile, constantly update the brightness and the attraction degree, so as to realize the goal of optimization. From a mathematical perspective describes the Firefly algorithm optimization mechanism, the description as follows<sup>[9,10]</sup>.

(1) The relative fluorescence intensity of fireflies as:

$$h = \frac{h_0}{(1 + zr_{ij}^2)}$$
(1)

Among them:  $h_0$  for the maximum fluorescence intensity of fireflies, that is itself (*r*=0) fluorescence brightness, associated with the value of the objective function, the objective function value is better, its brightness is higher; *z* as the light absorption coefficient, because the fluorescence decreases with the increasing of distance and media absorption <sup>[11]</sup>, so set the light absorption

coefficient to reflect this characteristic, it can set to constant;  $r_{ij}$  is the space distance of the firefly *i* and the firefly *j*.

(2) The firefly attraction degree is:

$$\rho = \rho_0 * e^{-zr_{ij}^2} \tag{2}$$

Among them:  $\rho_0$  as the biggest attraction, namely the light source (r=0) attraction degree; z,

 $r_{ij}$  meaning as above.

(3) The firefly *i* is attracted to move to the firefly *j*, the location update by formula (3) decision.  $x_i = x_i + \rho^*(x_j - x_i) + \alpha^*(rand - 1/2)$ (3)

Among them:  $x_i^{'}$  is the individual i towards a brighter individual j update location,  $x_i, x_j$  are the firefly *i* and the firefly *j* located before the entire population renewal;  $\alpha$  is the step factor, it is constant between 0 and 1, *rand* is a random factor between 0 and 1, it obeys uniform distribution.

The process of optimization algorithm is: Firstly, the firefly populations randomly scattered in the solution space, each firefly has different fluorescence brightness at different positions, by comparing Eq.1, high brightness fireflies can attract low brightness fireflies to move, moving distance depending on the size of the attraction (according to Eq.2). In order to increase the search area, avoid getting into local optimization, and updated measures, according to the formula 3 to calculate the updated position. So by repeatedly movement, all individuals will be gathered in the highest brightness firefly position, so as to realize the optimization.

At present, it has been found that many insects existing Lévy flight <sup>[12]</sup>, and Lévy flight has been used in the field of optimization, and achieve the expected effect. Simulation track of Lévy flights is shown in figure 1.In order to enhance the algorithm global search performance, avoid the population into a local optimum in the search process, in the firefly algorithm, if the individual is no better than their individual, choose to Lévy flight instead of random flight in the original algorithm. In addition, the non - optimal those individuals in a population, the flight formula was improved: when they find more bright than their individual, first generates a random number q by the system, if q is less than 0.5, the formula (4) is updated; otherwise, still use the formula (3) to update the individual position.

$$x_{i}^{*} = x_{i}^{*} + \rho^{*}(x_{j} - x_{i}) + \alpha^{*}(rand - 1/2)$$
(4)

Among them,:  $x_j$  still expresses renewal position of individual *j* before the entire population,  $x_i$  expresses the individual i toward the front of *j*-1 individuals than their bright individual after the

update new position,  $x_i^{'}$  expresses  $x_i^{'}$  toward than their bright individual *j* after update location,  $\rho^{'}$  expresses that individual *j* appeal to the individual *i*. As can be seen, formula (4) is updated in real time, formula (3) depends only on the entire population before moving. This flight update can increase the randomness of flight, It is helpful to keep the diversity of population, increasing the population search space.

In order to accelerate the convergence of the population, the paper proposes a method of  $\alpha$  updating, which  $\alpha$  gradually decreases with increasing number of iterations. Update formula is as follows:

$$\alpha = \alpha_0 - e^{-0.001*t} \tag{5}$$

In the formula (5):  $\alpha_0$  is 0.9, t is number of iterations.

The concrete steps of the algorithm are as follows:

1) Initialization the basic parameter of algorithm. Set the number of fireflies is *s*, the biggest attraction is  $\rho_0$ , light absorption coefficient is *z*, the random parameters is  $\alpha_0$ , the maximum number of iterations is  $t_{\text{max}}$ =1000, the evolving algebra *t*=0.

2) Randomly initialization firefly position, calculate the objective function value of firefly as the respective maximum fluorescence intensity of  $h_0$ .

3) By formula 1 and formula 2 to calculate the relative brightness of h and attraction  $\rho$ , according to relative brightness of h to decide the movement direction of the firefly.

4) The formula 5 is used to update  $\alpha$ , the individual, if there is more lighter than its individual, in accordance with the update is improved, by formula (4) to update the individual position; otherwise, the Levy flight is used to update the individual location.

5) According to the updated firefly position, to recalculate brightness of the firefly.

6) The system generates a random number, if the random number is less than the local search probability *p*, then local search of individuals of the population, and regenerate population.

7) Determine whether meet the conditions of termination of the algorithm. Such as the maximum number of iterations  $t_{\text{max}}$  is 1000 or best solution stagnation does not change, turn to step 7, or t = t+1, turn to step 3.

8) Output global extreme value point and optimal individual value.



Fig.1 Simulation track of Lévy flights

#### 3. The analysis of simulation

In order to verify the validity and superiority of the method, 3 data sets with different phase states and different spectral ranges were used. All data sets are divided into training set, test set and validation set. The training set is used to establish the calibration model, and the validation set is used to calculate RMSE and feature selection, and the test set is used to evaluate the method.

The data set is as follows:

The near infrared spectra of maize and the content of vegetable oil. Spectrum acquisition range is  $1100 \sim 2498 \text{ } nm$ , step is 2 nm. Training set is 50 samples, test set is 15 samples and the validation set is 15 samples. The data set can be downloaded from Eigenvector company's website.

The experimental result of data set is as follows.

In the whole spectral range of the corn data set, a total of 700 wavelength points are included. The whole spectral range is divided into 70 sub regions, each of which contains 10 wavelengths. The basic algorithm initial parameters setting are as follows: bat populations m=30, the maximum number of iterations n = 50, pulse frequency range  $[f_{min}, f_{max}]=[0, 2]$ , the pulse intensity A=0.9, pulse frequency r=0.9, Lévy distribution scale parameter lamb  $\lambda = 1.5$ .

The characteristics of the first interval combination of the wave of the new algorithm is shown in Figure 2, a total of 9 sub range of and a total of 90 wavelengths. The wavelength of MC-VS, GA and UVE-PLS are 144, 134 and 291 in the Reference [1]. The search results of the algorithm cover the main absorption bands. Screening and optimizing of prediction and MC-VS C-H absorption telescopic 800 secondary C-H contrast can be found, such as 1 200 nm double frequency absorption area, 1 600-1800 *nm* C-H absorption telescopic , 2 000 *nm*, 2 300 *nm* frequency absorption area and 1 430 *nm* O-H stretching vibration of a double frequency absorption area, and further reduce the wavelength number, using the results established vegetable oil elm correction models is showed in Figure 3, the test set validation RMSEP is 0.121 6, the new algorithm ensures that the model prediction accuracy and reduce the model complexity.



Fig.3 Corn dataset' s predictive error of test set

#### 4. Analysis and discussion

#### 4.1 algorithm convergence rate analysis

Because the Reference [1] did not give the detailed configuration of the simulation platform, the new algorithm proposed in this paper can not be compared carefully. The author has written the MC-VS, GA and MATLAB SA procedures, and in the Windows 64 operating system, GHz CPU3.00, 4 RAM GB on the computer and LBA for comparison. Taking the corn data set as an example, the iterative evolution curve of a single run is shown in Figure 4.



Fig.4 Comparision of iterative evolution process

It can be directly seen from the graph, the speed of MC-VS is fastest, only 4 iterations (0.3 s) can be convergence, this algorithm of the paper speed times and need to 19 iterations (120 s) convergence is reached, GA need 28 times iteration (3.4 s) to converge, SA has the slowest convergence rate, it requires an iterative 35 (5.1 s). This is because, and compared to heuristic search algorithm MC-VS algorithm using Monte Carlo stochastic simulation, it has certain advantages in speed; the population size of the genetic algorithm due to the large slow speed of convergence, and simulated annealing cooling process is too long. At the same time, we can find that this algorithm of the paper is significantly better than MC-VS, GA and SA by comparing the optimal fitness function, and it can be concluded that LBA is an effective search method.

#### 5. Conclusions

Based on the basic firefly algorithm, this paper makes full use of the characteristics of the Lévy flight search strategy in the whole solution space, and puts forward a new method of infrared spectrum feature selection. The validation results show that, this method can find the global optimal feature combinations of wavelengths. Compared with the traditional algorithm, the algorithm has fast convergence and global optimization in excellent advantages, with a wide range of applications and practical value.

#### Acknowledgment

This work is supported by the Twelfth Five Year Plan of Educational Science in Gansu Province research project [2012](GSGXG039).

#### References

[1] Hong Mingjian, Wen Quan, Wen Zhiyu. New near infrared wavelength selection algorithm based on Monte-Carlo method[J]. Acta Optica Sinica, 2010, 30(12): 3637-3642. (in Chinese)

[2] Qin Yuhua, Ding Xiangqian, Gong Huili. High dimensional feature selection in near infrared spectroscopy classification[J]. Infrared and Laser Engineering, 2013, 42(5): 1355-1359.

[3] Guo Tengxiao, Ding Xuequan, Dong Xiaoqiang, et al. New method of preprocessing IR remote sensing spectrum signals based on EMD [J]. Infrared and Laser Engineering, 2013, 42(12): 3196-3200.

[4] Zou Xiaobo, Zhao Jiewen, Povey M J W, et al. Variables selection methods in near-infrared spectroscopy [J]. Analytica Chimica Acta, 2010, 667: 14-32.

[5] Liu Fei, Jiang Yihong, He Yong. Variable selection in visible/near infrared spectra for linear and nonlinear calibrations: a case study to determine soluble solids content of beer [J]. Analytica Chimica Acta, 2009, 635(1): 45-52.

[6] Cai Wensheng, Li Yankun, Shao Xueguang. A variable selection method based on uninformative variable elimination for multivariate calibration of near-infrared spectra [J]. Chemometrics and Intelligent Laboratory Systems, 2008, 90(2): 188-194.

[7] Givianrad M H, Saber-Tehrani M, Zarin S. Genetic algorithmbased wavelength selection in multicomponent spectrophotometric determinations by partial least square regression: application to a sulfamethoxazole and trimethoprim mixture in bovine milk [J]. Journal of The Serbian Chemical Society, 2013, 78(4): 555-564.

[8] Cramer J A, Kramer K E, Johnson K J, et al. Automated wavelength selection for spectroscopic fuel models by symmetrically contracting repeated unmoving window partial least squares [J]. Chemometrics and Intelligent Laboratory Systems, 2008, 92: 13-21.

[9] Yang Xinshe. Firefly algorithms for multimodal optimization [C] / /Proc of the 5th International Symposium on Stochastic Algorithms: Foundations and Applications. 2009: 169-178.

[10] Yang Xinshe, DEB S. Eagle strategy using lévy walk and firefly algorithms for stochasticoptimization [J]. Studies in Computational Intelligence, 2010, 284: 101-111.

[11] Yang Xinshe. Firefly algorithms for multimodal optimization[C]//Proceedings of the 5th International Conference on Stochastic Algorithms: Foundations and Applications. Berlin/Heidelberg, Germany: Springer-Verlag, 2009: 169-178.

[12] Brown C T, Liebovitch L S, Glendon R. Lévy flights in Dobe Ju' hoansi foraging patterns[J]. Human Ecology, 2007, 35(1): 129-138.