IEICE TRANS. FUNDAMENTALS, VOL.E105–A, NO.5 MAY 2022

LETTER

Feature Selection and Parameter Optimization of Support Vector Machines Based on a Local Search Based Firefly Algorithm for Classification of Formulas in Traditional Chinese Medicine

Wen SHI[†a)], ***Member***, **Jianling LIU**[†], **Jingyu ZHANG**[††], **Yuran MEN**[†], **Hongwei CHEN**[†], **Deke WANG**[†], ***and*** **Yang CAO**[†], ***Nonmembers***

SUMMARY Syndrome is a crucial principle of Traditional Chinese Medicine. Formula classification is an effective approach to discover herb combinations for the clinical treatment of syndromes. In this study, a local search based firefly algorithm (LSFA) for parameter optimization and feature selection of support vector machines (SVMs) for formula classification is proposed. Parameters C and γ of SVMs are optimized by LSFA. Meanwhile, the effectiveness of herbs in formula classification is adopted as a feature. LSFA searches for well-performing subsets of features to maximize classification accuracy. In LSFA, a local search of fireflies is developed to improve FA. Simulations demonstrate that the proposed LSFA-SVM algorithm outperforms other classification algorithms on different datasets. Parameters C and γ and the features are optimized by LSFA to obtain better classification performance. The performance of FA is enhanced by the proposed local search mechanism.

key words: *classification, firefly algorithm, feature selection, parameter optimization*

1. Introduction

Traditional Chinese medicine (TCM) has played an important role in health care in China for 2000 years. In TCM, the four aspects of the methods of diagnosis, called the four pillars, are observation (inspection), auscultation and olfaction, palpation, and inquiry. The four pillars are applied for syndrome differentiation (*Bian Zheng*). Syndrome is an important concept in TCM. It is a combination of signs and symptoms with internal relationships. In TCM, formulas can be classified based on the corresponding syndromes in a certain disease. Common herb combinations in the formulas in the same classes can be used for the clinical treatment of syndromes. In this paper, we focus on the formula classification problem in TCM.

There are many different types of classification algorithms, such as k nearest neighbour [1], decision tree (DT) [2], artificial neural network [3] and support vector machine (SVM) [4]. The SVM is a supervised method used for classification. It was first suggested by Vapnik in 1995 [5]. The SVM creates a hyperplane that separates the data into classes. To create the hyperplane, the SVM uses a kernel function to take the dimensional input space and transforms it into a better dimensional space for separation. As described in [6], optimal feature selection and optimal parameters are two crucial problems for SVMs. In [6], a genetic algorithm (GA) was presented to simultaneously optimize the parameters and feature subset for SVMs. The optimized parameters include the penalty parameter C and the kernel function parameter γ for the radial basis function kernel. Lin et al. [7] developed a particle swarm optimization (PSO) based approach for parameter determination and feature selection of the SVM.

The firefly algorithm (FA) [8] is a swarm intelligence-based algorithm for solving optimization problems effectively. FA uses three rules: (1) all fireflies are unisex so that one firefly is attracted to other fireflies regardless of their sex; (2) attractiveness is proportional to their brightness; thus, for any two flashing fireflies, the dimmer firefly will move towards the brighter one. The attractiveness is proportional to the brightness, and they both decrease as their distance increases. If no firefly is brighter than a particular firefly, it moves randomly; (3) the brightness or light intensity of a firefly is affected or determined by the landscape of the objective function to be optimized. FA is used to optimise the parameters [9] and feature subset [10] for SVM.

Information exchange between the top fireflies, or the optimal solutions during the process of the light intensity updating was developed in [11]. The Levy distribution was applied in FA instead of traditional uniform distribution to strengthen the exploration in global space in [12] and [13]. Xu et al. [14] combined the binary firefly algorithm with opposition-based learning to select features in classification.

In this paper, a novel approach is proposed for feature selection and parameter optimization of SVM based on an improved FA for formula classification of TCM. The effectiveness value of herbs is adopted as the feature for formula classification. The parameters optimized are C and γ. After the spatial coordinates of fireflies are designed, an improved FA algorithm based on local search (LSFA) is developed. In the LSFA, a local search mechanism is introduced to improve the performance of FA. After all fireflies update their positions and brightness, several of the brightest fireflies will perform a local search. If a greater intensity is found, the firefly will move to the new spatial coordinate.

Manuscript received September 1, 2021.
Manuscript revised October 26, 2021.
Manuscript publicized November 16, 2021.

[†]The authors are with the Tianjin University of Commerce, China.
[††]The author is with the Tianjin Nankai Hospital, China.
a) E-mail: shiwen@tjcu.edu.cn

DOI: 10.1587/transfun.2021EAL2075

2. Feature Definition for Formula Classification

In TCM, formulas are obtained for clinical treatment for different syndromes in a certain disease. Generally, one formula corresponds to one syndrome. In one formula, herbs play different roles with their effectiveness. Hence, a formula contains multiple effectiveness values. The main effectiveness of the formula is used to treat the syndrome.

Han et al. [15] proposed a model of TCM database for clustering. In this model, a TCM database is comprised of several herbal medicines. Each herbal medicine is represented as an attribute vector which includes the category, property, effectiveness and so on. The similarity measure for herb clustering is based on the attribute vectors of herbs. In this paper, we proposed a novel model of TCM formula based on the model in [15]. In our model, the formula is represented as an attribute vector based on the effectiveness of herbs in this formula and the attributes in the vectors of formulas are used as features for formula classification.

f^h is the herb set $\{h\}$ which constitutes the formula f.

$$f^h = \{h\} \tag{1}$$

h^e is the effectiveness set $\{e\}$ included in herb h.

$$h^e = \{e\} \tag{2}$$

N^e is the amount of effectiveness that is selected as the feature for formula classification. Then, the feature vector v of the formula f is:

$$v^f = [\sum e_1^b, \sum e_2^b, ..., \sum e_{N^e}^b] \tag{3}$$

$$e_i^b = \begin{cases} 1 & \text{if} \quad \exists h, h \in f^h \quad and \quad e_i \in h^e \\ 0 & otherwise. \end{cases}$$

For example, in the theory of TCM, the spleen is a comprehensive conception of structure and function that includes not only the spleen in modern anatomy but also the pancreas and lymphatic system. The spleen is a functional unit involved in multiple systems, such as digestion, absorption, energy conversion, and the immune system [16]. *Qi deficiency in the spleen* is a comprehensive manifestation of decreases in these systems [17]. It has been proven that diarrhea, myasthenia, and sub-health are closely related to *Qi deficiency in the spleen* [16]. The formula Four Gentlemen Decoction (*Sijunzi decoction*), which is a classic recipe of TCM to treat the syndrome of *Qi deficiency in the spleen* by strengthening the spleen and replenishing *Qi*, has been widely applied to treat disorders of gastrointestinal function [18], accompanied by poor appetite, reduced food intake and loose stools [19]. In modern pharmacological studies, *Sijunzi decoction* can also strengthen the immune system [20]. There are four herbs in this formula: the root of *Panax ginseng* C.A. Mey (*renshen*), the rhizome of *Atractylodes macrocephala* Koidz (*baizhu*), *Poria cocos* (Schw.) Wolf (*fuling*) and the root and rhizome of *Glycyrrhiza uralensis* Fisch (*gancao*) [21]. Effectiveness of the four herbs is shown in Table 1. The same effectiveness of the four herbs strengthens the spleen. Meanwhile, the same effectiveness of *renshen*, *baizhu* and *gancao* replenishes *Qi*. Hence, the main effectiveness of *Sijunzi decoction* is *replenishing Qi* and *strengthening the spleen*, which are used for the syndrome of *Qi deficiency in the spleen*. The feature vector of the formula *Sijunzi decoction* is shown in Fig. 1.

Table 1 Effectiveness of herbs in the formula *Sijunzi decoction*.

baizhu	*strengthening the spleen, replenishing Qi, promoting urination*
renshen	*strengthening the spleen, replenishing Qi, tranquilizing*
fuling	*strengthening the spleen, promoting urination, tranquilizing*
gancao	*strengthening the spleen, replenishing Qi, relieving cough*

strengthening the spleen	*replenishing Qi*	*promoting urination*	*tranquilizing*	*relieving cough*
4	3	2	2	1

Fig. 1 The feature vector of the formula *Sijunzi decoction*.

3. Parameter Optimization and Feature Selection by LSFA

In this paper, we proposed an algorithm for parameter optimization and feature selection in a SVM based on LSFA for formula classification by syndrome. In LSFA, a spatial coordinate of fireflies is proposed. The brightness of a firefly is based on the accuracy of SVM. Meanwhile, we developed a local search mechanism to improve FA. The attractiveness and movement of fireflies in the general FA described in [8] are adopted in LSFA. The flowchart of the proposed algorithm is shown in Fig. 2.

3.1 Spatial Coordinate of Fireflies

The spatial coordinate of a firefly in this paper is a binary array. It consists of three parts: parameters C and γ and features for classification. The first N^C bits of the array represent the value of C. The middle N^γ bits of the array represent the value of γ. The last N^e bits of the array represent the features. Set the C_{MAX} and the C_{MIN} as the maximum and minimum values of C. The C_{array} is the decimal value, which is calculated by the array. Then, the true value of C is:

$$C_{value} = \frac{C_{array}(C_{MAX} - C_{MIN})}{2^{N^C} - 1} + C_{MIN} \tag{4}$$

Similarly,

$$\gamma_{value} = \frac{\gamma_{array}(\gamma_{MAX} - \gamma_{MIN})}{2^{N^\gamma} - 1} + \gamma_{MIN} \tag{5}$$

3.2 Brightness of Fireflies

The brightness of fireflies is determined by the accuracy of the SVM based on the parameters of C and γ and the features selected in the spatial coordinate. For training and testing the SVM, a new dataset is generated by removing features

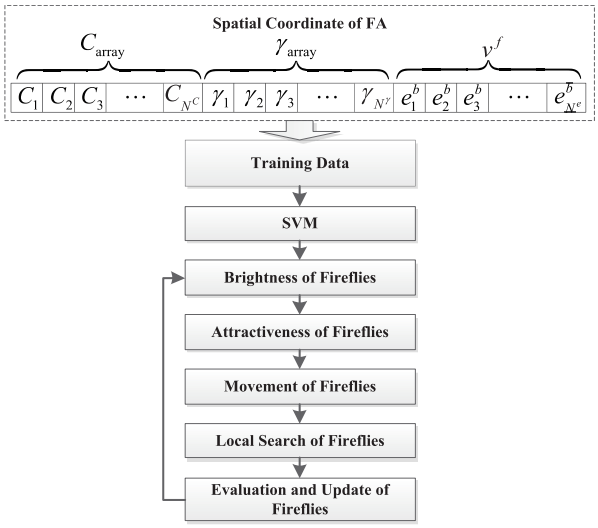

Fig. 2 Flowchart of the LSFA algorithm.

that were not selected from the original dataset. The SVM is trained and tested on the dataset. The classification accuracy is the brightness of the firefly.

3.3 Attractiveness of Fireflies

The attractiveness described in [8] is used as the attractiveness β of a firefly in this paper.

$$\beta = \beta_0 e^{-\gamma^F r^2} \tag{6}$$

where β_0 is the attractiveness at $r = 0$ and γ^F is the light absorption coefficient. The distance between fireflies i and j at x_i and x_j is defined as the Cartesian distance:

$$r_{ij} = ||x_i - x_j|| = \sqrt{\sum_{k=1}^{N^C+N^\gamma+N^e} (x_{i,k} - x_{j,k})} \tag{7}$$

where $x_{i,k}$ is the kth component of the spatial coordinate x_i of the ith firefly. If the brighter firefly is out of range of the field radius of the dimmer firefly, the dimmer firefly will not discover the brighter firefly, and will stay at the current spatial coordinate.

3.4 Movement of Fireflies

The movement of fireflies described in [8] is used in this paper. If firefly i is attracted to another brighter firefly j, the movement of firefly i is:

$$x_i = x_i + \beta_0 e^{-\gamma^F r_{ij}^2}(x_j - x_i) + \alpha\epsilon_i \tag{8}$$

where α is a randomization parameter and is in the range $[0, 1]$. ϵ is a vector of random variables drawn from either a Gaussian or uniform distribution. If the firefly is brighter than the others, it will move randomly according to Eq. (8), where $\beta_0 = 0$.

3.5 Local Search of Fireflies

In this paper, a local search mechanism is proposed in LSFA. After fireflies have moved, they are ranked according to brightness. N^b fireflies with the highest brightness will perform a local search. N^t candidate x are generated according to Eq. (8), where $\beta_0 = 0$. If the brightness of any candidate x is worse than the current x of the firefly, it will stay at the current spatial coordinate, or will move to the candidate x with the highest brightness.

4. Experimental Results

4.1 Computational Environment

In this section, experiments are implemented in MATLAB R2017b and LIBSVM version 3.21; all the experiments are performed on Windows 10 with a Pentium Dual-Core 2.5 GHz processor and 2.0 GB RAM.

Parkinson disease identification dataset in Kaggle dataset [22] is used to evaluate the general performance of LSFA. A formulas of ancient thoracic obstruction and heartache (ATOH) dataset in TCM is used to evaluate the performance for classification of formulas. In this dataset, we collected 291 formulas of ancient thoracic obstruction and heartache from "Dictionary of Traditional Chinese Medicine Prescription" [23] and "Chinese Medicine Encyclopedia" [24]. The specific features of two datasets are shown in Table 2. Parameters of LSFA are shown in Table 3. N^f is the number of fireflies.

4.2 Integral Performance Evaluation

In this study, we compare the ROC curve of the SVM based on the proposed LSFA (LSFA-SVM) with the other four algorithms, including the SVM based on GA (GA-SVM) [6], the SVM based on PSO (PSO-SVM) [7], DT [2] and the SVM without feature selection and parameter optimization with 5-fold cross-validation on the two datasets.

The performance of the various algorithms over the parkinson dataset and the ATOH formulas dataset is shown in Figs. 3 and 4. Best results over 30 replications of LSFA-SVM are compared with four other algorithms, including GA-SVM, PSO-SVM, DT and SVM. We can remark that the area under the ROC curve (AUC) of LSFA-SVM outperforms the other four algorithms in the two datasets. But the AUC of LSFA-SVM is not as outstanding on the parkinson dataset as that on the ATOH formulas dataset. It means that the LSFA-SVM proposed in this paper is more suitable for the formula classification problem in TCM.

4.3 Parameter Optimization, Feature Selection and Local Search Impact

The LSFA we proposed combines parameter optimization

Table 2 The specific features of two datasets.

Dataset	No. of samples	Dimensions	No. of labels
Parkinson	195	22	2
ATOH formulas	291	367	16

Table 3 Parameters of LSFA.

N^e	367	N^b	10	C_{MAX}		4000	C_{MIN}		1
N^c	12	N^t	5	γ_{MAX}		1	γ_{MIN}		0
N^γ	10	N^f	100	β_0	1	γ^F	1	α	0.4

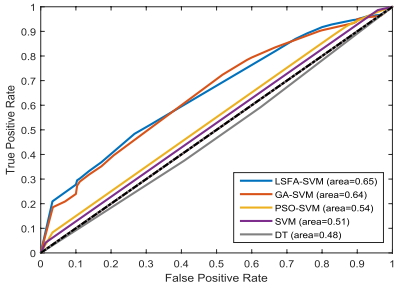

Fig. 3 ROC curve of different algorithm in the parkinson dataset.

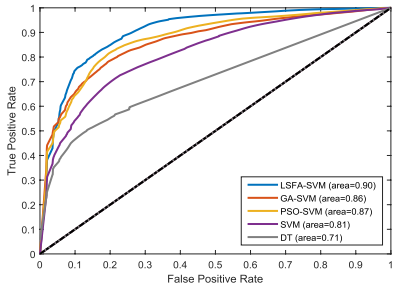

Fig. 4 ROC curve of different algorithm in the ATOH formulas dataset.

and feature selection for SVM to classify formulas. Figure 5 presents the average results of 30 runs of the other two contrast algorithms for the ATOH formulas dataset. As shown in Fig. 5, the obtained optimal solution of LSFA-SVM with parameter optimization and feature selection is superior to LSFA-PA-SVM only with parameter optimization and LSFA-FS-SVM only with feature selection from the first iteration to the 200th iteration. The highest average accuracy of LSFA-SVM reached nearly 70%, while the accuracies of LSFA-FS-SVM and LSFA-PA-SVM did not exceed 50%. The reason is that not all features are conducive to formula classification. We need to select a combination of features that play the most important role in formula classification. Furthermore, parameters such as C and γ of SVM can be effectively tuned through iteration to enhance the classification accuracy. Hence, LSFA-SVM achieved much better results than LSFA-PA-SVM and LSFA-FS-SVM.

Figure 6 shows that the local search mechanism of how optimal fireflies impact the FA. Compared with the FA without the local search mechanism, the performance of the FA with the local search mechanism is better. This is because the brightest fireflies have higher chances of finding better solutions based on their former optimal location through local random movement.

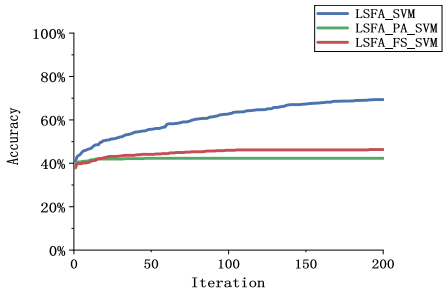

Fig. 5 Average optimal accuracy per iteration for LSFA-SVM, LSFA-PA-SVM and LSFA-FS-SVM.

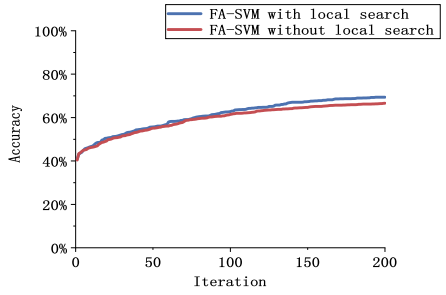

Fig. 6 Average optimal accuracy per iteration for FA-SVM with and without local search.

5. Conclusions

In this paper, we propose an improved firefly algorithm with a local optimal mechanism for parameter optimization and feature selection of SVM to classify formulas in TCM. The effectiveness of herbs is adopted as the feature for formula classification. Parameters C and γ are optimized for better classification performance of SVM. In the proposed LSFA, a local optimal mechanism is adopted to improve the performance of FA. The proposed algorithm LSFA-SVM outperforms other classification algorithms in two datasets. Parameters C and γ and the features are optimized by LSFA to obtain better classification performance. The performance of FA is enhanced by the local optimal mechanism. Further development may focus on the FA for formula clustering problems.

Acknowledgments

This study is supported by the Natural Science Foundation for Young Scientists of Tianjin (No.18JCQNJC70000) and National Student Training Program for Innovaton and Entrepreneurship of China (No.202010069019, No.JDS21009).

References

[1] S. Zhang, X. Li, M. Zong, X. Zhu, and R. Wang, "Efficient kNN classification with different numbers of nearest neighbors," IEEE Trans. Neural Netw. Learning Syst., vol.29, no.5, pp.1774–1785,

2017.
[2] S.L. Salzberg, C4.5: Programs for Machine Learning by J. Ross Quinlan, Morgan Kaufmann Publishers, San Francisco, 1993.
[3] M.J. El-Khatib, B.S. Abu-Nasser, and S.S. Abu-Naser, "Glass classification using artificial neural network," International Journal of Academic Pedagogical Research, vol.3, no.2, pp.25–31, 2019.
[4] P. Tao, Z. Sun, and Z. Sun, "An improved intrusion detection algorithm based on GA and SVM," IEEE Access, vol.6, pp.13624–13631, 2018.
[5] V. Vapnik, The Nature of Statistical Learning Theory, Springer Science & Business Media, 2013.
[6] C.L. Huang and C.J. Wang, "A GA-based feature selection and parameters optimizationfor support vector machines," Expert Syst. Appl., vol.31, no.2, pp.231–240, 2006.
[7] S.W. Lin, K.C. Ying, S.C. Chen, and Z.J. Lee, "Particle swarm optimization for parameter determination and feature selection of support vector machines," Expert Syst. Appl., vol.35, no.4, pp.1817–1824, 2008.
[8] X.S. Yang, "Firefly algorithm, stochastic test functions and design optimisation," International Journal of Bio-Inspired Computation, vol.2, no.2, pp.78–84, 2010.
[9] S. Styawati and K. Mustofa, "A support vector machine-firefly algorithm for movie opinion data classification," IJCCS (Indonesian Journal of Computing and Cybernetics Systems), vol.13, no.3, pp.219–230, 2019.
[10] B. Sahmadi, D. Boughaci, R. Rahmani, and N. Sissani, "A modified firefly algorithm with support vector machine for medical data classification," IFIP International Conference on Computational Intelligence and Its Applications, pp.232–243, Springer, 2018.
[11] G.-G. Wang, L. Guo, H. Duan, and H. Wang, "A new improved firefly algorithm for global numerical optimization," J. Computational and Theoretical Nanoscience, vol.11, no.2, pp.477–485, 2014.
[12] J. Wu, Y.G. Wang, K. Burrage, Y.-C. Tian, B. Lawson, and Z. Ding, "An improved firefly algorithm for global continuous optimization problems," Expert Syst. Appl., vol.149, no.113340, pp.1–12, 2020.
[13] N. Kardani, A. Bardhan, P. Samui, M. Nazem, A. Zhou, and D.J. Armaghani, "A novel technique based on the improved firefly algorithm coupled with extreme learning machine (ELM-IFF) for predicting the thermal conductivity of soil," Eng. Comput., pp.1–20, 2021.
[14] H. Xu, S. Yu, J. Chen, and X. Zuo, "An improved firefly algorithm for feature selection in classification," Wireless Pers. Commun., vol.102, no.4, pp.2823–2834, 2018.
[15] N. Han, S. Qiao, G. Yuan, P. Huang, D. Liu, and K. Yue, "A novel Chinese herbal medicine clustering algorithm via artificial bee colony optimization," Artif. Intell. Med., vol.101, 101760, 2019.
[16] X.-F. Zheng, J.-S. Tian, P. Liu, J. Xing, and X.-M. Qin, "Analysis of the restorative effect of Bu-zhong-yi-qi-tang in the spleen-qi deficiency rat model using ^{1}H-NMR-based metabonomics," J. Ethnopharmacol., vol.151, no.2, pp.912–920, 2014.
[17] B. Xiong and H. Qian, "Effects of Sijunzi decoction and Yupingfeng powder on expression of janus Kinase-signal transducer and activator of transcription signal pathway in the brain of spleen-deficiency model rats," J. Traditional Chinese Medicine, vol.33, no.1, pp.78–84, 2013.
[18] L. Liu, L. Han, D.Y.L. Wong, P.Y.K. Yue, W.Y. Ha, Y.H. Hu, P.X. Wang, and R.N.S. Wong, "Effects of Si-Jun-Zi decoction polysaccharides on cell migration and gene expression in wounded rat intestinal epithelial cells," Brit. J. Nutr., vol.93, no.1, pp.21–29, 2005.
[19] The Pharmacopoeia Commission of PRC, "Formula and single preparation," Pharmacopoeia of People's Republic of China, vol.1, pp.782–783, China Medical Science Press, Beijing, 2015.
[20] N. Zhang, S. Guo, H. Li, J. Li, X. Xu, C. Wan, H. Zhao, F. Liu, J. Zan, B. Wang, and J. Xu, "Effects of Sijunzi decoction on small intestinal T lymphocyte subsets differentiation in reserpine induced spleen deficiency rats," J. Animal and Veterinary Advances, vol.11, no.9, pp.1290–1298, 2012.
[21] B. Gao, R. Wang, Y. Peng, and X. Li, "Effects of a homogeneous polysaccharide from Sijunzi decoction on human intestinal microbes and short chain fatty acids in vitro," J. Ethnopharmacol., vol.224, pp.465–473, 2018.
[22] Maverick, "Parkinson Disease Identification," kaggle Inc., https://www.kaggle.com/vipulbahl/parkinson-disease-identification, accessed Aug. 30. 2021.
[23] H.R. Peng, Dictionary of Traditional Chinese Medicine Prescription, People's Medical Publishing House, Beijing, 1993.
[24] Chinese Medicine Encyclopedia, Hunan Electronic And Audio-visual Publishing House, Hunan, 2006.