International Journal of Hybrid Information Technology
Vol. 10, No. 9 (2017), pp.41-50
http//dx.doi.org/10.14257/ijhit.2017.10.9.04

Predicting Drug-target Interaction using Support Vector Machine and Invasive Tumor Growth Optimization

Deyu Tang[1], Dong Cao[2*] and Jie Zhao[3]

[1]School of Medical Information and Engineering, Guangdong Pharmaceutical University, Guangzhou, 510006, PR. China
[2]School of Medical Information and Engineering, Guangzhou University of Chinese Medicine, Guangzhou ,510006, PR China
[3]Department of Information Management Engineering, School of Management, Guangdong University of Technology, Guangzhou, 510520, PR China
[1]267495824@qq.com, [2]35079751@qq.com, [3]21117331@qq.com

Abstract

Prediction of drug-target interaction is a core problem in drug discovery. In these years, more machine learning methods have been used to solve this problem, but invalid due to the imbalanced data set. In this paper, we propose a new ensemble learning framework by the support vector machine(SVM) and invasive tumor growth optimization(ITGO) algorithm. ITGO is used to solve the penalty parameter optimization problem of SVM for the imbalanced data set. In order to verify the performance of our methods, four benchmark dataset are chosen to compare with the well-known methods. Experimental results show that our method has better effectiveness and robustness than other methods.

Keywords: *Support vector machine; Swarm intelligence; Drug-target interaction; Ensemble learning*

1. Introduction

The identification of drug-target interactions is a core problem of the genomic drug discovery[1]. In the past few years, many efforts have been made to discover new unknown target proteins of drugs due to the completion of the human genome project. However, few of the drug candidates can be approved to reach the market by Food and Drug Administration (FDA), which is caused by the unacceptable toxicity for those drug candidates. Many works in current system biology have shown that the toxicity of drug compounds are greatly caused by the interaction between drugs and some proteins. It is unpractical to achieve all the experiments for checking the drug-target interaction, because it is very expensive and time-consuming. Therefore, non-experimental methods are urgently needed to find the drug-target interactions, which helps the researchers find the potential drug for the old targets or find the potential targets for some old drugs. Computational methods can provide predictions for interactions between drugs and targets and speed up drug discovery [1,2].

In recent years, many machine learning methods have been developed for prediction of genome-wide drug-target interactions. Yamanishi[1] *et al.*, (2008) proposed a supervised bipartite graph learning method. In this method, the geometric space and the chemical space are mapped into a unified space in which those non-interacting drugs and targets are far away from each other and interacting drugs and targets are close to each other.

Received (May 25, 2017), Review Result (August 15, 2017), Accepted (August 20, 2017)
* Corresponding Author

ISSN: 1738-9968 IJHIT

Propose a learned mapping function by mapping the query pair of drug and target to that space, the probability of interaction between them is then calculated as their closeness in the mapped space. Bleakley *et al.*, [2] proposes a bipartite graph inference with local models, in which SVM (support vector machine) was used for the training local models. Gönen, M [3] proposes a novel bayesian formulation that combines matrix factorization, dimensionality reduction and binary classification for predicting drug-target interaction networks using only genomic similarity between target proteins and chemical similarity between drug compounds. Yamanishi *et al.*, [4] develop a new method to predict unknown drug− target interactions consists of inference of unknown drug− target interactions based on the pharmacological effect similarity in the framework of supervised bipartite graph inference. Chen, X. *et al.*, [5] proposes a method of network-based random walk with Restart on the heterogeneous network (NRWRH) to predict potential drug-target interactions. Mei *et al.*, [6] present a simple procedure called neighbor-based interaction-profile inferring (NII) and integrate it into the existing bipartite local model method to handle the new candidate problem., in which neigh information and SVM method are considered. Van Laarhoven *et al.*, [7] propose a simple weighted nearest neighbor procedure and integrate this procedure into a recent machine learning method for prediction of drug-target interaction. Van Laarhoven *et al.*, [8] propose the Gaussian interaction profile (GIP) kernel and use a simple classifier (kernel) regularized least squares (RLS), for prediction drug-target interactions. Shi *et al.*, [9] propose en enhanced similarity measures and super-target clustering method for prediction drug-target interactions.

Although, many machine learning method such as support vector machines (SVM) have been used to predict drug-target interaction, they are invalid due to the imbalance data set. For SVM, penalty parameter should be tuned for this problem, which is a multi-modal optimization problem. In these years, many swarm intelligence algorithms such as particle swarm optimization (PSO)[10], differential evolution (DE)[11,12,13,14], artificial bee colony algorithm (ABC)[15] *etc.*, have successfully used to solve these problems. But many these algorithms are easy to fall into the local optimum. Therefore, we proposed a ensemble learning method framework by SVM and the instructive tumor growth optimization algorithm [16,17].

In this paper, we proposed novel method for prediction of drug-target interaction. The rest of the paper was organized as follows. Section 2 introduces the our method by SVM and ITGO. Section 3 presents experimental simulations compared with well-known algorithms, results, and discussions. Finally, the work is concluded in Section 4.

2. Optimization Problem of Predicting Drug-target Interaction

Though, Bleakley, K. *etc.*, [2] proposed the bipartite local models by SVM, but they has not finished the parameter optimization. Therefore, we consider the local classifier to predict the interaction of drug and target by parameter optimization. According to the similarities between drug compounds and the similarities between target proteins, we can evaluate the interaction between a target protein and a drug compound, which can be formulated as a binary classification. Due to the state-of-the-art performance of SVM, we use it as the local classifier for drugs and targets. In order to achieve the global prediction task of drug-task interaction, we consider the ensemble learning method. Under the local classifier, we require to assign a score to the prediction of drugs or targets. In SVM, although we usually obtain the sign $\{-1,+1\}$ of the score , the value of the score itself contains some form of confidence in the prediction. We propose to rank all candidate samples by the sore value of their prediction according to each local SVM. In this case, we obtain two scores for drug and target candidate samples. Then, we desire choose a rule to convert these two scores into one score and rank the these aggregated scores.

The key problem is the parameter optimization of this ensemble learning method due to the imbalanced datasets from drug and target. In the prediction of drug-target interaction, the number of the positive samples and negative samples are not balance. Usually, the number of the negative samples is far exceed the number of the positive samples, which causes the lose efficacy of the ensemble learning method by SVM. In order to solve this problem, we consider the adjust method of the punishment parameter in our ensemble learning method. It is a continuous optimization problem, we firstly use the currently popular swarm intelligence algorithm, which is called the 'invasive tumor growth optimization algorithm'(ITGO). Figure 1 shows the structure of our method**.**

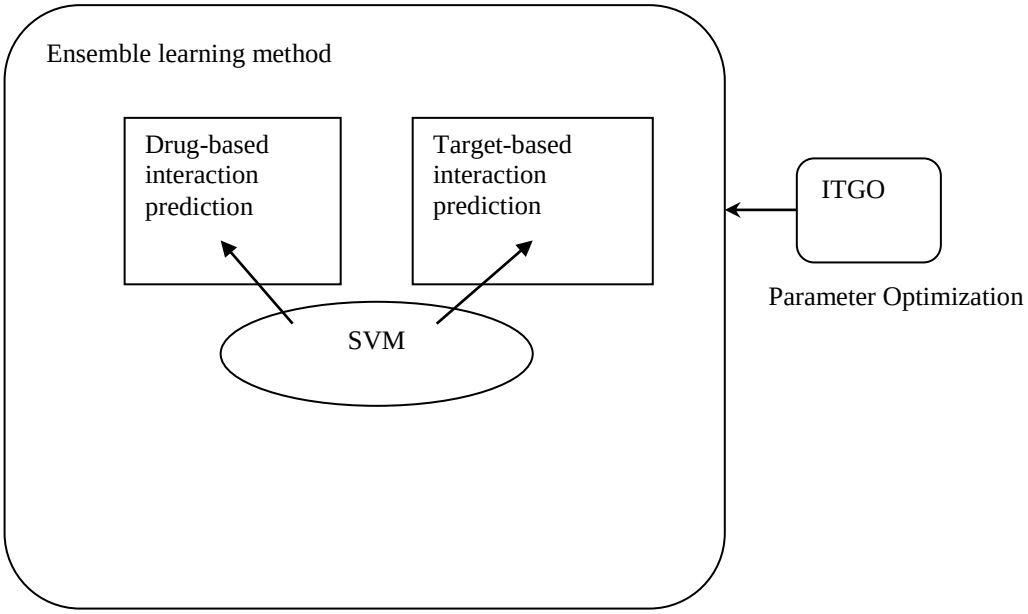

Figure 1. Optimization of Drug-target Interaction Prediction

The proposed approach can be seen in Figure 2. We use the invasive tumor growth optimization algorithm (ITGO) to solve the parameter optimization of SVM. The details of ITGO-based SVM drug-target interaction system is as follows:

2.1. Problem Formulation of Predicting Drug-target Interaction

We are given M_d drug samples denoted as $U_d = \{u_1, u_2, ...u_d\}$ and M_t target samples denoted as $W_t = \{w_1, w_2, ...w_t\}$. u_i means the i drug sample of the similarities between drug compounds and w_i means the i target sample of the similarities between similarities between target proteins. We are also given the interaction of drug and target dataset as the $M_d \times M_t$ adjacency matrix denoted as I. If the u_i drug sample interacts with the w_j target sample, I_{ij} equals to 1 otherwise -1. Then the prediction of drug-target interaction problem can be solved by a binary classifier.

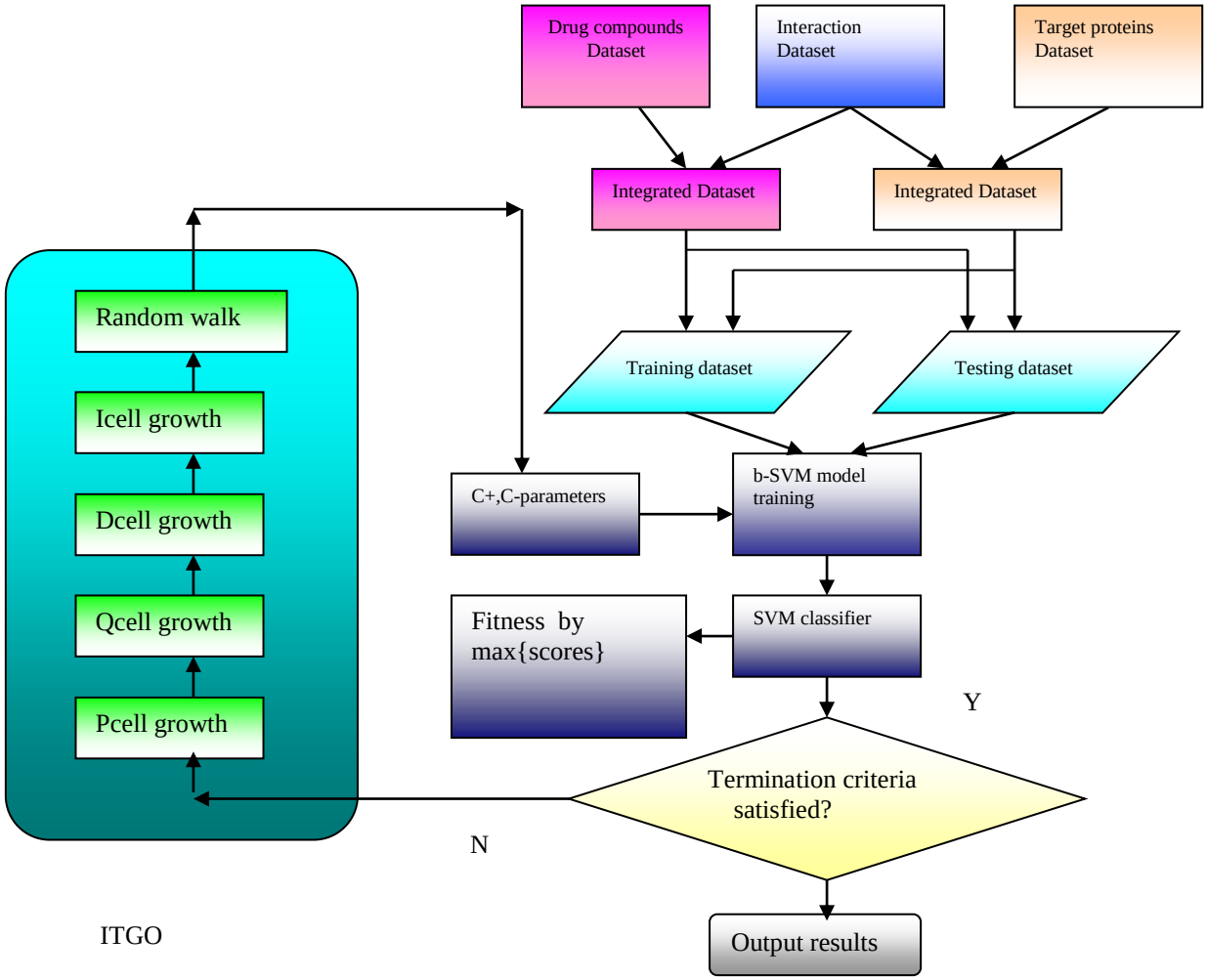

Figure 2. ITGO-based SVM Drug-target Interaction System

2.2 Punishment Parameter Selection of Imbalanced Support Vector Machines

Support vector machine (SVM), a state-of-the-art algorithm, was developed from statistical learning theory by Vapnik, which is based on structural risk minimization principle. SVM method is to find a linear hyperplane, in which the practice sample can not only be classified correctly, but also the classification interval of the two classes is the largest.

Given a training dataset $\{(x_1, y_1),(x_2, y_2),...,(x_l, y_l)\} \in (x \times y)^l\}$, $i = 1,2,...,l$. $x_i \in x \subset R^n$ is the input vector, $y_i = \{-1,+1\}$ is the class label. $w \bullet x + b = 0$ is the maximum margin separating hyper plane and w is a orthogonal vector to the hyper plane. In order to separate the training samples into different categories, we should maximize the margin $1/\|w\|$,that is to minimize $\|w\|^2$. SVM can be seen as a quadratic optimization problem:

$$\min \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}\xi_i \qquad (1)$$

$$s.t.\ y_i(w \cdot \phi(x_i) + b) + \xi_i \geq 1(\xi_i \geq 0, i = 1,2,...,l)$$

C is the penalty parameter which imposes a trade-off between training error and generalization, ξ_i is a slack variable.

In imbalanced data sets, the number of negative class samples are far more than the number of positive samples, so if the selected C value is not large, then all sample will be judged to be negative, so only shows less sample error in training process of SVM. There is an error in the data, the overall classification of the data set still shows a higher accuracy. But such classification results are invalid for imbalanced datasets.

In addition, when the imbalance ratio of positive and negative samples in train data set will increase, the gap of numbers of two classified support vectors from positive and negative samples will increase. It means that the classification of hyper plane in dimensional space will be more affected by negative class samples, the possibility of those points near the classification surface judged as negative class will increase. Therefore, the standard SVM method in dealing with imbalance data set, is apt to judge more sample as negative class, the prediction results are inclined to deviation. An effective way to solve this problem is to classify two classes in SVM model using different penalty parameters. Penalty parameter selection becomes a key optimization problems.

In order to solve the problem of the imbalanced data set, we can give different penalty parameter and then obtain imbalanced-SVM models as follows:

$$\begin{aligned} &\min \frac{1}{2}\sum_{i,j=1}^{l} y_i y_j \alpha_i \alpha_j K(x_i \cdot x_j) - \sum_{j=1}^{l} \alpha_j \\ &s.t. \sum_{i=1}^{l} y_i \alpha_i = 0(1,2,...,l) \\ &0 \le \alpha_i \le C_+, y_i = +1 \\ &0 \le \alpha_i \le C_-, y_i = -1 \end{aligned} \tag{2}$$

2.3. Invasive Tumor Growth Optimization for Penalty Parameter Selection of SVM

Invasive tumor growth optimization, a novel swarm intelligence algorithm, was proposed by Deyu Tang etc for solve continuous optimization in 2015[16,17]. It has been used to solve many machine learning problems and show better performance. Therefore, we consider it to solve the penalty parameter optimization problems in our ensemble learning method for predicting drug-target interaction. In ITGO, tumor cells grow according to the nutrient concentration (fitness). All the tumor cells are divided into five type according to the nutrient concentration from outside into inside of tumor, that is: invasive cells, proliferative cells, quiescent cells, dying cells and apoptotic cells. All the tumor cells grow by the nutrient concentration and interaction between them except apoptotic cells, because they have died. Figure 3 show the process of tumor cell growth.

In order to perform the optimization problems, there are five operations as follows:

1) The growth of proliferative cells

$Pcell_{i,j}$ presents the position of proliferative cell, α control the size of step, *step* presents the movement step by Levy distribution.

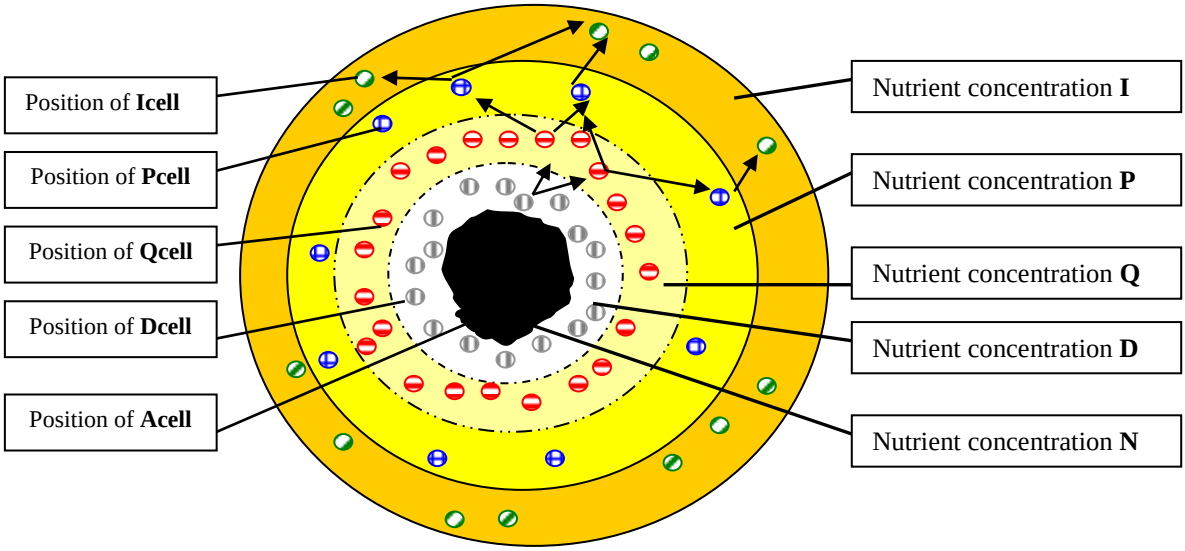

Figure 3. Invasive Tumor Growth Optimization

$$Pcell_{i,j}(t+1) = Pcell_{i,j}(t) + \alpha \bullet step \tag{3}$$

$$\alpha = rand \bullet (\frac{Fes}{Max\_fes}) \tag{4}$$

$$step = \frac{u}{|v|^{1/\omega}} \tag{5}$$

$$u \sim N(0, \sigma_u^2), v \sim N(0, \sigma_v^2) \tag{6}$$

$$\sigma_v = \{ \frac{\Gamma(1+\omega)\sin(\pi\omega/2)}{\Gamma[(1+\omega)/2]\omega 2^{(\omega-1)/2}} \}^{1/\omega} \tag{7}$$

$$\sigma_u = 1 \tag{8}$$

2) The growth of quiescent cells

$sQcell_{i,j}(t+1)$ is the current position of quiescent cell., β presents the control size of step, Here, $hPcell_{p,j}(t)$ is the historical best position of proliferative cell, $cPcell_{p,j}(t)$ is the current position of proliferative cell.

$$sQcell_{i,j}(t+1) = \begin{cases} Qcell_{i,j}(t) + \beta \bullet step \bullet (hPcell_{p,j}(t) - Qcell_{i,j}(t)) + \beta \bullet step \bullet ((Qcell_{x,j}(t) - Qcell_{y,j}(t))), rand < 0.5 \\ Qcell_{i,j}(t) + \beta \bullet step \bullet (cPcell_{p,j}(t) - Qcell_{i,j}(t)) + \beta \bullet step \bullet ((Qcell_{x,j}(t) - Qcell_{y,j}(t))), rand > 0.5 \end{cases} \tag{9}$$

$$Qcell_{i,j}(t+1) = \begin{cases} Qcell_{i,j}(t), rand < e^{(\frac{Fes}{Max\_fes} - 1)} \\ sQcell_{i,j}(t+1), else \end{cases} \tag{10}$$

$$\beta = rand(0,1) \bullet normal(0,1) \tag{11}$$

3) The growth of dying cells

$Dcell_{i,j}(t)$ indicates the position of the old dying cell, $Qcell_{x,j}(t)$ indicates the position of quiescent cell, $cPcell_{p,j}(t)$ indicates the position of proliferative cell (the current best position), $\gamma = rand[-1,1]$.

$$Dcell_{i,j}(t+1) = Dcell_{i,j}(t) + \gamma \bullet (cPcell_{p,j}(t) - Dcell_{i,j}(t)) + \gamma \bullet ((Qcell_{x,j}(t) - Dcell_{i,j}(t))) \tag{12}$$

4) The growth of invasive cells

$$newCell_{i,j}(t) = Random(i, D), D \in [D\min, D\max] \quad (13)$$

$$Dcell_{i,j}(t+1) = cPcell_{p,j}(t) + \eta \bullet (newCell_{i,j}(t) - cPcell_{p,j}(t)) \quad (14)$$

$$\eta \in rand[-1,1] \quad (15)$$

Where $newCell_{i,j}$ is new cell from population and $cPcell_{p,j}$ is the position of current proliferative cell.

5) Random walk of cells

$$Cell_{i,j}(t+1) = Cell_{i,j}(t) + \lambda \cdot (\frac{newCell_{i,j}(t)}{\| newCell_{i,1:D}(t) \|}) \quad (16)$$

Where $\| newCell_{i,j:D}(t) \|$ is an Euclidean norm. $\lambda \in rand[-1,1]$.

3. Experiments

In order to verify the performance of our methods, we use four benchmark data sets about drug-target interaction including enzymes, ion channels, GPCRs[2]. We test 10 trials of 10-fold cross-validation for all the experiments. In order to perform SVM, we also handled positivesemi-definite matrices of drug data set and target data set. We used the LIBSVM (v.2.88) (Chang and Lin, 2001) for implementation of SVM. we computed the ROC curve of true positives as a function of false positives when the threshold to predict interactions from the ranking varies. In addition, the area under this curve (AUPR for area under PR) is also computed for evaluation of the performance. Table 1-Table 4 show the comparison results of four benchmark data set for four algorithms including m(KRMd BLMd)[2], m(KRMt BLMt)[2], m(NNd,NNt)[2]and ITGO-SVM. We can see that the performance of our proposed method is better than other four methods except m(KRMd BLMd) for NR data set.

Table 1. NR Data Set

method	AUC	AUPR
m(KRMdBLMd)[2]	**85.4 45.0**	45.0
m(KRMt BLMt)[2]	53.6 36.0	36.0
m(NNd,NNt)[2]	85.1	53.6
ITGO-SVM	0.8450	**0.5672**

Table 2. NC Data Set

method	AUC	AUPR
m(KRMd,BLMd)[2]	73.9	33.9
m(KRMt,BLMt)[2]	93.5	81.3
m(NNd,NNt)[2]	91.7	53.8
ITGO-SVM	**0.9716**	**0.8292**

Table 3. GPCR Data Set

method	AUC	AUPR
m(KRMd,BLMd)[2]	88.2	41.4
m(KRMt,BLMt)[2]	86.7	57.4
m(NNd,NNt)[2]	88.5	48.5
ITGO-SVM	**0.9435**	**0.6335**

Table 4. ENZYMES Data Set

method	AUC	AUPR
m(KRMd,BLMd)[2]	86.9	39.4
m(KRMt,BLMt)[2]	94.4	80.7
m(NNd,NNt)[2]	93.0	63.8
ITGO-SVM	**0.9673**	**0.8474**

3. Conclusions

In this paper, we propose a new ensemble learning framework by SVM and ITGO for prediction of drug-target interaction. Drug data and interaction data are integrated as drug-based data set. Target data and interaction data are integrated as target-based data set. Two data set are training by SVM and scores are competed simultaneously. In addition, there are two penalty parameters C+ and C- in SVM are tuned by invasive tumor growth optimization. Then, we used four benchmark data set for comparison of many well known methods such as m(KRMd,BLMd), m(KRMt,BLMt), m(NNd,NNt) and Weighted profile, the results show that our proposed method has better performance than other method. In future, our proposed method can be widely used to solve other problems.

Acknowledgment

This work is partially supported by Guang Dong Provincial Natural fund project(2014A030313585, 2015A030310267, 2015A030310483), Drug-target interaction prediction method based on collaborative intelligent optimization (2016A030310300), NSFC (No. 61501128, 71401045).Guangdong provincial finance special project of clinical research and information integration platform for prevention and treatment of major diseases of traditional Chinese Medicine. Guangdong Provincial Department of education young creative talents project (2016KQNCX024); 2017 Cloud computing and big data national key special project (2017YFB1002300).

References

[1] Y. Yamanishi, “Prediction of Drug− Target Interaction Networks from the Integration of Chemical and Genomic Spaces”, Bioinformatics, vol. 24, **(2008)**, pp. i232-i240.
[2] K. Bleakley and Y. Yamanishi, “Supervised Prediction of Drug–target Interactions Using Bipartite Local Models, Bioinformatics, vol. 25, no. 18, **(2009)**, pp. 2397-2403.
[3] M. Gönen, “Predicting Drug–target Interactions from Chemical and Genomic Kernels Using Bayesian Matrix Factorization”, Bioinformatics, vol. 28, no. 18, **(2012)**, pp. 2304-2310.
[4] Y. Yamanishi, M. Kotera, M. Kanehisa and S. Goto, “Drug-target Interaction Prediction from Chemical, Genomic and Pharmacological Data in an Integrated Framework”, Bioinformatics, vol. 26, no. 12, **(2010)**, pp. i246-i254.
[5] X. Chen, X., M. X. Liu and G. Y. Yan, “Drug–target Interaction Prediction by Random Walk on the Heterogeneous Network”, Molecular BioSystems, vol. 8, no. 7, **(2012)**, pp. 1970-1978.
[6] J. P. Mei, C. K. Kwoh, P. Yang, X. L. Li and J. Zheng, “Drug–target Interaction Prediction by Learning from Local Information and Neighbors”, Bioinformatics, vol. 29, no. 2, **(2013)**, pp. 238-245.
[7] T. van Laarhoven and E. Marchiori, “Predicting Drug-target Interactions for New Drug Compounds Using a Weighted Nearest Neighbor Profile”, PloS one, vol. 8, no. 6, **(2013)**.
[8] T. van Laarhoven, S. B. Nabuurs and E. Marchiori, “Gaussian Interaction Profile Kernels for Predicting Drug–target Interaction”, Bioinformatics, vol. 27, no. 21, **(2011)**, pp. 3036-3043.
[9] J. Y. Shi, S. M. Yiu, Y. Li, H. C. Leung and F. Y. Chin, “Predicting Drug–target Interaction for New Drugs Using Enhanced Similarity Measures and Super-target Clustering”, Methods, vol. 83, **(2015)**, pp. 98-104.
[10] J. Kennedy and R. C. Eberhart, “Particle Swarm Optimization”, Proceeding IEEE International Conference Neural Network, **(1995)**; Perth, Western Australia.
[11] A. K. Qin and P. N. Suganthan, “Self-adaptive Differential Evolution Algorithm for Numerical Optimization”, Proc. IEEE Congr. Evol. Comput, **(2005)**, pp. 1785-1791.

[12] A. K. Qin, V. L. Huang and P. N. Suganthan, "Differential Evolution Algorithm with Strategy Adaptation for Global Numerical Optimization", IEEE Transactions on Evolutionary Computation, vol. 13, no. 2, **(2009)**, pp. 398-417.
[13] R. M. Storn and K. V. Price, "Differential Evolution -a Simple and Efficient Heuristic for Global Optimization over Continuous Spaces", Journal of Global Optimization, vol. 11, **(1997)**, pp. 341-359.
[14] J. Zhang and A. C. Sanderson, "JADE: Adaptive Differential Evolution with Optional External Archive", IEEE Transactions on Evolutionary Computation, vol. 13, no. 5, **(2009)**, pp. 945-958.
[15] D. Karaboga, "An Idea Based on Honey Bee Swarm for Numerical Optimization", Technical Report TR06, Erciyes University, **(2005)**.
[16] D.-Y. Tang, S.-B. Dong, Y. Jiang, H. Li and Y.-S. Huang, "ITGO: Invasive Tumor Growth Optimization Algorithm, Applied Soft Computing, vol. 36, **(2015)**, pp. 670-698
[17] D.-Y. Tang, S.-B. Dong, L.-F. He and Y. Jiang, "Intrusive Tumor Growth Inspired Optimization Algorithm for Data Clustering", Neural Computing & Applications, **(2016)**, pp. 27:349-374.

Authors

Deyu Tang, received the Ph.D. degree from the School of Computer Science and Technology of South China University of Technology, China, in 2015. He is now an associate professor in the School of Medical Information Engineering, Guangdong Pharmaceutical University, Guangzhou, China. He research interests are in the areas of swarm intelligence, machine learning, bioinformatics. He has published more than 14 articles in different journals including Information Sciences, Applied Soft Computing, Neural Computing & Applications etc.

Dong Cao, received the Ph.D. degree from the School of Mechanical Engineering, South China University of Technology, China. In 2013, he was a visiting scholar at the University of Dallas in Dezhou. He is now a professor in the University of Guangzhou Chinese Medicine. He research interests are in the areas of Traditional Chinese Medicine Informatics, medical large data analysis. He has published more than 50 articles.

Jie Zhao is an associate professor in the school of management at Guangdong University of Technology. She received her Ph.D. in Computer Science from South China University of Technology in 2010. Her main research interests include data mining, business intelligence, machine learning, and information fusion. Her research work has been published in renowned journals.