Hindawi
Mathematical Problems in Engineering
Volume 2021, Article ID 5585990, 10 pages
https://doi.org/10.1155/2021/5585990

Research Article

Virtual Screening of Drug Proteins Based on Imbalance Data Mining

Peng Li, Lili Yin, Bo Zhao, and Yuezhongyi Sun

School of Computer Science and Technology, Harbin University of Science and Technology, Harbin 150080, Heilongjiang, China

Correspondence should be addressed to Peng Li; printing3d@126.com

Received 30 January 2021; Accepted 11 May 2021; Published 24 May 2021

Academic Editor: Erik Cuevas

To address the imbalanced data problem in molecular docking-based virtual screening methods, this paper proposes a virtual screening method for drug proteins based on imbalanced data mining, which introduces machine learning technology into the virtual screening technology for drug proteins to deal with the imbalanced data problem in the virtual screening process and improve the accuracy of the virtual screening. First, to address the data imbalance problem caused by the large difference between the number of active compounds and the number of inactive compounds in the docking conformation generated by the actual virtual screening process, this paper proposes a way to improve the data imbalance problem using SMOTE combined with genetic algorithm to synthesize new active compounds artificially by upsampling active compounds. Then, in order to improve the accuracy in the virtual screening process of drug proteins, the idea of integrated learning is introduced, and the random forest (RF) extended from Bagging integrated learning technique is combined with the support vector machine (SVM) technique, and the virtual screening of molecular docking conformations using RF-SVM technique is proposed to improve the prediction accuracy of active compounds in docking conformations. To verify the effectiveness of the proposed technique, first, HIV-1 protease and SRC kinase were used as test data for the experiments, and then, CA II was used to validate the model of the test data. The virtual screening of drug proteins using the proposed method in this paper showed an improvement in both enrichment factor (EF) and AUC compared with the use of the traditional virtual screening, for the test dataset. Therefore, it can be shown that the proposed method can effectively improve the accuracy of drug virtual screening.

1. Introduction

In recent years, the continuous discovery of natural and synthetic compounds and other small molecules has provided a large amount of data validation resources for the drug development process, but since the traditional validation process in drug development is still mainly by means of clinical trials, nothing can be done in the face of the large number of clinical trial datasets nowadays [1]. This has led to the phenomenon that although a very large amount of money is invested in the drug development process each year, very few drugs are actually produced [2]. In the drug development process, the type of lead compound and the quality of the compound have a direct impact on the ease and duration of drug development; for example, the better the quality of the lead compound, the lower the elimination rate of the drug in the drug development process, and vice versa [3]. Since the 1980s, computer technology has been widely used in the drug development process due to the efficient computing power of computers [4]. The use of computer technology in drug development not only saves time and money but also, and most importantly, greatly improves the accuracy of the drug development process. One of the typical representatives of computer-aided drug design can be called virtual screening technology [5].

Virtual screening technology is actually a simulation of the drug development process on a computer, which makes it an efficient drug development aid due to the efficient performance of computers, and it can also improve the accuracy of drug development [6]. In experimental datasets, imbalance between inactive and active compound datasets can make the experimental data biased and seriously affect

the experimental results of virtual screening [7]. Therefore, the virtual screening process of drugs can be viewed as an imbalanced data classification problem. Traditional virtual screening techniques usually use two schemes, similarity search and sampling scoring, for the screening of lead compounds [8]. However, both schemes do not deal well with the problem of data imbalance caused by the disparity in the number of active and inactive compounds among the candidate compounds. In this paper, we introduce the problem of processed imbalanced data into the process of virtual screening of drug proteins, an idea that is still rare in the field of drug development. In the actual drug development process, we should focus more on the data of active compounds with less data volume in the dataset. Therefore, a detailed study of this imbalanced data prediction problem is carried out in this paper, and its study is applied to the drug virtual screening process, and a complete drug virtual screening scheme is established.

In the process of drug development, virtual screening of drug proteins using computer-aided drug design can improve the efficiency of drug development to a great extent [9]. To address the problem of imbalanced data arising from the large difference between the number of active and inactive compounds during the virtual screening approach, this paper proposes a genetic algorithm based on SMOTE to perform preprocessing on the imbalanced dataset as a way to reduce the imbalance ratio of the categorical data, by increasing the number of positive examples of data from a few classes, not only to ensure the integrity of the data but also reduce the imbalance ratio. In response to the problem that the traditional virtual screening technique is not effective in classification, this paper introduces the idea of integrated learning and proposes a more effective classification of the binding conformation of molecules based on RF-SVM classification algorithm, from which effective active compounds can be screened, and this method can not only improve the screening accuracy of lead compounds but also accelerate the drug development process and shorten the drug development cycle [10].

2. Genetic Algorithm Upsampling Technique Based on SMOTE

Data scarcity is evident during virtual screening experiments, where out-of-balance data are often required. This is because the crystal structures of many proteins have not yet been resolved. For example, the crystal structure of the resolved SRC kinase is only about 90, and if SRC is used as a target protein for virtual screening, it is easy to have a serious imbalance in the number of positive and negative example data samples in the dataset. The imbalance data problem in the classification problem is mainly manifested in the imbalance of the positive and negative data, often appearing in the classification process to the side of the majority of the negative data, and some may even cause the phenomenon of "data flooding." To address the imbalance problem, this paper combines the advantages of SMOTE algorithm and genetic algorithm and proposes a genetic algorithm upsampling technique based on SMOTE.

The SMOTE algorithm mainly aims to reduce the imbalance ratio by adding the number of data elements of the positive sample, but there is a certain blindness in generating new data. There are many uncertainties and errors in the process of actual use; for example, the data information around the positive sample is not considered in the process of sample synthesis, so the phenomenon of data confusion is easy to occur.

The genetic algorithm synthesizes new data samples by selecting the determined sample data by crossover variation, where the selection operator can be used to select the samples of the dataset to be manipulated and to control the number of new data to be synthesized in the end, the crossover operator of the genetic algorithm can keep the newly generated samples similar to the original sample data, and the diversity of the new samples can be maintained by the variation operator. By studying the advantages and disadvantages of these two algorithms in synthesizing new datasets, a new algorithm (GA-SMOTE) is proposed, which incorporates the ideas of the SMOTE algorithm into the genetic algorithm.

The specific steps of the new algorithm are as follows.

(1) The degree of influence of positive and negative examples in the training dataset is coded according to the eigenvalues of the sample dataset

(2) Two samples are selected from the sample data according to the fitness function

(3) The selected two samples are crossed or mutated to produce a new sample dataset

The flow of the new algorithm is as follows.

(1) Set the input original training sample TrainData; set the output path DataSet. Set the number of feature values in the sample dataset to n, the number of large class samples in the dataset to N, and the number of small class samples to T.

(2) Set the number of samples NUM, that is, the number of minority class samples that need to be synthesized in order to balance the minority class samples and the majority class samples.

(3) Coding of eigenvalues.

① A sample is randomly selected from a small number of classes T_i

② The k minority class sample sets $\{T_1, T_2, T_3, \ldots, T_k\}$ around T_i and the k majority class sample datasets $\{N_1, N_2, N_3, \ldots, N_k\}$ are selected by the Euclidean distance approach in the SMOTE algorithm, while the k sample data around T_i are selected, and if more than half of the k sample data are majority class samples, this data is assumed to be at the boundary position and no cross-variance operation is performed on it

③ Calculate the average value Lt of the distance from T_i to $\{T_1, T_2, T_3, \ldots, T_k\}$ and the average value Ln of the distance from T_i

to $\{N_1, N_2, N_3, \ldots, N_k\}$, and calculate $(Lt/Ln) = L_1$

④ Remove a random eigenvalue from the sample, and then, calculate the average value $Lt\prime$ of the distance from T_i to $\{T_1, T_2, T_3, \ldots, T_k\}$ and the average value $Ln\prime$ of the distance from T_i to $\{N_1, N_2, N_3, \ldots, N_k\}$, respectively, after removing this eigenvalue. $(Lt'/Ln') = L_2$ is calculated

⑤ If $L_1 > L_2$, then this eigenvalue can be characterized as majority class; otherwise, this eigenvalue can be characterized as minority class, and then use 0 to characterize the majority class eigenvalue and 1 to characterize the minority class eigenvalue

⑥ Repeat the previously mentioned steps until all n eigenvalues are encoded and finished

The pseudocode is as follows (see Algorithm 1).

(4) The minority class is ranked according to the fitness function of the minority class sample, and the top 50% is taken out, and then two samples of data are randomly selected from them, and the crossover variation is performed. If it is the majority class eigenvalue, the variation is performed by SMOTE algorithm, and if it is the minority class sample, the crossover is performed The schematic diagram is shown in Figure 1, where x represents a random variable between 0 and 1.

(5) Loop through the fourth step until the new sample data generated reaches NUM.

The flowchart of the improved genetic algorithm is shown in Figure 2.

2.1. Introduction of RF-SVM-Based Classification Algorithm. Currently, support vector machines (SVM) have been able to solve most of the data classification problems with relatively small amount of data, distinct eigenvalue identification, and relatively balanced data distribution [11]. However, for the classification problem of imbalanced datasets, the effectiveness of SVM decreases significantly. It is mainly because of the uneven distribution of the training dataset, which leads to a serious imbalance in the ratio of positive and negative example sample data, and the majority class of negative example data samples dominates, making the final classification hyperplane tilted towards the negative example data. Experimental studies have shown that SVM is more effective in dealing with classification of imbalanced data compared to other classification models. The flow of SVM in dealing with classification problems is as follows.

The given sample dataset is denoted as $L = \{(x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots, (x_n, y_n)\}$, where $x_i \in R^d$, $y_i \in \{-1, 1\}$, $i = 1, 2, 3, \ldots, n$. y_i denotes the class of sample x_i, d denotes the number of bits of sample data, and n denotes the number of sample data for training. The representation of the hyperplane is as follows:

$$w \cdot x + b = 0, \tag{1}$$

where w is the normal vector, representing the direction of the hyperplane, and b is the offset, representing the distance between the hyperplane and the origin. Therefore, a hyperplane can be represented by (w, b). The distance of any individual x in the sample space to the hyperplane can be expressed as r:

$$r = \frac{|wx + b|}{\|w\|}. \tag{2}$$

The SVM algorithm learning problem can be formulated as a constrained optimality problem, as shown in the following equation:

$$\min \frac{\|w\|^2}{2}, \tag{3}$$

and $(\|w\|^2/2)$ denotes the maximization decision edge, which has the following constraints:

$$y_i * (w \cdot x + b) \geq 1, \quad i = 1, 2, 3, \ldots, n. \tag{4}$$

This is a convex optimization problem, which can be solved by introducing the standard form of Lagrange multipliers with a modification of the objective function, as shown in the following equation:

$$L_p = \frac{\|w\|^2}{2} - \sum_{i=1}^{n} \lambda_i \left(y_i * (w \cdot x + b) - 1\right), \tag{5}$$

where λ_i denotes the introduced Lagrange multiplier. In order to minimize λ_i, w and b can be made equal to zero, and then, the derivative can be obtained as follows:

$$\begin{aligned} \frac{\partial L_p}{\partial w} = 0 &\Longrightarrow w = \sum_{i=1}^{n} \lambda_i y_i x_i, \\ \frac{\partial L_p}{\partial b} = 0 &\Longrightarrow \sum_{i=1}^{n} \lambda_i y_i = 0. \end{aligned} \tag{6}$$

The problem can be transformed into the pairwise function formula shown in the following equation:

$$L_D = \sum_{i=1}^{n} \lambda_i - \frac{1}{2} \sum_{i,j} \lambda_i \lambda_i y_i y_i x_i x_i. \tag{7}$$

The decision boundary can also be expressed as shown in the following equation:

$$\left(\sum_{i=1}^{n} \lambda_i y_i x_i \cdot x\right) + b = 0. \tag{8}$$

After determining the parameters of the decision boundary, the final classifier formula is obtained as shown in the following equation:

$$f(z) = \operatorname{sign}(w \cdot x + b) = \operatorname{sign}\left(\sum_{i=1}^{n} \lambda_i y_i x_i \cdot z + b\right), \tag{9}$$

where sign denotes the symbolic function and z denotes the instance data to be detected.

Input: number of minority class samples T; number of attributes n; K nearest nearby data
Output: attributes coding
(1) while $x < n$
(2) Randomize the minority class samples T_i
(3) k = Number of nearest Nearby
(4) $\{T_1, T_2, T_3, \ldots, T_k\}$: k minority classes around T_i
(5) $\{N_1, N_2, N_3, \ldots, N_k\}$: k majority classes around T_i
(6) Classes around T_i
(7) if majority classes > minority classes
(8) continue
(9) End
(10) Lt = Average distance from T_i to $\{T_1, T_2, T_3, \ldots, T_k\}$
(11) Ln = Average distance from T_i to $\{N_1, N_2, N_3, \ldots, N_k\}$
(12) $(Lt/Ln) = L_1$
(13) Randomize delete an attributes
(14) Lt' = Average distance from T_i to $\{T_1, T_2, T_3, \ldots, T_k\}$
(15) Ln' = Average distance from T_i to $\{N_1, N_2, N_3, \ldots, N_k\}$
(16) $(Lt'/Ln') = L_2$
(17) if $L_1 > L_2$
(18) Then the attribute representative the majority classes
(19) End
(20) The attribute representative the majority classes
(21) End

Algorithm 1: GA_SMOTE (T, n, k, x).

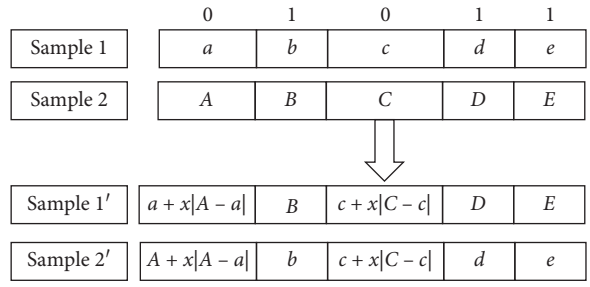

Figure 1: Eigenvalue cross-variation graph.

To address the shortcomings of SVM in dealing with imbalanced data problems, this paper introduces the idea of integrated learning. Since the Bagging classifier does not target a specific instance data in the training set, each sample is selected with equal probability. Bagging has a better fitting performance, and the independence between individual classifiers of Bagging is higher. Therefore, the classifiers can run in parallel with each other, and the speed is faster, almost the same as the speed of individual classifiers. The random forest algorithm is an extension of Bagging, which not only retains all the advantages of Bagging and adds automatic feature selection but is also simpler to implement than the Boosting family of algorithms. SVM shows good performance for processing small sample data, which is in line with the realistic data situation in the virtual screening process, and can also solve the high-dimensional data classification problem, so this paper chooses the combination of RF and SVM to build a classification model for the data of drug proteins.

The algorithmic idea is as follows: random forest is actually a combined classifier, combining individual classifiers together using the voting method to get the final classifier, because RF is not sensitive to the noise data in the dataset; therefore, the noise data in the data are first filtered by the SVM algorithm, that is, the misclassified data or the data that are not easily distinguished are extracted and selected, and then the data are learned and trained by the RF algorithm.

Input data are as follows: the training sample dataset $L = \{(x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots, (x_n, y_n)\}$, $x_i \in R^d$, $y_i \in \{-1, 1\}$, $i = 1, 2, 3, \ldots, n$ and the number of input iterations N.

Phase I:

(1) First, a portion of the data in L is selected for training, and then, the remaining data is iteratively added
(2) Classify the data through the hyperplane and remove the assumed noisy dataset according to a certain strategy
(3) Repeat the above steps N times

Phase II: training sample dataset N; number of base classifiers of RF X

(4) The training sample data N is randomly divided into X points, and the subdata samples are assumed to be T_i, $i = 1, 2, 3, \ldots, X$

(5) Train the base classifier by base classifier CART decision tree

(6) Repeat the previously mentioned steps X times

(7) The final strong classifier is selected based on the voting method

It is demonstrated experimentally that the best training effect of RF is achieved when X is taken as 100, and the sampling ratio of training data can be taken as 0.63. The value of N not only affects the accuracy of SVM algorithm and the speed of SVM but also affects the judgment of noisy data. Therefore, the final classification result is optimal when the value of N is about 40 through the experimental constant tuning test.

3. Experimental Verification and Analysis

3.1. Evaluation Criteria. In this paper, we chose the enrichment factor (EF) as one of the evaluation indexes. EF is an important criterion for evaluating the effectiveness of virtual screening, and it is also the most general criterion for evaluating virtual screening techniques. In the case of relatively small sample data, a larger value of the enrichment factor indicates a greater number of active compounds, which can be detected [10]. The formula for calculating EF is as follows:

$$\mathrm{EF} = \left(\frac{\mathrm{Hits}_{\mathrm{sampled}}^{x\%}}{N_{\mathrm{sampled}}^{x\%}}\right) \cdot \left(\frac{N_{\mathrm{total}}}{\mathrm{Hits}_{\mathrm{total}}}\right), \tag{10}$$

where $\mathrm{Hits}_{\mathrm{sample}}^{x\%}$ denotes the number of active compounds in x% of the test set database, $N_{\mathrm{sampled}}^{x\%}$ denotes the number of inactive decoys in x% of the test set database, N_{total} denotes the number of compounds in the entire test set, and $\mathrm{Hits}_{\mathrm{total}}$ denotes the number of active compounds in the entire test set.

In the actual drug screening process, only a small fraction of the compounds are screened by computer means. In this paper, EF is selected as the evaluation criteria for virtual screening, and ROC curve and AUC area are also introduced in the evaluation criteria of this paper. The ROC curve, also known as the "subject operating characteristic curve," is a visual graphical evaluation criterion that graphically represents the correlation between sensitivity and specificity. The false positive rate is used as the horizontal coordinate and the true positive rate as the vertical coordinate. The AUC area represents the area of the graph enclosed by the ROC curve and the x-coordinate axis, allowing a more visual analysis of the model [12]. The formula for the AUC area is as follows:

$$\mathrm{AUC}(f) = \sum_{t=1}^{n} \frac{1}{2}\left(y_{i-1} + y_i\right) \cdot \left(x_i - x_{i-1}\right). \tag{11}$$

The value of AUC area ranges from 0 to 1, but it is usually between 0.5 and 1. When the value of AUC reaches 1, it means that the classification effect of the classifier reaches the optimal situation, and the larger the value of AUC area is, the better the prediction effect of this model classifier is.

3.2. Experimental Data Acquisition. The two specific techniques proposed in this paper, "SMOTE-based genetic algorithm" and "RF-SVM-based classification algorithm," were validated using HIV-1 protease and SRC kinase as experimental data. HIV-1 protease is one of the more critical proteases for the treatment of human immunodeficiency syndrome, and HIV-1 protease has two identical peptide chains, a homodimer composed of two polypeptides [13]. SRC kinase is widely present in tissue cells and can react with important molecules in signal transduction pathways, participating in cellular metabolic processes and regulating cell growth, development, and differentiation. This kinase has been implicated in the development of several cancers [14]. Activation of the SRC pathway has been detected in approximately more than 50% of various cancers, so the treatment of cancer can be achieved by inhibiting the activity of SRC [12]. The crystal structures of HIV-1 protease and SRC kinase can be obtained directly from the PDB database. The HIV-1 protease has a PID of 1PRO with a resolution of 1.80 Å [15] and the structure is shown in Figure 3. The SRC kinase has a PID of 2H8H with a resolution of 2.20 Å and the structure is shown in Figure 4.

In this paper, the experiments are divided into four cases, in which the two proteins are compared before and after processing by the "SMOTE-based genetic algorithm" and the comparison experiments using the two methods of SVM and RF-SVM. The data structures before and after sampling are shown in Tables 1 and 2.

For both datasets, HIV-1 and SRC, the ROCs of the previously mentioned two datasets for the comparison experiments are shown in Figure 5.

In order to prove the validity of the experiments, two sets of representative data were used as experimental data for model reliability comparison experiments, which were conducted before and after sampling preprocessing, SVM algorithm and RF-SVM algorithm, respectively. The differences in ROC curve, AUC area, enrichment factor (EF), and accuracy were obtained by the experimental results, which showed that the RF-SVM algorithm proposed in this paper was improved for the virtual screening experimental results of drug proteins.

The data of HIV-1 protease and SRC kinase were observed: when HIV-1 protease was used as the target protein, the area of AUC increased from 0.795 to 0.839, the enrichment factor increased from 6.58 to 7.16, and the accuracy increased from 0.651 to 0.746, indicating that the results of the virtual screening of drug proteins by RF-SVM were improved after the sampling preprocessing. When the virtual screening simulations were performed by the basic SVM algorithm and RF-SVM algorithm before the sampling preprocessing, the area of the AUC increased from 0.729 to 0.785, the enrichment factor increased from 5.93 to 6.51, and

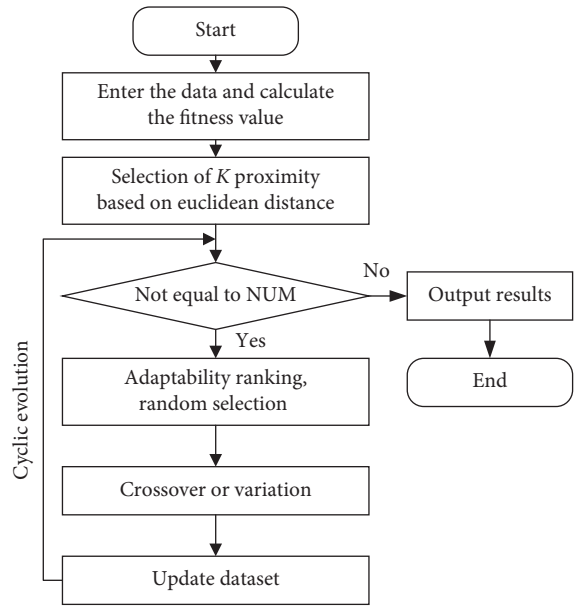

Figure 2: Improved genetic algorithm flowchart.

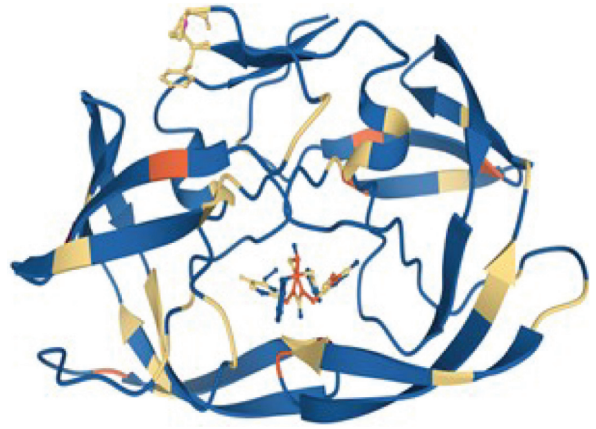

Figure 3: PID is 1RPO HIV-1 protease complex binding crystal.

accuracy increased from 0.547 to 0.647, indicating that the results of RF-SVM for virtual screening of drug proteins were improved when the sampling preprocessing was performed. When the virtual screening simulation experiments were performed using the basic SVM algorithm and the RF-SVM algorithm with SRC kinase as the target protein, the area of AUC increased from 0.741 to 0.828, the enrichment factor also increased from 6.25 to 6.96, and the accuracy also increased from 0.674 to 0.754 after sampling pretreatment. This indicates that, after sampling, the area of AUC increased from 0.722 to 0.769, the enrichment factor increased from 5.87 to 6.43, and accuracy increased from 0.544 to 0.679 when the basic SVM algorithm and RF-SVM algorithm were used for virtual screening simulation experiments before sampling pretreatment. The experimental results show that after pretreatment by sampling, the results of RF-SVM for virtual screening of drug proteins have been improved.

3.3. Experimental Verification. In order to verify the validity of the method proposed in this paper, the experimental results were verified using carbonic anhydrase II (CA II), a zinc-containing metalloenzyme that reversibly mediates the hydration of carbon dioxide, which maintains acid-base balance in blood and other tissues and helps tissues in the body to eliminate carbon dioxide. CA II has a PID of 1Z9Y and a resolution value of 1.66 Å. Its structure is shown in Figure 6.

The experimental results are summarized in Table 3.

The comparative results of the experiment are shown in Figure 7.

The stacking of ROC curves and AUC areas in the experimental results can be obtained: the virtual screening results have improved after sampling by sampling, which can verify the effectiveness of the method proposed in this paper. The AUC increased from 0.717 to 0.775, the enrichment factor increased from 5.83 to 6.36, and the

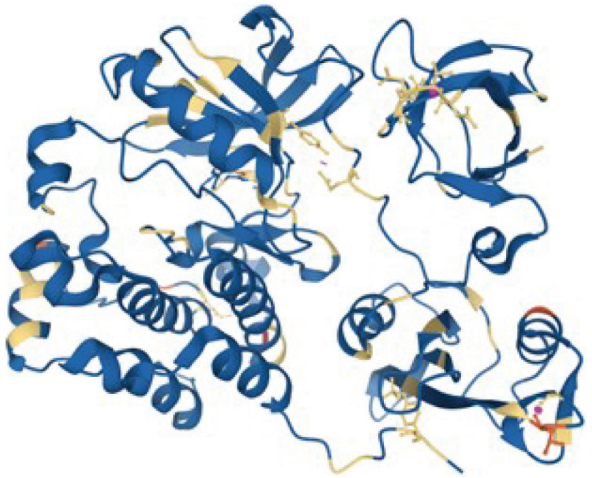

Figure 4: PID is 2H8H SRC kinase complex binding crystal.

Table 1: HIV-1 data classification results.

	Algorithm	Protein name	EF	AUC
Before sampling	SVM	HIV-1	5.93	0.729
Before sampling	RF-SVM	HIV-1	6.51	0.785
After sampling	SVM	HIV-1	6.58	0.795
After sampling	RF-SVM	HIV-1	7.16	0.839

Table 2: SRC data classification results.

	Algorithm	Protein name	EF	AUC
Before sampling	SVM	SRC	5.87	0.722
Before sampling	RF-SVM	SRC	6.43	0.769
After sampling	SVM	SRC	6.25	0.741
After sampling	RF-SVM	SRC	6.96	0.828

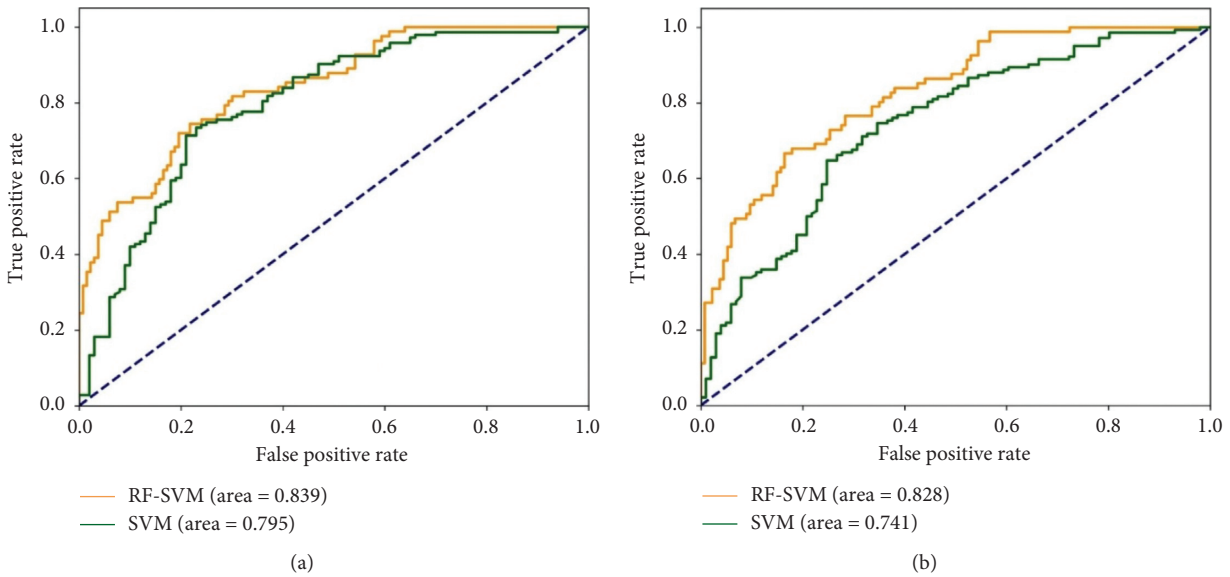

Figure 5: Continued.

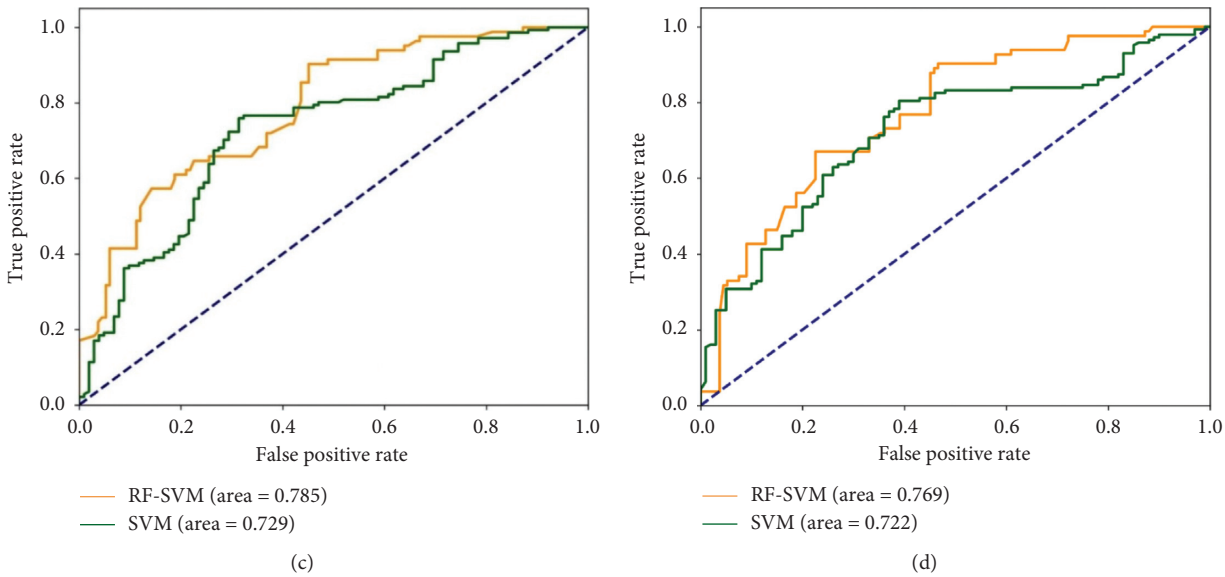

Figure 5: Graph of ROC comparison of two experimental datasets.

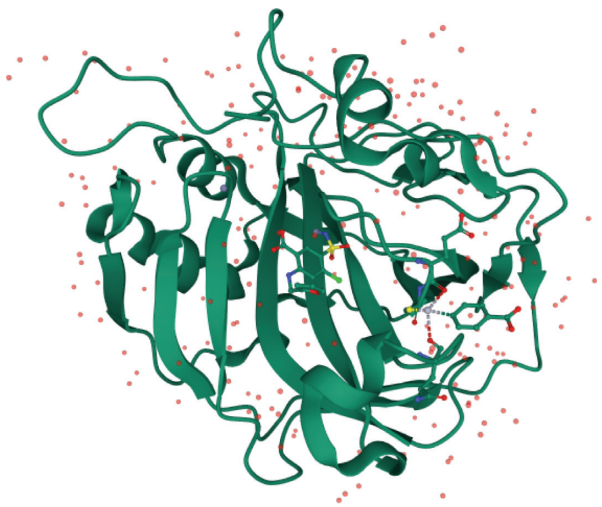

Figure 6: CA II with PID of 1Z9Y.

Table 3: CA II data classification results.

	Algorithm	Protein name	EF	AUC
Before sampling	SVM	CA II	5.83	0.717
After sampling	RF-SVM	CA II	6.36	0.775

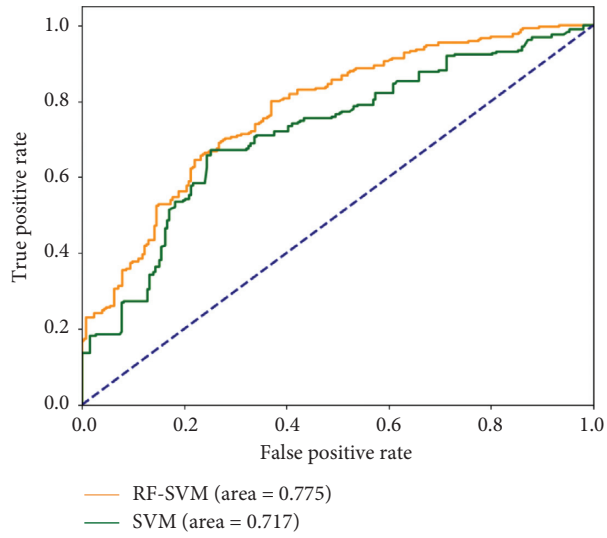

Figure 7: Comparative results of experimental data.

accuracy increased from 0.675 to 0.710 when SVM and RF-SVM were used for classification, indicating that the results of RF-SVM for virtual screening of drug proteins were improved.

4. Conclusion

In this paper, we studied the traditional molecular docking-based virtual screening technology and pointed out some problems of the present virtual screening methods, such as the accuracy of the scoring function and the data imbalance between inactive compounds and active compounds in docked conformations. In order to further improve the screening quality of the lead compounds from the virtual screening, this paper incorporates machine learning methods into the traditional virtual screening process. In order to solve the problem of imbalanced data in realistic virtual screening process, this paper proposes the preprocessing of active compounds by combining genetic algorithm with SMOTE. The solutions proposed in this paper provide new ideas for the study of actual virtual screening techniques. The main findings of this paper include the following.

First, for the situation that the number of active compounds is much less than the number of inactive compounds that occurs in the actual virtual screening process, this paper proposes a preprocessing method to deal with the imbalanced data problem by using a combination of SMOTE algorithm and genetic algorithm to upsample the data of minority class samples to balance the imbalanced dataset by increasing the number of minority class samples. This approach was chosen because the actual number of active compounds in the virtual screening process is very small, and this upsampling approach not only preserves the valid information of the data but also solves the data imbalance problem and avoids overfitting to some extent.

Second, in order to improve the quality of the lead compounds in the virtual screening process, this paper introduces the idea of integrated learning into the virtual screening process of drug proteins and proposes the RF-SVM method. The dataset is optimally selected by support vector machine, and then, the selected dataset is used as the data input of random forest. By this way, the insensitivity of random forest to noise can be effectively avoided, thus improving the generalization ability of the classification algorithm. The simulation experiments of virtual screening were performed by two more important target proteins, HIV-1 protease and SRC kinase, and then, the models were validated by CA II, which showed that the EF of these three datasets improved, on average, by about 0.95, all of which can show that the proposed method in this paper is effective for improving the virtual screening method.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This paper was supported by the Fundamental Research Fundation for Universities of Heilongjiang Province (no. LGYC2018JQ003) and University Nursing Program for Young Scholars with Creative Talents in Heilongjiang Province (no. UNPYSCT-2018208).

References

[1] X. Han, Z. Bao, and H. Zhao, "High throughout screening and selection methods for directed enzyme evolution," *Industrial & Engineering Chemistry Research*, vol. 54, no. 16, pp. 4011–4020, 2015.

[2] U. H. Manjunatha, A. T. Chao, F. J. Leong, and T. T. Diagana, "Cryptosporidiosis drug discovery: opportunities and challenges," *ACS Infectious Diseases*, vol. 2, no. 8, pp. 530–537, 2016.

[3] H. S. Roy, G. Dubey, V. K. Sharma, P. V. Bharatam, and D. Ghosh, "Molecular docking and molecular dynamics to identify collagenase inhibitors as lead compounds to address osteoarthritis," *Journal of Biomedical Structure and Dynamics*, vol. 1, pp. 1–13, 2020.

[4] J. Li, S. Fong, S. Mohammed, and J. Fiaidhi, "Improving the classification performance of biological imbalanced datasets by swarm optimization algorithms," *The Journal of Supercomputing*, vol. 72, no. 10, pp. 3708–3728, 2016.

[5] A. Roy, B. Srinivasan, and J. Skolnick, "PoLi: a virtual screening pipeline based on template pocket and ligand similarity," *Journal of Chemical Information and Modeling*, vol. 55, no. 8, pp. 1757–1770, 2015.

[6] Y. Wang, H. Guo, Z. Feng et al., "PD-1-targeted discovery of peptide inhibitors by virtual screening, molecular dynamics simulation, and surface plasmon resonance," *Molecules*, vol. 24, no. 20, pp. 3784–3791, 2019.

[7] T. Pan, J. Zhao, W. Wu, and J. Yang, "Learning imbalanced datasets based on SMOTE and Gaussian distribution," *Information Sciences*, vol. 512, pp. 1214–1233, 2020.

[8] K. A. Carpenter, D. S. Cohen, J. T. Jarrell, and X. Huang, "Deep learning and virtual drug screening," *Future Medicinal Chemistry*, vol. 10, no. 21, pp. 2557–2567, 2018.

[9] L. Isaias, K. Palacio-Rodríguez, C. N. Cavasotto, and P. Cossio, "Flexi-pharma: a molecule-ranking strategy for virtual screening using pharmacophores from ligand-free conformational ensembles," *Journal of Computer-Aided Molecular Design*, vol. 34, no. 10, pp. 1063–1077, 2020.

[10] A. Kumar and P. Kumar, "Identification of good and bad fragments of tricyclic triazinone analogues as potential PKC-θ inhibitors through SMILES-based QSAR and molecular docking," *Structural Chemistry*, vol. 32, pp. 149–165, 2020.

[11] A. K. Jain, "Data clustering: 50 years beyond *K*-means," *Pattern Recognition Letters*, vol. 31, no. 8, pp. 651–666, 2010.

[12] S. Martellucci, L. Clementi, S. Sabetta, V. Mattei, L. Botta, and A. Angelucci, "SRC family kinases as therapeutic targets in advanced solid tumors: what we have learned so far," *Cancers*, vol. 12, no. 6, pp. 1448–1475, 2020.

[13] M. M. Lawal, Z. K. Sanusi, T. Govender, G. E. M. Maguire, B. Honarparvar, and H. G. Kruger, "From recognition to reaction mechanism: an overview on the interactions between HIV-1 protease and its natural targets," *Current Medicinal Chemistry*, vol. 27, no. 15, pp. 2514–2549, 2020.

[14] M. Kästle, C. Merten, R. Hartig et al., "Tyrosine 192 within the SH2 domain of the SRC-protein tyrosine kinase p56Lck regulates T-cell activation independently of Lck/CD45 interactions," *Cell Communication and Signaling: CCS*, vol. 18, no. 1, p. 183, 2020.

[15] S. Chakraborty, M. Phu, T. Prado de Morais et al., "The PDB database is a rich source of alpha-helical anti-microbial peptides to combat disease causing pathogens," *F1000Research*, vol. 3, p. 295, 2015.