A Framework for Predicting Drug Effectiveness in Human Body

Linda Sara Mathew[1], Anusree P[2], Jislin Anna[3], Saranya V S[4]

[1]Assistant Professor, Dept. of computer science and Engineering, Mar Athanasius College, Kerala, India
[2,3,4]Student, Dept. of computer science and Engineering, Mar Athanasius College, Kerala, India

Abstract - *Drug target interaction prediction has great impact in medical field. It helps to identify new drugs for various diseases. Traditional drug target interaction has many disadvantages. Main disadvantage is costly and time consuming. To overcome this, a new technique machine learning was introduced. Machine learning technique provide an effective and efficient drug target interaction prediction. The pseudo-position specific scoring matrix (PsePSSM) is used to extract the features of protein, and FP2 molecular fingerprinting are used to extract the features of drug. Then combine the drug and target features. Then using Lasso, dimension of the extracted feature information is reduced. Dimensionality reduction is used to avoid over fitting. Based on various analyses it is found that lasso is the most efficient method for dimensionality reduction. To deal with unbalanced data, Synthetic Minority Oversampling Technique (SMOTE) method is used. The processed feature vectors are used as input into a support vector machine classifier for drug target interaction prediction.*

***Key Words*: SMOTE, Lasso, SVM, PsePSSm, FP2**

1. INTRODUCTION

The identification of drug-target interactions plays an important role in the development of medical field as it helps in finding new protein targeted drug and discovering new drug candidates. Targets are specific protein molecules like receptors, enzymes, etc., that are present in the cells of human tissues. They interact with drug molecules and confer the drug effects. Many drugs that are available today have incomplete target proteins. In the past few years, many methods have been introduced to verify drug-target interactions. Most of the methods that were introduced are time-consuming and expensive. Therefore, there is a pressing demand to develop new computational methods that can effectively identify these potential drug-target interactions. Traditional methods to predict drug-target interactions have been divided into the ligand-based methods and target-based methods. Ligand-based methods use ligand similarity to organize pharmacological features and associations between target proteins rather than sequence information or structural information of the target protein. Target-based methods are highly dependent on the accuracy of target structure information. Although more and more potential drug targets and ligands have been discovered, they still haven't proved to be efficient. Therefore, it is necessary to develop a more efficient method for drug-target interaction prediction based on machine learning.

The rapid development of machine learning techniques provides a more effective and efficient method for predicting drug-target interactions. Extraction of protein sequences is an important part of the use of machine learning to predict drug-target interactions. Protein feature extraction methods were mainly based on amino acid sequences. Amino acid sequence-based feature extraction methods mainly include amino acid composition (AAC), dipeptide composition (DC), and pseudo-amino acid composition (PseAAC). The processing of high-dimensional data is a complex issue in the process of predicting drug-target interactions. Excessive dimensions may cause the model to become less efficient. The data dimensionality reduction methods are classified into linear dimensionality reduction and nonlinear dimensionality reduction. The linear dimensionality reduction methods mainly used are principal component analysis (PCA), linear discriminant analysis (LDA) and linear discriminant analysis. We propose a drug-target interaction prediction method based on Lasso dimensionality reduction and support vector machine. First, we generate a pseudo-position specific scoring matrix (PsePSSM) from a position-specific scoring matrix (PSSM). FP2 fingerprint is used to represent each drug compound. Secondly, the positive and negative samples are constructed by using the drug-target interactions information on the extracted features. Then Lasso method is used to select best features from the extracted features. Then SMOTE method is used to balance the imbalanced data. Finally, the processed optimal feature vectors are input into a support vector machine for the prediction of drug-target interactions. Through 5-fold cross-validation, the optimal parameters of the model are selected. Now a drug-target interactions prediction model is established.

2. LITERATURE REVIEW

2.1 Pseudo-position specific scoring matrix (PsePSSM)

Pseudo-position specific scoring matrix(PsePSSM) is used to extract features of amino acid sequence of protiens. PsePSSM specifies the scores for observing a particular amino acid sequence. Based on the extraction of amino acid sequences of proteins on enzymes, ion channels, GPCRs, and nuclear receptors, in order to express the characteristic information in the amino acid sequence. For a target protein sequence *P* with *L* amino acid residues. The position-specific scoring matrix (PSSM) with a dimension of L×20 can be expressed as:

$$P_{PSSM}=\begin{bmatrix} E_{1\to1} & E_{1\to2} & \cdots & E_{1\to j} & \cdots & E_{1\to20} \\ E_{2\to1} & E_{2\to2} & \cdots & E_{2\to j} & \cdots & E_{2\to20} \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ E_{i\to1} & E_{i\to2} & \cdots & E_{i\to j} & \cdots & E_{i\to20} \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ E_{L\to1} & E_{L\to2} & \cdots & E_{L\to j} & \cdots & E_{L\to20} \end{bmatrix}$$

Where j is the 20 native amino acid types , *Ei→j* is the score of the residue of the *i*-th position in the amino acid sequence being mutated to the *j-t*h amino acid residue. A positive score means that the corresponding residue is mutated more frequently than expected, and the negative score is just the opposite. The PsePSSM algorithm converts the protein sequences with different lengths in the dataset into vectors with the same dimension after feature extraction. Each element in the original PSSM matrix is normalized to the interval(0, 1),

For normalization using the below equation.

$$E_{i\to j}=\frac{1}{1+\exp(-E_{i\to j})}$$

2.2 FP2 molecular fingerprint

Based on types of drug properties various types of descriptors are used to represent drug compounds. Based on specific studies it is found that molecular fingerprints can effectively represent the drug compound. Mole file format of a drug compound contains detailed description of the chemical structure of drug. Mol file format of drug molecule are converted into the FP2 format molecular fingerprint using the OpenBabel software. The output is hexadecimal sequence having length of 256, which is then converted to decimal digit sequence ranges between 0 and 15 as a drug molecule 256-dimensional vector.

2.3 Constructing positive and negative samples.

After combining pseudo-position specific matrix and FP2, construct bipartite graph. Bipartite graph represents a network. Each node in the network represents a target protein or drug molecule and each side indicates a drug-target interaction. Bipartite graph contains known and unknown drug-target interactions. All the known interactions are called as positive samples and all the unknown interactions are called as negative samples. Since the number of interacting drug-target pairs is less than the non-interacting drug-target pair, the positive and negative samples are unbalanced. For balancing the positive and negative samples use Synthetic Minority Oversampling Technique(SMOTE) method. SMOTE method perform random oversampling on positive samples and random under sampling on negative samples. So that the positive and negative samples are balanced.

2.4 Lasso method

Least Absolute Shrinkage And Selection Operator(LASSO) is a dimensionality reduction method, which is used to enhance the prediction accuracy. Each drug-target pair contain characteristics of the 476-dimensional data contained. It may contain some noise and redundant information. It will have a negative impact on the performance of the predictive model. We need to remove useless information and extract the most discriminating features from the descriptors of drug-target pairs. So we use the Lasso method to reduce the dimensionality of the original data features. The Lasso method is a compression estimation method. The basic idea of Lasso is to minimize the penalty function under the constraint that the sum of the absolute values of the regression coefficients is less than a constant.

2.5 SMOTE method

There is a high degree of imbalance in the samples used. The number of negative samples is significantly higher than the number of positive samples. So SMOTE method is used to make random oversampling on a small scale of positive samples and random under sampling on a large scale of negative samples so that the positive and negative samples are balanced. For each positive sample *x*, search its *k* nearest neighbors, if the sampling rate is *N*, then randomly select *N* samples from the *k* nearest neighbors, denoted as *y*1, *y*2, ..., *yN*. Random linear interpolation is performed on the line segments formed by *x* and *yi* (*i*=1,2,⋯,*N*) to obtain a new positive sample *zi*, as shown in the following equation:

$$Z_i = x + rand[0,1] * (y_i - x)$$

where *rand*(0,1) means to generate a random number between 0 and 1.

2.6 Support Vector Machine(SVM)

Support Vector Machine (SVM) is a robust classification and regression technique that maximizes the predictive accuracy of a model without overfitting the training data. Support Vector Machine is a machine learning method. It is a supervised learning model. It can be used for both classification and regression problems. SVM suited for two class problems. Using SVM, each data item can plot in n-dimensional space. Here given, n-dimension means number of features you have.

SVM works by mapping data to a high-dimensional feature space so that data points can be categorized, even when the data are not otherwise linearly separable. A separator between the categories is found, then the data are transformed in such a way that the separator could be drawn as a hyperplane. SVM algorithm is used for finding the right hyper plane that differentiates the two classes. For finding right hyper plane use margin, which hyper plane have maximum margin choose that. Margin means the distance between nearest data point. When you have a dataset, select two hyper planes which separate the data with no points between them. Maximize their distance (the margin). The region bounded by the two hyper planes will be the biggest possible margin.

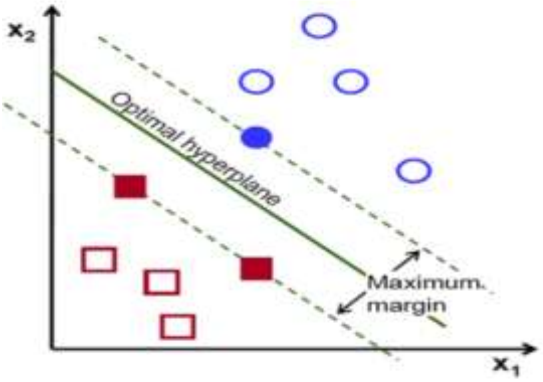

Fig -1: Support Vectors

3. DESIGN

3.1 System Design

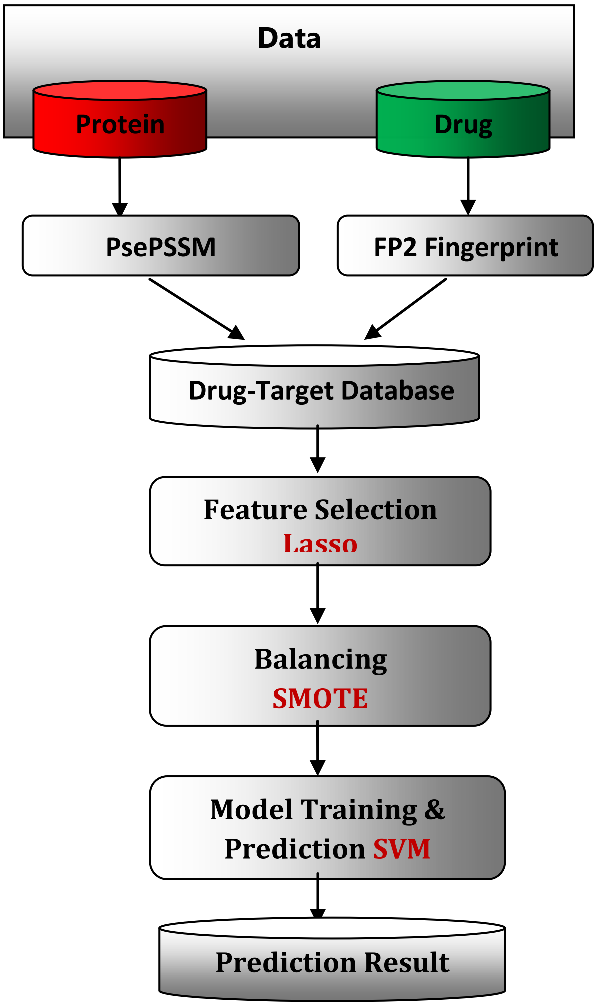

Fig -2: System Design

3.2 Process

Enter the protein sequence and drug Mol file format chemical structure for each of the drug-target datasets. Use Open Babel software to extract FP2 fingerprint features of drug molecules. And the PsePSSM is used to extract features of amino acid sequence of the target protein. Combine features of FP2 and PsePSSM to represent drug-target pairs, generating a dimensional feature vector. Lasso method is used to eliminate the redundant information and noise in the drug-target pair. Using Lasso we can enhance the prediction accuracy. Apply the SMOTE method to balance the positive and negative samples in the drug-target datasets. And the Support Vector Machine(SVM) method is used to predict the accurate drug-target interaction.

The pseudo-position specific scoring matrix (PsePSSM) and fingerprint information is combined with the four drug-target data sets to extract the features of the protein and drug sequences. After combining the drug and target, the redundant information in the dataset is processed based on the Lasso dimensionality reduction method. In the dataset many positive and negative samples are present. The SMOTE method is used to balance the positive and negative samples. Finally, the processed data is given as input to the Support vector machine algorithm to predict the drug-target interaction. PsePSSM feature extraction can generate features that describe the original information of a protein, while FP2 molecular fingerprints describe the molecular structure information. Lasso dimension reduction can remove the redundant information in the drug-target dataset effectively, making the characteristics of the drug-target pairs more obvious. Dimensionality reduction have great significance because it avoids overfiting. Overfitting happens when a model learn the detail and noise in the training data, which causes negative impact on the performance of model. The SMOTE method can avoid over fitting of the model and make full use of valid data. In addition, the selection of optimal parameters in PsePSSM can significantly improve the prediction accuracy rate. Support Vector Machine algorithms can handle two data types, have faster learning speeds, effectively handle noise data, and build highly accurate classifiers. SVM algorithm maps the data to a high-dimensional feature space.

4. RESULTS

4.1 PsePSSM method selection

For prediction of drug-target interaction, we need to extract features of amino acid sequence from target protein. PsePSSM is used to extract the features of amino acid sequence. Here given four datasets of target protein are enzymes, ion channels, GPCRs, nuclear receptors. Features of amino acid sequence are extracted from these four datasets of protein. PsePSSM specifies the scores for observing a particular amino acid sequence. The selection of features of amino acid sequence can directly affect the prediction accuracy of the model. using the PsePSSM method to extract features of each target protein can get 220-dimensional feature vectors.

4.2 Effect of dimensionality reduction

The dimensionality reduction algorithm is used to remove redundant information and noise in the feature vector, which can improve the prediction accuracy to some extent. Based on the 220-dimensional feature vectors obtained in PsePSSM and the 256-dimensional feature vectors generated after extracting the FP2 molecular fingerprint, selecting an appropriate method to improve the accuracy rate. For the dimensionality reduction, here we use Lasso method. The Lasso method has more advantages than the other dimensionality reduction method. Using the Lasso method to reduce dimension can effectively reduce information redundancy and delete some unimportant features, which helps to avoid overfitting of the model and reduce the complexity of the training model. So, Lasso is the best dimensionality reduction algorithm.

4.3 Imbalanced and balanced dataset

After combining the drug and target pair, there is a problem of imbalanced dataset. Known and unknown interactions are occur. The number of drug-target pairs that their interactions are known is smaller than the number of drug target pairs that are not interacting with each other, which will lead to a decrease in the prediction accuracy of prediction models. For balancing the dataset, use SMOTE method. In SMOTE method random oversampling on a small scale of positive samples and random undersampling on a large scale of negative samples so that the positive and negative samples are balanced. So, this method is very helpful to balancing the dataset.

4.4 Prediction of drug-target interaction

Many prediction methods are proposed for the prediction of drug-target interaction. we compared the prediction performance with the other prediction method using the same datasets. This paper use Support vector machine(SVM) algorithm for the drug-target interaction prediction.

SVM works by mapping data to a high-dimensional feature space. SVM algorithm is used for finding the right hyper plane. Margin is used for finding the right hyper plane. Margin which means maximize the distance between the data points. This method give better performance than the other method.

5. CONCLUSION

Identification of drug target interaction discover new drug candidate. But the relationship between drug and target is very difficult to understand. Identification of Traditional drug target interaction is time consuming and expensive. So, we use machine learning methods for the identification of drug target interaction. Through this discover new unknown drug target pair. For the drug target interaction prediction, using four datasets enzyme, ion channel(IC), G-protein-coupled receptor (GPCR) and nuclear receptor(NR).

Here, we use pseudo-position specific scoring matrix (PsePSSM) for extract the features of protein amino acid sequence. And also use FP2 molecular fingerprint for representing the drug compound. Then combining the extracted features of drug and protein. For removing redundant information in the dataset, using Lasso dimensionality reduction method. SMOTE method used to balance the positive and negative samples. Positive samples means known target interaction and Negative samples means unknown target interaction. With the help of smote method can avoid over fitting of the model and improve generalization performance. After Smote method, use SVM for dug target interaction prediction. SVM give better performance for prediction. Compare with the other prediction model, this model is very effective and efficient.

REFERENCES

[1] M.A. Yildirim, K.I. Goh, M.E. Cusick, A.L. Barabasi, M. Vidal, Drug-target network,Nat. Biotechnol. 25 (10) (2007)

[2] S.C. Janga, A. Tzakos, Structure and organization of drug-target networks: insights from genomic approaches for drug discovery, Mol. BioSyst. 5 (2009)

[3] M. Kuhn, M. Campillos, P. Bork, I.P. Gonza, L.J. Jensen, Large-scale prediction of drug-target relationships, FEBS Lett. 582 (8) (2008)

[4] Y.C. Wang, Z.X. Yang, Y. Wang, N.Y. Deng, Computationally probing drug-protein interactions via support vector machine, Lett. Drug Des. Discov. 7 (2010) 370–378.

[5] [5] Z. Xia, L.Y. Wu, X. Zhou, S.T.C. Wong, Semi-supervised drug-protein interaction prediction from heterogeneous biological spaces, BMC Syst. Biol. 4 (2010) S6.

[6] M. Hao, Y. Wang, S.H. Bryant, Improved prediction of drug-target interactions using regularized least squares integrating with kernel fusion technique, Anal. Chim. Acta 909 (2016)

[7] M. Takarabe, M. Kotera, Y. Nishimura, S. Goto, Y. Yamanishi, Drug target prediction using adverse event report systems: a pharmacogenomic approach, Bioinformatics 28 (18) (2012)

[8] Y.F. Dai, X.M. Zhao, A survey on the computational approaches to identify drug targets in the postgenomic era, Biomed. Res. Int. 239654 (2015)

[9] A. Ezzata, M. Wu, X. Li, C.K. Kwoh, Drug-target interaction prediction using ensemble learning and dimensionality reduction, Methods 129 (2017) 81–88.