JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 25, Number 4, 2018
© Mary Ann Liebert, Inc.
Pp. 435–443
DOI: 10.1089/cmb.2017.0149

Prediction of Central Nervous System Side Effects Through Drug Permeability to Blood–Brain Barrier and Recommendation Algorithm

JUN FAN,[1,2] JING YANG,[2] and ZHENRAN JIANG[2]

ABSTRACT

Drug side effects are one of the public health concerns. Using powerful machine-learning methods to predict potential side effects before the drugs reach the clinical stages is of great importance to reduce time consumption and protect the security of patients. Recently, researchers have proved that the central nervous system (CNS) side effects of a drug are closely related to its permeability to the blood–brain barrier (BBB). Inspired by this, we proposed an extended neighborhood-based recommendation method to predict CNS side effects using drug permeability to the BBB and other known features of drug. To the best of our knowledge, this is the first attempt to predict CNS side effects considering drug permeability to the BBB. Computational experiments demonstrated that drug permeability to the BBB is an important factor in CNS side effects prediction. Moreover, we built an ensemble recommendation model and obtained higher AUC score (area under the receiver operating characteristic curve) and AUPR score (area under the precision-recall curve) on the data set of CNS side effects by integrating various features of drug.

Keywords: blood–brain barrier, central nervous system, drug side effects, recommender system.

1. INTRODUCTION

Drugs can help patients in treating different diseases, but they are usually accompanied by a number of potential side effects. Since the side effects of drugs may bring out failures in drug development or drug withdrawal, it is a critical issue to identify potential side effects of drugs. Considering the cost and time consumption in wet experiments, computational methods were usually proposed to predict drug side effects.

In recent years, many machine-learning methods were popularly adopted in the study of drug side effects prediction due to the powerful learning abilities. To utilize the machine-learning methods to predict side effects, extracting the features of drugs is a necessary and important step. Huang et al. (2011) utilized drug targets, protein–protein interaction networks, and gene ontology annotations as features, and adopted support vector machine (SVM) and logistic regression as classification algorithms. Pauwels et al. (2011) used chemical structures and four machine-learning methods (SVM, k-nearest neighbor, sparse canonical correlation analysis, and ordinary canonical correlation analysis) to build prediction models. Mizutani et al.

[1] Shanghai Key Laboratory of Multidimensional Information Processing, East China Normal University, Shanghai, China.
[2] Department of Computer Science and Technology, East China Normal University, Shanghai, China.

(2012) made use of chemical substructures and target proteins about drugs, and applied sparse canonical correlation analysis for side effects prediction. Liu et al. (2012) integrated a large number of features, including phenotypic information, chemical information, and biological information, and used different classifiers (naive Bayes, k-nearest neighbor, logistic regression, SVM, and random forest) to build prediction models. Huang et al. (2013) combined chemical structures and protein–protein interactions as feature profiles, and applied SVM for prediction. Zhang et al. (2015) integrated a variety of drug-related information as features, and adopted multilabel k-nearest neighbor method to build side effects prediction models. And Zhang et al. (2016) proposed a drug side effect prediction model through linear neighborhoods and multiple data source integration.

Recently, Gao et al. (2017) used the central nervous system (CNS) side effects as a feature to predict drug permeability to the blood–brain barrier (BBB) and improved the accuracy of prediction. The BBB is a highly selective semipermeable membrane barrier that separates the circulating blood from the brain extracellular fluid in the CNS, and it also protects the brain from most pathogens. Gao's research proved that the CNS side effects of drugs are closely related to drug permeability to the BBB. Inspired by this, it is reasonable to relate drug permeability to the BBB with CNS side effects prediction. In addition, due to the dimensionality and sparsity of data set of CNS side effects, we make efforts to solve the problem in the frame of recommender system.

In this article, we considered the drug–CNS side effects prediction as a user items recommender system, and our task was to recommend CNS side effects to a given drug. Therefore, we proposed an extended neighborhood-based recommendation method (ENRM) by considering the relationship between CNS side effects and drug permeability to the BBB. Computational experiments demonstrate that on our data set of CNS side effects, the information about drug permeability to the BBB can effectively improve the accuracy of CNS side effects prediction. Moreover, although a large number of drug-related profiles were extracted as features, each feature has a difference in the validity of the prediction. To integrate different features of drugs, we built an ensemble model with ENRM. Compared with the state-of-the-art methods, our ensemble model obtained higher AUC scores and AUPR scores on the data set of CNS side effects.

2. MATERIALS AND METHODS

2.1. Data set

This study focuses on 201 drugs that have log[brain]/[blood] (logBB) data obtained from different academic articles (Subramanian and Kitchen, 2003; Winkler and Burden, 2004; Li et al., 2005; Abraham et al., 2006; Wang et al., 2015), and 476 CNS side effects extracted from SIDER database (Kuhn et al., 2010). The drug side effects in SIDER database are formatted according to the Medical Dictionary for Regulatory Activities (MedDRA). To verify the effectiveness of the BBB permeability (BP) for CNS side effects prediction, we extracted 476 CNS side effects under the SOC (system organ classes) of nervous system disorders and psychiatric disorders according to MedDRA.

About the drug information, there are several public databases. PubChem Compound Database (Wang et al., 2009) contains validated chemical depiction information to describe substances. KEGG (Kanehisa et al., 2009) is a database resource for approved drugs in Japan, United States, and Europe. DrugBank database (Law et al., 2013) is a bioinformatics and cheminformatics resource that combines detailed drug data with comprehensive drug target information. Table 1 displays the details of drug features in this study.

Table 1. Details of Drug Features in This Study

Feature	*Dimension*	*Source*
Substructures	510	PubChem
Enzymes	100	DrugBank
Targets	414	DrugBank
Transporters	55	DrugBank
Pathways	235	KEGG
Indications	1224	SIDER

2.2. Methods

2.2.1. Problem definition. Recommender system is a kind of information filtering system that seeks to recommend items (music, movies, books, etc.) to users by predicting the ''ratings'' that users would give to the items.

Predicting the CNS side effects of a new drug can be regarded as a task of recommending items to a user. Given a data set with *n* drugs and *m* CNS side effects, we can construct an $n \times m$ adjacent matrix based on the association of drugs and CNS side effects. In the adjacent matrix *A*, if A_{ij} equals 1, it means that the *i*th drug has *j*th CNS side effects, on the contrary, it is opposite. For a new drug, our goal is to recommend some CNS side effects with probabilities, and the results are represented as $P = \{p_{0,} p_1, p_2, \ldots, p_m\}$, in which p_j indicates the probability that the drug has the *j*th CNS side effects.

2.2.2. Extended neighborhood-based recommendation method. The neighborhood-based recommendation method (NRM) (Su and Khoshgoftaar, 2009) is one of the most popular recommendation algorithms, which recommends items according to preferences of similar users. In recommender systems, the neighbors are determined by the behaviors of users. However, for a new user without history behaviors, how to find its neighbors is a key issue called ''Cold Start Problem'' (Schein et al., 2002). To solve this problem, we always make use of the profiles of users, such as ages, hobbies, and professions.

In drug side effects prediction, we built a drug side effects recommender system, and profiles of drugs were represented by several drug information (*substructures*, *enzymes*, *targets*, *transporters*, *pathways*, and *indications*). Next, we encoded the profiles as binary vectors and made use of them to calculate drug–drug similarity. To verify the effectiveness of the BP on CNS side effects prediction, we analyzed the correlation between the BP and CNS side effects, and proposed an ENRM.

As shown in Figure 1, when making prediction for a new drug *d*, we first calculate drug–drug similarity and find its k neighbors *N(d)* with feature vectors. Here, we use the cosine similarity:

$$S(i, j) = cos\left(v_i, v_j\right) = \frac{v_i \cdot v_j}{\| v_i \| \times \| v_j \|}. \tag{1}$$

Next, the k neighbors recommend CNS side effects denoted as *C(d)*, which are associated with the k neighbors. Let *A* represent the adjacent matrix of association between drugs and CNS side effects in the training set. Each recommended CNS side effect has a recommendation score, defined as

$$Score(c) = \frac{\sum_{n \in N(d)} A_{nc}}{k}, \ c \in C(d). \tag{2}$$

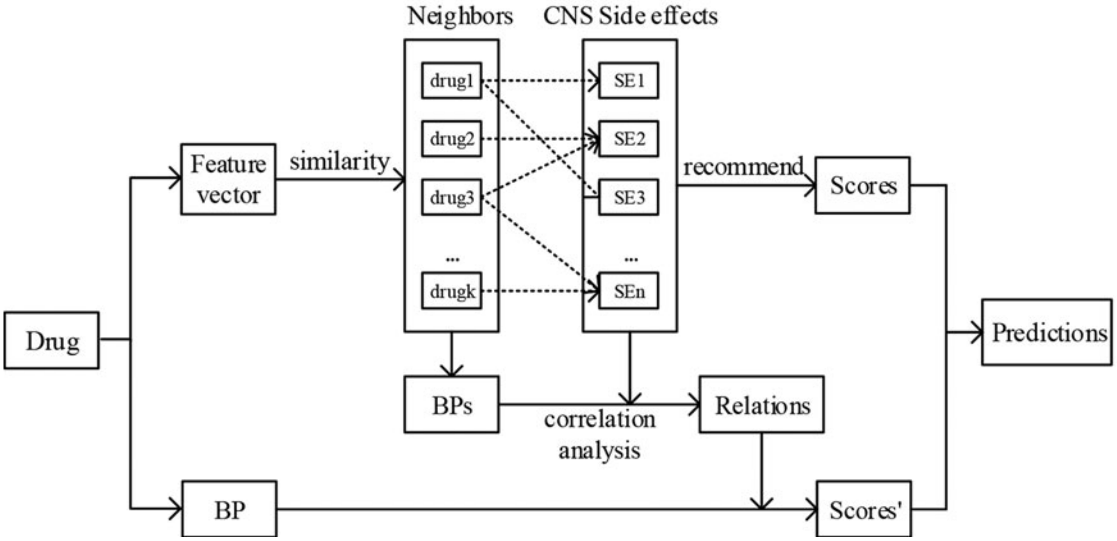

FIG. 1. Structure of extended neighborhood-based recommendation method.

To make use of the drug BP, we correlate the neighbors' permeability with recommended CNS side effects, and the relations R is determined by the following principle:

$$R_0^c = P(H_c|E_0) = \frac{\sum_{n \in N(d)} A_{nc}(1 - B_n)}{\sum_{n \in N(d)} 1 - B_n}, \; c \in C(d), \tag{3}$$

$$R_1^c = P(H_c|E_1) = \frac{\sum_{n \in N(d)} A_{nc} B_n}{\sum_{n \in N(d)} B_n}, \; c \in C(d), \tag{4}$$

where H_c denotes the event that a drug has CNS side effect c. Let E_1 represent the event that the drug can penetrate the BBB, while E_0 is the opposite. B_n indicates the permeability of neighbor n. For CNS side effect c, if drug d cannot penetrate the BBB, the Score$'$ (c) equals R_0^c, else, it equals R_1^c. Finally, we combine *Score* and *Score$'$* to obtain the prediction:

$$Prediction = \lambda \cdot Score + (1 - \lambda) \cdot Score'. \tag{5}$$

After several trials of contrast, we set $k = 10$ and $\lambda = 0.7$.

2.2.3. Ensemble recommendation method. There are several different drug-related features for current CNS side effects prediction. Therefore, we attempt to integrate various valuable features to achieve better performances. In machine learning, ensemble learning (Dietterich, 2000) is a methodology that has an ability to combine different features, and can get better results in many fields. Here, we designed an ensemble recommendation method based on ENRM. The flowchart of the ensemble recommendation method is shown in Figure 2.

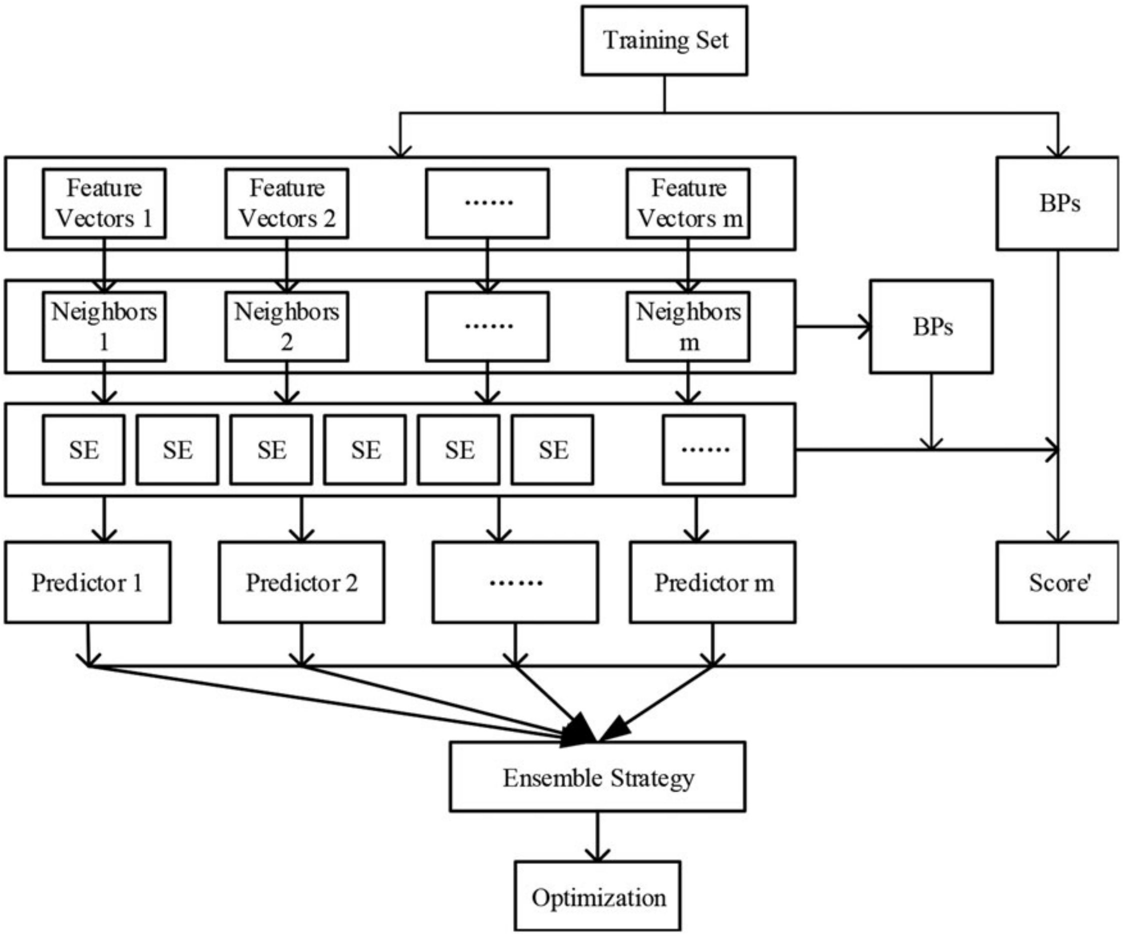

FIG. 2. Flowchart of ensemble recommendation method.

Given a training set T and m features, we find m groups of k neighbors, and make m predictors according to ENRM. Since different features make different contributes in CNS side effect predictions, we adopt weighted scoring ensemble strategy, and the ensemble predictions are produced by the following function:

$$Pred = \sum_{i=1}^{m} w_i(\lambda \cdot Predictor(i) + (1-\lambda) \cdot Score'), \tag{6}$$

where w_i denotes the weight of ith features. Let A_t represent the actual CNS side effects of drug t, the value of w is calculated by optimization.

$$[w_1, w_2, \dots w_m] = argmin \sum_{t \in T} \| A_t - Pred_t \| . \tag{7}$$

3. RESULTS AND DISCUSSION

This article adopted 10-fold cross-validation (10-CV) to test the performances of models. For the data set of CNS side effects, we randomly split all drugs into 10 subsets with equal size. Each time, let nine subsets be the training set, and the remaining one is the test set. About the metrics, we used the AUC and the AUPR to evaluate models.

3.1. Effectiveness of the BP

To verify the effectiveness of the BP on CNS side effects prediction, we proposed ENRM. By taking the BP into consideration, we obtained several experimental results using different features. The results compared with NRM are listed in Table 2.

The NRM and ENRM independently used different drug-related profiles (*substructures*, *enzymes*, *targets*, *transporters*, *pathways*, and *indications*) as features to identify neighbors and make recommendations. In NRM, the recommendation scores only depended on the k neighbors. However, whether a drug penetrates the BBB is an important factor in whether the drug has the side effect. By analyzing the correlation between the BP and CNS side effects, ENRM combined *Score* and *Score′* and achieved 2.37%–5.37% improvements on AUC and 1.17%–9.88% improvements on AUPR.

3.2. Performance of ensemble recommendation method

According to Table 2, we can see that different features may make different performances in CNS side effect predictions, and the ‘‘indications’’ appeared to be the most informative (highest AUC of 0.8794 and highest AUPR of 0.4831) for prediction, and ‘‘transporters’’ and ‘‘enzymes’’ achieved similar AUC and AUPR. Although some features lead to relatively lower AUC scores and AUPR scores, we cannot neglect the information of them. To integrate these features and achieve better performance in experiments, we built an ensemble recommendation model. In the ensemble recommendation model, we regarded recommendation

Table 2. The Average of 10-Fold Cross-Validation Results of Neighborhood-Based Recommendation Method and Extended Neighborhood-Based Recommendation Method

	NRM		*ENRM*	
Features	*AUC*	*AUPR*	*AUC*	*AUPR*
Substructures	0.8275	0.4179	0.8719	0.4497
Enzymes	0.8273	0.4081	0.8641	0.4425
Targets	0.8499	0.4477	0.8743	0.4671
Transporters	0.8214	0.3969	0.8645	0.4361
Pathways	0.8402	0.4333	0.8738	0.4625
Indications	0.8590	0.4775	0.8794	0.4831

AUC, area under receiver operating characteristic curve; AUPR, area under the precision–recall curve; ENRM, extended neighborhood-based recommendation method; NRM, neighborhood-based recommendation method.

Table 3. The Top 20 Recommended Drug–Central Nervous System Side Effects Interactions

DrugID (PubChem)	*Drug name*	*CNS side effect*	p	*Validated database*
4539	Norfloxacin	Ataxia	0.9204	SIDER;PubChem
5556	Triazolam	Ataxia	0.9132	SIDER
2764	Ciprofloxacin	Hallucination	0.9092	SIDER;PubChem
5391	Temazepam	Hallucination	0.9063	SIDER
4192	Midazolam	Hallucination	0.9061	SIDER;PubChem
5257	Sparfloxacin	Tremor	0.9043	SIDER
2118	Alprazolam	Tremor	0.9037	SIDER;PubChem
3016	Diazepam	Tremor	0.9005	SIDER;PubChem
3345	Fentanyl	Headache	0.8984	SIDER;PubChem
2269	Azithromycin	Headache	0.8975	SIDER;PubChem
3826	Ketorolac	Headache	0.8943	SIDER;PubChem
5408	Testosterone	Headache	0.8917	SIDER
5215	Sulfadiazine	Headache	0.8906	SIDER
444	Bupropion	Headache	0.8883	SIDER;PubChem
4205	Mirtazapine	Headache	0.8863	SIDER
4196	Mifepristone	Headache	0.8851	SIDER
298	Chloramphenicol	Headache	0.8839	SIDER;PubChem
5257	Sparfloxacin	Headache	0.8784	SIDER
4927	Promethazine	Headache	0.8753	SIDER
4889	Pravastatin	Headache	0.8721	SIDER

CNS, central nervous system.

results based on different features as predictors. With weight parameter *w*, we combined each individual predictor linearly to get the final recommendation score. Through 10-CV test, the ensemble recommendation model obtained an AUC of 0.9013 and an AUPR of 0.5142. The ROC curves and AUPR scores of individual features and integrated features are shown in Figures 3 and 4, respectively.

3.3. Comparison with benchmark methods

To the best of our knowledge, some state-of-the-art methods are always compared as benchmark methods. Pauwels et al. (2011) used sparse canonical correlation analysis to build prediction models. Liu

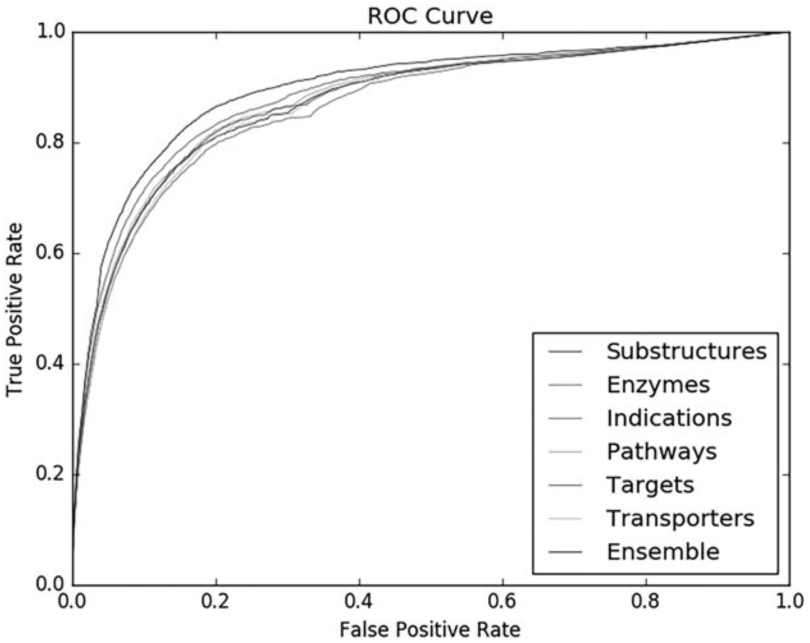

FIG. 3. ROC curves of individual features and integrated feature.

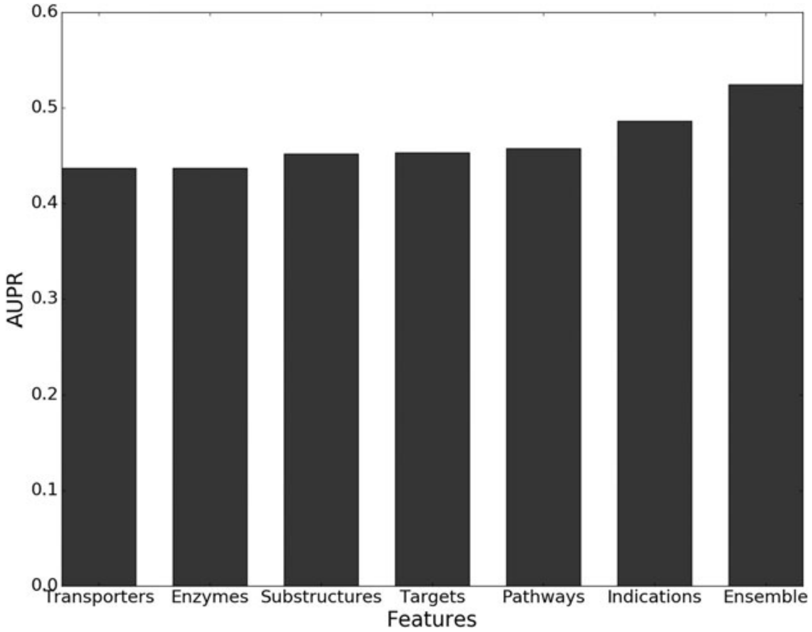

FIG. 4. AUPR scores of individual features and integrated feature. AUPR, area under the precision–recall curve.

et al. (2012) made a good performance using SVM classifier. Zhang et al. (2015) proposed feature selection-based multilabel k-nearest neighbor (FS-MLKNN) method to predict side effects. In this article, these three methods are compared as benchmark methods and the experiment results are obtained with their source code using default parameters. The ROC (receiver operating characteristic) curves and PR (precision-recall) curves of different methods are shown in Figures 5 and 6.

As shown in Figures 5 and 6, among the three benchmark methods, Liu's method obtained a low performance on AUC scores due to the sparsity of data set, whereas in his article, the data set was pretreated (each side effect was associated with >50 drugs) to satisfy the SVM classifier, which is affected by the proportion of positive and negative samples. However, neighborhood-based methods can avoid this

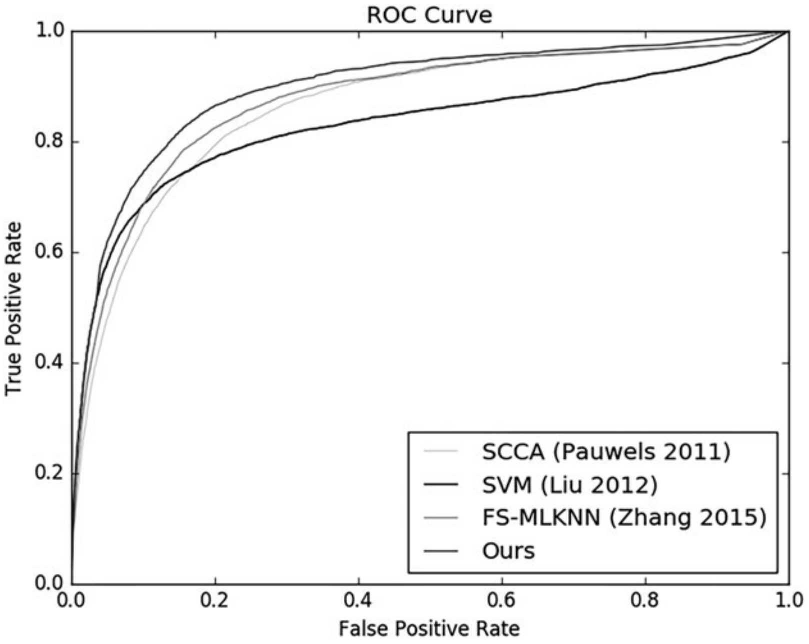

FIG. 5. ROC curves of four methods.

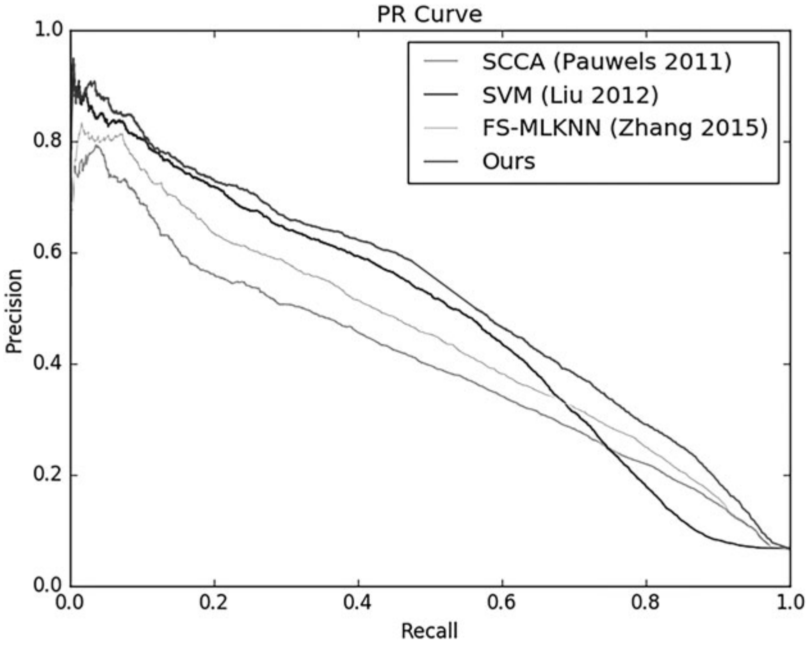

FIG. 6. PR curves of four methods.

problem; therefore, FS-MLKNN had a good performance (AUC of 0.8741 and AUPR of 0.4486) on the data sets of all side effects. In addition, FS-MLKNN also integrates several drug features to predict side effects, but it did not take the BP into account. Our ensemble recommendation model added the analysis of relationship between side effects and the BP on the basis of NRM, and improved the performance of side effects prediction. The results also demonstrated that the BBB is an interesting part of the brain, whose functional mechanism that mainly keeps things out of the brain may provide clues for disease treatment.

To validate the recommended CNS side effects, we obtained the top 20 recommendation results, and confirmed them in SIDER database and PubChem database. The information is listed in Table 3.

4. CONCLUSION

In this article, we proposed an ENRM to predict CNS side effects, considering drug permeability to the BBB. Experiment results demonstrated that the drug permeability to the BBB is effective to improve the AUC score and AUPR score of CNS side effects prediction. Furthermore, we built an ensemble recommendation model to integrate different features for CNS side effects prediction. Compared with benchmark methods, our ensemble recommendation model obtained a better AUC score and AUPR score on our data set.

Although our method has a good performance on our data set, there is still tremendous room for improvement. There is a concurrent relationship between side effects, which we can use to further improve the side effects prediction. In addition, deep learning techniques has better performances in various prediction situations than traditional machine-learning methods. Therefore, our future study will focus on the use of concurrent relationship between side effects and applying deep learning techniques for side effects prediction.

ACKNOWLEDGMENTS

This research is funded by the National Key Technology Support Program (No. 2015BAH12F01) and the National Nature Science Foundation of China (No. 61602179) and the Shanghai Municipal Commission of Economy and Information (No. 201602024) and the Xiaoi Research.

AUTHOR DISCLOSURE STATEMENT

No competing financial interests exist.

REFERENCES

Abraham, M.H., Ibrahim, A., Zhao, Y., et al. 2006. A data base for partition of volatile organic compounds and drugs from blood/plasma/serum to brain, and an LFER analysis of the data. *J. Pharm. Sci.* 95, 2091–2100.

Dietterich, T.G. 2000. Ensemble methods in machine learning. *Mult. Classif. Syst*. 1857, 1–15.

Gao, Z., Chen, Y., Cai, X., et al. 2017. Predict drug permeability to blood–brain-barrier from clinical phenotypes: Drug side effects and drug indications. *Bioinformatics* 33, 901–908.

Huang, L.C., Wu, X., and Chen, J.Y. 2011. Predicting adverse side effects of drugs. *BMC Genomics* 12, S11.

Huang, L.C., Wu, X., and Chen, J.Y. 2013. Predicting adverse drug reaction profiles by integrating protein interaction networks with drug structures. *Proteomics* 13, 313–324.

Kanehisa, M., Goto, S., Furumichi, M., et al. 2009. KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res*. 38(suppl_1), D355–D360.

Kuhn, M., Campillos, M., Letunic, I., et al. 2010. A side effect resource to capture phenotypic effects of drugs. *Mol. Syst. Biol.* 6, 343.

Law, V., Knox, C., Djoumbou, Y., et al. 2013. DrugBank 4.0: Shedding new light on drug metabolism. *Nucleic Acids Res*. 42(D1), D1091–D1097.

Li, H., Yap, C.W., Ung, C.Y., et al. 2005. Effect of selection of molecular descriptors on the prediction of blood–brain barrier penetrating and nonpenetrating agents by statistical learning methods. *J. Chem. Inf. Model.* 45, 1376–1384.

Liu, M., Wu, Y., Chen, Y., et al. 2012. Large-scale prediction of adverse drug reactions using chemical, biological, and phenotypic properties of drugs. *J. Am. Med. Inform. Assoc.* 19(e1), e28–e35.

Mizutani, S., Pauwels, E., Stoven, V., et al. 2012. Relating drug–protein interaction network with drug side effects. *Bioinformatics* 28, i522–i528.

Pauwels, E., Stoven, V., and Yamanishi, Y. 2011. Predicting drug side-effect profiles: A chemical fragment-based approach. *BMC Bioinformatics* 12, 169.

Schein, A.I., Popescul, A., Ungar, L.H., et al. 2002. Methods and metrics for cold-start recommendations, 253–260. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, Tampere, Finland.

Su, X., and Khoshgoftaar, T.M. 2009. A survey of collaborative filtering techniques. *Adv. Artif. Intell.* 2009, 4.

Subramanian, G., and Kitchen, D.B. 2003. Computational models to predict blood–brain barrier permeation and CNS activity. *J. Comp. Aided Mol. Des.* 17, 643–664.

Wang, W., Kim, M.T., Sedykh, A., et al. 2015. Developing enhanced blood–brain barrier permeability models: Integrating external bio-assay data in QSAR modeling. *Pharm. Res*. 32, 3055–3065.

Wang, Y., Xiao, J., Suzek, T.O., et al. 2009. PubChem: A public information system for analyzing bioactivities of small molecules. *Nucleic Acids Res*. 37(suppl_2), W623–W633.

Winkler, D.A., and Burden, F.R. 2004. Modelling blood–brain barrier partitioning using Bayesian neural nets. *J. Mol. Graph. Model*. 22, 499–505.

Zhang, W., Chen, Y., Tu, S., et al. 2016. Drug side effect prediction through linear neighborhoods and multiple data source integration, 427–434. In *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, Shenzhen, China.

Zhang, W., Liu, F., Luo, L., et al. 2015. Predicting drug side effects by multi-label learning and ensemble learning. *BMC Bioinformatics* 16, 365.

Address correspondence to:
Dr. Jing Yang
Department of Computer Science and Technology
East China Normal University
Shanghai 200262
China

E-mail: jyang@cs.ecnu.edu.cn

Dr. Zhenran Jiang
Department of Computer Science and Technology
East China Normal University
Shanghai 200262
China

E-mail: zrjiang@cs.ecnu.edu.cn