Mathematical Biosciences and Engineering

MBE, 20(1): 383–401.
DOI: 10.3934/mbe.2023018
Received: 29 August 2022
Revised: 26 September 2022
Accepted: 28 September 2022
Published: 09 October 2022

http://www.aimspress.com/journal/MBE

Research article

A model with deep analysis on a large drug network for drug classification

Chenhao Wu and Lei Chen*

College of Information Engineering, Shanghai Maritime University, Shanghai 201306, China

* **Correspondence:** Email: chen_lei1@163.com; Tel: +862138282825; Fax: +862138282800.

Abstract: Drugs are an important means to treat various diseases. They are classified into several classes to indicate their properties and effects. Those in the same class always share some important features. The Kyoto Encyclopedia of Genes and Genomes (KEGG) DRUG recently reported a new drug classification system that classifies drugs into 14 classes. Correct identification of the class for any possible drug-like compound is helpful to roughly determine its effects for a particular type of disease. Experiments could be conducted to confirm such latent effects, thus accelerating the procedures for discovering novel drugs. In this study, this classification system was investigated. A classification model was proposed to assign one of the classes in the system to any given drug for the first time. Different from traditional fingerprint features, which indicated essential drug properties alone and were very popular in investigating drug-related problems, drugs were represented by novel features derived from a large drug network via a well-known network embedding algorithm called Node2vec. These features abstracted the drug associations generated from their essential properties, and they could overview each drug with all drugs as background. As class sizes were of great differences, synthetic minority over-sampling technique (SMOTE) was employed to tackle the imbalance problem. A balanced dataset was fed into the support vector machine to build the model. The 10-fold cross-validation results suggested the excellent performance of the model. This model was also superior to models using other drug features, including those generated by another network embedding algorithm and fingerprint features. Furthermore, this model provided more balanced performance across all classes than that without SMOTE.

Keywords: drug classification; drug network; chemical-chemical interaction; Node2vec; random forest; support vector machine

1. Introduction

Disease is one of the greatest threats to human health. Many people die due to various diseases each year. With the development of medical science, several efforts have been given to various diseases [1–3], with an ultimate purpose of uncovering the underlying mechanism and the essential characteristics of different diseases to design efficient treatments. To date, several treatments have been designed to deal with different diseases, and drug is deemed as one of the most important parts. For fast indication of the utility of drugs, they are divided into several classes. Drugs in one class always have some common features. Correctly identifying the class of a candidate drug-like compound is helpful to uncover its effects. The classification procedure could be completed by traditional experiments. However, such methods are always inefficient and expensive. Designing quick and cheap methods that could accurately predict the class of a given drug is urgent. With the development of computer science, many advanced computational methods have been proposed in recent years and applied to deal with various practical problems, such as artificial neural networks [4–6], feature selection algorithms [7], statistical test methods [8–10]. These methods provide strong support for designing efficient methods.

Designing computational methods to classify drugs have recently become relatively popular. The most representative studies focused on predicting drug classes in the Anatomical Therapeutic Chemical (ATC) classification system. Such system is recommended and maintained by the World Health Organization (WHO). It has five levels, and several classes are contained in each level. Many computational methods have been proposed to predict classes in the first level of the ATC classification system. Most of them were based on machine learning algorithms, such as deep learning algorithms [11–14], network embedding algorithms [15], and multi-label classification algorithms [15–17]. Meanwhile, several types of drug properties were adopted to build methods that include drug fingerprints [16,17] and chemical–chemical similarity/interaction information [13,18–21]. These methods provide helpful insights into investigating other drug classification systems.

The Kyoto Encyclopedia of Genes and Genomes (KEGG) [22] recently reported a new drug classification system that classified drugs into the following 14 classes: (I) anti-allergic agent; (II) anti-inflammatory; (III) antibacterial; (IV) antidiabetic agent; (V) antifungal; (VI) antineoplastic; (VII) antiviral; (VIII) cardiovascular agent; (IX) endocrine and hormonal agent; (X) gastrointestinal agent; (XI) hypolipidemic agent; (XII) immunological agent; (XIII) neuropsychiatric agent; (XIV) ophthalmic agent. Under this system, drugs are roughly classified in accordance with the diseases they could treat. Identifying the class of any latent drug-like compound is helpful to preliminarily determine its effects to a certain type of disease. Targeted experiments for this type of disease could subsequently be conducted for further confirmation, which could accelerate the procedures for discovering novel drugs. Thus, for the first time, such a drug classification system was investigated in this study by developing a classification model, which could assign one of the above 14 classes to given drugs. A large drug network was constructed on the basis of chemical–chemical interaction (CCI) information reported in Search Tool for Interactions of Chemicals (STITCH) to extract informative drug features [23,24]. It was fed into a powerful network embedding algorithm called Node2vec [25] to produce drug features. These features were quite different from traditional drug fingerprint features, which could be obtained by considering each drug individually. The features adopted in this study abstracted drug associations, and they could overview each drug with all drugs as background, thus representing drugs at a system level. They were learnt by support vector machine (SVM) [26] or random forest (RF) [27] to construct the model. The cross-validation results indicated the good

performance of the model. The model was also superior to those with classic drug fingerprint features or embedding features yielded by another network embedding algorithm. The proposed model adopted synthetic minority oversampling technique (SMOTE) [28] to tackle the imbalance problem, and the results suggested that it could balance the performance across different classes.

2. Materials and methods

2.1. Dataset

Drugs and their information were retrieved from KEGG DRUG (https://www.genome.jp/kegg/drug/) [22]. According to https://www.kegg.jp/brite/br08332 (accessed on 10 January 2022), 820 drugs encoded by KEGG IDs were classified into 14 classes, as mentioned in Section 1 and listed in column 2 of Table 1. For convenience, these classes were tagged as C_1–C_{14}. The correspondence between tags and classes could be found in columns 1 and 2 of Table 1. Among these drugs, only two belong to two classes, while the others are in one exact class. Thus, these two drugs were removed. As the CCI information was employed to build the model, where drugs are represented by PubChem IDs, the KEGG IDs with unavailable PubChem IDs were also removed. If the PubChem ID of one KEGG ID was not included in the drug network (Section 2.2), it was also discarded. Finally, 579 drugs encoded by KEGG IDs were obtained. These drugs were classified into 14 classes, as listed in column 2 of Table 1. The number of drugs in each class is also listed in Table 1. The detailed drugs in each class are provided in S1.

Table 1. Breakdown of the drug dataset.

Tag	Drug class	Number of drugs
C_1	Anti-allergic agent	35
C_2	Anti-inflammatory	116
C_3	Antibacterial	99
C_4	Antidiabetic agent	25
C_5	Antifungal	21
C_6	Antineoplastic	51
C_7	Antiviral	44
C_8	Cardiovascular agent	54
C_9	Endocrine and hormonal agent	19
C_{10}	Gastrointestinal agent	29
C_{11}	Hypolipidemic agent	13
C_{12}	Immunological agent	8
C_{13}	Neuropsychiatric agent	37
C_{14}	Ophthalmic agent	28
Total		579

2.2. Construction of drug network

Recently, network is a popular form to investigate various complex biological and medical problems [15,29–31], because it could overview all objects (nodes in the network) and evaluate them

at a system level. In the present study, network was utilized to organize drugs.

The constructed network defined drugs reported in KEGG as nodes. The relationship between nodes should be determined to form the network. Several public databases provide the existing associations of drugs. Here, the CCI information reported in STITCH (http://stitch.embl.de/, version 4.0) was employed [23,24]. The file named "chemical_chemical.links.v4.0.tsv.gz" was downloaded from STITCH, which contains a large number of CCIs. Each CCI consists of two chemicals, represented by PubChem IDs, and one confidence score. The confidence score is obtained by measuring several aspects of chemicals, such as structures, reactions, activities, and literature descriptions. It ranges between 1 and 999 with a meaning that the higher the confidence score, the stronger the association between two chemicals. For formulation, the confidence score on the CCI between chemicals c_1 and c_2 was denoted as $S(c_1, c_2)$. For the constructed network, two nodes were connected by an edge if their corresponding drugs could comprise a CCI with a confidence score larger than zero. Evidently, each edge in the network indicated a CCI. Each edge was assigned a weight, which was defined as $S(c_1, c_2)/1000$, to reflect different strengths of CCIs. Such network was denoted by N_d, and it contained 17,956 nodes and 3,134,797 edges.

2.3. Drug network embedding features

Encoding samples into continuous vectors that could be processed by most computer algorithms is one of the most important steps to build efficient classification models. As mentioned in Section 2.2, a drug network was constructed, which contained informative associations of drugs. This information could be used to encode drugs. In recent years, several network embedding algorithms, such as DeepWalk [32], Mashup [33], and LINE [34], have been proposed to assign continuous vectors to nodes in one or more networks. Here, Node2vec [25] was selected to extract drug features from N_d.

Node2vec is a powerful network embedding algorithm. In fact, it could be deemed as an improved version of DeepWalk. Similar to DeepWalk, it samples many paths from a given network. The procedures of generating paths are greatly improved compared with those in DeepWalk. For each node u, Node2vec produces a predefined number of paths starting at u. Suppose that $n_{i\text{-}1}$ is the (i-1)-th node in one path starting at u. This path is extended to the i-th node, denoted by n_i, by selecting one neighbor of $n_{i\text{-}1}$. The selection probability from $n_{i\text{-}1}$ to any other node is defined as

$$P(n_i = w \mid n_{i-1} = v) = \begin{cases} \pi_{vw} / Z & if\ w \text{ is adjacent to } v \\ 0 & Otherwise \end{cases}, \tag{1}$$

where π_{vw} stands for the transition probability from v to w, defined by $\pi_{vw} = \alpha_{pq}(t,w) \cdot w_{vw}$; Z is the sum of transition probabilities from v to all its neighbors; and w_{vw} denotes the weight on edge connecting nodes v and w. Moreover, $\alpha_{pq}(t,w)$ could be determined by

$$\alpha_{pq}(t,w) = \begin{cases} 1/p & \text{if } d_{tw} = 0 \\ 1 & \text{if } d_{tw} = 1 \\ 1/q & \text{if } d_{tw} = 2 \end{cases}, \tag{2}$$

where t is the (i-2)-th node in the path, and d_{tw} represents the distance between t and w. For the path starting at u, it is iteratively extended by selecting a neighbor of the current end-point in accordance with the probabilities calculated using Eq (1) until the length of the path reaches the predefined length. After all paths are produced, they are fed into the word2vec algorithm with SkipGram to produce vectors of nodes, where the paths are deemed as sentences and nodes are considered as words.

In this study, the program of Node2vec was retrieved from https://snap.stanford.edu/node2vec/. Such program was performed on N_d by using its default parameters. Afterwards, the feature vectors of 579 drugs were picked up to construct models.

2.4. Classification algorithms

Besides the representation of samples, another key step to build efficient classification models is to select a proper classification algorithm. In this study, two classic classification algorithms were used: SVM [26] and RF [27]. These two algorithms have been widely used to set up classification models with excellent performance for investigating various biological [35–39] and medical problems [40–43].

SVM is a statistical learning-based classification algorithm. Its original version could only process binary classification problems. Its idea is to determine an optimal hyperplane that could separate samples into two classes as perfect as possible. However, such hyperplane is not easy to obtain in the original feature subspace. It generally employs kernel trick to map samples into a high-dimension space, in which the hyperplane, designed to separate samples in two classes, could be easily constructed. A test sample is also mapped into the high-dimension space. The class of the test sample is determined in accordance with the side of the hyperplane it is located. Subsequently, SVM could tackle multi-class classification problems by using "one-versus-rest" or "one-versus-one".

RF is another powerful and widely used classification algorithm, which is quite different from SVM. In fact, it is a type of ensemble algorithm. Several decision trees are constructed and comprise RF. Some features are randomly selected from all features to learn a decision tree. Samples are randomly selected, with replacement, from the original dataset. Then, a decision tree is built on the basis of these randomly selected features and samples. RF gives the prediction by majority voting on predictions generated by the decision trees it contains.

In this study, the tools "SMO" and "RandomForest" in Weka were directly used [44], and they implemented the above mentioned SVM and RF, respectively. These tools were used with their default parameters.

2.5. Synthetic minority oversampling technique

A total of 14 drug classes were involved in this study. Some classes contained much less samples than the others, implying the dataset was imbalanced. The classification model directly built on such dataset may provide excellent performance on the majority class and poor performance on the minority class. SMOTE [28] was employed to tackle such problem. For each class, except the largest, one sample is randomly selected, denoted by x. The k nearest neighbors of x in the same class are found, and one is randomly selected, say y. On the basis of x and y, a new sample z is generated, which is defined as the linear combination of x and y with randomly produced combination coefficients. Given that the new sample is highly related to x and y, it is placed into the same class of x and y. The above procedures are executed several times until the size of the minority class is the same as that of the

largest class.

This study adopted the tool "SMOTE" in Weka [44] to balance the investigated dataset, and default parameters were used.

2.6. Performance assessment

As a multi-class classification problem, the performance of the models could be measured by overall accuracy (ACC), which is defined as the proportion of all correctly predicted samples.

Recall, precision, F-measure, and Matthew's correlation coefficient (MCC) [45] were also computed for each class as follows:

$$Precision = \frac{TP}{TP + FP}, \tag{3}$$

$$Recall = \frac{TP}{TP + FN}, \tag{4}$$

$$F - measure = \frac{2 \times Precision \times Recall}{Precision + Recall}, \tag{5}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TN + FN) \times (TN + FP) \times (TP + FN) \times (TP + FP)}}, \tag{6}$$

where TP, FP, FN, and TN stand for true positive, false positive, false negative, and true negative, respectively. In multi-class classification, their specific definitions for one class are as follows: TP is the number of samples in the class that are also classified into this class, FP is the number of samples not in the class that are classified into this class, FN is the number of samples in the class that are classified into other classes, and TN is the number of samples not in the class that are classified into other classes. Eqs (3)–(6) could only assess the performance under a certain threshold. For full assessment, the receiver operating characteristic (ROC) and precision-recall (PR) curve analyses were further employed on the predicted results for each class. The area under the ROC curve, which was denoted by AUROC, and the PR curve, which was represented by AUPR, were calculated to fully evaluate the performance of different models on each class. For easy description, the six measurements for class C_i were denoted by precision(i), recall(i), F-measure(i), MCC(i), AUROC(i), and AUPR(i). Furthermore, the weighted precision, recall, F-measure, MCC, AUROC, and AUPR were calculated to fully evaluate the overall performance of all classification models.

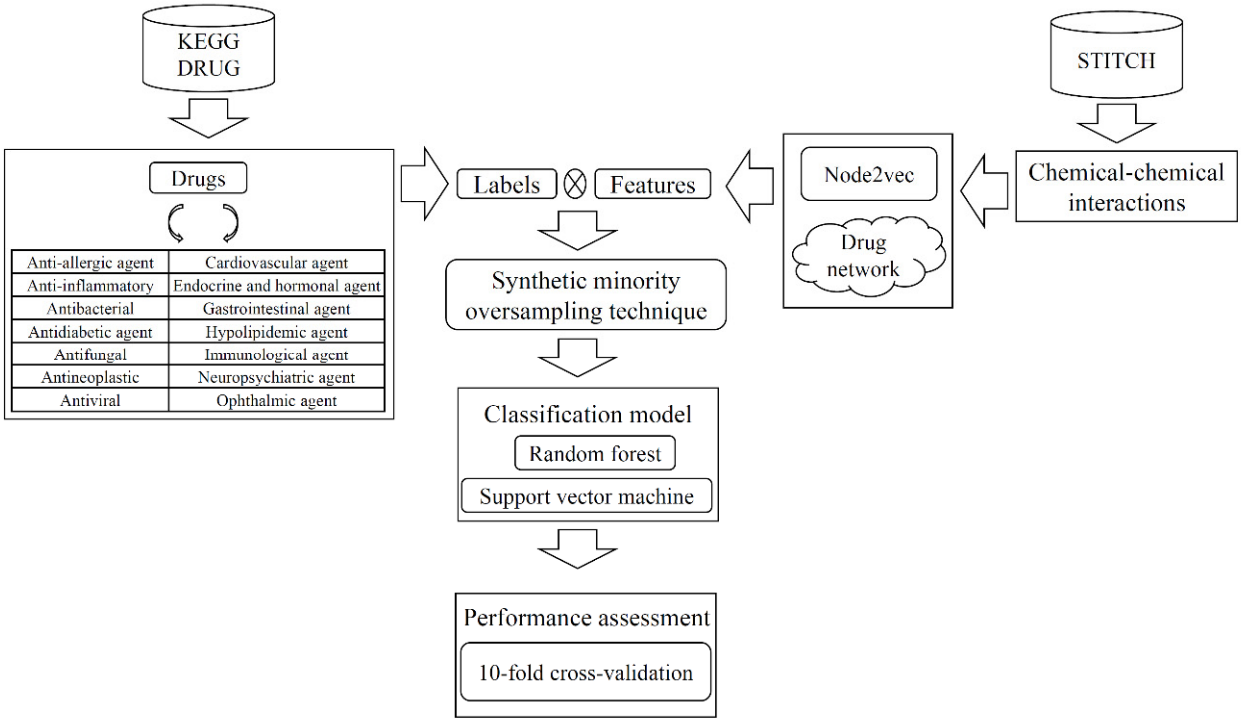

Figure 1. Entire procedures for constructing and evaluating the model. Drugs were classified into 14 classes reported in KEGG DRUG. The chemical-chemical interaction information retrieved from STITCH was used to build a large drug network, from which drug features were extracted via Node2vec. These features indicated the associations between drugs, and they could overview each drug with all drugs as background. After the dataset was processed by the synthetic minority oversampling technique, it was fed into one classification algorithm to build the model. The model was finally assessed by 10-fold cross-validation. This figure was generated in PowerPoint.

3. Results

In this study, a novel model was proposed to classify drugs in accordance with the drug classification system recently reported in KEGG. Its construction and evaluation procedures are illustrated in Figure 1. In this section, detailed evaluation results were provided, and further comparisons were conducted.

3.1. Performance of the proposed model

The proposed model adopted the drug features derived from a large drug network via Node2vec and SMOTE to tackle the imbalance problem. The best dimension of the features was found to be 128. Two classic classification algorithms (RF and SVM) were used as the prediction engine. For easy description, if RF (SVM) was selected as the prediction engine, the model was called RF (SVM) model. Each model was evaluated five times using a 10-fold cross-validation [46]. The confusion matrix of each model under each 10-fold cross-validation is provided in S2. The average performance is listed

in Table 2.

Table 2. Overall performance of the models with different classification algorithms under 10-fold cross-validation.

Classification algorithm	SMOTE	ACC	Weighted Precision	Weighted F-measure	Weighted MCC	Weighted AUROC	Weighted AUPR
Random forest	√	0.9486	0.9631	0.9511	0.9484	0.9958	0.9830
Support vector machine	√	0.9583	0.9752	0.9640	0.9625	0.9830	0.9350
Random forest	×	0.9410	-	-	-	0.9930	0.9710
Support vector machine	×	0.9600	0.9610	0.9600	0.9550	0.9840	0.9320

For the RF model, the ACC was 0.9486, and the weighted precision, F-measure, MCC, AUROC, and AUPR were 0.9631, 0.9511, 0.9484, 0.9958, and 0.9830, respectively. The result of each measurement was relatively high, suggesting the excellent performance of the RF model. The model's performance on 14 classes is provided in Table 3, and the ROC and PR curves under one 10-fold cross-validation are shown in Figure 2(A),(B), respectively. Table 3 shows that most measurements were higher than 0.9. The number of classes on which the recall values were higher than 0.9 was 12, and this numbers for precision, F-measure, MCC, AUPRO, and AUPR were 11, 12, 12, 14, and 13, respectively. The results indicated the good performance of the RF model on most classes, conforming to its overall performance. Careful examination of the measurements in Table 3 showed that the RF model provided the lowest performance on class C_{12} (immunological agent) for each measurement. For example, the recall on this class was only 0.2726. This class only contained eight drugs, which was not only the smallest among all 14 classes, but was considered the lowest performing class. Although the RF model adopted SMOTE to tackle the imbalance problem, the class sizes could still have influence. In the following text, we would elaborate that such influence has been decreased by SMOTE.

Table 3. Performance of the random forest model on each class under 10-fold cross-validation.

Tag of class	**Recall(*i*)**	**Precision(*i*)**	**F-measure(*i*)**	**MCC(*i*)**	**AUROC(*i*)**	**AUPR(*i*)**
C_1	0.9052	0.9319	0.9141	0.9162	0.9975	0.9627
C_2	0.9550	0.9449	0.9469	0.9366	0.9972	0.9879
C_3	0.9983	0.9761	0.9865	0.9845	0.9998	0.9991
C_4	0.9750	0.9749	0.9690	0.9708	0.9994	0.9921
C_5	0.9676	1.0000	0.9804	0.9812	0.9943	0.9780
C_6	0.9473	0.9113	0.9179	0.9182	0.9915	0.9770
C_7	0.9434	0.9611	0.9450	0.9453	0.9932	0.9769
C_8	0.9377	0.9829	0.9548	0.9537	0.9940	0.9709
C_9	0.9191	0.9294	0.9499	0.9511	0.9972	0.9659
C_{10}	0.9132	0.8882	0.8858	0.8886	0.9969	0.9602
C_{11}	0.8870	0.8528	0.9582	0.9622	1.0000	1.0000
C_{12}	0.2726	0.5183	0.5834	0.5852	0.9127	0.7654
C_{13}	0.9781	0.9456	0.9552	0.9557	0.9997	0.9958
C_{14}	0.9831	0.9738	0.9768	0.9764	0.9998	0.9967

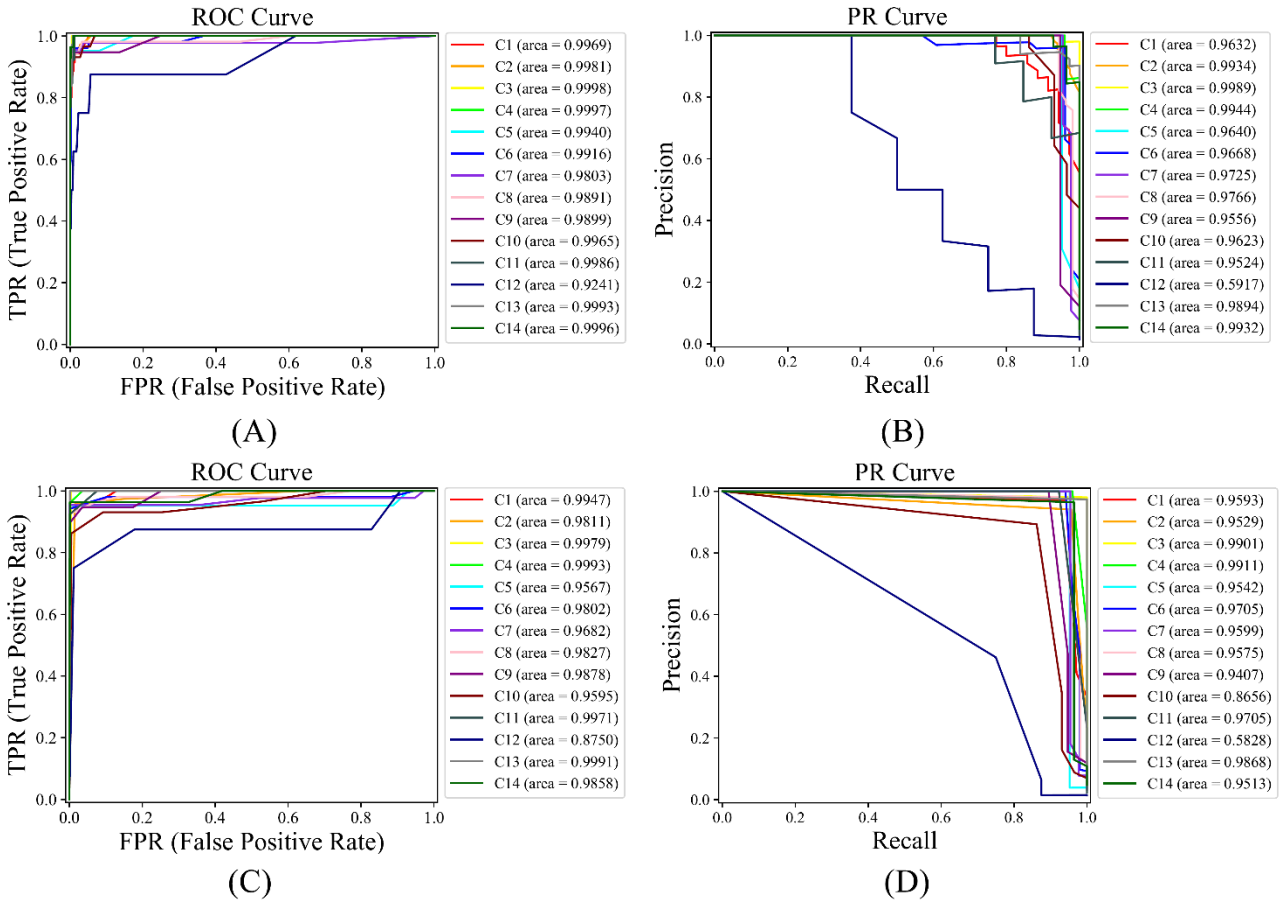

Figure 2. ROC and PR curves of two models on 14 classes under one 10-fold cross-validation. (A) ROC curves of the RF model; (B) PR curves of the RF model; (C) ROC curves of the SVM model; (D) PR curves of the SVM model. The two models provide nearly perfect ROC and PR curves on most classes, indicating that they could efficiently classify drugs. This figure was generated by matplotlib package in Python.

As for the SVM model, the ACC reached 0.9583, which was slightly higher than that of the RF model. As for the other five measurements, they were 0.9752, 0.9640, 0.9625, 0.9830 and 0.9350 (Table 2), respectively. Compared with the measurements of RF model, SVM model yielded higher weighted precision, F-measure, MCC, whereas lower weighted AUROC and AUPR. These indicated that the RF and SVM model provided almost equal performance. The detailed performance of SVM model on fourteen classes is listed in Table 4. Furthermore, the ROC and PR curves of this model under one 10-fold cross-validation are illustrated in Figure 2(C),(D), respectively. Similar to the performance of RF model, most measurements listed in this table were higher than 0.9. The numbers of classes on which recall, precision, F-measure, MCC, AUROC and AUPR, respectively, were higher than 0.9 were 12, 13, 13, 13, 14 and 11, which were also similar to those of the RF model. Again, the performance of SVM model on the smallest class C_{12} (immunological agent) was also lowest for each measurement.

Table 4. Performance of the support vector machine model on each class under 10-fold cross-validation.

Tag of class	Recall(*i*)	Precision(*i*)	F-measure(*i*)	MCC(*i*)	AUROC(*i*)	AUPR(*i*)
C_1	0.9444	0.9152	0.9211	0.9215	0.9934	0.8984
C_2	0.9635	0.9584	0.9593	0.9513	0.9882	0.9428
C_3	0.9967	0.9803	0.9877	0.9860	0.9982	0.9803
C_4	0.9717	1.0000	0.9831	0.9838	0.9982	0.9836
C_5	0.9676	0.9867	0.9704	0.9725	0.9916	0.9603
C_6	0.9556	0.9656	0.9549	0.9540	0.9857	0.9604
C_7	0.9384	0.9598	0.9407	0.9412	0.9700	0.9197
C_8	0.9569	0.9805	0.9650	0.9637	0.9905	0.9546
C_9	0.9109	1.0000	0.9756	0.9759	0.9964	0.9644
C_{10}	0.8567	0.9282	0.9049	0.9096	0.9623	0.8393
C_{11}	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
C_{12}	0.7373	0.4740	0.6904	0.7134	0.9270	0.5556
C_{13}	0.9790	0.9843	0.9786	0.9788	0.9993	0.9814
C_{14}	0.9841	0.9944	0.9874	0.9875	0.9985	0.9874

According to Table 2, the SVM model provided higher performance on four measurements than the RF model, whereas the RF model yielded higher performance on the other two measurements. The SVM model was slightly superior to the RF model. Several box plots were drawn on all measurements for the performance of the two models on 14 classes, as shown in Figure 3. For recall, the SVM model varied in a smaller range than the RF model. The same case occurred for F-measure, MCC, and AUROC. As for the remaining two measurements (precision and AUPR), the RF model changed in a smaller interval. On the whole, the stability of the performance of these two models on different classes was at the same level. However, the SVM model was slightly more stable. These results showed that the SVM model is a more suitable tool for classifying drugs than the RF model.

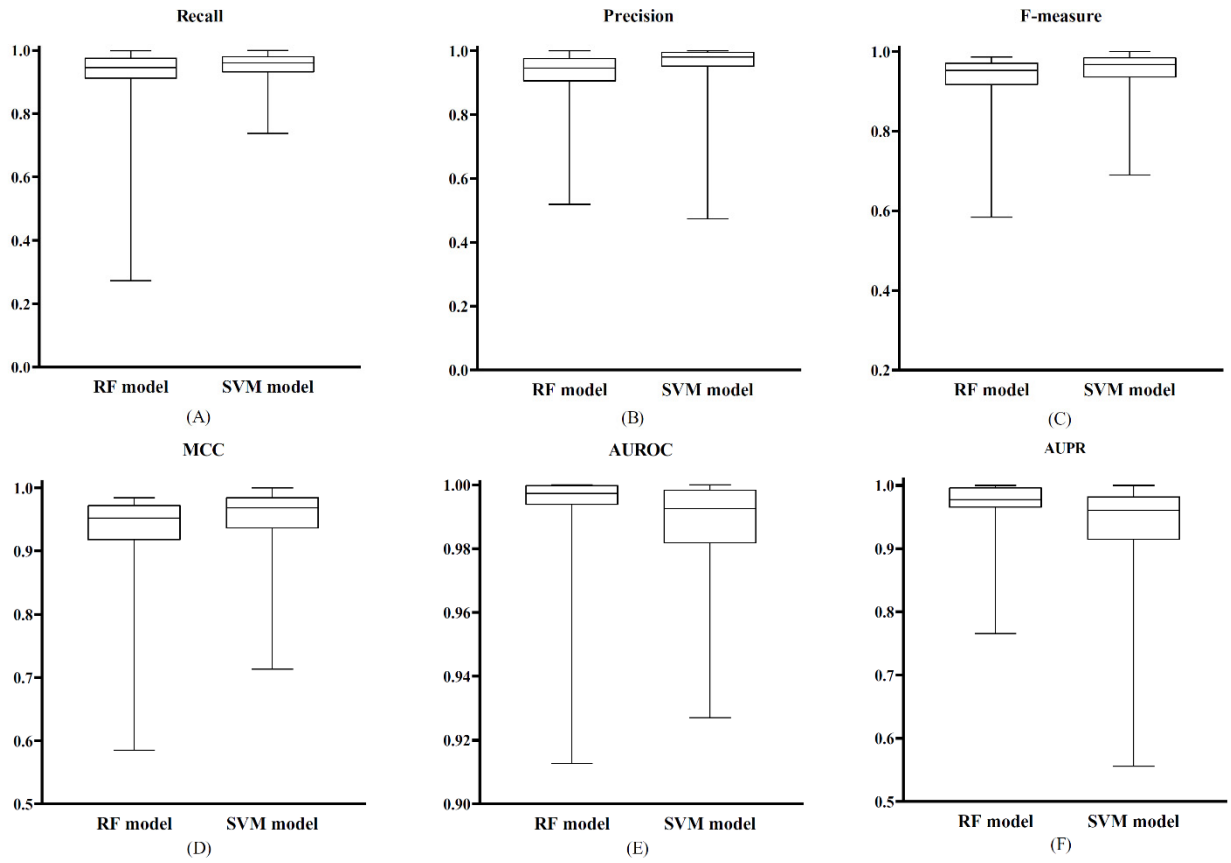

Figure 3. Box plot showing the performance change of RF and SVM models on 14 classes. (A) Performance change on recall; (B) Performance change on precision; (C) Performance change on F-measure; (D) Performance change on MCC; (E) Performance change on AUROC; (F) Performance change on AUPR. The SVM model provides more stable performance on 14 classes than the RF model, indicating that the former is a more suitable tool for drug classification. This figure was generated by GraphPad Prism.

3.2. Comparison of the model without SMOTE

In this study, SMOTE was adopted to tackle the imbalance problem. The performance of the model without SMOTE must be investigated and compared with that of the model with SMOTE. In view of this, RF and SVM were directly applied on the features yielded by Node2vec to build the RF and SVM models without SMOTE. Their performance was also evaluated five times using a 10-fold cross-validation. The average performance is listed in Table 2. For the RF model without SMOTE, the ACC, weighted AUROC, and AUPR were 0.9410, 0.9930, and 0.9710, respectively. As no drugs were classified into C_{12}, the precision, F-measure, and MCC for this class could not be computed, further inducing no results for weighted precision, F-measure, and MCC. According to the computed measurements, the performance of this model was only slightly lower than that of the RF model with SMOTE. As for the SVM model without SMOTE, the six measurements were 0.9600, 0.9610, 0.9600, 0.9550, 0.9840, and 0.9320, which were relatively similar to those of the SVM model with SMOTE. This finding suggested that the employment of SMOTE did not improve the overall performance of the model at all. However, the utilization of SMOTE was reflected not only on the improvement of the

overall performance of the models but also on the balance of performance across different classes. Thus, the accuracy, i.e., recall, on the 14 classes yielded by the models with or without SMOTE was further investigated. A box plot was drawn for each model to clearly show the change in recall on all classes, as shown in Figure 4. When the same prediction engine was used (RF or SVM), the recall values of the model with SMOTE varied in a smaller interval than those of the model without SMOTE. This result implied that the employment of SMOTE balanced the performance across different classes. For example, on C_{12}, the RF and SVM models without SMOTE produced recall values of 0.0000 and 0.5000, respectively, and these values were improved to 0.2726 and 0.7373, respectively, by employing SMOTE. For practical applications, the model with SMOTE could avoid the extremely good performance on the majority class and extremely low performance on the minority class.

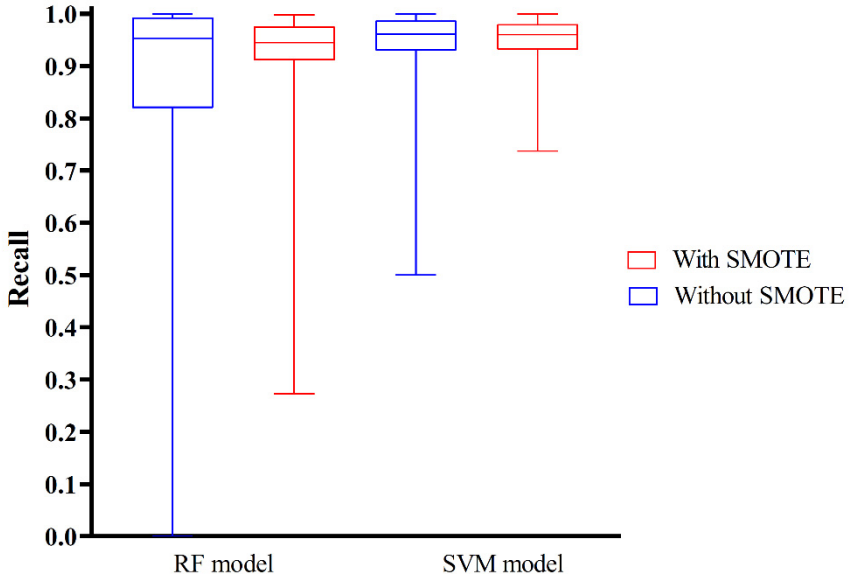

Figure 4. Box plot showing the performance change of models with or without SMOTE on 14 classes. The models with SMOTE evidently provided more balanced performance on all classes than those without SMOTE. Through SMOTE, the drug samples in each class, including real and synthesized samples, are almost similar in number. The bias of the model based on the dataset processed by SMOTE could be reduced. This figure was generated by GraphPad Prism.

3.3. Comparison of models using other drug features

In this study, the drug features derived from a network via Node2vec were used to build the model. Two other feature types were adopted to build models, which were compared with the proposed model.

The first feature type contained features derived from the drug network N_d via another powerful network embedding algorithm, Mashup [33]. The dimension was also set to 128. RF and SVM were applied on these features to construct models. SMOTE was employed to tackle the imbalance problem. The models were also assessed five times using a 10-fold cross-validation. For clear description, they were called Mashup-based RF and SVM models, whereas the proposed models were termed as Node2vec-based RF and SVM models. The average performance of the Mashup-based models is

provided in Table 5. When RF was selected as the prediction engine, the Node2vec-based model provided evidently higher performance on all measurements than the Mashup-based model. In detail, the improvement on ACC, weighted F-measure, MCC, and AUPR was about 0.05 and such improvement on weighted precision and AUROC was about 0.03 and 0.01, respectively. Meanwhile, the superiority of the Node2vec-based SVM model to the Mashup-based SVM model was not very evident. However, the Node2vec-based SVM model provided higher performance on five measurements, implied that it was slightly better than the Mashup-based model.

Table 5. Performance of models with other drug features.

Drug features	Classification algorithm	ACC	Weighted Precision	Weighted F-measure	Weighted MCC	Weighted AUROC	Weighted AUPR
Drug features produced by Mashup	Random forest	0.8891	0.9282	0.9069	0.9006	0.9819	0.9359
	Support vector machine	0.9526	0.9724	0.9617	0.9579	0.9880	0.9280
Drug fingerprint features	Random forest	0.8432	0.8768	0.8431	0.8366	0.9748	0.9107
	Support vector machine	0.8426	0.8791	0.8489	0.8420	0.9476	0.7943

The second feature type included the drug fingerprint features, which were widely used to deal with various drug- or chemical-related problems [16,38,47–49]. Here, RDKit [50] was adopted to extract the ECFP fingerprints of all 579 drugs. The fingerprints of each drug were represented by a binary vector with dimension of 1024. These fingerprint features were fed into RF and SVM to construct models. SMOTE was also employed. The models were called fingerprint-based models. All models were evaluated five times using a 10-fold cross-validation. The average performance is provided in Table 5. The results showed that most measurements were lower than 0.9, indicating low performance of such models. Compared with the performance of the proposed model (Table 2), the proposed model clearly provided much higher performance.

Finally, the ROC and PR curves on each class for some models were plotted. For the Node2vec-based model, the SVM model was selected because it was slightly better and more stable than the RF model. For the Mashup-based model, the SVM model was chosen as it was superior to the RF model (Table 5). As for the fingerprint-based model, the RF model was selected due to its superiority to the SVM model (Table 5). The ROC and PR curves of these models on all classes are shown in Figure 5. A notable detail that these curves were produced by one of the five 10-fold cross-validations. The areas under the ROC curves of the fingerprint-based RF model were much smaller than those of other two models. The same results were obtained for PR curves. As for the Node2vec-based and Mashup-based SVM models, the areas under the ROC or PR curves were relatively similar. However, the areas under the ROC or PR curves of the Mashup-based SVM model on some classes (e.g., C_{12}) were relatively low, whereas they were improved by the Node2vec-based SVM model.

Therefore, the proposed model (Node2vec-based model) was better than the models using other drug features, indicating its superiority in classifying drugs.

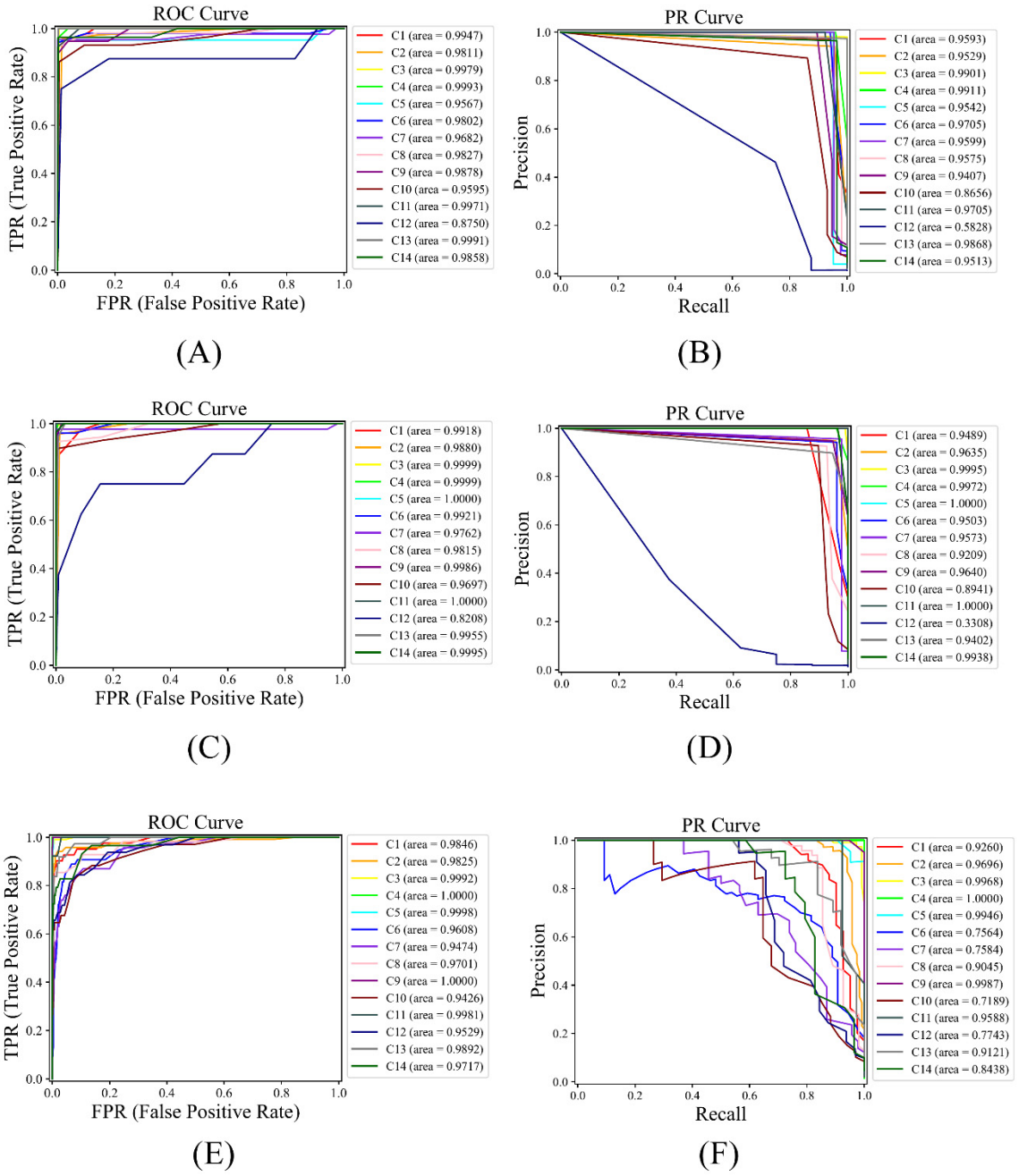

Figure 5. ROC and PR curves of some models on 14 classes. (A) ROC curves of the Node2vec-based SVM model; (B) PR curves of the Node2vec-based SVM model; (C) ROC curves of the Mashup-based SVM model; (D) PR curves of the Mashup-based SVM model; (E) ROC curves of the fingerprint-based RF model; (F) PR curves of the fingerprint-based RF model. Models with features derived from the drug network are superior to those with fingerprint features, suggesting network features are more related to and informative for drug classification than traditional drug fingerprint features. This figure was generated by matplotlib package in Python.

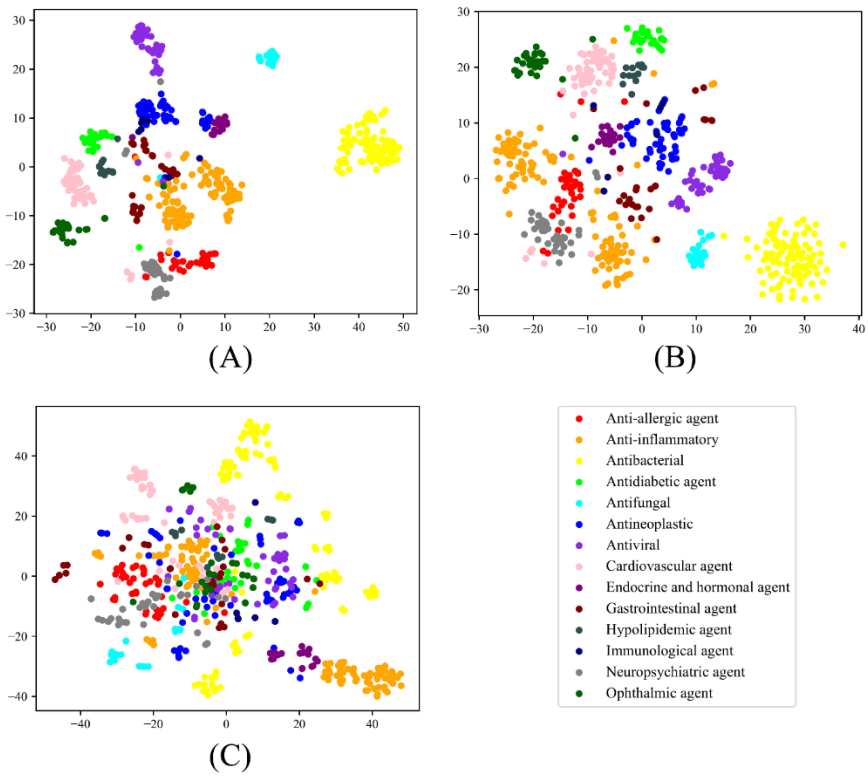

Figure 6. Visualization of three drug feature types in 2D space by using t-SNE. (A) Drug features derived from the drug network via Node2vec; (2) Drug features derived from the drug network via Mashup; (3) Drug fingerprint features. The drug features derived from the drug network via Node2vec could evidently cluster drugs in different classes the best, implying that such features are the excellent representations for drug classification. This figure was generated by matplotlib package in Python.

3.4. Superiority of features derived from the drug network

In Section 3.3, the proposed model and two other models with different drug features were compared. The proposed model provided better performance than the two other models, suggesting that the features used in the proposed model were more efficient than those adopted in other models for drug classification. The t-SNE [51] was adopted to analyze three feature types used in the Node2vec-based, Mashup-based, and fingerprint-based models to further confirm these findings and provide intuitive evidence. Such tool could reduce the dimensions of features to two and display them in a 2D space, where samples in different classes are in different colors. The results of t-SNE are illustrated in Figure 6. The features derived from the drug network [Figure 6(A),(B)] could cluster drugs in different classes evidently better than the drug fingerprints [Figure 6(C)]. As for two feature types derived from the drug network, the features yielded by Node2vec [Figure 6(A)] could cluster drugs in different classes more compactly than those generated by Mashup. These results implied that the features derived from the drug network via Node2vec, which were used in the proposed model, were more helpful in classifying drugs. This finding was the main reason why the Node2vec-based model could provide better performance.

4. Conclusions

In this study, a new drug classification system reported in KEGG was investigated by proposing classification models for assigning class in this system to any given drug. Informative drug features were derived from a large drug network via a powerful network embedding algorithm. The evaluation and comparison results indicated that the features were more related to the drug classification system in KEGG and more informative than classic fingerprint features. Employing more drug information (e.g., drug side effects and indications) or more complex and advanced feature learning algorithms (e.g., graph convolutional network and convolutional neural network) may be helpful to access more informative drug features, thereby building more powerful classification models. These aspects could be investigated in future work. Furthermore, the employment of SMOTE guaranteed that the performance on all classes was high. The model could be a useful tool to classify drugs, discover novel effects of existing drugs and determine the effects of candidate drug-like compounds.

Conflict of interest

The authors declare there is no conflict of interest.

References

1. P. A. Naik, M. Yavuz, S. Qureshi, J. Zu, S. Townley, Modeling and analysis of COVID-19 epidemics with treatment in fractional derivatives using real data from Pakistan, *Eur. Phys. J. Plus*, **135** (2020), 795. https://doi.org/10.1140/epjp/s13360-020-00819-5
2. P. A. Naik, J. Zu, K. M. Owolabi, Modeling the mechanics of viral kinetics under immune control during primary infection of HIV-1 with treatment in fractional order, *Phys. A*, **545** (2020), 123816. https://doi.org/10.1016/j.physa.2019.123816
3. P. A. Naik, J. Zu, M. Ghoreishi, Stability analysis and approximate solution of SIR epidemic model with Crowley-Martin type functional response and holling type-II treatment rate by using homotopy analysis method, *J. Appl. Anal. Comput.*, **10** (2020), 1482–1515. https://doi.org/10.11948/20190239
4. B. Wang, J. F. Gomez-Aguilar, Z. Sabir, M. A. Z. Raja, W. F. Xia, H. Jahanshahi, et al., Numerical computing to solve the nonlinear corneal system of eye surgery using the capability of morlet wavelet artificial neural networks, *Fractals*, **30** (2022), 1–19. https://doi.org/10.1142/S0218348X22401478
5. J. E. Solís-Pérez, J. A. Hernández, A. Parrales, J. F. Gómez-Aguilar, A. Huicochea, Artificial neural networks with conformable transfer function for improving the performance in thermal and environmental processes, *Neural Networks*, **152** (2022), 44–56. https://doi.org/10.1016/j.neunet.2022.04.016
6. M. Umar, Z. Sabir, M. A. Z. Raja, J. F. G. Aguilar, F. Amin, M. Shoaib, Neuro-swarm intelligent computing paradigm for nonlinear HIV infection model with CD4+ T-cells, *Math. Comput. Simulat.*, **188** (2021), 241–253. https://doi.org/10.1016/j.matcom.2021.04.008
7. A. A. Mostafa, A. A. Alhossary, S. A. Salem, A. E. Mohamed, GBO-kNN a new framework for enhancing the performance of ligand-based virtual screening for drug discovery, *Expert Syst. Appl.*, **197** (2022), 116723. https://doi.org/10.1016/j.eswa.2022.116723

8. Q. Dai, C. Bao, Y. Hai, S. Ma, T. Zhou, C. Wang, et al., MTGIpick allows robust identification of genomic islands from a single genome, *Brief. Bioinf.*, **19** (2016), 361–373. https://doi.org/10.1093/bib/bbw118
9. R. Kong, X. Xu, X. Liu, P. He, M. Q. Zhang, Q. Dai, 2SigFinder: the combined use of small-scale and large-scale statistical testing for genomic island detection from a single genome, *BMC Bioinf.*, **21** (2020), 159. https://doi.org/10.1186/s12859-020-3501-2
10. S. Yang, Y. Wang, Y. Chen, Q. Dai, MASQC: Next generation sequencing assists third generation sequencing for quality control in N6-Methyladenine DNA identification, *Front. Genet.*, **11** (2020), 269. https://doi.org/10.3389/fgene.2020.00269
11. Z. Lu, K. C. Chou, iATC_Deep-mISF: a multi-label classifier for predicting the classes of anatomical therapeutic chemicals by deep learning, *Adv. Biosci. Biotechnol.*, **11** (2020) ,153–159. https://doi.org/10.4236/abb.2020.115012
12. A. Lumini, L. Nanni, Convolutional neural networks for ATC classification, *Curr. Pharm. Design*, **24** (2018), 4007–4012. https://doi.org/10.2174/1381612824666181112113438
13. H. Zhao, Y. Li, J. Wang, A convolutional neural network and graph convolutional network-based method for predicting the classification of anatomical therapeutic chemicals, *Bioinformatics*, **37** (2021), 2841–2847. https://doi.org/10.1093/bioinformatics/btab204
14. Y. Cao, Z. Q. Yang, X. L. Zhang, W. Fan, Y. Wang, J. Shen, et al., Identifying the kind behind SMILES—anatomical therapeutic chemical classification using structure-only representations, *Brief. Bioinf.*, (2022), bbac346. https://doi.org/10.1093/bib/bbac346
15. J. P. Zhou, L. Chen, Z. H. Guo, iATC-NRAKEL: An efficient multi-label classifier for recognizing anatomical therapeutic chemical classes of drugs, *Bioinformatics*, **36** (2020), 1391–1396. https://doi.org/10.1093/bioinformatics/btz757
16. J. P. Zhou, L. Chen, T. Wang, M. Liu, iATC-FRAKEL: A simple multi-label web-server for recognizing anatomical therapeutic chemical classes of drugs with their fingerprints only, *Bioinformatics*, **36** (2020), 3568–3569. https://doi.org/10.1093/bioinformatics/btaa166
17. S. Tang, L. Chen, iATC-NFMLP: Identifying classes of anatomical therapeutic chemicals based on drug networks, fingerprints and multilayer perceptron, *Curr. Bioinf.*, (2022), in press. https://doi.org/10.2174/1574893617666220318093000
18. X. Cheng, S. G. Zhao, X. Xiao, K. C. Chou, iATC-mISF: a multi-label classifier for predicting the classes of anatomical therapeutic chemicals, *Bioinformatics*, **33** (2016), 341–346. https://doi.org/10.1093/bioinformatics/btw644
19. L. Nanni, S. Brahnam, Multi-label classifier based on histogram of gradients for predicting the anatomical therapeutic chemical class/classes of a given compound, *Bioinformatics*, **33** (2017), 2837–2841. https://doi.org/10.1093/bioinformatics/btx278
20. X. Cheng, S. G. Zhao, X. Xiao, K. C. Chou, iATC-mHyb: a hybrid multi-label classifier for predicting the classification of anatomical therapeutic chemicals, *Oncotarget*, **8** (2017), 58494–58503. https://doi.org/10.18632/oncotarget.17028
21. X. Wang, Y. Wang, Z. Xu, Y. Xiong, D. Q. Wei, ATC-NLSP: Prediction of the classes of anatomical therapeutic chemicals using a network-based label space partition method, *Front. Pharmacol.*, **10** (2019), 971. https://doi.org/10.3389/fphar.2019.00971
22. H. Ogata, S. Goto, K. Sato, W. Fujibuchi, H. Bono, M. Kanehisa, KEGG: Kyoto encyclopedia of genes and genomes, *Nucleic Acids Res.*, **27** (1999), 29–34. https://doi.org/10.1093/nar/28.1.27

23. M. Kuhn, C. von Mering, M. Campillos, L. J. Jensen, P. Bork, STITCH: interaction networks of chemicals and proteins, *Nucleic Acids Res.*, **36** (2007), D684–D688. https://doi.org/10.1093/nar/gkm795
24. M. Kuhn, D. Szklarczyk, S. Pletscher-Frankild, T. H. Blicher, C. von Mering, L. J. Jensen, et al., STITCH 4: integration of protein-chemical interactions with user data, *Nucleic Acids Res.*, **42** (2014), D401–407. https://doi.org/10.1093/nar/gkt1207
25. A. Grover, J. Leskovec, node2vec: Scalable feature learning for networks, in *the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (2016), 855–864. https://doi.org/10.1145/2939672.2939754
26. C. Cortes, V. Vapnik, Support-vector networks, *Mach. Learn.*, **20** (1995), 273–297. https://doi.org/10.1007/BF00994018
27. L. Breiman, Random forests, *Mach. Learn.*, **45** (2001), 5–32. https://doi.org/10.1023/A:1010933404324
28. N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, SMOTE: Synthetic minority over-sampling technique, *J. Artif. Intell. Res.*, **16** (2002), 321–357. https://doi.org/10.1613/jair.953
29. X. Zhao, L. Chen, Z. H. Guo, T. Liu, Predicting drug side effects with compact integration of heterogeneous networks, *Curr. Bioinform.*, **14** (2019), 709–720. https://doi.org/10.2174/1574893614666190220114644
30. W. Zhang, X. Yue, F. Liu, Y. L. Chen, S. K. Tu, X. N. Zhang, A unified frame of predicting side effects of drugs by using linear neighborhood similarity, *BMC Syst. Biol.*, **11** (2017), 101. https://doi.org/10.1186/s12918-017-0477-2
31. G. Li, T. Fang, Y. Zhang, C. Liang, Q. Xiao, J. Luo, Predicting miRNA-disease associations based on graph attention network with multi-source information, *BMC Bioinf.*, **23** (2022), 244. https://doi.org/10.1186/s12859-022-04796-7
32. B. Perozzi, R. Al-Rfou, S. Skiena, Deepwalk: Online learning of social representations, in *the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, (2014), 701–710. https://doi.org/10.1145/2623330.2623732
33. H. Cho, B. Berger, J. Peng, Compact integration of multi-network topology for functional analysis of genes, *Cell Syst.*, **3** (2016), 540–548. https://doi.org/10.1016/j.cels.2016.10.017
34. J. Tang, M. Qu, M. Wang, M. Zhang, J. Yan, Q. Mei, Line: Large-scale information network embedding, in *the 24th international conference on world wide web*, (2015), 1067–1077. https://doi.org/10.1145/2736277.2741093
35. L. Chen, Z. Li, S. Zhang, Y. H. Zhang, T. Huang, Y. D. Cai, Predicting RNA 5-methylcytosine sites by using essential sequence features and distributions, *BioMed. Res. Int.*, **2022** (2022), 4035462. https://doi.org/10.1155/2022/4035462
36. Y. Wang, Y. Xu, Z. Yang, X. Liu, Q. Dai, Using recursive feature selection with random forest to improve protein structural class prediction for low-similarity sequences, *Comput. Math. Method M.*, **2021** (2021), 5529389. https://doi.org/10.1155/2021/5529389
37. Z. Wu, L. Chen, Similarity-based method with multiple-feature sampling for predicting drug side effects, *Comput. Math. Method M.*, **2022** (2022), 9547317. https://doi.org/10.1155/2022/9547317
38. B. Ran, L. Chen, M. Li, Y. Han, Q. Dai, Drug-Drug interactions prediction using fingerprint only, *Comput. Math. Method M.*, **2022** (2022), 7818480. https://doi.org/10.1155/2022/7818480

39. A. Kastrin, P. Ferk, B. Leskosek, Predicting potential drug-drug interactions on topological and semantic similarity features using statistical learning, *PloS One*, **13** (2018), e196865. https://doi.org/10.1371/journal.pone.0196865
40. S. Ding, D. Wang, X. Zhou, L. Chen, K. Feng, X. Xu, et al., Predicting heart cell types by using transcriptome profiles and a machine learning method, *Life*, **12** (2022), 228. https://doi.org/10.3390/life12020228
41. X. Zhou, S. Ding, D. Wang, L. Chen, K. Feng, T. Huang, et al., Identification of cell markers and their expression patterns in skin based on single-cell RNA-sequencing profiles, *Life*, **12** (2022), 550. https://doi.org/10.3390/life12040550
42. F. Ahmad, A. Farooq, M. U. G. Khan, M. Z. Shabbir, M. Rabbani, I. Hussain, Identification of most relevant features for classification of francisella tularensis using machine learning, *Curr. Bioinf.*, **15** (2020), 1197–1212. https://doi.org/10.2174/1574893615666200219113900
43. M. Onesime, Z. Yang, Q. Dai, Genomic island prediction via chi-square test and random forest algorithm, *Comput. Math. Method M.*, **2021** (2021), 9969751. https://doi.org/10.1155/2021/9969751
44. E. Frank, M. Hall, L. Trigg, G. Holmes, I. H. Witten, Data mining in bioinformatics using Weka, *Bioinformatics*, **20** (2004), 2479–2481. https://doi.org/10.1093/bioinformatics/bth261
45. B. Matthews, Comparison of the predicted and observed secondary structure of T4 phage lysozyme, *BBA-Protein Struct.*, **405** (1975), 442–451. https://doi.org/10.1016/0005-2795(75)90109-9
46. R. Kohavi, A study of cross-validation and bootstrap for accuracy estimation and model selection, in *IJCAI'95: Proceedings of the 14th International Joint Conference on Artificial Intelligence*, (1995), 1137–1145.
47. W. Zhang, F. Liu, L. Luo, J. Zhang, Predicting drug side effects by multi-label learning and ensemble learning, *BMC Bioinf.*, **16** (2015), 365. https://doi.org/10.1186/s12859-015-0774-y
48. Y. Tabei, E. Pauwels, V. Stoven, K. Takemoto, Y. Yamanishi, Identification of chemogenomic features from drug-target interaction networks using interpretable classifiers, *Bioinformatics*, **28** (2012), i487–i494. https://doi.org/10.1093/bioinformatics/bts412
49. T. Pahikkala, A. Airola, S. Pietila, S. Shakyawar, A. Szwajda, J. Tang, et al., Toward more realistic drug-target interaction predictions, *Brief Bioinf.*, **16** (2015), 325–337. https://doi.org/10.1093/bib/bbu010
50. G. Landrum, RDKit: Open-source cheminformatics, 2006. Available from: http://www.rdkit.org.
51. M. LJPvd, G. Hinton, Visualizing high-dimensional data using t-SNE, *J. Mach. Learn. Res.*, **9** (2008), 2579–2605.

AIMS Press

©2023 the Author(s), licensee AIMS Press. This is an open access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0)