IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received July 31, 2017, accepted September 2, 2017, date of publication September 21, 2017, date of current version October 25, 2017.

Digital Object Identifier 10.1109/ACCESS.2017.2755045

Predicting Drug Side Effects Using Data Analytics and the Integration of Multiple Data Sources

WEI-PO LEE[1], JHIH-YUAN HUANG[1], HSUAN-HAO CHANG[1], KING-TEH LEE[2], AND CHAO-TI LAI[3]

[1]Department of Information Management, National Sun Yat-sen University, Kaohsiung 80424, Taiwan
[2]Department of Healthcare Administration and Medical Informatics, Kaohsiung Medical University, Kaohsiung 80708, Taiwan
[3]Department of Nursing, Kaohsiung Medical University, Chung-Ho Memorial Hospital, Kaohsiung 80708, Taiwan

Corresponding author: Wei-Po Lee (wplee@mail.nsysu.edu.tw)

This work was supported in part by the Joint Project of National Sun Yat-sen University and Kaohsiung Medical University under Grant NSYSU-KMU-105-P-020.

ABSTRACT The development of automated approaches employing computational methods using data from publicly available drugs datasets for the prediction of drug side effects has been proposed. This paper presents the use of a hybrid machine learning approach to construct side effect classifiers using an appropriate set of data features. The presented approach utilizes the perspective of data analytics to investigate the effect of drug distribution in the feature space, categorize side effects into several intervals, adopt suitable strategies for each interval, and construct data models accordingly. To verify the applicability of the presented method in side effect prediction, a series of experiments were conducted. The results showed that this approach was able to take into account the characteristics of different types of side effects, thereby achieve better predictive performance. Moreover, different feature selection schemes were coupled with the modeling methods to examine the corresponding effects. In addition, analyses were performed to investigate the task difficulty in terms of data distance and similarity. Examples of visualized networks of associations between drugs and side effects are also discussed to further evaluate the results.

INDEX TERMS Drug side effect, data analytics, machine learning, predictive modeling, feature selection.

I. INTRODUCTION

Drug side effects (SEs) are a major cause of failure during drug development. Adverse side effects of medication can affect the quality of life among patients. Every year, many Food and Drug Administration (FDA) approved drugs are recalled because of their side effects, particularly when side effects are unexpected but discovered to be major concerns [1]. This process of post-market drug withdrawal is costly. Therefore, the ability to evaluate the potential side effects of drugs as early as possible is imperative during the drug design and development processes. The results of these evaluations can be used as guidance in the effort to reduce side effects and provide safe therapies in the clinical setting.

The assessment of potential SEs can be implemented during different stages of the drug development cycle, including the early stages of drug design, different phases of clinical trials, and post-marketing surveillance. Traditionally, preclinical *in vitro* safety pharmacology profiling has been used to predict side effects by testing compounds using biochemical and cellular assays. However, the experimental detection of SEs using extensive *in vitro* profiling remains challenging, mainly due to the cost and efficiency required [2], [3]. Post-market surveillance especially relies upon the spontaneous reports provided by physicians and patients through the Adverse Events Reporting system of the FDA, and it usually takes time to accumulate these reports in the form of a formal record [4], [5]. To address the problems associated with the cost and efficiency of SE detection during the drug discovery process, *in silico* approaches have been proposed to predict SEs by developing computational methods to assess the available large public datasets of drugs at both the preclinical and post-market stages (e.g., [6]–[8]).

When investigating drug-SE relationships, drugs can be regarded as molecules that introduce perturbations to a biological system consisting of various molecular interactions. These interactions include protein-protein interactions and metabolic and signal transduction pathways [9], [10]. The interactions of a drug with its targets may produce the anticipated therapeutic effects; however, off-target interactions may also occur and cause previously unexpected side effects. These comprehensive interactions are often difficult to predict, because both SEs and therapeutic effects occur are the result of the emergence of complex relationships between a drug and a biological system. Existing studies of SE prediction have focused on using the chemical structures or molecular pathways of the drugs for prediction or explaining side effects based on known drug targets and their pathways. Some SEs can be understood through analyzing the biological properties of their modulated targets, whereas others are better explained by considering only the chemical properties of the drug compound. Low affinity binding to proteins that are not usually considered drug targets and not normally associated with drug responses may also lead to side effects [11], [12]. Moreover, the complex effects of the inhibition of multiple targets often cannot be predicted based on a simple drug interaction profile [13], [14]. These studies suggest that no definitive methodology has been developed thus far to evaluate drug SEs, and the separation of biological and chemical factors often leads to incomplete models that are unable to provide a unified view of SEs [2], [15]. Therefore, different drug features have to be considered simultaneously in the prediction of side effects.

To predict side effects using computational methods, two major issues have to be addressed: the information used to characterize the drugs and the computational techniques used for making the prediction. Regarding the drug-related information, chemical and biological features are the two types of data that have been most frequently used in the relevant studies described above. Methods are developed based on the selected drug features to investigate the correlations between drug features and SEs. During the application of chemical structure-based approaches, drug side effects are usually evaluated in association with their chemical structures. For example, Scheiber *et al.* [16] performed a global analysis to identify the chemical substructures associated with known side effects. Yamanishi *et al.* [17] proposed a method to predict pharmacological and side effect data using chemical structures; however, their approach cannot be applied to predict high-dimensional side effect profiles. To achieve this goal, Pauwels *et al.* [9] developed a sparse canonical correlation analysis method to predict the high-dimensional side effect profiles of drug molecules based on their chemical structures. However, it may be difficult to select appropriate sparsity parameters and an appropriate number of components.

Approaches that consider biological information often use protein-target as features. The principle underlying these approaches is the idea that drugs with similar *in vitro* protein-binding profiles tend to exhibit similar side effects [18], as reported previously [10], [19]. Some methods have been developed to determine the association between drug SEs and perturbed biological pathways because these pathways shared the proteins that the drugs targeted. For example, Xie *et al.* developed a chemical systems biology approach to identify the off-targets of drugs. Then, the drug-protein interaction pair with the best score was mapped to the known biological pathways to identify the potential off-target binding networks of a drug [12]. However, these methods rely upon the availability of gene-expression data gathered during the chemical perturbations produced by the drugs. The performance of these methods depends heavily on the availability of data regarding the three-dimensional structures of proteins and the known biological pathways. These requirements therefore limit the applicability of these approaches in small-scale studies.

In addition to the aforementioned chemical and biological drug information, phenotypic information (e.g., indication) has been shown to be useful in drug-related studies, even though it was not often considered. Liu *et al.* [3] investigated the use of phenotypic information, together with chemical and biological properties, in SEs prediction. They also comprehensively evaluated different combinations of features to see how each feature set contributed to the prediction accuracy. The results showed that approaches involving the integration of chemical, biological, and phenotypic properties outperformed methods using only individual information. Wang *et al.* also conducted an experimental study, which showed that therapeutic indications are the information source most useful in the prediction of drug side effects [20].

In the prediction of side effects, the most relevant computational methods are those that employ machine learning techniques to build classifiers based on known drug-side effect associations. These methods have often been applied to determining the associations between different drug features, such as chemical structures, protein targets, molecular pathways, and phenotypic information. For example, Pauwels *et al.* used chemical structures as features and then applied popular machine-learning methods (including the k-nearest neighbor, the support vector machine, the ordinary canonical correlation analysis, and the sparse canonical correlation analysis) to train models for prediction [9]. Mizutani *et al.* combined chemical structures and target proteins as features, and adopted the sparse canonical correlation method to build prediction models [21], [22]. Additionally, Liu *et al.* [3] integrated a wide variety drug-related information as features and then used machine-learning techniques to train classifiers (including the logistic regression, the naïve Bayes, the k-nearest neighbor, the random forests and the support vector machine), and causality analysis was used to determine the molecular predictors of adverse drug reactions in [23].

In addition, some researchers have attempted to predict potential side effects based on the known side effects. For example, Cheng *et al.* [24] plotted drugs, side effect terms and the known side effects on a bipartite graph. The authors used the known side effects as initial resources and applied a network inference method based on resource allocation to infer potential side effects. Additionally, in the most recent work by Zhang *et al.*, the authors mapped approved drugs, side effect terms and drug-side effect associations to users, items, and user-to-item ratings and incorporated the derived drug predictions into recommendation tasks [25]. They presented two recommendation methods for predicting side effects, an extended neighborhood-based method (INBM) and a revised Boltzmann machine-based method (RBMBM). Extensive surveys of computational methods on side effect predictions can be found in [2] and [26].

Drawing on data-enriched web databases, this work attempts to develop a hybrid machine learning approach to construct side effect classifiers with an appropriate set of data features via the integration of different types of online knowledge resources. To investigate the effect of data (i.e., drug) distribution in the feature space, we categorized side effects into several types depending on the distribution of data in different classes and then adopted suitable strategies to build data models accordingly. To verify the presented approach, a series of experiments were conducted. The results showed that the presented approach could apply a data analytics perspective to the consideration of the characteristics of different types of side effects, thereby leading to better predictive performance. Different feature selections schemes were also evaluated to examine their effects. The use of *in silico* side effect prediction based on various drug features provides a prospective area of drug research that may facilitate the improvement of drug safety during and after clinical trials. In addition, examples of visualized networks of associations between drugs and side effects are also analyzed and discussed to further inspect the quantitative experimental results. These results confirm the feasibility and effectiveness of the approach developed herein.

II. MATERIALS AND METHODS

As mentioned above, the increasing number of drug side effects in the pharmaceutical industry indicates the need to determine the contributing factors underlying drug side effects and to develop automated methods to predict side effects. Recent studies have shown that SEs are most frequently caused by interactions between drugs and off-target proteins, and *in silico* approaches have thus focused on exploiting drug-target profiles to estimate the probability of clinical SEs. However, it is also known that some side effects occur as a consequence of non-specific interactions between drugs and reactive metabolites or enhanced cellular production of reactive oxygen or nitrogen species [27]. These reactions may be triggered by well-defined chemical features. In addition, though it is generally true that molecules with similar structure may exhibit similar biological activity, it is also known that small modifications of active compounds can improve (or decrease) their potency and these active compounds may be distinguished from inactive compounds by the presence of small chemical differences.

The aforementioned factors reveal the importance of using data mining or machine learning methods to explore these integrated effects. Drug, target and disease spaces may be evaluated in association to study the effect of drugs on different spaces. Therefore, we considered and evaluated multiple drug features to represent each drug data as a feature vector and adopted different data mining or machine learning methods to infer associations between drugs and side effects. From both perspectives of data science and medical science, the drug SE prediction is a challenge. The main focus of this work is to take the viewpoint of data analytics to develop a more effective approach to enhance the overall prediction performance. This computational method/model can be used to analyze chosen diseases with specific knowledge resources and clinical experiences [28]. The linkages between clinical drugs and meaningful side effects for a target disease can be further investigated, and then special attention can be paid to controlling and dosing of the drug in the clinical trials. For example, individual studies have been performed for neurological diseases emphasizing on Alzheimer [15], and cutaneous diseases on psoriasis [29]. Details of the data features and the predictive methods used are described in the following subsections.

A. DATA REPRESENTATION

In this work, three types of drug information were used to describe drugs from the chemical and biological perspectives, including their chemical substructures, associated proteins, and indications. According to their corresponding functions, the proteins were further categorized into four different types: target, enzyme, transporter, and carrier proteins. That is, there were six types of data features used to describe the drugs. Though pathways may also serve as useful biological features, they were not included in the binary vector representation defined here for prediction, as their inclusion would involve utilizing drug perturbed gene expression profiles, thereby limiting the large-scale applicability of the prediction model. The overall data representation is illustrated in Fig. 1.

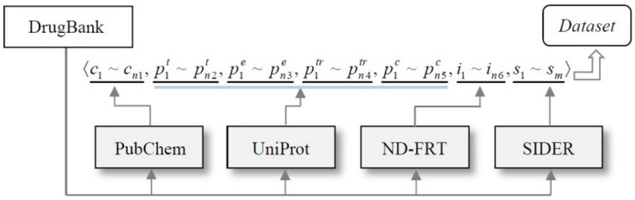

DrugBank
Dataset
$\langle c_1 \sim c_{n1}, p_1^t \sim p_{n2}^t, p_1^e \sim p_{n3}^e, p_1^{tr} \sim p_{n4}^{tr}, p_1^c \sim p_{n5}^c, i_1 \sim i_{n6}, s_1 \sim s_m \rangle$
PubChem
UniProt
ND-FRT
SIDER

FIGURE 1. Representation of drug data, in which n_1 to n_6 are the numbers of different types of features, and m is the number of side effects considered.

To retrieve chemical information regarding the substructures of the drugs, we used the popular public drug information database, DrugBank ([30]), to collect data on FDA-approved small-molecule drugs (with PubChem Compound ID) and map them to PubChem ([31]). Here, we used SIMES (Simplified Molecular Input Line Entry Specification) to translate information regarding the substructures of the drugs. The substructures were defined based on segment rules (obtained using the chemical toolbox, Open Babel) and converted to PF2 format (that uses integers 0 to 1020 to encode different substructures). In this way, the chemical substructures defined in PubChem could be encoded as binary features (i.e., $c_1 \sim c_{n1}$ in Fig. 1) as follows: the entry was 1 if the corresponding PubChem substructure is present in the drug; otherwise the entry was 0.

We also retrieved protein data for each drug, which were collected from DrugBank (with UniProt ID). The proteins were mapped to UniProt Knowledgebase [32], which is the knowledgebase includes the most comprehensive and complete information for proteins. Similar to the aforementioned binary feature representation used for chemical substructures, proteins (target, enzyme, transporter, and carrier) were encoded as binary features for each drug (p^t, p^e, p^{tr}, p^c in Fig. 1, respectively) to indicate the presence or absence of the corresponding proteins.

The third type of data we collected was regarding the therapeutic indications for the drugs. These indications were obtained by mapping the drug names from DrugBank to the treatment relationships between drugs and diseases. These relationships were extracted from the National Drug File-Reference Terminology, part of the Unified Medical Language System (UMLS) [33]. Again, the retrieved indications were encoded as binary features for each drug (i.e., $i_1 \sim i_{n6}$ in Fig. 1), with each feature indicating the presence or absence of the corresponding therapeutic indication.

In addition to organizing the aforementioned information, we extracted data on side effect (keywords) from the SIDER database ([34]), which contains information about medicines in market and their associated adverse drug reactions. SIDER uses STITCH compound identifiers to represent drugs (http://stitch.embl.de/cgi/show_download_page.pl), which can be mapped into PubChem compound identifiers to ensure consistency with other drug relevant data. Each side effect was regarded as a binary target class (i.e., with a label of positive or negative) to indicate and individually predict its occurrence in association with the data for each drug.

B. CLASSIFICATION METHODS FOR SIDE EFFECT PREDICTION

Using the aforementioned data representation, we regarded the drug side effect prediction task as a binary classification problem in the predictive modeling phase. That is, a binary classifier was built in the training phase for each side effect of the drugs (with or without causing this specific side effect), and the classifier was then used in the testing phase to predict the occurrence of side effects in association with new drugs. For every classifier, the input was drug related information (features), as described in the above section, and the output indicated the occurrence of the side effect of consideration. In this study, three representative data modeling methods were adopted and applied to the collected data, due to their computational efficiency. These approaches included a statistical-based algorithm, (i.e., Bayesian classifier), a distance-based algorithm, (i.e., k-nearest neighbor), and an ensemble learning algorithm (i.e., random forest). These methods have been most frequently used to compare the performance of prediction models in different experimental situations, and their results were analyzed and discussed. Though the SVM is also a popular data classification method, it was not adopted for use in this study because the existing literature has shown that compared to other methods, the prediction performance of the SVM is not robust and is side effect dependent [9]. This method is also inefficient (i.e., time-consuming) for classification cases that include a large number of data features.

The first method, the Bayesian classifier, is a probabilistic model where a classification is generated to relate a latent variable probabilistically to the observed variables (i.e., side effects). The classification then becomes an inference in the probabilistic model to predict class membership probabilities (e.g., the probability a data vector belonging to a particular class). The second method, the nearest neighbor method, generates predictions for a drug based on the conclusions (often, the voting results) of its nearest neighbors. Therefore, in the application of this approach, the measurement of similarity between data instances (i.e., drugs) is most important. In this method, the Euclidean distance is used as a measure of similarity. The third method, random forest, is a type of ensemble machine learning algorithm called Bootstrap Aggregation (or bagging). The main principle behind this type of method is that a group of weak classifiers (decision trees) can be used together to form a classifier with better performance. It combines predictions from multiple models in ensembles and has been found to perform better if the predictions from the sub-models are uncorrelated or, at most, weakly correlated. The details of the aforementioned classification methods are described in [35]. These approaches were adopted to perform side effect prediction, and the development of a hybrid approach for performance enhancement is presented in the next subsection.

Though the machine learning procedure is an efficient and convenient method to construct data models based on known data instances that are subsequently used to predict unknown data, to ensure the success of this approach, the problem of class imbalance must be overcome [36], [37]. Class imbalance means that in a classification task, when the numbers of data instances within each class are quite different, the classification performance of the standard classifier may be damaged. Class imbalance is a crucial problem in the machine learning community since data are often distributed unequally in real world applications. Corresponding to a classification task, in this study, the input was a set of drug features and the output indicated the classes of side effects (i.e., positive or negative for drugs with or without side effect, respectively). For a specific side effect, the number of FDA-approved drugs associated with this side effect is often smaller than that of those that are not associated, which means that the data instances for a specific side effect may distribute unequally in different classes of data, and consequently, the models cannot be successfully learned. An additional data balancing procedure was thus required to conciliate this problem.

C. A DELIBERATE METHOD FOR SIDE EFFECT PREDICTION

Taking into account both predictive performance of the approach and the realistic data distribution, we divided all side effects into three intervals and applied different prediction strategies based on the number of drugs causing each side effect. Fig. 2 illustrates these situations. The first category included side effects currently known to be caused by a relatively small number of drugs. The prediction of a side effect in this category for a specific (new) drug d was determined based on the average distance (dissimilarity) between this drug and the two classes of its nearest neighbors (i.e., drugs with and without this side effect). That is, the nearest neighbors of drug d were first selected, and the two average distances d_p and d_n (between drug d and the selected drugs with and without the side effect of consideration, respectively) were calculated. If the neighbors belonged to the same class, a default maximum value (representing infinite) was assigned as the average distance for the missing class. Then, the two distances, d_p and d_n, were compared, and the shorter distance was used to determine the classification result (i.e., with or without this side effect). The rationale for using this strategy was that side effects belonging to this interval are caused by a small number of drugs, and the two classes of drugs with and without SEs were likely to be seriously imbalanced. Neither over-sampling nor under-sampling techniques could be used to effectively solve this problem. Thus, a simple, reasonable and workable distance-comparison strategy was adopted, and the experimental results confirmed its feasibility.

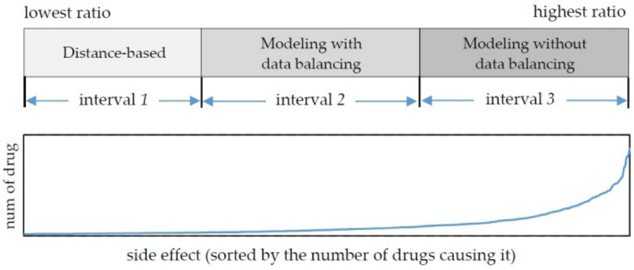


FIGURE 2. Ratio of the two classes of drug data (without over with) for each side effect (colors indicate intervals).

In contrast, for side effects included in the second and the third intervals, classification models were built. For side effects in the second category, the aforementioned data balancing strategy was used because while the number of side effect-causing drugs reached a certain level of representativeness, the two classes of data were not yet balanced. No further data processing was needed for the side effects included in the third category since the two classes contained approximately the same number of data records.

D. FEATURE SELECTION

With the above four types of features, a feature selection scheme can be performed to choose a subset of the original features to maximize the performance of a model-learning algorithm. In this way, the dimension of feature vectors can be reduced, and which often can reduce overfitting and the computational effort in learning a model (classifier). In this study, three popular features selection methods were applied to the original data, including univariate selection (chi-squared and Pearson correlation coefficient), feature importance, and principal component analysis. They were used to eliminate irrelevant and redundant attributes. Each feature selection method has its specific characteristics. Univariate selection performs statistical test for non-negative features to select k of the best features. Principal component analysis uses linear algebra to transform the dataset into a compressed form. Feature importance adopts bagged decision trees, such as random forest and extra trees to estimate the importance of features. More details on the above three algorithms can be found in [17], and the evaluation results are presented in the experimental section.

E. SIMILARITY MEASURES OF THE DRUGS

In addition to the descriptions of the data modeling and prediction methods, this section briefly describes how data similarity was calculated. Two types of measurement methods were developed to quantitatively evaluate the similarity of drug-pairs. The first method applied was to directly measure this distance (dissimilarity) using a binary representation, as previously described. This type of similarity was used in the data modeling procedure and result analysis (the k-nearest neighbor method) due to its succinctness and simplicity. The other method applied was to separately measure the similarities of different types of features with their corresponding domain-dependent strategies (from chemical to biological perspectives) and then aggregate the results. This type of similarity was not used in the data modeling procedure because it involves additional calculations and feature weighting. Rather, it was only used in the experimental section of the result analysis to provide an alternative perspective regarding the distribution of data.

In the first measurement method, a valid distance measure should be symmetric and have an obtainable minimum value (usually zero) to obtain the distance between two drugs represented by binary feature vectors. The distance between two data instances can be calculated using the Minkowski distance, and the most commonly used value of order in this distance formula is 2 (Euclidean distance). The Euclidean distance measure was applied during the performance of drug-drug dissimilarity calculations. As indicated in previous studies, using chemical features and biological features alone is not enough to derive the relationships between drugs and side effects; thus, we took all drug features into account. That is, for any two drugs, d_x, d_y, represented as binary feature vectors, the similarity between them was calculated using the following equation:

$$distance\ (d_x, d_y) = \sqrt{\sum\nolimits_{k=1}^{m} \left|d_{x,k} - d_{y,k}\right|} \tag{1}$$

Here, $d_{x,k}$, $d_{y,k}$ represent the drug features and m is the total number of drug features described above.

In contrast, in the second type of measurement, the similarity between any two drugs was obtained by individually calculating the similarity between each type of feature considered (see below) and then summing them with weighting factors. The first type of critical features indicated that the presence of a chemical substructure that could be used alone to calculate the similarity between two drugs. To measure the similarity in terms of their chemical substructures, we retrieved the chemical information for these drugs and translated the information into substructures, as described in Section II.*A*. In this method, each drug (d) could be represented as a set of integers ($h(d)$) indicating which substructures it included. Then, the similarity of two drugs, d_x and d_y, could be determined using the following equation:

$$Sim_{chem}(d_x, d_y) = \frac{\left|h(d_x) \cap h(d_y)\right|}{\left|h(d_x) \cup h(d_y)\right|} \tag{2}$$

in which $|h(d_x) \cap h(d_y)|$ is the number of substructures contained in both d_x and d_y; and $|h(d_x) \cup h(d_y)|$ represents the total number of different substructures included in the two drugs.

The second type of critical drug feature included was protein, and the proteins related to each drug were retrieved from the DrugBank database. As indicated above, the drug-related proteins were categorized into the following four types according to their roles and functions in the therapeutic mechanism: target, enzyme, transporter, and carrier. To calculate the similarity between two drugs, d_x, d_y, in terms of the relevant proteins required the measurement of the similarity of the gene sequences of the proteins, which was achieved using the equation below:

$$Sim_{protein}(d_x, d_y) = \frac{1}{|P(d_x)|\,|P(d_y)|} \sum_{i=1}^{|P(d_x)|} \sum_{j=1}^{|P(d_y)|} g\left(P_i(d_x), P_j(d_y)\right) \tag{3}$$

In the above equation, the protein could be any one of the four types of proteins mentioned; $P(d_x)$ and $P(d_y)$ were the sets of proteins related to d_x and d_y, respectively, $|P(d_x)|$ and $|P(d_y)|$ were the numbers of proteins in the corresponding sets; and the function g represented the similarity of the gene sequences included in the two proteins specified (i.e., $P_i(d_x)$ and $P_j(d_y)$). In this study, we calculated the Smith-Waterman Sequence Alignment Score (a dynamic programming method through which the longest common subsequence for two sequences is identified to obtain an alignment score, the details of which are described in [38]). The aforementioned measurement process was applied to all the four types of proteins (target, enzyme, transporter, and carrier) to obtain a similarity value for each type of protein (i.e., Sim_{target}, Sim_{enzyme}, $Sim_{transporter}$, $Sim_{carrier}$).

The third type of critical drug features included was the therapeutic indication. As described in Section II.*A*, the indications constituted a list of therapeutic relationships between drugs and diseases. Therefore, the binary distance measurement described in equation (1) could be directly used to calculate the indication similarity between two drugs.

The fourth type of feature included was the biological pathway. Though the pathway information was not encoded in the drug representation for the prediction task, this feature provided an alternative way to measure task difficulty and was used for the experimental analysis described in a later section. We thus describe how the pathway similarity is calculated here. Pathway information was obtained from the SMPDB (The Small Molecule Pathway Database, [39]). To measure the similarity between two drugs, d_x and d_y, we first retrieved and compared the pathways involved in d_x and d_y. Given the two sets of pathways $path(d_x)$ and $path(d_y)$ for d_x and d_y, respectively, the Jaccard similarity equation was used to calculate the pathway similarity of d_x and d_y, as described below:

$$Sim_{path}(d_x, d_y) = \frac{\left|path(d_x) \cap path(d_y)\right|}{\left|path(d_x) \cup path(d_y)\right|} \tag{4}$$

After separately calculating the aforementioned similarity measures, we integrated the effects of all critical factors to determine the overall drug similarity. Here, the linear combination (i.e., weighted sum) of all factors was adopted, as described below:

$$Sim_{all}(d_x, d_y) = \sum_{i \in F} w_i \times Sim_i(d_x, d_y) \tag{5}$$

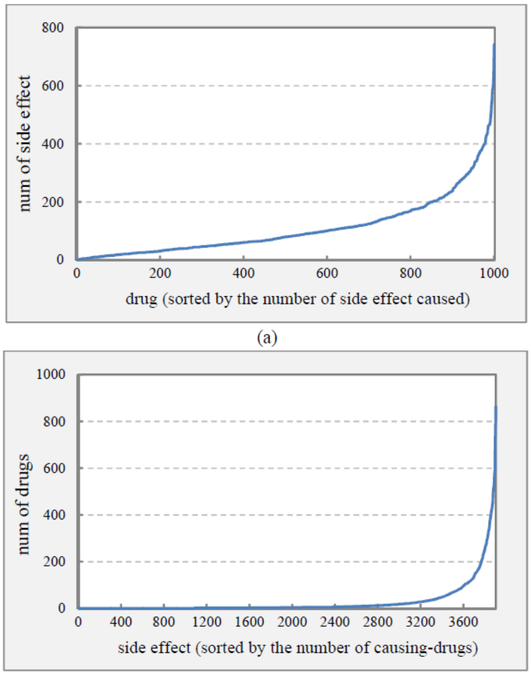
Here, in equation (5), F was the set of factors considered as mentioned above, which included $\{chem, target, enzyme, transporter, carrier, path\}$, and the weights of these factors were determined by applying a preliminary test procedure to each dataset.

III. EVALUATIONS AND RESULTS

A. DATASETS

To evaluate the performance of our proposed drug side effect prediction approach, four datasets were collected and used. The first (called dataset-A) included the FDA-approved small-molecule drugs obtained from DrugBank, with the chemical substructures defined in PubChem and corresponding drug side effects retrieved from SIDER. This dataset was used in the first series of experiments, which included the performance of extensive trials to verify the proposed approach, and then in-depth analyses were performed to assess the proposed approach from different perspectives. Before conducting the experiments, we briefly analyzed the dataset used for the first series of experiments (i.e., dataset-A), which included 1002 different drugs and 3903 side effects, and a total of 7257 features were used to characterize the drugs.

To observe the distribution of drug data over side effects (and vice versa), we provided an overview of dataset-A by plotting the number of side effects caused by each drug (sorted in decreasing order), which are described in Fig. 3(a), and the number of drugs causing each side effect, which are illustrated in Fig. 3 (b). We observed that 4.49% of drugs had less than 10 (inclusively) side effects; 55.29% of drugs had between 10 and 100 different side effects; 39.32% of drugs had between 100 and 500 side effects; and 0.9% of drugs had more than 500 side effects. Additionally,



(a)

(b)

FIGURE 3. Distribution of drugs and side effects included in dataset-A.

Fig. 3(b) shows that 0.64% of side effects had a high rate of occurrence (occurring in association with more than 500 drugs); 6.87% of side effects occurred in association with 100-500 drugs; 22.85% of all side effects occurred in association with 10-100 drugs; and 69.64% of side effects occurred in association with less than 10 drugs. These figures, which illustrate the characteristics of dataset-A, were used to determine the data intervals using the previously presented hybrid approach.

The other three datasets included the Pauwels's dataset [9], Mizutani's dataset [21] and Liu's dataset [3]. These datasets were selected because they are popular and publicly available, and thus adopted in the second series of experiments for further performance evaluation and comparison. The aforementioned three datasets (i.e., Pauwels's, Mizutani's and Liu's datasets) used have 888, 658, 832 drugs and 1385, 1339, 1385 side effects, respectively. Details of these datasets are described in the original studies [3], [9], [21].

B. PERFORMANCE METRICS

In the experiments, we employed several criteria that are frequently used in binary classification to evaluate the utility of the different methods in the prediction of side effect. We first measured the true positive (TP), false positive (FP), true negative (TN), and false negative (FN) rates and then use them to calculate various performance metrics. The first performance metric was accuracy, defined as the proportion of correctly predicted instances relative to all predicted instances. The second performance metric was precision, defined as the proportion of retrieved instances that were relevant. The third performance metric was recall, defined as the proportion of relevant instances that were retrieved. Although often in conflicting in nature, the measures of precision and recall are both important in evaluating the performance of a prediction approach. Therefore, these two measures can be combined with equal weights to obtain a single metric, the F-measure. The four previously described performance evaluation metrics were defined using the following equations:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

$$precision = \frac{TP}{TP + FP} \tag{7}$$

$$recall = \frac{TP}{TP + FN} \tag{8}$$

$$F - measure = 2 \times \frac{precision \times recall}{precision + recall} \tag{9}$$

In addition, the area under the ROC (Receiver Operating Characteristic) curve (called AUC) was calculated to evaluate the predictive performance. AUC score has now been widely used as a classification performance measure in biomedical informatics. The ROC curve is obtained by plotting the true positive rate (y-axis) against false positive rate (x-axis) at different discrimination threshold values for prediction score. Then, using the generated curve, results may be positive (above the curve) or negative (below the curve). After calculating the aforementioned performance metrics for each side effect classifier, we summarized the performance across all of the considered side effects. The measures over all SEs (non-SE were excluded) classifiers were averaged to obtain an overall score, which is reported in the following sections.

TABLE 1. Results of different classifiers.

Dataset	Method	accuracy	precision	recall	F	AUC
Dataset-A	NB	0.919	0.377	0.431	0.402	0.700
	k-NN	0.930	0.615	0.235	0.340	0.745
	RF	0.951	0.710	0.304	0.426	0.894
	Three intervals	0.952	0.711	0.714	0.713	0.987

C. EVALUATIONS OF SIDE EFFECT PREDICTION

At the first stage of evaluation, we adopted three popular computationally efficient methods, Bayes, k-NN, and RF, and applied them to the integrated dataset (i.e., Dataset-A) to examine the performance of the approach using the combined features for side effect prediction. The 10-fold cross validation method was used to perform a more objective assessment. Table 1 summarizes the results in terms of the five performance metrics in classification, including accuracy, precision, recall, F-measures and AUCs. Notably that the results reported here were averaged over all SE classifiers only (not including non-SE classifiers that usually gave high values in a dataset dominated by non-SE data); therefore

TABLE 2. Comparison of the performance of the proposed method with those of other approaches for the three available datasets.

Dataset	Method	accuracy	precision	recall	F	AUC
Pauwels	Pauwels's method [8]	0.931	0.361	0.517	0.425	0.897
	Liu's method [2]	0.934	0.400	0.643	0.493	0.920
	Cheng's method [22]	0.955	0.547	0.587	0.566	0.922
	RBMBM [23]	0.958	0.579	0.605	0.592	0.941
	INBM [23]	0.961	0.605	0.608	0.607	0.934
	Avg. scoring ensemble model [23]	0.962	0.612	0.621	0.616	0.949
	Three intervals (this work)	0.916	0.590	0.673	0.629	0.972
Mizutani	Mizutani's method [19]	0.927	0.387	0.527	0.446	0.890
	Liu's method [2]	0.930	0.418	0.637	0.505	0.918
	Cheng's method [22]	0.951	0.560	0.593	0.576	0.923
	RBMBM [23]	0.954	0.581	0.614	0.597	0.939
	INBM [23]	0.956	0.605	0.616	0.611	0.932
	Avg. scoring ensemble model [23]	0.958	0.619	0.624	0.622	0.946
	Three intervals (this work)	0.909	0.522	0.749	0.615	0.970
Liu	Liu's method [2]	0.917	0.341	0.669	0.452	0.907
	Cheng's method [22]	0.954	0.550	0.589	0.569	0.922
	RBMBM [23]	0.957	0.581	0.608	0.594	0.941
	INBM [23]	0.959	0.606	0.607	0.606	0.934
	Avg. scoring ensemble model [23]	0.960	0.611	0.623	0.617	0.948
	Three intervals (this work)	0.908	0.554	0.657	0.601	0.976

the performance became relatively low, compared to those studies aggregating both SE and non-SE classifiers to obtain the final performance. From this table, we observed that all the three methods had relative high accuracy but low precision, recall, and F-measures. These results indicate that the performance of the classifiers had been overestimated. This has been a problem in drug side effect prediction, and required extra effort to tackle [20], [40]. The other problem was that datasets used for side effect prediction are often not balanced because they contain few SE-causing drugs and many non-SE-causing drugs. In the study conducted by [20], the authors incorporated a sample balancing strategy into their data modeling method to alleviate this problem. The results of this study showed that though the situation could be improved by the application of such a strategy, recall remained at a low level (with a rate of 0.24-0.52, depending on the information resources used in the training phase).

To investigate the effect of data balancing, we also conducted a series of trials for a comparison of performance. Similar to the results reported in relevant studies, the overall predictive performance was improved after the training dataset was balanced by randomly sampling the original data within certain classes with a smaller number of records (here, classes of data with side effects). However, it is notable that the overall performance of the approach may not truly reflect its real world performance. That is, for side effects caused by a very small number of drugs (for example, only drug Ropinirole (DrugBank identifier DB00268) was associated with a side effect of abdominal adhesions according to the database), randomly sampling the data from the positive class resulted in the repetitive duplication of this small set of data. As a result, the data in the training and testing sets became consistent, leading to the identification of an unrealistically high predictive performance.

As described in Section II.*C*, we develop a deliberate and practical interval-based method to predict the drug side effects associated with each interval using corresponding strategies. In the aforementioned dataset, side effects were separated into the following three intervals: those caused by less than 100 drugs, those caused by 100 to 500 drugs, and those caused by more than 500 drugs, respectively. The number of intervals and boundary values depended upon the dataset used. The values indicated above were simply used as an example to demonstrate the effect of our proposed method. The last row in Table 1 shows the prediction results of the presented method (i.e., three intervals). As shown, our method appeared to perform better than the standard approaches, especially regarding the most important performance measures, recall and the F-measure. In addition to the performance issue, our hybrid approach could address extreme cases of data imbalance (that were not taken into account in relevant studies) in which side effects are caused only by very few drugs, as described above. These results demonstrated the effectiveness and usefulness of our data analytics method in practice.

To further examine the generalizability of the presented hybrid method, we applied this method to the prediction of drug side effects for the three additional datasets mentioned in Section III.*A*. Table 2 summaries the results of our approach and results obtained using various approaches proposed in well-known studies, which were used to compare their performance with that of our method. The results of the other approaches are extracted directly from the original studies. These results indicated that, on average, the presented

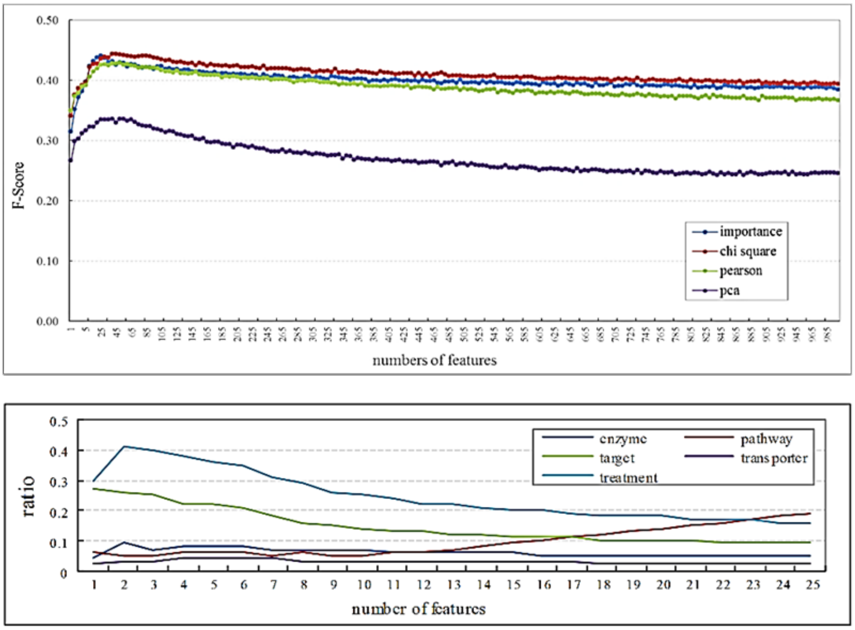


FIGURE 4. F-measure of prediction based on selected features and ratios of different types of feature.

approach demonstrated better results in terms of recall and the F-measure, the two most important factors in prediction, while maintaining an accuracy above an appropriate level of 0.9. These results, again, verified the performance of the proposed method for predicting side effects in practical applications.

D. EFFECTS OF FEATURE SELECTION

The second set of experiments was to investigate the effect of applying feature selection schemes to the original data for dimension reduction. As mentioned before, in the experiments, three features selection methods were evaluated: univariate selection (chi-squared and Pearson correlation coefficient), feature importance, and principal component analysis. Based on the experimental results shown in the above section, the random forest method (which gave the best performance among the original methods) was selected to work with the feature selection methods for evaluation.

Theoretically speaking, each feature contributes to the classification process. However, this does not mean that more features equal better results. To further examine the effect of different numbers of features selected by the three methods, we performed a feature profiling trial. Fig. 4 presents the representative results when the feature selection methods were applied to the dataset with a largest number of data features (i.e., Liu's dataset). Similar results were obtained for other datasets. The upper part of the figure shows the F-measure (that concluded both precision and recall together) under different feature combinations when the number of selected features increased (indicated in the x-axis). The lower part of the figure provides the ratio of each type of features to be selected during the selection of the first twenty-five features. These results show that increasing the number of features enhances the prediction performance. As shown, in this example the best result (i.e., 0.45) is obtained when thirty-eight features were selected, which is better than the result (i.e., 0.41) of using all features. It can also be seen that there was no obvious improvement when the number of features selected was increased to a certain extent. For instance, for each feature selection method, there were only slight changes when the number of features selected reached twenty. However, it is notable that the performance of F-measure declined when more and more features were selected. Here, a feature selection scheme aims to find the subset of useful features (variables) and exclude the redundant ones for building a good predictor. From a viewpoint of machine learning, redundant features often provide no extra information about the classes and thus are taken as noises by the predictor. The predictors that use them in the training phase will have poor generalization and performance.

IV. ANALYSIS AND DISCUSSION

A. MEASURING PREDICTION DIFFICULTY THROUGH DATA DISTANCE

Notably, in the results presented in the above section, the recall associated with all of the methods was not as precise as the accuracy obtained, though the proposed approach demonstrated enhanced performance. To investigate the reasons for this phenomenon, we further inspected and analyzed the results of all side effects in detail. After examining the results, we discovered that certain side effects (those caused by approximately 50-70 drugs) were more difficult to predict than others. Using the data balance strategy improved the classification results for these side effects in all performance

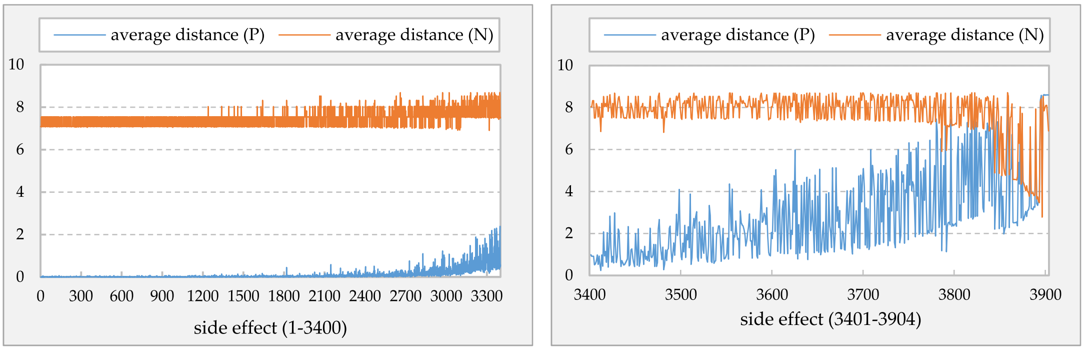


FIGURE 5. The average distance between drugs with and without certain side effects.

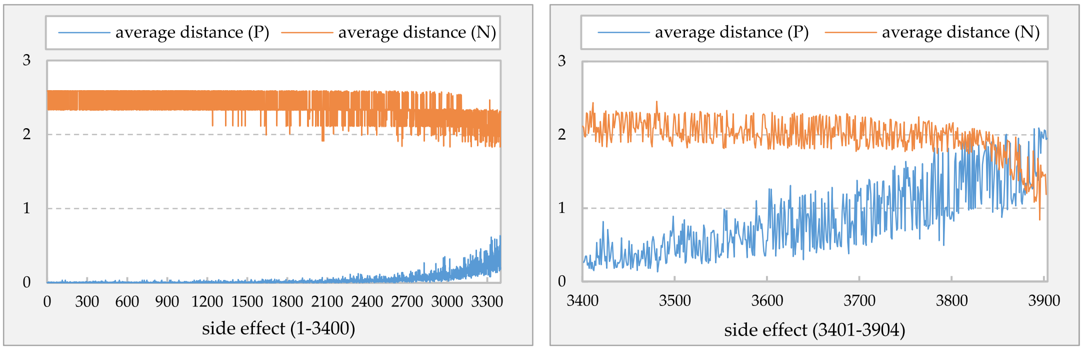


FIGURE 6. The similarity-based average distance between drugs with and without certain side effects.

metrics, while the success rates remained relatively low compared with those of most of the other side effects in the same dataset.

Using a data analysis perspective, we considered the aforementioned problem as a data differentiation issue and, thus, investigated the drug distribution within the data feature space to measure the task difficulty. For each side effect, we calculated the distance (and similarity) between drugs to determine how close drugs with and without a given side effect were in the feature space. The two measurement methods described in Section II.*E* were adopted. The first method was used to compare two types of average distances (one is the distance between the test drug and drugs with a specific SE; and the other, the distance between the test drug and drugs without this SE), during the process by using the k-nearest neighbor method to generate predictions. The distance was measured in terms of binary features using equation (1). Fig. 5 illustrates the results for dataset-A as an example, in which the x-axis represents the side effects sorted by the number of drugs causing them, and the y-axis represents the average distance between the test drug and the already known drugs. In the figure, P and N represent positive and negative neighbors (with and without a specific side effect) of the test drug, respectively. As seen in the figure, for side effects caused by only a few drugs (shown in Fig. 5 (left)), the two types of average distances were quite different. This makes the drugs with and without SEs distinguishable, suggesting that the distance measures could be used to make a successful prediction. In contrast, for side effects caused by a relatively large number of drugs, as shown in Fig. 5 (right), the two types of average distances were similar. That is, it became more difficult to distinguish between the two types of drugs, especially for SEs presented on the right-hand side of Fig. 5 (right). This innate data characteristic introduced difficulty in building a successful classifier.

In addition to binary vector-based distance measurement, the second method was used to calculate the average similarity-based distance between drugs to differentiate between drugs with and without side effects. The similarity was determined as the weighted combination of drug features described in Section II.*E*, including substructure, protein, pathway, and indication (i.e., equation (5)). Without losing generalizability, the results reported here are based on the combination of weights determined by the results of a preliminary test. Fig. 6 shows the combined similarity-based distance. As seen in this figure, the results were very similar to those illustrated in Fig. 5. That is, if it was difficult to distinguish between the two types of drugs using their average distance, it was also difficult to achieve the same task using their average similarity. The aforementioned analysis clarified the reason why certain side effects were more difficult to predict than others from a quantitative measurement of data distance perspective. In other words, it was difficult to distinguish between the two types of drugs with and without

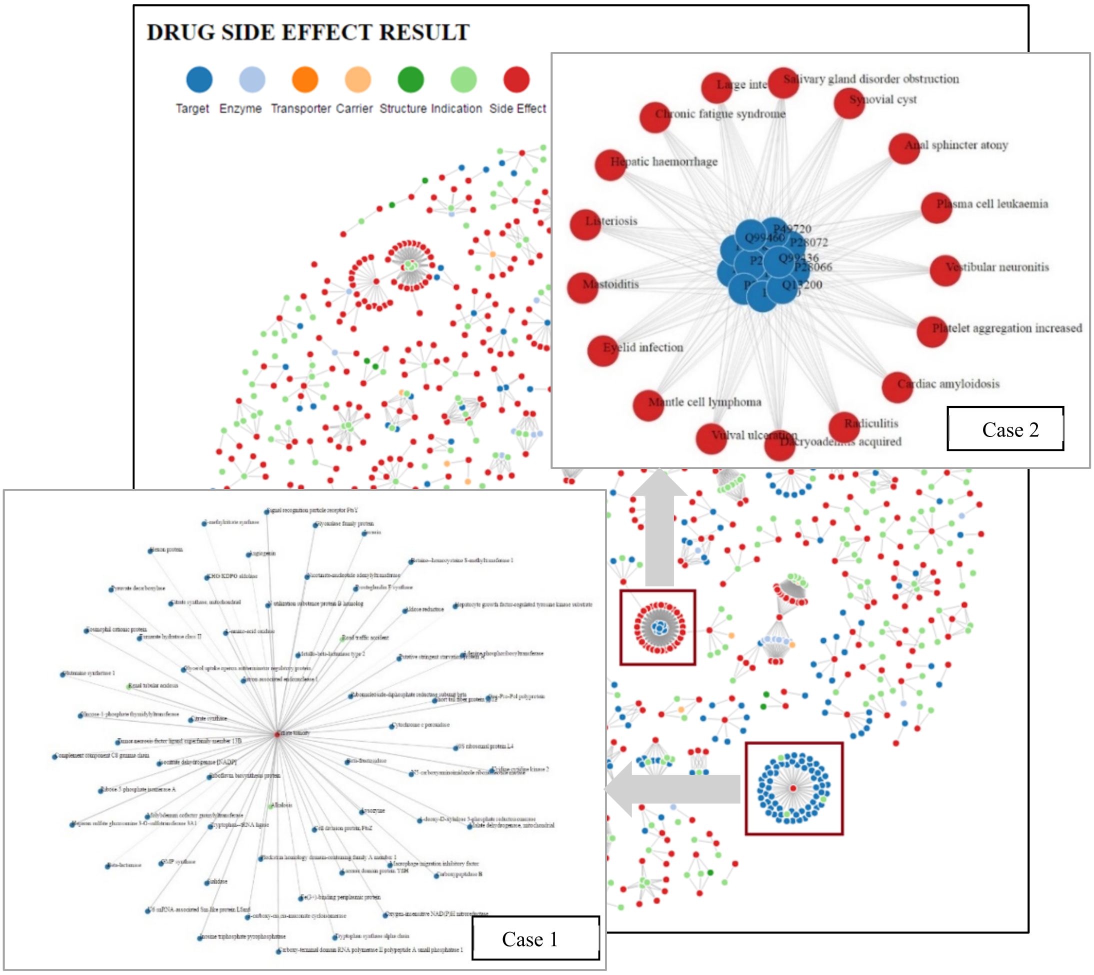


FIGURE 7. Visualization of the associations between drugs and side effects (the two cases described in the text are illustrated only in part to fit in the figure size).

these side effects in terms of the numerical features available. To enhance the performance, additional data features should be investigated.

B. CORRELATION AND VISUALIZATION

To further investigate the factors contributing to and co-occurring with drug side effects (i.e., the drug features directly related to each side effect), we measured the Pearson correlation coefficient for each drug feature and side effect pair (as mentioned previously, 7257 features and 3903 side effects were considered in this study). Based on the correlations derived from the drug features and side effects, we obtained 2927 direct relationships with a correlation coefficient of 1. Among these associations, 994 (out of 4003) were target proteins, 109 (out of 226) were enzymes, 8 (out of 24) were carrier proteins, 15 (out of 881) were substructures and 1744 (out of 2005) were indications. These critical factors should be taken into consideration in rational drug design. The results also showed that target proteins and indications for the drugs were the most important determining factors when predicting the occurrence of side effects. These results are consistent with those of previous studies that performed extensive experiments to confirm the importance of targets [41]–[43] and indications [20] in side effect prediction.

We constructed a graph including the side effects and the six types of drug features as nodes with links representing strong associations measured, as described above (i.e., with a Pearson correlation coefficient of 1). Fig. 7 illustrates the graph, in which the six types of features are indicated with different colors. The figure indicates that biological features were more directly correlated with drug side effects than were chemical features. We also observed that in this dataset, the nodes with links connecting to side effects were mostly targets (proteins), and the nodes for the substructures were less frequently linked to side effects. In fact, other studies have shown that drugs with shared targets, or those close in the interactome network, often share similar side effects [44], [45]. With the aid of this data visualization, it was much easier for users to determine the associations between drugs and side effects.

Two cases were taken from the figure to serve as representatives to illustrate the advantages and practicability of data visualization, rather than providing detailed case studies and applications. The first example (marked as case 1 in the figure, in which a single red node is surrounded by a group of blue nodes) is the node for the side effect of citrate toxicity, which had a large number of links connected to target nodes of drugs, such as citrate synthase, isocitrate dehydrogenase, and malate dehydrogenase. Citrate engages in the tricarboxylic acid cycle. It is a series of chemical reactions used by all aerobic organisms to generate energy through the oxidation of acetyl-CoA, beginning with the transfer of acetyl group from acetyl–CoA to oxaloacetate to form citrate. Citrate toxicity is primarily a result of hypocalcemia (decreased Ca as normal to bound to citrate) and metabolic effects. The metabolism occurs predominately in the livers and kidneys, and dysfunctions in these organs tend to cause problems in citrate clearance therefore become risk factors for citrate toxicity.

In contrast, the second example is the drug node bortezomib (Velcade, formerly PS-341), which causes several side effects. This anticancer drug is a type of proteasome inhibitor that is used to treat plasma cell myeloma by inhibiting and depleting the malignant myeloma cells. However, it often changes the levels of certain proteins, potentially causing many side effects (i.e., strongly related to these side effects), such as radiculitis (peripheral neuropathy), chronic fatigue syndrome, leukopenia, gastrointestinal symptoms, and cutaneous eruption. The associations between this drug and the side effects it may cause can be observed clearly from the figure. These figures, which illustrate the associations between drugs and side effects, provide not only the overall perspective but also helpful references for users to consult explanations in detail.

V. CONCLUSION

Given the cost and efficiency required for the prediction of SEs during the drug discovery process, automated approaches have been proposed to predict SEs through the application of computational methods using data from the available large public datasets of drugs at both the preclinical and post-market stages. Because simply using chemical or biological information alone may not be sufficient to capture the interactions and relationships between drugs and proteins, a deliberate and effective approach should take into account both types of features. In this work, we utilized a hybrid machine learning approach to construct side effect classifiers using an appropriate set of data features by integrating different types of online knowledge resources. This approach utilized the perspective of data analytics to investigate the effect of drug distribution in the feature space and categorize side effects into several intervals depending on the distribution of different classes of data. Then, we adopted suitable strategies for each interval to build data models accordingly. To verify the presented method, a series of experiments were conducted to demonstrate its utility in side effect prediction. The results showed that the presented approach was able to take into account the characteristics of different types of side effects, thereby leading to better predictive performance. Additional analyses were performed to investigate task difficulty in terms of data similarity for each side effect. Moreover, examples of visualized networks of associations between drugs and side effects were analyzed and discussed to further evaluate the quantitative experimental results. These results confirmed the feasibility and effectiveness of the developed approach.

Overall, the development of computational approaches for the prediction of SEs based on collective drug features is a trend that could significantly improve drug safety and decrease attrition rates in the future. However, the results of this study indicate that with the drug features currently used, some side effects were more difficult to predict than others. This result suggests that some essential targets governing the occurrence of these side effects must be carefully analyzed, and additional deterministic features need to be extracted and defined. These specific features should require resources of knowledge and reports of clinical trials of a certain disease domain. We are investigating how to capture the features of these essential targets to generate even more precise predictions. In addition to the prediction performance, for the near future we aim to examine the model interpretability, another issue also important from the perspective of clinicians. All mean to make the inferred models with their parameters statistically relevant as well as clinically meaningful. After coupling with specific resources of target diseases, drugs, and side effects as mentioned in Section II, relatively simple models with some intrinsic properties (such as the multiple regression models and the nearest neighbor models) provide potentials to support the capabilities of interpretability.

REFERENCES

[1] J. C. Veeren and M. Weiss, "Trends in emergency hospital admissions in England due to adverse drug reactions: 2008–2015," *J. Pharmaceutical Health Services Res.*, vol. 8, no. 1, pp. 5–11, 2016.

[2] S. M. Ivanov, A. A. Lagunin, and V. V. Poroikov, "In silico assessment of adverse drug reactions and associated mechanisms," *Drug Discovery Today*, vol. 21, no. 1, pp. 58–71, 2016.

[3] M. Liu *et al.*, "Large-scale prediction of adverse drug reactions using chemical, biological, and phenotypic properties of drugs," *J. Amer. Med. Inf. Assoc.*, vol. 19, pp. e28–e35, Jun. 2012.

[4] R. Harpaz, H. Perez, H. S. Chase, R. Rabadan, G. Hripcsak, and C. Friedman, "Biclustering of adverse drug events in the FDA's spontaneous reporting system," *Clin. Pharmacol. Therapeutics*, vol. 89, no. 2, pp. 243–250, 2011.

[5] N. P. Tatonetti *et al.*, "Detecting drug interactions from adverse-event reports: Interaction between paroxetine and pravastatin increases blood glucose levels," *Clin. Pharmacol. Therapeutics*, vol. 90, no. 1, pp. 133–142, 2011.

[6] E. Bresso *et al.*, "Integrative relational machine-learning for understanding drug side-effect profiles," *BMC Bioinf.*, vol. 14, Jun. 2013, Art. no. 207.

[7] R. Garcia-Serna, D. Vidal, N. Remez, and J. Mestres, "Large-scale predictive drug safety: From structural alerts to biological mechanisms," *Chem. Res. Toxicol.*, vol. 28, no. 10, pp. 1875–1887, 2015.

[8] H. Iwata, R. Sawada, S. Mizutani, M. Kotera, and Y. Yamanishi, "Large-scale prediction of beneficial drug combinations using drug efficacy and target profiles," *J. Chem. Inf. Model.*, vol. 55, no. 12, pp. 2705–2716, 2015.

[9] E. Pauwels, V. Stoven, and Y. Yamanishi, "Predicting drug side-effect profiles: A chemical fragment-based approach," *BMC Bioinf.*, vol. 12, May 2011, Art. no. 169.

[10] V. I. Pérez-Nueno, M. Souchet, A. S. Karaboga, and D. W. Ritchie, "GESSE: Predicting drug side effects from drug–target relationships," *J. Chem. Inf. Model.*, vol. 55, no. 9, pp. 1804–1823, 2015.

[11] R. L. Chang, L. Xie, P. E. Bourne, and B. O. Palsson, "Drug off-target effects predicted using structural analysis in the context of a metabolic network model," *PLoS Comput. Biol.*, vol. 6, no. 9, p. e1000938, 2010.

[12] L. Xie, J. Li, and P. E. Bourne, "Drug discovery using chemical systems biology: Identification of the protein-ligand binding network to explain the side effects of CETP inhibitors," *PLoS Comput. Biol.*, vol. 5, no. 5 p. e1000387, 2009.

[13] M. Duran-Frigola and P. Aloy, "Analysis of chemical and biological features yields mechanistic insights into drug side effects," *Chem. Biol.*, vol. 20, no. 4, pp. 594–603, 2013.

[14] T. Liu and R. B. Altman, "Relating essential proteins to drug side-effects using canonical component analysis: A structure-based approach," *J. Chem. Inf. Model.*, vol. 55, no. 7, pp. 1483–1494, 2015.

[15] S. Jamal, S. Goyal, A. Shanker, and A. Grover, "Predicting neurological adverse drug reactions based on biological, chemical and phenotypic properties of drugs using machine learning models," *Sci. Rep.*, vol. 7, Apr. 2017, Art. no. 872.

[16] J. Scheiber *et al.*, "Mapping adverse drug reactions in chemical space," *J. Med. Chem.*, vol. 52, no. 9, pp. 3103–3107, 2009.

[17] Y. Yamanishi, M. Kotera, M. Kanehisa, and S. Goto, "Drug-target interaction prediction from chemical, genomic and pharmacological data in an integrated framework," *Bioinformatics*, vol. 26, no. 12, pp. i246–i254, 2010.

[18] A. F. Fliri, W. T. Loging, P. F. Thadeio, and R. A. Volkmann, "Analysis of drug-induced effect patterns to link structure and side effects of medicines," *Nature Chem. Biol.*, vol. 1, no. 7, pp. 389–397, 2005.

[19] J. Scheiber *et al.*, "Gaining insight into off-target mediated effects of drug candidates with a comprehensive systems chemical biology analysis," *J. Chem. Inf. Model.*, vol. 49, no. 2, pp. 308–317, 2009.

[20] F. Wang, P. Zhang, N. Cao, J. Hu, and R. Sorrentino, "Exploring the associations between drug side-effects and therapeutic indications," *J. Biomed. Inform.*, vol. 51, pp. 15–23, Oct. 2014.

[21] S. Mizutani, E. Pauwels, V. Stoven, S. Goto, and Y. Yamanishi, "Relating drug–protein interaction network with drug side effects," *Bioinformatics*, vol. 28, no. 18, pp. i522–i528, 2012.

[22] Y. Yamanishi, E. Pauwels, and M. Kotera, "Drug side-effect prediction based on the integration of chemical and biological spaces," *J. Chem. Inf. Model.*, vol. 52, no. 12, pp. 3284–3292, 2012.

[23] M. Liu *et al.*, "Determining molecular predictors of adverse drug reactions with causality analysis based on structure learning," *J. Amer. Med. Inform. Assoc.*, vol. 21, no. 2, pp. 245–251, 2014.

[24] F. Cheng *et al.*, "Adverse drug events: Database construction and in silico prediction," *J. Chem. Inf. Model.*, vol. 53, no. 4, pp. 744–752, 2013.

[25] W. Zhang, H. Zou, L. Luo, Q. Liu, W. Wu, and W. Xiao, "Predicting potential side effects of drugs by recommender methods and ensemble learning," *Neurocomputing*, vol. 173, pp. 979–987, Jan. 2016.

[26] Y.-G. Chen, Y.-Y. Wang, and X.-M. Zhao, "A survey on computational approaches to predicting adverse drug reactions," *Current Topics Med. Chem.*, vol. 16, no. 30, pp. 3629–3635, 2016.

[27] D. P. Williams and B. K. Park, "Idiosyncratic toxicity: The role of toxicophores and bioactivation," *Drug Discovery Today*, vol. 8, no. 22, pp. 1044–1050, 2003.

[28] C. Federer, M. Yoo, and A. C. Tan, "Big data mining and adverse event pattern analysis in clinical drug trials," *ASSAY Drug Develop. Technol.*, vol. 14, no. 10, pp. 557–566, 2016.

[29] K. Raja, M. Patrick, J. T. Elder, and L. C. Tsoi, "Machine learning workflow to enhance predictions of adverse drug reactions (ADRs) through drug-gene interactions: Application to drugs for cutaneous diseases," *Sci. Rep.*, vol. 7, Jun. 2017, Art. no. 3690.

[30] D. S. Wishart *et al.*, "DrugBank: A knowledgebase for drugs, drug actions and drug targets," *Nucl. Acids Res.*, vol. 36, pp. D901–D906, Nov. 2008.

[31] Y. Wang, J. Xiao, T. O. Suzek, J. Zhang, J. Wang, and S. H. Bryant, "PubChem: A public information system for analyzing bioactivities of small molecules," *Nucl. Acids Res.*, vol. 37, pp. W623–W633, Jul. 2009.

[32] R. Apweiler *et al.*, "UniProt: The universal protein knowledgebase," *Nucl. Acids Res.*, vol. 32, pp. D115–D119, Jan. 2004.

[33] O. Bodenreider, "The unified medical language system (UMLS): Integrating biomedical terminology," *Nucl. Acids Res.*, vol. 32, pp. D267–D270, Jan. 2004.

[34] M. Kuhn, M. Campillos, I. Letunic, L. J. Letunic, and P. Bork, "Side effect resource to capture phenotypic effects of drugs," *Mol. Syst. Biol.*, vol. 6, pp. 343–348, 2010.

[35] P. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*. Reading, MA, USA: Addison-Wesley, 2005.

[36] P. Branco, L. Torgo, and R. P. Ribeiro, "A survey of predictive modeling on imbalanced domains," *ACM Comput. Surv.*, vol. 49, no. 2, 2016, Art. no. 31.

[37] N. Japkowicz and S. Stephen, "The class imbalance problem: A systematic study," *Intell. Data Anal.*, vol. 6, no. 5, pp. 429–450, 2002.

[38] T. F. Smith, M. S. Waterman, and C. Burks, "The statistical distribution of nucleic acid similarities," *Nucl. Acids Res.*, vol. 13, no. 2, pp. 645–656, 1985.

[39] A. Frolkis *et al.*, "SMPDB: The small molecule pathway database," *Nucl. Acids Res.*, vol. 38, pp. D480–D487, Jan. 2010.

[40] A. Gottlieb, G. Y. Stein, E. Ruppin, and R. Sharan, "PREDICT: A method for inferring novel drug indications with application to personalized medicine," *Mol. Syst. Biol.*, vol. 7, 2011, Art. no. 496.

[41] J. Bowes *et al.*, "Reducing safety-related drug attrition: The use of in vitro pharmacological profiling," *Nature Rev. Drug Discovery*, vol. 11, no. 12, pp. 909–922, 2012.

[42] X. Wang, B. Thijssen, and H. Yu, "Target essentiality and centrality characterize drug side effects," *PLoS Comput. Biol.*, vol. 9, no. 7, p. e1003119, 2013.

[43] S. Whitebread, J. Hamon, D. Bojanic, and L. Urban, "Keynote review: In vitro safety pharmacology profiling: An essential tool for successful drug development," *Drug Discovery Today*, vol. 10, no. 21, pp. 1421–1433, 2005.

[44] L. Brouwers, M. Iskar, G. Zeller, V. van Noort, and P. Bork, "Network neighbors of drug targets contribute to drug side-effect similarity," *PLoS ONE*, vol. 6, no. 7, p. e22187, 2011.

[45] M. Campillos, M. Kuhn, A. C. Gavin, L. J. Jensen, and P. Bork, "Drug target identification using side-effect similarity," *Science*, vol. 321, pp. 263–266, Jul. 2008.

WEI-PO LEE received his Ph.D. degree in artificial intelligence from The University of Edinburgh, U.K. He is currently a Professor at the Department of Information Management, National Sun Yat-sen University, Kaohsiung, Taiwan. He is interested in systems biology, medical informatics, machine learning, and data mining.

JHIH-YUAN HUANG received his M.A. degree from the Department of Information Management, National Sun Yat-sen University, Kaohsiung, Taiwan, where he is currently pursuing his Ph.D. degree. His research interests include artificial intelligence, data mining, and medical informatics.

HSUAN-HAO CHANG received his M.A. degree from the Department of Information Management, National Sun Yat-sen University, Kaohsiung, Taiwan. He is interested in system modeling, simulation and analysis, data mining, and knowledge management.

KING-TEH LEE received his Ph.D. degree from the College of Medicine, Kaohsiung Medical University. He is currently a Professor at the Department of Healthcare Administration and Medical Informatics, Kaohsiung Medical University, Taiwan, and a M.D. at Kaohsiung Medical University, Chung-Ho Memorial Hospital. He was a Vice President of Kaohsiung Medical University, Chung-Ho Memorial Hospital from 2010 to 2014.

CHAO-TI LAI received her M.A. degree from the Department of Healthcare Administration and Medical Informatics, Kaohsiung Medical University, Taiwan. She is the Head Nurse at Kaohsiung Medical University, Chung-Ho Memorial Hospital. She is interested in medical informatics and data analytics.