Prediction of Adverse Drug Reactions using Gene Expression and Chemical Fingerprint Data

Euisoon Hwang Seyed Mehdi Jazayeri

Abstract

Adverse drug reactions (ADRs) have become one of the major public health concerns. In addition, it is one of the main causes of failure in the process of drug development, and of drug withdrawal from the market. It has become critical to successfully predict such adverse drug reactions beforehand in order to increase benefits of such drugs while avoiding life-threatening side effects. In this study, we extend our study of prediction of ADRs from previous assignments by data fusion of gene expression and chemical fingerprint data. After the selection of interested ADRs, genes, and drugs, we attempt to build prediction models for ADR prediction and achieve successful prediction with six combination of machine learning techniques. The acquired results indicate that using support vector machine is the most suitable classification method for ADR prediction. In addition, the data fusion/integration resulted in higher accuracy, implying that the information on drugs is valuable for ADR prediction. Moreover, integrating different aspects of drugs (biological, chemical, phenotypic properties, etc.) with correct approach would result in successful ADR prediction.

1. Introduction

Being the fourth largest cause of death in the United States [1], adverse drug reactions (ADRs) has been one of the major healthcare issues that are dealt throughout the world [2]. Around 2 million patients in the US experience ADR, and the severe ADRs resulted over 100,000 deaths. In addition to the public health concerns, ADR creates several negative impacts in the economy. Increased length of stay in hospital and drug withdrawal from the market due to side effects caused loss of millions of dollars [3]. Thus, in order to increase efficiency of medications and reduce such loss, accurate predictions on ADR is very critical.

In this paper, we focus on the biological aspect of drugs as an extension of the previous assignments, while we also bring the chemical aspect of drugs, which were not introduced before. In result, we attempt to predict ADR based on the processed integrated data.

2. Methods

2.1 Data description

There are three datasets that were used in this particular study. (i) Drug side effect data from FDA Adverse Event Report System (FAERS) using Propensity Score Matching (PSM). (ii) Gene expression signatures for drugs/small molecule compounds in the landmark gene space (iii) 166-bit MACCS chemical fingerprint matrix for drugs/small molecule compounds

2.1.1 FAERS ADR Data

FAERS ADR data is acquired from FAERS, a database containing information on adverse event and medication error reports that are submitted to Food and Drug Administration. It contains binary information on which drugs result in certain side effects. For this dataset, there are 684 drugs associated with 9405 ADRs, resulting in 6433020 drug-ADR associations.

2.1.2 Gene Expression Signatures (GE)

The Gene expression signature data is acquired from the Library of Integrated Network-based Cellular Signatures (LINCS) L1000 dataset. The representing data are measurement of changes in gene expressions pre and post-treatment of FDA-approved drugs.

2.1.3 MACCS Chemical Fingerprint (CF)

166-bit MACCS Chemical Fingerprint data represents chemical fingerprints that are associated with each drug. Chemical fingerprint breaks each compound down to give characteristics of each chemical such as ring structures, functional groups, and bond types. It provides information about drugs on how organic compounds would react in human body at a chemical level [4]. The main reason why this data was selected is that it introduces more information about the drugs themselves and may result in higher accuracy. In addition, with fingerprints, we can isolate and predict which specific segment of structures would result in ADR.

2.2 Data integration

With the datasets prepared as described above, we attempted to filter, reduce, and integrate the datasets in order to produce results in a given time. As GE contains probes instead of gene symbols that we are familiar with, we introduced a meta-data called meta_probes_info to map probes to its representing gene symbols. In terms of pert_ids, it is difficult to convert into common drug names since there are more than one names that can represent pert_ids. Thus, in order to avoid confusion in the data, we decided to use pert_ids as is. During the step of data preparation, we found 681 common drugs that are in all FAERS, GE, and CF. The number of drugs that are used for this study are more than 438 drugs that were used in the previous assignments, implying that it would bring more information to possibly produce a higher accuracy in ADR predictions. In CF, there were columns that are either all 0's or 1's. This fingerprint columns were neglected since it would not provide the differences among drugs. On the other hand, all the gene expressions with low difference in their maximum and minimum values were removed (ex: Max(X1) - Min(X1) < threshold). The intuition is we are interested in variables that fluctuate in different experiments and consequently affect the final outcome. Furthermore, we removed columns (variables) that had all 0's or 1's from MACCS dataset. Finally, a set of 946 features were selected.

2.3 Feature Selection

Feature selection is one of the most important part in any machine learning task. Simplification of models, shorter training times and enhanced generalization by reducing overfitting are reasons for feature selection. Since obtaining information and measuring several features in the domain of bioinformatics is costly, feature selection plays a crucial role. Below is explanation of used methods.

2.3.1 t-test Feature Selection

By definition, t-test is any statistical hypothesis test. It can be used to determine if two sets of data are significantly different from each other, and is most commonly applied when the test statistic would follow a normal distribution if the value of a scaling term in the test statistic were known. In our scenario, we separate each variable's value into two distinct set based on the corresponding ADR responses. One set contains values that caused a ADR (1's) and while the other contained binary information that did not cause the ADR (0's). Then we run t-test to check whether these two set are significantly different from each other. If the variable passes a t-test we concluded that the variable has a predictive power and can be selected as a feature. To decide on what threshold to use, 10 histograms of p-values of 10 different features were plotted. From histograms, a p-value equal to 0.05 seemed promising because higher values led to a large set of variables and lower values led to a small feature set.

2.3.3 Correlation Feature Selection:

The idea behind this approach is "a good feature subset is one that contains features highly correlated with (predictive of) the class, yet uncorrelated with (not predictive of) each other". Thus, we sought gene expressions (GE) and chemical structures (MACCS) that pass the above condition. To do so, we exploit a package called FSelector. The *linear.correlation* function in the package, takes class values as well as features in form of R formula. Then, it provides the importance rank of each feature. Again, using histogram of 10 different features we selected attribute_importance equal to 0.12 as threshold.

2.4 Classification

Prediction of adverse drug reaction can be defined as a binary classification. Three supervised learning methods, Support Vector Machine (SVM), Decision Tree and, K Nearest, Neighbor (KNN) were selected to classify ADRs. We examined combination of each method with feature selection methods. Below is a brief description of each method and the reason why they are selected.

2.4.1 Support Vector Machine

An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided clearly by a gap that is as wide as possible. Since we dealt with only two categories (0 and 1), this method has the potential of making a strong predictive model. In general, we hoped to find a margin that can divide our data space into two distinct regions so that, points on each side belongs to one class. SVM implementation of "e1071" package is used.

2.4.2 k-NN

In k-NN classification, the output is a class membership. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors. This method is selected to test if we can exploit similarity aspect in our dataset to classify new data points. We decided to check similarity among 10 neighbors. One disadvantage of this method was its time inefficient nature as it is a lazy learner method. We used more reliable implementation of this method presented in "impute" package.

2.4.3 Decision Tree

Decision tree is a method that uses inductive inference to approximate a target function, which will produce discrete values. This method is widely used when there is a doubt about the presence of noise in the dataset. Thus, it is selected as one of the options to address problems caused by noise data points. However, extra care must be taken to avoid overfitting as this method is prone to overfitting. We utilized implementation of "rpart" package. Furthermore, a control variable was defined with minsplit=2, minbucket=1, cp=0.0001 setting, to avoid building overfitted tree. This configuration is selected by observing the outcome of different configurations.

2.5 Cross-validation

After reviewing several publication in the literature, we realized small number of fold for cross-validation is used frequently by domain experts [4][5][6][7]. In addition, by decreasing the size of the fold in cross-validation, the required computational time is reduced, which eventually lead to obtaining the result within given time. Furthermore, the number of folds in cross-validation does not significantly affect the evaluation [5]. Thus, 5-fold cross-validation was used for this study. In 5-fold cross-validation, the final dataset is randomly partitioned into 5 equal sized subsets (each subset contains 137 except the last one). To do so, a sample of size equal to the number of drugs is selected at random (using sample() function in R) and divided by 5. A single subset is held as the test data, and the remaining subsets are used as training data. The cross-validation process is then repeated 5 times, with each of the 5 subsets used exactly once as the test and training data. The results from each run is averaged to produce a single evaluation and this process is repeated five times. Also, feature selection methods were called at each run.

2.6 Area under curve (AUC)

To create ROC plots we utilized "ROCR" package. Since probabilities of predictions were required to plot a ROC, we could not make plots for the decision tree and k-NN for which we only have final predicted values. By making a few changes in SVM, however, we could plot ROC for combination of SVM and t-test. Plots bellow are examples of ROC plots for predicting ADR that returned the lowest error rate using combination of SVM and t-test and GE, MACCS and combination of these two dataset. The intuition behind ROC curve is to show the global AUC for the best prediction in order to verify whether integrated dataset had a better accuracy in prediction than the separate dataset since our implementation did not deal with the class imbalance directly.

3. Results

Table 1 Average Error Rate of 10 ADRs

Avg Error Rate	knn_cor	knn_t.test	svm_cor	svm_t.test	dtree_cor	dtree_t.test
Cardiac.failure	0.34088	0.32834	0.33150	0.29948	0.42933	0.41594
Jaundice	0.29832	0.31418	0.29668	0.27901	0.40576	0.42161
Sleep.apnoea.syndrome	0.35564	0.36857	0.34203	0.31631	0.42759	0.43895
Muscle.spasms	0.37740	0.37260	0.36414	0.32768	0.44006	0.44549
Hernia.hiatal	0.37125	0.35903	0.34833	0.31753	0.42433	0.44531
Thrombocytopenia	0.44153	0.48386	0.44294	0.46169	0.49011	0.49690
Weight.decreased	0.48481	0.46901	0.47157	0.45469	0.47811	0.48131
Gait.disturbance	0.42902	0.42265	0.40675	0.38955	0.47233	0.47977
Deep.vein.thrombosis	0.37770	0.38321	0.37388	0.34048	0.46633	0.46117
Csf.protein	0.28753	0.27492	0.26224	0.25112	0.38680	0.38228

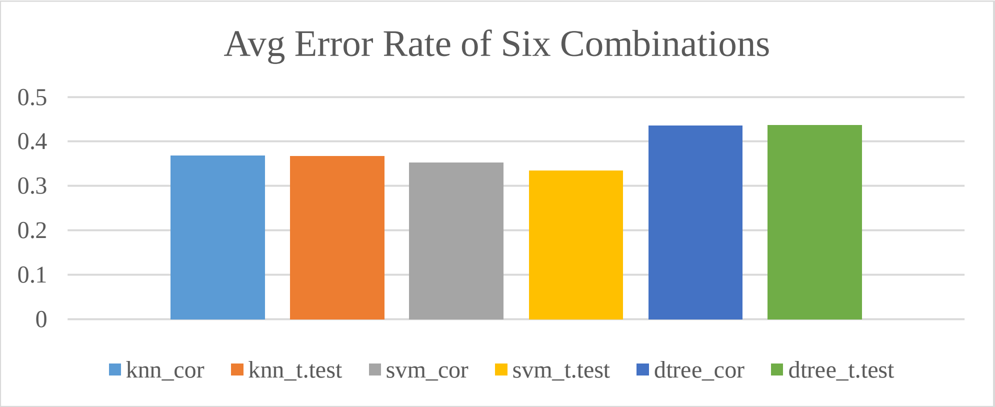

Figure 1 Average Error Rate for Six Combinations

Table 1 represents average error rate of 10 ADR predictions with 6 combinations that are used in this study. Based on the table, we can see that the prediction Csf.protein with SVM & t-test combination had the lowest error rate. Figure 1 shows the average error rate of all 263 ADR prediction across six combinations of feature selection and classification. SVM with t-test feature selection returned to have the highest accuracy out of all.

Table 2 Top 10 overall lowest error rate of ADRs in a given combination

	knn_cor	**knn_t.test**	**svm_cor**
1	Diabetic.neuropathy	International.normalised.ratio.increased	Dialysis
2	Blood.bilirubin.increased	Blood.albumin.decreased	Ventricular.tachycardia
3	Disseminated.intravascular.coagulation	Exostosis	International.normalised.ratio.increased
4	Blood.albumin.decreased	Toothache	Blood.albumin.decreased
5	Hepatic.function.abnormal	Csf.protein	Exostosis
6	International.normalised.ratio.increased	Ventricular.tachycardia	Csf.protein
7	Ear.pain	Ear.pain	Blood.bilirubin.increased
8	Dialysis	Agranulocytosis	Mental.disorder
9	Lung.cancer.metastatic	Skin.lesion	Dental.caries
10	Endodontic.procedure	Vertigo	Agranulocytosis
	svm_t.test	**dtree_cor**	**dtree_t.test**
1	Csf.protein	Blood.bilirubin.increased	Dental.caries
2	Blood.albumin.decreased	Diabetic.neuropathy	Fluid.overload
3	Exostosis	Haemodialysis	Ear.pain
4	Ear.pain	Ear.pain	Diabetic.neuropathy
5	International.normalised.ratio.increased	Urinary.retention	Musculoskeletal.stiffness
6	Agranulocytosis	Agitation	Debridement
7	Diabetic.neuropathy	Exostosis	Ventricular.tachycardia
8	Vertigo	Rales	Csf.protein
9	Mental.disorder	Orthostatic.hypotension	Speech.disorder
10	Dialysis	Oesophagitis	Blood.bilirubin.increased

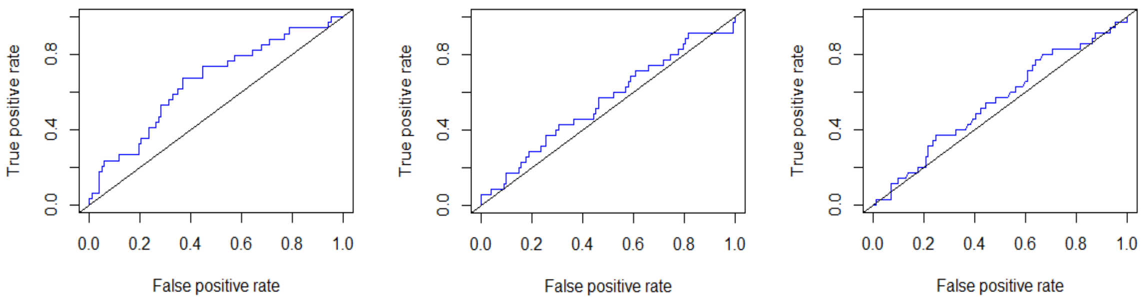

Figure 2 ROC Curve of Integrated Data (left), GE (middle), and CF (right)

Despite the fact that overall accuracy of this study was not plausible, we can see from the three plots (Figure 2) that integration of two datasets produced a better prediction than considering two datasets separately. According to this study, CSF protein side effect had the highest prediction accuracy (i.e. the lowest error rate) among all possible 263 ADRs that were used in the study. CSF protein side effect also appeared when CF was only considered. Meanwhile, exostosis had the highest prediction accuracy when considering only GE. The obtained data for each dataset (combined, GE, CF) for 6 combinations of feature selections and classifications are provided as separate text files.

4. Discussion and Conclusion

As introduced earlier, prediction of ADR is critical not only in healthcare, but also in economy. We extended our studies further from assignments in order to increase the accuracy of prediction with scalable computational approach and integration of chemical aspect that were not present in previous studies. By bringing the chemical structure to the approach, we could see the increase in overall accuracy in ADR prediction. It implies introducing the behavior of each chemical compound/molecule offers the insight of chemical reaction mechanism of each drug. The difference in chemical structures induces electrons to “attack” certain spots based on charge or attractions, which produces distinct resulting compound. In result, there are more significant features that enables machine learning method to predict various ADRs more accurately. Although the resulting ROC curve from this study did not have the “best look”, it could be said that there is a gain and potential in this study. Furthermore, addition of datasets related to drugs and genes could increase the prediction accuracy.

There were several shortcomings and challenges encountered during this study. One of the shortcomings was the structure of the script. The R script used in this study contains many functions instead of naïve implementation. There are advantages and disadvantages of using various functions. The main reason why we used functions for feature selection and classification is to include such methods into cross-validation method. Also, functions requires less line of code and input parameters, providing faster runtime. On the other hand, the disadvantage was that it is hard to extract intermediate information unless the parameters are specified. Thus, it was a challenging task to determine what is returned in each step despite the knowledge of the script. This could be overcome in the future by using different structure of code to store detailed information each step. Another shortcoming is that the ROC curve indicates that prediction model for this study does not provide sufficient accuracy of ADR prediction. Gathering information on suitable thresholds may relieve certain issue. Moreover, the computational time was one of the biggest challenge. The FAERS ADR data contains more than 9000 ADRs and GE contains approximately 20,000 drugs. Since we had to produce results in a given time while resolving unexpected errors and interpreting data, we unfortunately had to narrow the given data to a size that can be implemented and run in a given time. In spite of the fact that the code was written in order to reduce runtime as much as it could, each dataset had to be reduced to a reasonable size. By doing so, there would be significant information loss that produced infeasible outcome.

Though this study could be told that it is unfinished, we have discovered some remarkable attributes and potentials throughout this project. Expansion of data with new related data showed a significant increased trend in ADR predictions. Among all combinations in this study, SVM with t-test feature selection outperformed the others. While it is a good idea to compare with other type of classification methods, we could see that SVM is an optimal approach in terms of prediction of

ADRs. Based on the acquired results and interpretations, approaching to ADR prediction with correct data integration is very promising.

5. References

[1] Giacomini, K. M., Krauss, R. M., Roden, D. M., Eichelbaum, M., Hayden, M. R., & Nakamura, Y. (2007). When good drugs go bad. *Nature, 446*(7139), 975-977.

[2] Pirmohamed, M., Breckenridge, A. M., Kitteringham, N. R., & Park, B. K. (1998). Adverse drug reactions. *British Medical Journal*, *316*(7140), 1295.

[3] Lazarou, J., Pomeranz, B. H., & Corey, P. N. (1998). Incidence of adverse drug reactions in hospitalized patients: a meta-analysis of prospective studies. *Jama*, *279*(15), 1200-1205.

[4] Pauwels, E., Stoven, V., & Yamanishi, Y. (2011). Predicting drug side-effect profiles: a chemical fragment-based approach. *BMC Bioinformatics*, *12*(1), 169.

[5] Wang, Z., Clark, N. R., & Ma'ayan, A. (2016). Drug-induced adverse events prediction with the LINCS L1000 data. *Bioinformatics*, btw168.

[6] Liu, M., Cai, R., Hu, Y., Matheny, M. E., Sun, J., Hu, J., & Xu, H. (2014). Determining molecular predictors of adverse drug reactions with causality analysis based on structure learning. *Journal of the American Medical Informatics Association*, *21*(2), 245-251.

[7] Liu, M., Wu, Y., Chen, Y., Sun, J., Zhao, Z., Chen, X. W., ... & Xu, H. (2012). Large-scale prediction of adverse drug reactions using chemical, biological, and phenotypic properties of drugs. *Journal of the American Medical Informatics Association*, *19*(e1), e28-e35.