ISSN No:-2456-2165

Survey on Pharmacovigilance System for Predicting Drug Indications and Side Effects

D. Mohanapriya
Assistant professor
Department of Computer Science
PSG College of Arts & Science
Research Scholar of Kongunadu
Arts and Science College
Coimbatore, Tamil Nadu, India

Dr.R. Beena
Associate Professor
Department of Computer Science
Kongunadu Arts and Science College
Coimbatore, Tamil Nadu, India

Abstract:- Drugs are chemicals that treat, prevent or diagnosis diseases which is also known as medicine or medication. The task of discovering a new drug for a specific disease requires high cost and long time, instead the prediction of remedial and side effects of available drugs preciously supports to recommend candidate drugs for specific diseases with low cost and time. The recommended drugs for diseases are predicted by analyzing relationship between the drugs-side effects, drugs-genes and drugs-diseases. The extraction of terms which are related to drug, genes, disease and side effects from textual documents, literatures and biomedical repositories are used to mine relationship among them. Many text mining based drug recommendation approaches were proposed in the literature. The drug indication and side effects were mined mostly based on topic modeling, machine learning, Feature dependency graph and Similarity based approaches. This article presents a detailed survey of predicting drug indications and side effects using text mining approaches. At first, different approaches are analyzed in depth, established from previous research. Furthermore, a comparative study is performed to identify the limitations of current methods and provide a suggestion for further progress in the estimation of signs of drugs and side effect

Keywords:- *Drug discovery, Drug indication, Side effects, Natural language processing, Text mining.*

I. INTRODUCTION

A drug is any substance that alters the function of the body, physically and/or psychologically, when taken into the body. A Drug is discovered to cure the disease. Drug discovery is the method of identifying potential new medicines. This covers a wide range of scientific subjects, including biology, chemistry, and pharmacology. Most of the drug discovery is failed, because of the drug development cost and project failure. Almost all drugs have impact, and unexpected symptoms can harm people and lead to severe effects. So, it is more necessary to find the side effects for reducing the sever effects. The severe effects will be reduced by drug repositioning. The process of finding a new indication from the existing drug is termed as drug repositioning. The importance of drug repositioning has grown dramatically due to the massive increase in the cost of new drug production [1]. So it reduces the drug development cost and time. Various methods have been proposed for drug repositioning based on computational methods was proposed due to the exponential increase in available genomic / phenotypic data and the appearance of various methods for data analysis, such as machine learning, text mining [2]. Text mining is a technique of data mining and is used to extract meaningful knowledge from huge unstructured text data. Text mining is more successful as it leverages secondary data resources to help meet the goal of detecting side effects as soon as proper drug advice contributes to enhanced drug safety is accurate. Various techniques were implemented using this method. By using this method, various techniques were performed. In this study, various techniques are investigated for predicting drug indication and its side effects.

II. VARIOUS TECHNIQUES FOR DRUG INDICATION AND SIDE EFFECT PREDICTIONS

Finally, complete content and organizational editing before formatting. Please take note of the following items when proofreading spelling and grammar: Gottlieb et al. [3] proposed a method to predict the relationship between drug and disease. According to the relationship between disease-disease and drug-drug similarity measures were constructed and then classification rules were constructed using these similarity measures to study the drug-disease relationship. A relationship between new drug and disease was predicted from the classification rules.

Pauwels et al. [4] proposed a new method for predicting the potential side-effects of drug candidate molecules on the basis of their chemical structures. The main goal of this method was extracting the correlated sets of chemical substructures and side-effects. For this purpose, five different methods were proposed such as random assignment, nearest neighbor, Support Vector Machine (SVM), ordinary canonical correlation analysis and sparse canonical correlation analysis.

Lee et al. [5] proposed a novel process-drug-side effect network for finding the association between biological processes and side effects. In this method, three phases for systematically searching associations between drugs and biological processes i.e., enrichment scores computations, t-score computation and threshold-based filtering. Then, the side effect-related biological processes were discovered by combining the drug-biological process network and the drug-side effect network.

Chen et al. [6] proposed a computational method for predicting drug side effects based on chemical-chemical interactions and protein-chemical interactions. The chemical-chemical and protein-chemical interaction data was obtained from the STITCH data. A new scheme was introduced to measure the likelihood of interaction of two chemicals and between the proteins and drug. A training dataset was described based on the interaction between drug and side effects. The training dataset and likelihood measure was used to predict drug side effects.

Xu et al. [7] presented an automatic learning approach for automatic construction of a large-scale and accurate drug-side-effects association knowledge. Initially, parse trees and relevant sentences were determined from the knowledge of Food and Drug Administration (FDA) drug labels. Then, the syntactic patterns related to the drug-side-effect pairs were extracted from the resulting set of parse trees. The specific patterns were ordered based on the pattern-ranking algorithm and a set of patterns with high recall and precision were chosen to extract the drug-side-effects pairs from the text corpus.

Bravo et al. [8] proposed a BeFree system for the identification of relationships between drugs, diseases and genes. This system was the combination of the shallow linguistic kernel and dependency kernel. Shallow syntactic information was used in the shallow linguistic kernel for the extraction of adverse drug reactions from drug-drug interaction and clinical reports. The dependency kernel exploited the syntactic information of the sentence using the walk-weighted subsequence kernels. Based on the information obtained from shallow linguistic kernel and dependency kernel, the relationships between drugs, diseases and genes were predicted.

Zhang et al. [9] proposed an algorithm to predict the side effects. Initially, mutual knowledge between the feature dimension and undesirable effects was explored to choose the feature dimensionality. Then, the Generic Algorithm (GA) and the Multi label K-Nearest Neighbor Method (MLKNN) were processed to choose the optimal features and to predict the drug side effects.

Zhang et al. [10] proposed an integrated method for prediction of side effects and its association with the disease. The integrated method was used known side effects to predict the possible or missing side effects of approved drugs. Integrated Neighborhood- Based Method (INBM) was used to predict the side effects, by using similarity between the drugs. Restricted Boltzmann Machine-Based Method (RBMBM) learned the distribution of probabilities governing associations with side effects.

Kim et al [11] proposed for predicting unintended effects of the drug by examining the off-target tissue effects. For identifying the off target, matrix was constructed using Anatomical Therapeutic Chemical code (ACT) & target proteins of the drug. By using the matrix, relations between tissue proteins and symptoms were predicted.

Zhang & Chen [12] Proposed machine learning methods to predict the side effect problem. The drug-drug similarity was extracted by constructing linear neighborhood in the feature space. Then, similarity based graph was constructed to predict the side effects of new drugs based on the known undesirable effect information. The graph based method was further improved which processed heterogeneous data.

Jang et al. [13] proposed a computational framework to predict drug unintended effects. In this framework, statistical analysis was processed to find the disease signature vectors and literature mining was applied to determine clinical drug effect vectors. At last, each disease drug pair was allocated the repositioning possibility rate by calculating the balancing and relationship between the clinical drug effect and the clinical status of signatures of disease.

Ibrahim et al. [14] proposed a hybrid algorithm such as an optimized and modified association rule mining technique for extracting the important association patterns of drug interaction-adverse event. In this algorithm, two different triage features such as three-element taxonomy and three performance metrics were used for assessing the resulting drug interaction-adverse event. Also, logistic regression method was applied for quantifying the magnitude and direction of interactions.

Lee et al. [15] proposed a hybrid machine learning method for predicting the drug side effects based on the suitable data set feature. In this approach, data analytic techniques were used for analyzing the impact of drug distribution in the feature space, and classifying side effects into many types according to the distribution of classes. After that, domain-dependent strategies for each type were adopted for constructing the data models.

Jang et al. [16] identified a drug-based gene regulation and a biomedical-literature based phenotype. To extract sentence from a gene and a drug, co-occur and sentence where a gene and a phenotype co-occur from abstract, then construct the extracted sentence's dependence graph. Identifying gene-drug and gene-phenotype relationships in the dependency graphs. The relationship is defined by the activation or inhibition of gene regulation. By using the relationships, calculate the effect of a drug on a gene and the effect of a phenotype on a gene.

Dimitri & Lio [17] proposed a machine learning algorithm to predict the drugs side effects. Initially, the cluster methods such as K-mean, PAM and K-seeds were applied to cluster the drugs with respect to the feature profiles. Bayesian method has been used for each cluster to measure the matrix of probability score in which each dimension contains the score to indicate the probability of particular side effects for a drug belonging to the same cluster.

Abdelaziz et al. [18] presented a large-scale similarity-based framework for prediction of drug-drug interaction. Initially, a drug-related data and its knowledge were semantically integrated that returned a knowledge graph. It defined the drug attributes and its association with other associated objects such as pathways, enzymes and chemical structures. The various similarity measures between all the drugs were calculated in a distributed and scalable framework with the help of knowledge graph. The resulting similarity metrics were utilized to develop features for a large-scale logistic regression model for prediction of drug-drug interaction.

Zhang et al. [19] proposed a novel measure of drug-drug similarity, namely linear neighborhood similarity computed in the drug feature space by exploring the linear neighborhood association. After that, the similarity was transferred from the feature space to the side effect space and predicting the drug side effects by propagating known side effect information via a similarity-based graph. Also, this method was extended using similarity matrix integration and missing side effects for predicting the side effects of new drugs and unobserved side effects of approved drugs.

Ma et al. [20] proposed a novel and effective unsupervised model, namely Sifter for automatically finding the drug side effects. In this model, the estimation on drug side effects was enhanced by learning and measuring platform-level and user-level quality simultaneously. Zheng et al. [21] proposed an optimized drug similarity framework for increasing the efficiency of side effect prediction. Initially, four different drug similarities were integrated into the comprehensive similarity and optimized by clustering. Then, the optimized similarity was enhanced by the indirect drug similarity for predicting the side effects.

Zhang et al. [22] proposed computational method for predicting the unintended signs in drugs according to the features of determined available drug and relationship between drug and unintended signs. Computational method developed in low-dimensional space, which extracted features of unintended signs and drugs. This method was differs from the conventional matrix factorization approach, and can determined the biomedical context into account. The matrix factorization was very powerful technique, and can find unobserved associations based on the known association-based matrix.

Zhao et al. [23] proposed a binary classification model for predicting the drug side effects using heterogeneous information of drugs. Similarity based method was applied to encode the each drug-side effect pair and random forest was adapted to predicting the drugs side effects. If the prediction outcome was positive, then the considered drug has side effects. Otherwise, the drug does not have any side effects.

Zhang et al. [24] proposed a Network Topological Similarity-based Classification method (NTSIM-C) for prediction of drug-disease association. It differentiated the therapeutic associations from others. In NTSIM-C, a drug-disease association was described as a feature vector. It consisted of similarity scores between drugs and other drugs and the vector corresponded to a row vector in drug-drug similarity matrix. A linear neighborhood similarity matrix was developed for drugs and the linear neighborhood similarity matrix was calculated for a disease. For the association between the drug and disease, a vector was created which merged the similarity vector of drug and similarity vector of disease. Based on the vector, the association between drug and disease was predicted.

Jang et al. [25] proposed an algorithm to predict drug-phenotype and drug-side effect relationship. A Natural Language Processing (NLP) technique was used in this algorithm to extract and identified the words which denote the drug and gene relationship. According to the property of the words, identify the up-regulating and down-regulating effect on the gene. Then, a probability matrix was constructed based on a topic model where probability of each drug in topic was computed statistically. It was normalized and ordered for each drug to build a classifier for prediction.

Methods Used	Merits	Demerits
Similarity measures & classification feature [3]	High degree of accuracy and sensitivity.	Biological interpretation difficult, risk of over training the system does not provide mechanistic details.
Random assignment, nearest neighbor, SVM, ordinary canonical correlation analysis and sparse canonical correlation analysis [4]	Computationally efficient.	Only applicable on large datasets.
Novel process-drug-side effect network [5]	Efficient and useful to find the association between biological processes and side effects.	A loss of information was caused to obtain more accurate association.
Computational method [6]	Effective in identifying drugs side effects.	Face the problem of coverage scope because of using a benchmark dataset.
Automatic learning approach [7]	High precision.	It requires manual examination of many more patterns.
BeFree [8]	Highly efficient to extraction relationship among drug, disease and gene.	The large scale analysis of gene-disease association based on BeFree leads to some issues regarding data prioritization and curation.
GA, MLKNN [9]	The computational complexity of multi-label learning was less.	Ensemble learning model does not integrated individual feature-based FS-MLKNN models.
RBM [10]	High in performances.	low AUC value.
Off-target tissue effect, Side-effect prediction [11]	The side effects can be reduced by predicting off target tissue effect.	High computational complexity for construction of matrix.
Side effects, Data Integration [12]	High accuracy performances.	This method does not support prediction of side effect caused by variety of drug.
Drug repositioning, Clinical drug effects [13]	Have a high probability of biological validity.	Due to the small sample size, different conditions of the disease are not treated.
Optimized and modified association rule mining technique [14]	Better sensitivity and specificity.	It has high percentage of missing information in the attribute of the drug in the dataset.
Hybrid Machine Learning Approach [15]	Better predictive performance.	Difficult to predict side effects.
Biomedical literature [16]	High specificity.	Does not consider external factors including gene expression etc...
Bayesian approach [17]	Simple to use tool to predict undesirable effects.	Existing undesirable effects, are based on this assumption.
Large-scale similarity-based framework [18]	Combination of locally and globally generated drugs similarity features improves the performance of large-scale similarity-based framework.	Less F-score.
Linear neighborhood similarity graph method [19]	Better accuracy for predicting the side effects of new drugs.	Computational complexity was high.
Sifter model [20]	High recall.	High computational complexity.
Optimized drug similarity framework [21]	It can able to find the drug features from different perspectives.	Less F1-score.
FGRMF [22]	Result is more satisfied.	Does not predict the SIDER database.
Binary Classification Model [23]	Good performances.	Not possible to detect the side effect in early stage.
NTSIM-C [24]	More promising for predicting drug-disease associations and their therapeutic function.	Not applicable for large datasets.
NLP [25]	NLP is more accurate and robust.	Does not support the computational machine learning approach.

Table 1:- illustrates an overview of merits and demerits of above discussed drug indications and side effect prediction techniques.

III. CONCLUSION

In this article, a detailed view of predicting drug indication and its side effect was presented with their merits and demerits. Based on this analysis, it is observed that the algorithm can predict drug indication and side effect efficiently than other methods. Algorithm can predict not only associations of drug phenotypes, but also new associations of drug-side effects. The major advantage of the algorithm [25] is that it can predict therapeutic and negative effects using a single algorithm. However, algorithm does not consider the emergence of a phenotype, so long as there are known associations of drug phenotype, candidate drugs may be derived.

REFERENCES

[1]. De Oliveira, E. A. M., & Lang, K. L. (2018). Drug repositioning: concept, classification, methodology, and importance in rare/orphans and neglected diseases. Journal of Applied Pharmaceutical Science, 8(08), 157-165.
[2]. Edwards, I. R., & Aronson, J. K. (2000). Adverse drug reactions: definitions, diagnosis, and management. The lancet, 356(9237), 1255-1259.
[3]. Gottlieb, A., Stein, G. Y., Ruppin, E., & Sharan, R. (2011). PREDICT: a method for inferring novel drug indications with application to personalized medicine. Molecular systems biology, 7(1)
[4]. Pauwels, E., Stoven, V., & Yamanishi, Y. (2011). Predicting drug side-effect profiles: a chemical fragment-based approach. BMC bioinformatics, 12(1), 169.
[5]. Lee, S., Lee, K. H., Song, M., & Lee, D. (2011). Building the process-drug–side effect network to discover the relationship between biological Processes and side effects. In BMC bioinformatics (Vol. 12, No. 2, p. S2). BioMed Central.
[6]. Chen, L., Huang, T., Zhang, J., Zheng, M. Y., Feng, K. Y., Cai, Y. D., & Chou, K. C. (2013). Predicting drugs side effects based on chemical-chemical interactions and protein-chemical interactions. BioMed research international, 2013.
[7]. Xu, R., & Wang, Q. (2014). Automatic construction of a large-scale and accurate drug-side-effect association knowledge base from biomedical literature. Journal of biomedical informatics, 51, 191-199.
[8]. Bravo, À., Piñero, J., Queralt-Rosinach, N., Rautschka, M., & Furlong, L. I. (2015). Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research. BMC bioinformatics, 16(1), 55.
[9]. Zhang, W., Liu, F., Luo, L., & Zhang, J. (2015). Predicting drug side effects by multi-label learning and ensemble learning. BMC bioinformatics, 16(1), 365)
[10]. Zhang, W., Zou, H., Luo, L., Liu, Q., Wu, W., & Xiao, W. (2016). Predicting potential side effects of drugs by recommender methods and ensemble learning. Neurocomputing, 173, 979-987.
[11]. Kim, D., Lee, J., Lee, S., Park, J., & Lee, D. (2016). Predicting unintended effects of drugs based on off-target tissue effects. Biochemical and biophysical research communications, 469(3), 399-404
[12]. Zhang, W., Chen, Y., Tu, S., Liu, F., & Qu, Q. (2016, December). Drug side effect prediction through linear neighborhoods and multiple data source integration. In 2016 IEEE international conference on bioinformatics and biomedicine (BIBM) (pp. 427-434). IEEE)
[13]. Jang, D., Lee, S., Lee, J., Kim, K., & Lee, D. (2016). Inferring new drug indications using the complementarity between clinical disease signatures and drug effects. Journal of biomedical informatics, 59, 248-257
[14]. Ibrahim, H., Saad, A., Abdo, A., & Eldin, A. S. (2016). Mining association patterns of drug-interactions using post marketing FDA's spontaneous reporting data. Journal of biomedical informatics, 60, 294-308.
[15]. Lee, W. P., Huang, J. Y., Chang, H. H., Lee, K. T., & Lai, C. T. (2017). Predicting drug side effects using data analytics and the integration of multiple data sources. IEEE Access, 5, 20449-20462
[16]. Jang, G., Lee, T., Lee, B. M., & Yoon, Y. (2017). Literature-based prediction of novel drug indications considering relationships between entities. Molecular BioSystems, 13(7), 1399-1405
[17]. Dimitri, G. M., & Lió, P. (2017). DrugClust: a machine learning approach for drugs side effects prediction. Computational biology and chemistry, 68, 204-210
[18]. Abdelaziz, I., Fokoue, A., Hassanzadeh, O., Zhang, P., & Sadoghi, M. (2017). Large-scale structural and textual similarity-based mining of knowledge graph to predict drug–drug interactions. Journal of Web Semantics, 44, 104-117.
[19]. Zhang, W., Yue, X., Liu, F., Chen, Y., Tu, S., & Zhang, X. (2017). A unified frame of predicting side effects of drugs by using linear neighborhood similarity. BMC systems biology, 11(6), 101.
[20]. Ma, F., Meng, C., Xiao, H., Li, Q., Gao, J., Su, L., & Zhang, A. (2017). Unsupervised discovery of drug side-effects from heterogeneous data sources. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 967-976).
[21]. Zheng, Y., Ghosh, S., & Li, J. (2017). An optimized drug similarity framework for side-effect prediction. In 2017 Computing in Cardiology (CinC) (pp. 1-4). IEEE.

[22]. Zhang, W., Liu, X., Chen, Y., Wu, W., Wang, W., & Li, X. (2018). Feature-derived graph regularized matrix factorization for predicting drug side effects. Neurocomputing, 287, 154-162
[23]. Zhao, X., Chen, L., & Lu, J. (2018). A similarity-based method for prediction of drug side effects with heterogeneous information. Mathematical biosciences, 306, 136-144
[24]. Zhang, W., Yue, X., Huang, F., Liu, R., Chen, Y., & Ruan, C. (2018). Predicting drug-disease associations and their therapeutic function based on the drug-disease association bipartite network. Methods.
[25]. Jang, G., Lee, T., Hwang, S., Park, C., Ahn, J., Seo, S., & Yoon, Y. (2018). PISTON: Predicting drug indications and side effects using topic modeling and natural language processing. Journal of biomedical informatics, 87, 96-107