Personalized Drug-Disease prediction using Multiple Linear Regression with ReLU

S.Vijaya
Assistant Professor, KG College of Arts and Science, Coimbatore, TamilNadu, India 641035
s.vijaya@kgcas.com

Abstract. Predicting models for personalized Drugs related to specific disease are essential, as traditional methods are expensive and time consuming. The most challenging task in personalized medicine is predicting the status of disease from high dimensionality data. In the biomedical domain the association between drugs and disease plays a vital role as the same drug may treat similar diseases. For the good adaptability to complex and nonlinear behaviour data, Multiple Linear Regression method with ReLU Activation function is used for calculation and to fit the model with Drug –Disease dataset. Based on the results the drug or combination of drugs that treat a specific disease is predicted efficiently.

1. Introduction

The discovery of drug is very tedious, expensive and very lengthy process. Even it can take more than an era and hundreds of millions to introduce a new drug and the probability of getting approval for the new drug is very less. Drug repository method becomes popular method to overcome the challenges faced in Traditional methods of drug discovery. Using Drug repository new indications can be discovered for old drugs. Recently, Machine Learning methods are majorly used to identify the interactions of drug-disease potentially. Review sites and social media forum has a huge volume of information about the user preferences and ratings of multiple products based on their experiences. Valuable insights can be obtained by applying approaches like Sentiment Analysis. The review of drugs can be collected with the features like Effectiveness, Side Effects, Benefits, over all Comments. Based on the information Sentiment Analysis can be done, to predict effective drug for any disease. Most of the Sentiment Analysis approaches used lexicons for identifying positive, negative and neutral words. The opinions collected from social media forum would be given in highly informal, ambiguous medical terms. Hence, such information have to be trained using Machine Learning approaches.

Recommendation of drug plays a vital role as millions of individuals are affected due to the wrong prescription of medicines, every year. Combinations of drugs, various drugs used for similar type of diseases may give side effects[10][11]. This problem can be rectified, if personalized drug is recommended based on the revises of people who had used and gave positive rating for the drug. In this work, annotated dataset has been taken for predicting personalized drug for similarity diseases. The dataset has been retrieved from UCI Machine Learning Repository https://archive.ics.uci.edu/ml/datasets.php.
2. Related Work

With the rapid growth of advance in biomedical domain, a huge volume of data stored in several databases [1]. Jianlin. Network-based method is popularly used in the discovery of drug repositioning. Drug and disease Data set are available in public open access databases. A new model CMRF has been proposed by Jianlin Wang et al. [1] to find the benefits of old drugs, using computational drug repositioning method. To circulate association scores WKNKN method was used in the pre-processing step. LPRIA, NMFRIA and NCPRIA methods were adopted in their work to obtain predictive association possibilities. Their model can be utilized to find the single drug-drug and disease-disease similarity.

Hui Cui et al. [2] used filtering algorithm to find the combination of drugs, the disease associated with the drugs. They had identified drug-disease relationship using Micro array data 67 using RNA seq data. To know the relationship between drug and disease Wen Zhang et al. [3] proposed the similarity constrained matrix factorization methods. For this they have considered biological context and tested the performance with 5-fold cross validation. Similar type of drugs may treat disease of same kind. The similarity measures between diseases and drugs can be measured using computational methods. Guangshang Wu et al. [4] proposed Ensemble Meta Paths and Singular Value Decomposition (EMP-SVD) to find the association between drug and disease using negative samples instead of having similarity information. As a first step they generated Community matrix using EMP-SVD method and each matrix was factorized using SVD method. From the case studies they have concluded that their proposed method EMP-SVD can be used to extract more drug-disease associations.

Personalized medicines prediction is the major challenging task and getting high performance also difficult. The advancement of Deep Learning algorithms Qianfan Wu et al., [5] performed a literature review on four articles related to personalized medicine. Auto-encoders method is used as common in those four articles to differentiate high-performance data from low dimensional data. The status of disease prediction done by applying Machine Learning Algorithm. Computational methods has been attracted the growing attention in prediction of drug-disease associations and the computational methods reduce the cost spent for library screening. Feng Huang et al., [6] proposed Collective Matrix Factorization-based multi-task learning method (CMFMTL) to predict two different types of association such as Therapeutic and non-Therapeutic based on bipartite network. Their case studies concluded that their method helped to predict the new drug-disease associations that were not added in the Comparative Toxicogenomics Database (CTD).

The Computational methods to predict drug-disease associations with high-throughput are time consuming and cost effective. Han Jing Jiang et al., [7] presented Sparse Auto-Encoder and Rotation Forest (SAEROF) method to predict drug-disease associations. Using Rotation Forest classifier with Sparse Auto-encoder, the drug similarity based on structure similarity and similarity of diseases based on semantic were predicted. They compared the performance of SAEROF method with SVM method. Three types of diseases and 20 drugs were used in the experimental study and concluded that their method SAEROF predicted the drug-disease associations effectively. Harleen Kaur et al., [9] analyzed Twitter data related to COVID-19, virus details, death rates, tested positive cases, recovered cases, etc., based on hashtag keywords. To perform Sentiment Classification, Hybrid Heterogeneous Support Vector Machine (H-SVM) method had been used. The performance calculated based on Precision, Recall, F1 Score and Accuracy. When compared their H-SVM method with RNN based these performance measures H-SVM obtained better results.

For the prediction of Kwiatkowski-Phillips-Schmidt Shin (KPSS) score in education, Gulden Kay Uyanik [12] used Multiple Linear Regression Analysis and to analyse the assumptions the features such as Measurements used for Evaluation, the techniques used for instructions, Counselling results
and Educational Psychology were estimated with respective to KPSS. The decision making for Public Health based on estimating the Invasive Meningococcal Disease (IMD) on epidemiologic data is a critical and tremendous task as these data estimate ambiguous population. To create simple patterns of infection dynamics within Ukraine, Hennadii Mokhort[13] had used a logic model Multiple Linear Regression and found the informativeness of this model was sufficient to identify the main factors that form the incidence of Invasive Meningococcal Disease(IMD) in Ukraine. Buza.K,Peska.L and Koller.J(2020)[14] used Asymmetric Loss Model(ALM) with conventional Regression methods to predict Drug_Target Interactions(DTI) and stated that their approach outperformed the other DTI methods and also they had recommended their system for the Drug Recommendations system task.

To predict the active cases of corona virus disease, Smita Rath,Alakananda T,Alok Ranjan T[15] developed a Multi Linear Regression(MLR) model with a global data consists of total confirmed cases, active cases, deceased and positive cases. The same dataset had been tested with Linear Regression and Multiple Linear Regression (MLR). The MLR showed strong prediction results with R2 value between 0.99 and 1.0.

3. Methodology

3.1. Multiple Linear Regression
The dataset consist of multiple features to predict personalized medicine. Hence, Multiple Linear Regression method is used. Each feature is considered as a dimension and the formula for Multiple Linear Regression is poised of multiple x values. Each x value is weighted by beta value. From the DrugLib dataset we have considered only four features to predict the output using equation 1. This formula yields a hyper plane rather a plane and its plane individuates the response feature values for every combination in the multidimensional features.

\[ Y = x_1 \beta_1 + x_2 \beta_2 + x_3 \beta_3 + x_4 \beta_4 + \beta \]  

Where \( x_1 \) -> DrugID , \( x_2 \) ->Rating , \( x_3 \) -> Effectiveness , \( x_4 \) -> SideEffects , \( y \) -> Whether the drug can be referred for specific disease of not.

3.2. Rectified Linear Unit (ReLU)
The model has been built and trained with Multiple Linear Regression and tested. Due to the vanishing gradient problem the result was not up to the expected level. The activation function is needed in order to train neural network methods as these model use speculative gradient descent with back propagation of errors. To improve and achieve better results different activation functions such as Sigmoid, tanh and Rectified Linear Unit(ReLU) had been tried with Multiple Linear Regression. When compared with the results ReLU is computationally performed better results. Rectified Linear Unit(ReLU) is an activation function used to transform the sum of weighted input features from that particular node into the output node. The purpose of using ReLU is, it will output the given input directly if the input has positive value otherwise the output will be zero.

\[ F(x) = \begin{cases} 
0 & \text{for } x < 0 \\
x & \text{for } x \geq 0 
\end{cases} \]  

Due to its better performance, many of the Neural Networks use this ReLU function. The block diagram for the workflow is given in Figure 1.
4. Dataset

Drug Review Dataset is downloaded from UCI Machine Learning Repository. Patient reviews for specific disease with associated conditions were provided in the data set. The ratings and patients reviews are combined into Benefits, Side Effects and Overall Comments. This Data set is multivariate, the attributes are integer data type, consists of 3107 rows and 5 attributes. The attributes considered for this work are, DrugID, Rating, Effectiveness, SideEffects, Result. The experiences of 3107 peoples own experience of a specific drug they took for diseases like Cancer, panic attacks and depression, asthma and chronic pain. The dataset had been annotated based on their reviews and performing Sentiment Analysis.

5. Experimental Results

The DrugLib dataset consists of 3107 rows, 8 features in the training set and 1036 rows, 8 features in the testing dataset. From this only 5 features namely DrugID, DrugName, Rating, Effectiveness and SideEffects were selected for prediction of the personalized drug. The rest of the features Condition, BenefitsReview, SideEffectsReview were not considered as these features values were given as sentences. These kind of values were not be transformed into numerical forms. Hence these features were removed from the dataset manually before modeling. First the categorical data were transformed in to numerical data to perform Multiple Linear Regression. Figure 1 shows the ratings given by users for various drugs based on their experience. These rating given for different DrugIDs are shown in Figure 2.
Figure 2 Drug names and the rating given by users.

Figure 3 Features selected for personalized Drug

Figure 4 Pre-processed Data (Categorical Data converted into Numerical Data)
Table 1  Sample Dataset before preprocessing

| DrugID | urlDrugName | Rating | Effectiveness       | Side Effects       |
|--------|-------------|--------|---------------------|--------------------|
| 1366   | Biaxin      | 9      | Considerably Effective | Mild Side Effects |
| 3724   | Lamictal    | 9      | Highly Effective    | Mild Side Effects  |
| 3824   | Depakene    | 4      | Moderately Effective| Severe Side Effects|
| 969    | Sarafem     | 10     | Highly Effective    | No Side Effects    |
| 696    | Accutane    | 10     | Highly Effective    | Mild Side Effects  |
| 1380   | Biaxin      | 2      | Marginally Effective| No Side Effects    |
| 45     | carbamazepine | 8 | Considerably Effective | Moderate Side Effects|
| 1939   | ultram-er  | 10     | Highly Effective    | Mild Side Effects  |
| 2576   | Klonopin    | 10     | Highly Effective    | No Side Effects    |
| 1093   | Effexor     | 1      | Marginally Effective| Extremely Severe Side Effects |
| 1962   | Lyrica      | 3      | Moderately Effective| Moderate Side Effects|
| 142    | Valtrex     | 9      | Highly Effective    | Mild Side Effects  |
| 1269   | Provigil    | 6      | Moderately Effective| No Side Effects    |
| 3499   | Maxalt      | 10     | Highly Effective    | Mild Side Effects  |
| 1598   | Xanax       | 5      | Moderately Effective| Mild Side Effects  |
| 4095   | Chantix     | 5      | Moderately Effective| Moderate Side Effects|
| 3707   | Ranitidine  | 7      | Moderately Effective| No Side Effects    |
| 3168   | Prednisone  | 7      | Considerably Effective | Moderate Side Effects|
| 1866   | Augmentin   | 8      | Highly Effective    | Severe Side Effects|
| 2150   | Zyrtec      | 10     | Highly Effective    | No Side Effects    |

The model has been trained and tested using Python. In f-test initial screening is done to find at least one variable as a significant predictor. Once a variable is found as response variable for prediction t-test is performed. For the accurate prediction ReLU activation function is used. ReLU activation function takes t-test and p-test values and give positive result when the values are greater than 0. The result of ReLU function based on f-test and p-test values is shown in Figure 5.

![Figure 5](image-url)  The result of ReLU function
As an initial step the model was developed with Multiple Linear Regression (MLR) in Python. To improve and achieve better results different activation functions such as Sigmoid, TanH, Maxout and ReLU had been tried with MLR. The accuracy of drug prediction for a specific disease using Multiple Linear Regression (MLR) with TanH Activation function is 90.5%, MLR with Maxout Activation function is 88.7%, MLR with Sigmoid function is 91.4% and MLR with ReLU Activation function is 93.8%. When compared with the prediction results for various diseases, the Rectified Linear Unit activation function with Multiple Linear Regression is computationally performed better results in prediction Drugs efficiently. The prediction values given for some drug names by TanH, Maxout, Sigmoid and ReLU function is given in Table 2.

| Drug Name | Tanh Prediction | Maxout Prediction | Sigmoid Prediction | ReLU Prediction |
|-----------|-----------------|-------------------|-------------------|----------------|
| Enalapril  | 0.94            | 0.92              | 0.94              | 0.96           |
| Ponstel   | 0.93            | 0.91              | 0.92              | 0.95           |
| Vyvanse   | 0.86            | 0.84              | 0.90              | 0.92           |
| elavil    | 0.91            | 0.90              | 0.92              | 0.94           |
| xanax     | 0.90            | 0.89              | 0.91              | 0.93           |
| Flagyl    | 0.90            | 0.87              | 0.91              | 0.94           |
| sarafem   | 0.90            | 0.89              | 0.92              | 0.94           |
| latisse   | 0.92            | 0.9                | 0.92              | 0.94           |
| Aldeara   | 0.93            | 0.91              | 0.92              | 0.95           |
| effexor-xr| 0.86            | 0.84              | 0.88              | 0.91           |

6. Conclusion

Nowadays reviews on any products in social media sources and other sources are very common. These reviews help in taking right decisions in most of the cases. Motivated by this, the personalized Drug–Disease prediction system using Multiple Linear Regression (MLR) with ReLU activation function was developed. As an initial step the model was developed with Multiple Linear Regression (MLR). To improve and achieve better results different activation functions such as Sigmoid, TanH, Maxout and ReLU had been tried with MLR. The accuracy of drug prediction for a specific disease using Multiple Linear Regression (MLR) with TanH Activation function is 90.5%, MLR with Maxout Activation function is 88.7%, MLR with Sigmoid function is 91.4% and MLR with ReLU Activation function is 93.8%. When compared with the prediction results for various diseases, the Rectified Linear Unit activation function with Multiple Linear Regression is computationally performed better results in prediction Drugs efficiently.

References

[1] Jianlin Wang et al., “Predicting Drug-disease Association Based on Ensemble Strategy”,
Frontiers in Genetics, Original Research, published on 03 May 2021, doi:10.3389/fgene:21.666575.

[2] Hui Cui et al., “The Prediction of Drug-Disease correlation Based on Gene Expression Data”, Research Article, Hindawi, BioMed Research International, Volume 2018, Article ID 402843, 6 pages, https://doi.org/10.1155/2018/4028473.

[3] Wen Zhang et al., “Predicting drug-disease associations by using similarity constrained matrix factorization”, BMC Bioinformatics (2018) 19:233 https://doi.org/10.1186/s12859-018-2220-4.

[4] Guangshang Wu et al., “Prediction of drug-disease associations based on ensemble meta paths and Singular Value Decomposition”. From the 17th Asia Pacific Bioinformatics Conference (APBC 2018), Wuhan, China, 14-16 January 2019.

[5] Qianfan Wu et al., “Deep Learning methods for Predicting Disease Status using Genomic Data”, HHS Public Access, JBiom Biostat, Author Manuscript, available in PMC 2019 May 22.

[6] Feng Huang et al., “Predicting Drug-Disease Associations via Multi-Task Learning based on Collective Matrix Factorization”, Frontiers in Bioengineering and BioTechnology.

[7] Han-Jing Jiang et al., “SAEROF: an ensemble approach for large-scale drug-disease association prediction by incorporating rotation forest and sparse auto encoder deep neural network”, Scientific Reports, Nature Research (2020) 10:4972 | https://doi.org/10.1038/s41598-020-61616-9.

[8] Felix Gräßer, Surya Kallumadi, Hagen Malberg, and Sebastian Zaunseder 2018. Aspect-Based Sentiment Analysis of Drug Reviews Applying Cross-Domain and Cross-Data Learning. In DH’18: 2018 International Digital Health Conference, April 23–26, 2018, Lyon, France. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3194658.3194677.

[9] Harleen Kaur et al., “A Proposed Sentiment Analysis Deep Learning Algorithm for analyzing COVID-19 Tweets”, Information Systems Frontiers, https://doi.org/10.1007/s10796-021-10135-7.

[10] Wittich CM et al., “Medication errors- an overview for clinicians”, Mayo Clin, Proc, 2014 Aug. 89(8):1116-25.

[11] Chen.M.R and Wang H.F (2013), “The reason and prevention of hospital medication error, Practical Journal of Clinical Medicine 4.

[12] Gulden Kay Uyanik, Nese Guler, “A Study on Multiple Linear Regression Analysis”, Procedia-Social Behavioral Sciences 106(2013) 234-240, Published by Elsevier Ltd.

[13] Hennadii Mokhort , “Multiple Linear Regression Model of Meningococcal Disease in Ukraine:1992-2015”, Computational and Mathematical Methods in Medicine, Volume 2020, Article ID 5105120, 7 page, https://doi.org/10.1155/2020/5105120.

[14] Buza.K,Peska.L, Koller.J (2020), “Modified Linear Regression predicts drug-target interactions accurately”, PLoS ONE 15(4):e0230726, https://doi.org/10.1371/journal.pone.0230726.

[15] Smita Rath, Alakananda Tripathy, Alox Ranjan Tripathy, “Prediction of new active cases of corona virus disease(COVID-19) using Multiple Linear Regression Model”, Diabetes & Metabolic Syndrome: Clinical Research & Reviews, Journal Homepage: www.elsevier.com/locate/dsx.