Enhancing Prediction of Drug Indication and Side Effects through Named Entity Recognition and Jointly Learning of Syntactic Structures of Sentences

D. Mohanapriya[1], Dr.R. Beena[2]

[1]Assistant professor, Department of Computer Science PSG College of Arts & Science Research Scholar of Kongunadu, Arts and Science College, Coimbatore, Tamil Nadu, India
[2]Associate Professor, Department of Computer Science, Kongunadu Arts and Science College, Coimbatore, Tamil Nadu, India

E-mail: mohanapriyasekar08@icloud.com[1], beenamridula@yahoo.co.in[2]

Abstract: The drug discovery process needs long time and cost to discover proper drug for treating the patients effectively. The unintended effects of drugs and the beneficial impact of drugs must be recognized because they may inflict severe patient's injuries due to unforeseen acts of the produced candidate drugs. One of the effective techniques is text mining it can find the hidden relation between genes, diseases and drugs from the huge volume of data. Predict drug Indications and Side effects using TOpic modeling and Natural language processing (PISTON) was a text mining method which used to find the association between drug-disease and drug-side effects. Natural Language Processing (NLP) is used to identify words which relate association among drugs and genes from the sentences which are collected from literatures where words represent drugs and genes co-occurred. The relation between drugs and genes is represented through building drug-topic probability matrix by topic modeling. From the drug-topic probability matrix, the drugs for phenotypes can be identified by training a classifier for high-rank topics of drugs. It also predicted the association between drug and side effects. However, expressive power of named entities and their potential for enhancing the quality of discovered topics has not received much attention in PISTON. So in this paper, an Improved PISTON (IPISTON) is proposed which enhance the quality of discovered topics through named entity recognition system and inducing the syntactic structure from unannotated sentences. Initially, the sentences from the collected literature data are extracted and a dependency graph is constructed using NLP. After that, a Gene Regulation Score (GRS) of each sentence is calculated to define the relationship between gene and diseases. The topic modeling is enhanced by finding the biomedical entities in the biomedical repository using Conditional Random Field (CRF) and Bi-directional Long-Short Term Memory-CRF (BLSTM-CRF). CRF is a sequence modeling framework which finds the biomedical entities through the conditional probability distributions of biomedical entities on collected documents. BLSTM-CRF is a deep learning technique which is used to enhance the performance of CRF based named entity recognition. Moreover, the syntactic structure of sentences is calculated through syntactic distance measure. The syntactic structure, biomedical entities and the drug-topic probability matrix is given as input to CRF, BLSTM-CRF, Naïve Bayes, CART and Logistic for prediction of drug-phenotype and drug-side effects associations.

Keywords: Drug discovery, drug-phenotype association, drug-side effect association, named entity recognition, Conditional Random Field, Bi-directional Long-Short Term Memory, syntactic structure.

1. INTRODUCTION

The interdisciplinary scientific area of medicinal science combines numerous areas of science and engineering, aiming at finding a new drug. Drug [1] may be regarded as molecules communicating with an appropriate target protein in order to disturb various biological interaction networks, for particular the signal transduction network, the metabolic pathway and the network for protein interaction. Drugs are utilized for infectious prevention and management to protect and improve safety. Drug discovery [2] is a more complicated process to define and possible goals for drugs. Most of the drug discovery is failed, due to the project failure and drug development cost. Nearly, all drugs have an affect so unexpected signs (i.e., side effects) may hurt and have serious consequences. So, it is more necessary to find the side effects for reducing the sever effects.

The extreme consequences are minimized by drug repositioning [3]. The significance of drug repositioning has risen significantly as the expense of new drug development has increased drastically. This also decreases time and expense for drug development. Various methods has been developed for drug repositioning according to the computational methods was proposed because of the exponential increase in available phenotypic or genomic data and the appearance of various methods for data analysis including machine learning and text mining. Text mining [4] is used to extract useful knowledge from high dimensional unstructured text data. Text mining works easier when using secondary data resources in order to help predict harmful outcomes as correct drug guidance leads to increased protection of drugs works reliable.

A text mining model called PISTON [5] for predicting the relation between drug-phenotype and drug-side effects pairs. Initially, sentences from literature where drugs and genes co-occurred were collected. After that, a dependency graph was constructed using Natural Language Processing (NLP). A Gene Regulation Score (GRS) was calculated to recognize the outcome of the drug on gene regulation. According to the topic modeling, grouped the regulatory relationships between genes and drugs which are often co-occur were grouped into one topic and drug-topic matrix was constructed through probability calculation of each drug occurred under various topics. Finally, a classifier was developed and learned from the of identified phenotype-drug-side effects matrix to predict unknown associations among phenotype-drug-side effects. However, expressive power of named entities and their potential for enhancing the quality of discovered topics has not received much attention in PISTON.

In this paper, the named entities is used as domain-specific terms for biomedical text content and classifiers such as Conditional Random Field (CRF) and Bi-directional Long-Short Term Memory (BLSTM-CRF) are used for named entity recognition. The recognition of named entities supports the topic modeling to provide high precision topics for disease, drug, gene and side effects. Furthermore, syntactic structure is induced from unannotated sentences in the biomedical context and leverages the inferred structure to learn a better language model. A syntactic distance is calculated between the topic and words to find the syntactic structure. It is given as additional input to the classifiers such as CRF, BLSTM-CRF, Naïve Bayes, Classification and Regression Tree (CART) and Logistic to predict the drug-phenotye and drug-side effect association effectively. This whole work is named as Improved PISTON (IPISTON).

2. LITERATURE SURVEY

A semi-supervised graph cut algorithm and three layer data integration [6] were proposed to predict the drug-disease interactions. The heterogeneous data were integrated into three layers based on the hierarchical fashion. Here, a novel weighted drug-disease pair network was built where a node was act as drug-disease pair which was weighted with the similarity score between two pairs. Then, the similar drug-disease pairs were was obtained to find an optimal graph cut of the network. The drug-disease pair with unknown relation was considered to have similar diagnosis relation within the same cut. However, multiple sources of data were not fused properly and reasonably.

A hybrid machine learning method [7] was proposed to predict the drug side effects according to the appropriate dataset features. In this method, data analytic techniques were employed to analysis the impact of drug distribution in the feature space and categorizing side effects based on the distribution of classes. Finally, domain-dependent strategies for each type were adopted to build the data models. However, this method finds difficult to predict the complex side effects.

A machine learning algorithm [8] was proposed for prediction of drugs side effects. The drug side effect prediction process was started with clustering the drugs with respect to the feature profiles using K-mean, Partitioning Around Medoids (PAM) and K-seeds techniques. Bayesian method has been used for each cluster to measure the matrix of probability score in which each dimension contains the score for indicating the probability of particular side effects for a drug belonging to the same cluster. However, the convergence speed of the clustering algorithm depends on the initial clusters.

A large-scale similarity-based framework [9] was presented to predict the interaction between drugs. A drug-related data and its knowledge were semantically combined that returned a knowledge graph. It described the drug attributes and its relationship with other associated objects including chemical structures, pathways and enzymes. The different similarity measures between all the drugs were computed in a scalable and distributed framework with the aid of knowledge graph. The resulting similarity metrics were used to develop features for a large-scale logistic regression model to predict the interactions between drugs. However, the logistic regression model was difficult to capture complex relationships.

An optimized drug similarity framework [10] was proposed to enhance the performance of side effect prediction. The process of this framework was started with combining four various drug similarities as the comprehensive similarity and fine tuned by clustering. After that, the optimized similarity was improved by the indirect drug similarity to predict the side effects of drugs. However, it has low F1-score.

A computational method [11] was proposed to predict the side effects in drugs based on the features of determined available drug and association between side effects and drugs. Computational method developed in low-dimensional space, which extracted features of side effects and drugs. This method was differs from the traditional conventional matrix factorization approach, and can found the biomedical context into account. The matrix factorization was an efficient technique, and determined the undetermined relationship according to the known association-based matrix. However, it does not predict the association between drug and side effects in the SIDER database.

A binary classification model [12] was proposed to predict the drug side effects through heterogeneous information of drugs. In order to encode the each drug-side effects, similarity based method was applied. Also, random forest was adapted to predict the drug side effects. It was considered that the drug has side effects when the prediction outcome was positive and wise versa. However, this model not possible to detect the side effect in early stage.

A Network Topological Similarity-based Classification (NTSIM-C) [13] method was proposed to predict the relationship between drug and disease. In NTSIM-C, relationship between drug and disease was defined as a feature vector. It consisted of similarity scores between drugs and the other drugs and the vector related to the row vector in drug-drug similarity matrix. A linear neighborhood similarity matrix was constructed for drugs and the linear neighborhood similarity matrix was calculated for a disease. A vector was created for the relationship between the disease and the drug. The vector combined the similarity vector of both drug and disease. The relationship between disease and drug was predicted based on the vector. However, this method is not applicable for large datasets.

3. PROPOSED METHODOLOGY

In this section, the proposed Improved PISTON (IPISTON) is described in detail for prediction of drug-phenotype association and drug-side effect association.

Initially, the sentences are collected where drugs and genes are co-occurred. Then, a word dependency graph was built using Natural Language Processing (NLP) for every sentence. The identified words that represent connections between drug and genes are used to find side effects of drugs based on gene regulation. A Gene Regulation Score (GRS) of sentences are computed from the words occurred in the sentences that indicates up and down regulation of genes. The up and down regulation represented by words are identified by comparing these words with the gene keyword dictionary. GRS of a drug-gene pair in a sentence is calculated as,

$$GRS(Dis, Gene) = (+1)^{up}(-1)^{down} \qquad (3.1)$$

In Eq. (3.1), Dis denotes the disease, $Gene$ denotes the gene, up denotes the quantity of recognized up-regulation in a dependency graph and $down$ denotes the quantity of recognized down-regulation in a dependency graph. $GRS(Dis, Gene)$ of the sentence is not computed for the sentences those are not containing words represents up or down regulation. The $GRS(Dis, Gene)$ represents score of sentences as +1 for having more number of up-regulation words,-1 for having more number of up-regulation words.

The optimal number of topics is required to build a drug-topic matrix that is determined using log-likelihood. The named entities and GRS are used as features in the drug-probability matrix and construct a matrix using genes that affects drugs and their regulatory relationship based on topic modeling. The drug-topic probability matrix is constructed by grouping genes and their regulatory association that co-occur frequently from all drugs into optimal number of topics. Then, the genes of each drug with their regulatory association are compared with the genes in optimal number of topics and computed the probability that the drug contained the topic.

3.1 Topic Modeling and named entity recognition

The document collected from the literature comprises of words related with drugs and several topics about drugs. The individual topics comprises of words for representing gene, phenotypes, drugs and side effects. In this paper, each document refers a drug; the topic refers set of similar genes and their regulatory relationships and the word refers a gene.

The recognition of named entities in the biomedical repository supports the topic modeling to provide more accurate topics for drug-phenotype and drug-side effect association. CRF and BLSTM-CRF are used for named entity recognition.

a) CRF based named entity recognition

For building probabilistic models, CRF is a sequence modeling framework, conditional probability distributions on an undirected graph model. A linear-chain CRF is applied for named entity recognition. The conditional probability of linear chain CRF determined on observation x (i.e., document) and a random variable y (i.e., biomedical entities) as follows:

$$P(y|x) = \frac{1}{Z(x)} \exp\left(\sum_{j}\sum_{i=1}^{n} \delta_j trans_j(y_{i-1}, y_i, x, i) + \sum_{k}\sum_{i=1}^{n} \beta_k state_k(y_i, x, i)\right) \qquad (3.2)$$

In Eq. (3.2), $Z(x)$ is a denominator scaling factor that makes sure sum of the posterior probabilities is exactly equal to one, $trans_j(y_{i-1}, y_i, x, i)$ represents the transition function of $i-1$ and i-th label sequences for all observation sequences, $state_k(y_i, x, i)$ denotes a state of i-th observation sequence feature. δ_j and β_k parameters are approximated by cross-entropy computed from training data.

b) Bi-directional LSTM-CRF based named entity recognition

A deep learning technique is applied to enhance the performance of named entity recognition. The bidirectional BLSTM-CRF consists of input layer for getting inputs, hidden layer for feature weight updating, CRF layer and output layer for classifcation. The input feature is a word vector that was extracted from sentences of each document. The extracted words are converted in to a vector representation by a word embedding method BioASQ [14].

The output layer provides the label for entity which relates to drug, phenotype or side effects with probability values. Named entity recognition start with little quantity of physically interpreted corpus and after that construct a classifier learned with the annotated corpus. The constructed classifier then analyzes and provides the label for unlabeled data. The prediction results are included to the training data for the retraining of classifier for further predictions. Thus the performance of the classifier progressively increased by re-training with the machine-labeled corpus and manfully labeled data. The named entity recognition is trained by using back propagation method and the dropout rate for learning from training data is fixed as 0.5. The basic structure of BLSTM-CRF is depicted in Figure 3.2.

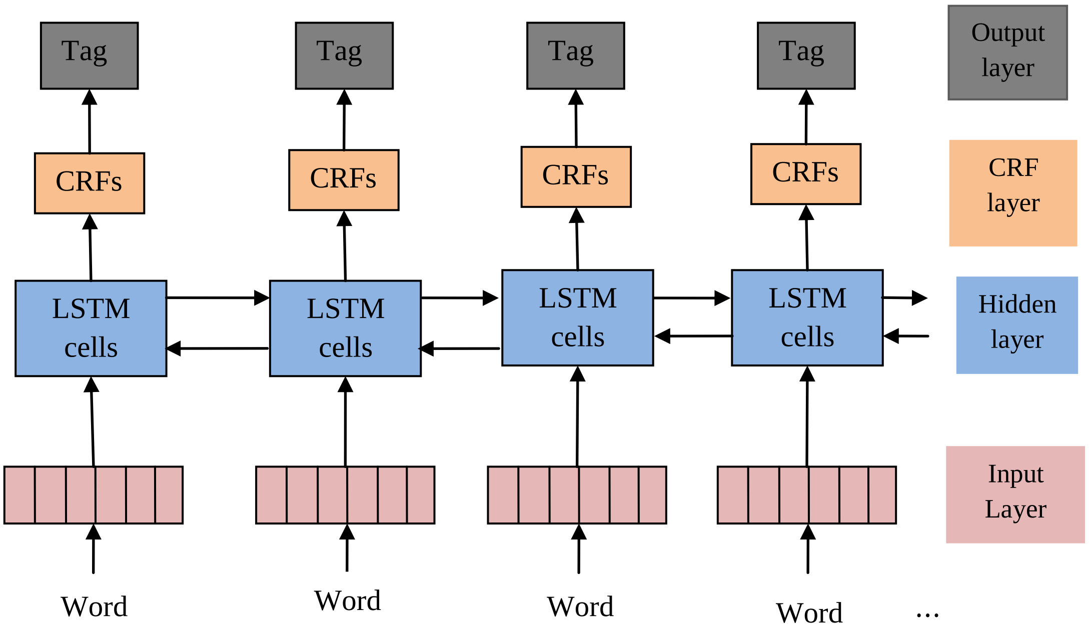

Figure.3.2 Basic structure of BLSTM-CRF for named entity recognition

3.2 Inducing syntactic structure from unannotated document

Long-Short Term Memory (LSTM) is used to induce the syntactic structure from unannotated document and leverage the inferred structure to learn a better language model. The syntactic structure is modeled by using stick-breaking process. It calculates the probability distribution between topics and genes. LSTM is a kind of recurrent neural network which effectively handle the sequential data. In LSTM, the current hidden state depends on the last hidden state. The hidden state holds the syntactic structure of the unannotated document for the last time step. However, it suffers from ignoring the real dependency relation that dominates the structure of genes. A skip-connection is used to integrate structure dependency relation with LSTM. By integrating the structure dependency relation with LSTM, the current hidden state depends on the last hidden state, earlier hidden states which have a syntactic relation to the current one. The skip connections are controlled by gates. It is defined by introducing a latent variable l_t to denote local structural context of document x_t:

- The location of the left most child belonging to the left most sibling y_i is l_t while the left most child of a subtree y_i is x_t.
- The location of x_t's left most sibling is l_t while the left most child of any subtree is not x_t. The gates are represented by,

$$g_i^t = \begin{cases} 1, & l_t \leq i < t \\ 0, & 0 < i < l_t \end{cases} \qquad (3.3)$$

By using this structure, the sibling dependency relation is replicated by at least one skip-connect. Through the skip-connect relation between the nodes, the parent-to-child relation will be implicitly modeled. The following Eq. (3.4) shows the model newly updates the hidden states.

$$m_t = h(x_t, m_0, \dots, m_{t-1}, g_0^t, \dots, g_{t-1}^t) \qquad (3.4)$$

The probability distribution for next word is approximated by:

$$P(x_{t+1}|x_0, x_1, \dots x_t) \approx P\big(x_{t+1}; f(m_0, \dots, m_{t-1}, g_0^t, \dots, g_{t-1}^t)\big) \quad (3.5)$$

In Eq. (3.5), g_i^t are gates that control skip-connections. Both f and h have a structured mechanism which obtain g_i^t as input and compels the model to focus on the most correlated information. In order to model the local structure of document, a probabilistic view is used. At time step t, $P(l_t|x_0, x_1, \dots x_t)$ denotes the probability of selecting one out of t possible local structures. The Stick-Breaking process is used to model the distribution and it given as follows,

$$P(l_t|x_0, x_1, \dots x_t) = (1 - \alpha_i^t) \prod_{j=i+1}^{t-1} \alpha_j^t \quad (3.6)$$

After the time step $i + 1, \dots, t - 1$ have their probabilities assigned, $\prod_{j=i+1}^{t-1} \alpha_j^t$ is remaining probability, $1 - \alpha_i^t$ is the portion of rest of the probability that assigned to time step i. The expectation of gate value g_i^t is the cumulative distribution function of $P(l_t|x_0, x_1, \dots x_t)$. Hence, the discrete gate value is replaced by its expectation:

$$g_i^t = P(l_t \leq i) = \prod_{j=i+1}^{t-1} \alpha_j^t \quad (3.7)$$

A soft gating vector is used to approximate Eq. (3.4) and (3.5) with the above relaxations for updating the hidden state (syntactic structure). A hypothesis is considered to parameterize α_j^t. In the hypothesis, it is considered that genes which are co-occur frequently in all drugs should have a closer syntactic relation within themselves, and that this syntactical proximity can be denoted by scalar value. A syntactic distance is introduced to model the syntactical proximity. A set of K real valued scalar variables $d_0, \dots d_{K-1}$ with d_i denoting a measure of the syntactic relation between the genes which are frequently co-occur in all drugs and the genes. For time t, the closest document x_j is determined which have largest syntactic distance than d_t. Eq. (3.8) defines the α_j^t

$$\alpha_j^t = \frac{hardtanh\left(\left(d_t - d_j\right).\gamma\right) + 1}{2} \quad (3.8)$$

In Eq. (3.8), $hardtanh\,(x) = \max(-1, \min(1, x))$ and γ is the temperature parameter that controls the sensitivity of α_j^t to the differences between distances. The values of syntactic distance have more conceptual definition. If two adjacent genes are peers of each other, the syntactic difference should be around zero; whereas if they belong to different sub-trees, they should have a greater syntactic distance. In the worst scenario, the syntactic distance is closer to 1 if the two genes do not have a subtree in general. The drug-topic matrix is concatenated with syntactic distance and it is given as input to the classifiers such as CRF, BLSTM-CRF, Naïve Bayes, CART and Logistic for prediction of drug-phenotype and drug-side effects associations. The classifiers are trained using known drug-phenotype and drug-side effect associations which are collected from CTD and SIDER respectively.

3.3 IPISTON

The overall flow of the IPISTON is shown in Figure 3.1.

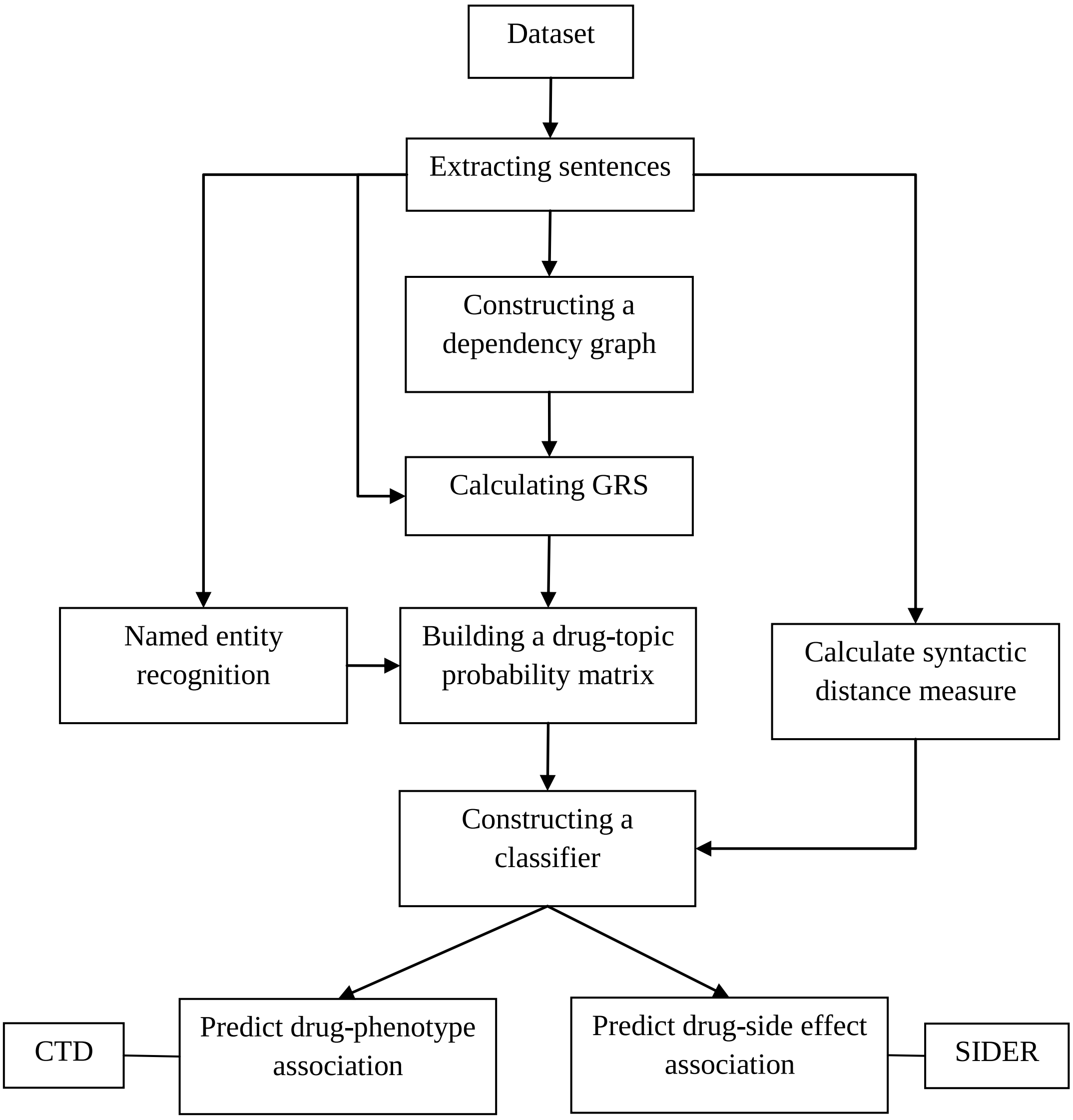
Dataset
Extracting sentences
Constructing a dependency graph
Calculating GRS
Named entity recognition
Building a drug-topic probability matrix
Calculate syntactic distance measure
Constructing a classifier
CTD
Predict drug-phenotype association
Predict drug-side effect association
SIDER

Figure 3.1 Overall flow of IPISTON

IPISTON Algorithm

Step 1: Collect the literature data from biomedical repository.

Step 2: Extract the sentences in which drugs and genes co-occur from the abstract of literature data.

Step 3: Find the biomedical entities in the sentences as tag using CRF and BLSTM-CRF.

Step 4: Construct a dependency graph and identified words defining the association between drug and gene.

Step 5: Calculate GRS of each sentence to recognize the effect of the drug on gene regulation.

Step 6: Model the topic (gene) from the document (drug) by considering the regulatory association that frequently co-occur in different drugs.

Step 7: Calculate the syntactic distance of topic and word using Eq. (3.8).

Step 8: Construct a drug-topic probability matrix using GRS and biomedical entities.

Step 9: Train the CRF, BLSTM-CRF, Naïve Bayes, CART and Logistic classifiers with known association of drug-phenotype and drug-side effect along with the probability matrix and syntactic distance to predict the unknown association of drug-phenotype and drug-side effect.

4. EXPERIMENTAL RESULTS

In this section, the efficiency of PISTON and IPISTON are tested in terms of accuracy, sensitivity, specificity and z-score. For the experimental purpose, PubMed, DrugBank, KEGG DRUG and PharmGKB datasets are used. PubMed is a database which provides biology literature and it collects 1,454,763 abstracts from 6975 journals. The official names of the drugs are collected from DrugBank to find the names of drugs in sentences. DrugBank provides comprehensive drug data in cheminformatics and bioinformatics. KEGG provides for approved drugs in U.S., Japan and Europe. From the DrugBank, the official names of 2196 approved drugs are obtained and their synonyms are collected from KEGG DRUG. PharmGKB is a database that provides information on genetic variation in drug responses. 26,886 gene symbols are obtained from PharmGKB database. The drug-side effect associations are collected from SIDER and 411 out of 684 drugs are used those have a unique MeSH id in SIDER. Table.4.1 shows the phenotypes and side effects which are considered in the experiment.

Table.4.1 Phenotypes and side effects

S.No.	Phenotypes	Side effects
1.	Myelogenous leukemia	Cheilitis
2.	Colitis	Redness
3.	Small cell lung cancer	Ulcer
4.	Ovarian cancer	Hypothermia
5.	Pulmonary edema	Hyperlipidaemia
6.	Cystitis	Sleep disturbance
7.	Non-small cell lung cancer	Heartburn
8.	Melanoma	Inflammation
9.	Bladder cancer	Laryngitis
10.	Heart disease	Eruption
11.	Hyperglycemia	Cataract
12.	Cerebrovascular disease	Ageusia
13.	Prostate cancer	Gout
14.	Breast cancer	Delirium
15.	Bradycardia	Gastritis
16.	Hypotension	Eczema
17.	Tachycardia	Amnesia
18.	Proteinuria	Diplopia
19.	Depressive disorder	Ataxia
20.	Anxiety disorder	Fatigue

Table 4.2 shows the candidate drugs which are considered in the experiment.

Table.4.2 Candidate Drugs

S.No.	**Candidate Drugs**
1.	Adenine
2.	Adenosine
3.	Caffeine
4.	Cocaine
5.	Enoxaparin
6.	Glucose
7.	Cisplatin
8.	Dexamethasone
9.	Gemcitabine
10.	Glutathione
11.	Glycine
12.	Metformin
13.	Mifepristone
14.	Nitric oxide
15.	Oxygen
16.	Oxaliplatin
17.	Phenol
18.	Paclitaxel
19.	Phenobarbital
20.	Progesterone
21.	Progesterone
22.	Simvastatin
23.	Sorafenib
24.	Temozolomide
25.	Testosterone
26.	Tetracycline
27.	Urea
28.	Water

4.1 Accuracy

Accuracy is defined as the number of all correct prediction of drug-phenotype (side effects) association divided by the total number of phenotype (side effects) prediction made. This is defined as a ratio of appropriately classified data to overall classified data.

$$Accuracy = \frac{True\ Positive\ (TP) + True\ Negative\ (TN)}{TP + False\ Positive\ (FP) + TN + False\ Negative\ (FN)}$$

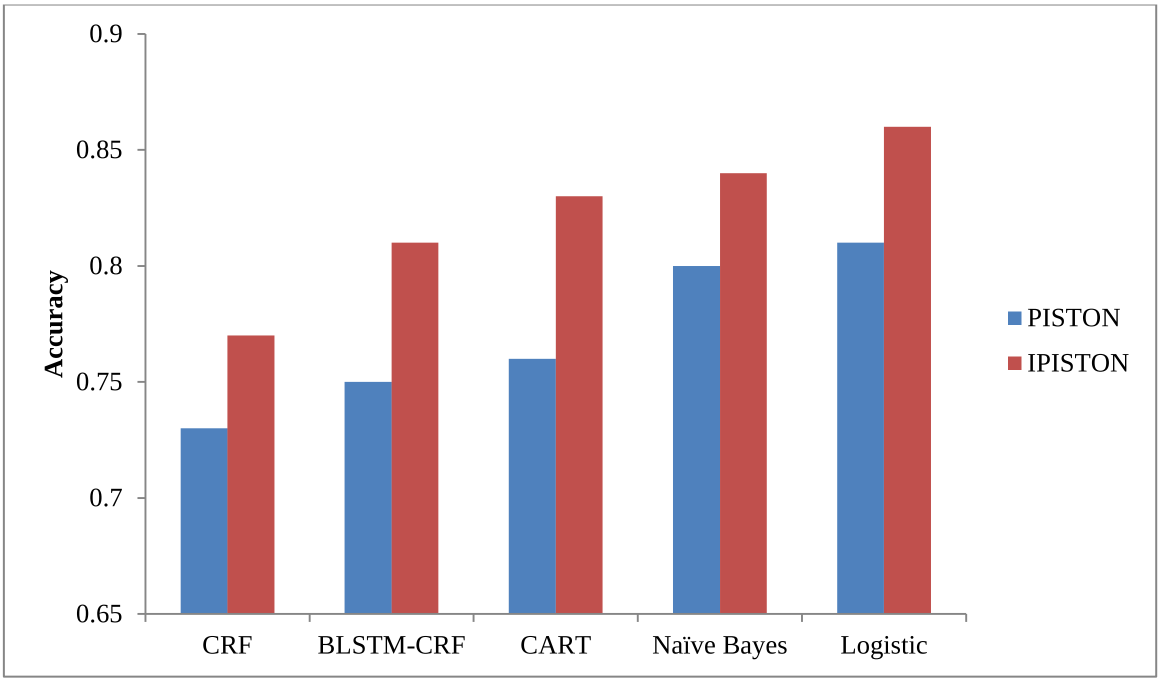

Figure.4.1 Comparison of Accuracy for phenotypes

Figure 4.1 shows the comparison of accuracy of PISTON and IPISTON for prediction of drug-phenotype association with different classifiers. X axis denotes the drug-phenotype association prediction methods and Y axis denotes the accuracy. The accuracy of IPISTON is 6.17% greater than PISTON for logistic classifier. From the Figure 4.1, it is proved that the proposed IPISTON has high accuracy than PISTON for drug-phenotype association with CRF, BLSTM-CRF, CART, Naïve Bayes and logistic classifiers.

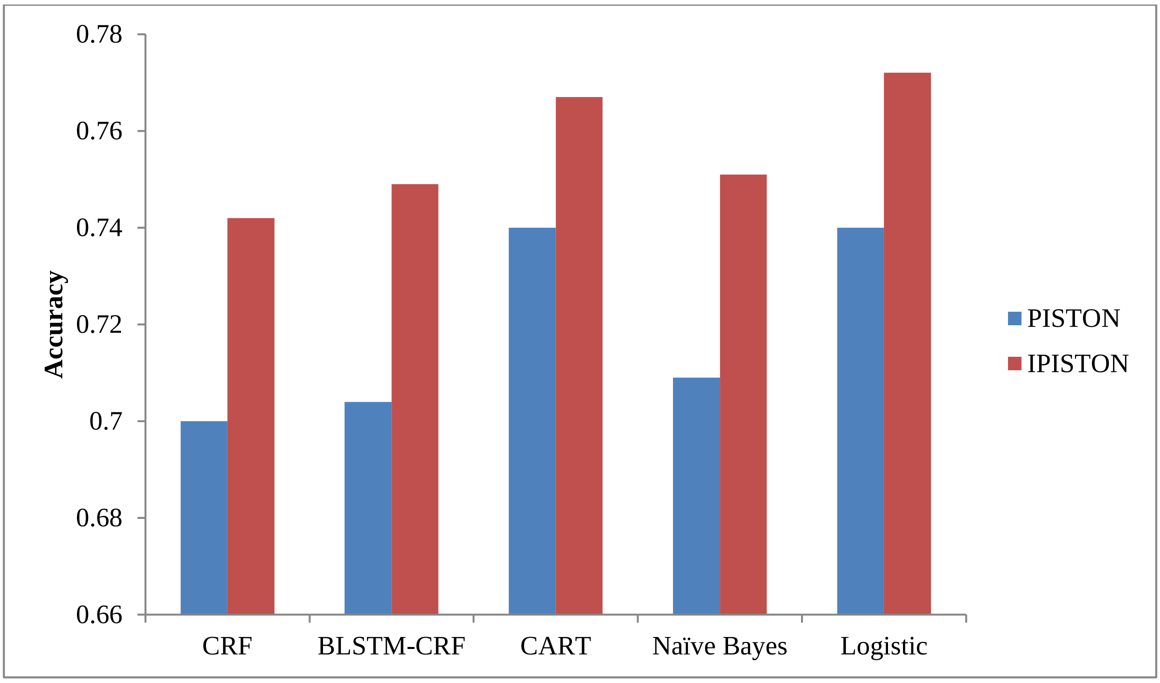

Figure.4.2 Comparison of Accuracy for side effects

Figure 4.2 shows the comparison of accuracy of PISTON and IPISTON for prediction of drug-side effects association with different classifiers. X axis denotes the drug- side effects association prediction methods and Y axis denotes the accuracy. The accuracy of IPISTON is 4.32% greater than PISTON for logistic classifier. From the Figure 4.2, it is proved that the proposed IPISTON has high accuracy than PISTON for drug-side effects association with CRF, BLSTM-CRF, CART, Naïve Bayes and logistic classifiers.

4.2 Sensitivity

It is used to measure the fraction of positive patterns that are correctly predicted. It is calculated as,

$$Sensitivity = \frac{TP}{(TP + FN)}$$

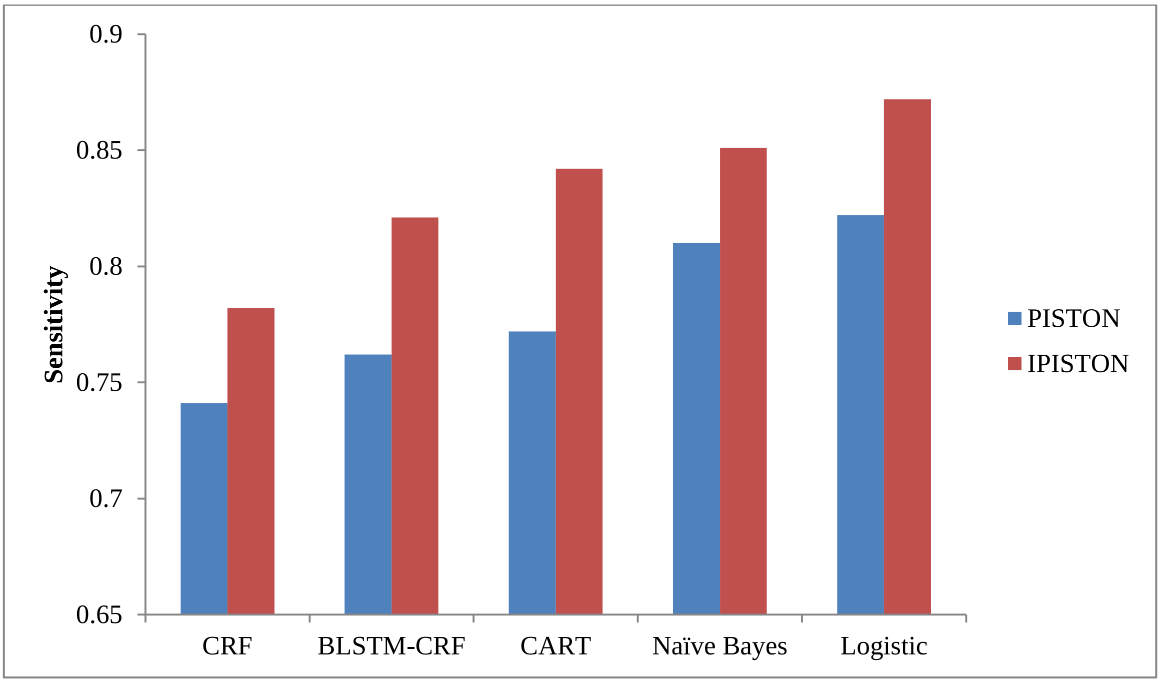

Figure.4.3 Comparison of Sensitivity for phenotypes

Figure 4.3 shows the comparison of sensitivity of PISTON and IPISTON for prediction of drug-phenotype association with different classifiers. X axis denotes the drug-phenotype association prediction methods and Y axis denotes the sensitivity. The sensitivity of IPISTON is 6.08% greater than PISTON for logistic classifier. From the Figure 4.3, it is proved that the proposed IPISTON has high sensitivity than PISTON for drug-phenotype association with CRF, BLSTM-CRF, CART, Naïve Bayes and logistic classifiers.

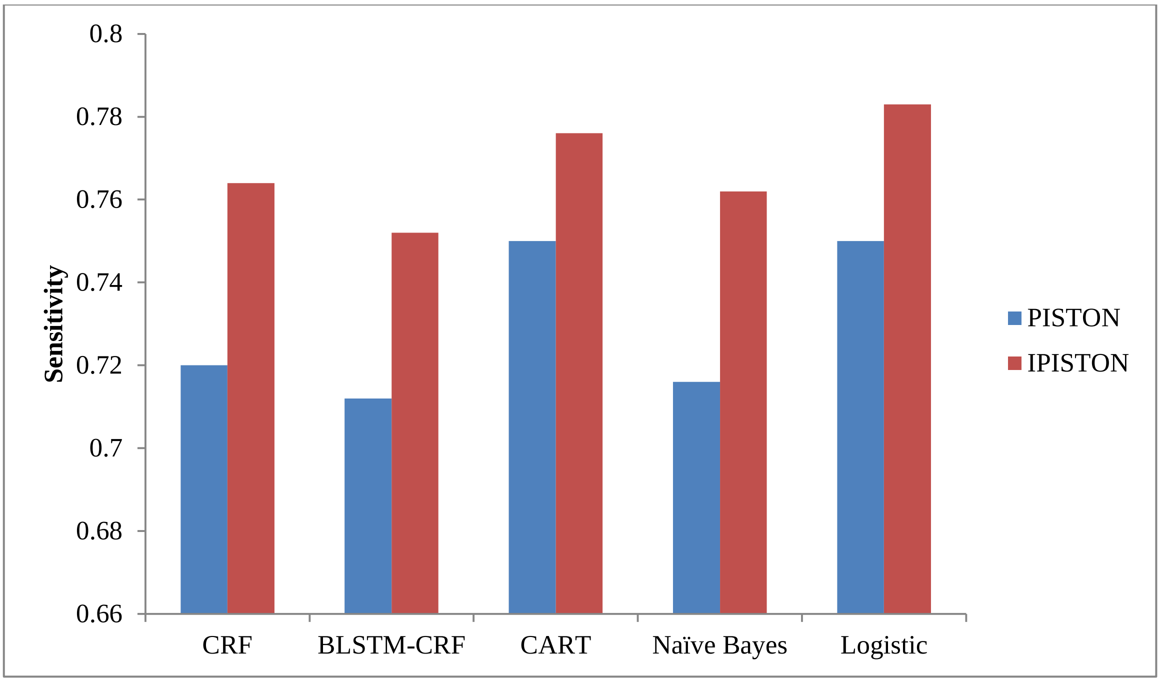

Figure.4.4 Comparison of Sensitivity for side effects

Figure 4.4 shows the comparison of sensitivity of PISTON and IPISTON for prediction of drug-side effects association with different classifiers. X axis denotes the drug- side effects association prediction methods and Y axis denotes the sensitivity. The sensitivity of IPISTON is 4.4% greater than PISTON for logistic classifier. From the Figure 4.4, it is proved that the proposed IPISTON has high sensitivity than PISTON for drug-side effects association with CRF, BLSTM-CRF, CART, Naïve Bayes and logistic classifiers.

4.3 Specificity

Specificity of a test is the proportion of correctly predicted drug-phenotype (side effect) association with the summation of correctly predicted drug-phenotype (side effect) association and wrongly predicted drug-phenotype (side effect) association.

$$Specificity = \frac{TN}{TN + FP}$$

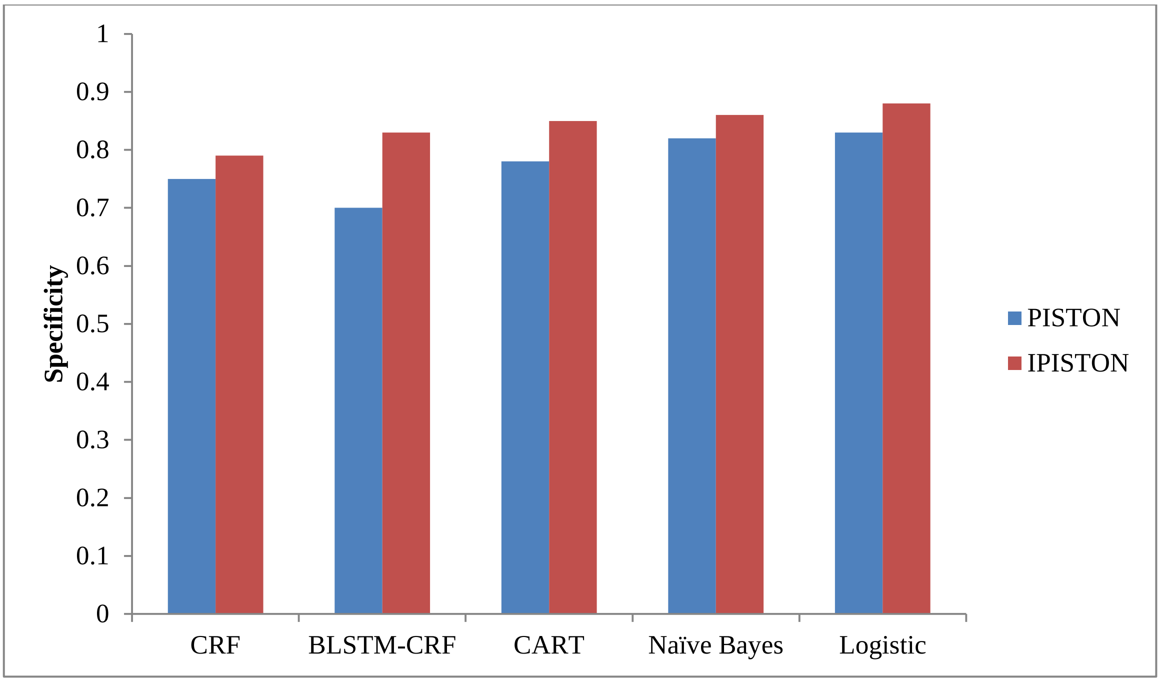

Figure.4.5 Comparison of Specificity for phenotypes

Figure 4.5 shows the comparison of specificity of PISTON and IPISTON for prediction of drug-phenotype association with different classifiers. X axis denotes the drug-phenotype association prediction methods and Y axis denotes the specificity. The specificity of IPISTON is 6.02% greater than PISTON for logistic classifier. From the Figure 4.5, it is proved that the proposed IPISTON has high specificity than PISTON for drug-phenotype association with CRF, BLSTM-CRF, CART, Naïve Bayes and logistic classifiers.

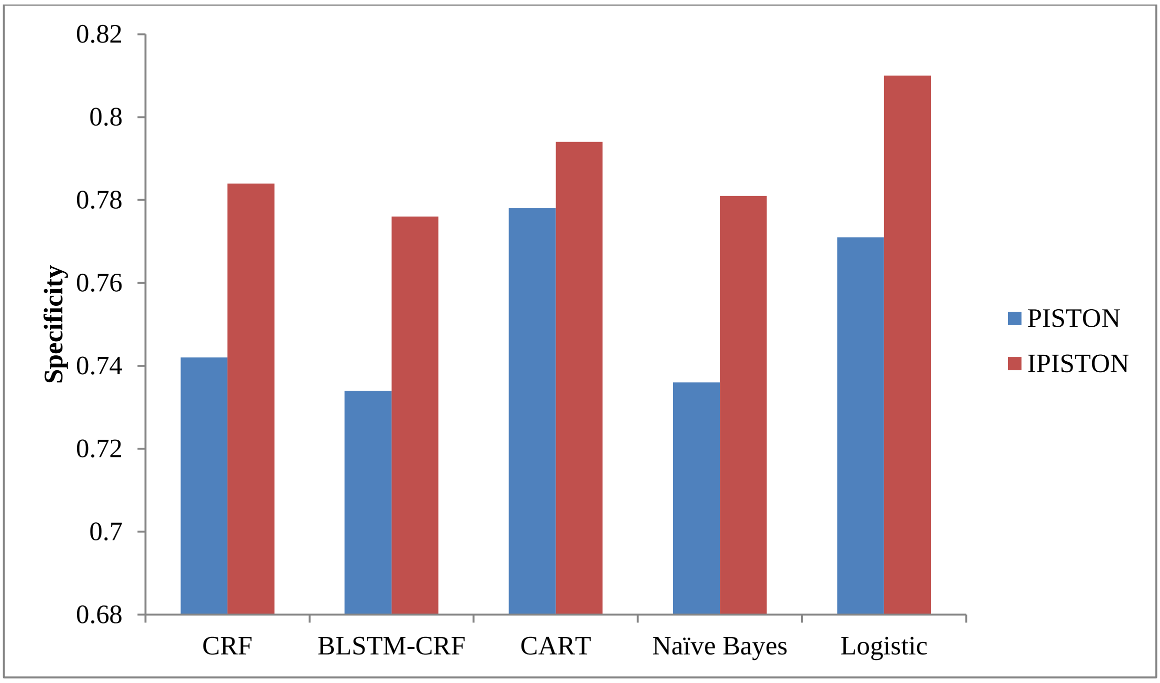

Figure.4.6 Comparison of Specificity for side effects

Figure 4.6 shows the comparison of specificity of PISTON and IPISTON for prediction of drug-side effects association with different classifiers. X axis denotes the drug- side effects association prediction methods and Y axis denotes the specificity. The specificity of IPISTON is 5.06% greater than PISTON for logistic classifier. From the Figure 4.6, it is proved that the proposed IPISTON has high specificity than PISTON for drug-side effects association with CRF, BLSTM-CRF, CART, Naïve Bayes and logistic classifiers.

4.4 Z-score

Z-score is a numerical measurement that describes closeness between drug and phenotype (side effects). It can be calculated as,

$$Z - score\ (T,P) = \frac{d_{short}(T,P) - \mu_{d_{short}(T,P)}}{\sigma_{d_{short}(T,P)}}$$

$$Z - score\ (T,S) = \frac{d_{short}(T,S) - \mu_{d_{short}(T,S)}}{\sigma_{d_{short}(T,S)}}$$

Where, $d_{short}(T,P)$ is the shortest distance between T (drug) and P (phenotype), $d_{short}(T,S)$ is the shortest distance between T (drug) and S (side effect), $\mu_{d_{short}(T,P)}$ is mean of $d_{short}(T,P)$ values calculated for all drugs, $\mu_{d_{short}(T,S)}$ is mean of $d_{short}(T,S)$ values calculated for all drugs, $\sigma_{d_{short}(T,P)}$ is the standard deviation of $d_{short}(T,P)$ values calculated for all drugs and $\sigma_{d_{short}(T,S)}$ is the standard deviation of $d_{short}(T,S)$ values calculated for all drugs.

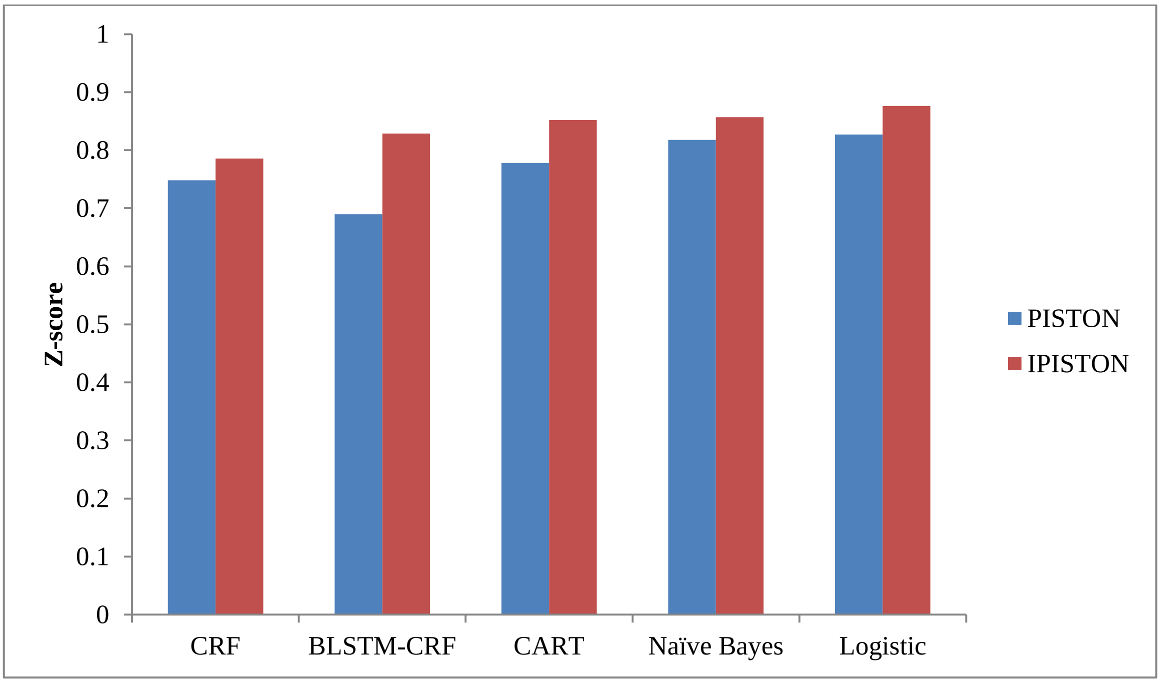

Figure.4.7 Comparison of Z-score for phenotypes

Figure 4.7 shows the comparison of z-score of PISTON and IPISTON for prediction of drug-phenotype association with different classifiers. X axis denotes the drug-phenotype association prediction methods and Y axis denotes the z-score. The z-score of IPISTON is 5.93% greater than PISTON for logistic classifier. From the Figure 4.7, it is proved that the proposed IPISTON has high z-score than PISTON for drug-phenotype association with CRF, BLSTM-CRF, CART, Naïve Bayes and logistic classifiers.

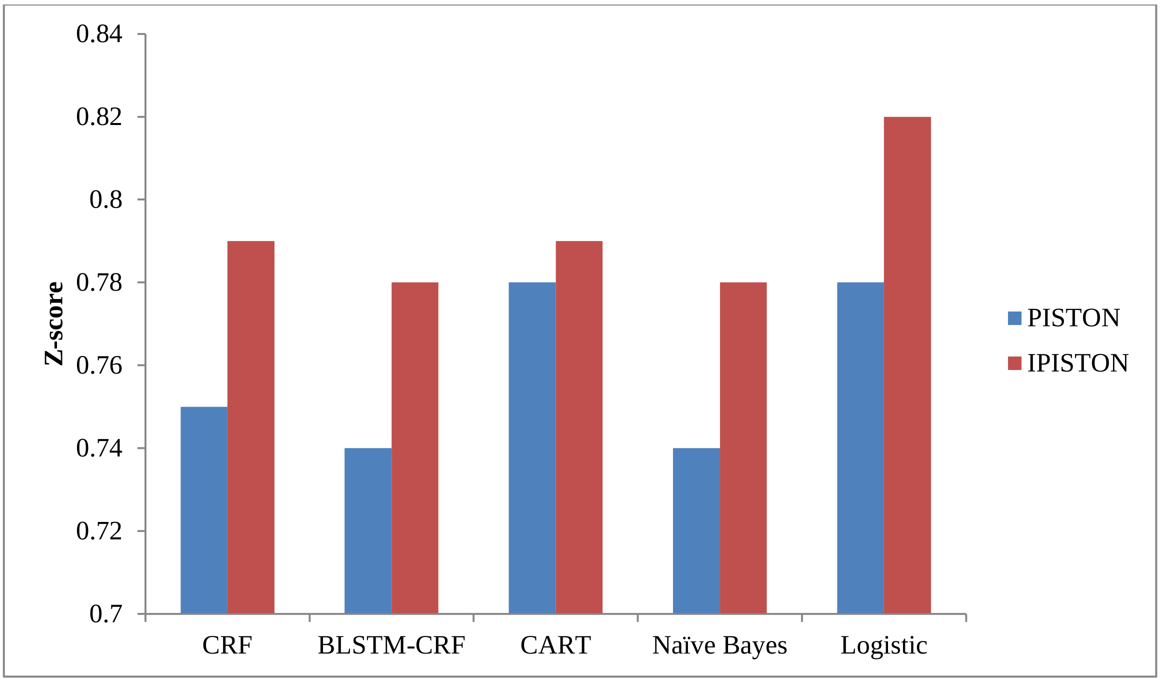

Figure.4.8 Comparison of Z-score for side effects

Figure 4.8 shows the comparison of z-score of PISTON and IPISTON for prediction of drug-side effects association with different classifiers. X axis denotes the drug- side effects association prediction methods and Y axis denotes the z-score. The z-score of IPISTON is 5.13% greater than PISTON for logistic classifier. From the Figure 4.8, it is proved that the proposed IPISTON has high z-score than PISTON for drug-side effects association with CRF, BLSTM-CRF, CART, Naïve Bayes and logistic classifiers.

5. CONCLUSION

In this paper, an IPISTON is proposed for prediction of drug-phenotype and drug-side effect association using text mining techniques. Initially, data related to drug, gene and side effects are collected from biomedical repository. Then, the sentences in the collected document are extracted and a dependency graph is constructed using NLP. A GRS is calculated for each sentence to find the effect of the drug on gene regulation. In the collected documents, the topics are modeled and then the biomedical entities are determined using CRF and BLSTM-CRF which enhance the quality of topic modeling. The syntactic distance between the topic and words are computed which refines the syntactic structure of the sentences. The syntactic distance and drug-topic probability matrix are given as input to the CRF, BLSTM-CRF, Naïve Bayes, CART and Logistic to predict the drug-phenotye and drug-side effect association effectively. The experimental results prove that the proposed IPISTON has better accuracy, sensitivity, specificity and z-score than PISTON for prediction of drug-phenotype and drug-side effect associations.

6. REFERENCES

[1] Jahid, M. J., & Ruan, J. (2013, December). An ensemble approach for drug side effect prediction. In *Bioinformatics and Biomedicine (BIBM), 2013 IEEE International Conference on* (pp. 440-445). IEEE.

[2] Sharma, A., & Rani, R. (2018). BE-DTI': Ensemble Framework for Drug Target Interaction Prediction using Dimensionality Reduction and Active Learning. *Computer Methods and Programs in Biomedicine.*

[3] Pillaiyar, T., Meenakshisundaram, S., Manickam, M., & Sankaranarayanan, M. (2020). A medicinal chemistry perspective of drug repositioning: Recent advances and challenges in drug discovery. *European Journal of Medicinal Chemistry*, 112275.

[4] Edwards, I. R., & Aronson, J. K. (2000). Adverse drug reactions: definitions, diagnosis, and management. The lancet, 356(9237), 1255-1259.

[5] 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.

[6] Wu, G., Liu, J., & Wang, C. (2017). Predicting drug-disease interactions by semi-supervised graph cut algorithm and three-layer data integration. *BMC medical genomics*, *10*(5), 79.

[7] 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.

[8] Dimitri, G. M., & Lió, P. (2017). DrugClust: a machine learning approach for drugs side effects prediction. *Computational biology and chemistry*, *68*, 204-210.

[9] 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.

[10] 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.

[11] 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

[12] 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.

[13] 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*, *145*, 51-59.

[14] Sarrouti, M., & El Alaoui, S. O. (2017, August). A biomedical question answering system in BioASQ 2017. In *BioNLP 2017* (pp. 296-301).