Research Article

A Comparative Analysis of Adverse Drug Reaction Prediction Techniques

Lakshmi K S[a*], Vadivu G[b]

[a]Department of Information Technology, Rajagiri School of Engineering & Technology, Kochi, Kerala, India.

[b]Department of Information Technology, SRM Institute of Science and Technology, Kattankulathur, Chennai, India.

***Corresponding author's E-mail:** lekshmy.shalu@gmail.com

Received: 30-12-2018; **Revised:** 14-02-2019; **Accepted:** 25-02-2019.

ABSTRACT

Patients subjected to medical treatments are often susceptible to adverse drug reactions. Many comorbid disease conditions arise as a result of adverse drug reactions which in turn increase the complexity of treatment procedures. Usually the side effects of a drug are identified prior to its introduction into market. But many of the adverse reactions can be identified only after the post-marketing phase. Once the possible adverse reactions of a drug are known, then effective steps can be taken to alleviate the reaction. The objective of this research is to compare the existing techniques used in the prediction of adverse reactions of drug by combining information from electronic health records and online healthcare forums. Adverse reactions of drug are predicted using 4 separate classifiers: SVM, Naïve Bayes, Decision Tree and KNN. Individual classifiers were tested and the accuracy seemed to be increased when using an ensemble of these classifiers. Among the 4 classifiers, SVM showed an accuracy of 84%. Usage of ensemble classifier improved the accuracy to 87%.

Keywords: Support vector machine, Naïve Bayes, Decision Tree, KNN, LSTM, Adverse Drug Reaction.

INTRODUCTION

Adverse drug reaction (ADR) is defined as a harmful reaction caused by the intake of a medication. ADRs may occur following a single dose or prolonged administration of a drug or result from the combination of two or more drugs. Large amount of money and energy have to be invested for studying the adverse effects of drugs. Adverse drug reactions are different from side effects. Side effects of a drug can be positive or negative; negative side-effects are considered as Adverse Drug Reactions. If ADRs are not properly diagnosed and treated, it can lead to harmful and fatal health conditions.

Pharmacovigilance is the science and activities related to understanding, detecting, assessing and preventing adverse effects of drug-related problems. Usually the adverse reactions of drugs are conducted in the post-marketing phases, even though it can be done during pre-marketing phase also. Recent advancements in technology contributed large scale availability of text data in social media and health related forums. Electronic medical health records are also sources of text data. These data sources can be used for pharmacovigilance that helps in early detection and prevention of adverse drug reactions which in turn can contribute to enhanced survival rate.

Various researches have been conducted in the field of pharmacovigilance during the past few years. During early stages, data mining techniques were widely used for the prediction of adverse drug reaction from electronic health records and clinical reports. Ioannis Korkontzelos[1] et al. examined the scope of sentiment analysis based feature extraction for locating ADR mentions from social media and online health forums. They proved that adding sentiment analysis features can slightly improve the performance of ADR identification method. Hariprasad Sampathkumar[2] et.al presented a method for mining ADRs from online health forums using Hidden Markov Model. In this approach, they extracted messages from online healthcare forums containing ADRs and further processed these messages for finding out adverse drug reactions. A.Pushpa[3] et.al., utilized LSTM method to overcome the difficulties in extracting the user post having misspellings, abbreviations. This method proved better than the existing machine learning techniques for ADR prediction. Andy W. Chen[4] et al. compared the performance of different machine learning models for predicting the ADR outcome. He used machine learning techniques for predicting ADR based on patient characteristics and drug usage. WenchengSun[5] et.al. discussed in detail about the processing of Electronic Medical Records and the challenges involved in it. They conducted a detailed study on the different kinds of data mining techniques to extract information from EMR database. Trung Huynh et al. proposed two neural network models, Convolutional Recurrent Neural Network (CRNN) and Convolutional Neural Network with Attention (CNNA) for ADR classification[6]. Experimental results showed that all the Neural Network architectures outperform the traditional Maximum Entropy classifiers trained from n-grams with different weighting strategies on ADE datasets. Stefanie Friedrich[7] et al. conducted a comparative study of machine learning algorithms for adverse drug event classification from health records. Three different supervised machine learning algorithms: decision tree, random forest and LibSVM were compared and found that LibSVM outperformed the other two

algorithms. MertTiftikci[8] et al. proposed hybrid deep learning and dictionary-based approaches for extracting Adverse Drug Reactions. They employed two different methods for detecting mentions of type ADR, drug class, animal, severity, factor, and negations from drug labels. The neural network-based approach outperformed the dictionary- and rule-based approach in extracting ADRs.

MATERIALS AND METHODS

Materials

Electronic Health Records

Electronic Health Record (EHR) stores the detailed information of a patient in digital format. It includes the demographics, history of past illness, past medications, present problems, current medications, treatment plans, radiology images and laboratory results. EHR allows computerized access to information and has the potential to streamline the clinician's workflow. It helps in instant availability of information to authorized users in a secured manner. Electronic Health Record serves as rich sources of medical information which can be used for the extraction of disease comorbidities, prediction of various diseases and adverse drug reactions. In this work, electronic health records collected from nearby hospitals of Kochi has been used for finding adverse reactions of drugs.

Online Health Forums

With the advancement in internet technology, information sharing becomes effortless andrapid. People use to share their information over internet through various discussion forums. Nowadays, there are many online health discussion forums where patients and medical practitioners can share their knowledge and experiences.Online health forums serves as a source of health information, providing patients with a safe environment to share experiences, seek information, and improve their health knowledge. Careful analysis of information from these health forums can used for better treatment and patient care. Recently, many research works have been conducted for extracting information from online healthcare forums. Steadyhealth.com, medications.com etc. are some of the online healthcare forums which can be used for ADR prediction. Here, we have used messages from medications.com.

Methods

Machine Learning Approach

The main sources of data are electronic health records, collected from nearby hospitals in Kochi. Electronic health records were collected after undergoing an ethics committee approval. For collecting data from online healthcare forum, the site that we have referred is www.medications.com where people talk only about the reactions that they had on having a certain drug for a period of time.

The overall architecture is given in Fig.1. EHRs and online data are processed separately. For processing data from EHRs, we have used the approach done in our previous work[9-10]. In order to extract the required information from the online health forum, an HTML parser is used. Beautiful Soup is a python package used for parsing html and xml documents. It creates a parse tree for parsed pages that can be used to extract data from HTML tags, which is useful for web scraping. The data collected from the online healthcare forums are publicly available data. Personally identifiable information of the forum users were not collected or used in this study. After extracting raw data from the data sources, data pre-processing is done. Raw data contains a large amount of irrelevant data like stop words (is, was, the etc.), punctuations which needs to be removed. Here we used the natural language toolkit, NLTK to filter out the irrelevant data. NLTK is a suit of libraries and programs for symbolic and statistical natural language processing (NLP) for English and is written in the python programming language. After stop word removal, punctuations are removed. This scraped data is then stored in a text document format.

In the next phase, i.e. Named Entity Recognition phase, entities of interest are identified and extracted. Here the entities of interest are names of drugs and terms denoting side effects. The online healthcare forums are used by regular people without a medical background. Therefore the terms used may not be those used by experts in the medical field. In this regard, we constructed a custom list of drugs and side effects that consist of terms that are used by people who are not experts in the field of medicine.

The 2 main phases in machine learning approach are: Feature Extraction and Classification. Feature extraction module is used to identify the presence of relationships between the named entities in a given text.

Feature Extraction

Five types of features are extracted for performing classification.

1. N-grams: These are continuous sequences of n tokens taken together at a time from the tweets for analysis. We have chosen n as three in our experiment.
2. POS: Part of Speech (POS) of tokens indicates the syntactic function of each token.
3. Negation: Words mentioned in negative context indicating adverse events
4. Relative positions of drug names and terms denoting adverse events
5. TF-IDF weight

This weight is a statistical measure used in information retrieval to estimate the importance of a word in a given document. The importance increases proportionally to the number of times a word appears in the document but

is offset by the frequency of the word in the corpus. The TF-IDF weight is composed of two terms: the first computes the normalized Term Frequency (TF), the second term is the Inverse Document Frequency (IDF). The TF-IDF weight is calculated as the product of these 2 terms. In our experiment, while considering Electronic Health Records, each record (record of a single patient) is taken as a document and words in the document is considered as terms and the entire records available are taken as corpus. For the online health forum data, each tweet/message is treated as a document and the entire messages related to a particular medicine is treated as corpus.

Term frequency related to a term, t is given as:

tf(t) = (Number of times term t appears in a document) / (Total number of terms in the document)

Inverse Document frequency related to a term, t is given as:

IDF(t) = $\log_e$ (Total number of documents / Number of documents with term t in it)

Classification

After extracting features, classification is done using 4 types of classifiers: Naïve Bayes, Decision Tree, KNN and SVM. The overall architecture is shown in Fig.1.

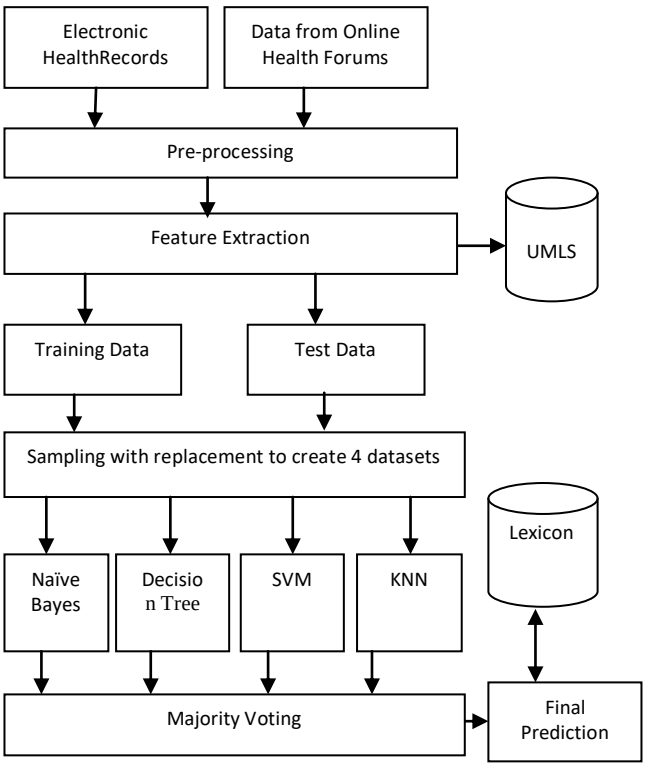

Figure 1: Overall Architecture

Naïve Bayes classification is based on Bayes theorem. It uses strong (naive) independence assumptions between the features, which presume that an attribute value on a given class is independent of the values of other attributes. Naïve Bayes classifier predicts the probability of outcome of an event as:

$$P\left(\frac{H}{X}\right) = \frac{P\left(\frac{X}{H}\right) * P(H)}{P(X)}$$

where P(H) is the probability that the hypothesis H holds for the observed data tuple X, P(X) is the prior probability of training data X, P(H/X) is the probability of H given X and P(X/H) is the probability of X given H.

K-nearest neighbour's algorithm (k-NN) is a non-parametric method used for classification. Input consists of *k* closest training examples in the feature space. The output is a class membership. An object is classified by a majority vote of its neighbours, with the object being assigned to the class most common among its *k* nearest neighbours (*k* is a positive integer, typically small). If *k* = 1, then the object is simply assigned to the class of that single nearest neighbour.

Decision Trees (DTs) are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. A decision tree is a graph that uses a branching method to illustrate every possible outcome of a decision. Each node of the tree denotes an attribute.

Support Vector Machine is a supervised classification learning algorithm. It is based on the concept of decision planes that define decision boundaries. A decision plane is one that separates between a set of objects having different class memberships. SVM classifier finds the best hyperplane for separating data points that belong to a particular class from the data points of other classes.

An ensemble of these classifiers was used for enhancing the accuracy of classification.The ensemble algorithm is given in Fig.2:

Input: Dataset divided into training set (Tr) and test data (Ts)
Step1: Set j=1
Step2: Set a counter for each of the classifier denoted as c_i initialized to 0.
Step3: Four random samples are created from Tr with replacement
Step4: Each sample dataset of Tr is fed as input to the classifier
Step5: Test the classifier with input Ts
Step6:If output of classifier$_i$ is positive increment c_i
Step7: Shuffle the sample dataset of Tr among each classifier and increment j
Step8: Repeat steps 2 to 8 till j=4
Step9: Set vote=0
Step10: If c_i>2,then set output of classifier$_i$ as positive otherwise negative
Step11: Increment the vote for each positive prediction
Step12: If vote>2 then the resultant class is positive
Elseif vote<2 then the resultant class is negative else choose the output with the highest accuracy

Figure 2: Ensemble Algorithm

Deep Learning Approach

Deep Learning is a machine leaning approach based on artificial neural network. A deep learning network consists of an input layer, multiple hidden layers and an output layer. Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), Deep belief networks and Generative adversarial networks are some of the popular models of Deep Learning networks. RNN is widely used for text analysis and hence we have chosen LSTM, an improved version of RNN for ADR prediction. RNNs make use of sequential information for text processing. For each element in a sequence, RNNs repeat the same task and hence they are called recurrent networks. The output of the network depends on previous computations. RNNs have a memory to store information about previous computations. The architecture for ADR prediction using LSTM is given in Fig.3.

RESULTS AND DISCUSSION

The performance of machine learning algorithms has been compared. Table 1 shows the performance of various algorithms. SVM is having higher accuracy compared to Decision tree, Naive Bayes and KNN classifiers. Using ensemble method has considerably increased the performance. Fig.3. shows the graphical comparison of the algorithms. LSTM showed an accuracy of 86%.

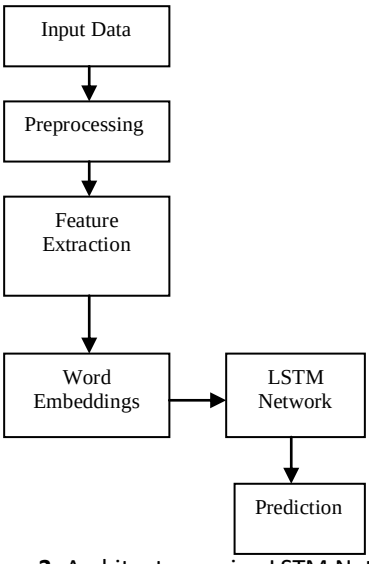

Figure 3: Architecture using LSTM Network

Table 1: Performance of Machine Learning Algorithms

Parameter	Naïve Bayes	Decision Tree	KNN	SVM	Ensemble Method
Accuracy	0.822	0.827	0.81	0.84	0.87
Precision	0.75	0.71	0.68	0.77	0.81
Recall	0.48	0.50	0.45	0.54	0.59
F-Score	0.59	0.58	0.54	0.64	0.68

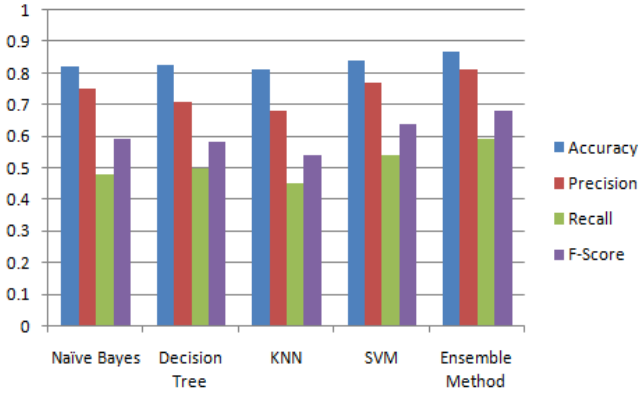

Figure 4: Comparison of algorithms

CONCLUSION

In this work, we have explored the different possibilities for mining adverse drug reactions. Diverse data sets including online health forums and electronic health records are considered for prediction tasks. Data sets were divided into training set and test set. Machine learning techniques, SVM, Naive Bayes, Decision tree and KNN methods are used for classification. An ensemble of these classifiers was also developed which showed an increased performance compared to individual classifier. Latest technique of deep learning is also used for ADR prediction using LSTM network. LSTM performed well compared to other classifiers, but the accuracy was less compared to ensemble approach. As future work, we could develop a dictionary with the predictions which can be further analysed for finding disease comorbidities.

REFERENCES

1. Ioannis Korkontzelos, Azadeh Nikfarjam, Matthew Shardlow, Abeed Sarker, Sophia Ananiadou, Graciela H. Gonzalez “Analysis of the effect of sentiment analysis on extracting adverse drug reactions from tweets and forum posts” Journal of Biomedical Informatics, 62, 2016, 148–158.

2. Hariprasad Sampathkumar, Xue-wen Chen and Bo Luo, “Mining Adverse Drug Reactions from onlinehealthcare forums using Hidden Markov Model”, *BMC Medical Informatics and DecisionMaking,* 14, 2014, 91.

3. A.Pushpa, S.Kamakshi, “Enhancing the extraction of adverse drug reactions using deep learningtechnique” International Journal of Pure and Applied Mathematics, Volume 118 No. 20, 2018, 495-504.

4. Andy W. Chen “Predicting adverse drug reaction outcomes with machine learning” International Journal of Community Medicine and Public Health, 5(3), 2018 Mar, 901-904.

5. Wencheng Sun, ZhipingCai, Yangyang Li, Fang Liu, Shengqun Fang and Guoyan Wang “Data Processing and Text Mining Technologies on Electronic Medical Records: A Review” Hindawi Journal of Healthcare

Engineering, Volume 2018, Article ID 4302425, 9 pages.

6. Trung Huynh, Yulan He, Alistair Willis and Stefan Ruger "Adverse Drug Reaction Classification With Deep Neural Networks" Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, Osaka, Japan, December 11-17, 2016, pages 877–887.
7. Stefanie Friedrich and Hercules Dalianis 'Adverse drug event classification of health records using dictionary-based pre-processing and machine learning" Proceedings of the Sixth International Workshop on Health Text Mining and Information Analysis (Louhi), pages 121–130, Lisbon, Portugal, 17 September 2015. Association for Computational Linguistics.
8. MertTiftikci, ArzucanÖzgür, Yongqun He, and JungukHur "Extracting Adverse Drug Reactions using DeepLearning and Dictionary Based Approaches" TAC, 2017.
9. Lakshmi K.S, G.Vadivu, "Extracting Association Rules from Medical Health Records Using Multi-Criteria Decision Analysis", Procedia Computer Science, Volume 115, 2017, Pages 290-295.
10. Lakshmi K.S, G. Santhosh Kumar, "Association rule extraction from medical transcripts of diabetic patients", Proceedings of the fifth International Conference on the Applications of Digital Information and Web Technologies (ICADIWT 2014).
11. Mingyong Liuand Jiangang Yang, "An improvement of TFIDF weighting in text categorization", International Conference on Computer Technology and Science (ICCTS 2012) IPCSIT vol. 47, (2012).
12. Hyon Hee Kimand Kiyon Rhew "Analysis of Adverse Drug Reaction Reports using Text Mining" Korean J Clin Pharm, Vol. 27, No. 4, 2017.

Source of Support: Nil, **Conflict of Interest:** None.