International Journal of Computer Engineering & Technology (IJCET)
Volume 10, Issue 2, March–April, 2019, pp. 83-90. Article ID: IJCET_10_02_010
Available online at
http://iaeme.com/Home/issue/IJCET?Volume=10&Issue=2
Journal Impact Factor (2019): 10.5167 (Calculated by GISI) www.jifactor.com
ISSN Print: 0976-6367 and ISSN Online: 0976–6375

PREDICTING SAFETY INFORMATION OF DRUGS USING DATA MINING TECHNIQUE

Dr. V. Umarani*
Department of Computer Science,
Sri Ramakrishna College of Arts and Science for Women, Coimbatore, India

C. Rathika
Department of Computer Science,
Sri Ramakrishna College of Arts and Science for Women, Coimbatore
*Corresponding Author

ABSTRACT

Data Classification is the application of data mining techniques to discover patterns from the micro array and biological datasets. This research entitled "PREDICTING SAFTEY INFORMATION OF DRUGS USING DATA MINING TECHNIQUE" incorporates information theory, which is the process of deriving the information from the unsupervised dataset through feature selection. Finding the best features that are similar to a test data is challenging task in current data era. This research presents a framework for discovering best feature selection from unsupervised datasets. The proposed research work presents a new approach to measure the features (attributes) in drug prediction dataset using the methodologies namely, data cleaning, Adaptive Relevance Feature Discovery and Random Forest Classification. There are number of pharmacy companies are available in the market with multiple medicines for same problem.This prediction of drugs is used to prescribe the medicines for the patient's disease by analyzing the history of the patient's health. Feature selection and dimensionality reduction is characterized by a regularity analysis where the feature values correspond to the number times that term appears in the dataset. The relevance feature discovery method gives a useful measure is used to find the similarity features between data points are likely to be in terms of their features property. Some of the challenges faced in finding the best feature selection include positive, negative and inconsistency. This Proposed work proposes an enhanced Drug prediction based on Random Forest classification method to estimate the feature searching that is measured using minimal redundancy optimization method corresponding to drug prediction dataset.

Keywords: Random forest, data mining, healthcare, association rules, feature selection.

Cite this Article: Dr. V.Umarani and C. Rathika, Predicting Safety Information of Drugs using Data Mining Technique, *International Journal of Computer Engineering and Technology,* 10(2), 2019, pp. 83-90.
http://iaeme.com/Home/issue/IJCET?Volume=10&Issue=2

1. INTRODUCTION

Due to the rapid growth of digital data made available in recent years, knowledge discovery and data mining have attracted a great deal of attention with an imminent need for turning such data into useful information and knowledge. Many applications, such as market analysis and business management, can benefit by the use of the information and knowledge extracted from a large amount of data. Knowledge discovery can be viewed as the process of nontrivial extraction of information from large databases, information that is implicitly presented in the data, previously unknown and potentially useful for users.[1]

Worsening of a drug's risk-to-benefit ratio can lead regulatory authorities to take actions aiming at mitigating or withdrawing the use of the drug. A very first step in such actions is the issue of a safety warning that has modest impact on the usage of a drug. One important way by which regulatory authorities may act on safety warning issuance is through prediction of side effects.[2] Prediction of drug's side effects can help the Doctor and Pharmacist to prescribe the patients with exact medicine to the complaints. Therefore, predicting the side effects of drugs based on quantity of drugs can have a strong impact on the way of drugs used by physicians and patients.[3]

Based on indication, the side effects of drug is predicted to deliver the exact reaction from different information. This work highly aids the Physicians and Pharmacist to prescribe the drugs according to patient's complaint. [4]

In section 2, the existing work pertaining to the study is discussed. Section 3 discusses about the phases involved in the proposed work and deals with the results. Section 4 concludes the work and presents the future directions of this research work.

2. EXISTING WORK

Drug is any substance that when taken into a living organism may modify one or more of its functions. Drugs can provide temporary relief from unhealthy symptoms and/or permanently supply the body with necessary substances the body can no longer make. Some drugs lead to an unhealthy dependency that has both physiological and behavioural roots [5]. Drug addiction can cause serious, long-term consequences, including problems with physical and mental health, relationships, employment, and the law [6]. Decision tree uses the simple divide-and conquer algorithm. In these tree structures, leaves represent classes and branches signify conjunctions of features that lead to those classes.

2.1. DRAWBACKS

- For data including categorical variables with different number of levels, information gain in decision trees is biased in favor of those attributes with more levels.
- Calculations can get very complex, particularly if many values are uncertain and/or if many outcomes are linked.
- Computing probabilities of different possible branches, determining the best split of each node, and selecting optimal combining weights to prune algorithms contained in the decision tree are complicated tasks that require much expertise and experience
- It is difficult for the doctor to remember and to prescribe the number of medicines from different pharmaceutical companies.

- Time consuming.
- Needs manual calculations.

2.2. PROPOSED WORK

Data classification problems often have a large number of attributes, but not all of them are useful for classification. Irrelevant and redundant features may even reduce the classification accuracy. Drug prediction feature selection is a process of selecting a subset of relevant features, which can decrease the dimensionality, shorten the running time, and/or improve the classification accuracy. Feature selection (FS) refers to the problem of selecting those input attributes that are most predictive of a given outcome; a problem encountered in many areas such as machine learning, pattern recognition and signal processing [7]. The proposed work attempts to use the uncertain information to improve the performance of drug prediction based on the real time information from FAERS. The utility of the approaches is demonstrated and compared empirically with several other dimensionality reduction techniques. [8]

2.2.1. FEATURES

- It is one of the most accurate learning algorithms available. For many data sets, it produces a highly accurate classifier.
- It runs efficiently on large databases.
- The accuracy of side effects of drugs based on indication number
- It can find the most accurate detection using this technique.
- A proper database is maintained for the drugs in the pharmacy to prescribe the OTC drugs.
- To reduce the tedious work of a manual systems.
- Helpful to Pharmacist and Physicians

The figure 2.1 shows a clear picture of the proposed work.

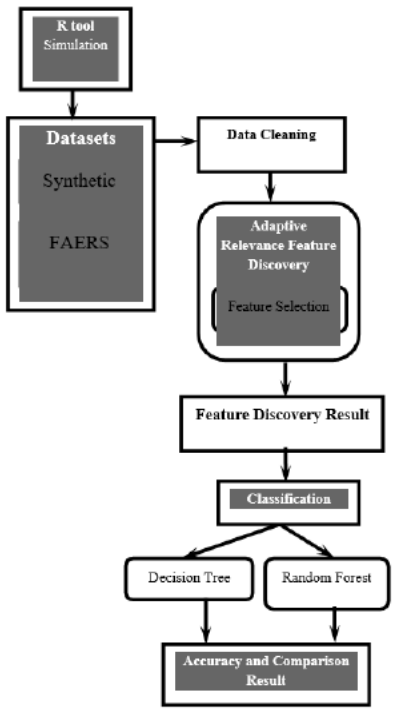

Fig 2.1

3. METHODLOGY ANALYSIS

The proposed architecture accepts the data classification parameters as input which contains the R tool simulation where the novel prediction drug label changes in random forest classification algorithm is applied to the FDA Adverse Event Reporting System (FAERS) dataset. The users initialize the dataset instances, attributes and classes as initial parameters in which the classification process is to be evaluated.

The following methodology is listed below,

- Data Cleaning
- Adaptive Relevance Feature Discovery
- Classification and Regression Training

3.1. DATA CLEANING

Data cleaning method is kind of pre-processing technique it plays a very important role in data classification techniques and applications. It is the first step in the adaptive relevance feature discovery mining process. In this data cleaning process there are three key steps of procedures namely, Training set extraction, Feature Attribute selection and Filtering methods.

The data pre-processing of untrained raw dataset is first partitioned into three groups: (1) a predetermined set of instance initiation, (2) the group of attributes (features, variables) and (3) the class of attribute. For each groups in the dataset, a reduction decision classification is constructed. For each reduction system is consequently divided into two parts: the training dataset and the testing dataset. Each training dataset uses the corresponding input features and fall into two classes: normal (+1) and abnormal (−1).

The training set feature set process to compute the cross validation classification error for a large number of features and find a relatively stable range of small error. The Training feature selection is to select n (a preset large number) sequential features from the input X.

The Feature attribute selection is a statistical technique that can reduce the dimensionality of data as a by-product of transforming the original attribute space. Transformed attributes are formed by first computing the covariance matrix of the original features, and then extracting its sorting. The attribute selection defines a linear transformation from the original attribute space to a new space in which attributes are uncorrelated.

The Filtering approach has much lower complexity than wrappers; the features thus selected often yield comparable classification errors for different classifiers, because such features often form intrinsic clusters in the respective subspace.

3.2. ADAPTIVE RELEVANCE ROUGHSET FEATURE DISCOVERY

The adaptive relevance feature discovery process considers the mutual-information-based feature selection for both supervised and unsupervised data. For discrete feature variables, the integral operation in (1) reduces to summation. In this case, computing mutual information is straightforward, because both joint and marginal probability tables can be estimated by tallying the samples of categorical variables in the data.

3.3. RANDOM FOREST CLASSIFICATION

Random forests are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. The generalization error for forests converges (i.e.,). to a limit as the number of trees in the forest becomes large. The generalization error of a forest of tree classifiers depends on the strength of the individual trees in the forest and the correlation between them.

Random forest classifier creates a set of decision trees from randomly selected subset of training set. It then aggregates the votes from different decision trees to decide the final class of the test object. There are two stages in Random Forest algorithm, one is random forest creation, and other is to make a prediction from the random forest classifier created in the first stage.[9]

3.4. CLASSIFICATION AND REGRESSION TRAINING

Classification And Regression Training which is possibly the biggest project in R. This package alone is all you need to know for solve almost any supervised machine learning problem. It provides a uniform interface to several machine learning algorithms and standardizes various other tasks such as

- Data splitting
- Pre-processing
- Feature selection
- Variable importance estimation

In this work, data splitting is carried down in the FEARS dataset and Feature selection is done to extract only a specific type of drugs.

3.5. DATASET

A database is an organized mechanism that has the capability of storing information through which a user can retrieve stored information in an effective and efficient manner. The data is the purpose of any database and must be protected.

The database design is a two level process. In the first step, user requirements are gathered together and a database is designed which will meet these requirements as clearly as possible.

The dataset used for this project is a real data taken from the FDA Adverse Event Reporting System (FAERS).The data has been gathered from U.S. Food Drug Administration (FDA)[10]

This data set consists of ASCII data files. These ASCII data files are '$' delimited; that is, a '$' is used to separate the data fields. These files can be imported into spreadsheet programs such as earlier versions of MS Excel. This excel sheet is converted into CSV format and implemented in R tool for classification of drugs according to the indication level which helps in predicting the side effects of the drugs.[11][12]

The following figures 3.1, 3.2, 3.3shows the execution of the work in R environment.

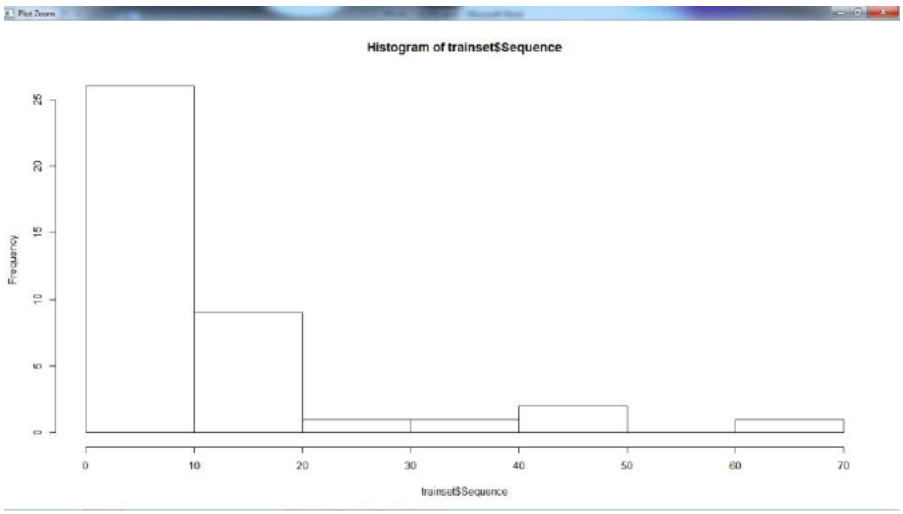

Fig 3.1

The above Fig represents the sequence of the drug that is consumed and the frequency of the consumption.

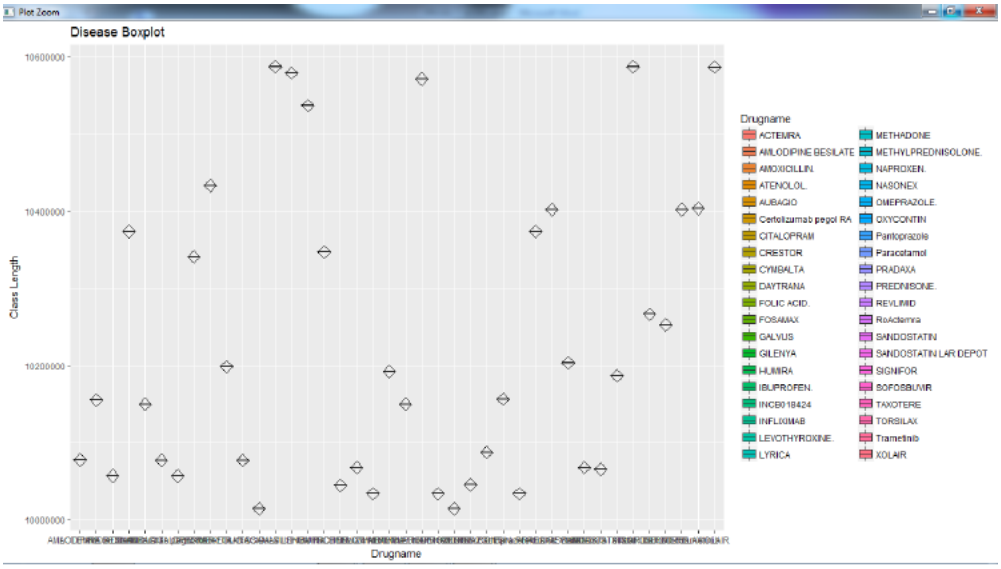

Fig 3.2

Fig 3.2 represents the drugs consumed for the respective diseases.

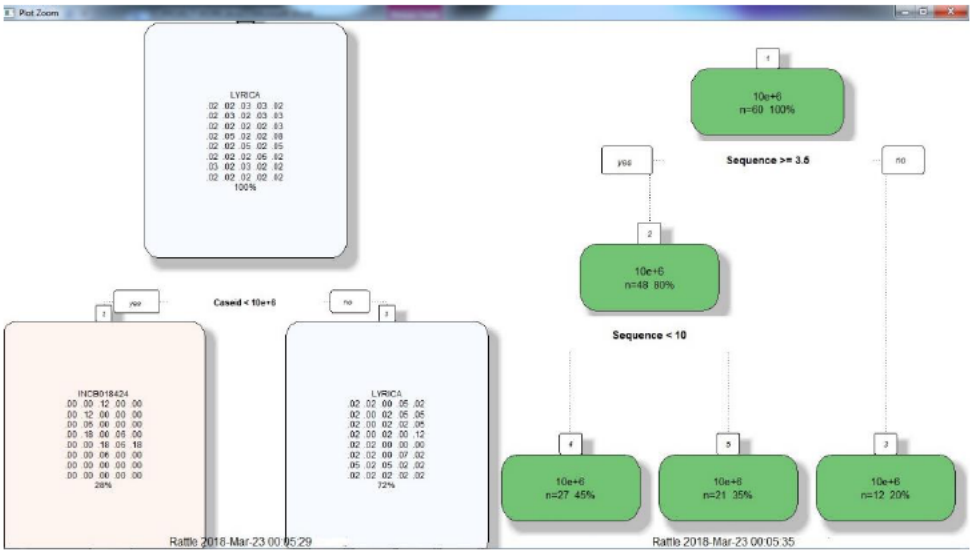

Fig 3.3

Fig 3.3 represents the side effects caused by the consumption of a particular drug with respect to the frequency of the drug consumed.

4. CONCLUSION AND FUTURE ENHANCEMENTS

The proposed work entitled “PREDICTING SAFTEY INFORMATION OF DRUGS USING DATA MINING TECHNIQUE” is an R tool based application. This provides facility for guarantees to predict drug labels with sequence, drug name and reactions is able to detect drug predictions by considering a data mining techniques. The proposed system presented an application of drug prediction using random forest classification using adaptive relevance feature selection. The different steps in predictions are represented as the underlying data mining process of a classification.

4.1. SCOPE OF FURTHER ENHANCEMENT

This work shows a new way of predicting the side effects of drugs based on real data registered by the common people. Our future work will focus on the following:

- For educating the pharmacist and public to lead better medical practices.
- Also this data can be integrated with other types of datasets such as the patient's history in hospital database health record.
- By integrating data from multiple available information sources, more effective prediction may be achieved.

ACKNOWLEDGEMENT

The authors would like to thank the Management for providing the seed funding assistance to carry out this research work.

REFERENCES

[1] Andy W. Chen, Predicting adverse drug reaction outcomes with machine learning, International Journal of Community Medicine and Public Health, Chen AW. Int J Community Med Public Health. 2018 Mar;5(3):901-904

[2] Rajeev Kumar, Shivani Singh, Sudhir Arora, Savita Bhati,, Adverse Drug Reactions: A Comprehensive Review, Journal of Drug Delivery & Therapeutics. 2018; 8(1):103-107

[3] Daniel M. Bean, Honghan Wu, Ehtesham Iqbal, Olubanke Dzahini, Zina M. Ibrahim, Matthew Broadbent, Robert Stewart & Richard J. B. Dobson, Knowledge graph prediction of unknown adverse drug reactions and validation in electronic health records,www.nature.com/scientific reports 7: 16416 | DOI:10.1038/s41598-017-16674-x,2017

[4] Gurulingappa H, Mateen-Rajput A, Toldo L. Extraction of Adverse Drug Effects from Medical Case Reports. J Biomed Semantics 2012; 3: 15.

[5] Gurulingappa H, Rajput AM, Roberts A, et al., Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. J Biomed Inform 2012; 45(5): 885–892.

[6] Tu-Bao Ho1,2*, Ly Le3, Dang Tran Thai1 and Siriwon Taewijit1,4, Data-driven Approach to Detect and Predict Adverse Drug Reactions, Current Pharmaceutical Design, , 22, 000-000,2016

[7] Kalpana Raja, Matthew Patrick, James T. Elder & Lam C. Tsoi, Machine learning workflow to enhance predictions of Adverse Drug Reactions (ADRs) through drug-gene interactions: application to drugs for cutaneous diseases,www.nature.com/Scientific Reports | 7: 3690 | DOI:10.1038/s41598-017-03914-3

[8] Kuhn M, Campillos M, Letunic I, et al. A side effect resource to capture phenotypic effects of drugs. Mol Syst Biol 2010.

[9] S.Preethi, C.Rathika, An Efficient First Order Logical Casual Decision Tree in High Dimensional Dataset, International Journal of Computer Sciences and Engineering, Volume-6 , Issue-2 , Page no. 73-78, Feb-2018

[10] DuMouchelW, Bayesian DataMining in Large Frequency Tables, with an Application to the FDA Spontaneous Reporting System. Am Stat 1999; 53(3): 177–190.

[11] Ahmed I, Thiessard F, Miremont-Salamé G, et al. Pharmacovigilance Data Mining With Methods Based on False Discovery Rates: A Comparative Simulation Study. Clin Pharmacol Ther 2010; 88(4): 492–498.

[12] Ahmed I, Poncet A. PhViD: a R package for PharmacoVigilance signal Detection. 2011.

AUTHOR'S PROFILE

Dr. V. Umarani received her Ph.D Degree in Science and Humanities (Computer Applications) from Anna University, Chennai in 2013. She is currently working as Associate Professor in Sri Ramakrishna College of Arts and Science for Women, Coimbatore. She has 19 years of experience in teaching. Her research interests are Data mining, Big data analytics and machine learning. She has published twenty research papers in Journals and presented many research papers in National and International Conferences in India and abroad. She is currently guiding 2 Ph.D scholars. She has guided 12 M.Phil scholars, She is a life member of International Association of Engineers.(IAENG).

Mrs. C. Rathika is currently working as an Assistant Professor in the Department of Computer Science at Sri Ramakrishna College of Arts and Science for Women. in year 2009. She is currently pursuing her Doctorate degree in computer science at Sri Ramakrishna College of Arts and Science for Women. She has 17 years of experience in teaching. Her research areas include Data mining and Big Data analytics. She has guided 6 M.Phil scholars. She acted as an Adjunct Faculty for a period of one month in Texila American University, South America to develop the curriculum for the course Diploma in Information Technology. She published 8 papers in International Journals and presented papers in international and national Conferences.