Adverse Drug Reaction Detection Using Data Mining Approaches: A Survey

Naresh Poloju¹  Purushotham Muniganti²

¹Assistant Professor, ²Associate Professor ¹²Department of Computer Science and Engineering
Sreyas Institute of Engineering and Technology
Beside Indu Aranya, GSI, Bandlaguda, Nagole, Hyderabad, Telangana

Abstract— Adverse Drug Reaction (ADR) is a major problem faced by medical practitioners all over the world with respect to drug safety. Adverse drug reactions (ADRs) are the harmful reactions of the drugs caused to humans due to overdose or chemical reactions between two or more chemicals in the medicines, etc. Discovering unknown ADRs as early as possible is highly desirable because they affect large number of people and can help in raising early warning against adverse effects of drugs and help medical experts in making treatment effective and timely. In today’s digital era a huge amount of data correlated to adverse effects of drugs is being collected at hospitals, drug retail stores and by drug producers. We use technology to supervise the complex data produced by the health care system. Data mining is one of the technologies used to find the interesting knowledge from the vast data produced by the health care system and used to analyze the patterns in large sets of data. This data can be used for finding out the secreted relationships between drugs and their adverse reactions. This paper provides a survey on how data mining techniques can be utilized in detection of adverse effects of drugs which is useful for new entrants as well as researcher in this filed.

Keywords— Adverse Drug Reaction, Data Mining, Adverse Effects, Knowledge Discovery

I. INTRODUCTION

Adverse Drug Reaction (ADR) is defined as “a response to a drug that is noxious and unintended and occurs at doses normally used in man for the prophylaxis, diagnosis or therapy of disease, or for modification of physiological function” [51]. ADRs cause Death, Hospitalization, Life threatening, Disability - significant, persistent, or permanent change, impairment, damage or disruption in the patient’s body structure/ function, physical activities or quality of life. They require intervention to prevent permanent impairment or damage. Surveys are showing that 5%-20% of total hospital admissions are due to Adverse Drug Reactions [39]. Because of the ADRs 1.8% of the deaths happened in India [1]. ADRs cause increase of morbidity, mortality and health care costs worldwide. In United States 6.7% of total admissions are due to ADRs in 1998, out of which fatalities were 0.32% [33]. In United Kingdom(UK) ADRs resulted in approximately 250,000 admissions a year which costs 870 million dollars to the United Kingdom's National Health Service [13][39]. The WHO Classified the severity of the adverse reaction as Incidental, Mild, Moderate, Severe [57]. The Adverse Drug Reactions (ADR) are classified into Dose related and non-dose related. Non-Dose related ADRs are further classified as Type A (Augmented), and Type B (Bizarre). The Dose related ADRs are categorized as reactions related to both dose and time (Type C), and delayed reactions (Type D). Always it is not possible to classify an ADR into one of these categories, as more becomes known about the specific adverse drug effects, this classification will be revised and classification of currently unclassifiable reactions will become easier [41].

If we are able to predict ADRs then the fatal causes to the human beings can be avoided. To do this the information about ADRs needs to be digitalized. In this “Data Age” almost all of the hospitals are maintaining the digital information about the patient treatment. This information needs to be converted into the knowledge for the prediction of ADRs. Data mining is the process of discovering interesting patterns and knowledge from large amounts of data. Data mining algorithms are very effective for retrieving the useful information’s from the very huge databases stored in the repositories. The data sources can include databases, data warehouses, the Web, other information repositories, or data that are streamed into the system dynamically. A robust data mining ADRs prediction model can identifies patients who are trending toward a high-risk condition. This information gives a head start in identifying high-risk patients so that steps can be taken to improve the patient’s health and to prevent health problems in the future [37].
II. RELATED WORK FOR ADR DETECTION

One of the first works to detect ADEs used traditional techniques like medical information processing, patient surveys and laboratory results are available in [46][10][36].

A Bayesian confidence propagation neural network (BCPNN) which is used to find drug-ADR combinations with other variables, which are highly associated, compared to the generality of the stored data, or a section of the stored data was described in [3]. The BCPNN used for early signal detection of ADR and to avoid false positives where a common drug and ADRs occur in the database. The BCPNN suspected drug-ADR combinations reached the 97.5% confidence level of difference from the generality.

In [48] the authors used Multi-Item Gamma Poisson Shrinker (MGPS) data mining method. In addition to detecting possible serious single-drug adverse event problems, MGPS is currently being evaluated to detect the effect of multiple drug interactions, adverse events (syndromes), and to detect differences among subgroups defined by gender and by age, such as pediatrics and geriatrics. From the 2.4 millions reports, the algorithm detected adverse event signals with <95% confidence.

The description of manual and automated adverse event detection systems was available at [12]. The manual system requires the intervention of the physician where the automated systems can automatically detect adverse events from the available data. Finally the authors stated that there is a lot research needs to be done in the area of automated detection systems to get the accurate results.

In [26][29][28] Based on the domain-driven knowledge representation, Unexpected Temporal Association Rules(UTAR), the authors have introduced two interestingness measures, unexlev and rankRatio, in the context of signaling unexpected and infrequent patterns characteristic of ADRs from administrative health databases. They have presented two simple but effective mining algorithms MUTARA and HUNT to identify pair-wise UTARs from the linked administrative health databases QLDS (Queensland Hospital morbidity data, commonly referred as the Queensland Linked Data Set)[50]. A series of experiments on the real-world data substantiate that both MUTARA and HUNT can shortlist some ADR signals. HUNT has shortlisted statistically significantly more suspected ADR signals than MUTARA. The authors have only concentrated on highlighting pairwise UTARs in [26][29][28]. The presented mining techniques require further accuracy performance examination and refinement on more databases when they become available. It includes setting domain-specific parameters, integrating with ADR validation techniques together, and taking into account of spontaneous ADR case reports and adverse severity levels. Extending the UTAR concept and mining techniques to other applications is another interesting research direction.

An automated ADE monitoring system to improve the detection and characterization of adverse drug events (ADEs) in hospital patients was developed in [5]. Over 18 months they monitored 36,653 hospitalized patients. There were 731 verified ADEs identified in 648 patients. Out of which 92 are voluntarily reported by Physicians, pharmacists, and nurses. The other 631 ADEs were detected from automated signals.

In [30] the authors stated that the Classification and association rule mining algorithms are producing more number of risk patterns so they are not suitable for medical data mining. They implemented optimal risk pattern mining algorithm by using anti-monotone property. This algorithm produced less number of risk patterns when compared with association rule mining algorithms. From this resulting optimal risk patterns the authors selected the top risk pattern, they applied this method on medical data to detect adverse drug events.

User posts in online health forums can be extracted for the detection of adverse effects. This concept was experimented in [34]. The posts were extracted and drug names are mapped with adverse effect. The extracted drug adverse effects are compared with the adverse effects of the drugs available in FDA drug database. The results have given correlation between the extracted and predefined adverse effect. The main goal of the research is to show that, the adverse effects can be extracted from online forums. This extracted information can be used for post market surveillance of the drugs. R Leaman et al.,[34] stated that the work further can be extended by considering the comments associated with the user posts.

In [54] the authors have introduced a knowledge representation called Potential Causal Association Rules (PCAR) to represent the potential causal relationship between a drug and ICD-9 (International Classification of Diseases, Ninth Revision)[17], and developed a new interestingness measure, potential causal leverage, to quantify the degree of possible causality of a PCAR. Based on this novel measure and an experience-based fuzzy RPD model [31], they have developed a data mining algorithm to search an electronic health database for potential ADR signals. They have taken example drug as enalapril[19].
Experimental results shown that, still improvements are required to get more accurate results.

An extensible and re-usable environment for conducting Adverse Drug Reaction (ADR) studies based on mining large-scale primary care databases was designed by Curcin V et al., [6].

The method described in [27] collected seven years data of nine drugs from eight different parenting Web sites. The authors depended only on free-text comments in each post. User's comment contains one or more drugs, side effects and possibly some personal information (like age, gender, etc...) in addition to narrative and text data. These reports were extracted and stored into a database and applied disproportionality analysis methods for drug safety measure: reporting ratio (RR) [8], proportional reporting ratios (PRR) [22], reporting odds ratios (ROB) [43] and information component (IC) [25]. To generate affiliation measure between a given drug and side effect in the database, they generated statistical scores. If the score is higher, then they have strong statistical association. The ultimate goal of the authors is to prove that “Adverse drug reactions can be identified by harnessing the parenting websites”.

The authors proposed an approach in [56] to detect the ADRs using feature matrix and feature selection. The experiments are carried out on the drug Pioglitazone[15]. Major side effects for the drug are detected and better performance is achieved compared to other computerized methods.

The ADR detection systems are suffering from the underreporting of adverse events. To provide assistance for automatic ADR detection systems, user posts from the online medical help forums can be extracted and ADR’s can be detected as stated in [34]. In addition to the work done in [34], the approach of [44] used Hidden Markov Model [40] to extract adverse side-effects of drugs. The results of [44] given an F-Score of 0.86, which is higher than the F-score 0.73 of manual method for the extraction of adverse side-effect relationship with the drug.

Harnessing online healthcare communities to mine DDIs (Drug-Drug Interactions) was proposed in [11]. The Association mining was used in the study to detect DDI signals. They have collected thread contents of thirteen drugs from MedHelp using Consumer Health Vocabulary (CHV), selected five reported DDI associations from DrugBank, and computed the values of leverage[14] and lift[21] to identify DDIs. They also proposed another measure called interaction ratio to capture drug safety signals. The results were shown that the DDIs reported by DrugBank are effectively detected.

The patient data was retrieved from the datasets and then from the patient’s history, the formation of drug table and symptom table taken place. A new interestingness measure called as exclusive causal leverage can be used based on fuzzy Recognition Primed Decision model (RPD) [32] which was described in [7]. On the basis of this measure the relationship between the drug and associated drug reactions can be mined and the support count is calculated. Then the causal leverage is evaluated and mapped with the causality categories according to which the Adverse Drug Reactions are detected. The authors worked on ADRs specific to type-2 diabetes.

The traditional spontaneous reporting systems are suffering from under reporting of adverse drug reactions. To overcome this drawback observational medical outcomes partnership (OMOP) dataset was used. This dataset consists of predefined adverse effect. Among the applied Proportional Reporting Ratio(PRR), Chi-square, Bayesian confidence propagation neural network (BCPNN), Likelihood Ratio(LR), Bayesian network(BN) and Ensemble algorithms, experiments shown that the Likelihood method which was exposed in [35] stood third place .

Feature based similarity model is used to measure the association strength between a drug and adverse reaction. For the prediction of ADR’s for particular patients, the patient’s data will be taken and compared with the features of already known ADR cases and it will display the top possible predictions. This feature based model briefed in [52] will help the doctors in early warning and decision making of adverse reactions.

Food and Drug Administration (FDA) ADRs dataset was taken and implemented multi-relational association mining algorithm to find the associations between the adverse events. Unlike the traditional way of taking the entire dataset as a single relation, the information was stored in multiple relations and then applied association mining algorithm which was described in [53]. To do this mining Proportional Reporting Ratio (PRR) [22] used to assess the strength or interestingness of an association. The results are shown that the algorithm is scalable in terms of both the number of the association rules that are evaluated and the size of the data.

Association rule mining for extracting ADR’s from electronic healthcare data was used in [42]. To find the rules that truly corresponds to ADR’s; the authors proposed a method for learning association rules between drugs and adverse reactions and then refined these associations by removing those rules that are caused by confounding. Using minimum left support and confidence measures, strength of association rules is measured.
Proportional Reporting Ratio (PRR) [22] method, in combination with an estimator of the precision of point estimate such as the Chi-square test [24] used to mine the different associations between drugs and adverse reactions which were exposed in [2]. Amiya Kumar T et al., proposed a system [2] for the detection of Adverse Drug Reactions (ADRs) allowing an interactive discovery of associations between drugs and symptoms, called a drug-ADR association. By the prediction of possible ADRs for a particular patient this system is providing a chance for the doctors to change their prescription. For this analysis the authors used 5000 records of ADR cases, but the results are not 100% accurate, further works are needed to provide better accuracy.

A novel method for successful detection of ADRs by introducing feature matrix and feature selection to detect the significant changes after patients consumes the drugs was proposed in [55]. A feature matrix, which characterizes the medical events before or after patients consumes the drugs, is created from THIN (The Health Improvement Network) database. Feature matrix transfers the unstructured medical data collected from daily basis into feature matrix of standard saving format, and is a foundation to perform feature selection methods. Feature selection methods based on Student’s t-test and Wilcoxon rank-sum test are used to detect the significant features from high dimensional feature matrix. The significant ADRs, which are corresponding to significant features are detected. Experiments are performed on three drugs: Atorvastatin, Alendronate, and Metoclopramide. Compared to other computerized methods available at [26][29][28], proposed method of [55] achieved better performance.

In [45] the authors proposed a method to identify adverse drug effects from tweets by modeling it as a link classification problem in graphs. A drug symptom graph is built by extracting mentioned drug and symptoms from the tweet history of each user, where nodes represent drugs or symptoms and edges are labeled positive or negative, for desired or adverse drug effects respectively. A link classification model is then applied to identify negative edges i.e. adverse drug effects. Using 10-fold cross validation with sizer’s dataset [20] as ground truth, the authors tested the model on 864 users. This model was able to achieve an F-Score [18] of 0.77 compared to the best baseline model with an F-Score of 0.58.

A novel framework was proposed in [38] to detect ADRs from web forums. The authors of [38] have used Latent Dirichlet Allocation (LDA) [4] to construct a topic space over the consumer corpus, then they used Gibbs sampling inference algorithm [48] for Approximation, to retrieve the useful comments they used nearest neighbor search with KL divergence distance[16]. Then FWSs extraction Algorithm was used to find frequent-word-sequence which is having the possibility of ADR.

The method available at [47] collected the information of patient’s medical history from pharmacists, MR’s, Doctors, etc..., and converted the data into structured form, then applied Naive Bayes classifier algorithm for classification of Adverse Drug Reactions. After classification of the data he applied Association Rule Mining algorithms for prediction of appropriate rules to prevent the Adverse Drug Reactions.

According to the discussion available in [49], Clinical Data Mining (CDM) was applied to detect Psychiatric ADR’s. CDM involves Association Rules, Clustering models, Classification Algorithms, and Outlier detection. Viveka S et al., [49] have taken a dataset of 51,332 records of various Adverse Reactions from this 2442 records are mapped various psychiatric disorders. For this study they had taken the drug substance Clopidogrel [23]. By this research they found that the drugs used to treat HIV patients, chemotherapy, heart disease and severe chronic pain in people who cannot use or do not respond to standard pain relieving medications creates many adverse effects including depression and suicide behavior.

III. CONCLUSION:

In this paper we have presented a survey on ADR’s with applications and analysis. Key research aspects in the field were identified, particularly concerning with modeling, Data Analysis, and Data Mining Algorithms, and existing works were surveyed in this framework.

Finally, we identified Nearest Neighborhood, Distance-Based, Decision Trees, and Interference algorithms are the disciplines that are most relevant to ADR’s but not limited to, which are most likely to continue to provide valuable influences.

Also the extensive bibliography provided with this paper will be helpful to the researchers who are working in this field and as well as to the new entrants.

REFERENCES


[57] www.who.int/medicines/technical_briefing/tbs/04-Drug-Safety_final-08.ppt