INTERNATIONAL JOURNAL OF RESEARCH IN COMPUTER APPLICATIONS AND ROBOTICS
ISSN 2320-7345

A DATA MINING APPROACH FOR EFFICIENT EVENT ANALYSIS TO FIND ADVERSE DRUG REACTIONS

R. Sindhulakshmi[1], D. SaravanaPriya[2]
[1]Department of CSE, P.A College of Engineering and Technology, Coimbatore, Tamilnadu
(sindhusarchive@gmail.com)
[2] Department of IT, P.A College of Engineering and Technology, Coimbatore, Tamilnadu
(dspriyapacet@gmail.com)

Abstract

With the early detection of the unknown Adverse Drug Reaction (ADR) saves the lives of the human. The recognition-primed decision (RPD)Model, Interestingness measure and causal leverage plays an important role in data mining.RPD using fuzzy system is a decision making approach, how human takes decisions while handling complex task. Interestingness measures are often used in association rule discovery for mining and ranking patterns. Causal leverage assesses the strength by association of drug symptom. The exclusive causal-leverage is used to rank the potential causal associations between drug and recorded symptoms, which corresponded to a potential ADR.

Keywords - Data mining, adverse drug reactions, recognition-primed decision, interestingness measure, causal association.

I. INTRODUCTION

Medications have brought better health and longer life to the human race. Every day, hundreds of millions of people from all over the world are affected by the medicines. The term “adverse effect” is preferable to other term such as side effect [1]. Signal is the reported information regarding the causal relation between the adverse event and the drug. Postmarketing surveillance deals with the drug safety. Interestingness measure is called potential causal leverage. This causal leverage deals with the causal association between the drug and symptom. The drug symptom pair is evaluated by the computational fuzzy recognition primed decision model [2]. RPD model performs decision making process in order to handle the complex problems. The decisions are made based on the experience. Fuzzy interpretations are incorporated by the fuzzy logic. Fuzzy sets are employed by cue values. Cue is the key concept in the RPD model. The RPD model similarity measure is needed to assess the degree of likeness between the current situation and the past situation. Fuzzy reasoning is employed to abstract high level cues from elementary data. In FP Growth algorithm number of patterns discovered is large. Therefore there is need to rank the discovered patterns according to their degree of interestingness measures

A. Data mining

Data mining deals with analysis of large volumes of data to automatically discover interesting regularities or relationships which in turn leads to better understanding of the underlying processes [13]. The main goal is to discover hidden patterns, unexpected trends in the data. Data mining is the combination of techniques from database technologies, artificial intelligence statistics and machine learning. Over the last two decades data mining has emerged as a significant research area. The actual data mining task is the automatic or semi-automatic analysis of large quantities of data to extract previously unknown interesting patterns. This is primary due to the inter-disciplinary nature of the subject and the diverse range of application domains in which data mining based products and techniques are being employed. This includes bioinformatics, genetics, clinical research, medicine, education and marketing research.

2. COMPUTATIONAL RPD MODEL

In the computation RPD model some of the cognitive processes that a physician would employ when making decisions under uncertainty regarding suspected adverse drug reactions (ADRs) [3]. For this discussion an ADR will refer to the unanticipated drug-associated adverse incident(s) that follow the administration of a drug when it is used properly and at an appropriate dosage [1]. ADRs represent a significant public health problem. Also, the concept of an adverse reaction following drug administration does not require extensive knowledge and expertise in order to understand the significance for decision making. Computational RPD are limited and do not detect the ADR problem. Fuzzy based Computational RPD provides the effective paradigm to detect the ADR.

Also [8] used intelligent agents with a fuzzy decision model is used to develop a distributed adverse drug reaction. Adverse drug reaction is detected by utilizing distributed electronic patient data. The Recognition-Primed Decision model is furnished to a fuzzy RPD model and it uses fuzzy logic technology to represent, compute imprecise, interpret and subjective cues. They were commonly encountered in ADRs and they retrieve prior experiences that are reported.

3. CUE TYPES

Cue is a key concept in the RPD model for both the situation awareness and action evaluation processes. Cues are usually abstracted from elementary data and their types could be linear, nominal or fuzzy. A linear cue refers to a variable whose values are described by or related to a straight line. For example, the blood pressure of a patient is a linear cue. A nominal cue is represented as discrete cue and their values are not in any linear order. For example, a variable representing the gender of a patient has a value such as male, female, or unknown. A fuzzy cue is inspired by the need of representing information that is imprecise in nature. The fuzzy set theory provides a breakthrough, relatively new information processing technology for describing such imprecision, uncertainty, and subjectivity which are common issue for more real-world applications, especially those in medicine [4].

Detecting an unknown ADR is a complex process which includes spontaneous ADR reporting, expert clinical reviews and studies related on epidemiology. For example, consider how a physician assesses the evidence for causality between a drug and an adverse event in an individual patient case, which may result in filing a report and trigger further investigation. Cues Temporal association refers to the temporal relationship between taking the occurrence of the adverse reaction and drug. Rechallenge describes the relationship between re-introduction of the drug and recurrence of the adverse event. Dechallenge is defined as the relationship between withdrawal of the drug and reduction of the adverse effect. Cues temporal association, rechallenge and dechallenge are all fuzzy variables and they represent fuzzy sets and they are achieved through fuzzy reasoning.

4. MINING METHOD FOR PHARMACOVIGILANCE

Pharmacovigilance is the science which deals with the detection, assessment and prevention of ADRs. In the pre-marketing stages of a drug, pharmacovigilance deals with predicting ADRs using preclinical characters of the compounds [7]. For example, it consists of chemical structure, drug targets. In the postmarketing stage, pharmacovigilance has traditionally involved in mining spontaneous reports. The research focus towards the use of data generated from platforms outside the framework such as biomedical literature, electronic medical records

(EMRs), and patient-reported data in online health forums. The emerging of pharmacovigilance is to link preclinical data with human safety information observed in the postmarketing phase. This provides a general overview of the current computational methods applied for Pharmacovigilance at different stages of drug development and concludes with future directions and challenges.

The Pre-Marketing surveillance at the stage has been devoted to predict or assess ADRs early in the pipeline drug development. One of the fundamental methods in mining ADRs is the application of preclinical Safety Pharmacology Profiling (SPP) by testing compounds with biochemical and cellular assays. The symptom is that if a compound binds to a certain target, then its effect may translate into possible occurrence of an ADR in humans. However, experimental detection of ADRs remains challenging in terms of cost and efficiency [6]. There has been large amount of research activities devoted to developing computational approaches to predict potential ADRs using preclinical character of the screening data. Most of the existing research can be categorized into chemical structure and target-based approaches.

5. POTENTIAL CAUSAL-LEVERAGE MEASURE

Early detection of Adverse Drug Reaction protects the life of the patient [11]. It provides the relationship between drug and ICD-9(International Classification of Drug-Ninth Revision) Potential causal leverage assesses the strength of the association of a drug symptom pair [3]. Drug safety depends heavily on the post marketing surveillance. FDA (Food and Drug Administration) currently adopts data mining algorithm called Multi Item Gamma Poisson Shrinker. Support measures the statistical significance of a pattern. Minimum support based pruning is good strategy for positively correlated association rules. Confidence measures the strength of the association rule. Measuring the interestingness of discovered patterns is an active and important area of data mining. An association rule is an implication expression in the form of $X \rightarrow Y$, where X and Y are two event sets and they are disjoint (i.e., $X \cap Y = \varphi$), meaning that they share no common events .An association rule indicates that the if X is present it implies that Y will also present. X and Y temporal constraint is often applied to the association rule. It provides the relationship for temporal association. It is represented as $X,T\rightarrow Y$, is called the temporal association rule, where $T\rightarrow$ denotes that Y occurs after X within a time window T in the same event sequence. In this paper, association rule mining is based on two measures that are support and confidence. The association rule for support is represented as supp($X \rightarrow Y$). It is the proportion to the sequences in which both X and Y among all the event sequence occur s at least once.

In the context of ADR signal generation using electronic patient data, a health database often contains many patient records, each of which can be considered as an event sequence where various events such as drug prescription, occurrence of a symptom, and laboratory test occur at different times. The above mentioned frequency-based measures cannot be used for finding the association between a drug and an Adverse Drug Reaction because of the low frequency of ADRs.

6. INTERESTINGNESS MEASURES OF DEPENDENCY FROM DATAMINING

Many of the measures can be adapt to association patterns. It is a well-known fact that the data mining process can generate many hundreds and often thousands of patterns from data [5]. The task for the data miner becomes one of determining the most useful patterns from those that are trivial or are already well known to the organization. It is important to remove out those patterns through the use of few measures. This paper provides measures devised for evaluating and ranking the discovered patterns produced by the data mining process. The interestingness measures are generally divided into two categories, they are objective measures based on the statistical strengths or properties of the discovered patterns and subjective measures that are derived from the user's beliefs or expectations of their particular problem domain. Existing subjective and semantics-based measures employ various representations of the user’s background knowledge, which lead to different measures and procedures for determining interestingness. A general framework for representing knowledge that is related to data mining would be useful for defining a unifying view of subjective and semantics-based measures. Selecting interestingness measures that reflect to real human interest remains an open problem.

A. Framework for Finding Unexpected Patterns

Unexpectedness of a pattern is quantified as the impact the pattern has on a prespecified set of prior beliefs (the background knowledge). Beliefs are divided into hard and soft. Hard beliefs cannot change with new evidence. A pattern that violates a hard belief is by definition interesting as it implies incorrect data or incorrect acquisition of

patterns. A degree is assigned to each soft belief. In [9], the authors provide five methods for formalizing this degree of belief using tools from Bayesian and Dempster–Schafer theory. The difference between the degrees of the belief system before and after presenting a pattern is used as a measure of pattern's unexpectedness.

B. Using a Bayesian Network as Background Knowledge

The use of a Bayesian Network (BN) is to encode the prior knowledge of the support itemsets. Interestingness is the difference between the support of an itemset calculated from the actual data . The construction of the Bayesian Network, however, may be a nontrivial procedure for the user. Furthermore, performing inference in a Bayesian Network can be hard; it depends on the structure of the Bayesian Network. To address this issue, the authors in [10] propose an approximate inference scheme under the same rationale regarding unexpectedness. One approach is to use meta learning to automatically select or to combine appropriate measures. Another method is to develop an interactive user interface which is interpreting the data by using a selected measure to assist the selection process. Experiments comparing the results of interestingness measures with actual human interest could be used as another approach to analysis. User interactions are indispensable in the determination of rule interestingness, it is important to develop new theories, methods, and tools to facilitate the user's involvement. To reduce the number of mined results, many interestingness measures have been proposed for various kinds of pattern. Based on the form of the patterns produced by the data mining method, we distinguished measures for association rules, classification rules, and summaries. Survey distinguishes objective, subjective, and semantics-based measures. Objective interestingness measures are based on statistics, probability theory, and information theory. Therefore, they have strict principles and foundations and their properties can be formally analyzed and compared. However, objective measures take into account neither the context of the domain of application nor the goals and background knowledge of the user. Subjective and semantics-based measures incorporate the user's background knowledge and goals, respectively, and are suitable both for more experienced users and interactive data mining. It is widely accepted that no single measure is superior to all others or suitable for all applications.

7. FREQUENT PATTERN MINING

The notion of frequent itemsets was introduced by Agrawal et al [14]. Itemset can be defined as a non-empty set of items. An itemset with k different items is termed as a k-itemset. In frequent itemsets the transactions appear frequently. Frequent itemset mining goal is to identify all the itemsets in a transaction dataset. The support value of an itemset is the percentage of transactions that contain the itemset. Frequent itemset mining plays an essential role in the theory and practice of many important data mining tasks, such as mining association rules, long pattern, emerging patterns, and dependency rules. Frequent itemset mining is based on the rationale that the itemsets which appear more frequently in the transaction databases are of more importance to the user .It has been that in many real applications that the itemsets that contribute the most in terms of some user defined utility function (for e.g. profit) are not necessarily frequent itemsets. Frequent itemset mining identifies high utility item combinations. However the practical usefulness of mining the frequent itemset by considering only the frequency of appearance of the itemsets is challenged in many application domains such as retail research. The algorithm is designed to find segments of data defined through the combinations of drugs (rules) which satisfy certain conditions as a group and maximize a predefined objective function. Efficiency of mining is achieved with three techniques:

Step1: a large database is compressed into a highly condensed, much smaller data structure, which avoids costly, repeated database scans.

Step2: FP-tree-based mining adopts a pattern fragment growth method to avoid the costly generation of a large number of candidate sets, and a partitioning-based divide-and-conquer method is used to dramatically reduce the search space. FP-growth method is efficient and scalable for mining both long and short frequent patterns, and faster than the Apriori algorithm. The advantages of FP-growth over other approaches are it constructs a highly compact FP-tree. It is usually substantially smaller than the original database and it saves the costly database scans in the subsequent mining processes. It applies a partitioning-based divide-and-conquer method which dramatically reduces the size of the subsequent conditional pattern bases and conditional FP-trees. Several other optimization techniques, including ordering of frequent items, and employing the least frequent events as x, also contributes to the efficiency of the method. There are many research issues related to FP-tree-based mining, including implementation of SQL-based, highly scalable FP-tree structure, constraint-based mining of frequent patterns using FP-trees, and the extension of the FP-tree-based mining method for mining other interesting frequent patterns.

8. EXPERIMENTS

8.1 Experiment Data

The experiment data were from adverse drug reaction data. The electronic data of the patients, who received at least one of the drugs of the interest, were retrieved. The drugs allow examining the effectiveness of the data. The mining framework in identifies the potential ADR for drugs both in the same drug class and in different drug classes. ICD–9 code represents the potential ADR that is causally associated with any drug. SQL is used to query the tables and obtain desired information. The data are relational tables in a Microsoft Access database. The query results were returned to Java programs that implements the data mining algorithm. Fuzzy sets are coded using Fuzzyjess [12], a Java-based fuzzy inference engine.

8.2 Experiment Results

TABLE I

Calculating Adverse Drug Reactions

ICD 9 codes	ICD 9 codes description	Exclusive causal leverage	Causal leverage	Leverage	Risk ratio	ADR
784.0	Headache	2	18	267	1142	Present
585.5	Kidney diseases	4	28	1176	1173	Not Present
427.32	Atrial flutter	7	37	122	618	Present

Table gives the causal-leverage, exclusive causal leverage and risk ratio between drug enalapril and each of seven ICD-9 codes: 784.0 (congestive heart failure), 585 (Urinary System failure), 427 (Circulatory failure).The table shows that its exclusive causal leverage is closer to its causal leverage. The table indicates that its causal-leverage is smaller than its reverse causal-leverage, which results in a negative exclusive causal leverage value. The reverse causal-leverage is 0, and then there will be no cases present reasonable temporal patterns for the reverse pair. The code ranking is high; this indicates that some unrelated codes could still be ranked high due to the complexity of the data. Ranking these seven ICD-9 codes using their causal-leverage values. That the proposed algorithm can indeed effectively remove some undesirable effects caused by frequent events. For selected drugs compute the exclusive causal-leverage values for all the pairs between the drug and ICD-9 codes. The ranking of the codes is a decreasing order based on to each pair's exclusive causal-leverage value. Ranking the association between the drug and each ICD-9 code using three other measures. The three measures are causal-leverage, traditional leverage without considering causality and risk ratio. The value of the parameter needs to be chosen carefully. If the parameter is too large, then some unrelated symptoms could form false temporal relationships with the drug. If it is too small, signals may be excluded because some ADRs happen after a drug has been taken for as long as tens of days. The table also indicates that the results generated by these two measures (i.e., risk ratio and leverage) are much more similar than those generated by the other two measures. The ranks of the eight ICD-9 codes, the causal-leverage measure gives the much better performance as compared with the risk ratio and leverage measures since it could capture the causal relationship between a pair. With capability of reducing the undesirable effects of frequent ICD-9 codes exclusive causal-leverage measure achieves the best performance. The causal-leverage measure gives higher ranks for all these seven ICD-9 codes. The table1 show the details of the ADRs present or not present. By calculating and comparing the ranks of Exclusive causal leverage, leverage, causal leverage the ADR is find out. The above experiment results showed that the exclusive causal-leverage measure outperformed traditional interestingness measures like risk ratio and leverage because of its ability to capture suspect causal relationships and exclude undesirable effects caused by frequent events. Due to the complexity, incompleteness, and potential bias (e.g., tolerable ADRs sometimes may not be recorded) of the data, some ICD-9 codes may be falsely ranked high based on the exclusive causal-leverage. Clinical trials revealed some similar ADRs between the two statins including nausea, headaches, myalgia, and dizziness. As a result, the ADRs derived from the exclusive causal-

leverage measure represent more naturalistic and real patients in clinical care. Based on the drug intake and the dosage the Adverse Reaction can be calculated for the future scope.

9. RESULT AND DISCUSSION

ADR extraction was evaluated using our annotated data set. The training data set was used to mining the adverse drug reactions. The drug signal pairs provides the pair generation. Based on the fuzzy RPD the pair is generated. Generated pair detects the adverse reaction based on the drug consumed by the patient. Adverse effect of the patient is determined using the consumption of the drug and dosage. The focus of the proposed paper is to mine the adverse drug reaction (i.e., finding interesting ADR signal pairs) and mine all possible rare event sets and association rules based on minimal support. In contrast, in this work data from a relational (medical) database composed of several related tables. Developing an efficient algorithm is suitable for analyzing its complexity and efficiency in the future. The table1 show the details of the ADRs present or not present. By calculating and comparing the ranks of Exclusive causal leverage, leverage, causal leverage the ADR is found out. Based on the drug intake and the dosage the Adverse Reaction can be calculated for the future scope.

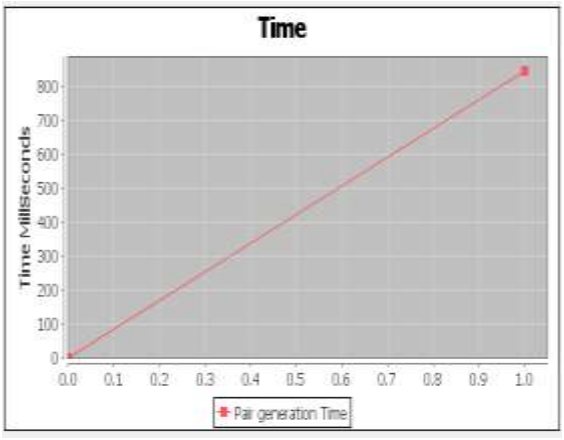

Fig 1: Pair Generation

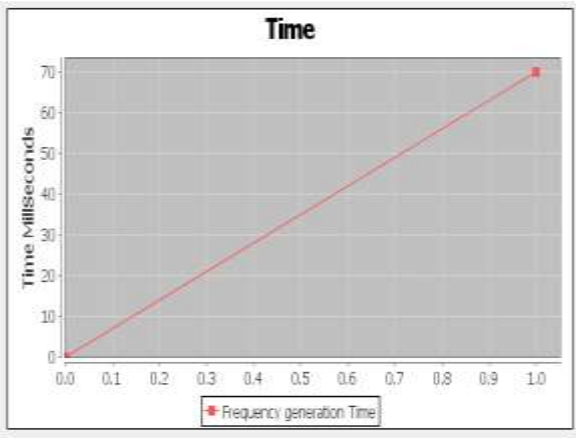

Fig 2: Frequency Generation

The figure 1 represents the generation of drug pair in 845.0 milliseconds and figure 2 represents the generation of frequency pair in 70.0 milliseconds. The method allows us to examine and evaluate complicated samples of patients to draw useful causal information. The focus of this paper is to design a novel interestingness measure and apply the result to the important adverse drug reaction problem. In contrast, the work must use data from a relational (medical) database composed of several related tables. Developing an efficient algorithm is suitable for relational databases and analyzes its complexity and efficiency in the future.

10. CONCLUSION

This paper describes various approaches to handle the ADR in data mining. Many significant works has been done in the field of fuzzy based RPD, interestingness measures and causal leverage. Potential causal leverage and experience based fuzzy RPD model have developed a data mining algorithm. Fuzzy rule based approach can be used in post marketing surveillance system to enhance the detection of ADR signal pairs.. Mining the causal association between two events is important. It is useful in many real applications. It can avoid potential adverse effects and help to discover the causality of a type of events. Mining the associations is very difficult especially when events of interest occur infrequently. The proposed method deals with interestingness measure, exclusive causal-leverage, experience-based fuzzy RPD model along with the FP Growth algorithm. The algorithm was developed to search a real electronic patient database for potential ADR signals. Experimental results showed that the algorithm could effectively make known ADRs rank high among all the symptoms in the database. These approaches discussed the merits and demerits of each technique. Many significant works has been made in the field of mining adverse drug event, yet there is more scope for future research in this field.

REFERENCES

[1] J. Talbot and P. Waller, Stephens, "Detection of New Adverse Drug Reactions", fifth ed. John Wiley & Sons, 2004.
[2] Y. Ji, H. Ying, P. Dews, M.S. Farber, A. Mansour, J. Tran, R.E. Miller, and R.M. Massanari, "A Fuzzy Recognition-Primed Decision Model-Based Causal Association Mining Algorithm for Detecting Adverse Drug Reactions in Post marketing Surveillance," Proc. IEEE Int'l Conf. Fuzzy Systems, 2010.
[3] Y. Ji, H. Ying, P. Dews, A. Mansour, J. Tran, R.E. Miller, and R.M. Massanari, "A Potential Causal Association Mining Algorithm for Screening Adverse Drug Reactions in Post marketing Surveillance,"
IEEE Trans. Information Technology in Biomedicine, vol. 15, no. 32, pp. 428-437, May 2011.
[4] G.A. Klein, "A Recognition-Primed Decision Making Model of Rapid Decision Making," Decision Making in Action: Models and Methods, pp. 138-147, Ablex Publishing, 1993.
[5] P.-N. Tan and V. Kumar, "Interestingness Measures for Association Patterns: A Perspective," Department of Computer Science, Univ. of Minnesota, 2000.
[6] Whitebread, S., Hamon, J., Bojanic, D. and Urban, L.Keynote reviews: in vitro safety pharmacology profiling: an essential tool for successful drug development. Drug Discovery Today, 10, 21 (Nov 1 2005), 1421-1433.
[7] Harpaz, R., Dumouchel, W., Shah, N. H., Madigan, D., Ryan,P. and Friedman, C. Novel Data-Mining Methodologies for Adverse Drug Event Discovery and Analysis. Clin Pharmacol Ther (May 2 2012).
[8] Yanqing, J; Y. Hao, S. Margo, S. Farber, Y. John, D. Peter, E. Richardand M. Michael. A Distributed, Collaborative Intelligent Agent SystemApproach for Proactive Post marketing Drug Safety Surveillance. IEEETransactions on information technology in Biomedicine, 14 (3), pp. 826 –837, May 2010.
[9] Silberschatz A, Tuzhilin A. What makes patterns interesting in knowledge discovery systems IEEE TransKnowl Data Eng 1996, 8:970–974.
[10] Jaroszewicz S, Scheffer T. Fast discovery of unexpected patterns in data, relative to a Bayesian network. In: Proceedings of the Eleventh ACM SIGKDD International conference on Knowledge Discovery in Data Mining. Chicago, IL: ACM; 2005, 118–127.
[11] A Method for Mining Infrequent Causal Associations and Its Application in Finding Adverse Drug Reaction Signal Pairs April 2013 (vol. 25 no. 4) pp. 721-733.
[12] R. Orchard, "Fuzzy Reasoning in Jess: The Fuzzy Toolkit and FuzzyJess," Proc. Third Int'l Conf. Enterprise Information Systems,pp. 533-542, 2001.
[13] S Laxman, P S Sastry, A survey of temporal data mining, Sadhana , Vol. 31 , part 2 , April 2006, pp. 173-198.
[14] R. Agrawal, T. Imielinski, A. Swami, 1993, mining association rules between sets of items in large databases, in: proceedings of the ACM SIGMOD International Conference on Management of data, pp. 207-216.