ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

Post Market Drug Analysis using Irregular Pattern Mining Scheme

Mr. S. Prakash[1], Ms. S. Kanjanadevi[2]

Department of CSE, Velalar College of Engineering & Technology, Erode[1]

Assistant Prof, Department of CSE, Velalar College of Engineering & Technology, Erode[2]

Abstract: Frequent pattern mining is performed using association rule mining algorithms. Candidate-set and item-set are prepared using the attribute name and its associated values. Minimum support and confidence values are used to select frequent patterns. Frequent pattern mining methods produces better performance in sparse or low dimensional data values. Dense and high-dimensional data sets have to use high thresholds to produce results within limited time and low support patterns. Rule mining methods are used to identify the drug reactions on patients.

Drug reaction analysis is performed to find out the casual associations between two set in low frequency levels. Knowledge-based approach is used to capture the degree of causality of an event pair within each sequence with application or domain dependent. Interestingness measure incorporates the causalities across all the sequences in a database. Premarketing analysis is not sufficient to detect rare areas. A data mining framework is used to mine causal associations in patient data sets where the drug-related events of interest occur infrequently. A computational fuzzy Recognition-primed Decision (RPD) model is used to estimate the interestingness measure. Support count estimation algorithm is used to estimate support count for each drug. Pair generation algorithm is used to prepare candidate set pairs. Class leverage detection algorithm is applied to mine the causal relationship between drugs and their associated adverse drug reactions (ADRs).

Scalability feature is provided in the enhanced drug reaction analysis system. Class leverage detection algorithm is enhanced with SQL functions. Rule summary analysis mechanism is integrated with the system to improve the accuracy levels. Support estimation and candidate pair generation algorithm are enhanced with aggregation functions.

I. INTRODUCTION

Association rules are if/then statements that help uncover relationships between seemingly unrelated data in a relational database or other information repository. An example of an association rule would be "If a customer buys a dozen eggs, he is 80% likely to also purchase milk." An association rule has two parts, an antecedent and a consequent. An antecedent is an item found in the data. A consequent is an item that is found in combination with the antecedent. Association rules are created by analyzing data for frequent if/then patterns and using the criteria support and confidence to identify the most important relationships. Support is an indication of how frequently the items appear in the database. Confidence indicates the number of times the if/then statements have been found to be true. In data mining, association rules are useful for analyzing and predicting customer behavior. They play an important part in shopping basket data analysis, product clustering, catalog design and store layout. Programmers use association rules to build programs capable of machine learning. Machine learning is a type of artificial intelligence (AI) that seeks to build programs with the ability to become more efficient without being explicitly programmed.

Adverse drug reactions (ADRs) refer to the drug-associated adverse incidents in which drugs are used at an appropriate dose and indication. They can complicate a patient's medical condition or contribute to increased morbidity, even death. Drug-induced morbidity and mortality occurs on a daily basis. The latest data that we can find in the literature show that in year 2000 there were about 100,000 deaths in the U.S. due to medical errors, of which about 7,000 were attributed to drug reactions. 1975 and 1999, 548 new drugs were approved by the FDA, 16 of which were subsequently

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

withdrawn from the market because of ADRs. Forty-five of the 548 drugs acquired at least 1 block box warning for an ADR that was not known when the drug was approved by the FDA for marketing. The authors pointed out that "Many serious ADRs are discovered only after a drug has been on the market for years. Only half of newly discovered serious ADRs are detected and documented in the Physician's Desk Reference within 7 years after drug approval".

Drug safety depends heavily on postmarketing surveillance - the systematic detection and evaluation of medicines once they have been marketed. Current postmarketing methods largely rely on FDA's spontaneous reporting system MedWatchTM. The limitations of this system are well described. MedWatchTM is a passive system in that it depends on voluntary, spontaneous reports of suspected ADRs to be filed by healthcare professionals, drug manufactures, and/or consumers using the system's online forms. Detection of an ADR generally relies on FDA's retrospective or concurrent review of patient cases. Because ADR reports are filed at the discretion of the users of the system, there is gross underreporting. It was estimated that less than 10% of all ADR cases were reported to MedWatchTM. Moreover, it depends on human recognition of a potential link between a drug and an apparent adverse reaction and on the time and will to report the observation. In addition, the rate at which cases are reported is dependent on many factors, including the time period since the drug was released into the market place, pharmacovigilance-related regulatory activity, the indications for use of the drug and media attention. Finally, the passive surveillance system is limited by latency and inconsistent. Consequently, the current approach may require years to identify and withdraw problematic drugs from the market, and result in unnecessary mortality, morbidity, and cost of healthcare.

II. RELATED WORK

There exist some works on rare eventset mining in the literature. A straightforward approach to discovering infrequent eventset is to relax the uniform minimal support criterion. One drawback of this mining approach is the extremely high computational cost. In addition, if this approach was used to discover rare events like ADRs, thresholds minsupp and minconf would have to be set very small, possibly leading to a lot of false associations. A couple of studies attempted to use less uniform support criteria. Yun et al. adopted a relative support approach, while Liu and his colleagues proposed a method that relied on multiple minimal support thresholds specified by users. These approaches output all frequent eventsets and association rules together with a subset of all infrequent ones. Several other studies explored rare eventsets with support lower than the threshold [6]. These studies implemented different strategies to traverse the power set lattice of a data set. For example, some of them take a bottom-up approach to moving across the lattice, while another algorithm named Rarity takes a top-down method [1]. Koh and Rountree proposed an algorithm called Apriori-Inverse where sporadic rules are discovered by simply discarding all eventsets above a support threshold.

The above approaches to detection of rare association rules are based on traditional interestingness measures like support and confidence. One pitfall of these measures is that they simply find the statistical correlation between X and Y . For example, the confidence of an association rule $X \rightarrow Y$ determines how frequently Y appears in those event sequences that contain X. That is, traditional measures like confidence try to find the statistical significance of the coexistence of two eventsets. They do not indicate any temporal relationship between X and Y. In addition, they are not able to capture the causal relationships between two event sets. They do not specify whether X causes Y , or vice versa, or a third event causes the coexistence of X and Y . Hence, statistical associations established by traditional measures may or may not represent temporal or causal associations.

Causal modeling and inference have been widely studied in the field of machine learning and traditional probability and statistics theory. Compared with statistical co-existence, causal relationships have a couple of unique properties. First, there is an intrinsic asymmetry in a cause-effect relationship. That is, when saying event X causes event Y, one would not expect X to be influenced by Y. Second, the concept of causality is linked to time dependencies. That is, the causes must precede their effects. Researchers have proposed various models to model causal relationships for the purpose of prediction or data modeling. These models include artificial neural networks, Markov models [7], various graphical models [4], etc. One of the most famous models is Pearl's causal model where causal modeling and inferences are based on representations by means of directed acyclic graphs (DAGs). For a more extensive and detailed review about the state of the art of causality, readers are referred to a recent survey paper [5].

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

With a surge of interest in association rule mining in recent years, some of the above models have been borrowed to mine the causal relationships between two events or eventsets. For instance, Bayesian network and its variants have been adapted to mine causal structures in a couple of studies. One criticism of Bayesian analysis is its requirement to specify prior distributions for all variables, which can be very difficult since, in many cases, prior knowledge is vague or nonexistent, or can be tedious if the number of variables is large. Silverstein et al. proposed a constraint-based algorithm that narrowed down the search space and thus made the mining process more scalable. In addition, Hidden Markov Models (HMMs) have been utilized to predict protein structures and analyze genome sequences based on massive amounts of observed data. However, these approaches are inherently probability based and designed to find frequent patterns. Thus, they have limited capability in discovering infrequent associations.

III. DRUG REACTION ANALYSIS

In this paper, we try to employ a knowledge-based approach to capture the degree of causality of an event pair within each sequence since the determination of causality is often ultimately application or domain dependent. We then develop an interestingness measure that incorporates the causalities across all the sequences in a database. Our study was motivated by the need of discovering ADR signals in
postmarketing surveillance, even though the proposed framework can be applied to many different applications. ADRs represent a serious world-wide problem. They can complicate a patient's medical condition or contribute to increased morbidity, even death. Studies have shown that ADRs contribute to about 5 percent of all hospital admissions and represent the fifth commonest cause of death in hospitals.

Even though premarketing clinical trials are required for all new drugs before they are approved for marketing, these trials are necessarily limited in sample-size and duration, and thus are not capable of detecting rare ADRs. In general, an ADR cannot be recognized by these trials if its occurrence rate is less than 0.1 percent. Therefore, drug safety depends heavily on postmarketing surveillance; that is, the monitoring of impacts of medicines once they have been made available to consumers. In the US, current postmarketing surveillance methods primarily rely on the FDA's spontaneous reporting system MedWatch. Because ADR reports are filed at the discretion of the users of the system, there is gross underreporting. Consequently, the current approach may require years to identify and withdraw problematic drugs from the market, and result in unnecessary mortality, morbidity, and cost of healthcare. Studies have shown that only half of newly discovered serious ADRs are detected and documented in the Physician's Desk Reference within 7 years after drug approval.

As electronic patient records become more and more easily accessible in various health organizations such as hospitals, medical centers, and insurance companies, they provide a new source of information that has great potential to generate ADR signals much earlier [8]. Note that each patient case can be considered as an event sequence where various events such as drug prescription, occurrence of a symptom and lab test occur at different times. In the literature, there exist a couple of studies [2] that attempted to find the associations between drugs and potential ADRs by mining their temporal relationships. That is, they tried to mine temporal association rules (represented as $X \xrightarrow{T} Y$) where Y occurs after X within a time window of length T.

These studies obtained promising results based on administrative health data. However, temporal association was the only parameter used for linking a symptom with a drug within each patient case in their work. Temporal association assumes that cause precedes effect. Other parameters such as dechallenge and rechallenge can also give direct or indirect cues of the potential causal association of a drug-symptom pair. Dechallenge is defined as the relationship between withdrawal of the drug and abatement of the adverse effect. Rechallenge describes the relationship between reintroduction of the drug followed by recurrence of the adverse event. In addition, their approaches suffer from the sharp boundary problem. On the one hand, the symptom events near the time boundaries are either ignored or overemphasized. On the other hand, two symptom events contribute equally to the interestingness measure as long as they occur within the hazard period T. That is, the length of the time duration between exposure to the drug and occurrence of the symptom has no effect on the interestingness measure. This is not true in reality because if an ADR symptom occurs within a shorter period, it is usually more likely to be caused by the drug. To more effectively mine infrequent causal associations, it is necessary to develop a new data mining framework. The this paper is a substantial extension of our previous work [3] where an interestingness

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

measure called causal-leverage was developed on the basis of a computational fuzzy recognition-primed decision (RPD) model we previously developed.

3.1. Recognition-Primed Decision Model (RPD)

The RPD model represents a popular cognitive decision model. It is particularly useful for modeling how human experts make decisions based on their prior experiences. It was found that about 50 to 80 percent of all decisions were made in this way [7]. The original RPD model is descriptive and is not directly implementable on a computer. Hence, we developed a fuzzy RPD model, which is not only computational but also capable of handling vague and subjective information using fuzzy logic.

Experiences play a key role in the RPD model. The ADR detection experiences were acquired through the joint efforts of our engineering and medical team members after careful analysis of the relevant literature. According to the classification scheme, a particular pattern of cue values characterizes a specific degree of causality which may require certain courses of action to handle the ADR. Therefore, we can define four experiences, each of which is associated with a degree of causality. These experiences were stored in an experience knowledge base (EKB). An experience consists of four components—cues, goals, actions, and expectancies. Cues represent the higher level information that a decision maker must pay attention to. Expectancies describe what will happen next as the current situation continues to evolve in a changing context. Goals represent an end state that the decision maker is trying to achieve. Actions represent a set of potential decisions that the decision maker can take in the current situation. Cues are used to match the current situation with prior experiences and determine which experience can be reused to solve a new problem. This sample experience has four cues: temporal association, dechallenge, rechallenge, and other explanation.

After representing the experiences, the next step is to extract the cue values from elementary data for the current situation. Fuzzy rules and fuzzy reasoning are used to achieve this task. For example, to obtain the fuzzy value of temporal association, one of the fuzzy rules of the type "if the time duration between taking the drug and the occurrence of the adverse event is short, then temporal association is likely" is used. An embedded fuzzy inference engine is employed to perform fuzzy reasoning. The inference engine is what drives the RPD model, updating cues once new information is detected, and monitoring expectancies and goals. After the cue values of the current situation are known, similarity measures are applied to measure the degree of matching between the current situation and prior experiences. The actions in the most matching experience are then used to solve the current problem. For more details of the fuzzy RPD model as well as concrete examples, the reader is referred to our previous work.

3.2. Fuzzy Recognition-Primed Decision Model (RPD)

The fuzzy RPD model was preliminarily validated in our previous study using it to calculate the extent of causality between cisapride and some of its adverse effects for 100 simulated patients created based on the profiles of 1,015 patients of the VA Medical Center. The simulated patients were used because we found that the number of apparent signal pairs of potential interesting symptoms in the real patient data was very limited. Therefore, we used the real patients to create simulated patient cases, all of which containing drug-symptom pairs of interest with various degrees of causality. The model's validity was then established by comparing the decisions made by the model and those by two independent experienced physicians for the 100 simulated patients. The levels of agreements were measured by the weighted Kappa statistic, which is a measure of agreement between two raters after chance agreement is controlled. The results suggested good to excellent agreements.

IV. ISSUES ON DRUG REACTION IDENTIFICATION PROCESS

Finding causal associations between two events or sets of events with relatively low frequency is very useful for drug reaction analysis. Knowledge-based approach is used to capture the degree of causality of an event pair within each sequence with application or domain dependent. Interestingness measure incorporates the causalities across all the sequences in a database. Premarketing analysis is not sufficient to detect rare areas. A data mining framework is used to mine causal associations in patient data sets where the drug-related events of interest occur infrequently. A computational

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

fuzzy Recognition-primed Decision (RPD) model is used to estimate the interestingness measure. Support count estimation algorithm is used to estimate support count for each drug. Pair generation algorithm is used to prepare candidate set pairs. Class leverage detection algorithm is applied to mine the causal relationship between drugs and their associated adverse drug reactions (ADRs). The following problems are identified from the current drug reaction analysis schemes. They are computational complexity is high, supports limited data only, relational database environment is not supported by the system and detection latency is high.

V. POST MARKET DRUG ANALYSIS

The drug reaction identification system is improved to support large volume of data. Class leverage detection algorithm is enhanced with SQL functions. Rule summary analysis mechanism is integrated with the system to improve the accuracy levels. Support estimation and candidate pair generation algorithm are enhanced with aggregation functions. The system is designed to analyze the post drug reactions with low support levels. Candidate sets are generated with drug reaction information. Class leverage analysis is performed to detect drug reactions. The system is divided into five major modules. They are EPR analysis, candidate set identification, support analysis, frequent pattern mining and low support rule mining. The EPR analysis module is designed to perform data preprocessing on patient records. Candidate set construction module combines the attributes with associated weight values. Support analysis module is designed to estimate support count values. Frequent item sets are identified in the pattern mining process. Low support rule mining module is designed to detect rules with limited frequency.

5.1. EPR Analysis

Electronic Patient Record (EPR) is maintained in five different tables. Patients demographic data, visit data, diagnostic data, drug-related data and laboratory data are used for the EPR details. Patient diagnosis and associated drug reactions are analyzed in the EPR analysis. Diagnosis and drug reaction summary is estimated for each patient record.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

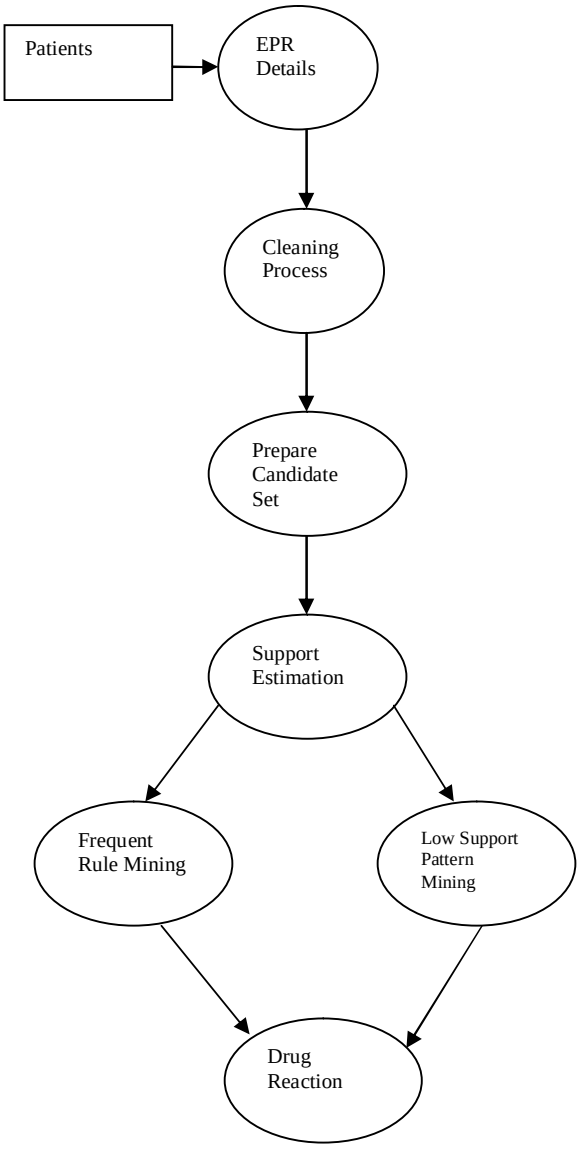

Fig. No: 5.1. Post Market Drug Analysis

5.2. Candidate Set Identification

The candidate sets are used to represent attributes and associated weight values. Candidate set generation process combines the attribute name and associated weight values. Candidate set identification is initiated for all attributes. Support count is estimated for all candidate sets.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

5.3. Support Analysis

Candidate set pairs are prepared to identify the candidate rules. Candidate rules are referred as item sets. Support count is estimated for the candidate set pairs. Probability ratio for each candidate set pair is estimated with support count values.

5.4. Frequent Pattern Mining

The frequent pattern mining is applied to filter most frequent rules. Interestingness measure is estimated using fuzzy recognition primed decisions (RPD) values. Support and confidence values are used for frequent rule identification process. Minimum support and confidence values are used to identify the frequent patterns.

5.5. Low Support Rule Mining

Class leverage detection algorithm is used to find frequent items. SQL commands are used to enhance the class leverage detection algorithm. Rule summary is used to identify the rules and their frequency values for each support level. Frequent rule selection is improved with rule summary analysis scheme.

VI. CONCLUSION

The drug reaction identification system is used to analysis the drug reactions on patients. Support confidence values are estimated for all item sets. Frequent rules are identified with reference to the support and confidence values. Casual reactions are identified with low support probability values. The system is enhanced to support high scalability with SQL support. The system is analyzed with rule precision and rule retrieval rate parameters. The computational complexity is also analyzed in the system. The frequent pattern mining (GPM) mechanism and low support pattern mining (LSPM) mechanism are analyzed in the system. Post mining drug reactions are extracted from the system with better rule precision levels.

REFERENCES

[1] L. Troiano and C. Birtolo, "A Fast Algorithm for Mining Rare Itemsets," Proc. Ninth Int'l Conf. Intelligent Systems Design and Applications, 2009.

[2]J. Hopstadius and I.R. Edwards, "Temporal Pattern Discovery in Longitudinal Electronic Patient Records" Data Mining and Knowledge Discovery, 2010.

[3] Y. Ji, Ying and assanari, "A Potential Causal Association Mining Algorithm for Screening Adverse Drug Reactions in Postmarketing Surveillance," IEEE Trans. Information Technology in Biomedicine, May 2011.

[4] J. Pearl, Causality: Models, Reasoning and Inference, Cambridge Univ. Press, 2009.

[5] I. Guyon and Janzing, "Causality: Objectives and Assessment," JMLR Machine Learning Research Workshop and Conf. Proc., 2010.

[6] L. Szathmary and A. Napoli, "Finding Minimal Rare Itemsets and Rare Association Rules," Proc. Fourth Int'l Conf. Knowledge Science, Eng. and Management, 2010.

[7] G.A. Fink, Markov Models for Pattern Recognition: From Theory to Applications. Springer, 2010.

[8] Y. Ji, H. Ying, and R.M. Massanari, "A Distributed, Collaborative Intelligent Agent System Approach for Proactive Postmarketing Drug Safety Surveillance," IEEE, Dec. 2010.