Mining Infrequent Causal Association in Finding Adverse Drug Reaction

P. Kiruthika[1], S. Suganya[2], N. Sundaravadivu[3], K. Madhubala4, A. Sheelavathi[5]
[1,2,3,4]B.Tech (IT)
[5]Assistant Professor/IT

***Abstract* — We developed and incorporated an exclusion mechanism that can effectively reduce the undesirable effects caused by frequent events. Our new measure is named exclusive causal-leverage measure. We proposed a data mining algorithm to mine ADR signal pairs from electronic patient database based on the new measure. The algorithm's computational complexity is analysed .We compared our new exclusive causal-leverage measure with our previously proposed causal-lever-age measure as well as two traditional measures in the literature: leverage and risk ratio. To establish the superiority of our new measure, we did extensive experiments. In our previous work, we tested the effectiveness of the causal-leverage measure using a single drug in the experiment. In this paper, we selected three drugs and evaluated the top 10 ICD-9 (International Classification of Diseases, ninth Revision) codes ranked by the exclusive causal-leverage measure for each drug .We propose an innovative data mining framework and apply it to mine potential causal associations in electronic patient data sets where the drug-related events of interest occur infrequently. A data mining algorithm was developed to mine the causal relationship between drugs and their associated adverse drug reactions (ADRs).**

I. INTRODUCTION

Adverse drug reactions (ADRs) represent a serious problem worldwide. They refer to 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. Discovering unknown adverse drug reactions (ADRs) in post marketing surveillance as early as possible is of great importance. The current approach to postmarketing surveillance primarily relies on spontaneous reporting. Drug safety depends heavily on postmarketing surveillance, the systematic detection and evaluation of medicines once they have been marketed. The infrequency makes existing frequent itemsets/sequential patterns mining techniques ineffective. Thus, finding these unexpected and infrequent episodes necessitates innovative knowledge representations and mining techniques. Finding causal associations between two events or sets of events with relatively low frequency is very useful for various real-world applications. For example, a drug used at an appropriate dose may cause one or more adverse drug reactions (ADRs), although the probability is low. Discovering this kind of causal relationships can help us prevent or correct negative outcomes caused by its antecedents. In this system, 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 post marketing surveillance, even though the proposed framework can be applied to many different applications. ADRs may happen from following a single drug or results from the mixture of two or more drugs. This ADRs may be a known reactions or side effects of the drugs otherwise it may be unknown or new side effects that are unrecognized previously. Medication errors may result to Adverse Drug Events (refers ADE) but most of them do not. Medication errors are accidents that happened during recommending, transcribing, screening, dispensing and supervising the drugs. Examples for ADE are misreading or miswriting the prescription. ADRs are one of the type of ADEs. The purpose for recording these ADRs to halt the later injuries for the patients. Causality refers to the relationship of a given adverse event to a specific drug. The assessment of causality determination is very difficult because the lack of reliable data. In this proposed system, we focus on mining infrequent causal associations. We developed and incorporated an exclusion mechanism that can effectively reduce the undesirable effects caused by frequent events. Our new measure is named exclusive causal-leverage measure. We proposed a data mining algorithm to mine ADR signal pairs from electronic patient database based on the new measure. The algorithm's computational complexity is analyzed. We compared our new exclusive causal-leverage measure with our previously proposed causal-leverage measure as well as two traditional measures in the literature: leverage and risk ratio. To establish the superiority of our new measure, we did extensive experiments. In our previous work, we tested the effectiveness of the causal-leverage measure using a single drug in the experiment.

II. LITERATURE REVIEW

A. *Data Mining*

Data Mining (the analysis step of " Knowledge Discovery in Database" process, or KDD), interdisciplinary subfield of Computer Science, is the computational process of discovering patterns in large datasets involving methods at the intersection of Artificial Intelligence, machine learning, statistics and database system. The overall goal of Data Mining process is to extract information from a dataset and transform it into an understandable structure of future use. Aside from the raw analysis step, it involves database and data management aspects, data pre-processing, model and interface consideration, interestingness metrics, complexity consideration, post-processing of discovered structures, visualization and online updating.

The term is a misnomer, because the goal is the extraction of patterns and knowledge from large amount of data, not the extraction of data itself. The actual data mining task is the automatic or semi-automatic analysis of large quantities of data to extract previously unknown interesting patterns such as group of data records (cluster analysis), unusual records (anamoly detection) and dependencies(association rule mining). This usually involves using Database techniques such as spatial indices. These patterns can then be seen as a kind of summary of input data, and may be used in further analysis. Before data mining algorithm can be used, a target data set must be assembled. As data mining can only uncovered patterns actually present in the data, the target data set must be large enough to contain these patterns while remaining concise enough to be mined within an acceptable time limit. A common source for data mart is data warehouse. Pre-processing is essential to analyze the multivariate data set before data mining. The target set is then cleaned. Data cleaning removes the observation containing noise and those with missing data.

Data Mining involves six common classes of tasks:

- Anamoly detection- the detection of unusual data records, that might be interesting or data errors that require further investigation.
- Association rule learning- searches for relationship between variables.
- Clustering- is the task of discovering groups and structures in the data that are in some way or another "similar", without using known structures in the data.
- Classification - is the task of generalizing known structure to apply to new data.
- Regression- attempts to find a function which models the data with the least error.
- Summarization - providing a more compact representation of the data set including visualization and report generation.

III. ALGORITHMS

Apriori is an algorithm for frequent item set and association rule learning over transactional database. It proceeds by identifying the frequent individual items in the database and extending them to larger and larger item sets as long as those item sets appear sufficiently often in the database. The frequent item sets determined by Apriori can be used to determine association rules which highlight general trends in the database.

Support:

The support supp(X) of an itemset X is defined as the proportion of transactions in the data set which supp(X)= no. of transactions which contain the itemset X / total no. of transactions contain the itemset.

Confidence:

The confidence of a rule is defined
Conf(x → y)=sup(x U y)/Supp(x).

Frequency:

Frequency is estimated by using the support and confidence values. Freq(x→y)=sup(x,y)/Conf(x,y).By using that we can identify the frequent itemset in the dataset.

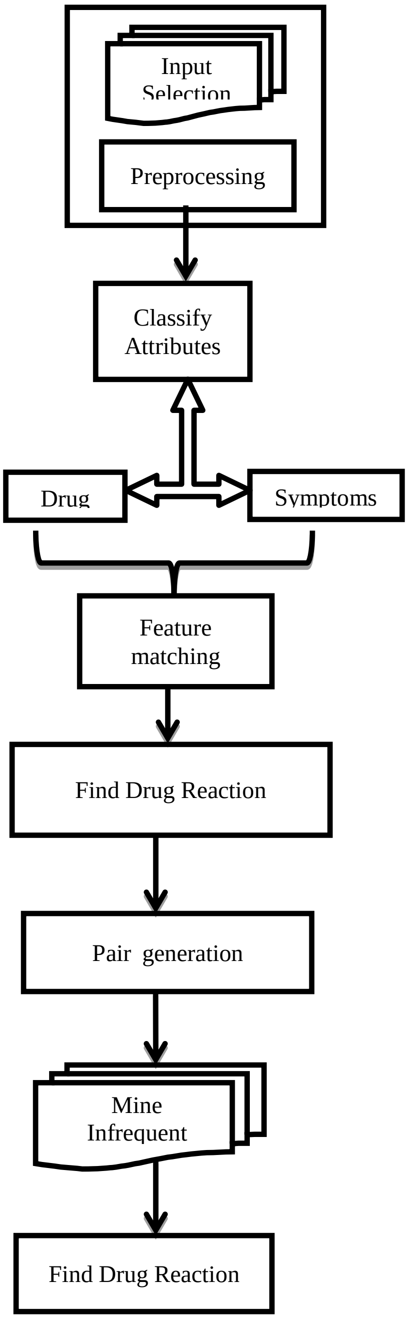

Figure 1.Architecture for Mining Rare ADRs

A. Implementation

Most commonly a data set corresponds to the contents of a single database table, or a single statistical data matrix, where every column of the table represents a particular variable, and each row corresponds to a given member of the dataset. In our process, the input is dataset which is collected from the website and embedded into the database. Data preprocessing is the method for cleaning the dataset. Normally data contains incomplete attribute values, or aggregate data noisy, or discrepancies in codes or names. Data cleaning eliminates this type of occurring in the dataset. Classification is a way of categorizing the data (records) for an attribute. The choice of classification system is critical to information displayed by a map. Classification can be used to enhance the information or to deliberately mislead. Attributes can use different classifications for the same data to change the nature of the display. Attributes in the dataset are patient id, drug name, symptoms, etc. By using this type of information, the attributes are classified.

Apriori is an algorithm for frequent item set and association rule learning over transactional database. It proceeds by identifying the frequent individual items in the database and extending them to larger item sets as long as those item sets appear sufficiently often in the database. Based on Apriori algorithm, the pair generation is predicted. Based on the ADR scores, the adverse drug reaction for every drug is predicted.

The parameters such as execution time, processing speed and memory are based on the resource utilization parameters such as efficiency and accuracy.

IV. CONCLUSION

We have developed a framework that aims to help safety experts and healthcare professionals in healthcare systems to achieve better post marketing surveillance and earlier detection of potential ADRs. Mining ADR signals is of great application value. ADR signals generated can be used, after validation, to prevent lots of unnecessary conditions or hospitalization worldwide. Causal association represents an important relationship between two events in many applications. Finding causal associations can help us prevent adverse effects caused by their antecedents. A data mining algorithm was developed to search a real electronic patient database for potential ADR signals. Experimental results showed that our algorithm could effectively make known ADRs rank high among all the symptoms in the database

V. FUTURE ENHANCEMENT

Further in future scope of the work may be directed towards the enhancement of desktop application as a web service. Eventhough the data records are increased the efficiency will be maintained. This technique can also be used in many real world applications. Nowadays this is used in many of the Diagnostics centers, whereas in future it can be used in every hospital to find drugs under ADR.

REFERENCES

[1] H. Jin, J. Chen, H. He, G. Williams, C.Kelman, and C. O'Keefe, " Mining unexpected temporal associations: applications in detecting adverse drug reactions, " *IEEE Transactions on Information Technology in Biomedicine*, vol. 12, pp. 488-500, 2008.

[2]H.SankaraVadivu,E.Manohar,R.Ravi,"Postmining Of Association Rule Using Ontologies And Rule Schemas", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 3 Issue 3, March 2014.

[3]L. Szathmary, P. Valtchev, and A. Napoli, "Finding Minimal Rare Itemsets and Rare Association Rules," Proc. Fourth Int'l Conf.Knowledge Science, Eng. and Management, pp. 16-27, 2010.

[4] Y.Ji , H. Ying, M. S. Farber, J. Yen, P. Dews, R. E. Miller, and R.M.Massanari , "A Distributed, Collaborative Intelligent Agent System Approach for Proactive Post marketing Drug Safety Surveillance," *IEEE Transactions on Information Technology in Biomedicine,* In press.

[5] Y. Ji, H. Ying, J. Yen, and R.M. Massanari, "A Fuzzy Logic-Based Computational Recognition-Primed Decision Model," Information Science, vol. 77, pp. 4338-4353, 2007.