EFFECTIVE ALGORITHM FOR MINING ADVERSE DRUG REACTIONS

[1]H.SANKARA VADIVU, [2]E.MANOHAR, [3]R.RAVI

[1]*(PG SCHOLAR, Department of CSE, Francis Xavier Engineering College)*
[2]*(ASSISTANT PROFESSOR, Department of CSE, Francis Xavier Engineering College)*
[3]*(PROFESSOR AND HEAD, Department of CSE, Francis Xavier Engineering College)*

Abstract : **One of the most important issues in the assessment of drug safety is Adverse Drug Reactions(ADR). In premarketing clinical trials, most of the ADRs are not discovered because of the limitations in size. But that are discovered in post-marketing surveillance that is, the impacts of medicines are monitored when they have been delivered to the user. Nowadays, many data mining techniques and methodologies have been developed to motivate the mining and detection of ADRs. These methods are inconvenient and inefficient for users and time consuming. We proposed a combined system platform for the detection of ADRs. This system platform proposed a new data mining algorithm, named as NM-Algorithm. It is based on the genetic algorithm with supervised learning. This proposed algorithm is completely different from the association rule mining. This proposed algorithm covers both similarity and non-similarity between the elements so that it is much more efficient than others.**

Index Terms : **Adverse Drug Event, Adverse Drug Reaction, NMRatio.**

I. INTRODUCTION

An **adverse drug reaction** (abbreviated **ADR**) is an expression that describes the harm associated with the use of given medications that experienced by the normal dosage during normal use. 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 a genetic algorithm, an initial population of candidate solutions to an optimization problem are evolved toward better solutions. It starts from a population that has the randomly generated individuals and the generation is an iterative process. Each individual's for each generation in the population are evaluated. Usually the value of the objective function is the fitness which is obtained by solving the problem. When a satisfactory fitness level has been reached and then the algorithm terminates. Supervised learning is the task of taking known set of inputs and build a predictable model that generates the predictable results. In this kind of machine learning, the input object consists of a pair and produces the desired output value. The training data are analyzed by this algorithm and produces the function which maps the new example.

In this proposed system, we try to employ an interactive approach to capture the adverse drug reactions between drugs and their reactions. To capture the relationship drugs and symptoms, first we concentrated on generating the initial population named as medical datasets. Based on genetic algorithm with supervised learning, we generate the training tuple. The fitness function that is NM ratio is calculated by comparing the initial population with the training instances. From the fitness function results, the adverse drug reactions are effectively mined.

II. REVIEW OF LITERATURE

Postmarketing surveillance was limited by gross underreporting (<10% reporting rate), latency, and inconsistent reporting. To overcome the limitations, an interestingness measure, *causal-leverage* was proposed [7] to signal potential adverse drug reactions (ADRs) from electronic health databases which are readily available in most modern hospitals. Another interestingness measure, exclusive causal-leverage, causal leverage were also developed to mine the relationship between drugs and symptoms [1], [2] and the measures were based on a fuzzy recognition-primed decision (RPD) model[1]. To mine the causality relationship between drugs and ADRs, a data mining algorithm was developed based on PCAR. An algorithm for discovering rare association rules in distributed environment was proposed in [2]. It utilized the idea of using statistic percentile to produce multiple minimum supports to mine rare association rules.

Drug safety signal detection was one of the growing interest from the observational data. In this paper, they proposed two novel algorithms [3] —a likelihood ratio model and a Bayesian network model—for adverse drug effect discovery. Knowing the drugs patients take is not only critical for understanding patient health (e.g., for drug-drug interactions or drug-enzyme interaction), but also for secondary uses, such as research on treatment effectiveness. An algorithm [4] *SPOT* was proposed which identifies the drug names that can be used as new dictionary entries from a large corpus, where a "drug" is defined as a substance intended for use in the diagnosis, cure, mitigation, treatment, or prevention of disease. Within this realm, they presented a novel temporal event matrix representation and learning framework [6] that discovers complex latent event patterns, which were easily interpretable by humans.

Early detection of adverse drug reactions (ADRs) which are unknown in postmarketing surveillance saves lives and prevents harmful consequences. A novel data mining approach [5] was proposed to signaling potential ADRs from electronic health databases. More specifically, they introduced potential causal association rules (PCARs) to represent the potential causal relationship between a drug and ICD-9 codes. To develop intelligent molecular biomarkers, a methodology and the related concepts was proposed [8] via knowledge mining and knowledge discovery in data, illustrated on prostate cancer diagnosis. Mining knowledge done an informed feature selection about pathways involved in prostate cancer and in specialized data bases. Rule Schema Formalism was proposed [9] to reduce the number of rules to several dozens or less. Moreover, the domain experts validated the quality of the filtered rules at various points in the interactive process.

ADR-Monitor was proposed [10] which represents a novel team-based intelligent agent software system approach for monitoring and detecting potential ADRs of interest from electronic patient records. A novel fuzzy subspace-based approach to hidden Markov model [11] was proposed and features extracted from patterns are considered as feature vectors in a multi-dimensional feature space. Unexpected temporal association rules (UTARs) was proposed [12] to describe unanticipated episodes where certain event patterns unexpectedly lead to outcomes.

III. PROPOSED SYSTEM FRAMEWORK

DESIGN STRATEGY

This proposed system establishes an interactive platform for the end user to allow the analysis and detection of suspicious ADR signals. The detection of ADR signal is time consuming and it has the same complexity as typical data mining tasks for example, association rule mining and classification analysis. A general concept commonly used to reduce the computation time in the context of query processing, that includes the preprocessing concepts. This executes the total or partial computation which are involved in the process to answer the query in advance.

SYSTEM OVERVIEW

In this proposed system, we have databases such as medical database and patient database. Patient information and medical information are collected from various web resources. Before proceeding into the algorithm, first we need to preprocess these databases. In this preprocessing step, we use the partitioning method. After preprocessing these databases, it is applied into the data mining engine. The data mining engine includes the NM Ratio generator which performs the NM ratio calculations. Based on this ratio, we mine the adverse drug reactions.

ALGORITHM TO MINE ADR SIGNALS

We use Γ_P and Γ_T to represent the medical datasets and patient datasets. Medical datasets are considered as the initial population and patient datasets are considered as the training datasets. Both medical datasets and patient datasets have many attributes regarding with drugs and their related symptoms. Medical datasets have the drugs related attributes such as drug name, dosage and its description. Patient datasets have the patient related attributes such as patient details with the usage of drugs. Initially, the datasets are initialized into the system. Both datasets have n number of attributes. $E_p \rightarrow E_{p1}, E_{p2}......E_{pn}$ represents the number of elements in the medical datasets and $E_T \rightarrow E_{T1}, E_{T2}......E_{Tn}$ indicates the number of elements in the patient datasets.

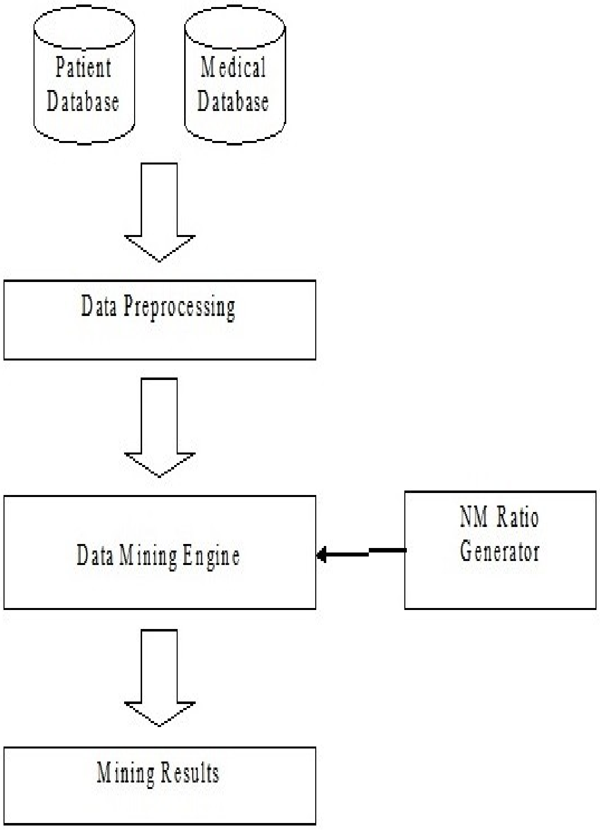

Fig. 1. System Architecture

TABLE I
SAMPLE MEDICAL DATASET

S.No	Age	Sex	Medicine	Dosage	Purpose
1	P	Male	Diazepam	100mg	Allergy
2	S	Female	Acetyle salicylic acid	50mg	Marasmos
3	T	Male	Diclofenac	200-250mg	Stomach pain
4	T	Male	Iboprofen	200mg	Headache
5	T	Female	Paracetamol	200mg	Pain relief

Algorithm : NM Ratio calculation

```
1. INITIALIZATION
        $\Gamma_P \rightarrow$ MedicalDataSet
        $\Gamma_T \rightarrow$ PatientDataSet
        $E_p \rightarrow E_{p1}, E_{p2}......E_{pn}$
        $E_T \rightarrow E_{T1}, E_{T2}......E_{Tn}$
2. CALCULATION OF N
for each elements in $\Gamma_P$ do
  if ( $E_p$ .key = $E_T$ .key) then
    for each attribute in $\Gamma_P$ do
      count = 0
        if ( $\Gamma_P$ .attributeValue = $\Gamma_T$ .attributeValue) then
          count = count + 1
        end if
     end for
  end if
end for
  N = count
return M
3. CALCULATION OF M
for each elements in $\Gamma_P$ do
  if ( $E_p$ .key = $E_T$ .key) then
    for each attribute in $\Gamma_P$ do
      count = 0
        if ( $\Gamma_P$ .attributeValue $\neq$ $\Gamma_T$ .attributeValue) then
          count = count + 1
        end if
     end for
  end if
end for
  N = count
return M
4. NM RATIO CALCULATION
for each elements in $\Gamma_P$ do
                  ratio = $\frac{N}{M}$
end for
5. MINING PROCESS
if ( ratio > dec-criteria )
        no adverse effect
else
        some adverse effect
end if
```

To proceed with the NM algorithm, initially we need to calculate the value for N and M. It is explained in the second and third step of the algorithm. To calculate the value for N and M, we need to consider the key attributes which are useful for finding the adverse drug reaction. For example, drugs and its purpose. These key attributes are compared with the training datasets which are considered as the patient datasets. Let N be the number of matches of the input attribute values of E with training instances of its own class. Let M be the number of input attribute value matches to all training instances from the competing classes and 1 is added to the resulting count. Divide the values of N by the values of M and the resulting value is denoted by Ratio. If the ratio passes the fitness criteria then it has no adverse drug reaction else it has some ADR.

IV. RESULTS

Sample medical datasets are described in Table1 and is referred as the initial population. Similarly the electronic patient datasets are collected and is considered as the training datasets. Assume that the initial population has 5 elements and the training tuples have 10 elements. Element 1 in the initial population is a member of the class medicine = Diazepam. Therefore N is computed as the number of matches element 1 has with the members in Table2 having medicine = Diazepam. That is, for element 1 we compute N using the following information.

- Age = P matches with matches with training instances 1and 4.
- Sex = Male matches with training instances 1,7 and 9.
- Dosage = 100 matches with training instances 1and 4.
- Purpose = Allergy matches with training instances 1,4 and 7.

TABLE II
SAMPLE PATIENT DATASETS

S.No	Age	Sex	Medicine	Dosage	Purpose
1	P	Male	Diazepam	100mg	Allergy
2	T	Female	Paracetamol	450mg	Fever
3	S	Male	Diclofenac	200mg	Cold
4	P	Female	Diazepam	100mg	Allergy
5	P	Male	Paracetamol	450mg	Stomach pain
6	S	Female	Iboprofen	200mg	Fever
7	T	Male	Diazepam	150mg	Allergy
8	P	Female	Paracetamol	450mg	Cold
9	S	Male	Diazepam	50mg	Stomach pain
10	T	Female	Acetyle salicylic acid	100mg	Marasmos

Therefore, the value of N computes to 10. Next, we compute M by comparing element 1 with the training instances 2,3,5,6,8 and 10 (competing classes). The computation is as follows:

- Age = P matches with matches with training instances 5 and 8.
- Sex = Male matches with training instances 5and 3.
- Dosage = 100 matches with training instance 10.
- There are no matches for purpose = allergy.

As there are 5 matches with the training instances where medicine ≠ Diazepam, the value of M for instance 1 is 4. Next we add 1 to M, giving us a final NM Ratio for element 1 is 10/6, or 1.67. Similarly we calculate the NM ratio for the other elements in the medical dataset. Thus, the following is the list of NM ratio values for the entire sample initial population. R(1) = 1.67, R(2) = 0.5, R(3) = 0.18, R(4) = .10 and R(5) = 1.14. From these ratio values, we conclude that the elements whose ratio is greater than zero has no adverse drug reaction. We now eliminate the subset of the population whose NM ratio is less than zero and this subset of population decided to have some adverse drug reactions. These subset of population needed to have some concentrations.

The physicians made decisions based on the Causal Association Rule [1]. It is denoted by $X \xrightarrow{C} Y$, meaning that X potentially causes Y. They defined the support of a CAR, $\text{supp}\left(X \xrightarrow{C} Y\right)$, as the accumulated votes over all sequences. Based on this definition of the support of a CAR, they defined the interestingness measure called causal leverage measure. The support measures for the elements in this sample medical datasets are 0.5, 0.1, 0.3, 0.1, 0.5 respectively. The figure 2 shows the results for the Association rule and the figure 3 shows the results for the NM algorithm. The figure 4 represents the comparison results for the NM algorithm with the existing causal association rule.

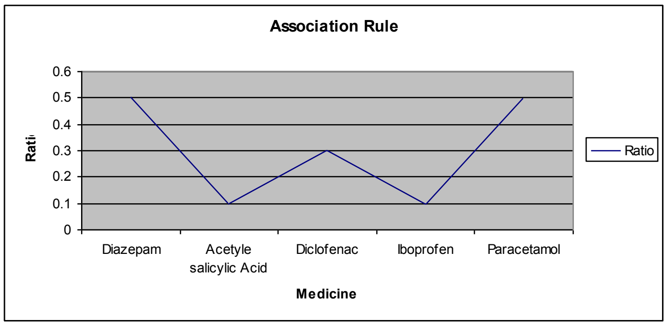

Fig. 2 . Results for the association rule

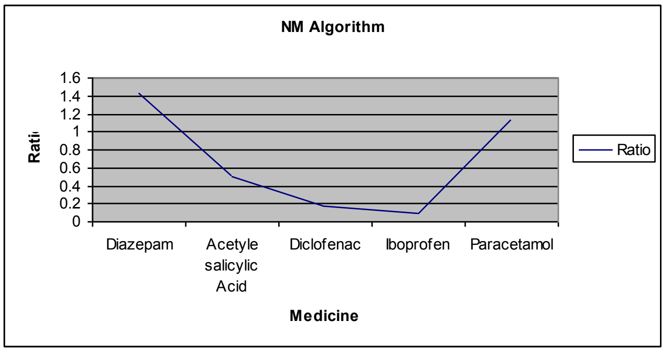

Fig. 3. Results for the NM algorithm

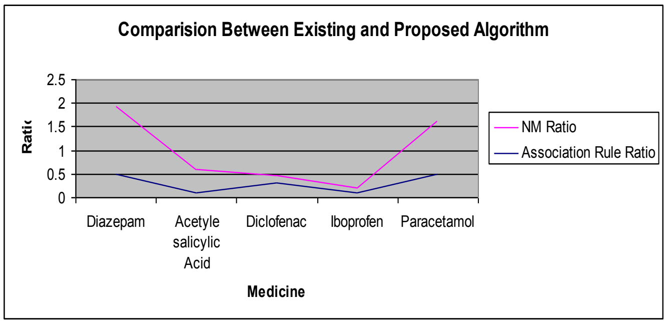

Fig. 4. Comparison between existing and NM algorithm

V. CONCLUSION

A system framework have been developed to achieve better postmarketing surveillance. In this framework, we have developed NM algorithm and it is based on the genetic algorithm. This provide the information that can help people to discover the causality of a type of events and avoid its potential adverse effects. . Users can interact this platform to examine various forms of ADR signals from different view points, by selecting and readjusting parameters of measures of interest.

One of the main problems is that the pharmocovigilance using computer systems is a lack of standard measures for signal detection. This paper presents a preliminary development of ADR detection and analysis and there is much scope for extending research, such as

- This system only discovers drug-ADR and multi-ADRs. The algorithms will be improved to consider about unsupervised learning methods.
- It is planned to include more iterations after finding the fitness functions because iterations reduces some unwanted measure from the results.

REFERENCES

1. Yanqing Ji, Hao Ying, Fellow, IEEE, John Tran, Peter Dews, Ayman Mansour, and R. Michael Massanari, ‘A Method for Mining Infrequent Causal Associations and Its Application in Finding Adverse Drug Reaction Signal Pairs’ IEEE Transactions on Knowledge and Data Engineering, Vol. 25, No. 4, April 2013.
2. Jutamas Tempaiboolkul, School of Engineering and Technology, Asian Institute of Technology, Thailand (2013) ‘Mining Rare Association Rules in a Distributed Environment using Multiple Minimum Supports’.
3. Lian Duan, Mohammad Khoshneshin, W. Nick Street, and Mei Liu (2013) ‘Adverse Drug Effect Detection’ IEEE Journal of Biomedical and Health Informatics, vol.17,No.2.
4. Anni Coden, Daniel Gruhl, Neal Lewis, Michael Tanenblatt, Joe Terdiman (2012) ‘SPOT the drug! An unsupervised pattern matching method to extract drug names from very large clinical corpora’ IEEE Second Conference on Healthcare Informatics, Imaging and Systems Biology.
5. Ji, H. Ying, P. Dews, A. Mansour, J. Tran, R.E. Miller, and R.M.Massanari (2011) ‘A Potential Causal Association

Mining Algorithm for Screening Adverse Drug Reactions in Postmarketing Surveillance' IEEE Transactions on Information Technology in Biomedicine, Vol 15, No.3.

6. Noah Lee, Andrew F. Laine, Jianying Hu, Fei Wang, Jimeng Sun, Shahram Ebadollahi (2011), 'Mining electronic medical records to explore the linkage between healthcare resource utilization and disease severity in diabetic patients' First IEEE International Conference on Healthcare Informatics, Imaging and Systems Biology.
7. Yanqing Ji, Hao Ying, Peter Dews, Margo S. Farber, Ayman Mansour, John Tran, Richard E. Miller, R. Michael Massanari (2010), 'A Fuzzy Recognition-Primed Decision Model-Based Causal Association Mining Algorithm for Detecting Adverse Drug Reactions in Postmarketing Surveillance'.
8. Alexandru Floares, Ovidiu Balacescu, Carmen Floares, Loredana Balacescu, Tiberiu Popa, Oana Vermesan (2010) 'Mining Knowledge and Data to Discover Intelligent Molecular Biomarkers: Prostate Cancer i-Biomarkers'.
9. C. Marinica and F. Guillet (June 2010) 'Knowledge-Based Interactive Postmining of Association Rules Using Ontologies' IEEE Transactions on Knowledge and Data Engineering, Vol 22, No.6.
10. Y. Ji, H. Ying, M.S. Farber, J. Yen, P. Dews, R.E. Miller, and R.M.Massanari (May2010) 'A Distributed, Collaborative Intelligent Agent System Approach for Proactive Postmarketing Drug Safety Surveillance' IEEE Transactions on Information Technology in Biomedicine, Vol 14, No.3.
11. Dat Tran, Wanli Ma, and Dharmendra Sharma (2009) 'Fuzzy Subspace Hidden Markov Models for Pattern Recognition'.
12. H. Jin, J. Chen, H. He, G. Williams, C. Kelman, and C. O'Keefe (2008) 'Mining Unexpected Temporal Associations: Applications in Detecting Adverse Drug Reactions' IEEE Transactions on Information Technology in Biomedicine, Vol 12, No.4.

H.Sankara Vadivu is doing her M.E in Francis Xavier Engineering College at Tirunelveli. She received her B.Tech degree in Information Technology from Sethu Institute of Technology, Madurai in 2012. She is an active member of the Computer Society of India(CSI) and IAENG. Her area of interest s are Data Mining and Data Base Management Systems.

E.Manohar is presently working as Assistant Professor in Francis Xavier Engineering College, Tirunelveli. He completed his B.E in Computer Science and Engineering from Manonmaniam Sundaranar University, Tirunelveli and completed his M.E in Computer Science and Engineering from Francis Xavier Engineering College, Tirunelveli. And he is currently doing Ph.D in Anna University, Chennai

Dr. R. Ravi is an Editor in International Journal of Security and its Applications (South Korea). He is presently working as a Professor & Head and Research Centre Head, Department of Computer Science and Engineering, Francis Xavier Engineering College, Tirunelveli. He completed his B.E in Computer Science and Engineering from Thiagarajar College of engineering, Madurai in the year 1994 and M.E in Computer Science and Engineering from Jadavpur Government research University, Kolkatta. He has completed his Ph.D in Networks from Anna University Chennai. He has 18 years of experience in teaching as Professor and Head of department in various colleges. He published 12 International Journals, 1 National Journal. He is also a full time recognized guide for various Universities. Currently he is guiding 18 research scholars. His areas of interest are Virtual Private networks, Networks, Natural Language Processing and Cyber security.