Effective Analysis of Medical Dataset using Infrequent Causal Association Mining

[1]Ms. M. Saranya M.E.,
[1]Assistant Professor
Department of Computer Science and Engineering
K.S.R.College of Engineering
Tiruchengode, India

[2]R. Nivetha, [3] G. Dhivya,
[4] V. Jivithkumar, [5] A. Kanagasapabathy
[2,3,4,5] U G Students
Department of Computer Science and Engineering
K.S.R.College of Engineering
Tiruchengode,India

Abstract— **The data mining is a process of analyzing a huge data from different perspectives and summarizing it into useful information. Data mining plays a significant role in the field of information technology. The data mining techniques are very useful to make medicinal decisions in curing diseases. The discovered knowledge can be used by the healthcare administrators to improve the quality of service. The mining task that focuses on discovering frequent paramedical from the medical databases is called frequent paramedical mining. Mining infrequent paramedical is a challenging endeavor because there is an enormous number of such medical that can be derived from a given data set. More specifically, the key issues in mining frequent paramedical are: (1) to identify interesting infrequent paramedical patterns (Symptoms and Side Effect) and (2) to efficiently discover them in large paramedical data sets. To get a different perspective on various types of interesting infrequent paramedical, two related symptom are negative action and side effect correlated disease.**

In this paper, focus on the medical data extraction part, particularly, on the very first step of selecting the optimal data mining workflow for automatic classification of paramedical dataset. Medical record is automatically group related classification using NBC model. In this classification frame work, focuses on correlations between the symptoms in the local area are maximized while the correlations between the side effect these area are minimized simultaneously.

Keywords — Mining Infrequent Mining, Medical dataset, NBC classification Model.

I. INTRODUCTION

Data mining automates the process of sifting through historical data in order to discover new information. This is one of the main differences between data mining and statistics, where a model is usually devised by a statistician to deal with a specific analysis problem. It also distinguishes data mining from expert systems, where the model is built by a knowledge engineer from rules extracted from the experience of an expert.

The emphasis on automated discovery also separates data mining from OLAP and simpler query and reporting tools, which are used to verify hypotheses formulated by the user.

Data mining does not rely on a user to define a specific query, merely to formulate a goal - such as the identification of fraudulent claims.

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. However, mining these relationships is challenging due to the difficulty of capturing causality among events and the infrequent nature of the events of interest in these applications.

In this thesis work, to best 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. Develop an interestingness measure that incorporates the causalities across all the sequences in a database. Our proposed 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 represent a serious world-wide problem. They can complicate a patient's medical condition or contribute to increased morbidity.

To more effectively mine infrequent causal associations, it is necessary to develop a new data mining framework. The this thesis work is a substantial extension of our previous work where an interestingness measure called causal-leverage was developed on the basis of a computational fuzzy recognition-primed decision (RPD) model previously developed.

In this research work is focus on mining infrequent causal associations. They are significantly extended our previous work in both theories and experiments.

- They are developed and incorporated an exclusion mechanism that can effectively reduce the undesirable effects caused by frequent events. A new measure is named exclusive causal-leverage measure.
- They are 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.

- A compared 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 and did extensive experiments. In our previous work, tested the effectiveness of the causal-leverage measure using a single drug in the experiment.

Mining the causal association between two events is very important and useful in many real applications. It can help people discover the causality of a type of events and avoid its potential adverse effects. However, mining these associations is very difficult especially when events of interest occur infrequently. We have developed a new interestingness measure, exclusive causal-leverage, based on an experience-based fuzzy RPD model.

This measure can be used to quantify the degree of association of a CAR. Moreover, the measure was designed to mask the undesirable effects caused by high-frequency events. They are applied and measure to detect the causal associations between each of the drugs (i.e., enalapril, pravastatin, and rosuvastatin) and 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.

Fuzzy set theory is being used more and more frequently in intelligent systems because of its simplicity and similarity to human reasoning. Therefore to use fuzzy sets in data mining, a mining approach that integrates fuzzy set concepts with the fuzzy casual association rule mining algorithm has been identified. It finds interesting dugs item set and fuzzy association rules in transaction data with quantitative values. The role of fuzzy sets helps transform quantitative values into linguistic terms, which reduces possible drugs item sets in the mining process. They are used in the fuzzy casual association data mining algorithm to discover useful association rules from quantitative values.

The fuzzy mining algorithm first transforms each quantitative value into a fuzzy set with linguistic terms using casual values. The algorithm then calculates the scalar cardinality of each linguistic term on all symptoms using the patient data set. Each attribute uses only the linguistic term with the maximum cardinality in later mining processes, which keeps the number of symptoms and side effect the same as that of the original patient data set attributes. The mining process based on fuzzy counts is then performed to find fuzzy association rules engineer from rules extracted from the experience of an expert.

- The fuzzy algorithm is applied on an encoded temporal database
- The data set casuals has weighted drugs details
- The weighted minimum support is used to calculate the support.
- Thus fuzzy mining leads to identifying association rules in terms of linguistic terms rather than quantitative values.

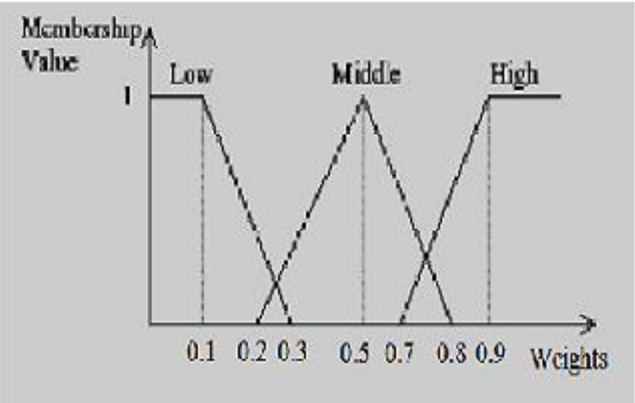


Fig 1.1 Fuzzy Set Algorithms

II.RELATED WORKS

Yanqing Ji and Hao Ying et al [2010] [1] is discovering unknown adverse drug reactions (ADRs) in post-marketing surveillance as early as possible is of great importance. The current approach to post-marketing surveillance primarily relies on spontaneous reporting. It is a passive surveillance system and limited by gross underreporting (<10% reporting rate), latency, and inconsistent reporting. We propose a novel team-based intelligent agent software system approach for proactively monitoring and detecting potential ADRs of interest using electronic patient records. We designed such a system and named it ADRMonitor. The intelligent agents, operating on computers located in different places, are capable of continuously and autonomously collaborating with each other and assisting the human users (e.g., the food and drug administration (FDA), drug safety professionals, and physicians).

Y. Ji, H. Ying, P. Dews, A. Mansour [2011] [2] is early detection of unknown adverse drug reactions (ADRs) in postmarketing surveillance saves lives and prevents harmful consequences. We propose a novel data mining approach to signaling potential ADRs from electronic health databases. Due to the infrequent nature of ADRs, the existing frequency-based data mining methods cannot effectively discover PCARs. We introduce a new interestingness measure, potential causal leverage, to quantify the degree of association of a PCAR. This measure is based on the computational, experience-based fuzzy recognition-primed decision (RPD) model that we developed previously. The potential causal leverage assesses the strength of the association of a drug-symptom pair given a collection of patient cases.

C. Marinica and F. Guillet [2010] [3] is describe the usefulness of association rules is strongly limited by the huge amount of delivered rules. To overcome this drawback, several methods were proposed in the literature such as itemset concise representations, redundancy reduction, and postprocessing. However, being generally based on statistical information, most of these methods do not guarantee that the extracted rules are interesting for the user. Thus, it is crucial to help the decision-maker with an efficient postprocessing step in order to reduce the number of rules. This paper proposes a new interactive approach to prune and filter discovered rules. First, we propose to use ontologies in order to improve the integration of user knowledge in the postprocessing task. Second, we propose the Rule Schema formalism extending the specification language proposed by Liu et al. for user expectations. Furthermore, an interactive framework is designed to assist the user throughout the analyzing task. Applying our new approach over voluminous sets of rules, we were able, by integrating domain expert knowledge in the postprocessing step, to reduce the number of rules to several dozens or less. Moreover, the quality of the filtered rules was validated by the domain expert at various points in the interactive process.

Isabelle Guyon [2009] [4] Machine learning has traditionally been focused on prediction: Given observations that have been generated by an unknown stochastic dependency, the goal is to infer a law that will be able to correctly predict future observations generated by the same dependency. Statistics, in contrast, has traditionally focused on “data modeling”, i.e., on the estimation of a probability law that has generated the data.During recent years, the boundaries between the two disciplines have become blurred and both communities have adopted methods from the other, however, it is probably fair to say that neither of them has yet fully embraced the field of causal modeling, i.e., the detection of causal structure underlying the data.

G. Niklas Norén [2010] [5] is Large collections of electronic patient records provide a vast but still under utilized source of information on the real world use of medicines. They are maintained primarily for the purpose of patient administration, but contain a broad range of clinical information highly relevant for data analysis. While they are a standard resource for epidemiological confirmatory studies, their use in the context of exploratory data analysis is still limited. In this paper, we present a framework for open-ended pattern discovery in large patient records repositories. At the core is a graphical statistical approach to summarizing and visualizing the temporal association between the prescription of a drug and the occurrence of a medical event.

M. Adda, L. Wu [2007] [6] is describe describe here a general approach for rare itemset mining. While mining literature has been almost exclusively focused on frequent itemsets, in many practical situations rare ones are of higher interest (e.g., in medical databases, rare combinations of symptoms might provide useful insights for the physicians). Based on an examination of the relevant substructures of the mining space, our approach splits the rare itemset mining task into two steps, i.e., frequent itemset part traversal and rare itemset listing. They are propose two algorithms for step one, a native and an optimized one, respectively, and another algorithm for step two. We also provide some empirical evidence about the performance gains due to the optimized traversal.

III. METHODOLOGY

Drugs and their associated ADRs have causal relationships. In this section, they examine how to search for potential ADR signal pairs from an electronic patient database using the above exclusive causal-leverage measure. We assume that patient data are stored in relational tables in a database and can be retrieved using database language like structured query language (SQL). These tables are linked through patient identification numbers (PIDs). We also assume that the drug-related data and symptom-related data are stored in two tables called Patient Drug Table and Patient Symptom Table, respectively.

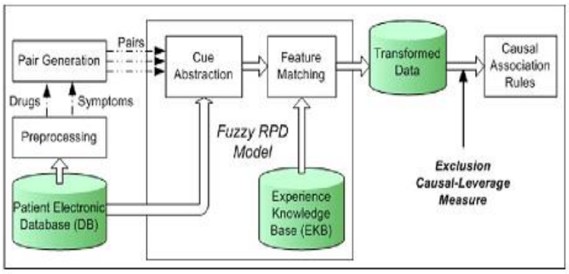


Fig 3.1 Algorithm for mining casual association rules

Fig. 3.1 shows an overall picture of the data mining algorithm. First, preprocessing is needed to get two types of information: 1) a list of all drugs D (d1, d2, ... ,dm) in the database and the support count for each drug; 2) a list of all symptoms S(s1, s2……, sm)in the database and the support count for each symptom. The lists of drugs and symptoms are needed to form all possible drug-symptom pairs whose causal strengths will be assessed. Since a patient database normally only contains a subset of all drugs on the market and a subset of all symptoms, it is necessary to search the Patient Drug Table and the Patient Symptom Table to get the drugs and symptoms covered by the database. Using the discovered d rugs and symptoms in the database (instead of all drugs on the market and all available symptoms) can avoid forming unnecessary pairs, which will reduce the computational complexity of the data mining algorithm. In addition, the support count for each drug or symptom will be used to calculate the exclusive causal-leverage value for related pairs. Hence, their values also need to be computed.

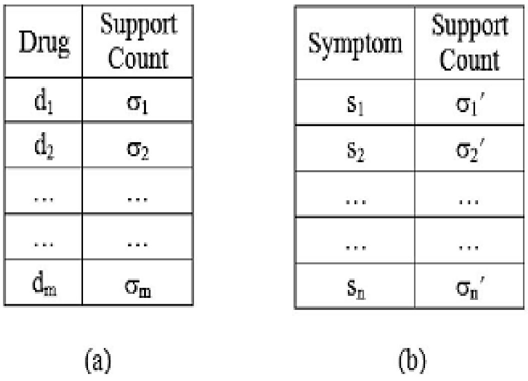
Algorithm 1 shows how to search a database (DB) for the list of drugs and the support count for each drug. The discovered drugs and their support counts are stored in a hash table as shown in Fig a. Initially, the hash table is empty. We use Dk to represent the set of drugs taken by the kth patient pk in the database. For each patient, we first retrieve all the drugs taken the patient. For each of these drugs, we then check whether the hash table contains the drug. If the hash table does not contain the drug, the drug’s support count is set as 1, and both the drug and its support count are added to the hash table. Otherwise, the drug’s support count is increased by 1. After all the patient cases are searched, the hash table is returned. Note that, when calculating the support count for a drug, it is counted only once for one patient case even if the drug appears several times in that patient case. One can see that the computational complexity of this searching process can be affected by the total number of patients N in the database and the average number of drugs taken by a patient. The later is normally determined by the type o f patients in the database. For instance, old patients often have multiple diseases and thus may take multiple drugs either at the same time or different times.

CASUAL ASSOCIATION RULE MINING

Algorithm 1. Searching for drugs and the support count for each drug

```
drugHashTable = null
for each patient Pk ε DB do
    retrieve all the drugs Dk taken by the patient
    for each drug dkl ε Dk do
        if ( drugHashTable:containsKey (dkl) == false do
            σ = a new drug dkl is found and set its
            support count as 1}
        else
            σ = rugHashTable:getValue(dkl) +1update
            support count}
        end if
        drugHashTable:putValue (dkl, σ)
    end for
end for
return (drugHashTable)
```

The algorithm used to search the Patient Symptom Table for a list of symptoms covered by the database as well as the support count for every symptom (Fig b) is similar to Algorithm 1. If users are only interested in mining the potential ADRs of a particular drug or a couple of drugs, the users can specify the drugs of interest. Similarly, the users can also specify the list of symptoms if they want to analyze which drugs can cause the symptoms of interest. In both cases, however, the Patient Drug Table and the Patient Symptom Table still need to be searched in order to get the support count for each drug or symptom.



Drug	Support Count
d_1	σ_1
d_2	σ_2
…	…
…	…
d_m	σ_m

(a)

Symptom	Support Count
s_1	σ_1'
s_2	σ_2'
…	…
…	…
s_n	σ_n'

(b)

Fig 4. 2 Data structure for storing drug/symptoms support counts.

a) Drug hash table b) Symptom hash table

After getting the list of drugs D(d1 ; d2 ;... ; dm) and the list of symptoms S(s1, s2……, sm), the next step is to generate all the possible pairs, each of which represents a CAR. Algorithm 2 shows the process for pair generation and evaluation. Please note that most existing data mining methods mine all interesting association rules that combine all possible events or items in a database. We are only interested in mining drug-symptom pairs that can be easily generated, given D and S. The interested in other patterns like drug-drug pairs, symptom-symptom pairs or combinations of multiple drugs and symptoms. Thus, our algorithm generates a much fewer number of candidate rules, which implies much less complexity. The complexity of this pair generation process is O (m n) where m and n are the number of drugs and symptoms, respectively. In addition, since we are interested in mining infrequent patterns, it is inappropriate to prune pairs using the support measure (i.e., support > minsupp).

However, in post marketing surveillance, a signal pair generated by a data mining method is generally not considered as valid if only one or two patient cases contain the pair. Therefore, we utilize a minimum support count mincount = 3 (instead of minsupp) to further reduce the number pairs that will be evaluated. Specifically, we retrieve the PIDs of those patient cases that contain both the drug and the symptom of each pair. This is done by sending a SQL statement to query the data base.

Modern database management systems can utilize optimization techniques like index to speed up this process. If the number of PIDs that support a pair is greater than or equal to mincount (Line 4 of Algorithm 2), the pair is evaluated. To evaluate the strength of the causal association of the pair, its causal-leverage value is first computed. Then, the pair is reversed. That is, the pair < di ; sj > becomes < sj ; di > The reverse causal-leverage value of pair < di ; sj > is equal to the causal-leverage value of its reversed pair < sj ; di > After that, the exclusive causal-leverage value of the pair is computed by subtracting its reverse causal-leverage value from its normal causal-leverage value.

CANDIDATE RULE GENERATION

Algorithm 2. Pair (Candidate Rule) Generation and Evaluation

```
1: for each drug di D do
2:     for each symptom sj  S do
3:         retrieve PIDs that support pair < di , sj >
           from database
4:         if (count (PIDs)>=mincount) then
5:             value1 = causal-leverage        (di , sj,PIDs)
6:             value2 = reverse causal-leverage (di , sj) =
               causal-leverage (sj , di, PIDs)
7:             exclusive causal-leverage value = value1 - value2
8:             output pair < di , sj > and its exclusive
               causal-leverage value
9:         end if
10:    end for
11: end for
```

Algorithm 3 shows how to compute the causal-leverage value of a general pair between event X and Y. Both X and Y could be either drug event or symptom event. First, the drug or symptom hash table is searched in order to get the support count for event Y. Then, for each PID that supports the pair, a process called cue abstraction is used to extract a set cue values

V from the related patient case. Specifically, a list of drug start dates and a list of symptom dates are retrieved from the Patient Drug Table and the Patient Symptom Table, respectively. Note that, since each patient case records the patient's history for a long period of time, the same drug may be prescribed many times and thus multiple drug start dates may exist within one patient case.

Similarly, there may exist multiple symptom dates for the symptom within the same patient case. Therefore, we must compare each start date of the drug with each symptom date of the symptom in order to obtain interested temporal patterns for the pair within the case. The complexity of this process is O(Ld * Ls), where Ld and Ls are the length of the list of drug start dates and the length of the symptom dates, respectively. Ld and Ls often depend on the characteristics of the patient. For example, a patient with chronic diseases tends to take the same drug for many times and have repeated symptoms. After getting the temporal patterns, cue values for temporal association, rechallenge, and dechal-lenge are derived from these patterns using fuzzy rules. Note that, in order to make our algorithm more generic, the cue other explanations was not utilized in this study since it is drug dependent. Readers are referred to our previous paper for specific fuzzy rules that were used in this process.

NBC CLASSIFICATION

Algorithm 3. Procedure causal-leverage (X, Y, PIDs)

```
search drug/symptom hash table to get support count for Y –σy
for each PID that supports the pair do
    V =cue-abstraction (PID)
    SV ={sv|sv =SG(V,V')^V'є EKB}
    {calculate similarity values}
    SV'=normalization(SV)
    C<X;Y > =weightedSum (SV';W) {W: weights calculated by Eq (8)}
    if (C<X;Y > > 0) then
        accumulatedVotes+=C<X; Y >
        contributionCases ++ {number of cases whose votes are greater than 0}
    end if
end for
$supp(X \overset{c}{\rightarrow} Y) = accumulatedVotes/N$
$supp(X \overset{c}{\rightarrow}) = contributionCases/N$
supp ( à Y) = σy / N
$\textbf{return}\ (supp(X \overset{c}{\rightarrow} Y) - supp(X \overset{c}{\rightarrow})_x supp(\rightarrow Y))$
```

After the set of cue values V of the pair are extracted from the related patient case, similarity values are computed between V and each set of cue values V0 in the experience knowledge base. This step is also called feature matching in the Fuzzy RPD model. These similarity values are then normalized so that they are transformed to the format shown in Table 2. After that, the degree of causality C<X, Y > within the current patient is computed. If C<X, Y> is greater than 0, it is added to the accumulated votes and the number of cases whose votes are greater than 0 is increased by one. After the above computation is done for all the supporting cases of the pair, the causal-leverage value of the pair is computed (refer to (10)) and returned.

STEP FOR PROPOSED ALGORITHMS

1. Add Symptom Details (Id, Name).
2. Add Side Affect Details (Id, Name).
3. Add Tablets (Tablet Id, Name).
4. Add Cue Type (Cue Type Id, Name).
5. Add Causality Degree with Membership Value [0.1 to 1.0] (Tablet Id, Side Affect Id, Cue Type Id and Membership Value).
6. Get Patient Profiles with Unique Patient Id.
7. Get Patient Symptoms (Patient Id, Entry Date, Symptom Id and Tablet Taken) with Multiple Records in various Dates.
8. Get Patient Side Affects (Patient Id, Entry Date, Tablet Taken, Side Affect Id and Cue Type Occurred) with Multiple Records in various Dates.
9. Find Drug and Support Count for the selected patient.
10. Generate Pair (Candidate Rule) with Given tablets and Side affects along with Exclusive value).
11. Generate membership of each causality category to find percentage of side affect occurred to all patients for the taken tablets.

A proposed system 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. They are developing an interestingness measure that incorporates the causalities across all the sequences in a database. Our thesis 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 Systematic methods for the detection of suspected safety problems from spontaneous reports have been studied and practically implemented.

First, preprocessing is needed to get two types of information:

- A list of all drugs D (d1, d2, ... ,dm) in the database and the support count for each drug with numerical value assign.
- A list of all symptoms S(s1, s2……, sm)in the database and the support count for each symptom for each drug with numerical value assign Process:

The lists of drugs and symptoms are needed to form all possible drug-symptom pairs whose causal strengths will be assessed by several causal numerical value data sets.

1. Since a patient database normally only contains a subset of all drugs on the market and a subset of all symptoms, it is necessary to search the Patient Drug Table and the Patient

Symptom Table to get the drugs and symptoms covered by the database.

2. Using the discovered d rugs and symptoms in the database (instead of all drugs on the market and all available symptoms for causal numerical value data sets) can avoid forming unnecessary pairs, which will reduce the computational complexity of the data mining algorithm.

3. In addition, the support count for each drug or symptom will be used to calculate the exclusive causal-leverage value for related pairs. Hence, their values also need to be computed

```
//NBC RULE MINING: PROPOSED PROCESS //
Notation:
Symptom spm
Side Affect saf
Tablets tab
Cue type cty
Causality Degree ∞
Causality Values pcv
Patient Record pada
Multi causality category mcc
Input: spm, saf, tab, cty and pada,i, j, n
Output: Effective mcc
For i=1; i<= n; i++
//where n total Number of Patient Records
do
pada □spm
pada□saf
pada□tab
set cty values
cty□pada based
set ∞ values
∞□pada based
while
End
Calculate pvc();
// Finding causality with membership values
If pada==pvc
Calculate support_count()  //Finding support Count Values
Else
Return pvc   //Return causality value
End
 Generate pair wise pvc in pada data sets   //Candidate Rule Generation
    For j=i; j<=pada; j++
    If pada==pvc then
        Calculate mcc values   // Finding MCC Values
Return mcc
Else
Sigle pair ∞ values
End
```

IV. EXPERIMENTAL RESULTS

Table 5.1 is describing the cue value analysis in local and global similarity value in proposed system. The table contains cue value 1 and cue value 2, cue value between similarity value and cue value between global similarities values are shown below.

S.NO	Cue Value1	Cue Value2	Local Similar	Global Similar
1	Very Likely	Very Likely	1	0.435
2	Very Likely	Probable	0.249	0.560
3	Very Likely	Possible	0.197	0.585
4	Very Likely	Unlikely	0.142	0.564

Table 5.1 NBC Cue Value Analysis [Cue 1, Cue 2]

Fig 5.1 is describing the cue value analysis in local and global similarity value in proposed system. The figure contains cue value 1 and cue value 2, cue value between similarity value and cue value between global similarities values are shown below.

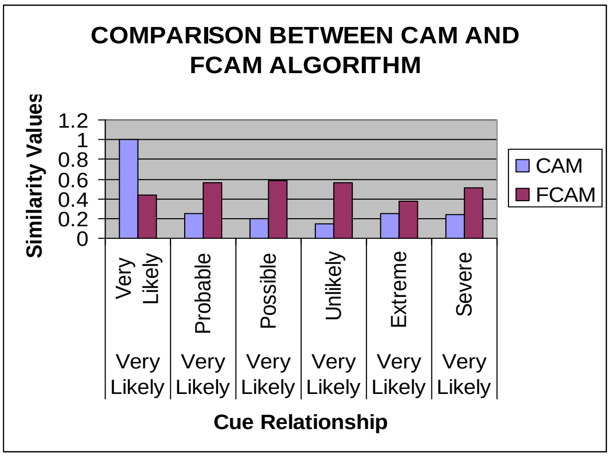


Fig 5.1 Comparison between CAM and FCAM

[Cue 1 & Cue 2]

Table 5.2 is describing the cue value analysis in local and global similarity value in proposed system. The table contains cue value 3 and cue value 4, cue value between similarity value and cue value between global similarities values are shown below.

S.NO	Cue Value 3	Cue Value 4	Local Similar	Global Similar
1	Probable	Very Likely	0.249	0.465
2	Probable	Possible	0.143	0.544
3	Probable	Unlikely	0.199	0.486
4	Probable	Extreme	0.239	0.367
5	Probable	Severe	0.227	0.506
6	Unlikely	Possible	0.201	0.638

Table 5.2 NBC-Cue Value Analysis
[Cue 3, Cue4]

Fig 5.2 is describing the cue value analysis in local and global similarity value in proposed system. The figure contains cue value 3 and cue value 4, cue value between similarity value and cue value between global similarities values are shown below.

- It is found that combination of drugs can be taken for cue identification and ADR.
- The relationship can be maintained between combination of drugs and combination of symptoms at a give time.
- Unlike existing system, 'N' degrees of causality (Likely, Probable, Possible, Unlikely and more) can be given.
- Different types of tablets make different impact on patients.
- It can help people discover the causality of a type of events and avoid its potential adverse effects.
- Mining the associations in drug consumption and its effect is very difficult especially when events of interest occur infrequently. But the proposed system solves the problem using Association Rule Mining.
- A new methodology named 'exclusive causal-leverage' is developed to solve the above mentioned problem.
- The thesis applied this measure to detect the causal associations between combination of drugs (i.e., Symptom-Drug and Drug-SideAffect occurred).
- The measure was designed to mask the undesirable effects caused by high-frequency events.
- The algorithms could effectively make known ADRs rank high among all the symptoms in the database.

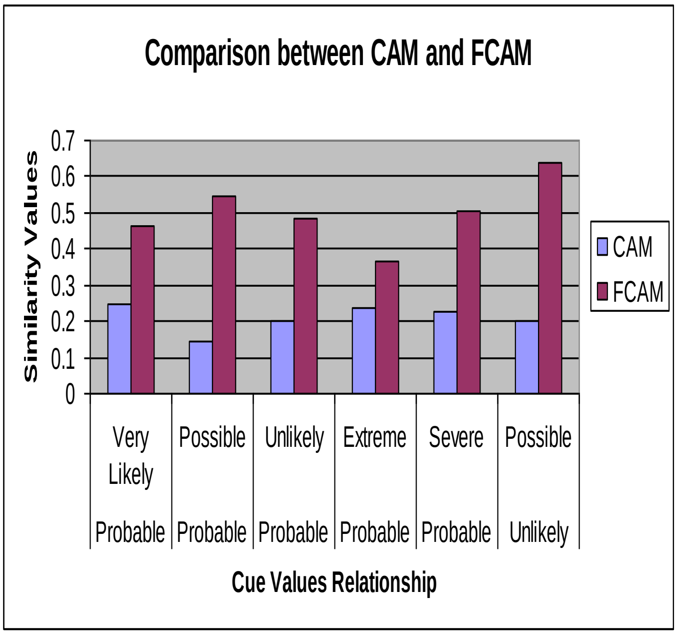


Fig 5.2 NBC-Comparison between CAM and FCAM
[Cue 3 & Cue 4]

V. CONCLUSION

Mining the causal association between two events is very important and useful in many real applications. It can help people discover the causality of a type of events and avoid its potential adverse effects. However, mining these associations is very difficult especially when events of interest occur infrequently. The research work contains a new interestingness measure, exclusive causal-leverage, based on an experience-based fuzzy RPD model. This measure can be used to quantify the degree of association of a CAR. Moreover, the measure was designed to mask the undesirable effects caused by high-frequency events. We have applied this measure to detect the causal associations between each of the drugs. 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. The project uses <X, Y> and $C_{<X, Y>}$to represent a pair of events and the degree of causality of the pair in a sequence, respectively. In addition, a drug event X may be modified to a list of drug combination and Y a single symptom event may be modified to a list of symptom events and the similarities are found out. Likewise, 'N' number of degrees is given for Cue set. The project contains the following features i) Combination drugs are taken for cue identification and ADR ii) Relationship between combination of drugs and combination of symptoms at a give time is considered iii) N' degrees of causality (Likely, Probable, Possible, Unlikely and more) can be given.

At present the research work uses data from a relational (medical) database composed of several related tables. Although designing efficient NBCstatements and data models to handle large volumes of data in a relational database is beyond the scope of this project, the future study will consider developing an efficient algorithm suitable for relational databases and analyze its complexity and efficiency. In addition, multiple symptoms occurred for combination of drug consumption is also to be studied.

V. REFERENCES

[1] 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 Postmarketing Drug Safety Surveillance," IEEE Trans. Information Technology in Biomedicine, vol. 14, no. 3, pp. 826-837, Dec. 2010.

[2] 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 Postmarketing Surveillance," IEEE Trans. Information Technology in Biomedicine, vol. 15, no. 32, pp. 428-437, May 2011

[3] C. Marinica and F. Guillet, "Knowledge-Based Interactive Post mining of Association Rules Using Ontologies," IEEE Trans Knowledge and Data Eng., vol. 22, no. 6, pp. 784-797, June 2010.

[4] Guyon, D. Janzing, and B. Scho ¨ lkopf, "Causality: Objectives and Assessment," JMLR Machine Learning Research Workshop and Conf. Proc., vol. 6, pp. 1-42, 2010.

[5] G.N. Nore ´n, J. Hopstadius, A. Bate, K. Star, and I.R. Edwards, "Temporal Pattern Discovery in Longitudinal Electronic Patient Records" Data Mining and Knowledge Discovery, vol. 20, pp. 361 -387, 2010.

[6] M. Adda, L. Wu, and Y. Feng, "Rare Itemset Mining," Proc. Sixth Int'l Conf. Machine Learning and Applications, pp. 73-80, 2007.

[7] J. Talbot and P. Waller, Stephens' Detection of New Adverse Drug Reactions, fifth ed. John Wiley & Sons, 2004

[8] L. Hazell and S.A.W. Shakir, "Under-Reporting of Adverse Drug Reactions - A Systematic Review," Drug Safety, vol. 29, pp. 385-396, 2006

[9] P.-N. Tan, M. Steinbach, and V. Kumar, Introduction to Data Mining. Addison-Wesley, 2005.

[10] G.M. Weiss, "Mining with Rarity: A Unifying Framework," ACMSIGKDD Explorations Newsletter, vol. 6, pp. 7-19, 2004.

[11] 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.

[12] Jiawei Han and Micheline Kamber, "Data Mining Concepts and Techniques",Elseveir.

[13] L.T. Kohn, J.M. Corrigan, and M.S. Donaldson, To Error is Human:Building a Safer Health System. Nat'l Academy Press, 2000

[14] Y. Ji, H. Ying, P. Dews, A. Mansour, J. Tran, R.E. Miller, and R.M.Massanari, "A Potential Causal Association Mining Algorithm forScreening Adverse Drug Reactions in Postmarketing Surveillance," IEEE Trans. Information Technology in Biomedicine, vol. 15, no. 32, pp. 428-437, May 2011.

[15] Y. Ji, R.M. Massanari, J. Ager, J. Yen, R.E. Miller, and H. Ying, "A Fuzzy Logic-Based Computational Recognition-Primed Decision Model," Information Science, vol. 177, pp. 4338-4353, 2007.