Temporal and Causal Relations on Evidence Theory: an Application on Adverse Drug Reactions

Luiz A. P. A. Ribeiro, Ana Cristina B. Garcia, Paulo S. M. Santos

Abstract

The use of big data and information fusion in electronic health records (EHR) allowed the identification of adverse drug reactions(ADR) through the integration of heterogeneous sources such as clinical notes (CN), medication prescriptions, and pathological examinations. This heterogeneity of data sources entails the need to address redundancy, conflict, and uncertainty caused by the high dimensionality present in EHR. The use of multisensor information fusion (MSIF) presents an ideal scenario to deal with uncertainty, especially when adding resources of the theory of evidence, also called Dempster–Shafer Theory (DST). In that scenario there is a challenge which is to specify the attribution of belief through the mass function, from the datasets, named basic probability assignment (BPA). The objective of the present work is to create a form of BPA generation using analysis of data regarding causal and time relationships between sources, entities and sensors, not only through correlation, but by causal inference.

Keywords: Information Fusion; Dempster & Shafer Theory; Machine Learning; Adverse Drug Reactions; Eletronic Health Records

1 Introduction

Data fusion is defined as the integration of data and knowledge from many sources (Castanedo 2013). In MSIF, data from different sensors are combined to provide a robust description of a situation of interest (Durrant-Whyte and Henderson 2016).

A literature review was conducted and the expressive use of DST in Artificial Intelligence (AI) was identified with strong growth over the last five years owing to the combination of machine learning (ML) techniques. This review identified a research opportunity related to one of the pillars and challenges of the DST, which is to specify the attribution of belief through the mass function, from the datasets. This assignment is called the basic probability assignment (BPA). In (Gordon and Shortliffe 1984) BPA is defined as the impact of each possible subset of belief in the sample space, and the number of possible subsets defined by two is high to the number of 43 distinct elements of the domain under evaluation . Assigning the BPA from the dataset is not a simple task.

One of the aspects observed in the literature review was the precarious use of data analysis from the temporal and causal point of view, or even dependence between sources and sensors. (Pei et al. 2017) observe the importance of analyzing time aspects in pre-processing. In (Zheng 2015), it is addressed that the probabilistic dependency-based data fusion method uses graph structure. The object of study of this research is care in the form of BPA generation for the use of DST. The main objective of this work is the analysis of data regarding causal and time relationships between sources, entities and sensors, not only through correlation, but by causal inference. The hypothesis is that based on heterogeneous sources, causative and time factors are identified from the data set in an MSIF context, which allows the BPA to be allocated in the construction of evidence of a DST discernment framework, respecting uncertainty.

The contribution of this article is the elaboration of an experiment that uses DST, performing a pre-processing with temporality and causality analysis.

The remainder of this paper is organized as follows. In section 2 fundamental topics such as DST, and temporal-series are presented. Section 3 presents the ADR extraction by means of EHR. Section 4 presents the preliminar results and section 5 presents a discussion. In Section 6 we discuss future work and finally concluding remarks are presented.

2 Foundations

2.1 Dempster and Shafer Theory

In DST there is a fixed set of N mutually exclusive and exhaustive elements, called the framework of discernment. Be it a set, indicated by $\Theta = \Theta_1, \Theta_2, \Theta_i, , \Theta_N$.

The Θ share set, represented by $P(\Theta)$, consists of all subsets of Θ. This involves $\emptyset$, the empty set , $P(Theta), \emptyset$ and the Θ itself. the 2^N composite set of Θ elements

$$P(\Theta) = \{\, \emptyset, \Theta_1, , \Theta_N, \{\Theta_1, \Theta_2\}, , \{\Theta_1, \Theta_2, \Theta_i\}, , \Theta\}.$$

Element A represents any of the elements that make up the set of $P(\Theta)$ parts. Mass m (A) represents how strongly the evidence supports A. When m (A) ¿ 0, A is called the focal element of the mass function.

Given evidence, the mass assigned to each element of P(Θ) is equivalent to an indicative value of the belief assigned to it. DST defines this function as mass m, called

Basic Probability Assignment (BPA), which gathers the following properties:

$$P(\Theta) \rightarrow [0,1], A \rightarrow [m(A)].$$

meeting the following conditions:

$$m : 2^{\Theta} \rightarrow [0,1](1)$$
$$m(\emptyset) = 0(2)$$
$$\sum A \in P(\Theta)m(A) = 1(3)$$

(1) Indicates that all subsets of Θ is assigned a belief value between 0 and 1;

(2) It means that a belief deposited in the empty set is always zero; And

(3) That the sum of all assigned values must be one.

2.2 Causality Analysis

Adverse reactions and symptoms caused by diseases To more accurately measure the probability of presenting a symptom or adverse reaction when taking a drug, it is necessary to isolate from the analysis the effect that the disease itself causes on the reaction or symptom. To this end, pearson's Chi-square Test was used, modeling the hypotheses:

H 0 : The disease causes the reaction/symptom vs. H 1 : The disease does not cause the reaction/symptom

If we assume that disease D causes the R reaction, then the reaction/symptom occurs in the same way, regardless of the medication administered to the patient. We then have a multinomial distribution with a uniform parameter (1 / k, ..., 1 / k) where k is the total of medicines and sum P (1 / k, ..., 1 / k) = 1 . The null hypothesis can then be specified in terms of a distribution such as:

$$H_0 = p = (1/k, ..., 1/k) vs H_1 = p \neq (1/k, ..., 1/k)$$

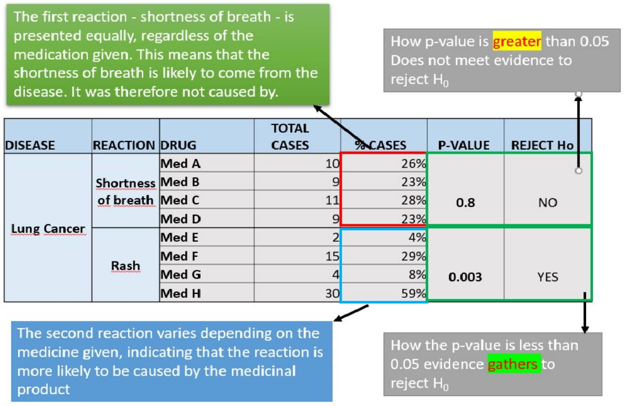
To perform the proposed hypothesis test, the data were segmented by disease and reaction and Pearson's 2 (Chi-square) test was applied. These combinations of diseases and reactions with a p-value ≤ 0 were rejected at the level of 5% significance 0.05 . Figure 1 displays a simulation of how Pearson's test works.

3 ADR from EHR using DST

3.1 Introduction

The most important record in EHR is the clinical note. Medications can be indicated to the patient undergoing outpatient treatment, in the office, by issuing prescriptions. In the case of the patient being hospitalized, undergoing treatment or receiving new dosages of medications, frequency, volume, type of administration and dosage are usually changed, characterizing contextual information and increased risk of ADR.

Clinical Notes and Knowledge Bases The structure for ADR analysis is formed from a domain model, which relates diseases, medications and adverse reactions. A knowledge bases were also used an ontology of adverse effects (Cai et al. 2015)), to search for known reactions, referred to in this project as medicine package leaflet*MPL*.



The first reaction - shortness of breath - is presented equally, regardless of the medication given. This means that the shortness of breath is likely to come from the disease. It was therefore not caused by.

How p-value is greater than 0.05 Does not meet evidence to reject H_0

DISEASE	REACTION	DRUG	TOTAL CASES	% CASES	P-VALUE	REJECT Ho
Lung Cancer	Shortness of breath	Med A	10	26%	0.8	NO
		Med B	9	23%		
		Med C	11	28%		
		Med D	9	23%		
	Rash	Med E	2	4%	0.003	YES
		Med F	15	29%		
		Med G	4	8%		
		Med H	30	59%		

The second reaction varies depending on the medicine given, indicating that the reaction is more likely to be caused by the medicinal product

How the p-value is less than 0.05 evidence gathers to reject H_0

Figure 1: Multinomial distribution frame - Pearson

3.2 Proposed Model

- Tokenization: Promotes word separation by neglecting punctuation and accent characters.
- Stop Words: Provides a list of words to be ignored in the process of analysis such as: articles, prepositions and pronouns.
- Steamming: Reduces each word of the lexicon and gives the terms to be compared (normalization of words).

NLP techniques are used, such as tokenization, stop words, stemming and bag of words. The ADR candidate records are selected by methods that scan unstructured texts in the clinical notes. Terms, tags ans entities that can be drugs and reactions or symptoms for further analysis with medicine package leaflet are selected for analysis. The proposed model implemented four functionalities:

1.-Causality Analysis. Incorporation of estimates obtained by Causal dependence in the calculation of belief functions. Initially, a study was done with Pearson Chi-square test to study dependence between reactions and diseases, avoiding attributing such reactions to medications.

2.- BPA probability estimation with basic probability prior based on frequency.

3- Analysis of time factors - use time series analysis, through linear regression obtained from the sensor of pathological laboratory tests. These tests report results of indicators such as urea, glucose, creatinine, leukocytosis, among others. The analysis aims to evaluate the impact of a given drug on the trend of that particular marker, in a period of time.

4 - credibility index - A PCA method was used to generate a linear combination, which was used as a credibility index.

3.3 Sensors and data sources

The heterogeneity and multimodality of sources bring with them a high degree of uncertainty, including redundant and ambiguous sources of information.

Figure 2 demonstrates the block diagrams of the fusion information process, where orange circles are represented by the 4 data sources that materialize by five sensors:

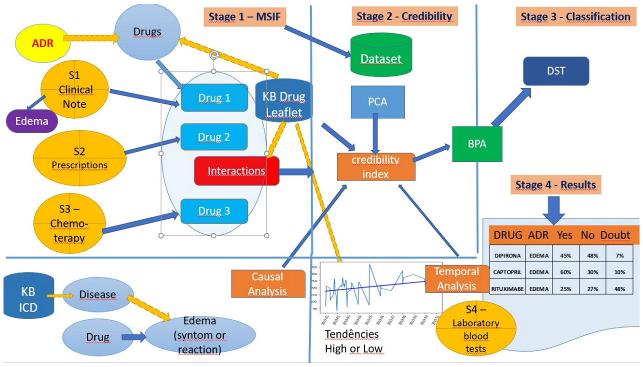


Figure 2: Block diagram of fusion information process

- Clinical Note - from a source of examination of the patient, is presented in an unstructured format that requires the use of NLP techniques for extraction of medications and symptoms that may indicate ADR.
- Prescription - from a source of Outpatient Prescriptions, gathers the dispensations of prescription for patients who are at home and outpatient clinic, who are not hospitalized in the hospital;
- Medical Prescription - from the source of Hospitalizations, gathers the prescriptions of medicines for hospitalized patients.
- Chemotherapy - from the sensor that gathers data on patients undergoing chemotherapy;
- Pathological Tests - With varied frequency, are obtained from the sensors of laboratory tests, the most usual being blood and urine tests. They report in structured information the reference values of the result of the markers.

ADR Base General features and volumes

The general characteristics of the database are described below.

- a) The database contains records that relate reactions and medications that have been administered to patients since 2019.
- b) The database consists of 81,740 such records and 11 columns describing, among other things, patient identification, date of birth, gender, clinical department, reaction and administered medications.
- c) The records reported in the database correspond to 5,937 patients treated in this period.

Model for Analysis of Pathological Examinations The methodology used is based on linear regression models and hypothesis tests, using F-Test.

Figure 3 demonstrates a temporal analysis of the impact of the drug under examination through a linear regression graph:

- 1.- An applied drug is selected, the medical history of a patient who received this treatment and a particular parameter(marker) of the blood test, for example, ureia.

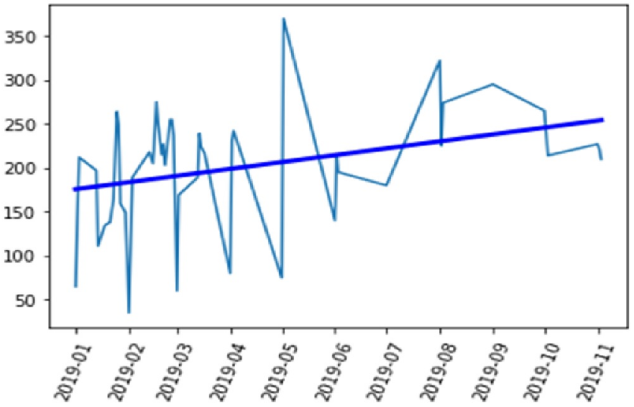


Figure 3: Linear Regression - temporal analysis of the impact of the drug under examination

- 2.- A linear regression model is adjusted to determine the trend of the monitored value during the application of the drug in a period of 30 days;
- 3.- To test the effect of the drug on the value of the study, a hypothesis test is performed; Ho: The medicine has no effect on the blood marker H1: The medicine has an effect on the marker;
- 4.- For each blood test is determined the probability of presenting variations up or down as an effect of the application of a drug.

Credibility Index The following criteria were established during this study, and will be reviewed throughout the research in order to assign measures on the credibility of an ADR. The functions intended to quantify these parameters are formally detailed.

- 1.- EFFECT SIZE : The size of the effect has been measured to consider that a ADR in which several drugs interacted is less believable than one in which the effect is isolated. Be n the number of medications that patient i took, then:

$$EffectSize_i = \frac{max(n) - n}{max(n)}$$

- 2.- PVALOR C : The p-value (*p-value*) obtained with pearson's test was used to quantify the effect of the disease on AMR. For each patient, a value of 1-p was calculated as a measure of ADR credibility when considering the effect of the adverse reaction.
- 3.- IndicatorBula: If the ADR potential is contained in the Bula base, the record is marked with the boolean value one. Otherwise, it's zero. This function is essentially an indicator boolean variable. This function is essentially an indicator variable.

4 Preliminary Results

Initially, 63,000 clinical notes were recovered and about 8.7% mentioned drugs or symptoms, with 5,500 records se-