Discovering Associative Patterns in Health Care Data

Diego de Castro Rodrigues[1,2] {0000-0001-8396-1947}, Vilson Siqueira[1,2] {0000-0003-0843-2839}, Fabiano Tavares[1] {0000-0003-2835-3062}, Márcio Lima[2] {000000032782386X}, Frederico Oliveira[2] {0000-0002-5885-6747}, Lucas Osco[3] {000000020258536X}, Wilmar Junior[3] {0000-0002-1876-6907}, Ronaldo Costa[2] {0000-0003-1892-9080}, and Rommel Barbosa[2] {0000-0002-2638-7026}

[1] Federal Institute of Tocantins, Palmas, Brazil,
diego.rodrigues@ifto.edu.br,
http://www.ifto.edu.br
[2] Federal University of Goiás - UFG
Goiânia, Brazil
[3] University of Western São Paulo - UNOESTE
São Paulo, Brazil

Abstract. Health care has several knowledge discovery techniques. Among them are association rules, which provide quick access to standards. However, classic algorithms can generate many patterns or fail to identify rare cases relevant to healthcare professionals. This study identified asymmetric associative patterns in health-related data using the Health Association Rules (HAR) algorithm. We use a combined strategy of six metrics to filter, select, and eliminate contradiction steps to find patterns and identify possible rare cases. The proposed solution uses adjustment mechanisms to increase the quality of standards with knowledge of the health professional. The HAR assists health researchers and decision support systems. A survey of 597 studies identified the primary needs and problems of associative patterns in the health context. The HAR identifies characteristics with the highest cause and effect relationship. The experiments were carried out on 13 datasets, where we identified the most pertinent patterns for the datasets without losing relevant knowledge.

Keywords: Medical Data Mining, Association rules, Health Care, Asymmetric Association Rule, Data Mining

1 Introdution

Numerous diseases have become obsolete with the development of medical technologies. Data analysis related to health care has combined medical knowledge with data mining technologies [8, 11, 4]. These studies indicate that it is vital to ensure safe analysis of patterns discovered in medical data. However, with the continuous growth in the volume of information and analysis techniques (Machine learning, Data Mining, Medical Data Mining), an imbalance has emerged

among data demand, analysis capacity, and pattern discovery for specific contexts in the dataset of health. Therefore, data mining techniques are becoming more generic to analyze any data, leading to dissatisfaction in finding specific patterns and higher costs of time and knowledge to identify trends.

The association rules demonstrate simplicity to understand the standards obtained in fields such as engineering [10], recommendation systems [6], and clinical diagnostics [5]. The discovery of associative patterns is one of the main tasks of data mining, with emphasis on the use of association rules based on the **Apriori model** in health care [3]. Elham Buxton [2] presents some of the limitations pointed out in the classic associative rules algorithms, initially proposed by Agrawal [1]. The limitations of traditional algorithms have insufficiency related to the amount of generated standards, redundancies, selection of better rules, and elimination of standards.

This article presents a study conducted on data from a systematic review and experiments on real datasets. We developed an associative analysis algorithm applied to the health care context. The algorithm identifies patterns using probabilistic metrics, pruning, filters, and custom metrics. The objective of this study is to identify asymmetric associative patterns in data related to health care, selecting the associative patterns based on a set of probabilistic asymmetric metrics. Thus obtaining better results in identifying trends, selection, and ranking, valuing the casual relationship, and identifying possible rare patterns. Some highlights of the study: (1) Health Association Rules Algorithm (HAR); (2) Use of alternative metrics to the Support/Confidence model; (3) Identification of rare patterns, and; (4) Application in real databases.

2 Approach of Method

To identify the best strategies for the research problem and build the HAR algorithm, we conducted a systematic review of the literature with articles that applied association rules algorithms to medical data published between the period 2015 to 2019.

The articles' selection was carried out in six bases: Science Direct, PubMed, ACM, Springer Link, IEEE, and Google Scholar. Initially selecting 597 papers and after applying the inclusion and exclusion criteria, and chose 51 studies. The systematic review protocol, as well as its details, are available in their entirety in the supplementary material (https://cutt.ly/3fQQw6E).

Based on the Apriori algorithm, the Health Association Rules (HAR) presents customized processes and adds steps in the original operation of the Apriori algorithm. Figure 1 indicates the levels of the HAR and their interactions.

Process I analyzes the database in the same way as the Apriori algorithm, maintaining its original functioning in this process. Process II, the value of minimum support and minimum confidence is calculated by $\frac{1}{N}$, where N is the number of elements in the database. Thus, the lowest possible value of Support will be assigned to the dataset, thus obtaining the most frequent items and association rules. The upper limit is defined, and these are informed through

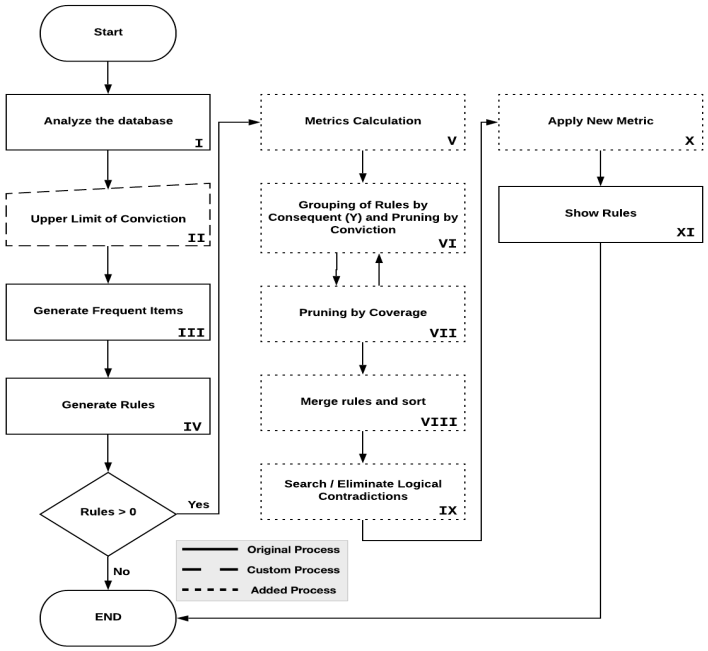

Fig. 1: HAR Operation Flowchart

the user of the HAR algorithm. Based on the *Conviction* metric, which varies between 0 and ∞.

The generation of frequent items, Process III, occurs similarly to Apriori, the difference is in using the minimum values of *Support* obtained automatically in Process II. It is thus generating the set of frequent items.

Process IV uses the set of frequent items from the previous step together with the minimum *Confidence* value to generate the set of association rules. In the original algorithm, only *Confidence* is calculated for each of the rules. However, the V Process is added to the HAR, which calculates the metrics for all association rules.

Pruning by *Conviction* Process VI, uses the upper limit determined in Process II and a lower limit, all association rules that do not respect the established values are eliminated. The pruning process's remaining rules are grouped in Process VI, where all the rules $X \rightarrow Y$ are separated into subsets of rules. The lower limit of *Conviction* is determined by the arithmetic mean of each of the subsets of Consequents.

Each subset is organized by its consequent Y value. They were dividing the set of rules resulting from Process IV and VI into N subsets. The top rules of each of the subsets are selected and checked with the other rules, recursively looking for rules covered by another, until all the rules and subsets are checked.

Based on the study H. Toivonen [9], which proposed the idea of eliminating redundant rules using structural rule coverage, Process VII selects in each subset all rules with the first X antecedents. As an example, the subset [$Dengue = yes$] with the rules [***Pain = strong*** *and PainEyes = yes* $\rightarrow$ *Dengue = yes*], displays two antecedents the second rule with a single antecedent [***Pain = strong*** $\rightarrow$ *dengue = yes*].

The rules with their first identical antecedents are selected and compared for presenting similar information, regardless of the second antecedent X_2 [*PainEyes = yes*] the first antecedent X represented in the two rules by [***Pain = strong***] always presents a relationship with [*Dengue = yes*].

Thus, the rules with the lowest average between *Hyper-Confidence* and *Mutual Information* are eliminated. The first metric ensures that rules are chosen with the least chance of being generated randomly. The second one measures the information gain of the consequent Y provided through the antecedent X.

The search and elimination of logical contradictions were implemented at HAR in the IX Process, seeking contradictions of meaning. The contradiction of meaning is determined in rules [Aches = strong $\rightarrow$ Dengue = yes] and [Dengue = yes $\rightarrow$ Aches = strong]. In both situations, it is not trivial to define which could be eliminated.

The average between the *Confidence* and *Kalczynski* metrics verifies the slope patterns that take into account the relationship of $X \rightarrow Y$ and $Y \rightarrow X$. Applying these metrics to choose the rules in contradictions of sense, selecting the rule with the highest slope value (average), and eliminating the other.

The X Process displays the Difference from Sample Means (DMA), which orders the rules with the greatest asymmetric relationship. The set of final association rules is displayed in the XI process, accompanied by its metrics and ordered. following the pattern of $X \rightarrow Y$ with the measurement values *Hyper-Confidence*, *Mutual Information*, *Imbalance Ratio*, *Kalczynski*, *Gini Index* and *DMA*.

2.1 Evaluation of the Proposed Method

The HAR method consists of four steps. In the first step, the Dataset is provided as an input; The second step (Algorithms), runs the classic Apriori algorithm for the generation of association rules. It also executes the HAR method with a configuration similar to its standard objective metrics; In the third step (Individual Result) compares the rules generated by each step algorithm (Algorithms). The results of each algorithm are analyzed in the step (Analytics) employing objective metrics to understand the reasons for a hypothetically good rule, not being selected in the HAR or the classic algorithm.

Custom Measure The ideal rule is composed of values of objective metrics (*Hyper-Confidence*, *Gini Index*, *Mutual Information*, *Imbalance Ratio*, *Kalczynski*) with the respective default values (0.95, 0.3, 1, 1, 0.6) that together define the rule (orange line) with the greatest potential for the data context. The gray

lines show the behavior of the rules discovered in the HAR. When comparing the (ideal) rule with the HAR rules, it is possible to rank the best rules not in two metrics as in the classic algorithm but a set of six rules.

The customized measurement is performed employing the distance from the rules, which is calculated by the Difference of the Sample Means (DMA) introduced by [7]. We use DMA to calculate the distances from the ideal rule with the dataset rules (Equation 1).

$$DMA = \bar{X}_i - \bar{M} \tag{1}$$

The $\bar{X}_i$ indicates the average of the Hyper-Confidence, Gini Index, Mutual Information, Imbalance Ratio and Kulczynski metrics. The arithmetic mean of the ideal rule is defined by $\bar{M}$, the closer to 0 the DMA, the better the rule ranking in the HAR.

The standard values of the metrics for calculating the DMA are defined to value the relationship $X \rightarrow Y$ such that the asymmetric relationship X and Y is shown. The default values can be customized to meet particulars of the dataset when necessary, DMA is used to rank the rules.

Data Organization The dataset is composed of data from health care, Parkinson's disease, heart disease, physiological complexity, mental health, and frequency of disorders. The databases chosen for the experiments in this study aim to diversify the tests, including related unidentified data the health (Table 1).

Table 1: Datasets

Id	Domain	Instance	Attributes	Name	Source
01	Clinical Data	24978	7	eICU Collaborative	PhysioNet
02	Intensive therapy	1761	5	MIMIC-III Clinical	PhysioNet
03	Parkinson's disease	16	52	Parkinson Disease	PhysioNet
04	Pharmacology	1934	30	CiPA ECG	PhysioNet
05	Oxygen Saturation	36	6	Oxygen Saturation	PhysioNet
06	Physiology	196	21	Tai Chi, Physiological	PhysioNet
07	ECG	4211	33	ECG Effects	PhysioNet
08	Electrophysiological	5232	32	ECG Ranolazine	PhysioNet
09	Heart disease	303	14	Heart Disease UCI	Kaggle
10	Heart disease	1025	14	Heart Disease Dataset	Kaggle
11	Disorders	1259	27	Mental Health	Kaggle
12	Healthcare	43400	12	Stroke Data	Kaggle
13	Healthcare	649	33	Student Alcohol	Kaggle

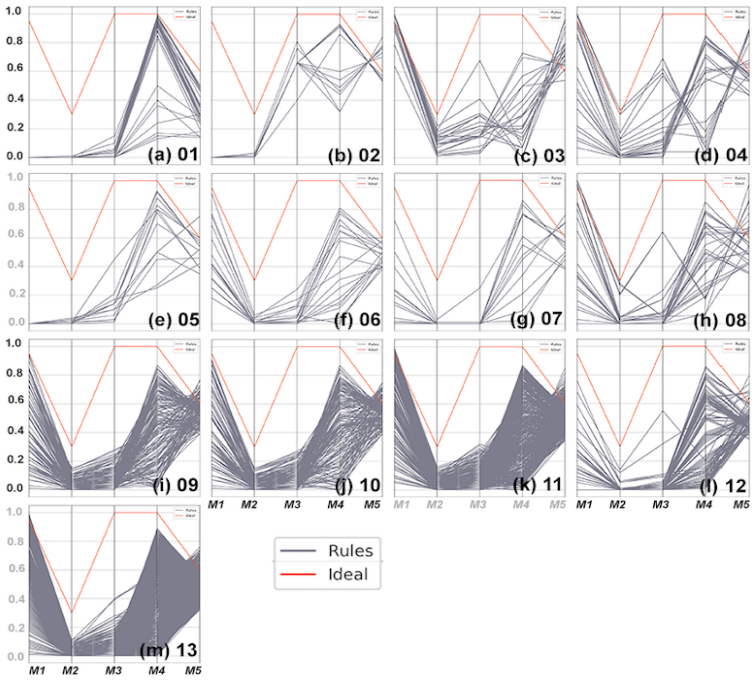

Fig. 2: HAR experiments

acronyms: M1 - *Hyper-Confidence* , M2 - *Gini Index* , M3 - *Mutual Information*, M4 - *Imbalance Ratio*, M5 - *Kalczynski* .

3 Results and discussions

We present the behavior of the patterns of the datasets of Table 1 individually. The datasets (01, 02, 05, 07, 08, and 12) stand out for presenting rules with a low value of *Hyper-Confidence*. This behavior is justified by low Support values found in real datasets (Figura 2).

We believe that the results presented through the execution of the HAR are related to its harmonious functioning with different metrics to validate relations between $X \rightarrow Y$. We present a reduced number of rules compared to the classic Apriori (Figure 3).

When a dataset presents values close to zero in most of the metrics used, it indicates that the dataset in question is not appropriate for associative patterns. The datasets (01, 02, 05, 07, 08, and 12) also indicate that these data may present rare patterns, which would be eliminated by the classic algorithm due to the low support values (Figures 4 Rules Distribution - Dataset 1 and 13).

Featured datasets, regardless of the number of instances, can generate a large number of patterns. Figure 3 presents information about the datasets and their execution in the Classic and HAR algorithm.

Fig. 3: Dataset

Datasets 03, 04, and 13 showed the highest number of rules generated in the Classic algorithm and the HAR, even without being the sets with higher instances as highlighted in the upper part of Figure 3. When comparing the classic algorithm with the HAR, we noticed that the number of generated rules had a strong relationship with the number of attributes (characteristics) present in each dataset.

We highlight dataset 01 and 13 for an individual analysis of its results in the classic algorithm and the HAR. All rules obtained in dataset 01 using the Classic algorithm Figure 4 (A) with a Support value of up to 10% have their knowledge to the left of the line eliminated in the classic approach. Studies with medical data presented in the supplementary material (`https://cutt.ly/3fQQw6E`). Use traditional algorithms and, by definition, need to determine a minimum Support value, so part of the knowledge can be eliminated.

The HAR in the Figure 4 (B) is composed of different metrics, such as *Conviction, Hyper-Confidence, Gini Index, Mutual Information, Imbalance Ratio, Kulczynski, and DMA*, seeks to identify the standards regardless of their occurrence. Thus, the HAR can identify common and rare patterns in the dataset, always seeking the most balanced rules.

When looking at Figure 4 (A) we notice that the classic selected rules with Confidence values less than 80% even the dataset showing situations with 100% Confidence, this occurrence is due to the Support being the main pattern cutting factor in the classic algorithm. In Figure 4, it is highlighted in (E and F) the rules selected by HAR in comparison to the classic algorithm.

The behavior of dataset 13 was presented because it is more significant in the number of rules. In this way, with a support limit of 20%, all knowledge to

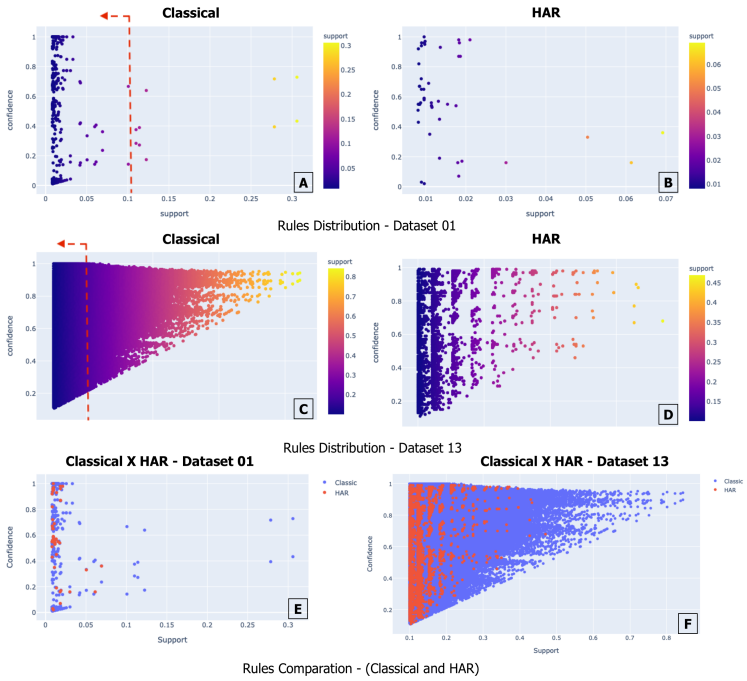

Fig. 4: Comparisons

the left of the red line would be eliminated using the classic algorithm (Figure 4 (C)).

When it comes to health care data, discarding patterns may not be an exciting solution, so HAR (D) searches for relevant patterns and uses metrics to validate the real implication among associative patterns.

In Figure 4 (D), it is possible to observe the points with the best-ranked rules. The classic algorithm rules have a very high support value, which can be expected or even irrelevant standards when observed only the Support/Confidence model.

By performing filters in different stages and using a set of metrics to validate the standards found, the HAR selects the rules of the dataset closest to the ideal through customized objective metrics. It is taking into account the particularities of each dataset. Identifying and selecting patterns that could be discarded when compared to a classic approach.

4 Conclusions

The algorithm (HAR), selects associative patterns in a set of data, for this, it chooses the best rules of the group of metrics, in order to identify the most appropriate relationship of the Antecedent (X) and Consequent (Y) in associative analyzes. HAR seeks to find more balanced rules through the composition

of six metrics (Hyper-Confidence, Conviction, Gini Index, Mutual Information, Imbalance Ratio, Kulczynski, and DMA). Together, select standards that value knowledge from the database, identifying rare and most common patterns, and eliminating redundancies and contradictions. Our algorithm values the meaning of the $X \rightarrow Y$ implication and eliminates potentially uninteresting rules, generating a smaller set of rules. As it is not limited to the Support/Confidence model of the classic algorithm (Apriori), HAR does not eliminate knowledge in real datasets, which may have a low Support value.

References

1. Agrawal, R., Srikant, R., et al.: Fast algorithms for mining association rules. In: Proc. 20th int. conf. very large data bases, VLDB, vol. 1215, pp. 487–499 (1994)
2. Buxton, E.K., Vohra, S., Guo, Y., Fogleman, A., Patel, R.: Pediatric population health analysis of southern and central illinois region: A cross sectional retrospective study using association rule mining and multiple logistic regression. Computer Methods and Programs in Biomedicine **178**, 145 – 153 (2019). DOI https://doi.org/10.1016/j.cmpb.2019.06.020
3. Han, J., Kamber, M., Pei, J.: 7 - advanced pattern mining. In: J. Han, M. Kamber, J. Pei (eds.) Data Mining (Third Edition), The Morgan Kaufmann Series in Data Management Systems, third edition edn., pp. 279 – 325. Morgan Kaufmann, Boston (2012). DOI https://doi.org/10.1016/B978-0-12-381479-1.00007-1
4. Jeong, H., Ohno, Y.: Cordless monitoring system for respiratory and heart rates in bed by using large-scale pressure sensor sheet. Smart Health **13**, 100,057 (2019). DOI https://doi.org/10.1016/j.smhl.2018.07.025
5. Lakshmi, K., Vadivu, G.: Extracting association rules from medical health records using multi-criteria decision analysis. Procedia Computer Science **115**, 290 – 295 (2017). DOI https://doi.org/10.1016/j.procs.2017.09.137. 7th International Conference on Advances in Computing & Communications, ICACC-2017, 22-24 August 2017, Cochin, India
6. Osadchiy, T., Poliakov, I., Olivier, P., Rowland, M., Foster, E.: Recommender system based on pairwise association rules. Expert Systems with Applications **115**, 535 – 542 (2019). DOI https://doi.org/10.1016/j.eswa.2018.07.077
7. Rothschild, M., Stiglitz, J.E.: Increasing risk: I. a definition. Journal of Economic Theory **2**(3), 225 – 243 (1970). DOI https://doi.org/10.1016/0022-0531(70)90038-4
8. Sethia, D., Gupta, D., Saran, H.: Smart health record management with secure nfc-enabled mobile devices. Smart Health **13**, 100,063 (2019). DOI https://doi.org/10.1016/j.smhl.2018.11.001
9. Toivonen, H., Klemettinen, M., Ronkainen, P., Hätönen, K., Mannila, H.: Pruning and grouping discovered association rules (1995)
10. Xu, C., Bao, J., Wang, C., Liu, P.: Association rule analysis of factors contributing to extraordinarily severe traffic crashes in china. Journal of Safety Research **67**, 65 – 75 (2018). DOI https://doi.org/10.1016/j.jsr.2018.09.013
11. Xue, Q., Chuah, M.C.: Explainable deep learning based medical diagnostic system. Smart Health **13**, 100,068 (2019). DOI https://doi.org/10.1016/j.smhl.2019.03.002