Filtered-ARN: Asymmetric objective measures applied to filter Association Rules Networks

Dario Brito Calçada and Solange Oliveira Rezende
Universidade de São Paulo
São Carlos, Brasil
dariobcalcada@usp.br, solange@icmc.usp.br

Abstract

In this paper, the Filtered-Association Rules Network (*Filtered*-ARN) is presented to structure, prune, and analyze a set of association rules in order to construct candidate hypotheses. The *Filtered*-ARN algorithm selects association rules with the use of asymmetric objective measures, *Added Value* and *Gain* then builds a network allowing more exploration information. The *Filtered*-ARN was validated using three datasets: Lenses, Hayes-roth, and Soybean Large, available online. We carried out a concept proof experiment using a real dataset with data on organic fertilization (*Green Manure*) for test the proposed method. The results were validated by comparing the *Filtered*-ARN with the conventional ARN and also comparing the results with the decision tree. The approach presented promising results, showing its ability to explain a set of objective items and the aid to build more consolidated hypotheses by guaranteeing statistical dependence with the use of objective measures.

Keywords: Association rules, Networks, Association rules networks, Data mining, Graphs, Objective measures.

1 Introduction

Data mining is often described as the process of discovering "interesting" patterns in large databases [1]. People in the world regularly generated and stored a large amount of digital data. Data mining offers a solution to the problem of quickly summarizing and searching for non-obvious relationships in data.

Data mining can be used as a methodology for discovering hypotheses or candidate theories. This approach allows us to explore recent advances in computationally efficient research techniques in conjunction with traditional statistical methods, which continue to be the basis for the verification and validation of theories [2].

The starting point of the mining process comes from observations (events) that trigger the researcher to accelerate conceptual studies and arrive at a structure in which the underlying process (which is generating the events) can be elucidated. The association rules mining [3, 4, 5] is used to find interesting patterns in the form of rules $A \Rightarrow B$, where A and B can be attributes, items or more generally "objects of data". Considering that it was known in advance that A and B are correlated, in a statistical sense, then the finding of rule $A \Rightarrow B$ only confirms previous knowledge and does not present new information. On the other hand, if the correlation between A and B has never been identified, the finding of the $A \Rightarrow B$ rule suggests that A and B are candidate pairs to be statistically validated (for correlation).

Since datasets are generally increasing in both quantity and dimensionality, listing all possible combinations of A and B and then verifying their correlation is not computationally feasible. Thus, the mining of association rules is a mechanism to offer candidate theories for validation. In this paper, an association rules mining method is presented that uses objective measures allied to a network structure to optimize hypothesis formation.

Algorithms that discover association rules use measures of interest that are capable of evaluating the quality of a rule. Among these measures, support and trust stand out, although *lift*, *gain*, *certainty factor*, *added value* or *leverage* are also indicators that provide information useful about the rules extracted [6].

Because association rule algorithms can extract all association rules according to minimum support and minimum trust value, the number of rules extracted usually exceeds the user's exploitability. Several

approaches have been proposed to guide the user in exploring rules. However, the vast majority of these approaches focus on the farm according to the entire rule and do not consider farms that can focus on a small set of items or a target item.

On the other hand, to facilitate the extraction of knowledge, many mining processes use networking techniques to visualize the data [7]. Combining objective measures with network structure for visualization, the *Filtered*-Association Rules Network (*Filtered*-ARN) is proposed in this article. *Filtered*-ARN uses objective measures asymmetric with the Association Rules Network (ARN) proposed by Pandey [8] to structure and assist in the analysis of extracted rules in a dataset.

By using asymmetric objective measures, *Filtered*-ARN allows for a more concise exploration of the rules by selecting only those where the predecessor of the rule statistically influences the successor. *Filtered*-ARN expands the features present in the ARN by adding some properties such as calculation of statistical influence and use of a measure of gain among the elements of the rules.

The structure of the network is the same as the ARN, but the selection of rules with proven statistical influence promotes the identification of hypotheses that are more likely to be true.

The central goal of *Filtered*-ARN is to present a graph with association rules that have a greater chance of being interesting for the analysis performed by the domain expert. The main difference between *Filtered*-ARN and conventional ARN is that in *Filtered*-ARN the user can view a set of items that promote statistical influence rather than elements that only relate to the target item. By allowing this, *Filtered*-ARN presents rules that indicate hypotheses with evidence of dependence between antecedent and consequent.

To validate the *Filtered*-ARN, we performed 3 case studies, and the results were compared with the conventional ARN and with a decision tree algorithm since they can be used to visualize degrees of dependence between elements of a dataset. A proof of concept was also carried out to prove the effectiveness of the proposed technique in data mining with a real dataset. The results demonstrated that the *Filtered*-ARN could describe the elements that influence the target item more concisely compared to ARN, allowing the user to observe cases where an object statistically interferes with a target item. Besides, we compared the *Filtered*-ARN to a decision tree algorithm, to analyze the explanation of the data. Such a comparison demonstrates that the *Filtered*-ARN has a more efficient structure than the decision tree, making it easier for the user to understand the extracted knowledge.

The paper is organized as follows. In Section 2 an overview of association rule mining is given, showing the definition and some objective measures known in the literature, in addition to the description of ARN. In Section 3, we presented some research that inspired the proposal of this article. The concept and structures of *Filtered*-ARN are presented with all its definitions and principles in Section 4. In Section 5, we presented the case studies, with the purpose of validating the *Filtered*-ARN and comparing it to the ARN as well as to the decision tree. The proof of concept is presented in Section 6. Finally, the conclusions and some future work are presented in Section 7.

2 Association Rules Mining

The purpose of data mining is to find models to predict the future or to understand the past [1]. The discovery of association rules is a data mining technique, which attempts to identify specific patterns of data in datasets, allowing, after its interpretation, to acquire specific knowledge about the problem under analysis [9]. Some techniques are designed as a black box, to obtain a model that is as accurate as possible, rather than to obtain a model that allows explanations [10]. However, various data mining techniques seek to uncover patterns in data that are understandable by humans.

Approaches to discovering patterns in data can be classified according to what they find out. Some common types of patterns found in datasets are clusters, sets of items, trends, and outliers [11].

Definition 1 Let $I = \{i_1, i_2, ..., i_n\}$ be a set of objects called items that can assume binary values 0 or 1 (false or true), which represent the presence or not of a particular object. Let T be a set of transactions, where each transaction D corresponds to a set of items such that $D \subseteq I$. It is also considered that a set of items A that is contained in a D transaction, if all items in the set have a "true" value in the transaction, that is, they are part of the same transaction. A R association rule can be represented by an expression of the form: $A \Rightarrow B$, with $A \subseteq I, B \subseteq I \; A \cap B = \emptyset$. It is also possible to treat the quantitative or qualitative variables, creating ranges of values, using them, later, as binary. A is called the antecedent (LHS - Left Hand Side) of the rule and B the consequent (RHS - Right Hand Side).

In the case of association rules mining, we can divided this process into three stages [12, 3, 13]:

- **Preprocessing:** preparation of the base for the extraction step and the removal of non-interesting items may occur.

- **Extraction of patterns:** calculation of the measurements, construction of frequent itemsets and elaboration of the association rules themselves.
- **Post-processing:** removes non-interesting rules and reduces the number of rules to be explored by the user.

Definition 2 For each rule ($LHS \Rightarrow RHS$), extracted from a set of transactions T, a support value (sup) is calculated, shown in Equation 1, which checks the strength of association between LHS and RHS (probability of occurrence of transaction $LHS \cup RHS$); and a confidence value ($conf$), Equation 2, which measures the force of the logical implication of the rule (conditional probability of RHS given LHS) [3].

$$sup(LHS \Rightarrow RHS) = P(LHS \cup RHS) \tag{1}$$

$$conf(LHS \Rightarrow RHS) = P(RHS|LHS) \tag{2}$$

Support can be described as the probability that any transaction satisfies both LHS and RHS, whereas confidence is the probability that a transaction satisfies RHS since it satisfies LHS.

According to Agrawal [3], the problem of extracting all association rules can be broken down into two parts:

- Find all sets of items that have transaction support above an informed threshold, called frequent itemsets.
- Generate membership rules from the frequent set of items. You should only select rules that have the minimum degree of support and confidence.

Thus, given a set of transactions, the problem of mining by association rules is to generate all rules that contain support and confidence equal to or higher than the minimum values determined by the user, referred to as minimal support ($minsup$) and minimal confidence ($minconf$), respectively.

The best-known algorithm for obtaining association rules is *Apriori* [3]. The algorithm employs deep search and generates sets of candidate items (patterns) of k elements from sets of $k-1$ elements items. The algorithm eliminates unfrequented patterns. The entire database is run, and the sets of frequent items are obtained from the sets of candidate items.

Given the size of the current databases, the number of rules discovered can be so high that it almost transforms its interpretation into a new mining problem. In this way, it is essential to understand the rules of association and the search for better ways to interpret them [14].

The selection of itemsets of interest makes the mining process of association rules complex, since the definition of interest is very subjective, in addition to being directly connected to the objective of the mining process and its respective dataset [4].

Measures of interest play an essential role in extracting and selecting interesting association rules. These measures are used to find patterns based on user need, since a large number of association rules generated by the pattern mining algorithm may not be useful as a whole. Therefore, there is a need to filter the rules [15].

In addition to the usual support measures (sup) and confidence ($conf$), other measures for the rule ($LHS \Rightarrow RHS$) can be calculated. Some symmetric measures are widely used as *Lift*[16], *Rule Interest* [17] e Test of χ^2 (chi-square) [18].

Definition 3 Objective measures of interest can be classified as symmetric or asymmetric, that is, a measure M is symmetric if $M(A \Rightarrow B) = M(B \Rightarrow A)$ [19].

One of the functions of objective measures of interest is to demonstrate how the items influence one another. This influence can be direct when the items vary directly or inversely when the increase of the incidence of one item leads to the decrease of another.

For this work, the measures *Added Value* (AV) and *Gain* were selected to construct the proposed approach because they are asymmetric measures and are directly related to the directed nature of the ARNs.

Definition 4 *Added Value* [-1;1]: the *Added Value* (AV) measure, described in Equation 3, indicates how much the frequency of consequent item increases in the presence of the antecedent item, i.e., it measures the gain of RHS in the presence of LHS [20]. If $AV > 0$, the frequency of RHS increases in the presence of LHS. Being $AV < 0$, the frequency of RHS decreases in the presence of LHS. If $AV = 0$, there is a random coincidence, that is, the frequency of LHS does not change the frequency of RHS.

$$AV = P(RHS|LHS) - P(RHS) = conf(LHS \Rightarrow RHS) - P(RHS) \tag{3}$$

Definition 5 ***Gain*** **[0;1]:** it is a measure proposed by Fukuda [21] (Equation 4) that forms a trade-off between support and confidence, assisting in the selection of rules according to their frequency and minimum confidence value.

$$Gain = [conf(LHS \Rightarrow RHS) - minconf].P(LHS) \tag{4}$$

By means of the equation 4, it can be seen that the objective measure *Gain* functions as a normalization of the confidence measure. When the gain value ($Gain = 0$) the rule confidence equals the minimum confidence ($conf(LHS \Rightarrow RHS) = minconf$).

The advantage of this measure over confidence measure is that one can more accurately calculate the influence of the predecessor element on the successor, and this action can also be used to select rules.

3 Related Work

As shown, membership rules are usually extracted using minimal support and confidence threshold.

Hahsler and Karpienko [22] present an interactive visualization method through a grouped matrix representation, which allows to explore and interpret highly complex scenarios intuitively. The generated rule sets are selected using the *Lift* measure and nested into a hierarchy that can be interactively explored up to the individual rule.

The work of Deng [23] demonstrates the use of associative classifiers that consist of a rule set ordered and represented as a tree model. Besides, he also proposed an algorithm to transform a tree into a set of ordered rules with concise rule conditions.

By incorporating negation and disjunction operators into rule antecedents, Kim [24] offers an expressive power in describing the interests of users as antecedents. In this work, the use of three elements is demonstrated: (1) a conceptual model; (2) three algorithms, called the family of CULTIVATION algorithms; (3) a system based, which is a Java implementation of the approach.

The presented works work by processing the rules and modeling them in a way that facilitates the understanding of the user. However, sometimes the user wants to analyze the behavior of a specific item. These jobs can reduce the number of rules to be parsed by the user, but do not explain how a particular item interacts with the entire dataset. This exploration item can be extremely useful in constructing hypotheses about the data.

Some approaches have been proposed to facilitate the acquisition of knowledge through association rules. These approaches combine the use of networks to the association rules mining in the three main mining phases (Preprocessing, Extraction, and Post-processing) [25].

The post-processing approach with the use of networks and transductive learning, in which some rules to be classified by the user are selected, directing efforts of the same with the rules considered of more significant impact in the network, according to with some network measure [26]. A necessary operation in the post-processing of rules is pruning, which consists of eliminating non-interest rules, so the use of targeted hypergraphs is an efficient approach to aid this process [27].

Combining the use of networks to aid in the mining of association rules, especially in its post-processing stage, as well as the pruning of the rules for optimization of knowledge extraction, Pandey [8] introduced the Association Rules Networks.

Association Rules Networks (ARNs) have a structure that allows synthesizing, pruning, and analyzing a set of association rules for the construction of candidate hypotheses.

The central idea of ARN is that the association rules discovered by the mining algorithm can be synthesized, pruned, and integrated into the context of specific research objectives. In particular, if there is a variable of interest ("objective"), a network can be formed with the most relevant variables related to the objective, and afterward, to elaborate a structure that can be tested using statistical methods.

As described by Chawla [27], ARNs use as a representation a reverse-oriented hypergraph (B-graph), which after the pruning processes, can transform the ARN according to the objective. For the creation of ARN, four steps are performed:

(1) Step A: Given a dataset D, minimum support and minimum confidence, we must first extract all association rules using a standard algorithm such as *Apriori* [4], *Apriori-Tid* [5] or *FP-Growth* [28].

(2) Step B: choose a frequent Z item, which will be represented in the rule set as the target node, and construct a *B-graph* that flows recursively to Z.

(3) Step C: Perform the pruning of the *B-graph* generated in Step B by removing hypercycles and reverse hyper-edges. The resulting "B-graph" is called ARN.

(4) Step D: Find shorter paths between the target node and the other nodes at the higher level (a variant of the distance between ends) of the ARN. The set of these paths represents the exploratory network for the target node.

4 *Filtered*-Association Rules Networks

In general, we represent a network R as $R = (V, E)$, in which V is a set of vertices (or nodes), and E is a set of edges (or links), which connect some pairs of vertices in V. Statistically, a graph can be characterized by derived values, such as the average degree of the nodes and the average length (path) between the nodes. Additional features such as network diameter, number of triangles, number of isomorphisms and clustering coefficient, can also be analyzed [29].

Given a network $R = (V, E)$, several links and auto-connections are not allowed depending on the type of network being implemented. If R is a directed network (DN), consider the universal set, denoted by U, containing all $|V| * (|V| - 1)$ potential directed links between a pair of nodes in V, in which —V— denotes the number of elements in V. If R is a non-directed network, the universal set U contains $|V| * (|V| - 1)$ links. In this way, the network representation is directly related to the type of data that it represents [30].

Newman [31] states that a variety of systems can be represented as networks in which some criterion can gather data. The function of the system that the graph represents can indicate the ideal form of the network. Some approaches promote the analysis of a network according to an objective item (node).

There are cases where the mining of association rules is done to explain predetermined items. The ARN presents an exploration guided by a single objective item. This exploit removes all the rules that are not interesting in the context of the target item, according to the minimum support and confidence metrics, showing the user only the relevant rules, but without certainty of the statistical dependence between the elements of the rules.

For example, consider the dataset *Lenses*[1]. If the user constructs an ARN with the "[lenses] = hard" attribute, to find out which symptoms lead to the patient using a rigid lens, the "[prescription] = myope" attribute is directly connected to the target node (Level = 1). This knowledge may direct the user to think that a patient with myopia is more likely to use a rigid lens.

However, the ARN can sometimes present relations that do not influence the elements of the rule. When we calculate the *Added Value* value of the rule "[prescription] = myope ⇒ [lenses] = hard" we find $AV = 0$, which affects a total independence between the constituent elements of this rule, therefore being a mistaken hypothesis regarding the behavior of patients who need rigid lenses.

To allow a complete exploration and taking into account the relationship between a set of objective items, this article proposes the *Filtered*-Association Rules Network (*Filtered*-ARN), which allows the exploration of an objective item with dependency analysis between the elements of the rules.

Definition 6 Given a set of rules of association R, containing unitary rules of *itemsets*, and a target item Z, *Filtered*-ARN is a DN that models all related rules to the item in Z, such as:

1. Each edge models a rule $r \subset R$.
2. From any point on the network, it is always possible to reach at least one vertex representing a Z item.
3. Given a vertex $v \subset$ *Filtered*-ARN, such as $v \not\subset Z$. There is no path of any Z item to v.
4. If there is a rule r such as $RHS(r) \subset Z$, then the rule $r \subset$ *Filtered*-ARN.

In the algorithm for constructing the *Filtered*-ARN, we make use of filters with asymmetric objective measures (*Added Value* and *Gain*), building the graph according to the selected rules.

The algorithm for generating the *Filtered*-ARN can be described in 3 steps:

(1) Step A: similar to the first step of all association rules mining processes. The extraction of rules with minimum support and confidence.

(2) Step B: calculate the asymmetric objective measures *Added Value* e *Gain*, and delete all rules with $AV = 0$ and values below the minimum gain (*mingain*).

(3) Step C: choose a frequent Z item, which will be represented in the rule set as the target node, and construct a *B-graph* that flows recursively to Z according to Pandey's methodology [8].

[1] `http://archive.ics.uci.edu/ml/datasets/Lenses`

The first step consists of the association rules mining phase. The only constraint added to this step, if compared to a conventional association rule mining, is that the rules must have unitary sets in the antecedent and consequent ($|LHS| = 1$ and $|RHS| = 1$). This restriction was added to facilitate *Filtered*-ARN modeling.

The second step is the filtering of the rules, which consists of the selection of rules that have elements with statistical dependence and the definition of the minimum gain of influence (*mingain*). This step will guide the entire exploration, as it will define the rules of interest that will be used with the objective item from which the network will be built.

In the last step, the user must select the item that he wants to understand in the dataset. Subsequently, the construction of the *Filtered*-ARN is performed. This step is responsible for getting all rules that are directly or indirectly related to the target item and modeling them. The construction of the *Filtered*-ARN is done recursively. First, the element selected as the target item is shaped on the chart (Level = 0). So, all the rules that the LHS item are not on the map and have the RHS item at Level 0 are modeled on the network. The same process is done for all items in Level 1, Level 2, and so on until there are no more rules to be modeled.

Besides, *Filtered*-ARN is constructed according to the levels of its vertices.

Definition 7 The level of a given vertex $v \subset$ *Filtered*-ARN is the number of edges needed to access item Z.

For example, item Z has Level zero (Level = 0) because it does not have to go through any edges to reach the item Z. Items in the LHS part of rules that have $RHS \subset Z$ will have Level one (Level = 1). They are an edge away from the items in Z.

As described, the algorithm constructs the *Filtered*-ARN connecting vertices on a graph according to the association rule it is representing. The *Filtered*-ARN is established as a DN, the proof is as follows:

Proof: the *Filtered*-ARN is constructed by connecting vertices from the X level to the vertices at the $X - 1$ level. Thus, all connections in *Filtered*-ARN are directed. Suppose there is a cycle $A \rightarrow B \rightarrow C \rightarrow A$, according to the rules of construction of the *Filtered*-ARN, Level (A) = Level (B) + 1, Level (B) = Level (C) + 1 and Level (C) = Level (A) + 1. Level (C) = Level (A) - 2 therefore it is not possible to create a cycle.

Another valuable property is that, from any vertex of *Filtered*-ARN, it is possible to reach a vertex at Level 0 (target item).

Proof: suppose that there exists a vertex v_x, at Level $N > 0$, which is not connected to any vertex Level zero (Level = 0). The *Filtered*-ARN is constructed from the zero level to the highest levels, always modeling the rules that have the RHS at a lower level.

Thus, the vertex v_x will only be modeled if a vertex at a level lower than that v_x is connected, is also connected to a vertex at Level 0. Then, all nodes in *Filtered*-ARNs are connected to at least one vertex at Level 0.

The exclusion of null statistical dependency rules ($AV = 0$) is a fundamental property of *Filtered*-ARNs since they guarantee that all the items that participate in the network generate some influence on the objective item.

Proof: the objective measure *Added Value* (Equation 3) is the result of the difference between the confidence $conf(LHS \Rightarrow RHS)$ and the $sup(RHS)$. This measure becomes zero ($AV = 0$) when rule confidence and consequent item support are equal. As rule confidence is a probability conditioned to the presence of RHS and support is the probability of RHS, it means that the antecedent item does not influence at any moment in the presence of the consequent, proving the total statistical independence of the same.

All of these properties are important for the analysis of *Filtered*-ARN, as they ensure that all modeled items always point to the selected target item with some influence relationship. Properties provide that the entire *Filtered*-ARN will be exploring the item defined in Z, allowing the user to understand its occurrence and construct hypotheses from the data.

For validation of the *Filtered*-ARN, three datasets of the UCI, [2] (*Lenses*, *Hayes-roth* and *SoyBean Large*), as well as a proof of concept with a real dataset (Green Manure) collected at Embrapa Meio Norte, in the city of Parnaíba, state of Piauí, Brazil. For the selection of the rules, a minimum gain value (*mingain*) was set equal to 0.1, filtering the rules that have $AV = 0$. For the construction of the network, we used the recursive process was described by Pandey [8].

Since the goal of *Filtered*-ARN is to analyze the correlations in the dataset so that the user obtains reliable information and to construct hypotheses likely to be true, the *Filtered*-ARN is evaluated with two other methods. First, a comparison is made between *Filtered*-ARN and conventional ARN by analyzing the differences between them and the pros and cons of using each of these approaches. After that, *Filtered*-ARN is compared to a decision tree algorithm. The purpose of this comparison is to examine the graphical results and to discuss which one produces a better structure for analyzing the dataset according to a set of items.

[2] `http://archive.ics.uci.edu/ml/`

5 Case Studies

To validate the *Filtered*-ARN approach and to present its capacity to explore datasets, we carried out some experiments on known datasets as well as a proof of concept with a real dataset. The tests were focused on presenting the differences between the methods, showing how the *Filtered*-ARN can allow more extensive exploration of the data, giving the user a complete understanding of them. Besides, a decision tree algorithm was used in the datasets to compare their output with the *Filtered*-ARN output.

The first study was using dataset *Lenses*, available at UCI [3]. In this dataset, each line represents the attributes of a patient and the contact lens that was prescribed for him. The main purpose of the dataset is to describe which features imply the prescription of each type of contact lens. It is important to remember that this dataset is a simplified version of the problem, the insights gained from it may not represent the actual correlations of the contact lens prescription scenario.

The second study was performed on the *Hayes-roth* dataset, available online [4]. The dataset brings information generated by personal data, such as age and level of education. This database contains six numeric value attributes, the first being only an index, which was deleted. The last attribute divides the instances into three classes.

The third study was done on the dataset *Soybean (Large)*, also available at UCI [5]. The dataset reports the results of a survey of disease characteristics in soybean plantations. There are 19 classes, and only the first 15 were used in previous work. The last four categories are little explored because they hae few examples. There are 35 categorical attributes, some nominal and some ordered. Attribute values are coded numerically, with the first value encoded as "0", the second as "1", and so on. The dataset has "empty" fields.

For the choice of datasets, a proposal for the calculation of the complexity rate was elaborated using the data found in Gupta [32]. In the present research, the author compared nine machine learning algorithms with 11 datasets from the UCI. With the results generated in this work, a calculation was proposed that allows comparing the degree of complexity of each dataset according to the definition.

Definition 8 Complexity Rate Proposal The Complexity Rate of a dataset is the ratio between the arithmetic mean of the complexity time (Tc) generated by machine learning algorithms and the arithmetic mean of the accuracy (Acc) of these same algorithms multiplied by the correction factor 1000 (Equation 5).

$$TxComplexity = \frac{Tc}{Acc} * 1000 \tag{5}$$

Thus, we classified the 11 UCI datasets according to Table 1.

Table 1: Complexity Rate

Dataset	**Complexity Rate**
LENSES	1,804
LABOR	2,305
IRIS	3,065
LUNG CANCER	4,765
VOTE	5,355
HAYES-ROTH	5,686
TEACHING ASSISTANT	6,030
STATLOG	8,256
GLASS IDENTIFICATION	8,897
SOYBEAN LARGE	123,469
ABALONE	1192,960

To perform the experiments presented in this paper three datasets were selected, one of low complexity (LENSES), one of medium complexity (HAYES-ROTH) and another of high complexity (SOYBEAN LARGE). With the three datasets the same steps were performed to construct the *Filtered*-ARNs.

In these experiments, the rules were extracted using the *Apriori-TID* algorithm implemented in Java ®.

[3] https://archive.ics.uci.edu/ml/datasets/lenses

[4] https://archive.ics.uci.edu/ml/datasets/Hayes-Roth

[5] http://archive.ics.uci.edu/ml/datasets/Soybean+%28Large%29

The extracted rules were filtered by the *Added Value* measured and all those with a null value ($AV = 0$) were excluded, indicating a lack of statistical influence, and a minimum gain of 0.01 was also selected so that the highest number possible rules could be analyzed.

The rule networks were constructed graphically with the use of the *Gephi* [33] software. *Gephi* is a specific open source application for networking and is available online [6].

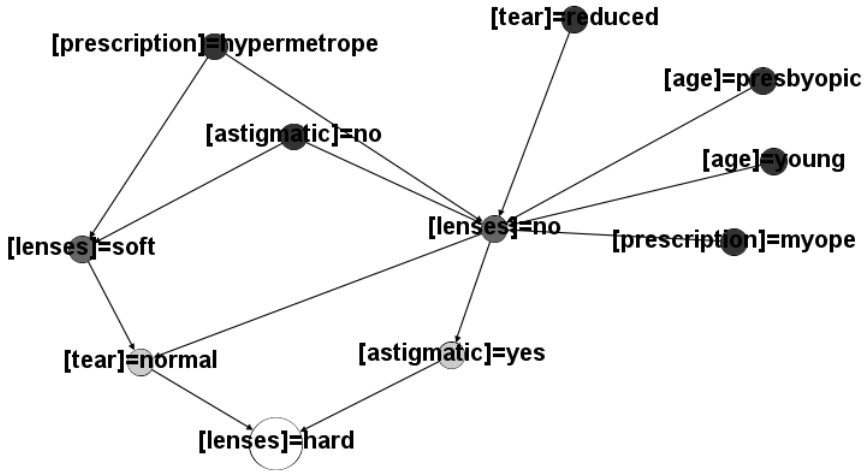
The algorithm J48, available in Weka [7] generated the decision trees, using the default configuration. In the following sections, the results are explained with each dataset.

5.1 Lenses dataset

Because the dataset has a small number of attributes, the minimum support was set to 0, so all values were considered. Minimum confidence was placed at 0.25, allowing the study of classes that occurred at least 1 in 4 times. Besides, the size of the rule was specified in two, considering an item in the LHS and an item in the RHS. Using this configuration, 99 candidate association rules for the dataset "*lenses*" were obtained. After the filtering stage, 60 rules remained.

The *Filtered*-ARN was generated considering as an objective item: "[lenses] = hard". The result can be seen in Figure 1. The generated network has four levels: Level 0, Level 1, Level 2 and Level 3. The *Filtered*-ARN has only 2 items related to the target item, which are: "[tear] = normal" and "[astigmatic] = yes". These rules may be able to influence the item "[lenses] = hard" and it is interesting to investigate, as they are the only parameters that generate an influence on the target item, therefore with a high degree of probability of generating true hypotheses.

Figure 1: *Filtered*-ARN with "[lenses=hard]" as target item



It can be seen in the generated *Filtered*-ARN that the Level 2 nodes are the other classes that indicate the lens type "[lenses] = soft" and "[lenses] = no", indicating Level 1 elements are subject to interference from these items. When we analyze the nodes of Level 3, it turns out that some rules stand out as: "[tear] = reduced" ⇒ "[lenses] = no", "[age]=presbyopic" ⇒ "[lenses] = no", "[age]=young" ⇒ "[lenses]=no", and "[prescription]=myope" ⇒ "[lenses]=no". These rules have a connection only to the "[lenses] = no" class, which creates hypotheses for the construction of a new *Filtered*-ARN, provoking a new direction in the exploration of knowledge.

In Figure 2 the ARN with "[lenses] = hard" is displayed as the target item. It can be observed that the network has only 3 levels, however with a structure totally different from the *Filtered*-ARN. As the construction of the networks is directly induced by the RHS item of the rules, it is noticed that Level 1 rules with no proven influence ($AV = 0$) appear, such as "[prescription] = myope" ⇒ "[lenses] = hard" and "[age] = young" ⇒ "[lenses] = hard", which leads to a generation of misconceptions regarding the use of rigid lenses.

Another notorious difference is the increase in the number of Level 3 nodes, without the distinction of dependencies between them, which makes it impossible to direct the studies, since the network demonstrates the same degree of importance for all items of the same level.

We show the decision tree in Figure 3. The J48 algorithm obtained 20/24 correct classifications (83.33%) and lost 4/24 examples (16.67%).

[6] https://gephi.org/users/download/

[7] http://www.cs.waikato.ac.nz/ml/weka/

Figure 2: ARN with "[lenses=hard]" as target item

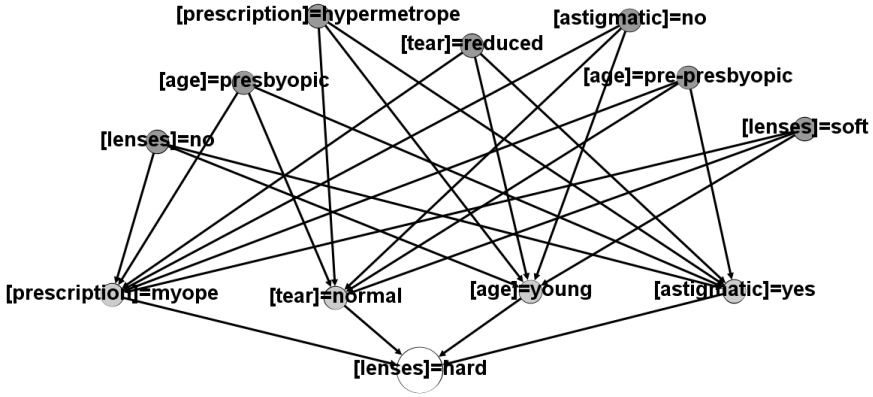
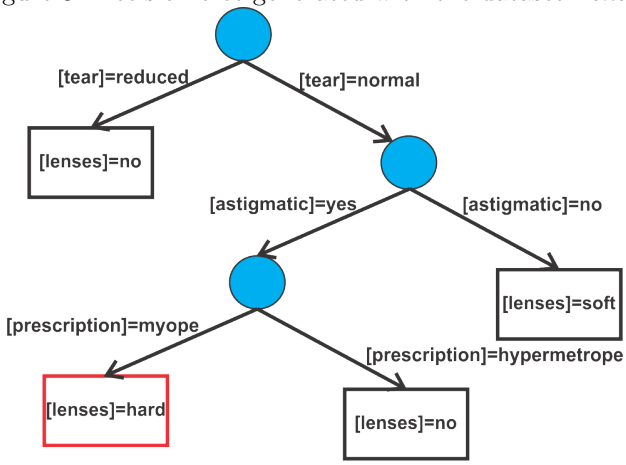


Figure 3: Decision tree generated with the dataset *Lenses*



Comparing the output of *Filtered*-ARN with the decision tree (Figure 3), we can see differences in the explanation of objective items. Both have "[tear] = reduced" and "[prescription] = hypermetrope" directly connected to "[lenses] = no", but in *Filtered*-ARN other rules arise as the condition of age.

In the tree, it is explained that "[prescription] = myope" binds directly to "[lenses] = hard", which is a misconception because it is a condition without influence ($AV = 0$), which can be inferred from the *Filtered*-ARN.

5.2 Hayes-roth dataset

The *Hayes-roth* dataset has six attributes, the first 2 of which were generated randomly, so they were removed. The size of rule was specified in two, considering 1 item in the LHS and 1 item in the RHS. We extract the association rules, and we construct the networks with the "[class] = 3" as target node because it is the class with the highest number of instances.

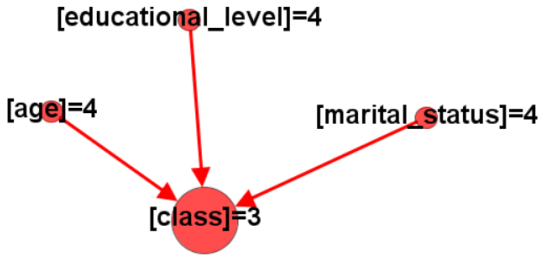
Because the dataset has only 162 instances, the minimum support was set to 0, so all values were considered. Minimum confidence was placed at 0.25, allowing the study of classes that occurred at least 1 in 4 times. Using this configuration, 167 candidate association rules for the *Hayes-roth* dataset were obtained. After the filtering stage, 123 rules remained.

In the Figure 4, the *Filtered*-ARN array with "[class]= 3" is displayed as the target item. The presence of only nodes in the whole network is observed, with 1 node being the target node and the other 3, nodes connected directly to the target node.

With the analysis of the *Filtered*-ARN of Figure 4, it is inferred that the profile of persons in class 3 of the dataset have higher ages ([age] = 4), higher educational levels [educational_level] = 4) and marital status of type 4 ([marital_status] = 4). This information is clear in the *Filtered*-ARN and represents connections with a high probability of being true.

The ARN with "[class]=3" as the target item was constructed (Figure 5) in order to compare with the *Filtered*-ARN. The structure of the ARN is different, and it has the three levels of nodes. Several connections

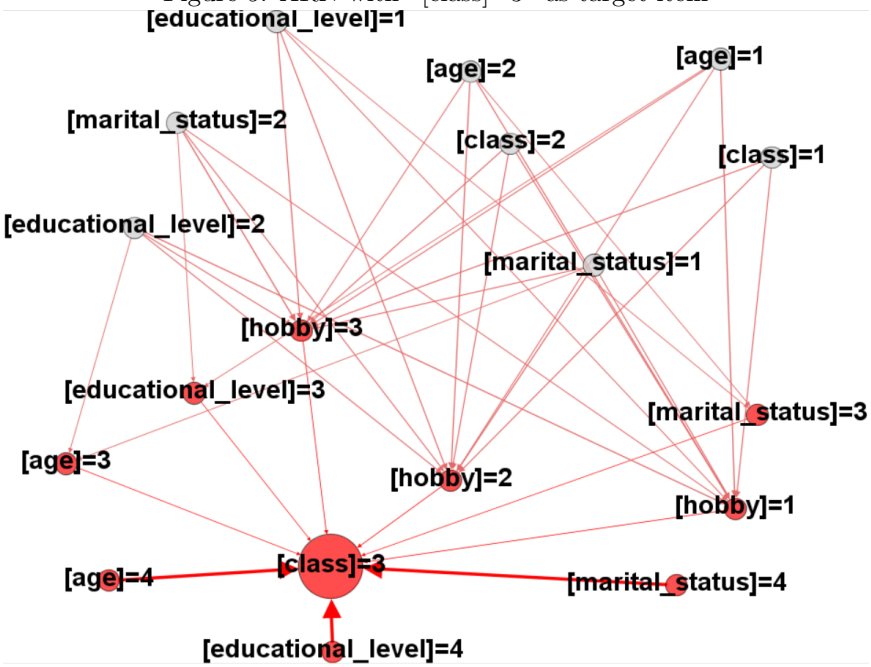
Figure 4: *Filtered*-ARN with "[class]=3" as target item



without influence are presented on the network. Unrelated edges lead to the generation of hypotheses with a high probability of being false.

Connections to the target item such as "[hobby] = 1", "[hobby] = 2" and "[hobby]=3" are inserted into the ARN, but these rules have AV = 0, which denotes the non-influence of the "hobby" in the selection of people from dataset class 3.

Figure 5: ARN with "[class]=3" as target item



The decision tree generated (Figure 6) for validation has 19 sheets and size 25. Comparing the output of the *Filtered*-ARN with the decision tree it is possible to perceive differences in the explanation of the objective item. When parsing the tree, one sees the same connections of the *Filtered*-ARN directly linked to "[class] = 3" of the dataset, "[age] = 4", "[educational_level] = 4" and "[marital_status] = 4", but there are dependencies of the educational level and matrimonial state with other age categories, which can generate inconsistency in the extracted knowledge and the elaboration of false hypotheses.

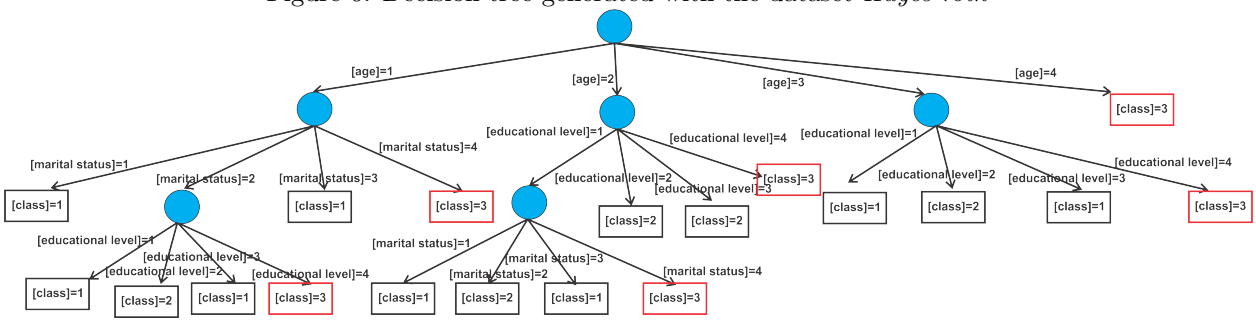
5.3 Soybean Large dataset

To continue validation of the benefits of *Filtered*-ARN some experiments were carried out on the soybean dataset (Soybean Large) because it is a dataset with a high degree of complexity (Table 1).

Since dataset has 19 classes, 36 attributes and 307 instances, some with empty fields, the minimum support has been set to 0.03 to avoid the appearance of classes with incomplete information. The minimum confidence was placed to 0.45, to generate a more coherent proportion with the objective item "[class] = rhizoctonia", which was selected, concerning the items in the other classes. Besides, the size of the rule was specified in two, considering an item in the LHS and an item in the RHS. Using this configuration, 4223 candidate association rules for the dataset "*Soybean*" were obtained. After the filtering stage, 4019 rules remained.

The *Filtered*-ARN was generated considering as an objective item: "class = rhizoctonia" which means

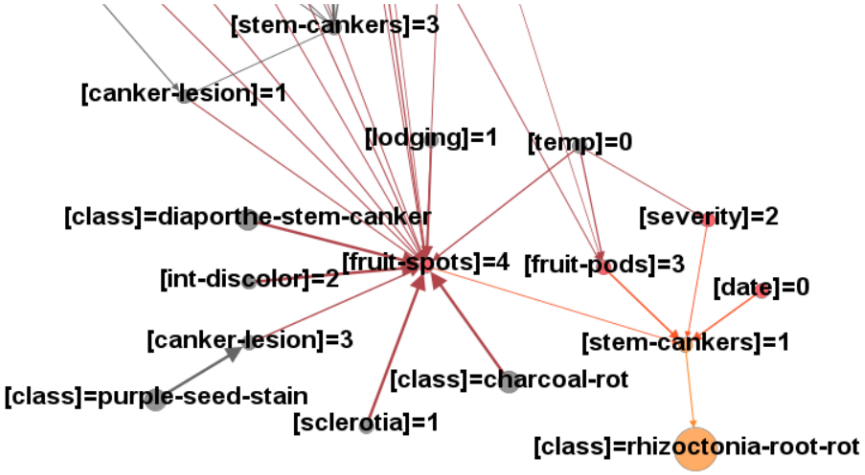
Figure 6: Decision tree generated with the dataset *Hayes-roth*



plants with a specific type of fungus. A highlight of the network result, with Level 1 and Level 2 nodes, can be seen in Figure 7. The complete *Filtered*-ARN is available on-line [8].

Analyzing Level 1 items of *Filtered*-ARN, we can notice that only one element causes direct influence on the target item, "[stem-cankers] = 1". This relation may be able to describe the objective item with a high degree of dependence, causing the formation of a hypothesis with a high probability of being correct.

Figure 7: *Filtered*-ARN with "[class]=rhizoctonia" as target item



In Figure 8 a cut of the ARN with "[class] = rhizoctonia" is shown as the target item and Level 1, Level 2 and Level 3 nodes without predecessors. The complete network is available online[9]. We can observe that the network has a much larger number of items connected directly to the target. In addition to the item "[stem-cankers] = 1", other four elements form Level 1 of the network, but without any guarantee of dependency, which can lead to false hypothesis formation.

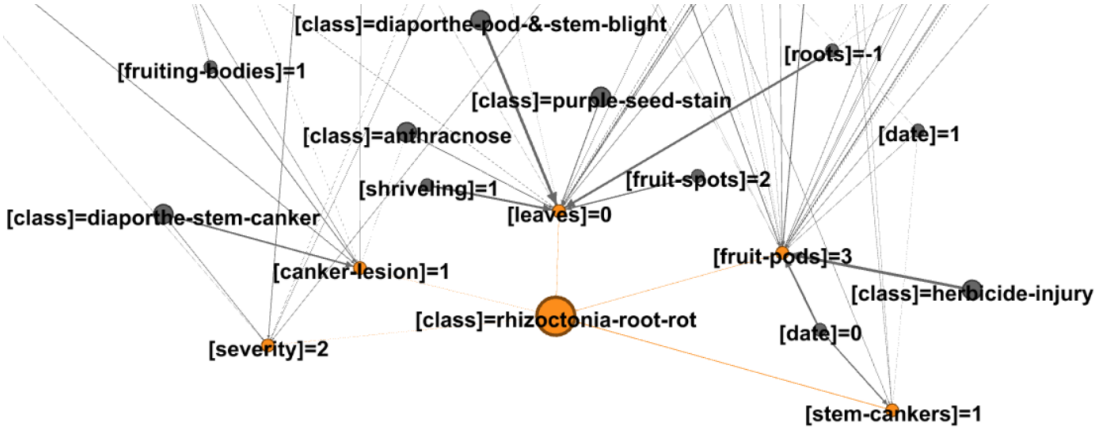
The elements "[canker-lesion] = 1" and "[severity] = 2", which are part of Level 1 of the ARN, are observed at Level 2 in *Filtered*-ARN, with which they can affect the target item, but in a more indirect way than indicated by traditional ARN. The other items observed at Level 1 of the ARN, "[leaves] = 0" and "[fruit-pods] = 3" are part of Level 3 of the *Filtered* -ARN, which drastically decreases the probability of a hypothesis being generated directly with the target item.

In *Filtered*-ARN you can see other Level 2 items, "[fruit-spots] = 4" and "[date] = 0", which influence the only condition directly linked to the target item .

The decision tree generated for validation had 69 sheets and obtained 87.58% accuracy. As the tree is very long, no image of it was inserted in this paper. The file containing the complete output is available online[10]. When analyzing the tree, it is difficult to even understand the behavior of the "[class] = rhizoctonia" item. The tree that explains the classes is more complicated to understand than *Filtered*-ARN.

[8] https://goo.gl/V882D2
[9] https://goo.gl/V882D2
[10] https://goo.gl/V882D2

Figure 8: ARN with "[class]=rhizoctonia" as target item



6 Proof of Concept - Green Manure dataset

To validate the benefits of using the *Filtered*-ARN in real data some experiments were carried out on the Organic Manure dataset collected from EMBRAPA Meio Norte. This dataset has already been used by [34] in the construction of ARNs for information exploitation. The extracted knowledge was analyzed and validated by specialists.The extracted knowledge was analyzed and validated by specialists. We purpose this proof of concept to verify the technique result in a dataset built with real data and without any treatment.

Since the dataset has six classes related to the legume decomposition half-life, 11 attributes, and 28 instances, the characteristics being formed by categories of values referring to the field research performed by EMBRAPA Meio Norte, we set the minimum support to 0.3, and the minimum confidence to 0.5 for benchmarking with the studies of [34]. Besides, the size of the rule was specified in two, considering one item in the LHS and one item in the RHS. Using this configuration 64 candidate binding rules were obtained for the dataset *Green Manure*. After the filtering stage, 50 rules remained.

The *Filtered*-ARN was generated considering as an objective item: "[HalfLife] = 6" which means legumes with higher half-life rates because they have a high decomposition rate. The result of the network can be seen in Figure 9.

Analyzing the Level 1 items of *Filtered*-ARN, we can see that five items are related to the target item, highlighting "[AP] = 4", a parameter related to plant height at flowering, which has no predecessor. These rules may be able to describe the probable characteristics of plants with a longer half-life, being the only parameters that generate an influence on the objective item, thus making hypotheses with a high probability of being correct.

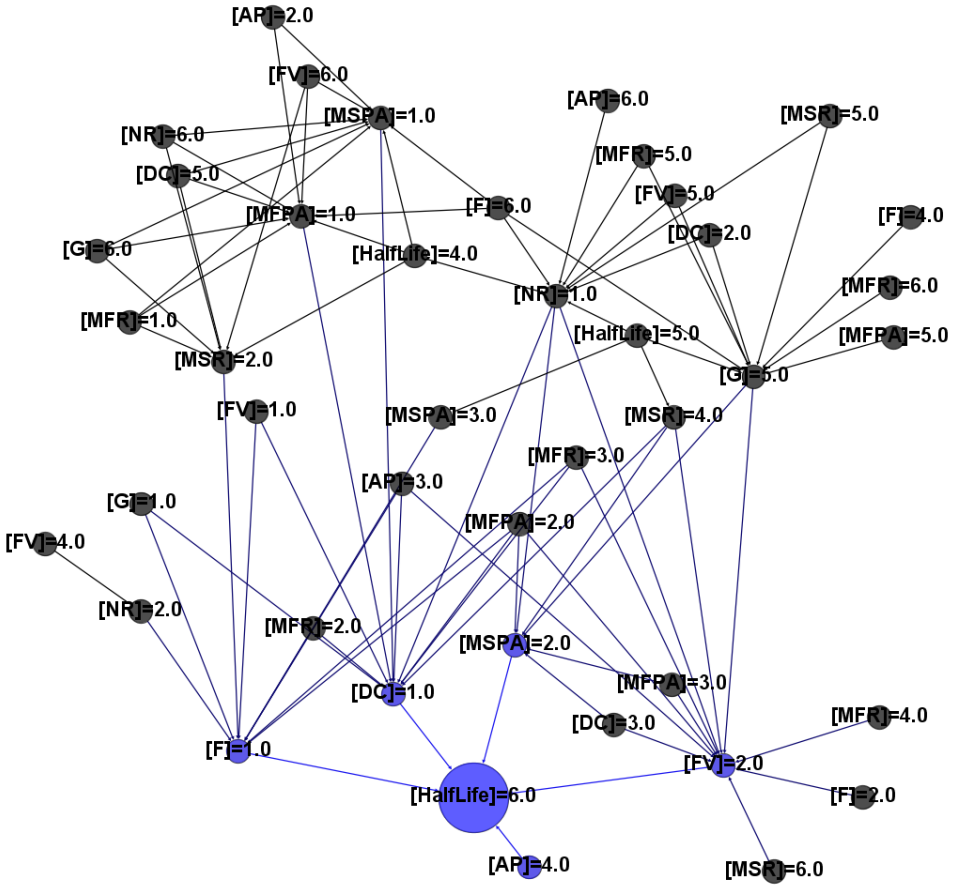
In Figure 10 the ARN with "[HalfLife] = 6" is displayed as the target item. We can see that the network has a structure other than the *Filtered*-ARN. The number of items connected directly to the target increased to 7. It is noticed that some elements are maintained in both networks connected to the target item: "[MSPA] = 2.0", "[F] = 1.0" and "[AP] = 4.0", the latter having no predecessors in the *Filtered*-ARN, increasing its importance for the study. The parameter "[FV]" underwent a category change, in the ARN it perceived a category "1.0" and in the *Filtered*-ARN, this value was changed to the " 2.0 " category, although there is a relation between these elements, the influence begins in the second range of values. The remaining rules with ARN level one were pruned in *Filtered*-ARN.

Another notable difference is the appearance of a new parameter of influence related to the collector diameter ([DC]). In the *Filtered*-ARN this parameter in the "1.0" category appears with direct impact on the target item, which could not be perceived by the conventional ARN.

In addition to comparing with the ARN, we also compared the *Filtered*-ARN to a decision tree. The decision tree generated by the algorithm J48 is presented in Figure 11. The J48 algorithm obtained only 39.28% of correct classifications forming a tree with 16 leaves.

Comparing the output of the *Filtered*-ARN with the decision tree (Figure 11), it is possible to perceive differences in the explanation of objective items. For the "[HalfLife] = 6.0" objective, only the direct possibility of the parameter "[F]" referring to the flowering of the plant is observed in the tree, and the values "[F] = 1.0" are obtained when "[MSPA] = 1.0" and "[F] = 1.0 or 2.0 or 4.0 or 5.0" when "[MSPA] = 2.0". The J48 algorithm completely ignores the other parametric conditions for obtaining higher half-lives. In this way, the application of the *Filtered*-ARN algorithm allows a broader study of parameters, optimizing the extraction of knowledge through the study of more relevant hypotheses and more likely to be true.

Figure 9: *Filtered*-ARN with "*HalfLife*=6" as target item



The EMBRAPA Meio Norte specialists evaluated all results, and only the *Filtered*-ARN information was validated in its entirety. The experts also reported the relationship between collector diameter ([DC]) and half-life rate of plants with the lowest percentage of decomposition is considered only as suspect. This result positively encourages the realization of new agricultural experiments to prove the knowledge provided by the mining of data with the use of *Filtered*-ARNs.

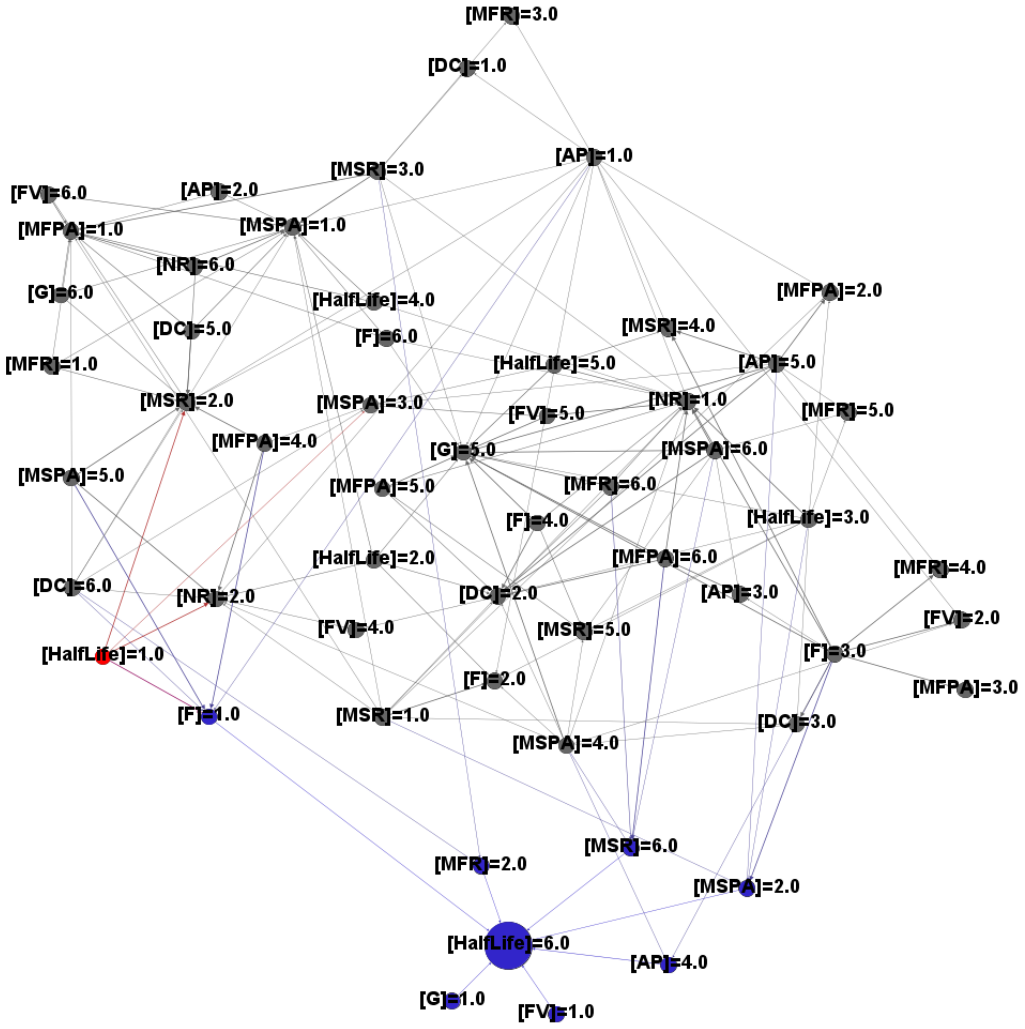
7 Conclusion and Future Works

In this article, we present the proposal of *Filtered*-ARN, a method capable of modeling association rules, previously selected by asymmetric objective measures, according to an already defined objective item. The rules chosen for the construction of the network are those that have a statistical dependence proven by the AV measure. The objective item is used as guiding the exploration and is chosen according to the problematic one that wishes to formulate hypotheses. The *Filtered*-ARN creates a directed hypergraph, modeling the membership rules that have the target item in RHS recursively. It aims to explain the correlation between items in the dataset with the target item.

Three case studies were developed to validate the ability of the *Filtered*-ARN. Artificial datasets were explored: *Lenses*, *Hayes-roth* and *Soybean Large*. A proof of concept with a real dataset with data on organic fertilization (*Green Manure*) obtained at EMBRAPA Meio-Norte was also carried out. Experts evaluated the proof of concept. The objective was to describe the occurrence of rules that statistically influence the desired class, aiming to find the items that best explain the occurrence of the class item. Besides, we applied the standard ARN and the J48 algorithm for comparison with the *Filtered*-ARN.

The results showed that ARN is useful for describing relationships with an actual item. However, it does not show the user which cases in which the items truly influence the target item statistically. It is important to highlight these cases because some explorations are done with an objective item and mistaken hypotheses are elaborated causing a delay in the extraction of knowledge, and the same can still be false. *Filtered*-ARN is very successful in presenting such cases, allowing the user to see the items that produce variations in the

Figure 10: ARN with "*HalfLife*=6" as target item. Adapted from [34].



target item.

We also compared the *Filtered*-ARN algorithm to J48, which is a decision tree algorithm available on Weka. Since the decision tree performs only the classification of the elements, it is not very good at explaining the relations. The decision tree generated in the dataset Soybean is an excellent example of this, as the tree achieved 87.58 % accuracy, but the tree became very large (69 leaves), it is complicated for a user to understand the correlations. *Filtered*-ARN has obtained excellent results, describing the data using the association rules extracted and selected, showing the user the items that are entirely responsible for the occurrence of the target item and that generate influence on it.

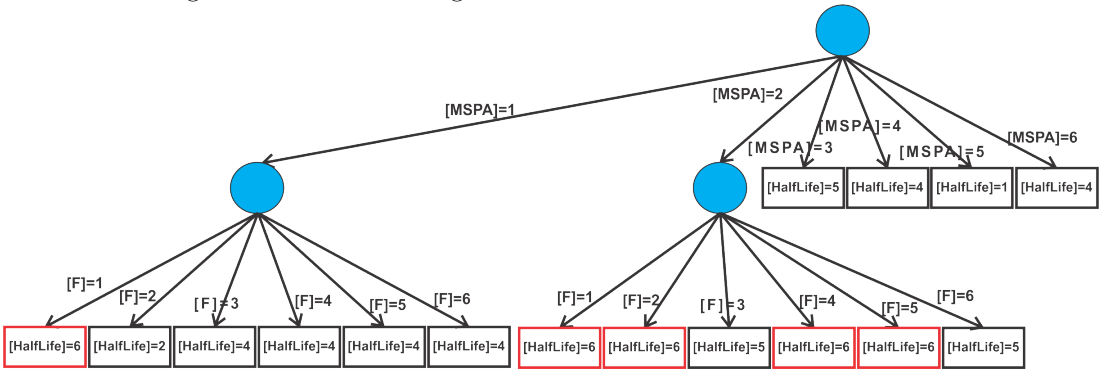
Experts validated the results. Only the output generated by the *Filtered*-ARN was confirmed in its entirety, and a new possibility of the agricultural study was also verified by the involved researchers.

In addition to presenting exciting insights, several improvements can be made to the *Filtered*-ARN to help the user identify items that are not interesting for their exploration. The development of an approach with more than one objective item to relate those elements that do not compete with each other can help identify possible exciting items. Another aspect is that the knowledge generated by the mining of association rules with the use of the *Filtered*-ARN can aid in the improvement of other classifiers.

It is important to emphasize the variability of the values of the minimum gain (*mingain*), this type of measurement interferes directly in the elaborated network structure, therefore a correlation of the objective measures of the association rules with measures of centrality of the network can be made generated in order to assist in the selection of the best extraction parameters.

About the *Filtered*-ARN structure, it is interesting to analyze the effect of some algorithms of network construction on the final result, to optimize specific characteristics and to allow the user to manipulate the creation of the *Filtered*-ARN.

Figure 11: Decision tree generated with the dataset *Green Manure*



Acknowledgment

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brazil (CAPES) - Finance Code 001.

References

[1] C. C. Aggarwal, *Data Mining: The Textbook*, 1st ed. New York, USA: Springer, 2015.

[2] M. Vinaya and K. Shah, "Performance Evaluation of Distributed Association Rule Mining Algorithms," *Procedia - Procedia Computer Science*, vol. 79, pp. 127–134, 2016.

[3] R. Agrawal, T. Imielinski, and A. Swami, "Mining Association Rules between Sets of Items in Large Databases," *Special Interest Group on Management of Data, 22(2)*, vol. 22, no. 2, pp. 207–216, 1994.

[4] R. Agrawal and R. Srikant., "Fast algorithms for mining association rules," *Jorge B. Bocca, Matthias Jarke, and Carlo Zaniolo, editors, Proceedings of Twentieth International Conference on Very Large Data Bases (VLDB)*, pp. 487–499, 1994.

[5] R. Agrawal and J. C. Shafer, "Parallel mining of association rules," *IEEE Transactions on Knowledge and Data Engineering*, vol. 8, no. 6, pp. 962–969, 1996.

[6] M. Martínez-Ballesteros, F. Martínez-Álvarez, A. Troncoso, and J. C. Riquelme, "Selecting the best measures to discover quantitative association rules," *Neurocomputing*, vol. 126, pp. 3–14, 2014.

[7] C. Kim, H. Lee, H. Seol, and C. Lee, "Identifying core technologies based on technological cross-impacts: An association rule mining (ARM) and analytic network process (ANP) approach," *Expert Systems with Applications*, vol. 38, no. 10, pp. 12 559–12 564, 2011.

[8] G. Pandey, S. Chawla, S. Poon, B. Arunasalam, and J. G. Davis, "Association Rules Network: Definition and Applications," *Statistical Analysis and Data Mining*, vol. 1, no. 4, pp. 260–179, 2009.

[9] T. Le and B. Vo, "The lattice-based approaches for mining association rules: a review," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 6, no. 4, pp. 140–151, 2016.

[10] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. Waltham, MA, USA: Morgan Kaufmann, 2012.

[11] P. Fournier-Viger, J. C. W. Lin, B. Vo, T. T. Chi, J. Zhang, and H. B. Le, "A survey of itemset mining," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 7, no. 4, pp. 1–18, 2017.

[12] C. C. Aggarwal, C. Procopiuc, and P. S. Yu, "Finding localized associations in market basket data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 14, no. 1, pp. 51–62, 2002.

[13] L. T. T. Nguyen and N. T. Nguyen, "Updating mined class association rules for record insertion," *Applied Intelligence*, vol. 42, no. 4, pp. 707–721, 2015.

[14] M. A. Domingues and S. O. Rezende, "Post-processing of Association Rules using Taxonomies," *Proceedings of the 12th Portuguese Conference on Artificial Intelligence (EPIA 2005)*, pp. 192–197, 2005.

[15] D. J. Prajapati, S. Garg, and N. C. Chauhan, "MapReduce Based Multilevel Consistent and Inconsistent Association Rule Detection from Big Data Using Interestingness Measures," *Big Data Research*, vol. 9, pp. 18–27, 2017.

[16] S. Brin, R. Motwani, J. D. Ulman, and S. Tsur, "Dynamic itemset counting and implication rules for market basket data," *Proc. of the ACM SIGMOD Intl. Conf. on Management of Data*, pp. 255–264, 1997.

[17] G. Piatetsky-Shapiro, "Discovery, Analysis and Presentation of Strong Rules," *Knowledge Discovery in Databases, AAAI/MIT Press*, pp. 229–248, 1991.

[18] B. Liu, W. Hsu, and Y. Ma, "Pruning and summarizing the discovered associations," *Proceedings of the 5th ACM International Conference on Knowledge Discovery and Data mining*, pp. 125–134, 1999.

[19] P.-N. Tan, M. Steinbach, and V. Kumar, "Association Analysis: Basic Concepts and Algorithms," in *Introduction to Data mining*, 2005, pp. 327–414.

[20] S. Sahar, "What Is Interesting: Studies on Interestingness in Knowledge Discovery," Phd Thes, Tel-Aviv University The, 2003.

[21] T. Fukuda, Y. Morimoto, S. Morishita, and T. Tokuyama, "Data Mining Using Two-Dimensional Optimized Association Rules: Scheme, Algorithms, and Visualization," in *SIGMOD '96 Proceedings of the 1996 ACM SIGMOD international conference on Management of data*, 1996, pp. 13–23.

[22] M. Hahsler and R. Karpienko, "Visualizing association rules in hierarchical groups," *Journal of Business Economics*, vol. 87, no. 3, pp. 317–335, 2017.

[23] H. Deng, G. Runger, E. Tuv, and W. Bannister, "CBC: An associative classifier with a small number of rules," *Decision Support Systems*, vol. 59, no. 1, pp. 163–170, 2014.

[24] E. H. Kim, H. G. Kim, S. H. Hwang, and S. I. Lee, "FARM: An FCA-based Association Rule Miner," *Knowledge-Based Systems*, vol. 85, pp. 277–297, 2015.

[25] M. Rashid, I. Gondal, and J. Kamruzzaman, "Mining Associated Patterns from Wireless Sensor Networks," *IEEE TRANSACTIONS ON COMPUTERS*, vol. 64, no. 7, pp. 1998–2011, 2015.

[26] R. d. Padua, S. O. Rezende, and V. O. D. Carvalho, "Post-processing association rules using networks and transductive learning," *Proceedings - 2014 13th International Conference on Machine Learning and Applications, ICMLA 2014*, pp. 318–323, 2014.

[27] S. Chawla, "Feature Selection, Association Rules Network and Theory Building." *JMLR: Workshop and Conference Proceedings - The Fourth Workshop on Feature Selection in Data Mining*, vol. 10, pp. 14–21, 2010.

[28] G. Grahne and J. Zhu, "Fast algorithms for frequent itemset mining using FP-trees," *IEEE Trans Knowl Data Eng*, vol. 17, pp. 1347–1362, 2005.

[29] D. F. Nettleton, "Data mining of social networks represented as graphs," *Computer Science Review*, vol. 7, no. 1, pp. 1–34, 2013.

[30] J. C. Valverde-Rebaza and A. De Andrade Lopes, "Link prediction in online social networks using group information," Ph.D. dissertation, Universidade de São Paulo, 2014.

[31] M. Newman, *Networks: An introduction*, 1st ed. New York, USA: Oxford University Press, 2010, vol. 55.

[32] A. Gupta, "Classification of Complex UCI Datasets Using Machine Learning Algorithms Using Hadoop," *International Journal of Scetific & Techology Research*, vol. 4, no. 05, pp. 85–94, 2015.

[33] M. Bastian, S. Heymann, and M. Jacomy, "Gephi: An Open Source Software for Exploring and Manipulating Networks," *Third International AAAI Conference on Weblogs and Social Media*, pp. 361–362, 2009.

[34] D. B. Calçada, S. O. Rezende, and M. S. Teodoro, "Analysis of decomposition parameters of green manure in the Brazilian Northeast with Association Rules Networks," in *I International Conference on Agro BigData and Decision Support Systems in Agriculture*, Montevideo, Uruguay, 2017, pp. 63–65.