Analyzing strengths and weaknesses of fuzzy association rules algorithms

Jorge Casillas, Daniel Moreno

J. Casillas is with the Department of Computer Science and Artificial Intelligence. University of Granada, 18071 Granada, Spain. e-mail: casillas@decsai.ugr.es

D. Moreno is with the Department of Industrial Engineering, University of Costa Rica, 11501-2060 UCR, San Jose, Costa Rica. e-mail: daniel.dmoreno@gmail.com

Abstract

A methodology to compare fuzzy association rules algorithms is presented in this paper, showing that allows the epithet on equal terms. This is motivated because of this is not a soft computing area with a experimental framework as developed as others, for example: Classification. One reason is that the problem itself is multi-objective. Thus, the diverse approaches are evaluated in different ways by authors. First at all, we propose how to classified (viewpoint of the final user) algorithms of fuzzy association rules and then used this classification map to select 3 algorithms and apply mentioned methodology

Keywords: Data mining, fuzzy association rules, comparison methodology

1 Introduction

Since its introduction in [1] the technique of association rules mining has received great interest by the data mining community. A lot of research has been done, resulting in the development of many different algorithms and an increasingly number of publications. But maybe the most important milestone occurred when the concept of fuzzy sets were included, giving rise to Fuzzy association rules. In last years this area has received increasing attention as shown in Figure 1

Despite this, is not a soft computing area with a benchmark or experimental framework as developed as others, for example: Classification. One reason is that the problem itself is multi-objective: the quality measures, quantity of
2 Classifying the algorithms

2.1 Brief history

Since association rules introduction [1] and Apriori and AprioriTID [2] publication, a lot of new algorithms was develop, trying to improve the previous approaches. But maybe the most important milestone occurred when the concept of fuzzy sets [38] were included, and linguistics variables [37] replace at traditional itemset. Some milestones in the history of Fuzzy rules algorithms
are shown in Figure 2 (is not a exhaustive list).

Since first fuzzy version of association rules algorithms were published between 1997 and 1999 (one of them is [32]), the work of researchers has been done in basically 4 fields (below is development each of these):

- Obtain the membership functions
- Itemsets selection to include as antecedent or consequent
- Better alternatives to the Support and Confidence measures
- Specific applications: classification, rare items, sub-groups discovery, others

2.1.1 Obtain the membership functions

In this research line is possible to distinguish two different aspects: one based on triangular or trapezoidal labels, and the other based on fuzzy clustering. The first one is the classical way to label the domain of every quantitative variable. Thus, the methodology consist in to find the best limits for each label, as is presented in

Examples of this line, using Genetic algorithms are in [13]. Other proposals using Ant colony algorithms are in [35] and [21].

The other line is focused on using the known fuzzy versions of clustering algorithms to label the domain with “Gaussian” functions. There are different ways to do this 1-dimensional clustering of the variables, in [23] authors performed a comparative analysis of 3 different algorithms:

- FCM (Fuzzy C Means)
- CURE
- CLARANS

Which shows that Fuzzy C Means, a fuzzy version for classical “\(k\)-Means” [4] [14], is the most efficient of these methods. This manner to labeling is advocated by some authors because it uses the distribution of the data to create the labels instead of using only the range as in the case of triangular labels. However, is still an open topic, there are not clear advantages of one over another. We discuss this in section 3.1.1. Algorithms using Fuzzy clustering to label appear in [29], [31].

2.1.2 Itemsets selection to include as antecedent or consequent

An interesting (and non-trivial) issue is: how to obtain the large-itemsets? As search proposed in "Apriori" algorithm (BFS technique) is very low scalable to large databases (more detail in Subsection 3.2.3), an important work has been done in faster way to combine itemsets, avoiding the combinatorial explosion.
2. Classifying the algorithms

Fig. 2: Timeline of milestones for Fuzzy Association Rules

- 1993: Mining association rules between sets of items in large databases
- 1994: Fast algorithms for mining association rules (Apriori and AprioriTID)
- 1999: Fuzzy transaction data-mining algorithm (FTDA)
- 2001: Trade-off between computational time and number of rules for fuzzy mining in quantitative data (F-Apriori)
- 2003: Discovering fuzzy association rules using fuzzy partition methods
- 2008: Extraction and optimization of fuzzy association rules using multi-objective genetic algorithm
- 2009: Fuzzy association rule mining algorithm for fast and efficient performance on very large datasets (F-Arbor)
- 2010: Automatic discovery of potential causal structures in marketing databases based on (FCsgn)

Approaches included in paper
2. Classifying the algorithms

Fig. 3: Membership functions for a specific variable

Fig. 4: Combining itemsets in large-itemsets

involved in testing all possible combinations. Figure 1 shows graphically the problem.

As many other combinatorial problems, metaheuristics (including multi-objective) are the par excellence tool to deal with it. Interesting suggestions could be found in [16] and [25]. One of the most cited papers related to this topic is [30].

2.1.3 Better alternatives to the Support and Confidence measures

This topic is seen from more practical point of view in Subsection 3.1.2. Here is just mentioned some of the measures proposed in literature (in not an exhaustive list):

- Support
- Confidence
- Significance factor [19]
- Certainty factor [19]
- Correlation [26]
- Adjusted difference [19]
- Weight of evidence [19]
- Unusualness of a rule [27]
- Dependence [20]
By far the most used are still Support and Confidence. So the methodology is suggested to use both, but emphases in Confidence and subsection 3.1.2 details the exact definition of each of the calculations. Obviously there is not a dependency from them. Apply another measure of quality could be as long as it is used for all algorithms to compare; in this case, Confidence is chosen to evaluate the quality.

2.1.4 Specific applications: classification, subgroups discovery

Fuzzy rules, as well as being a representation of knowledge a priori unknown, can be used to build other knowledge, if you will, more elaborate. For example, even before the birth of fuzzy association rules, classifiers were constructed based on fuzzy rules, as in [17] and [6]. But this is not the only specific purpose of extracting fuzzy rules, since [18], the fuzzy rules began to facilitate the extraction and interpretation of knowledge on the Subgroup Discovery problem. Related with Sub-groups discovery interesting proposal appear in [12].

On the other hand, some authors have been concerned with specific problems. One interesting is: how to deal with rare itemsets? explicitly when the rare rules are the objective en first place. Regarding the search of rare itemsets an analysis is found in [34].

2.2 Classifying algorithms

Classifying algorithms itself is not an exact area, even seems closer to art than science. For scope of this research the classification is made through the point of view of a final user, no necessarily an expert en fuzzy systems. So a good way to classify them, is through the capabilities that each possesses intrinsic. In this context, it is clear that choosing a Fuzzy association algorithm means to answer questions about these capabilities. We have defined at least 4 basic crucial questions:

1. Which is the purpose?
2. What is size of database?
3. Which structure is desired for final knowledge?
4. Is there previous knowledge?

Those user decisions are summary in Figure 5 and will support our classification methodology and the selection of the algorithms analyzed.

Is clear we can separate, in not a exclusive manner, the algorithm by the capability to answer positive o negative to those questions.

2.2.1 Purpose

There are two principal reasons to extract fuzzy association rules, to classify (predictive) or to discover knowledge or descriptive (to create profiles, to discover subgroups). Generally, a predictive nature is related with supervised
Fig. 5: Classification Map with decisions to select a specific algorithm
learning methods, while a descriptive nature is more close to non-supervised ones. Classifiers based on association rules usually are designed with multi-objective meta heuristics, as a mechanism to increase accuracy with the less quantity of rules possible. There are interesting proposals, some recently are [24] by Nojima, Kaisho and Ishibuchi; [31] by Sowan, Dahal, Hossain, and Alam; and applications to novelty topics as Intruders detection in [33] by Tajbakhsh, Rahmati and Mirzaei.

2.2.2 Size of database

This a is a cross-cutting issue to Data Mining, not just for Fuzzy association rules. The scalability of the algorithm makes an enormous difference, as pointed out by one of the 10 most important challenges in Data Mining [36]. Moreover, is a important problem no just for the quantity of tuples (instances) of the dataset, but including the potential combinatorial explosion of itemsets resulting of fuzzyfication of every variable (attribute). For example, a BFS strategy, as in Apriori, is more sensitive to the number of variables than the number of tuples. Given the importance of this in the performance of any algorithm, is included in the methodology of comparison and is discussed in subsection 3.2.3.

2.2.3 Structure for final knowledge

In some important applications is desirable or necessary, to define previously some aspects of final knowledge acquired. For example, could be desired to limit the number of antecedents in every rule, or find (or not) conjunction-disjunction in the rules, linguistics edge. To make proposals for desired structure of knowledge is needed to have greater knowledge in association rules (fuzzy or not). Typically, the average user would not know how to answer this question, so it is left to the expert or consultant to explain the meaning of each item and the need for either.

2.2.4 Previous knowledge

After, and very related with previous, the question lies in whether or not there is prior knowledge. This could be in the way of Membership functions (for example: triangular, trapezoidal, fuzzy C means, etc), Number of linguistics labels (3, 5, etc), Consequent known (commonly to identifier variables than explain others), and Number of antecedents desired. An example is to characterize a specific consequence in terms of all possible antecedents. A very interesting application of the above could be found in [28] by Sanchez, Vila, Cerda and Serrano. Other case is bounding the possible antecedents of the rules, as in [25] by Orriols and Casillas. The fastest way to answer this question would be “learning everything”.

2.3 Selected algorithms

Based on this Classification Map some algorithms are chosen to test our comparison methodology. The requirements were: Knowledge discovery + Not necessarily high scalability + Basic form of rules + Some requirements of previous knowledge. The algorithms were FARMOR [22] (which uses fuzzy C means for labeling), FGBRMA [15] (which uses triangular labeling), and FCSar [23] (permitted desired structure of knowledge even though is not addressed in this paper).

3 Methodology for comparison

3.1 Before the experiment

To propose a methodology of comparison there are at least 3 aspects to considered before the experiment. Henceforth it will be called experimental framework:

1. Guaranties an equivalent pre-processing method (or understand there is a bias using one or another)

2. Uses the same minimum support and confidence definition (and t-norm for itemsets intersections)

3. Applying the algorithms to databases public and accessible

In the next 3 subsections discuss each of them. This experimental framework is resume in Fig 6.

3.1.1 Pre-processing

Any Fuzzy Association Rules algorithm requires some pre-processing process which mainly involves creation of fuzzy partitions either using an expert-driven
approach or a data-driven approach. The way to do this, is not addressed or clear in many proposals, even though implies not a trivial questions and not for just select triangular membership functions is everything clear. For example: selecting 3 triangular functions a question arises:

- Which form do the triangles have?

This is related at form triangles are distributed in domain. In Figure [7] are the two principal different ways to distributed triangular label in the domain [1, 20]. The first are used for example in [3], while the second appears for example in [15]

The questions are not less when 1-dimension Fuzzy C Means algorithm [4] is applied to each quantitative variable:

- Which \( m \) value use? (see details of algorithm)

- When to stop the re-calculation of centroids (means)?

In Figure [8] is the fuzzy partions derived to applied Fuzzy C means to the same domain [1,20]. It was applied a \( m \) value = 2 and a re-calculation point equal to 0.0055

It is clear with the 3 different partions different membership values correspond to same singleton. In example, Table [1] recovered the memberships for the singleton “6” in the mentioned domain.

To compare graphically both labeling techniques, the result of labeling first variable (“Sepal length”) in dataset Iris is presented in Figure [9].

It is clear that different \( k \)-itemsets resulting from such membership values belonging to the every singleton. Even is not an issue to defined here which labels are better, is important to mention the possible differences results. So, is part of our methodology to guaranties the same conditions for all algorithms
To test the impact of one or another form of labeling we design the following experiment: comparing only Triangular labels type 2 (figure 7b) and Fuzzy C Means (figure 8), the domain of every variable in Iris dataset was labeled.

Applying a BFS technique to combined itemsets (as in [22]), 5 labels (Very-Low, Low, Middle, High, VeryHigh) and Support = 0.05 the results in quantity of itemsets found is clear: with triangular labels the number of itemsets is less than using fuzzy C means. The results are shown in Figure 10.

It demonstrates that very different results are expected in the future construction of the fuzzy rules. Then, in a comparison of proposals should be clear whether to use the original labels proposed by authors or if instead they will standardize. Understanding that the same algorithm can produce different results depending on the labeling.

### 3.1.2 Measures of quality

For make a fair comparison between algorithms, the second element to clearly define is how to assess the quality of the fuzzy rules. As mentioned in subsection 2.1.3 other measures of quality for fuzzy rules have been proposed, for example: significance factor and certainty factor. However, not being a theoretical analysis, but rather practical, choose to use support and confidence, because of its dissemination among researchers and its easier interpretation for non-experts in fuzzy logic.

Defined above, the next logical question is the definition of support and confidence to use. As Dubois, H"ullermeier and Prade mentioned in [9], the usual
Fig. 9: Labels *Triangular* and *Fuzzy C means* for Sepal length variable of Iris dataset

Fig. 10: Itemsets found with Triangular labels type 2 and Fuzzy C means for Iris dataset
t-norm $\otimes$ is the minimum, yet, in a previous analysis in [8], Dubois, Prade and Sudkamp demonstrated that, in a more intuitive way, the intersection by product uses information from both sets to intersect. Indeed, they show empirically that, using minimum, confidence measures will be independent of the value of one set in the whenever their membership values satisfied that are less than the membership values of the other set to intersect. By this analysis, is suggest to use the product t-norm

In similar way, based on analysis in [8] and continued in [9], is suggested to include a relevance of an object $x$ for the rule. For support and confidence definitions is follow the recommendation to use an implication, in this case Dienes implication: $\max \{1 - x, y\}$

Therefore, the definitions used are as follows:

$$Sup = \frac{1}{n} \sum \mu_A(x) \cdot \mu_C(x)$$ (1)

$$Conf = \frac{\sum \mu_A(x) \cdot \max\{1 - \mu_A(x), \mu_C(x)\}}{\sum \mu_A(x)}$$ (2)

Where $\mu_A(x)$ represents the membership of the itemsets in Antecedent and $\mu_C(x)$ the membership in the Consequent

### 3.1.3 Selected datasets

As stated above, a fair comparison requires other researchers could repeat the experiments. So, instead of using artificial databases, the experiments were undertaken with real datasets from UCI repository [10]. The properties of these datasets are presented in Table 2: identifier of the data set (Id), the number of features (#Fea), the number of instances (#Ins), the number of continuous features (#Re), the number of integer features (#In), the number of nominal features (#No)

In figure 11 complexity is approximated graphically for each of the databases. Clearly, the more complicated databases are in the upper right of the graph, where most of the variables are continuous, and also it increases with the number of instances (represented by bubble size)

### 3.2 After experiments

Once the experiments, the next step is the actual comparison of the results. This is divided in four:

1. Scalability of the algorithm
2. Compare the quality of the rules, understood as $Confidence$, (equation 2)
3. Compare the diversity of rules

\(^2\) All datasets was taken away the class attribute
Fig. 11: Difficulty of the databases (Size of fam95 appears in other scale for clear representation)
Tab. 2: Properties of the datasets considered for the experimental study

<table>
<thead>
<tr>
<th>Id.</th>
<th>#Fea</th>
<th>#Ins</th>
<th>#Re</th>
<th>#In</th>
<th>#No</th>
</tr>
</thead>
<tbody>
<tr>
<td>appendicitis</td>
<td>7</td>
<td>106</td>
<td>7</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>bupa</td>
<td>6</td>
<td>345</td>
<td>1</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>Fam95</td>
<td>10</td>
<td>63756</td>
<td>0</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>Glass</td>
<td>9</td>
<td>214</td>
<td>9</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>heart-statlog</td>
<td>13</td>
<td>270</td>
<td>1</td>
<td>12</td>
<td>0</td>
</tr>
<tr>
<td>iris</td>
<td>4</td>
<td>150</td>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>magic</td>
<td>10</td>
<td>19020</td>
<td>10</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>phoneme</td>
<td>5</td>
<td>5404</td>
<td>5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>pima_diabetes</td>
<td>8</td>
<td>768</td>
<td>8</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ring</td>
<td>20</td>
<td>7400</td>
<td>20</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>segment</td>
<td>19</td>
<td>2310</td>
<td>19</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>sonar</td>
<td>60</td>
<td>208</td>
<td>60</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>new-thyroid</td>
<td>5</td>
<td>215</td>
<td>4</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>fis</td>
<td>21</td>
<td>216</td>
<td>21</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>vehicle</td>
<td>18</td>
<td>846</td>
<td>0</td>
<td>18</td>
<td>0</td>
</tr>
<tr>
<td>waveform</td>
<td>40</td>
<td>5000</td>
<td>40</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>wdbc</td>
<td>30</td>
<td>369</td>
<td>30</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Wine</td>
<td>13</td>
<td>178</td>
<td>13</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>yeast</td>
<td>8</td>
<td>1484</td>
<td>8</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

4. Statistical tests to confirm or discard differences in performance

Which we call experimental analysis. In next subsections are described each of them. Figure 12 shows a schematic summarizing

3.2.1 Quality of rules

We apply the follow analysis: included the results of Quality in terms of Confidence average and Confidence median all. Additionally, rules generated are sorted based on their Confidence value. If are used, for example, 0.05 for Support itemsets and minimum Confidence of 0.7 for the rules, thousands of rules can be obtained; in results table, only 25% of the highest Confidence values are included (25% percentile)

3.2.2 Diversity of rules

For diversity approach, is applied the methodology defined in [31] by Sowan, Dahal, Hossain and Alam to quantifier the “distance” within fuzzy association rules. Exactly, the procedure is as follows: be $S_{ij}$ the similarity between two Fuzzy association rules (i and j) and $RAFS_i$ (Rule attribute fuzzy set) : the fuzzy set concerning an attribute within the Fuzzy Rule_i and $RAFS_j$: the fuzzy set concerning an attribute within the Fuzzy Rule_j, then:
3 Methodology for comparison

![Flowchart showing the steps of comparison: Compare algorithms scalability, Evaluate rules quality, Evaluate rules diversity, Statistical tests.]

Fig. 12: Experimental analysis

\[ S_{ij} = \frac{|RAFS_i \cap RAFS_j|}{|RAFS_i \cup RAFS_j|} \]  

(3)

And, the distance is calculated as the opposite, using equation:

\[ Distance_{ij} = 1 - S_{ij} \]  

(4)

These equations were taken by the authors of [7]. Were proposed to maintain diversity within the sharing function technique in multi-objective optimization. An example illustrates the calculation. Let the rules (i and j):

- “If High_Petal-width and Low_Sepal-width then High_Petal-length”
- “If High_Petal-width and Middle_Sepal-length then High_Petal-length”

From Iris dataset. So, \( RAFS_i \) is \([\text{High_Petal-width, Low_Sepal-width, High_Petal-length}]\) and \( RAFS_j \) is \([\text{High_Petal-width, Middle_Sepal-length, High_Petal-length}]\), then using Equation (3) the \( S_{ij} \) is:

\[ S_{ij} = \frac{|2|}{|4|} \]

Numerator comes from the number of \( RAFS_i \cap RAFS_j \); \([\text{High_Petal-width, High_Petal-length}]\) and denominator comes from all attributes in \( RAFS_i \cup RAFS_j \); \([\text{High_Petal-width, Low_Sepal-width, Middle_Sepal-length, High_Petal-length}]\). Therefore, the \( Distance_{ij} \) is: 0.5. For each rule, is calculated the average distance with all the rules found from same algorithm and with the same number of Antecedents, then is analyzed the 25% of highest Distance (average per rule) by algorithm.
3.2.3 Scalability of the algorithm

Related this aspect, is important not only to analyze the scalability in terms of number of tuples (instances) but also in terms of the number of variables. For this purpose is select a dataset and is progressively increased both the number of instances (until complete the original total of instances) as the number of variables (until complete the original total of variables).

4 Applying the comparison methodology (results)

As we mentioned in Subsection 2.3, to test the proposed methodology, a comparison between “Fuzzy grids based rules mining algorithm” (FGBRMA) [15] and “Fuzzy association rule mining algorithm for fast and efficient performance on very large datasets” (F-ARMOR) [22] and “Fuzzy Classifier System for Association Rule mining” (FCSar) [25] is presented. In Figure 2 the proposals selected to compare appear marked with a different symbol.

Each algorithm was tested using 5 labels, triangular and type 2 in FGBRMA and Fuzzy C Means in FARMOR (m = 2, stop the re-calculation in 0.005). Instead of Support and Confidence definitions proposed by the original authors, have been used the equations 1 and 2 detailed in subsection 3.1.2.

4.1 Scalability of algorithm

For this analysis, we proceed as described in subsection 3.2.3. The dataset chosen is Waveform and 5 turns were applied for every analysis. In case of FGBRMA and FARMOR, both were implemented in the Java language, so they are comparable in efficiency. The Figure 13 shows the behavior (time consuming) of FGBRMA and FARMOR when the number of tuples is increasing. The Support is 0.1 and the Confidence 0.7.

Clearly FGBRMA behavior is so different that does not reflect well the graph. Therefore, we proceed to present only their performance in Figure 14.

Then, proceed in similar way gradually increasing the number of variables until complete the 40 variables belonging to original database: For both algorithms, the performance is presented in Figure 15.

As in case of variables increase, it is notable again that FGBRMA is located on a different scale by which we proceed to plot separately. It appears in Figure 16.

It is clear that both algorithms, increases the time consumption linearly with the number of rows and exponentially with the number of variables (both apply BFS technique to find large-itemsets). However, is equally clear that the execution is faster for FGBRMA over FARMOR, the cause of this could be the difference in the number of itemsets we get to a form of labeling in the other (see Subsection 3.1.1) and the compression method apply to the Tid-list vectors in FARMOR, which maybe represents an extra time consumption.

The other compare algorithm FCSar was implemented in C++ language, so can not be compared with others in terms of efficiency (plotted together), but if
Fig. 13: FGBRMA and FARMOR behavior increasing the number of instances over Waveform dataset

Fig. 14: FGBRMA behavior increasing the number of instances over Waveform dataset
Fig. 15: FGBRMA and FARMOR behavior increasing the number of variables over Waveform dataset

Fig. 16: FGBRMA behavior increasing the number of variables over Waveform dataset
4 Applying the comparison methodology (results)

Fig. 17: FCSar behavior increasing the number of instances over Waveform dataset

their behavior can be seen in similar tests. Increasing the number of variables, presents a non-exponential growth as is shown in Figure 17.

Proceeding to rise the number of variables, the performance is very linear, this is submitted in Figure

4.2 Quality of the rules (by dataset)

An example with Appendicitis dataset

First, applying FARMOR and FGBRMA, the total of fuzzy rules obtained from app dataset the results are shown in Figure 19. There is indicated the percentile 25% of rules with highest Confidence.

And rules pertaining to the highest 25% percentile shown in more detail in Figure 20

Then presents results for some of the databases. This is done because the search technique in FGBRMA and FARMOR are the same (BFS technique). View of the foregoing, the results between the two are basically the results between the method of labeling. Therefore, both are applied over some of datasets to prove the results between label Triangular type2 and Fuzzy C Mean (m = 2, stop point = 0.005), Support = 0.05, Confidence = 0.7

Cleared that, the total results of Confidence appear in Table 3

The results show that FARMOR seems better results in terms of the quality, but it will be confirm or discard by statistical test in Subsection 4.4.

Following, is the experiment propose to compare FGBRMA and the third algorithm FCSar in all datasets (creation based on Consequent, number of explores = 100 000). Unlike the previous experiment, where basically comparing
Fig. 18: FCSar behavior increasing the number of variables over Waveform dataset

Fig. 19: Confidence for all rules obtained in Appendicitis dataset (FARMOR vs FGBRMA)
4 Applying the comparison methodology (results)

Fig. 20: Percentile 25% of higher confidence rules obtained in *appendicitis* dataset. See Figure 19

<table>
<thead>
<tr>
<th>Id.</th>
<th>Rules in highest 25% percentile</th>
<th>Confidence average</th>
<th>Confidence median</th>
</tr>
</thead>
<tbody>
<tr>
<td>appendicitis</td>
<td>2</td>
<td>0.765</td>
<td>0.763</td>
</tr>
<tr>
<td>bupa</td>
<td>8</td>
<td>0.753</td>
<td>0.747</td>
</tr>
<tr>
<td>Glass</td>
<td>222</td>
<td>0.831</td>
<td>0.822</td>
</tr>
<tr>
<td>iris</td>
<td>0</td>
<td>0.757</td>
<td>0.757</td>
</tr>
<tr>
<td>magic</td>
<td>200</td>
<td>0.787</td>
<td>0.787</td>
</tr>
<tr>
<td>phoneme</td>
<td>0</td>
<td>0.738</td>
<td>0.733</td>
</tr>
<tr>
<td>pima_diabetes</td>
<td>55</td>
<td>0.768</td>
<td>0.747</td>
</tr>
<tr>
<td>Wine</td>
<td>1</td>
<td>0.732</td>
<td>0.725</td>
</tr>
</tbody>
</table>

Tab. 3: Results of quality FGBRMA and FARMOR
forms of labeling, in this case is compared the way to perform the search within the space of itemsets. We proceed in the same way, to graph only Appendicitis for comparison FGBRMA vs FCSar, Figures 21 and 22, and complete results appear in Table 4.

4.3 Diversity of the rules (by dataset)

An example with Appendicitis dataset

Applying FARMOR and FGBRMA, the average of Distance between rules (for each rule) obtained from Appendicitis dataset are shown in Figure 23. There is indicated the percentile 25% of rules with highest Diversity.

And rules rules pertaining to the 25% percentile shown in more detail in Figure 24.

Then presents the results for some of the databases, for FGBRMA and FARMOR in the same way as Subsection 4.2. All appear in Table 5.

The results show that FARMOR get better results in terms of the diversity. Following, the experiment propose to compare FGBRMA and third algorithm FCSar in all datasets (creation based on Consequent, number of explores = 100 000). Graphs showing results are Figures 25 and 26 and results appear in Table 6.
4 Applying the comparison methodology (results)

Fig. 22: Percentile 25% of higher confidence rules obtained in *appendicitis* dataset (FGBRMA vs FCSar). See Figure 21

<table>
<thead>
<tr>
<th>Id.</th>
<th>Rules in highest 25% percentile</th>
<th>Confidence average</th>
<th>Confidence median</th>
</tr>
</thead>
<tbody>
<tr>
<td>appendicitis</td>
<td>12</td>
<td>0.765</td>
<td>0.763</td>
</tr>
<tr>
<td>bupa</td>
<td>6</td>
<td>0.753</td>
<td>0.747</td>
</tr>
<tr>
<td>Fam95</td>
<td>521</td>
<td>0.842</td>
<td>0.830</td>
</tr>
<tr>
<td>Glass</td>
<td>124</td>
<td>0.831</td>
<td>0.822</td>
</tr>
<tr>
<td>heart-statlog</td>
<td>2056</td>
<td>0.840</td>
<td>0.845</td>
</tr>
<tr>
<td>iris</td>
<td>0</td>
<td>0.757</td>
<td>0.757</td>
</tr>
<tr>
<td>magic</td>
<td>272</td>
<td>0.793</td>
<td>0.793</td>
</tr>
<tr>
<td>phoneme</td>
<td>0</td>
<td>0.738</td>
<td>0.733</td>
</tr>
<tr>
<td>pima_diabetes</td>
<td>21</td>
<td>0.768</td>
<td>0.747</td>
</tr>
<tr>
<td>ring</td>
<td>6920</td>
<td>0.764</td>
<td>0.801</td>
</tr>
<tr>
<td>segment</td>
<td>71500</td>
<td>0.904</td>
<td>0.920</td>
</tr>
<tr>
<td>sonar</td>
<td>15775</td>
<td>0.730</td>
<td>0.725</td>
</tr>
<tr>
<td>new-thyroid</td>
<td>28</td>
<td>0.921</td>
<td>0.816</td>
</tr>
<tr>
<td>fis</td>
<td>8451</td>
<td>0.832</td>
<td>0.848</td>
</tr>
<tr>
<td>vehicle</td>
<td>1</td>
<td>0.766</td>
<td>0.754</td>
</tr>
<tr>
<td>waveform</td>
<td>5735</td>
<td>0.713</td>
<td>0.711</td>
</tr>
<tr>
<td>wdbc</td>
<td>14952</td>
<td>0.780</td>
<td>0.773</td>
</tr>
<tr>
<td>Wine</td>
<td>20</td>
<td>0.732</td>
<td>0.725</td>
</tr>
<tr>
<td>yeast</td>
<td>239</td>
<td>0.883</td>
<td>0.856</td>
</tr>
</tbody>
</table>

Tab. 4: Results of quality FGBRMA and FCSar
Fig. 23: Distance (average) for all rules obtained in *appendicitis* dataset (FARMOR vs FGBRMA)

Fig. 24: Percentile 25% of higher Distance (average per rule) obtained in *appendicitis* dataset (See Figure 23)
Rules in 25% percentile  |  Confidence average  |  Confidence median
---|---|---
appendicitis | 21 | 519 | 0.812 | 0.878 | 0.839 | 0.899
bupa | 41 | 105 | 0.818 | 0.748 | 0.817 | 0.756
Glass | 73 | 1910 | 0.751 | 0.843 | 0.722 | 0.869
iris | 14 | 27 | 0.808 | 0.822 | 0.781 | 0.830
magic | 65 | 1294 | 0.828 | 0.888 | 0.884 | 0.893
phoneme | 1 | 56 | 0.771 | 0.859 | 0.771 | 0.875
pima_diabetes | 56 | 329 | 0.842 | 0.818 | 0.835 | 0.818
Wine | 7 | 2638 | 0.915 | 0.941 | 0.913 | 0.939

Tab. 5: Results of diversity FGBRMA and FARMOR

Fig. 25: Distance (average) for all rules obtained in *appendicitis* dataset (FCSar vs FGBRMA)
Fig. 26: Percentile 25% of higher Distance (average per rule) obtained in *appendicitis* dataset (FCSar vs FGBRMA) (See Figure 25)

<table>
<thead>
<tr>
<th></th>
<th>Rules in 25% percentile</th>
<th>Confidence average</th>
<th>Confidence median</th>
</tr>
</thead>
<tbody>
<tr>
<td>appendicitis</td>
<td>26  111</td>
<td>0.811</td>
<td>0.839</td>
</tr>
<tr>
<td>bupa</td>
<td>34  85</td>
<td>0.811</td>
<td>0.865</td>
</tr>
<tr>
<td>Fam95</td>
<td>411 148</td>
<td>0.753</td>
<td>0.733</td>
</tr>
<tr>
<td>Glass</td>
<td>101 311</td>
<td>0.751</td>
<td>0.722</td>
</tr>
<tr>
<td>heart-statlog</td>
<td>1959 138</td>
<td>0.821</td>
<td>0.844</td>
</tr>
<tr>
<td>iris</td>
<td>70  58</td>
<td>0.808</td>
<td>0.781</td>
</tr>
<tr>
<td>magic</td>
<td>166 194</td>
<td>0.828</td>
<td>0.884</td>
</tr>
<tr>
<td>phoneme</td>
<td>1 72</td>
<td>0.771</td>
<td>0.771</td>
</tr>
<tr>
<td>pima_diabetes</td>
<td>6  175</td>
<td>0.842</td>
<td>0.860</td>
</tr>
<tr>
<td>ring</td>
<td>6951 368</td>
<td>0.900</td>
<td>0.877</td>
</tr>
<tr>
<td>segment</td>
<td>43497 111</td>
<td>0.829</td>
<td>0.851</td>
</tr>
<tr>
<td>sonar</td>
<td>42 323</td>
<td>0.842</td>
<td>0.818</td>
</tr>
<tr>
<td>new-thyroid</td>
<td>11 66</td>
<td>0.727</td>
<td>0.757</td>
</tr>
<tr>
<td>fis</td>
<td>8084 450</td>
<td>0.923</td>
<td>0.894</td>
</tr>
<tr>
<td>vehicle</td>
<td>689 320</td>
<td>0.912</td>
<td>0.936</td>
</tr>
<tr>
<td>waveform</td>
<td>5995 367</td>
<td>0.968</td>
<td>0.970</td>
</tr>
<tr>
<td>wdbc</td>
<td>14781 429</td>
<td>0.927</td>
<td>0.918</td>
</tr>
<tr>
<td>Wine</td>
<td>36 225</td>
<td>0.915</td>
<td>0.913</td>
</tr>
<tr>
<td>yeast</td>
<td>160 99</td>
<td>0.781</td>
<td>0.787</td>
</tr>
</tbody>
</table>

Tab. 6: Results of quality FGBRMA and FCSar
4 Applying the comparison methodology (results)

<table>
<thead>
<tr>
<th>Comparison</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Confidence FGBRMA-FARMOR</td>
<td>Test reject null hypotheses</td>
</tr>
<tr>
<td>Diversity FGBRMA-FARMOR</td>
<td>Test accept null hypotheses</td>
</tr>
<tr>
<td>Confidence FGBRMA-FCSar</td>
<td>Test accept null hypotheses</td>
</tr>
<tr>
<td>Diversity FGBRMA-FCSar</td>
<td>Tests reject null hypotheses</td>
</tr>
</tbody>
</table>

Tab. 7: Main results of statistical tests

4.4 Statistical test

4.4.1 Main results and interpretation

As is increasingly clear in Data Mining field, only achieve good results is not enough, its required statistical certainty for claims between algorithms. In Subsection 3.2 is mentioned as part of the methodology compared using statistical test. For that purpose, is decided use statistical tests proposed (and associated software) by Garcia and Herrera in [11]. Basically the test rank the algorithms and then accept or reject the null hypotheses, which permitted to find performance differences between them.

The main question is which measure use to compare?, as is well known the median is measures of central tendency more appropriate when there are biases in the data or outliers present. This robustness leads to consider it as the summary measure of confidence and diversity of the rules found by the algorithms. Thus, we compared the Confidence median and Diversity median in the tests. All data appears are in Tables in Subsections 4.2 and 4.3.

In subsequent pages are the complete results in detail, but in Table 7 may be a summary of main results:

The interpretation of results is as follows: the are statistical differences between the algorithms FGBRMA and FARMOR in terms of quality, which leads to a difference between labeling method. In diversity, is not a statistical difference, even were closed to reject hypotheses, so, more experiments are needed here. There are not conclusive differences in quality of rules between FGBRMA and FCSar but is there in terms of Diversity. All this implies that labeling method itself influences the final results of rules found, and second, an evolutionary algorithm as FCSar (with a linear scalability) produces rules as good as a deterministic algorithm (exponential scalability) in terms of quality and better in terms of diversity (which can be interpreted as a broader knowledge found)

Completed results appear next
4.4.2 Confidence FGBRMA-FARMOR (Tables of Friedman, Bonferroni-Dunn, Holm, Hochberg and Hommel Tests)

Complete tests appear next:

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Ranking</th>
</tr>
</thead>
<tbody>
<tr>
<td>FGBRMA</td>
<td>1.875</td>
</tr>
<tr>
<td>FARMOR</td>
<td>1.125</td>
</tr>
</tbody>
</table>

Tab. 8: Average Rankings of the algorithms

Friedman statistic considering reduction performance (distributed according to chi-square with 1 degrees of freedom: 4.5. P-value computed by Friedman Test: 0.03389485352630239.

Iman and Davenport statistic considering reduction performance (distributed according to F-distribution with 1 and 7 degrees of freedom: 9.0. P-value computed by Iman and Davenport Test: 0.01994212613199242.

<table>
<thead>
<tr>
<th>i</th>
<th>algorithm</th>
<th>(z = (R_0 - R_i)/SE)</th>
<th>(p)</th>
<th>Holm/Hochberg/Hommel</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>FGBRMA</td>
<td>2.1213203435596424</td>
<td>0.03389</td>
<td>0.05</td>
</tr>
</tbody>
</table>

Tab. 9: Holm / Hochberg Table for \(\alpha = 0.05\)

Bonferroni-Dunn’s procedure rejects those hypotheses that have a p-value \(\leq 0.05\.

Hochberg’s procedure rejects those hypotheses that have a p-value \(\leq 0.05\.

Hommel’s procedure rejects all hypotheses.

Bonferroni-Dunn’s procedure rejects those hypotheses that have a p-value \(\leq 0.1\.

Hochberg’s procedure rejects those hypotheses that have a p-value \(\leq 0.1\.

Hommel’s procedure rejects all hypotheses.
\( i \) algorithm \( z = (R_0 - R_i)/\text{SE} \) \( p \) Holm/Hochberg/Hommel

<table>
<thead>
<tr>
<th>( i )</th>
<th>algorithms</th>
<th>( z = (R_0 - R_i)/\text{SE} )</th>
<th>( p )</th>
<th>Holm/Hochberg/Hommel</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>FGBRMA</td>
<td>2.1213203435596424</td>
<td>0.03389</td>
<td>0.1</td>
</tr>
</tbody>
</table>

**Tab. 10: Holm / Hochberg Table for \( \alpha = 0.10 \)**

\( i \) algorithm unadjusted \( p \) \( p_{\text{Bonf}} \) \( p_{\text{Holm}} \) \( p_{\text{Hoch}} \) \( p_{\text{Homm}} \)

<table>
<thead>
<tr>
<th>( i )</th>
<th>algorithms</th>
<th>unadjusted ( p )</th>
<th>( p_{\text{Bonf}} )</th>
<th>( p_{\text{Holm}} )</th>
<th>( p_{\text{Hoch}} )</th>
<th>( p_{\text{Homm}} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>FGBRMA</td>
<td>0.03389</td>
<td>0.03389</td>
<td>0.03389</td>
<td>0.03389</td>
<td>0.03389</td>
</tr>
</tbody>
</table>

**Tab. 11: Adjusted \( p \)-values**

\( i \) algorithms \( z = (R_0 - R_i)/\text{SE} \) \( p \) Holm Shaffer

<table>
<thead>
<tr>
<th>( i )</th>
<th>algorithms</th>
<th>( z = (R_0 - R_i)/\text{SE} )</th>
<th>( p )</th>
<th>Holm</th>
<th>Shaffer</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>FGBRMA vs. FARMOR</td>
<td>2.1213203435596424</td>
<td>0.03389</td>
<td>0.1</td>
<td>0.1</td>
</tr>
</tbody>
</table>

**Tab. 12: Holm / Shaffer Table for \( \alpha = 0.05 \)**

Nemenyi's procedure rejects those hypotheses that have a \( p \)-value \( \leq 0.05 \).
Shaffer’s procedure rejects those hypotheses that have a \( p \)-value \( \leq 0.05 \).
Bergmann’s procedure rejects these hypotheses:

- FGBRMA vs. FARMOR

**Tab. 13: Holm / Shaffer Table for \( \alpha = 0.10 \)**

Nemenyi's procedure rejects those hypotheses that have a \( p \)-value \( \leq 0.1 \).
Shaffer’s procedure rejects those hypotheses that have a \( p \)-value \( \leq 0.1 \).
Bergmann’s procedure rejects these hypotheses:
- **FGBRMA vs. FARMOR**

<table>
<thead>
<tr>
<th>i</th>
<th>hypothesis</th>
<th>unadjusted $p$</th>
<th>$p_{Norm}$</th>
<th>$p_{Holm}$</th>
<th>$p_{Haf}$</th>
<th>$p_{Erg}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>FGBRMA vs. FARMOR</td>
<td>0.03389</td>
<td>0.03389</td>
<td>0.03389</td>
<td>0.03389</td>
<td>0.03389</td>
</tr>
</tbody>
</table>

Tab. 14: Adjusted $p$-values
4.4.3 Diversity FGBRMA-FARMOR (Tables of Friedman, Bonferroni-Dunn, Holm, Hochberg and Hommel Tests)

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Ranking</th>
</tr>
</thead>
<tbody>
<tr>
<td>FGBRMA</td>
<td>1.75</td>
</tr>
<tr>
<td>FARMOR</td>
<td>1.25</td>
</tr>
</tbody>
</table>

Tab. 15: Average Rankings of the algorithms

Friedman statistic considering reduction performance (distributed according to chi-square with 1 degrees of freedom: 2 0).
P-value computed by Friedman Test: 0.15729920705145928.

Iman and Davenport statistic considering reduction performance (distributed according to F-distribution with 1 and 7 degrees of freedom: 2.3333333333333335.
P-value computed by Iman and Davenport Test: 0.17047066078705386.

<table>
<thead>
<tr>
<th>algorithm</th>
<th>z</th>
<th>p</th>
<th>Holm/Hochberg/Hommel</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>FGBRMA</td>
<td>1.414213562373095</td>
<td>0.15729</td>
</tr>
</tbody>
</table>

Tab. 16: Holm / Hochberg Table for $\alpha = 0.05$

Bonferroni-Dunn’s procedure rejects those hypotheses that have a p-value $\leq 0.05$.
Holm’s procedure rejects those hypotheses that have a p-value $\leq 0.05$.
Hommel’s procedure rejects those hypotheses that have a p-value $\leq 0.05$.
Bonferroni-Dunn’s procedure rejects those hypotheses that have a p-value $\leq 0.1$.
Holm’s procedure rejects those hypotheses that have a p-value $\leq 0.1$.
Hommel’s procedure rejects those hypotheses that have a p-value $\leq 0.05$.
Nemenyi’s procedure rejects those hypotheses that have a p-value $\leq 0.05$. 
Applying the comparison methodology (results)

\[ i \text{ algorithm } z = (R_0 - R_i)/SE \quad p \quad \text{Holm/Hochberg/Hommel} \]

<table>
<thead>
<tr>
<th>( i )</th>
<th>FGBRMA</th>
<th>1.414213562373095</th>
<th>0.15729</th>
<th>0.1</th>
</tr>
</thead>
</table>

**Tab. 17:** Holm / Hochberg Table for \( \alpha = 0.10 \)

\[ i \text{ algorithm } \text{unadjusted } p \quad PB_{Bonf} \quad PH_{Holm} \quad PH_{Hoch} \quad PH_{Homm} \]

<table>
<thead>
<tr>
<th>( i )</th>
<th>FGBRMA</th>
<th>0.15729</th>
<th>0.15729</th>
<th>0.15729</th>
<th>0.15729</th>
</tr>
</thead>
</table>

**Tab. 18:** Adjusted \( p \)-values

\[ i \text{ algorithms } z = (R_0 - R_i)/SE \quad p \quad \text{Holm} \quad \text{Shaffer} \]

<table>
<thead>
<tr>
<th>( i )</th>
<th>FGBRMA vs. FARMOR</th>
<th>1.414213562373095</th>
<th>0.15729</th>
<th>0.05</th>
<th>0.05</th>
</tr>
</thead>
</table>

**Tab. 19:** Holm / Shaffer Table for \( \alpha = 0.05 \)

Holm’s procedure rejects those hypotheses that have a \( p \)-value \( \leq 0.05 \).

Bergmann’s procedure rejects these hypotheses:

\[ i \text{ algorithms } \]

<table>
<thead>
<tr>
<th>( i )</th>
<th>FGBRMA vs. FARMOR</th>
<th>1.414213562373095</th>
<th>0.15729</th>
<th>0.1</th>
<th>0.1</th>
</tr>
</thead>
</table>

**Tab. 20:** Holm / Shaffer Table for \( \alpha = 0.10 \)

Nemenyi’s procedure rejects those hypotheses that have a \( p \)-value \( \leq 0.1 \).

Holm’s procedure rejects those hypotheses that have a \( p \)-value \( \leq 0.1 \).

Bergmann’s procedure rejects these hypotheses:
### Table 21: Adjusted *p*-values

<table>
<thead>
<tr>
<th>i</th>
<th>hypothesis</th>
<th>unadjusted <em>p</em></th>
<th><em>PNmc</em></th>
<th><em>PHolm</em></th>
<th><em>PS1-tuf</em></th>
<th><em>PBerg</em></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>FGBRMA vs. FARMOR</td>
<td>0.15729</td>
<td>0.15729</td>
<td>0.15729</td>
<td>0.15729</td>
<td>0.15729</td>
</tr>
</tbody>
</table>

Applying the comparison methodology (results)
4.4.4 Confidence FGBRMA-FCSar (Tables of Friedman, Bonferroni-Dunn, Holm, Hochberg and Hommel Tests)

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Ranking</th>
</tr>
</thead>
<tbody>
<tr>
<td>FGBRMA</td>
<td>1.578947368421052</td>
</tr>
<tr>
<td>FCSar</td>
<td>1.421052631578947</td>
</tr>
</tbody>
</table>

Tab. 22: Average Rankings of the algorithms

Friedman statistic considering reduction performance (distributed according to chi-square with 1 degrees of freedom: 0.4736842105261889.
P-value computed by Friedman Test: 0.4912971242167641.

Iman and Davenport statistic considering reduction performance (distributed according to F-distribution with 1 and 18 degrees of freedom: 0.46022727272714625.
P-value computed by Iman and Davenport Test: 0.5061465835202206.

<table>
<thead>
<tr>
<th>i</th>
<th>algorithm</th>
<th>$z = (R_o - R_i)/SE$</th>
<th>$p$</th>
<th>Holm/Hochberg/Hommel</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>FGBRMA</td>
<td>0.6882472016116845</td>
<td>0.4912971242167641</td>
<td>0.05</td>
</tr>
</tbody>
</table>

Tab. 23: Holm / Hochberg Table for $\alpha = 0.05$

Bonferroni-Dunn’s procedure rejects those hypotheses that have a p-value $\leq 0.05$.
Holm’s procedure rejects those hypotheses that have a p-value $\leq 0.05$.
Hommel’s procedure rejects those hypotheses that have a p-value $\leq 0.05$.
Bonferroni-Dunn’s procedure rejects those hypotheses that have a p-value $\leq 0.1$.
Holm’s procedure rejects those hypotheses that have a p-value $\leq 0.1$.
Hommel’s procedure rejects those hypotheses that have a p-value $\leq 0.05$. 
Applying the comparison methodology (results)

\[ z = \frac{(R_0 - R_i)}{SE} \]

<table>
<thead>
<tr>
<th>i</th>
<th>algorithm</th>
<th>( z )</th>
<th>( p )</th>
<th>Holm/Hochberg/Hommel</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>FGBRMA</td>
<td>0.6882472016116845</td>
<td>0.4912971242158936</td>
<td>0.1</td>
</tr>
</tbody>
</table>

Tab. 24: Holm / Hochberg Table for \( \alpha = 0.10 \)

<table>
<thead>
<tr>
<th>i</th>
<th>algorithm</th>
<th>unadjusted ( p )</th>
<th>( p_{Bonf} )</th>
<th>( p_{Holm} )</th>
<th>( p_{Hoch} )</th>
<th>( p_{Homm} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>FGBRMA</td>
<td>0.4912971242158936</td>
<td>0.49129</td>
<td>0.49129</td>
<td>0.49129</td>
<td>0.49129</td>
</tr>
</tbody>
</table>

Tab. 25: Adjusted \( p \)-values

<table>
<thead>
<tr>
<th>i</th>
<th>algorithms</th>
<th>( z )</th>
<th>( p )</th>
<th>Holm</th>
<th>Shaffer</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>FGBRMA vs. FCSar</td>
<td>0.6882472016116845</td>
<td>0.49129</td>
<td>0.1</td>
<td>0.1</td>
</tr>
</tbody>
</table>

Tab. 26: Holm / Shaffer Table for \( \alpha = 0.05 \)

Nemenyi's procedure rejects those hypotheses that have a \( p \)-value \( \leq 0.05 \).
Holm’s procedure rejects those hypotheses that have a \( p \)-value \( \leq 0.05 \).
Bergmann’s procedure rejects these hypotheses:

<table>
<thead>
<tr>
<th>i</th>
<th>algorithms</th>
<th>( z )</th>
<th>( p )</th>
<th>Holm</th>
<th>Shaffer</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>FGBRMA vs. FCSar</td>
<td>0.6882472016116845</td>
<td>0.49129</td>
<td>0.1</td>
<td>0.1</td>
</tr>
</tbody>
</table>

Tab. 27: Holm / Shaffer Table for \( \alpha = 0.10 \)

Nemenyi’s procedure rejects those hypotheses that have a \( p \)-value \( \leq 0.1 \).
Holm’s procedure rejects those hypotheses that have a \( p \)-value \( \leq 0.1 \).
Bergmann’s procedure rejects these hypotheses:
<table>
<thead>
<tr>
<th>i</th>
<th>hypothesis</th>
<th>unadjusted $p$</th>
<th>$PN_{ume}$</th>
<th>$PH_{olm}$</th>
<th>$PS_{haf}$</th>
<th>$PB_{erg}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>FGBRMA vs. FCSar</td>
<td>0.49129</td>
<td>0.49129</td>
<td>0.49129</td>
<td>0.49129</td>
<td>0.49129</td>
</tr>
</tbody>
</table>

Tab. 28: Adjusted $p$-values
4.4.5 Diversity FGBRMA-FCSar (Tables of Friedman, Bonferroni-Dunn, Holm, Hochberg and Hommel Tests)

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Ranking</th>
</tr>
</thead>
<tbody>
<tr>
<td>FGBRMA</td>
<td>1.9999999999999991</td>
</tr>
<tr>
<td>FCSar</td>
<td>0.9999999999999996</td>
</tr>
</tbody>
</table>

Tab. 29: Average Rankings of the algorithms

Friedman statistic considering reduction performance (distributed according to chi-square with 1 degrees of freedom: 18.99999999999983.

P-value computed by Friedman Test: 1.3071866922786946E-5.

Iman and Davenport statistic considering reduction performance (distributed according to F-distribution with 1 and 18 degrees of freedom: 2.005509209063406E15.

P-value computed by Iman and Davenport Test: 1.1102230246251565E-16.

<table>
<thead>
<tr>
<th>i</th>
<th>algorithm</th>
<th>( z = (R_0 - R_i)/SE )</th>
<th>( p )</th>
<th>Holm/Hochberg/Hommel</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>FGBRMA</td>
<td>4.358898943540671</td>
<td>1.30718 E-5</td>
<td>0.05</td>
</tr>
</tbody>
</table>

Tab. 30: Holm / Hochberg Table for \( \alpha = 0.05 \)

Bonferroni-Dunn’s procedure rejects those hypotheses that have a p-value \( \leq 0.05 \).
Hochberg’s procedure rejects those hypotheses that have a p-value \( \leq 0.05 \).
Hommel’s procedure rejects all hypotheses.
Bonferroni-Dunn’s procedure rejects those hypotheses that have a p-value \( \leq 0.1 \).
Hochberg’s procedure rejects those hypotheses that have a p-value \( \leq 0.1 \).
Hommel’s procedure rejects all hypotheses.
Applying the comparison methodology (results)

\[ z = \frac{(R_0 - R_i)}{SE} \]

Tab. 31: Holm / Hochberg Table for \( \alpha = 0.10 \)

<table>
<thead>
<tr>
<th>algorithm</th>
<th>( p ) Holm/Hochberg/Hommel</th>
</tr>
</thead>
<tbody>
<tr>
<td>FGBRMA</td>
<td>4.358898943540671 1.30718 E-5 0.1</td>
</tr>
</tbody>
</table>

Tab. 32: Adjusted \( p \)-values

<table>
<thead>
<tr>
<th>algorithms</th>
<th>( z = \frac{(R_0 - R_i)}{SE} )</th>
<th>( p ) Holm Shaffer</th>
</tr>
</thead>
<tbody>
<tr>
<td>FGBRMA vs. FCSar</td>
<td>4.358898943540671 1.30718 E-5 0.05 0.05</td>
<td></td>
</tr>
</tbody>
</table>

Tab. 33: Holm / Shaffer Table for \( \alpha = 0.05 \)

Nemenyi’s procedure rejects those hypotheses that have a \( p \)-value \( \leq 0.05 \).
Shaffer’s procedure rejects those hypotheses that have a \( p \)-value \( \leq 0.05 \).
Bergmann’s procedure rejects these hypotheses:

• FGBRMA vs. FCSar

<table>
<thead>
<tr>
<th>algorithms</th>
<th>( z = \frac{(R_0 - R_i)}{SE} )</th>
<th>( p ) Holm Shaffer</th>
</tr>
</thead>
<tbody>
<tr>
<td>FGBRMA vs. FCSar</td>
<td>4.358898943540671 1.30718 E-5 0.1 0.1</td>
<td></td>
</tr>
</tbody>
</table>

Tab. 34: Holm / Shaffer Table for \( \alpha = 0.10 \)

Nemenyi’s procedure rejects those hypotheses that have a \( p \)-value \( \leq 0.1 \).
Shaffer’s procedure rejects those hypotheses that have a \( p \)-value \( \leq 0.1 \).
Bergmann’s procedure rejects these hypotheses:
- FGBRMA vs. FCSar

<table>
<thead>
<tr>
<th></th>
<th>hypothesis</th>
<th>unadjusted $p$</th>
<th>$p_{Nemeny}$</th>
<th>$p_{Holm}$</th>
<th>$p_{Shaffer}$</th>
<th>$p_{Berg}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>FGBRMA vs. FCSar</td>
<td>1.30718 E-5</td>
<td>1.30718 E-5</td>
<td>1.30718 E-5</td>
<td>1.30718 E-5</td>
<td>1.30718 E-5</td>
</tr>
</tbody>
</table>

**Tab. 35:** Adjusted $p$-values
5 Conclusions

We investigated the problem of comparing Fuzzy Association Rules algorithms from a point of view of final user, rather than from a theoretical point of view. It included how to select one, within the everyday more and more extensive list of algorithms. Were proposed three steps to follow to ensure equivalence in the subsequent comparison and again four steps to make post experiments comparison.

In an experimental analysis, three algorithms were compared, and applying the proposed methodology could be determined, the advantages and disadvantages of each in terms of scalability, quality, and diversity. Main conclusion is this methodology can be applied and / or adapt to any pair of algorithms, pursuing the possibility of deciding between the two, the one that best suits the final user needs.

Thanks to proposed methodology, was demonstrated that Fuzzy C means labeling tends to produce more large-itemsets and found better quality (and probably diversity) of rules. Also, making a comparison between algorithms with the same triangular form of the label, it was found that an evolutionary algorithm is more advantageous to maintain the diversity of the rules even if this is not a explicit goal. And maintenance same level of quality even in its most basic parametrization. All results are supported by statistical tests.

6 Future work

Concerning this study, there are still some promising directions for the future work. A possibility is to use different measures to assess the quality of an algorithm. For example, an algorithm that leads to another using the Confidence, also excel using measures such as Correlation or Certainty?. In the same line, other diversity measures may be applied. In example, some algorithms are able to find rules with disjunction structure, but the measure of diversity used here can not compare this type of rules. Finally, other important focus of attention would be the comparison of a greater variety of algorithms to extract rules. For example, an multi-objective algorithm which uses the Diversity as an objective, truly get better results on issues of diversity?

Acknowledgments

This research was possible by the Ministry of Science and Technology of Costa Rica (MICIT) and the National Commission for Scientific and Technological Research (CONICIT).

References

6 Future work


