Inference processes in rule cluster knowledge base - various approaches

Igor Gaibei *igor.gaibei@us.edu.pl*
Agnieszka Nowak-Brzezińska *agnieszka.nowak-brzezinska@us.edu.pl*

University of Silesia
Faculty of Science and Technology
Institute of Computer Science
Katowice, Poland

Abstract

This paper presents a novel approach to optimize inference in rule-based knowledge systems by introducing a clustering mechanism for rule organization. Rules are clustered using K-Means or Agglomerative Hierarchical Clustering (AHC) algorithms, with different distance measures and clustering strategies. We propose and evaluate four inference strategies based on different group representation methods (mean or median) and rule activation strategies (activation of one or all matching rules). Experimental studies on real knowledge bases show that clustering significantly improves inference performance while maintaining a satisfied inference success rate.

Keywords: inference algorithm, rule-based knowledge bases, clustering algorithms

1. Introduction

The increasing complexity and volume of knowledge bases, especially those in rule form, poses a serious challenge to efficient and effective inference processes. Classical expert systems rely on sequentially searching the entire rule set for those rules that apply to the current set of facts. As the size of the knowledge base increases, this method becomes inefficient and computationally expensive. Furthermore, traditional inference mechanisms often do not take into account structural relationships between rules that could be used to optimize the inference process.

In this paper, we propose an improved inference approach based on rule clustering and rule cluster representatives. By clustering similar rules and creating appropriate representatives for these groups, we aim to significantly reduce the rule search area during inference, improve inference performance and maintain or even increase inference success cases.

The main contribution of this paper is to develop and evaluate four modified inference strategies based on different cluster representation techniques (mean or median) and different rule activation mechanisms (activation of a single rule or all relevant rules). We investigate how these strategies affect inference time, inference completion successfully based on different real-world knowledge bases with different data structures.

1.1. State of the Arts

Reynolds et al. (2006) present a detailed comparative study of clustering algorithms applied to rule sets, analyzing partitioning methods and hierarchical clustering techniques [5]. Their research discusses how different algorithms and similarity measures affect the formation of rule clusters. The study also emphasises the importance of choosing appropriate distance measures and association criteria tailored to the specific characteristics of the rule data. The authors systematically evaluate how rule clustering integrates with numerical association rule mining. They reveal that the quality of clusters depends heavily on the underlying representation of rules and pre-processing steps [3].

2. Inference processes in rule cluster knowledge base - description of the idea

In rule-based knowledge bases, rule representation structures knowledge as "if-then" statements with premises and conclusion. Inference performance decreases as rule sets increase, due to increased search complexity.

The idea behind our approach is as follows. After loading any knowledge base with a rule representation, we cluster the rules using K-Means or Agglomerative Hierarchical Clustering (AHC) algorithms. We use different distance measures, different methods of combining clusters, different number of clusters. Before clustering, we decide whether rules will be clustered by premise, conclusion or both. We evaluate the structure of the rule clusters created in this way using the Calinski-Harabasz index and the Silhouette index to determine whether the clustering has been done correctly or not. We can optimise the inference process by shortening the inference time without reducing its quality, which we understand as the effective finding of a rule or rules to be activated if they exist in the knowledge base. If there is more than one rule that can be activated, we decide whether to activate only one, selected in accordance with the inference control strategy used (this topic is beyond the scope of this work), or all relevant rules. Of course, activating all of them will increase the inference time, but it will allow us to generate all possible knowledge. When we activate only one of the rules, the inference time will be as short as possible, but we will not derive all possible knowledge from the system (this means that the inference process was not complete). Inference may fail if, during the review of rule clusters, we cannot find a relevant rule even though one exists in the knowledge base.

What do we gain with our approach? Firstly, we review only the most relevant rule cluster in the inference process. We further compare the facts with the rules within that rule cluster. If we have clustered the rules well, it should be the case that this selected group contains all the theoretically sought rules.

3. Inference process

Inference is the process of deriving new information or statements (conclusions) from a set of premises, according to certain rules of logic. In this work, we have dealt with the forward inference method (from premises to conclusions). For a given set of facts, we look for rules whose premises cover a given set of facts. We activate these rules, by which we derive new knowledge (the conclusions of these activated rules become new facts, new knowledge). The forward inference process runs iteratively until the goal is reached or the rules are exhausted.

3.1. Description of inference on the rule clusters

We developed four modified approaches for inference on rule clusters: *meanOneRule*, *medianOneRule*, *meanAllRules* and *medianAllRules*. They differ in the way the selected rules are activated and in the method of forming a representative. The approaches used assume activation of all found rules or only the first one in the list. We assume, for current research, that the knowledge base contains rules described by attributes on a numerical scale. This gives us the ability to use the mean or median to represent a group of attribute values.

4. Rule clustering

Clustering is the process of partitioning a set of rules into the groups (clusters) such that objects within the same cluster are more similar to each other than to those in other clusters. In our approach, we selected two well-known clustering algorithms: K-Means and AHC [4].

To assess the quality of clustering, we use two internal validation indicators: Calinski–Harabasz Index (CH) and Silhouette Score (Sil). In the experiments, we used the term correct/incorrect to describe the correctness of the clustering. Knowing what values of both indices indicate optimal

clusterings, we set rules for evaluating whether a clustering is correct or incorrect.

5. Experiments

This section explains how rule clustering was performed, the approach to generating the initial fact sets, the inference strategies and the methods used to evaluate the effectiveness and efficiency of the inference process.

5.1. Knowledge bases

In the experiments, we included knowledge bases containing different structures [1]. The parameters of these sources are presented in Table 1. The source datasets were loaded into the RSES tool, where decision rules were generated using the LEM 2 algorithm [2].

Table 1. Knowledge bases and rules description

KB	Items	Attr	Data source description	#Rules	Range of Val	Av. Attr.	Min-Max Attr.
db1	4435	37	Statlog landsat satellite	937	29.0 - 128.0	4.3	1-13
db2	7027	65	Polish companies bankruptcy	4125	0 - 3532.8	3.0	1-27
db3	527	38	Water treatment plant	123	0 - 53012	3.0	1-6
db4	17898	9	Pulsar candidates	6432	0.02 - 190.42	1.08	1-2

Explanations of abbreviations:
KB – Knowledge Base; Items – Number of examples/instances in the data set; Attr – Number of input features describing each item; #Rules – Number of generated decision rules; Range of Val – Minimum and maximum observed attribute values; Av. Attr. - Average number of attributes (conditions) in a single rule; Min-Max Attr. – Minimum and maximum number of attributes in a single rule.

5.2. Methodology of experiments

We perform clustering and inference sequentially for each rule set, using the algorithms K-Means and AHC for $K = 2, 3, \ldots, 22$ and optimal K for each rule set. Optimal K is defined as $K \sim \sqrt{N}$, where: K is the estimated optimal number of clusters and N is the number of rules in a given knowledge base - according to the literature.

For each algorithm, we perform clustering using Euclidean, Chebyshev, and Manhattan distances. For the AHC, we repeated the algorithm for single, complete, and average linkages. We repeat each algorithm for three different inputs: the conditions (W) alone, the conclusions (D) alone, and the conditions and conclusions (W+D) together. For each algorithm, we generate the initial facts constituting 5%, 25%, and 50% of all unique descriptors in the premises. We used four approaches for creating representative - meanOneRule, medianOneRule, meanAllRules, medianllRules. For 4 knowledge bases, this gives a total of 37584 for experiments.

5.3. Facts generation for inference experiments

In a typical real-world expert system scenario, it is the user who provides initial information, known as facts, to the system. These facts represent specific observations or known values about the problem domain. Based on this set of input facts, the inference engine searches for applicable rules whose premises match the provided information and activate these rules to derive conclusions or new knowledge.

To replicate this process systematically in experimental conditions we developed a method to simulate fact generation automatically. This allows us to test the inference performance under various conditions and predict system behavior in diverse real-world scenarios. We generate fact sets that cover approximately 5%, 25% and 50% of all possible facts from the dictionary. This variability allows us to simulate different levels of user input completeness.

5.4. Results of the experiments

The Table 2 shows that successful inference is most often accompanied by correct clustering - i.e. obtaining clusters that are internally consistent and externally well separated.

Table 2. Cluster quality

	Clusters quality - Sil		Clusters quality - CH	
Inference state	correct	incorrect	correct	incorrect
false	12340 (80.93%)	2908 (19.07%)	5507 (36.12%)	9741 (63.88%)
true	19095 (85.49%)	3241 (14.51%)	13745 (61.54%)	8591 (38.46%)

We see in the Table 3 that inference was most often successful for db2 and db4. Dataset db3 looks very problematic. But one only has to look at the description of this database to guess that its structure is complex. The database contains a small number of rules (123) and an extreme range of values (0 - 53k). This range of attribute values may cause some distance measures (e.g. Euclidean distance) to be dominated by single attributes. We also observe a high number of attributes relative to the number of rules (38 attributes per 123 rules).

Table 3. Percentage of successful inference depending on the ruleset

KB	**Inference state**		**All**
	false	**true**	
db1	3974 (41.81%)	5530 (58.19%)	9504
db2	3393 (35.70%)	7611 (64.30%)	9504
db3	4502 (49.63%)	4570 (50.37%)	9072
db4	3379 (35.55%)	6125 (64.45%)	9504
All	15248	22336	37584

Table 4. Inference time (average) [s]

Inference method	**db1**	**db2**	**db3**	**db4**	**All**	**Inference state = true**
meanOneRule	0.012103	0.07346	0.002336	0.07293	0.04064	59.23%
medianOneRule	0.012453	0.07579	0.001652	0.03996	0.03282	59.41%
meanAllRules	0.192027	28.31245	0.002266	44.67265	18.50511	59.48%
medianAllRules	0.211920	68.74963	0.002618	41.18935	27.85488	59.60%

The Table 4 shows that the inference methods differ only slightly in terms of inference success. They also differ slightly in inference time. The Table 5 shows that if we cluster rules by conditions, inference is successful in 70% of cases regardless of the algorithm, but this is usually slightly more frequent for the K-Means algorithm than for AHC. On the other hand, if we cluster rules by decisions, on average, the frequency of successful inference is low, but when we look at which clustering algorithm we use. If it is K-Means, there is a much higher

chance of successful inference (nearly 90% of cases), while in the case of AHC, this efficiency of inference is unfortunately low at the level of several percent. Interestingly, when we cluster rules by both premises and conclusions, inference is successful almost as often, regardless of whether we cluster using the K-Means or AHC algorithm (in 3 out of 4 cases).

Table 5. Percentage of successful inference depending on clustering object and method

Clustering Object	True All	True K-Means	True AHC
W	8792 (70.18%)	2216 (70.75%)	6576 (69.99%)
D	4317 (34.46%)	2769 (88.41%)	1548 (16.48%)
W + D	9227 (73.65%)	2333 (74.49%)	6894 (73.37%)

6. Summary

This paper presents four different approaches of inference algorithm for knowledge bases with a rule cluster structure. The research includes the study of the inference efficiency, which is measured by the number of cases in which the inference was successful. Across all datasets, inference based on K-Means clustering outperformed AHC, achieving nearly 90% success compared to significantly lower rates with AHC. When clustering was based on conditions or both conditions and decisions, inference success remained generally high, with the latter approach yielding consistent results across all algorithms. A key finding is the strong influence of knowledge base structure on inference success. Knowledge bases with numerous rules and balanced attribute distributions (e.g., db2 and db4) achieved the highest success rates. In contrast, db3, characterized by a small number of rules, high dimensionality (38 attributes), and a wide value range (0–53k), demonstrated poor performance. Additionally, different inference strategies (e.g., meanOneRule vs. medianOneRule) exhibited only minor differences in success rate and execution time. Overall, K-Means - especially when combined with decision-based grouping - proved to be the most effective in terms of inference performance.

References

[1] HTRU2, Water Treatment Plant, Polish Companies Bankruptcy, Statlog (Landsat Satellite). UCI Machine Learning Repository, https://archive.ics.uci.edu/, Accessed April 20, 2025

[2] RSES Tool, URL: https://www.roughsets.org/roughsets/software/, Accessed April 20, 2025

[3] Kaushik, M., e.a.: A systematic assessment of numerical association rule mining methods. SN Computer Science 2(5), pp. 348 (2021), https://doi.org/10.1007/s42979-021-00725-2

[4] Nowak-Brzezińska, A., Gaibei, I.: Inference algorithm for knowledge bases with rule cluster structure. In: Computational Science – ICCS 2024. pp. 71–78. Springer Nature Switzerland, Cham (2024), https://doi.org/10.1007/978-3-031-63759-9_9

[5] Reynolds, A., e.a.: Clustering rules: A comparison of partitioning and hierarchical clustering algorithms. Journal of Mathematical Modelling and Algorithms 5(4), pp. 475–504 (2006), https://doi.org/10.1007/s10852-005-9022-1