Outlier Mining in Rule-Based Knowledge Bases

Agnieszka Nowak - Brzezińska
Institute of Computer Science
Silesian University
Bankowa 12, 40-007 Katowice, Poland
Email: agnieszka.nowak@us.edu.pl

***Abstract*—This paper introduces an approach to outlier mining in the context of rule-based knowledge bases. Rules in knowledge bases are a very specific type of data representation and it is necessary to analyze them carefully, especially when they differ from each other. The goal of the paper is to analyze the influence of using different similarity measures and clustering methods on the number of outliers discovered during the mining process. The results of the experiments are presented in Section V in order to discuss the significance of the analyzed parameters.**

***Index Terms*—outlier detection, similarity analysis, clustering, knowledge-based systems.**

I. Introduction

Outlier detection is a fundamental issue in data mining, it has been specifically used to detect and remove anomalous objects from data. Data mining, in general, deals with the discovery of nontrivial, hidden and interesting knowledge from different types of data. With the development of information technologies, the number of databases and their dimensions and complexity, grow rapidly. One of the basic problems of data mining is outlier detection. The identification of an outlier is affected by various factors, many of which have become the subject of practical applications such as public health or credit card transactions. In the first case (public health), outlier detection techniques help to detect anomalous patterns in patient medical data which could be symptoms of an ailment. Generally, outliers are the points which are different from or inconsistent with the rest of the data. It can be novel, new, abnormal, unusual or noisy information thus it is often more interesting than the majority of the actual data.

The main goal of the article is to present recent approaches to outlier mining in rule-based knowledge bases (KBs). Such data sources are very popular in the area of decision support systems (DSS). It is very important to have the possibility to explore the KBs, especially because they often consist of many correlations, dependancies, or even unusual cases (rules). A few years ago, outlier detection was just one of several steps of data preparation procedure. Frequently, the data that have been denoted as outliers would have been removed from the dataset and treated like errors. Nowadays outliers are no longer (or at least not entirely) seen as errors. When they are discovered in a given dataset, they might become a foundation for a deeper exploration as far as they are able to contain some important (yet to be discovered) knowledge. Unfortunately, if such type of data is not removed from the dataset, it has a negative influence on other (further) processes of data analysis. Outliers are capable of decreasing the quality of knowledge mined from a given dataset. Especially, if we rely on the knowledge mined from the clusters, inducted from the data with outliers inside. Of course it also depends on the clustering algorithm used to group the original data. However, there are some algorithms (e.g. k-means) which are suspectible to the outliers. The author works with hierarchical algorithms which are, fortunately, resistant to the outliers. Simply, when there are some unusual data, dissimilar to all the other in the dataset, they are clustered at the end of the clustering procedure. Thus it is enough to set a specified stop criterion (i.e. the moment in which the similarity between the merged clusters is smaller than a given threshold value) which will break the clustering at an exact time. Then, all the nodes (clusters) in a created hierarchical structure are being treated as outliers. The article presents the analysis of the influence of different clustering parameters on the results of the final clusters' structure and their ability to mine the outliers in KBs.

978-1-5090-5795-5/17/$31.00 © 2017 IEEE

A. Outliers in a Knowledge Base

There are numerous papers on mining outliers in data but there seem to be no publications which cover the issue of finding outliers in a specific kind of data such as rules usually stored in KBs. When we think about outliers in rules we should think about rules which represent some very crucial, but also rare, part of domain knowledge. As a matter of fact, a rule can be defined as a formula with two parts: conditional (premises) and decisional (conclusion). In this context an outlier can be a rule which is dissimilar to all other rules in a given KB, because the set of premises and/or conclusion is dissimilar to the rest of the data (rules). Additionally, outliers can be described as all the rules which contain a much smaller or greater number of premisses when compared with the others. We may also say that an outlier-based rule is a rule, which consists of unusual attributes and values of such attributes in the conditional part of it as well as in the decisional one. It means that the both types of rules: non-deterministic and the ones with different premises can be treated as outliers. An outlier is not only a single object as, in this case, a single rule in a KB. When we group rules together, based on the similarity criteria (as it is done when using data mining algorithms like cluster analysis), as a result we get as a small group (quite often singleton) while other (more similar) rules create bigger groups. Therefore, during the analysis it is possible that such small group will be not taken into account and it is not a desirable solution. If we cluster rules in order to optimize the exploration process, outlier-type rules, in a given KB, may take the form of small clusters or a single rule which has not been merged with the others.

B. The Structure of the Article

The rest of the paper is organized as follows. In Section II the definition and the assumptions about discovering outliers in KBs are presented. Section III includes the pseudocode of the rules clustering algorithm algorithm as well as the description of various similarity and clustering methods which then are examined and presented with the results in Section V. Section IV introduces the outlier detection method for a hierarchical structure of rules in KBs.

II. Outliers: the Definition, the Meaning and the Types of Outliers

The issue of outlier detection has numerous important uses in many applications which are high-dimensional domains and the data therein may consist of hundreds of dimensions. In paper [4] the authors present a new approach to the summarization of databases containing both numerical and partly standardized textual records. The described method enables the detection of outlier information using linguistic summaries. It is a primary step in many data mining applications. Although outliers are often considered as an error or noise, they may carry important information. Hawkins defines an outlier as an observation that deviates so much from other observations as to arouse suspicion that it has been generated by a different mechanism [7]. Other researchers indicate that an outlying observation, is one that appears to deviate markedly from other members of the sample in which it occurs or an observation in a dataset which appears to be inconsistent with the remainder of that set of data. In case of complex data using statistical methods, based on regression analysis is impossible. That is why we need to find methods which deal with the complexity of the data being analyzed, in order to identify all possible outliers.

A. The Meaning of Outliers

Outliers may contain important information, therefore they should be investigated carefully. It is well known that quite often they contain valuable information about the process being investigated or the data gathering and recording process. Before considering possible elimination of these points from the data, one should try to understand why they have appeared in the first place and whether it is likely that similar values will continue to appear. Of course, outliers are often bad data points and such as they should be removed from the entire dataset. Outliers can actually represent unexpected factors of practical importance and can therefore contain valuable information. In these situations, the influence of outliers should be emphasized rather than limited or minimized. Any attempts to reduce the influence of outliers without due consideration in these applications can lead to a loss of information which may often be crucial for problem solving.

B. Types of Outlier

Outliers can be presented with *scores* as well as with a *label*. Outliers with scores give us the information about a degree of outlierness of each data. Labelling the outliers means that we include the information whether the data is anomalous or not. Thus, scores are more informative to analysts and they can be readily converted into labels by choosing a particular threshold.

C. Data Mining Approach in the Outlier Detection Process

Most of the approaches to anomaly detection in data mining utilize the distance and similarity to the nearest neighbors and label observations as outliers or non-outliers [7]. Clustering can be a great example of such operations as it works by grouping similar objects into clusters and assumes that anomalies either do not belong to any cluster or they are distant from their cluster center or they belong to small and sparse clusters (generally unsupervised). Outlier mining could be defined as the process of grouping sets of records that behave in a different or deviant manner in comparison to the rest or majority of the data. It could also be viewed as the process of clustering, but with the difference that here clusters look out for objects or records that show a different behaviour when compared to the rest of the data. As outlier mining seems to be so easy to manage using the cluster analysis method it is necessary to make some notes about this kind of data mining techniques in this paper. Thus, in the next section, the general idea, the pseudocode and the most important aspects of clustering algorithms are given. Then, the pseudocode of the outlier mining algorithm is given in Section IV.

III. Clustering Algorithms

A cluster is a collection of objects which are similar to one another and dissimilar to the objects belonging to other clusters. Moreover, a clustering algorithm aims to find a natural structure or relationship in an unlabeled data set. There are several categories of clustering algorithms. Some of the algorithms are hierarchical and probabilistic. In this paper the author presents a hierarchical clustering algorithm which is based on the connection between the two nearest clusters. The starting condition is carried out by setting every object as a separate cluster. In each step, the two most similar objects are merged, and a new cluster is created with a proper representative for it. After a specified stop condition is reached the clustering process for the rules (or their groups) is finished. There are many possible ways for defining the stop condition. For instance it can be reaching the specified number of groups, or reaching the moment in which the highest similarity is under minimal required threshold (which means the groups of rules are now more differential than similar one another).

The pseudocode of the hierarchical clustering algorithm - namely Classic AHC (agglomerative hierarchical clustering) algorithm [3] - is presented as Pseudocode 1.

Pseudocode 1. Classic AHC Algorithm.
Input: stop condition *sc*, ungrouped set of objects *s*
Output: grouped tree-like structure of objects

```
1) Place each object o from s into a separate cluster.
2) Build a similarity matrix M that consists of every clusters
   pair similarity value.
3) Using M find the most similar pair of clusters and merge
   them into one.
4) Update M.
5) IF sc was met end the procedure.
6) ELSE REPEAT from step 3.
7) RETURN the resultant structure.
```

The most important step is the second one, in which the similarity matrix M is created based on the selected similarity measure and a pair of the two most similar rules (or groups of rules) are merged. In this (one) step two parameters are given by a user: the similarity measure and the clustering method. Eventually, both of them result in achieving different clusterings. For this reason the author decided to compare similarity measures in this research. In order to do that, the author choose three different similarity measures and repeated the clustering algorithm many times for every similarity measure while changing the number of groups [1] as well as the clustering method.

The main advantage of hierarchical clustering is that it does not impose any special methods of describing the clusters similarity.

A. Rules Clustering Algorithms

Many papers show the results of clustering a large set of data but rarely for such a specific type of data like rule-based knowledge representation. Clustering algorithms allow to organize the rules in a smart way [3]. To achieve groups of similar rules it is necessary to propose some method of deciding which rules are the most similar in a given step of the clustering process. Because rules are a specific type of data, usually attributed with short descriptions, the differences between rules are difficult to be noticed. Hence it is so important to find a similarity measure which is able to find all the differences and, as a result, decides about the order of rules clustering in an optimal way.

B. Similarity Analysis

In the literature there are numerous methods of describing similarity between objects [2] that can be modified to work with rules as well. The similarity measure used to find a pair of rules or groups of rules that are the most similar in a given moment is called the *intra-cluster similarity measure*. The authors studied the following five measures: Simple Matching Coefficient (SMC), based on it - the Jaccard Index sometimes also called the weighted similarity or the weighted similarity coefficient [11] and Gower measure (widely known in the literature) [6]. It is crucial to answer the question if a given similarity measure influences the shape of grouped KB's structure. Measuring similarity or distance between two data points is a core requirement for several data mining and knowledge discovery tasks that involve distance computation. The notion of similarity or distance for categorical data is not as straightforward as for continuous data. When data consists of objects that aggregate both types at once the problem is much more complicated.

For a set of attributes A and their values V, rules premises and conclusions are built using pairs (a_i, v_i). In this approach $a_i \in A, v_i \in V_a$ and a pair (a_i, v_i) is called a descriptor. In a vector of such pairs, i-th position denotes the value of the i-th attribute of a rule. Most of the rules do not consist of all attributes in A, thus constructed vectors (describing the rules) are of different lengths.

Almost all similarity measures assign a similarity value between two rules r_j and r_k belonging to the set of rules R as follows:

$$S(r_j, r_k) = \sum_{i=1}^{N} w_i s(r_{ji}, r_{ki})$$

where $s_i(r_{ji}, r_{ki})$ is the per-attribute (for i-th attribute) similarity between two values of descriptors of the rules r_j and r_k. The quantity w_i denotes the weight assigned to the attribute a_i and usually $w_i = \frac{1}{d}$, for $i = 1, \ldots, d$. In all the definitions, the s_{ijk} denotes the contribution provided by the k-th variable, and w_{ijk} is usually 1 or 0 depending upon whether or not the comparison is valid for the k-th variable; if differential variable weights are specified it is the weight of the k-th variable or 0 if the comparison is not valid.

The simplest measure is SMC (Simple Matching Coefficient) [2] - which calculates the number of attributes that match in the two rules in the following way:

$$s_{SMC}(r_{ji}, r_{ki}) = s_{jki} = 1 if r_{ji} = r_{ki} \texttt{ else } 0.$$

The range of the per-attribute SMC is $\{0; 1\}$. It treats all types of attributes in the same way. Unfortunately it tends to favour longer rules thus it is better to use the $Jaccard$ measure, which is similar to SMC however it is more advanced as it also divides the result by the number of attributes of both objects so longer rules are not favoured any more. It can be defined in the following form:

$$s_{Jaccard}(r_{ji}, r_{ki}) = s_{jki} = \frac{1}{n} \texttt{ if } r_{ji} = r_{ki} \texttt{ else } 0$$

where n is the number of attributes considered. The $Gower$ similarity coefficient is the most complex of the all used inter-cluster similarity measures as it handles numeric attributes and symbolic attributes differently. For ordinal and continuous

[1] In this work clustering is stopped when given number of clusters is generated.

[2] If both compared objects have the same attribute and this attribute has the same value for both objects then add 1 to a given similarity measure. If otherwise, do nothing. To eliminate one of the problems of SMC, which favours the longest rules, the author has also used the Jaccard Index.

variables it defines the value of s_{jki} as $s_{jki} = 1 - \frac{|r_{ji} - r_{ki}|}{range(i)}$, where: $range(i)$ is the range of values for the i-th variable. For continuous variables s_{jki} ranges between 1, for identical values r_{ji} = r_{ki} and 0 for the two extreme values r_{max} - r_{min}.

C. Clustering Methods

The distance among clusters can be computed using different methods, between which the following four methods are the most popular: *Single Linkage* (SL), *Complete Linkage* (CoL), *Average Linkage* (AL) and *Centroid-based Linkage* (CL). SL is a method that focuses on the minimum distances or the nearest neighbor between clusters meanwhile CoL concentrates on the maximum distance or the furthest neighbor between clusters. AL is a compromise between the sensitivity of CoL to outliers and the tendency of SL to form long chains that do not correspond to the intuitive notion of clusters as compact, spherical objects. In the CL method, the centroid is the mean of all points in a cluster.

D. Cluster Validity

Cluster validity seems to be an important feature in checking whether the created structure has got a good quality. If not, it may treat some data as outlier even it is not one. There are two criteria which are necessary to meet when good clustering results have to be achieved: separation and cohesion. There are many different measures to check if both of this condition are met. One of the most popular is the MDI index, which has been examined in this research.

IV. Outlier Mining Using Clustering Algorithms

An outlier or a noise point is an observation which appears to be inconsistens with the remainder of the data. Outliers may be considered as noise points lying outside a set of defined clusters. Generally, existing techniques work well in the absence of noise. When there is noise in the dataset, the clusters identified by these techniques include the surrounding noise points too. The presence of noise disrupts the process of clustering. Noise has to be separeted from the dataset to enhance the quality of the clustering results.

A. Outlier Detection Algorithm

The pseudocode of the outlier detection algorithm based on clustering approach is presented as Pseudocode 2.

```
Pseudocode 2. Outlier detection algorithm.
Input: $M$ - Number of clusters, $G = \{g_1, g_2, \ldots, g_M\}$.
Output: $UngroupedCount$ - a list of ungrouped rules.
  1) Group data objects $X$ using the $AHC$ algorithm in order
     to achieve $M$ clusters grouped in tree-like structure,
     $UngroupedCount = 0$.
  2) For each cluster $g_i$ from the $G$ set
  3) IF $sizeOf(g_i) == 1$ THEN $++UngroupedCount$
  4) RETURN $UngroupedCount$.
```

The goal of the algorithm is to discover ungrouped rules - individual rules which could not be joined with the others, because they do not have any common feature (neither premises nor conclusion). Only such rules (individual objects) represent something interesting to be found in the KB as it means that there are unique (so-far unexamined) areas of domain knowledge which are totally different from the rest of knowledge already stored in our KB. The most important step is examining each of the created cluster g_i in order to check if its size is equal to 1 as it would mean that it is an outlier - an ungrouped rule. The results of this research are given in Section V.

V. Experiments

The goal of the experiments has been to check whether choosing of similarity and/or clustering methods influences the possibility of finding outlier-type rules in KBs. Thus, in this section, experimental evaluation of 3 similarity measures and 4 clustering methods (described in Section III) on 7 different KBs [9] is presented. Decision rules were generated from the original data using RSES software and $LEM2$ algorithm [1]. The smallest number of attributes was 5, the greatest 280. The smallest number of rules was 42, the greatest 490. The first experiment is based on comparison of three similarity measures: SMC, $Jaccard$ and $Gower$ and checks if it influences the possibility of finding outliers in rules (the results are included in Table II). The aim of thr second experiment is to compare four clustering methods: SL, CL, CoL and AL in order to find some correlation between a given clustering method and any of the parameters important to describe the structure of clusters of rules and the number of outliers. The results of the second experiment are presented in Table III.

All the details of the analyzed datasets are included in table I.

The meaning of the columns in tables I, II and III is as follows:

- $AttrN$ - number of different attributes occuring in permises or conclusions of rules in a given knowledge base.
- $RulesN$ - number of rules in an examined knowledge base.
- $ClustersN$ - number of nodes in a dendrogram representing the resultant structure.
- U - number of singular clusters in the resultant structure of grouping.
- BRS - Biggest representative size - number of descriptors used to describe the longest representative.
- ARS - Average representative's size - average number of descriptors used to describe a cluster's representatives.
- $wARS$ - Weighted Average representative's size (AttrNumber) - division of average number of descriptors used to describe a cluster's representative in a given data set and the number of attributes in this data set.
- BRL - Biggest representative length - number of descriptors in the biggest cluster's representative.
- BCS - Biggest cluster's size - number of rules to have been used in the cluster.

Table II confirms the lack of any significant differences between using the analyzed similarity measures in the context

TABLE I
DATA GATHERED DURING EXPERIMENTS.

	Total	arythmia	audiology	autos	balance	Breast cancer	diab	diabetes
ClustersN	$16,5 \pm 14,1$ 4-49	$12,5 \pm 2,5$ 10-15	$6,9 \pm 3,0$ 4-10	$7,8 \pm 2,4$ 4-10	$19 \pm 9,1$ 10-28	$11 \pm 1,0$ 10-12	29 ± 19 10-48	$29,5 \pm 19,7$ 10-49
AttrN	$58,4 \pm 93$ 5-280	$280,0 \pm 0,0$ 280-280	$70,0 \pm 0,0$ 70-70	$26,0 \pm 0,0$ 26-26	$5,0 \pm 0,01$ 5-5	$10 \pm 0,0E-01$ 10-10	$9,0 \pm 0,1$ 9-9	$9,0 \pm 0,0$ 9-9
RulesN	234 ± 176 42-490	$154,0 \pm 0,0$ 154-154	$42,0 \pm 0,0$ 42-42	$60,0 \pm 0,0$ 60-60	$290 \pm 0,1$ 290-290	$130 \pm 0,1$ 130-130	$480 \pm 0,1$ 480-480	$490,0 \pm 0,0$ 490-490
NodesN	452 ± 343 74-970	$295,5 \pm 2,5$ 293-298	$77,2 \pm 3,0$ 74-80	$112,2 \pm 2,4$ 110-116	$560 \pm 9,1$ 550-560	$240 \pm 1,0$ 240-240	940 ± 19 920-960	$950,5 \pm 19,7$ 931-970
BCS	155 ± 144 $12 - 480$	$111,5 \pm 42,0$ $23 - 145$	$30,0 \pm 7,9$ $12 - 39$	$37,9 \pm 14,3$ $12 - 57$	170 ± 95 $21 - 280$	76 ± 32 $21 - 120$	320 ± 130 $37 - 470$	$336,6 \pm 135,2$ $45 - 479$
BRL	$35,4 \pm 50,4$ $3 - 150$	$147,4 \pm 1,9$ $145 - 154$	$66,8 \pm 0,5$ $66 - 68$	$10,7 \pm 0,6$ $10 - 12$	$4,0 \pm 0,01$ $4 - 4$	$9,0 \pm 0,01$ $9 - 9$	$4,9 \pm 0,46$ $3 - 5$	$4,9 \pm 0,3$ $4 - 5$
U	$6,9 \pm 9,0$ $0 - 41$	$5,8 \pm 4,7$ 0-14	$3,6 \pm 2,7$ 0-9	$3,8 \pm 3,0$ 0-9	$6,9 \pm 9,0$ 0,01-27	$5,1 \pm 3,4$ 0,01-10	11 ± 13 0,1-36	$12,5 \pm 14,4$ 0-41
BRS	$36,4 \pm 51,81$ 4-170	$152,1 \pm 4,9$ 147-165	$67,0 \pm 0,5$ 66-68	$11,6 \pm 1,8$ 10-18	$4,0 \pm 0,01$ 4-4	$9,0 \pm 0,01$ 9-9	$5,4 \pm 0,49$ 5-6	$5,5 \pm 0,6$ 5-7
ARS	$29,9 \pm 45,6$ 2,4-150	$134,0 \pm 11,1$ 101,3-148,9	$49,8 \pm 9,1$ 29,8-64	$8,5 \pm 1,7$ 5,7-11,5	$3,6 \pm 0,43$ 3-4	$7,1 \pm 1,1$ 5,4-8,9	$3,4 \pm 0,77$ 2,4-4,9	$3,3 \pm 0,8$ 2,4-4,8
wARS	$2,2 \pm 8,45E-01$ 1,1-4,6	$2,1 \pm 0,2$ 1,9-2,8	$1,5 \pm 0,32$ 1,1-2,4	$3,2 \pm 0,6$ 2,3-4,6	$1,4 \pm 0,15$ 1,3-1,6	$1,4 \pm 0,22$ 1,1-1,9	$2,8 \pm 0,63$ 1,8-3,8	$2,8 \pm 0,7$ 1,9-3,8

TABLE II
OUTLIER DETECTION VS. SIMILARITY MEASURES.

	ClusterN	BCS	BRL	U	BRS	ARS	wARS	MDI
p	Ns	Ns	Ns	Ns	Ns	Ns	Ns	Ns
Gower	$16,64 \pm 14,14$	$157,23 \pm 147,43$	$35,68 \pm 51,20$	$8,05 \pm 9,92$	$36,38 \pm 52,26$	$30,72 \pm 47,12$	$2,22 \pm 0,91$	$3,61 \pm 10,38$
SMC	$16,43 \pm 14,30$	$155,63 \pm 143,62$	$35,43 \pm 50,78$	$5,57 \pm 6,72$	$36,32 \pm 52,43$	$30,20 \pm 45,56$	$2,11 \pm 0,75$	$3,26 \pm 10,51$
Jaccard	$16,50 \pm 14,24$	$145,16 \pm 141,09$	$35,36 \pm 50,69$	$5,21 \pm 7,71$	$36,59 \pm 52,77$	$30,49 \pm 46,66$	$2,06 \pm 0,75$	$2,95 \pm 9,60$
Total	$16,52 \pm 14,14$	$152,67 \pm 143,31$	$35,49 \pm 50,58$	$6,28 \pm 8,27$	$36,43 \pm 52,17$	$30,47 \pm 46,17$	$2,13 \pm 0,81$	$3,27 \pm 10,12$

TABLE III
OUTLIER DETECTION VS. CLUSTERING METHODS.

	ClusterN	BCS	BRL	U	BRS	ARS	wARS	MDI
p	Ns	$< 0,05$	Ns	$< 0,05$	Ns	Ns	$< 0,05$	Ns
SL	$16,48 \pm 14,30$	$211,45 \pm 168,03$	$35,33 \pm 50,77$	$11,64 \pm 11,41$	$35,95 \pm 52,12$	$29,29 \pm 46,39$	$2,48 \pm 1,00$	$2,98 \pm 9,07$
CoL	$16,48 \pm 14,30$	$121,21 \pm 126,66$	$35,52 \pm 50,99$	$2,38 \pm 3,33$	$36,36 \pm 51,81$	$31,77 \pm 45,93$	$1,80 \pm 0,51$	$2,83 \pm 8,08$
AL	$16,52 \pm 14,26$	$142,95 \pm 134,26$	$35,60 \pm 51,25$	$3,48 \pm 3,34$	$36,50 \pm 52,79$	$31,19 \pm 46,89$	$2,00 \pm 0,69$	$3,30 \pm 9,89$
CL	$16,62 \pm 14,21$	$135,07 \pm 127,99$	$35,50 \pm 51,16$	$7,62 \pm 8,45$	$36,90 \pm 53,84$	$29,63 \pm 47,10$	$2,24 \pm 0,81$	$3,99 \pm 13,06$
Total	$16,52 \pm 14,14$	$152,67 \pm 143,31$	$35,49 \pm 50,58$	$6,28 \pm 8,27$	$36,43 \pm 52,17$	$30,47 \pm 46,17$	$2,13 \pm 0,81$	$3,27 \pm 10,12$

of their influence on the number of clusters, the number of ungrouped rules (outliers) and other parameters. The only conclusion we can give is the following: using SMC and $Jaccard$ measures gives similar results, while using $Gower$ measure gives us a higher number of outliers (U), and a slightly greater biggest cluster's size (BCS).

Table III, confirms the existance of statistically significant differences between using the analyzed clustering methods (SL, CL, CoL, AL in the context of their influence on the examined parameters: U, BCS and $wARS$. At the significance level of $p < 0,05$ it is possible to conclude that the SL method brings the biggest BCS, U and $wARS$ in comparison to the other clustering methods, while the CoL method produced the smallest. It confirms all the features these methods have as it can be widely found in the literature. Thus to get higher number of outliers we should choose SL instead of the other methods.

Figures 1, 2 and 3 confirm all the remarks made in the results of the experiments. The idea of CoL method develops the smallest size of the biggest cluster, and thus a smaller average number of descriptors included in the representative of each cluster ($wARS$). Additionally it results in the smallest number of ungrouped rules. Therefore the U values are the smallest when using the CoL method.

VI. SUMMARY

This article presents the evaluation of three different similarity measures and four clustering methods used for comparing the results of clustering of rules in KBs. The experiments have been carried out for seven different KBs from different domains and such datasets tend to differ in many parameters. Rules have been clustered using the AHC algorithm presented in Section III. The results taken from the experiments have been compared by the following parameters: the number of clusters, the number of ungrouped rules, and the parameters related to the cluster's representatives. They are included in

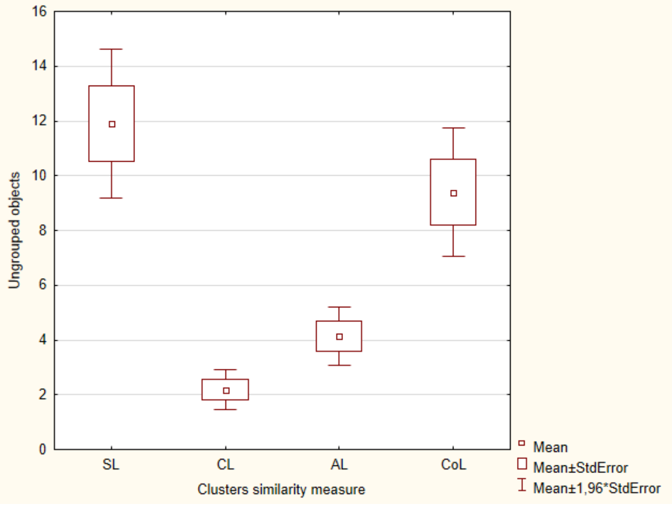

Fig. 1. Ungrouped rules number vs. clustering methods.

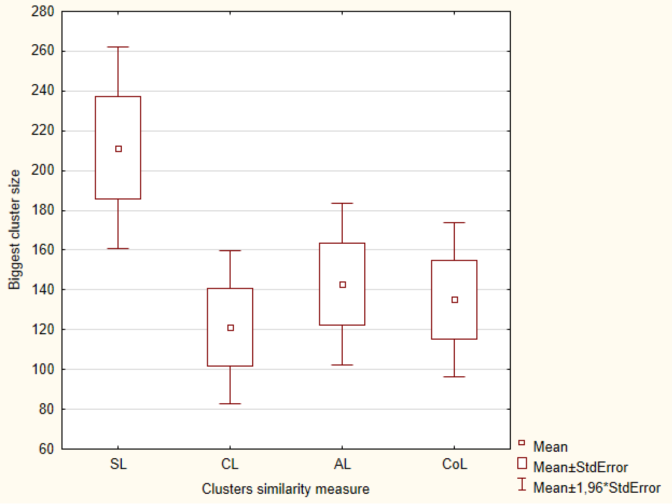

Fig. 2. Biggest cluster's size vs. clustering methods.

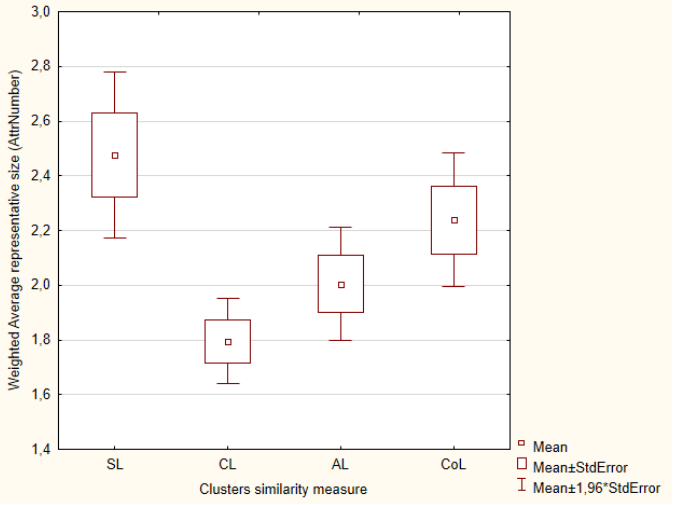

Fig. 3. Weighted average representative size vs. clustering methods.

Tables I,II, III and in Figures 1,2,3. The experiments show that the examined similarity measures produce comparable values of the analyzed parameters. The selection of clustering method to be used is crucial and there is a strong correlation between using a given clustering method and the value of the following parameters: the biggest clusters size, the number of outliers in rules as well as the weighted average representative's size. As it can be found in the literature, the SL results in achieving a higher number of outliers, while the CoL has a tendency to give the smallest number of outliers. In future research the author plans to examine other, much varied, similarity measures, and check if the size of the input data: the number of rules, the length of the rules, the type of the attribute used to describe rules in KBs have got any influence on the efficiency of the outlier mining process. There has been no evaluation of the outlier detection algorithm as there was no apriori knowledge provided by domain experts - such knowledge would have facilitated the labelling of the rules as outliers or normal data with the use of popular measures such as True/False Positive/Negative.

References

[1] Bazan J.G., Szczuka M.S., Wróblewski J. A new version of rough set exploration system. Rough Sets and Current Trends in Computing, Springer-Verlag, Berlin, pp. 397 404, 2002.

[2] Boriah S., Chandola V., Kumar V. Similarity Measures for Categorical Data: A Comparative Evaluation, Proceedings of the 8th SIAM International Conference on Data Mining, pp. 243-254, 2008.

[3] Dubes R., Jain A.K. Clustering techniques: The user's dilemma, Pattern Recognition, vol. 8, nr 4, 1976.

[4] Duraj A., Szczepaniak P., Ochelska-Mierzejewska J. Detection of Outlier Information Using Linguistic Summarization, Flexible Query Answering Systems 2015; Advances in Intelligent Systems and Computing 400, (Eds.: Andreasen T., et al.), Proceedings of the 11th International Conference FQAS 2015, Cracow, Poland; Springer 2016, pp.101–113.

[5] Goodall D.W. A new similarity index based on probability, Biometrics, vol.22, pp. 882-907, 1966.

[6] Gower J.C. A general coefficient of similarity and some of its properties, Biometrics, vol.27, International Biometric Society, Washington, 1971,pp.857–871

[7] Hawkins D. M. Identification of Outliers, Monographs on Applied Probability and Statistics, ISBN: 978-94-015-3996-8 (Print) 978-94-015-3994-4 (Online), 1980.

[8] Lee O., Gray P. Knowledge base clustering for KBS maintenance, Journal of Software Maintenance and Evolution, vol.10, nr 6, pp. 395-414, 1998.

[9] Lichman M. UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. University of California, 2013.

[10] Nowak-Brzeziska A. Mining rule-based knowledge bases, Advanced Technologies for Data Mining and Knowledge Discovery, CCIS, Springer, Volume 613, pp. 94-108, 2016.

[11] Nowak-Brzezińska A., Rybotycki T. Visualization of medical rule-based knowledge bases, Journal of Medical Informatics & Technologies, Vol.24, pp. 91-98, 2015.

[12] Nowak-Brzezińska A. Mining Rule-based Knowledge Bases Inspired by Rough Set Theory, Fundamenta Informaticae 148, pp. 3550, 35 DOI 10.3233/FI-2016-1421, IOS Press, 2016.

[13] Ramaswamy S., Rastogi R., Shim K. Efficient algorithms for mining outliers from large data sets, Proceedings of the 2000 ACM SIGMOD international conference on Management of data - SIGMOD, doi:10.1145/342009.335437. ISBN 1581132174, 2000, pp. 427.

[14] Wierzchoń S. T., Kłopotek M.A. Algorithms of Cluster Analysis, Wydawnictwo IPI PAN, 2015, Warszawa