Outliers mining in knowledge based systems

Agnieszka Nowak-Brzezińska

Institute of Computer Science, Silesian University, Poland

The aim of the research is to develop methods for mining complex knowledge bases (KBs) [1], especially in the context of discovering outliers in rules [2, 3]. Such data sets need efficient techniques (i.e. based on cluster analysis) for their exploration. Finding unusual and influential rules is crucial for exploring a given KB thus the article proposes an $AHCOB$ [2] algorithm dedicated to detecting the rare and interesting rules in complex KBs.

The last decade resulted in significant development of data mining methods, tools, and applications. After the rule sets are induced from data (i.e. using rough set theory) they can be applied in several ways[4]. Unfortunately, in many areas, the incomprehensible results of the data mining process become a problem. This situation often gets even more complicated because of the number of generated rules which is typically very high. Hundreds or even thousands of rules are a common result of the data mining process applied to real-world data sets. One of the basic problems of data mining is the outlier detection. An *outlier*[1] is an observation in data that deviates from other observations so much that it arouses suspicions of being generated by a different mechanism. In recent years, it has been applied in practical ways, i.e. to detect unusual signs of disease, unauthorized tampering to servers or to find new articles on the given subject. Many data mining methods will not allow for useful analysis of data without previous detection of outliers or mistakes. In literature following methods of finding outliers are well known: distribution-based, distance-based, density-based and clustering-based. Cluster analysis enables dividing all rules into smaller groups in which the rules are the most similar (in their conditional and/or decision part). The most desired result is a set or rules clusters with a good quality. Unfortunately, it depends mostly on the data distribution and the type of the clustering algorithm (*agglomerative hierarchical algorithm* AHC or its modification $mAHC$)[5].

In the context of KBs, the *outlier* is the rule, which is unusual. Therefore, it is difficult to cluster with other rules, because it would be a rule that conflicts with other rules in the KB. Unusual rules affect significantly the quality of the cluster or the form of the cluster's representative. Thus it is worth to find such rules before the clustering procedure because it improves the quality of clusters and makes the exploration of the unusual cases more detailed. Unusual does not mean erroneous. It is very important to establish different conditions to

[1] A complex KB consists of a large number of rules. Such data sets have hierarchical relationships between rules (the premise of one rule may be the conclusion of another rule). The inference processes for these sets would be time-consuming and complicated to interpret.

[2] Agglomerative Hierarchical Clustering Outlier Based

decide what parameter should be examined in order to get the best clustering results. It can be said that typical rules should be included in some cluster whereas unusual cases do not belong to any cluster. A typical rule lies close to the nearest centroid whereas an unusual one lies far from it (the nearest centroid). Typical rules belong to large clusters whereas unusual rules to the very small clusters. Unusual rules in KBs can be a subject of a more in-depth research in the field, especially by experts, in order to expand KBs as much as possible. Such rare rules may represent exceptional and specific cases or may be just the result of a modification a KB which was thoughtless. In case of finding outliers as the result of clustering we may say that a single object or a small cluster is a potential outlier. The $AHCOB$ algorithm detects unusual rules in three steps: (a) detection of conflicting rules takes place[3], (b) finding small clusters [4], (c) detection so-called *influential* rules. A rule is *influential* for a group if the quality of its cluster changes significantly. Rules that are influential should not be the representatives of created rules' clusters. During the experiments, small KBs were used (23, 29 and 49 rules), mainly because such set of rules are easy to analyse. For each KB, the full $AHCOB$ process was performed[6]. It should be emphasized that after using the $AHCOB$ algorithm the domain expert should decide if they will be further isolated or still included in the created clusters of rules. Furthermore, the efficiency of a given complex decision support system improves because the time of the inference process is reduced, the number of rules analysed in this process is smaller, and the quality of rules clusters is higher.

Ack. This work is a part of the project ,,Exploration of rule knowledge bases" founded by the Polish National Science Centre (NCN: 2011/03/D/ST6/03027).

References

1. Cherednichenko S.: Outlier Detection in Clustering, Master's Thesis, University of Joensuu, Department of Computer Science, (2005).
2. Jackson P.: Introduction to Expert Systems. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, (1998).
3. Grzymała-Busse J.W.: Managing Uncertainty in Expert Systems, Springer Science & Business Media, Vol. 143, (1991).
4. Skowron A., Komorowski J., Pawlak Z., Polkowski L.: Rough Sets Perspective on Data and Knowledge, Handbook of Data Mining and Knowledge Discovery, Oxford University Press, Inc., New York, USA, (134-149), (2002).
5. Nowak-Brzezińska A., Jach T., Simiński R., Xieski T.: Towards a practical approach to discover internal dependencies in rule-based knowledge bases, LNCS, vol.6954, Springer, 232-237, (2011).

[3] Rules are conflicted if the same conditional parts of these rules do not result in the same conclusion.

[4] It has to be the rule, which is so different from other rules that it is impossible to cluster it with them. There are many solutions to define the minimal size of a cluster[5]

[6] As a result, groups of unusual rules were achieved: conflicted rules, small cluster and influential rules. Comparing the information given by the domain expert with the set of unusual rules (that were found by the $AHCOB$ algorithm) enables measuring the quality of the proposed algorithm using the measures of *sensitivity* and *specificity.*